

MACHAKOS UNIVERSITY
FACULTY OF PURE AND APPLIED SCIENCES
BACHELOR OF SCIENCE IN PROGRAMMING AND STATISTICS
THIRD YEAR FIRST SEMESTER
SST 301: PROGRAMMING FOR STATISTICS I
SUPPLEMENTARY EXAMINATION

DATE:

TIME: 2 Hours

INSTRUCTIONS: Answer question one and any other two

Question 1 (30 Marks)

- a) A simple random sample of 70 coffee drinkers was taken. Each sampled coffee drinker was asked to taste two unmarked cups of coffee, one of which is actually **Brand A** and the other is **Brand B**, and was asked which one they *preferred*. 44 coffee drinkers preferred Brand A, and the other 26 preferred Brand B.
- i. Does a majority of all coffee-drinkers (that is, more than 50% of them) prefer brand A? Carry out a suitable test of significance to assess the evidence. What is the P-value of your test? **(5 Marks)**
- ii. It turned out that the coffee drinkers had always been given an unmarked cup containing Brand A coffee first, and Brand B coffee second.
How do you react to this knowledge? **(4 Marks)**
- b) A *report* states that 3% of all births are “multiple births” (that is, twins, triplets, etc.). A random sample of 5000 births is taken. Assuming the report to be correct, what is the **probability** that 2.5% or less of the births in the sample are multiple births? **(4 Marks)**
- c) Using relevant examples, differentiate between the following **types** of *research*
- i. Exploratory research **(3 Marks)**
- ii. Descriptive research **(3 Marks)**
- iii. Explanatory research **(3 Marks)**
- iv. Predictive research **(3 Marks)**
- d) A doctor collected data on how many hours **children** of different ages *slept* in a day. Data for 5 participants are shown below.

Age(years)	1	2	3	5	8
------------	---	---	---	---	---

Sleep(hours)	13	11	10	9	8
---------------------	----	----	----	---	---

- i. **Plot** the data in a scatter plot. **(1 Mark)**
- ii. **Describe** *clearly* and *concisely* what the scatter diagram indicates. **(2 Marks)**
- iii. **Discuss** whether or not it would be appropriate to calculate a *linear regression* for y on x for these data. **(2 Marks)**

Question 2 (20 Marks)

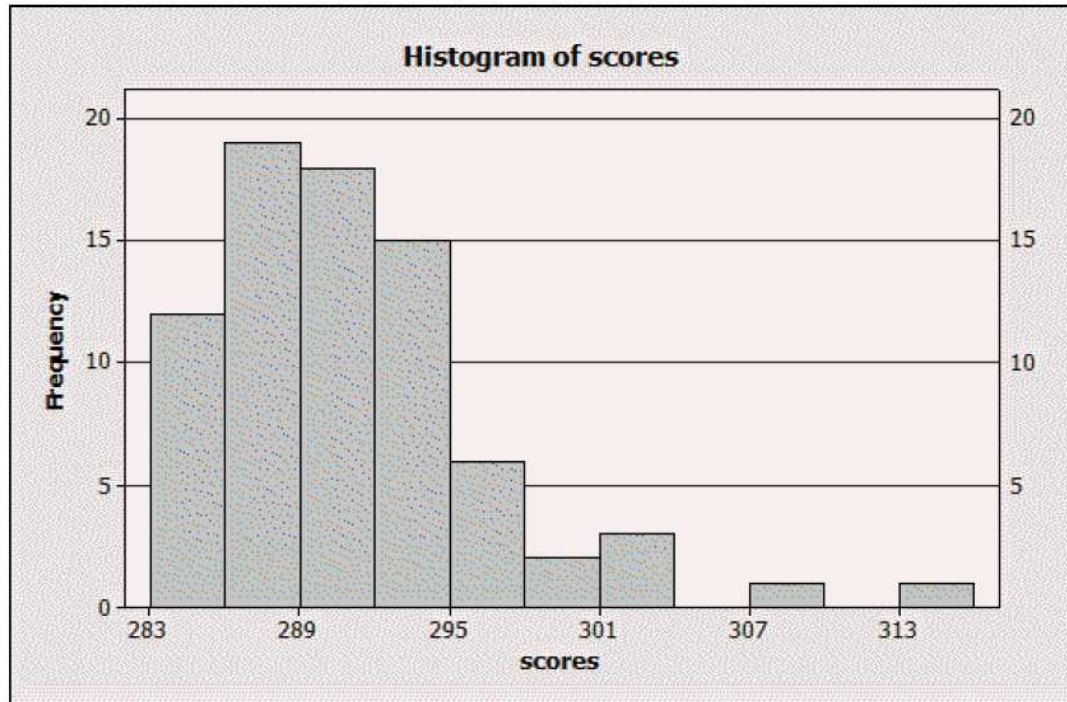
- a) A student's mark in a Java exam has a normal distribution with mean 65 and SD 10. The same student's mark in a Probability & Statistics exam has a normal distribution with mean 60 and SD 15.

The two marks are independent of each other.

- i. What is the probability that the student scores over 80 in both exams? **(4 Marks)**
 - ii. What is the probability that the student scores over 80 in exactly one of the two exams? **(4 Marks)**
 - iii. What is the probability that the student's mean mark in the two courses is over 80? **(4 Marks)**
- b) The correlation between two variables x and y is -0.7 .
Clearly explain the relationship between variable x and variable y. **(4 Marks)**
 - c) A University has 2000 male and 500 female faculty members. The Human Resource department is commissioning a survey on childcare and wants to sample 100 faculty members.
What kind of sampling method would best represent the possibly differing opinions of males and females? Justify your answer. **(4 Marks)**

Question 3 (20 Marks)

- a) The histogram below shows the scores of 77 players in the recently held *Lower Eastern golf tournament*.



- i. What is the median score for these golfers? **(2 Marks)**
- ii. How does the mean score compare with the median? **(3 Marks)**
- b) Using appropriate examples, explain the following terms as they relate to Statistical computing:
 - i. Correlation **(3 Marks)**
 - ii. Simple linear regression **(3 Marks)**
 - iii. Multiple regression **(3 Marks)**
 - iv. ANOVA **(3 Marks)**
 - v. Two-way ANOVA **(3 Marks)**

Question 4 (20 Marks)

- a) It is believed that applying zinc to soil can affect the amount of herbage (leaves and stems) of plants grown. In an experiment, the regression line for predicting the amount of herbage (y) from the amount of zinc applied(x) is $y = 3.29 - 3.825x$
 - i. What can you conclude from this regression line? **(3 Marks)**
 - ii. What is the predicted amount of herbage when the amount of zinc is 0.5? **(3 Marks)**
- b) Compare and contrast between Reliability and Validity as used in research. **(4 Marks)**

- c) Explain the research process. (2 Marks)
- d) Explain a scenario where generalizing research findings from a sample to the population is most likely to be justifiable (2 Marks)
- e) Discuss the following concepts:
 - i. Null hypothesis (2 Marks)
 - ii. Alternative hypothesis (2 Marks)
 - iii. Variables (2 Marks)

Question 5 (20 Marks)

- a) Investigators are interested in the relationship between levels of stress at work and self-reports of overall health. The researchers conduct a study with 102 participants and determines that the **Pearson correlation** coefficient between self-reported levels of job stress and self-reports of overall health is **+0.41**.
The investigators find that the mean score for stress is 14.3. The standard deviation for stress is 3.5. The mean score for health is 4.3. The standard deviation for stress is 1.3.
 - i. Are the investigators justified in saying that there is a significant relationship between these two variables? (4 Marks)
 - ii. Provide the regression equation to predict scores for health from scores for stress. (4 Marks)
 - iii. What is the regression equation to predict standard scores for health from standard scores for stress? (What are the slope and y-intercept for this equation when you're working in standard score units?) (4 Marks)
 - iv. Calculate the standard error of estimate for this example. What does this number tell you about the predicted scores for health? (4 Marks)
 - v. You know that three clients have scores of 9.1, 14.0, and 17.3, respectively, on the measure of stress. What are their predicted scores for health? (4 Marks)