CHUKA  **UNIVERSITY**

UNIVERSITY EXAMINATIONS

SECOND YEAR EXAMINATION FOR THE AWARD OF DEGREE OF BACHELOR OF SCIENCE IN PHYSICS

GPHY 251: STATISTICAL DATA ANALYSIS

STREAMS: GPHY **TIME: 2 HOURS**

DAY/DATE: FRIDAY 12/4/2024 **2.30P.M. – 4.30 P.M.**

INSTRUCTIONS: Answer question ONE and any other TWO questions

QUESTION ONE (30marks)

a. Give reasons why we use exploratory data analysis (EDA) [2marks]
b. Distinguish between statistics and parameters [2marks]
c. What is a random sampling error? [1mark]
d. Explain what we mean by significance level in hypothesis testing [1 mark]
e. What is a Gaussian distribution in statistics? [1mark]
f. Define an outlier in data analysis [1mark]
g. Explain the meaning of dependent, independent and control variables as used in statistical data analysis [3marks]
h. Discuss any two types of experimental designs. [4marks]
i. Explain two aspects to be considered when creating a test group. [2marks]
j. State the central limit theorem [1marks]
k. list any two statistical checks for a normal distribution [2marks]
l. Explain the importance of inferential statistics [2 marks]
m. In a survey of 100 people who had recently purchased motorcycles, data on the following variables was recorded: Gender of purchaser, Brand of motorcycle purchased,

Number of previous motorcycles owned by purchaser, Telephone area code of purchaser, Weight of motorcycle as equipped at purchase

i. Which of these variables are categorical? [1mark]

ii. Which of these variables are discrete numerical? [1mark]

iii. Which type of graphical display would be an appropriate choice for summarizing the gender data [1 mark]

iv. Which type of graphical display would be an appropriate choice for summarizing the weight data. [1mark]

h. Suppose you have the following data where, n = 10,

1.2, 1.5, 2.6, 3.8, 2.4, 1.9, 3.5, 2.5, 2.4, 3.0.

Determine;

i. Median [1 mark]

ii. Mode [1mark]

iii. Interquartile range [2marks]

QUESTION TWO (20 marks)

a. Consider the following n = 20 scores of the written part of a driving licence examination maximum of 100 points could be achieved:28, 35, 42, 90, 70, 56, 75, 66, 30, 89, 75, 64, 81, 69, 55, 83, 72, 68, 73, 16. summarize the scores in class intervals such as 0–20, 21–40, 41–60, 61–80, and 81-100,

i. make a table indicating the frequency, relative frequency and percentage frequency of the scores [5 marks]

ii. Construct a corresponding relative frequency histogram for the scores.

iii. Describe the shape of the histogram? [2marks]

b. Given the following sorted data:1.2, 1.5, 1.9, 2.4, 2.4, 2.5, 2.6, 3.0, 3.5, 3.8,

Compute an estimate for;

i. Mean [2 marks]

ii. Variance [5marks]

iii. Standard deviation [3marks]

QUESTION THREE (20marks)

a. Distinguish between probability and non-probability sampling, giving an example of each. [4 marks]

b. Define a hypothesis [2 marks]

c. Given the research question, "Do female and male grade point averages prior to entering a course differ significantly? Construct a null and alternative hypothesis. [4 marks]

d. Discuss any two data requirements for regression analysis. [4marks]

e. Explain what is meant by correlation in statistical data analysis. [2 marks]

f. Giving examples distinguish between the parametric and non-parametric tests [4marks]

QUESTION FOUR (20marks)

a. You are provided with the following data set: 23, 42, 12, 10, 15, 14, 9. Find

i. The maximum, minimum, median, first quartile, third quartile, and IQR [5 marks]

ii. Draw a box plot for the data. [5marks]

b. The results of 41 students' math tests (with a best possible score of 70) are recorded below:

31, 49, 19, 62, 50, 24, 45, 23, 51, 32, 48, 55, 60, 40, 35, 54, 26, 57, 37, 43, 65, 50, 55, 18, 53, 41, 50, 34, 67, 56, 44, 4, 54, 57, 39, 52, 45, 35, 51, 63, 42

i. Is the variable discrete or continuous? Explain. [2 marks]

ii. Prepare an ordered stem and leaf plot for the data and briefly describe what it shows. [3 marks]

iii. Are there any outliers? If so, which scores? [2 marks]

iv. Look at the stem and leaf plot from the side. Describe the distribution's main features such as:

a) number of peaks [1 mark]

b) symmetry [1 mark]

c) value at the centre of the distribution [1 mark

QUESTION FIVE (20Marks)

a. The following data on x = score on a measure of test anxiety and y = exam score for a sample of $n = 9$ students are consistent with summary quantities given in the paper "Effects of Humour on Test Anxiety and Performance"

x	23	14	14	0	17	20	20	15	21
y	43	59	48	77	50	52	46	51	51

Higher values for x indicate higher levels of anxiety.

i. Prepare a scatter plot for these data. [5marks]

ii. What does the scatter plot show about the relationship between x and y. [2 marks]

iii. Compute the value of correlation coefficient between x and y. Is the value of r consistent with your answer in (ii) above? [5 marks]

iv. Is it reasonable to conclude that test anxiety caused poor exam performance? Explain. [2marks]

v. Determine the regression line associating exam score with the test anxiety [4 marks]

vi. Use the equation to predict the exam score when the measure of test anxiety is 10. [2marks]

……