



MACHAKOS UNIVERSITY

University Examinations 2018/2019

SCHOOL OF PURE AND APPLIED SCIENCES

DEPARTMENT OF MATHEMATICS, STATISTICS AND ACTUARIAL SCIENCE

THIRD YEAR SECOND SEMESTER EXAMINATION FOR
BACHELOR OF SCIENCE IN MATHEMATICS

SMA 367: EXPLORORY DATA ANALYSIS

DATE: 24/4/2019

TIME: 8:30 – 10:30 AM

INSTRUCTION:

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

QUESTION ONE (COMPULSORY) (30 MARKS)

- a) Distinguish the following terms as they apply in data analysis
- i) A variable and a case
 - ii) Descriptive statistics and inferential statistics
 - iii) Parametric and Non-parametric test statistics (6 marks)
- b) Highlight three ways of transforming a non-normally distributed data (3 marks)
- c) Two different types of drugs A and B were tried on some patients for weight gain, 7 were given drug A while 13 were given drug B. the weight gain was as summarized below.

Drug A	10	8	8	12	14	16	2	5	4	-	-	-	-
Drug B	3	6	8	9	11	1	3	5	8	13	10	3	6

By using a computer or otherwise test whether the two drugs differ significantly with regard to their effect in increasing weight at $\alpha = 5\%$? (6 marks)

- d) The following data represents the sampled of 15 BSc mathematics students marks, present it in a stem and leaf plot:
23, 35, 25, 22, 25, 43, 27, 54, 28, 55, 32, 36, 36, 45, 50 (3 marks)

- d) A cohort study of smoking and bladder cancer was conducted in a small island population. There were a total of 1000 people on the island where 400 were smokers while 600 were not. Fifty of the smokers and ten of the non-smokers developed bladder cancer. Using this information; construct a two by two table and label the cells and margins and hence determine

- i) Relative Risk (risk ratio)
- ii) Attributable Risk
- iii) Odds Ratio (6 marks)

- d) 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. (2-tailed)	Mean difference	95% CI of the 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 (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 poison likelihood is gamma distribution. (6 marks)
- b) Discuss five ways of handling an outlier at data analysis stage (10 marks)

QUESTION THREE (20 MARKS)

- a) Highlight four sources of data for scientific research (4 marks)
- b) Show that the conjugate prior for binomial likelihood is beta distribution. (6 marks)

The data below relates to average score in end of the semester basic mathematic for BSc computer science students from a Kenyan university.

Female	10	38	48	12	14	16	22	25	44	32	37	28	03
Male	32	65	80	91	11	51	43	59	78	13	10	33	56

Explore the score distribution by gender, represent the scores in a box plots and interpret (10 marks)

QUESTION FOUR (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 FIVE (20 MARKS)

- a) Discuss four assumptions of parametric test statistics (8 marks)
- b) A random sample of ten mango trees had each the following sacks of fruits (y) and the tree spacing in metres

Yield (y)	20	30	33	40	15	13	26	38	35	43
Space (x)	7	9	8	11	5	4	8	10	9	10

- Plot the data and commend on the graph.
 - Estimate the regression equation of yield on space and interpret your results.
 - If the assumptions of the linear model are valid, test for the statistical significance of the parameter estimates at 5% level.
 - What are the predicted values?
 - Compute the residuals and the adjusted residuals and interpret the results.
 - Plot a graph of residuals against the predicted values.
 - Compute the 95% confidence intervals for b_0 and b_1 .
- (12 marks)