



MACHAKOS UNIVERSITY

University Examinations 2022/2023 Academic Year

SCHOOL OF PURE AND APPLIED SCIENCES

DEPARTMENT OF MATHEMATICS AND STATISTICS

THIRD YEAR SECOND SEMESTER EXAMINATION FOR

BACHELOR OF SCIENCE (MATHEMATICS AND COMPUTER SCIENCE)

SMA 367 EXPLORATORY DATA ANALYSIS

DATE: 12/3/2023

TIME: 2:00 – 4:00 PM

INSTRUCTIONS

Answer Question One and Any Other Two Questions

QUESTION ONE (30 MARKS)

- a) Distinguish between the following terms as used in statistics
- i. Descriptive statistics and inferential statistics (2 marks)
 - ii. Parameter and statistic (2 marks)
 - iii. Prior probabilities and posterior probabilities (2 marks)
- b) Suppose the fat contents (in grams) for eating English breakfasts and meat sandwiches are to be compared. The fat contents are shown below:

Sandwiches	26	17	12	34	7	18	20	21	21	24	6	28	30	13
English breakfast	12	41	25	16	18	45	15	25	36	36	38	14	44	23

Construct a stem and leaf plot to represent this information and interpret the display. (6 marks)

- c) Highlight four benefits of graphical representation of data. (4 marks)
- d) Differentiate between Type I and Type II errors as used in hypothesis testing. (4 marks)
- e) A doctor knows that the disease meningitis causes the patient to have a stiff neck, 70% of the time. The doctor also knows that the prior probability that a patient has meningitis is $1/50000$ and the prior probability that any patient has a stiff neck is 1%. Let s be the proposition that the patient has a stiff neck and m be the proposition that the patient has meningitis. Determine $P(s|m)$. (5 marks)

- f) Highlight five ways of normalizing skewed data for analysis. (5 marks)

QUESTION TWO (20 MARKS)

- a) Explain the five steps involved in Bayesian analysis of data. (5 marks)
- b) Show that the conjugate prior for a Poisson likelihood is the gamma distribution. (6 marks)
- c) In a forest area of Northern Europe there may be wild lynx. At a particular time the number X of lynx can be between 0 and 5 with

$$P(X = x) = \binom{5}{x} 0.6^x 0.4^{5-x}, x = 0, 1, \dots, 5$$

A survey is made but the lynx is difficult to spot and, given that the number present is x , the number Y observed has a probability distribution with

$$P(Y = y|X = x) = \begin{cases} \binom{x}{y} 0.3^y 0.7^{x-y} & 0 \leq y \leq x \\ 0 & x < y \end{cases}$$

Determine the conditional probability distribution of X given that $Y = 2$. (9 marks)

QUESTION THREE (20 MARKS)

- a) Outline the procedure used for hypothesis testing. (4 marks)
- b) An investigation was carried out into the number of hours (*hours*) lost due to stress-related illnesses at various companies (*company*). The null hypothesis stated that the means for the five companies sampled were equal while the alternative hypothesis stated that not all five companies' means are equal. ANOVA was used to test this hypothesis and the R output is as below.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
company	4	6507	1626.7	12.78	0.000138 ***
Residuals	14	1782	127.3		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

- i. Interpret the output and give your conclusion for the hypothesis under test. (3 marks)
- ii. Explain what can be done to determine which groups are different from each other. (3 marks)

- c) A study was carried out to assess the value of houses in a certain neighbourhood. A multiple regression was carried out to determine whether the variables *age* (average age of houses) and *lstat* (percent of houses with low socioeconomic status) had an effect on *medv* (median house value). The following output was obtained from R:

Residuals:

Min	1Q	Median	3Q	Max
-15.981	-3.978	-1.283	1.968	23.158

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	33.22276	0.73085	45.458	< 2e-16 ***
age	0.03454	0.01223	2.826	0.00491 **
lstat	-1.03207	0.04819	-21.416	< 2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.173 on 503 degrees of freedom

Multiple R-squared: 0.5513, Adjusted R-squared: 0.5495

F-statistic: 309 on 2 and 503 DF, p-value: < 2.2e-16

- Write down the regression line. (1 mark)
- Write down the hypothesis that you would test to check whether the explanatory variables are contributing significantly in predicting the response variable. (2 marks)
- From the output above, what would be the result of your tests in ii above at 5% significant level? (3 marks)
- What is the difference between the Multiple R squared and Adjusted R squared as given in R output? (2 marks)
- According to the output, would you say that the model is a good fit? Explain your answer. (2 marks)

QUESTION FOUR (20 MARKS)

- a) Show that the conjugate prior for a binomial likelihood is the beta distribution. (6 marks)

- b) A study was conducted to investigate the relationship between the price of a plot of land and the distance from the nearest tarmac road. A random sample of 12 plots was taken and the data relating the distance from the nearest tarmac road in hundred metres and the price of the plot in million Kenya shillings is as shown in the table below.

Plot	1	2	3	4	5	6	7	8	9	10	11	12
Distance	48	78	35	72	10	60	40	22	84	25	12	48
Price	1.2	0.4	1.4	0.8	1.8	0.6	1.5	1.2	0.8	1.5	1.6	1.2

- i. Compute the Pearsons product moment coefficient of correlation and interpret your result. (6 marks)
- ii. Write R code to compute the Pearson's product moment co-efficient of correlation between the distance from the tarmac and the price of the plot. (2 marks)
- c) A farmer wished to evaluate the effect of fertilizer PHK on his 10-acre farm. He noted the yield per acre before and after applying the fertilizer as recorded below:

Before	111	113	114	116	116	115	115	114	113	113
After	150	150	155	160	165	165	165	160	160	150

The output analysis is summarized below:

data: after and before

t = 26.833, df = 9, p-value = 6.711e-10

alternative hypothesis: true mean difference is not equal to 0

95 percent confidence interval:

40.29055 47.70945

sample estimates:

mean difference

44

- i. Based on the t-value and the sig.(2-tailed) value make statistical conclusion.
- ii. Interpret the 95% CI lower and upper difference values.
- iii. Highlight the procedure in getting the output using R.

(6 marks)

QUESTION FIVE (20 MARKS)

- a) Highlight three ways of transforming a non-normally distributed data to normal. (3 marks)
- b) In order to increase the efficiency with which employees in a certain organisation can carry out a task, 5 employees are sent on a training course. The time in seconds to carry out the task both before and after the training course is given below for the 5 employees:

Employee	1	2	3	4	5
Before	42	51	37	43	45
After	38	37	32	40	48

- i) Assuming the data is already saved in the workspace as *employeedata*, with the variables *before* and *after*, write R code to test whether the training course has had the desired effect. (2 marks)
- ii) Assuming you run the code in R and part of the results are as given below:
 $t = 1.6842$, $df = 4$, $p\text{-value} = 0.08372$
State and give the interpretation of your hypothesis, taking $\alpha = 0.05$. (3 marks)
- c) Outline the six important types of departures from the simple linear regression model that can be examined by use of residuals. In each case indicate the plot to be used. (12 marks)