UNIVERSITY OF FORT HARE

**STM 313**

FINAL EXAM

JUNE 2023

.....

SUBJECT : INTRODUCTORY APPLIED STATISTICS B1

MARKS : 100

**TIME:** 3 hours


Internal Examiners

**Mr. L. Kondlo**

**Dr. A.S. Odeyemi**

External Examiners

**Ms. J. Batidzirai**


**Instructions**

Show all your working including R programming codes and outputs

# QUESTION ONE [20 Marks]

Please choose the correct answer.

1.1 A ..... is a set of elements appearing in rows and columns where the elements are of the same mode whether they are logical, numeric (integer or double), complex or character. [1]

A. Vector
B. Matrix
C. List
D. Data Frames

1.2 The four most frequently used types of data objects in R are vectors, matrices, data frames and..... [1]

A. Function
B. Lists
C. Packages
D. Interfaces

1.3 What is the simplest way of creating the vector? [1]

A. C-function
B. Create
C. Destroy
D. Invalid

1.4 Which function replicates elements of vectors? [1]

A. C
B. Rep
C. Crep
D. Grep

1.5 The ..... function creates a regular sequence of values to form a vector. [1]

A. sequel
B. Rep
C. Seq
D. Grep

1.6 Which function is used to enter in data at the terminal? [1]

A. Scanned
B. Scnn
C. Scan
D. Sccn

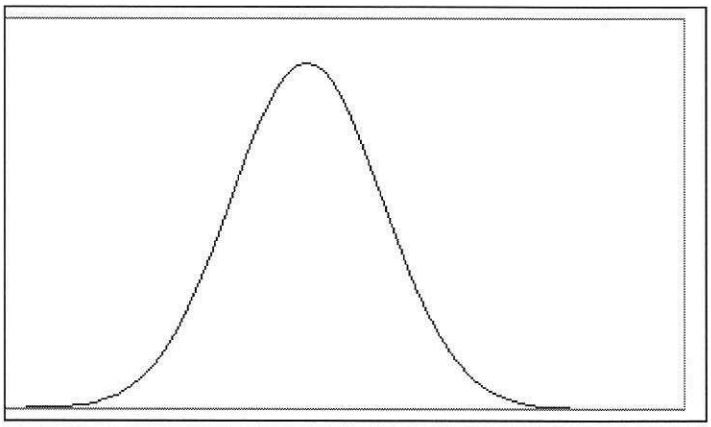
1.7 Computation with vectors is achieved using an element-by-element operation for avoiding_____ [1]

A. Loops
B. Functions
C. Packages
D. Interfaces

1.8 To bind a row onto an already existing matrix, the _____ function can be used. [1]

A. Rbind
B. Sbnd
C. Gbind
D. Sbind

1.9 The distribution of SBP in men, 20-29 years is shown below.



I. What is the best summary of a typical value [1]

A. Mean
B. Median
C. Interquartile range
D. Standard Deviation

II. The best summary of variability for this distribution is [1]

A. Mean
B. Median
C. Interquartile range
D. Standard Deviation

1.10 An RCT is planned to show the efficacy of a new drug (N) vs. placebo (P) to lower total cholesterol. What are the hypotheses? [1]

A. $H_0: m_N = m_P, H_A: m_N > m_P$

B. $H_0: m_N = m_P, H_A: m_N < m_P$

C. $H_0: m_N = m_P, H_A: m_N \neq m_P$

1.11 The null value of a difference in means is… [1]

A. 0

B. 0.5

C. 1

D. 2

1.12 The null value of an odds ratio is… [1]

A. 0

B. 0.5

C. 1

D. 2

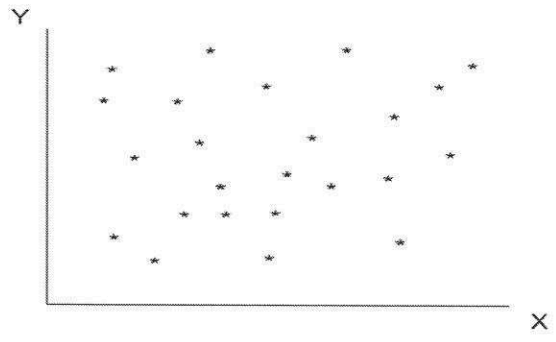
1.13 A two-sided test for the equality of means produces p=0.20. Reject H0? [1]

A. Yes

B. No

C. Maybe

1.14 Correlation (r)– measures the nature and strength of linear association between two variables at a time. What is the most likely value of r for the data shown below? [1]




A. r=-0.5

B. r=0

C. r=0.5

D. r=1

1.15 In Framingham Heart Study, we want to assess risk factors for Impaired Glucose. The outcome is Glucose and is grouped into 4 categories (Diabetes (glucose > 126), Impaired Fasting, Glucose (glucose 100-125), and Normal Glucose). The associated Risk Factors are Sex (male, female), Age (years), and BMI (normal weight, overweight, obese)

I. What kind of variable is Glucose? [1]

A. Binary,
B. Nominal,
C. Ordinal
D. Continuous

II. What statistical test would be used to assess whether age is associated with Glucose Category? [1]

A. ANOVA
B. Chi-Square GOF
C. Chi-Square test of independence
D. Test for equality of means
E. Other

1.16 Consider a Tertiary Outcome Diabetes Status (Diabetes / No Diabetes). The associated Risk Factors are Sex (male, female), Age (years), and BMI (normal weight, overweight, obese).

I. What test would be used to assess whether AGE is associated with Diabetes? [1]

A. ANOVA
B. Chi-Square GOF
C. Chi-Square test of independence
D. Test for equality of means
E. Other

II. What test would be used to assess whether BMI is associated with Diabetes? [1]

A. ANOVA
B. Chi-Square GOF
C. Chi-Square test of independence
D. Test for equality of means

E. Other

1.17 We want to study whether individuals over 45 years are at greater risk of diabetes than those younger than 45. What kind of variable is age? [1]

A. Binary
B. Nominal
C. Ordinal
D. Continuous

# QUESTION TWO [20 Marks]

2.1 Create a vector which is a sequence of 20 to 30, repeating 5 times [2]

2.2 Write a R program to create a sequence of numbers from 20 to 50 and find the mean and sum of the sequence. [3]

2.3 Write a R program to create a vector which contains 10 random integer values between -50 and +50. [2]

2.4 Write a R program to extract first 10 english letter in lower case and last 10 letters in upper case and extract letters between 22nd to 24th letters in upper case. Note: Use built-in datasets letters and LETTERS. [3]

2.5 Write a R program to create three vectors a,b,c with 3 integers. Combine the three vectors to become a 3×3 matrix where each column represents a vector. Print the content of the matrix. [2]

2.6 Write an R program to create the system's idea of the current date with and without time. [2]

2.7 Write an R program to create Data frames containing details of 5 students and display the data summary. [3]

| Student ID | Name | Gender | Age | PASS |
|---|---|---|---|---|
| 201716459 | Mkhontwana Y | Female | 19 | TRUE |
| 201819299 | Mtakazi S | Male | 17 | FALSE |
| 201907842 | Gomo S | Male | 20 | TRUE |
| 202016069 | Pezisa S | Female | 20 | TRUE |
| 202016676 | Mcoki Y | Male | 21 | FALSE |

2.8 Write a R program to create bell curve of a 1000 random normal distribution with mean equal to 50 and standard deviation of 10 and count occurrences of each value. [3]

# QUESTION THREE [20 Marks]

You work for a wastewater treatment plant and set up an experiment to understand whether a new bioremediation technique reduces the concentration of pollutants in the water. You set up 16 water tanks and filled each tank with sewage water. You add the bioremediation agent to eight randomly

selected tanks; the other eight are control. After a week, you measure the *E. coli* concentrations (measured as colony counts per 100ml) in the water tanks and obtain the following data:

| Replicate | Treatment | Ecolicounts |
|---|---|---|
| 1 | Control | 1840 |
| 2 | Control | 14981 |
| 3 | Biocontrol | 7 |
| 4 | Control | 26058 |
| 5 | Control | 331 |
| 6 | Biocontrol | 49 |
| 7 | Control | 2342 |
| 8 | Biocontrol | 79 |
| 9 | Biocontrol | 174 |
| 10 | Biocontrol | 833 |
| 11 | Control | 5 |
| 12 | Biocontrol | 149 |
| 13 | Control | 16 |
| 14 | Control | 691 |
| 15 | Biocontrol | 4 |
| 16 | Biocontrol | 11636 |

Use these data to complete the following tasks

3.1 Explain which are the *dependent* and *independent* variables [4]

3.2 Describe which statistical analysis would be appropriate to analyse these data and why? [4]

3.3 Read the data into R and perform the appropriate statistical test [5]

3.4 Report and explain the statistical outputs [4]

3.5 Present the data in an informative figure [3]

# QUESTION FOUR [20 Marks]

Use the data provided below to answer the questions:

| A | B | C | D | E |
|---|---|---|---|---|
| 0.40 | 0.26 | 0.24 | 1.04 | 0.74 |
| 1.50 | 0.47 | 0.25 | 2.78 | 0.99 |
| 0.98 | 0.42 | 1.01 | 0.82 | 1.26 |
| 0.33 | 0.64 | 0.77 | 1.65 | 1.50 |
| 0.75 | 0.32 | 0.47 | 0.49 | 0.30 |
| 1.48 | 0.65 | 0.47 | 0.97 | 0.34 |
| 1.18 | 0.43 | 0.46 | 1.39 | 0.77 |
| 0.33 | 0.67 | 0.65 | 3.24 | 1.94 |
| 1.42 | 0.43 | 0.41 | 1.12 | 2.62 |
| 2.09 | 0.70 | 0.81 | 2.82 | 1.42 |
| 1.37 | 0.79 | 1.20 | 1.27 | 0.73 |
| 1.23 | 0.89 | 1.08 | 1.60 | 2.09 |
| . | . | 0.34 | 1.98 | 1.52 |
| . | . | 1.98 | 9.32 | 1.67 |
| . | . | 1.39 | 2.31 | 3.40 |
| . | . | 1.12 | 4.19 | 2.16 |
| . | . | 3.14 | 1.73 | 2.31 |
| . | . | 2.78 | 5.16 | 1.32 |

**Question:** Difference in protein expression between 5 cell types?

4.1 Explain which are *dependent* and *independent* variables [2]

4.2 Restructure the file: wide to long

- Clue: use **melt()** ## reshape2 ## [2]

4.3 Rename the columns: "line" and "expression"

- Clue: use **colnames()** [2]

4.4 Remove the NAs

- Clue: use na.omit [2]

4.5 Plot the data using at least one type of graph: **stripchart, boxplot, beanplot, etc.** [2]

4.6 Check ALL the assumptions for the parametric test [3]

4.7 Perform ANOVA tests and apply ANY **post hoc** tests. [5]

4.8 State your null hypothesis and report your output results in details. [2]

## QUESTION FIVE [20 Marks]

Carina obtains cash from an ATM (cash machine). She suspects that the rate at which she spends cash is affected by the amount of cash she withdrew at her previous visit to an ATM. To investigate this, she deliberately varies the amounts she withdraws. She records, for each visit to an ATM, the amount, R$x$, withdrawn, and the number of hours, y, until her next visit to an ATM. Use the dataset provided to answer the following questions:

| withdrawal | amount | hours |
|---|---|---|
| 1 | 40 | 56 |
| 2 | 10 | 62 |
| 3 | 100 | 195 |
| 4 | 110 | 330 |
| 5 | 120 | 94 |
| 6 | 150 | 270 |
| 7 | 20 | 48 |
| 8 | 90 | 196 |
| 9 | 80 | 214 |
| 10 | 130 | 286 |

5.1 Explain which are the dependent and independent variables [4]

5.2 Describe which statistical analysis would be appropriate to analyse these data and why? [4]

5.3 Read the data into R and perform the appropriate statistical test [5]

5.4 Report and explain the statistical outputs [4]

5.5 Present the data in an informative figure [3]