



# MACHAKOS UNIVERSITY

University Examinations 2021/2022 Academic Year

SCHOOL OF PURE AND APPLIED SCIENCES

DEPARTMENT OF MATHEMATICS AND STATISTICS

THIRD YEAR SECOND SEMESTER EXAMINATION FOR

BACHELOR OF SCIENCE (MATHEMATICS)

SMA 367: EXPLORATORY DATA ANALYSIS

DATE: 6/12/2021

TIME: 8:30 – 10:30 AM

## INSTRUCTION:

*Answer Question One 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) The lecturer presumed that the average class attendance was 65 students. He did call register for 10 days and observed following: 66, 65, 69, 70, 69, 71, 70, 63, 64 and 68. The t-test output was as given

| The sample t-test output, for the test value=65 |       |    |                    |                    |                             |        |
|-------------------------------------------------|-------|----|--------------------|--------------------|-----------------------------|--------|
|                                                 | t     | df | Sig.<br>(2-tailed) | Mean<br>difference | 95% CI of the<br>difference |        |
|                                                 |       |    |                    |                    | Lower                       | Upper  |
| Attendance                                      | 2.825 | 9  | 0.020              | 2.500              | 0.4979                      | 4.5021 |

- i. State the hypothesis for the above scenario
  - ii. Based on the t-value and the sig.(2-tailed) value make statistical conclusion
  - iii. Interpret the 95% CI lower and upper difference values (7 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 (4 marks)
- d) Outline three main components captured when presenting the research findings (3 marks)
- e) 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 (6 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  $r, s, t, q, v, w, x$  and  $y$  (8 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                  | $x$                | 136            | 22.67               |           |           |
| Error               | $y$                | 210            |                     |           |           |
| Total               | 21                 | 325            |                     |           |           |

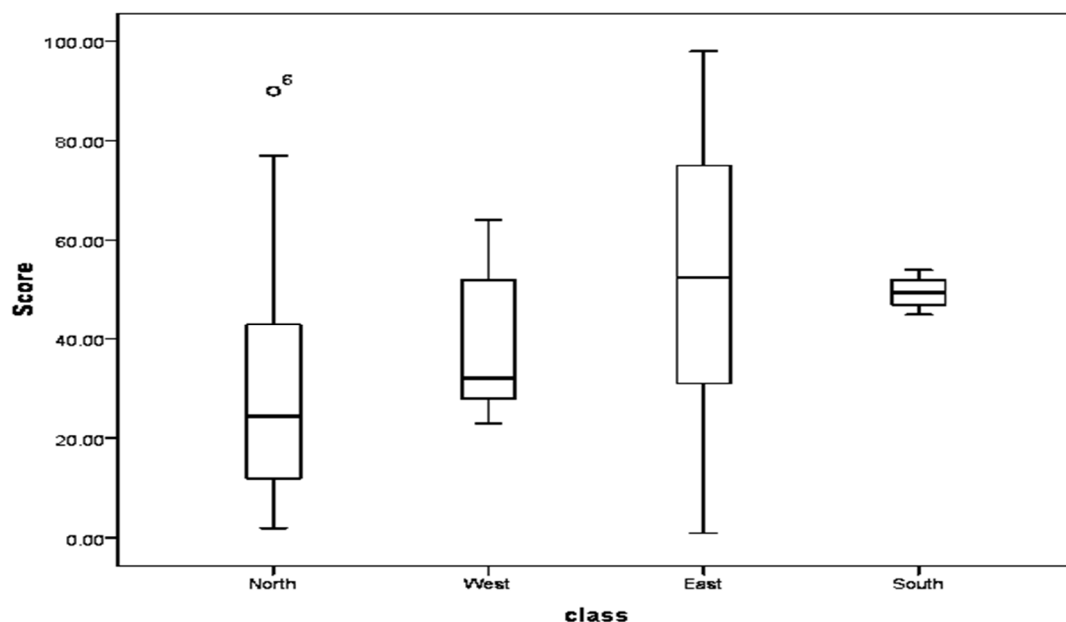
Commend on the significance of the findings represented in the ANOVA table at  $\alpha=5\%$

(2 marks)

- b) Based on the recent observations 3% of the MU male post graduate students have prostate cancer. The university clinic laboratory can detect 90% of the prostate cancer when present. It also gives 13% false positive results. A male post graduate student in MU volunteers for a prostate cancer test in the university clinic laboratory, determine the following probabilities;
- That the test result will be positive
  - That, given a positive result, he has a prostate cancer
  - That, given a negative result, he has no prostate cancer
  - That he was misclassified
- (10 marks)

#### QUESTION FOUR (20 MARKS)

- a) Show that the conjugate prior for Binomial likelihood is Beta distribution. (6 marks)
- b) The marks distribution for form four classes, Kalaba secondary school were represented in a boxplot as shown below. Interpret the marks distribution for each class (4 marks)



- c) The table below is part of an output of data analyzed whose response variable was whether or not (1=yes,0=no) a student has a fee balance {y}, the predictor variables included the students age {x1}, gender {x2}, marital status {x3}, family size {x4} and the religion coded such that (1=catholic,2=protestant,3=muslim,4=hindu) {x5}

Variables in the equation

| Step1 <sup>a</sup> | B       | S.E.   | Wald  | df | Sig. | Exp(B)  | 95% C.I. for EXP(B) |         |
|--------------------|---------|--------|-------|----|------|---------|---------------------|---------|
|                    |         |        |       |    |      |         | Lower               | Upper   |
| Family-size        | 1.338   | .494   | 4.419 | 1  | .036 | .354    | .135                | .932    |
| Age                | 1.197   | .120   | 2.710 | 1  | .043 | .821    | .649                | 1.038   |
| Marital status     | .201    | .173   | 1.342 | 1  | .247 | 1.222   | .870                | 1.716   |
| Religion           |         |        | 1.085 | 3  | .781 |         |                     |         |
| Religion(1)        | -.073   | 1.477  | .002  | 1  | .961 | .930    | .051                | 16.801  |
| Religion(2)        | 1.468   | 1.743  | 1.710 | 1  | .047 | 4.342   | .143                | 132.234 |
| Religion(3)        | -19.860 | 11.714 | .000  | 1  | .999 | .000    | .000                | .       |
| Gender(1)          | .645    | 1.343  | .231  | 1  | .631 | 1.906   | .137                | 26.476  |
| Constant           | 8.329   | 4.262  | 3.819 | 1  | .051 | 412.576 |                     |         |

- By citing the reasons highlight the variables that contributed significantly to the prediction of the fee balance status of a student. (4 marks)
- Interpret the betas,  $\text{Exp}(\beta)$  and the 95%CI for  $\text{Exp}(\beta)$  of the variables highlighted in (i). (4 marks)
- Write down the regression equation using only the significant variables. (2 marks)

#### QUESTION FIVE (20 MARKS)

- Highlight four ways of normalizing a skewed data for analysis (4 marks)
- Discuss four assumptions of parametric test statistics to apply (8 marks)
- The populism of two products, A and B were tested by carrying out an interview among the residents of the two main consumer blocks. The number of consumers per block is summarized in the table below.

| Block         | Products |    |
|---------------|----------|----|
|               | A        | B  |
| Highlands (H) | 13       | 60 |
| Lowlands (L)  | 54       | 27 |

Test whether the popularity levels differed significantly per block at  $\alpha = 5\%$  (8 marks)