

PRACTICE PAPER SET 1

Statistics for AI & Data Science (Paper Code: 48895)

T.E. Computer Engineering & AI-DS, Semester V

Duration: 3 Hours | Total Marks: 80

By: Nitin Gupta

Instructions to Candidates:

1. Question No. 1 is compulsory
2. Attempt any three questions from remaining five questions
3. All questions carry equal marks
4. Assume suitable data, if required and state it clearly

Q1. Attempt any FOUR: [20 Marks]

a. [5 Marks]

What is hypothesis testing? Explain type I and type II errors with examples.

b. [5 Marks]

Find the standard deviation for the following dataset: 15, 18, 22, 20, 16, 19, 21, 17

c. [5 Marks]

Explain the difference between Stratified and Cluster sampling with examples.

d. [5 Marks]

What do you mean by correlation and regression? How are they different?

e. [5 Marks]

Define Confidence Interval and explain its importance in statistics.

f. [5 Marks]

What is Fisher's exact test? When is it used instead of Chi-Square test?

Q2. [20 Marks]

a. [10 Marks]

X is a normally distributed variable with mean $\mu = 50$ and standard deviation $\sigma = 10$. Find:

1. $P(x < 65)$
2. $P(x > 40)$
3. $P(45 < x < 60)$

b. [10 Marks]

What is Chi-Square Test? A company wants to determine if there is a significant association between employee department and preference for remote vs office work. Use Chi-Square Test for Independence at $\alpha = 0.05$.

Department	Prefers Remote	Prefers Office	Total
IT	45	15	60
HR	20	30	50
Sales	25	35	60
Total	90	80	170

Q3. [20 Marks]

a. [10 Marks]

Find the correlation coefficient from the following data:

Student	Study Hours (X)	Test Score (Y)
1	2	55
2	4	65
3	6	75
4	8	85
5	10	90
6	12	95

Interpret the result.

b. [10 Marks]

Explain the concept of p-value in hypothesis testing. A pharmaceutical company claims their new drug reduces fever in 6 hours on average. The standard treatment takes 8 hours with SD = 2 hours. A sample of 25 patients using the new drug showed mean recovery time of 7 hours. Test the claim at $\alpha = 0.05$.

1. State null and alternative hypotheses
2. Calculate test statistic
3. Make decision

Q4. [20 Marks]**a. [10 Marks]**

Explain why ANOVA is used. Solve the following using one-way ANOVA to identify any difference between teaching methods:

Method A	Method B	Method C
75	82	88
80	78	92
85	85	90
78	88	85
82	80	95

b. [10 Marks]

Create a frequency distribution table for the following ages of 40 participants:

22, 25, 28, 30, 22, 35, 40, 22, 25, 28, 32, 35, 38, 40, 42, 25, 28, 30, 32, 35,
22, 25, 28, 30, 35, 38, 40, 42, 45, 22, 25, 30, 32, 35, 38, 28, 30, 32, 35, 40

Calculate:

1. Class mark for each interval
2. Range of data
3. Modal class

Q5. [20 Marks]

a. [10 Marks]

Find the simple linear regression equation for the given data. Calculate coefficient of determination (R^2):

Advertising Budget (X) in ₹1000	Sales (Y) in ₹10000
5	20
10	35
15	45
20	60
25	70
30	85

b. [10 Marks]

Calculate mean, standard deviation, standard error, and 95% confidence interval for the following sample data:

12, 15, 18, 14, 16, 13, 17, 15, 14, 16, 18, 12

Q6. [20 Marks]

a. [10 Marks]

A school conducted an aptitude test for three different classes. Use Kruskal-Wallis test at 0.05 significance level to determine if scores differ significantly:

Class A	Class B	Class C
85	78	92
88	82	88
90	75	95
82	80	90
87	85	93

b. [10 Marks]

Write short notes on (any two):

1. Box Plot and its components
2. Weibull Distribution
3. Bootstrapping vs Re-sampling
4. Central Limit Theorem

END OF PAPER

Best of luck!

Compiled by: Nitin Gupta