



# MACHAKOS UNIVERSITY

University Examinations 2020/2021 Academic Year

SCHOOL OF PURE AND APPLIED SCIENCES

DEPARTMENT OF MATHEMATICS AND STATISTICS

THIRD YEAR SPECIAL /SUPPLEMENTARY EXAMINATION FOR  
BACHELOR OF SCIENCE IN STATISTICS AND PROGRAMMING

SMA 367: EXPLORORY DATA ANALYSIS

DATE: 23/3/2021

TIME: 2.00-4.00 PM

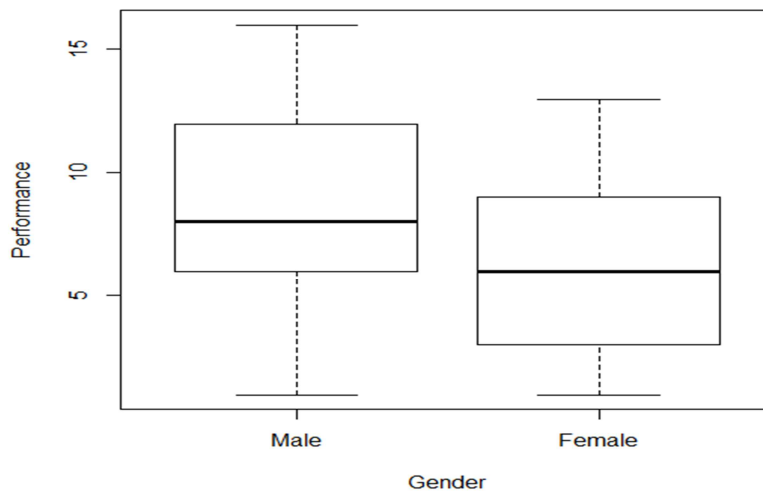
---

## INSTRUCTIONS TO CANDIDATES

Answer Question **ONE** which is compulsory and any other **TWO** Questions

### QUESTION ONE (30 MARKS)

- a) Distinguish the following terms as they apply in data analysis
- A variable and a case
  - Data mining and data screening
  - Descriptive statistics and inferential statistics
  - Parametric and Non-parametric test statistics (10 marks)
- b) A sample of 7 males and 13 females were assigned a task and their performance analyzed and summarized in figure below.



Using the figure estimate the; median and quartile deviation for each gender and commend on the gender with the higher dispersion. (6 marks)

- c) The following data represents the weight (KG) of newly born babies in Katu mission hospital, present it in a stem and leaf plot:

**23, 35, 25, 22, 25, 43, 27, 54, 28, 55, 32, 36, 36, 45, 50** (6 marks)

- d) Distinguish the following terms as they apply in exploratory data analysis

- i. Exploratory data analysis (EDA) and response surface methodology (RSM)
- ii. Saturated and parsimonious model
- iii. Confounded variable and noise variable
- iv. Fixed and random models

(8 marks)

## QUESTION TWO (20 MARKS)

- a) Distinguish between the terms

- i. prior and posterior probabilities (4 marks)
- ii. Show that the conjugate prior for poisson likelihood is gamma distribution.

(6 marks)

- b) Discuss five ways of handling a missing data

(10 marks)

**QUESTION THREE (20 MARKS)**

- a) Complete the two- way ANOVA table by inserting the right figures in the spaces labeled in letter  $p, q, r, s, t, u, v, w, x, y$  and  $z$  (10 marks)

**Two- way ANOVA**

| Source of Variation | Degrees of Freedom | Sum of Squares | Mean sum of Squares | F – value | F – table |
|---------------------|--------------------|----------------|---------------------|-----------|-----------|
| Factor A            | 3                  | 42             | $v$ -               | $t$       | $r$       |
| Factor B            | 2                  | $w$            | 16                  | $q$       | $s$       |
| AB                  | $x$                | $u$            | 22.67               | $z$       |           |
| Error               | $y$                | 210            | $p$                 |           |           |
| Total               | 21                 | 420            |                     |           |           |

Commend on the significance of the findings represented in the ANOVA table at  $\alpha=5\%$

(2 marks)

- b) Based on the recent observations 3% of students of huang university has coronavirus. The university clinic laboratory can detect 90% of the virus when present. It also gives 13% false positive results. If a student from the university volunteers for the virus test in the university clinic laboratory, determine the following probabilities;
- That the test result will be positive
  - That, given a positive result, he has the virus
  - That, given a negative result, he has no virus
  - That he was misclassified

(8 marks)

**QUESTION FOUR (20 MARKS)**

- a) Discuss four causes of asymmetry in data distribution structures (8 marks)
- b) Show that the conjugate prior for binomial likelihood is beta distribution (6 marks)
- c) A randomly sampled 13 students with a hearing impairment were given some test before and after being given a hearing aid. The test score for each student before and after were analyzed and the findings summarized in the output table below.

Table 1: Sample t-test output, for the test  $H_0 : \mu_A = \mu_B$

|            | t    | df | Sig.(2-tailed) | Mean difference | 95% CI of the difference |
|------------|------|----|----------------|-----------------|--------------------------|
|            |      |    |                |                 | Lower Upper              |
| Attendance | -0.5 | 1  | 0.6269         | 1.500           | -5.443 3.443             |
|            |      | 2  |                |                 |                          |

- State the hypothesis for the above scenario
- Based on the sample t-test output make statistical conclusion
- Interpret the 95% CI lower and upper difference values (6 marks)

#### QUESTION FIVE (20 MARKS)

- Highlight four methods of transforming a raw skewed data before analysis (4 marks)
- Discuss four non- parametric test statistics and highlight when appropriate to apply each one of them (8 marks)
- In an opinion survey regarding the BBI (Building Bridges Initiatives) 2500 individuals interviewed opined the initiative as summarized below.

| Age in Years | Opposed     | Undecided  | In support  | Total       |
|--------------|-------------|------------|-------------|-------------|
| Under 30     | 650         | 100        | 750         | 1500        |
| Over 30      | 350         | 400        | 250         | 1000        |
| <b>Total</b> | <b>1000</b> | <b>500</b> | <b>1000</b> | <b>2500</b> |

Test the hypothesis that age difference did no influence the opinion  $\alpha=5\%$  (8 marks)