



STRATHMORE INSTITUTE OF MATHEMATICAL SCIENCES (SIMS)
MASTER OF SCIENCE IN DATA SCIENCE AND ANALYTICS
END OF SEMESTER EXAMINATION
DSA 8104: FUNDAMENTAL STATISTICAL ANALYSIS

DATE: 7th Aug, 2021

TIME: 3 Hours

INSTRUCTIONS

1. This examination consists of **FOUR** questions.
 2. Answer Question **ONE (COMPULSORY)** and any other **TWO** questions.
 3. You may use a **SIMPLE CALCULATOR**. No **MOBILE PHONES** in the exams room.
-
-

Question One (30 Marks)

- (i) Explain four assumptions associated with a linear regression model. (4 marks)
- (ii) Define Bayes' Theorem (2 marks)
- (iii) A mean of 60 on seven exams is needed to pass a course. On her first six exams, Tina received grades of 51, 72, 80, 62, 57, and 69. What grade must she receive on her last exam to pass the course? (3 marks)
- (iv) The table below shows the quantity in hundred kgs of Wheat, Barley and Oats produced on a certain farm during the years 2011 to 2014.

Year	Wheat	Barley	Oats
2011	34	18	27
2012	43	14	24
2013	43	16	27
2014	45	13	34

Use the most appropriate bar chart to illustrate this data in terms of its percentage. (4 marks)

- (v) A random sample of 12 graduates of a certain secretarial school typed an average of 79.3 words per minute with a standard deviation of 7.8 words per minute. Assuming a normal distribution for the number of words typed per minute; find a 95% confidence interval for the average number of words typed by all graduates of this school. (5 marks)

- (vi) If 3% of the electric bulbs manufactured by a company are defective find the probability that in a sample of 100 bulbs exactly 5 bulbs are defective. (3 marks)
- (vii) The following information relates to absences in a statistics course and the final exam grade, for some 7 students. Find the correlation coefficient and interpret your result. (5 marks)

$$\sum x = 19, \quad \sum y = 565, \quad \sum x^2 = 75, \quad \sum y^2 = 46,775, \quad \sum xy = 1,380.$$

- (viii) It is believed that at least 60% of the residents in a certain area favor an annexation suit by a neighboring city. What conclusion would you draw if only 110 in a sample of 200 voters favor the suit? Use a 0.05 level of significance. (4 marks)

Question Two (15 Marks)

- (i) State and explain the most common goodness of fit tests (3 marks)
- (ii) A tobacco company statistician wishes to know whether heavy smoking is related to longevity. From a sample of recently deceased smokers, the number of cigarettes (estimated on a per day for their last five years after visits with their surviving relatives) is paired with the number of years that they lived.

Cigarettes	25	35	10	40	85	75	60	45	50
Years lived	63	68	72	62	65	46	51	60	55

- (a) Present the information using a scatter diagram. (3 marks)
- (b) Find the regression equation in the form $y = \beta_0 + \beta_1 x$. (3 marks)
- (c) Give a practical interpretation of the slope β_1 . (2 mark)
- The value of $\beta_1 = -0.2142$ implies that an increase in smoking by one unit of cigarette results to a reduction in the number of years live by 0.2142.
- (d) Using your answer in (b), find the number of years lived for someone who smoked 101 cigarettes. (2 marks)
- (e) Test the null hypothesis that the regression slope is zero at 95% confidence interval (2 marks)

Question Three (15 Marks)

- (i) In a retail market, fruit vendors were selling mangoes kept in packing boxes. These boxes contained varying number of mangoes. The following was the distribution of mangoes according to the number of boxes.

Number of mangoes	50-52	53-55	56-58	59-61	62-64
Number of boxes	15	110	135	115	25

- (a) Find the Harmonic mean number (HM) of mangoes kept in a packing box. (2 marks)
- (b) Find the Geometric mean (GM) of mangoes kept in a packing box. (2 marks)
- (c) Calculate and interpret the standard deviation of the distribution. (2 marks)
- (ii) The height of adult females are normally distributed with mean 160cm and standard deviation 8cm. Find the probability that a randomly selected adult female has a height greater than 170cm. (3 marks)
- (iii) A factory has three units A, B, and C. Unit A produces 50% of its products, and units B and C each produces 25% of the products. The percentage of defective items produced by A, B, and C units are 3%, 2% and 1%, respectively. If an item is selected at random from the total production of the factory is found defective, what is the probability that it is produced by:
- (a) Unit A (b) Unit B (c) Unit C (6 marks)

Question Four (15 Marks)

- (i) Let X denote the IQ of a randomly selected adult Kenyan. Assume, that X is normally distributed with unknown mean μ and standard deviation 16. Take a random sample of $n = 16$ students, so that, after setting the probability of committing a Type I error at $\alpha = 0.05$ we can test the null hypothesis

$$H_0 : \mu = 100 \text{ against the alternative hypothesis that } H_A : \mu > 100$$

What is the power of the hypothesis test if the true population mean were $\mu = 108$? (3 marks)

- (ii) The math teacher wants to determine the effectiveness of her statistics lesson and gives a pre-test and a post-test to 9 students in her class. Our hypothesis is that there is no difference between the means of the two samples and our alternative hypothesis is that the two means of the samples are not equal. In other words, we are testing whether or not these two samples are related or: (4 marks)

$$H_0 : \delta = \mu_1 - \mu_2 = 0$$

$$H_a : \delta = \mu_1 - \mu_2 \neq 0$$

The results for the pre-and post-tests are below:

Subject	1	2	3	4	5	6	7	8	9
Pre-test Score	78	67	56	78	96	82	84	90	87
Post-test Score	80	69	70	79	96	84	88	92	92

- (iii) The following are 15 measurements of the octane rating of a certain kind of gasoline: 97.5, 95.2, 97.3, 96.0, 96.8, 100.3, 97.4, 95.3, 93.2, 99.1, 96.1, 97.6, 98.2, 98.5 and 94.9. Use Signed-Rank test with $\alpha = 0.05$ (significance level) to test whether the mean octane rating of the given kind of gasoline is 98.5. (4 marks)

- (iv) Use the following data to test at the $\alpha = 0.01$ significance level whether a person's ability in mathematics is independent of his or her interest in computer science: (4 marks)

	Low Ability	Average Ability	High Ability
Low Interest	63	42	15
Average Interest	58	61	31
High Interest	14	47	29

END