



MACHAKOS UNIVERSITY

University Examinations 2022/2023 Academic Year

SCHOOL OF PURE AND APPLIED SCIENCES

DEPARTMENT OF MATHEMATICS AND STATISTICS

SECOND YEAR SECOND SEMESTER EXAMINATION FOR
BACHELOR OF SCIENCE (ANALYTICAL CHEMISTRY)

SAN 208/HFN 253: STATISTICAL DATA ANALYSIS

DATE: 18/4/2023

TIME: 2:00 – 4:00 PM

Instructions to the Candidate:

1. Answer **Question 1** and any other **two** questions.
2. You must have a scientific calculator for this paper.

QUESTION ONE (COMPULSORY) (30 MARKS)

- a) By citing examples distinguish between the control charts for variables and control charts for attributes. (4 marks)
- b) Differentiate the following terms as they apply in measurements and statistics
 - i. Descriptive statistics and Inferential statistics
 - ii. Parametric test statistics and Non- parametric test statistics
 - iii. Type I error and Type II error in hypothesis testing (6 marks)
- c) The population standard deviation for the amount of aspirin in a batch of analgesic tablets is known to be 7 mg of aspirin. If after randomly selecting ten tablets and analyzing it was found to contain 245 mg of aspirin, what is the 95% confidence interval for the population's mean? (6 marks)
- d) Explain three characteristics of a good estimator of population parameters (6 marks)

- e) Two different types of drugs A and B were tried on different couples for increasing fertility, 5 couples were given drug A and 7 couples were given drug B. The conception rate after a period of five years was as given below:

Drug A	2	4	5	3	3	-	-
Drug B	4	3	4	5	3	2	4

Its t-test output is summarized in the table below

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	-0.2802	10	0.785	0.17	-1.534	1.192

- Construct the hypothesis statements for the above scenario
- Based on the sample t-test output make statistical conclusion
- Interpret the 95% CI lower and upper mean difference values (8 marks)

QUESTION TWO (20 MARKS)

- a) Determine the regression equation of Y on X given that $\sum xy = 125$, $\sum x^2 = 2500$, $\bar{x} = 60$ and $\bar{y} = 4$. And hence calculate the value of Y when $X=20$ (6 marks)
- b) In a random sample of 150 tablets sachets taken from machine A, 50 are found to be defective. In another sample of 250 sachets taken from machine B, 100 are found to be defective. Do the data reveal significant difference between the two machines as far as the quality output is concerned, at $\alpha = 0.05$? (6 marks)
- c) The table below shows the KQ products' length mean and range for 10 randomly taken samples each of size 5

Sample no.	1	2	3	4	5	6	7	8	9	10
Range	13.5	12.9	11.4	10.8	11.6	9.8	11.6	11.3	10.9	11.2
Mean	7	4	8	5	6	4	7	4	7	8

By using mean and range charts determine whether the process is in control (8 marks)

QUESTION THREE (20 MARKS)

- a) In a random sample of 600 active antigens, 45 were found to have been expired. Determine the 95% confidence level the population proportion that expired. (5 marks)
- b) A contracting farming agency collected the data below according to the listed variables in order to predict the milk production as influenced by different dairy feeds in future:

X1	23	34	56	78	15	21	24
X2	34	56	78	79	90	65	14
X3	23	12	12	13	14	16	23
Y	4	5	12	23	4	5	8

Y= Milk production (Jericans of 5 litres)

X1= Pasture

X2= Fodder

X3 = Silage

The agency used ANOVA and regression, the following is part of the results:

SUMMARY OUTPUT

Regression Statistics		Coefficients	
Multiple R	0.625373	intercept	-15.33137
R Square	0.391092	X1	0.31518
Adjusted R Square	-0.21782	X2	0.06505
Standard Error	7.63801	X3	0.54998
Observations	7		

ANOVA

	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance</i>
Regression	3	<i>x</i>	37.47033	<i>z</i>	0.637562
Residual	3	175.0176	<i>y</i>		
Total	6	287.4286			

- i. State the statistical model in this analysis. (2 marks)
- ii. Complete for the values; *x*, *y* and *z* (3 marks)
- iii. Interpret the finding in this analysis and fit in the model (6 marks)
- iv. Explain the difference between R Square and R Adjusted (4 marks)

QUESTION FOUR (20 MARKS)

- a) Highlight two characteristics of a normally distributed data (2 marks)
- b) Highlight four types of non- parametrical test statistics (4 marks)
- c) The weight of the citizens who attended the slim possible audition was found to be normally distributed with an average weight of 100kg and a standard deviation of 16kg. Determine
- The probability that a participant picked at random weighed 132kg
 - The percentage of participants weighing between 92kg and 132kg
- (6 marks)
- d) A pharmaceutical company made the following sales and expenses in a certain year.

Firm:	A	B	C	D	E	F	G	H	I	J
Sales (Ksh, 000):	50	50	55	60	65	65	65	60	60	50
Expenses (Ksh. 000):	11	13	14	16	16	15	15	14	13	13

Determine;

- The correlation coefficient between the sales and expenses from the data
 - Determine the coefficient of determination and interpret it
- (8 marks)

QUESTION FIVE (20 MARKS)

- a) Distinguish the following terms as they apply in measurements and statistics
- Specificity and Sensitivity in diagnostic tests
 - Limits of Detection(LOD) and Limit of Quantification(LOQs)
 - Measurement errors and Personal errors of precision
 - Process control and product control
 - Assignable causes of variations and common causes of variations
- (10 marks)
- b) A random sample of 400 medical school community was selected from each of three age groups and each person was asked to specify which of the three covid-19 test s among; Polymerase Chain Reaction (PCR), Lateral Flow Tests (LFTs) and Antibody she/ he preferred. The results are shown in the table below:

Age group	Covid-19 Tests		
	PCR	LFTs	Antibody
Under 30	120	30	50
31 – 50	10	75	15
51 and above	10	30	60

Test the hypothesis that the populations were homogeneous with respect to the covid-19 tests they preferred despite their age difference.

(10 marks)