

PREDICTED PAPER 2025

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 real-world examples from AI/ML applications. What are the consequences of each error type?

b. [5 Marks]

What is Fisher's exact test? When should it be used instead of Chi-Square test? Give an example scenario from data science.

c. [5 Marks]

Explain the difference between Stratified and Cluster sampling. In what scenarios would each be preferred for collecting AI training data?

d. [5 Marks]

Explain Linear Regression and its applications in predictive analytics and machine learning.

e. [5 Marks]

Define standard deviation and interquartile range with examples. How are they used to detect anomalies in data?

f. [5 Marks]

Explain the concept of p-value in hypothesis testing. How do you interpret p-values in the context of statistical significance?

Q2. [20 Marks]

a. [10 Marks]

Find the correlation coefficient from the given data and interpret:

Data Scientist	Years Experience (X)	Annual Salary (Y) in ₹ Lakhs
1	2	6
2	4	8
3	6	12
4	8	15
5	10	18
6	12	22

1. Calculate Pearson correlation coefficient
2. Interpret the strength and direction
3. Can we conclude causation? Explain.

b. [10 Marks]

What is Chi-Square Test? An e-commerce company wants to determine if customer satisfaction is associated with delivery speed. Data from 300 customers:

Satisfaction	Fast Delivery	Standard Delivery	Slow Delivery	Total
Satisfied	80	50	20	150
Neutral	30	40	30	100
Dissatisfied	10	20	20	50
Total	120	110	70	300

Use Chi-Square Test at $\alpha = 0.05$ to determine association. Show all steps including:

1. Null and alternative hypotheses

2. Expected frequencies
3. Chi-square statistic
4. Degrees of freedom
5. Decision and interpretation

Q3. [20 Marks]

a. [10 Marks]

Explain the concept of p-value in hypothesis testing with examples. A machine learning model's accuracy on training data is claimed to be 85%. On a test set of 40 samples, accuracy is 80% with SD = 8%. Test at $\alpha = 0.05$ if the claim is valid:

1. State null and alternative hypotheses (two-tailed)
2. Calculate appropriate test statistic
3. Find p-value
4. Make decision
5. Interpret in context of model evaluation

b. [10 Marks]

A tech company conducted coding tests for three different training programs. Use Kruskal-Wallis test at $\alpha = 0.05$ to determine if median scores differ:

Program A	Program B	Program C
88	82	95
92	78	90
85	85	93
90	80	88
87	83	92
89	79	91

Show complete working including ranking procedure.

Q4. [20 Marks]

a. [10 Marks]

A data scientist is analyzing algorithm runtime. Sample mean runtime for 25 test cases is 145 ms, expected mean is 150 ms, and standard deviation is 20 ms. Calculate:

1. z-score for the sample mean
2. Probability of observing this or more extreme value
3. Interpret result

Then, create a frequency distribution table for the following runtimes (in ms) of 50 test cases:

120, 135, 145, 150, 125, 140, 155, 130, 145, 160, 135, 140, 150, 145, 155,
125, 130, 145, 150, 165, 140, 135, 150, 145, 155, 130, 140, 145, 160, 150,
125, 135, 140, 155, 150, 145, 135, 140, 150, 155, 130, 145, 150, 140, 135,
150, 145, 155, 140, 150

Answer: class intervals, frequencies, cumulative frequency, and relative frequency.

b. [10 Marks]

Explain QQ plots in detail and their use in checking normality assumptions. How do scatter plots help in exploratory data analysis for machine learning feature selection?

Q5. [20 Marks]

a. [10 Marks]

A cloud service provider claims their new optimization reduces average server response time from 120 ms to below 115 ms. Standard deviation is 15 ms. A sample of 35 requests shows mean response time of 112 ms. At $\alpha = 0.05$:

1. State null and alternative hypotheses (one-tailed test)
2. Calculate z-test statistic
3. Find critical value
4. Calculate p-value
5. Make decision using both methods
6. Interpret: Is the optimization effective?

b. [10 Marks]

Find the simple linear regression equation for the given data:

Training Data Size (X) in 1000s	Model Accuracy (Y) in %
5	70
10	75

Training Data Size (X) in 1000s	Model Accuracy (Y) in %
15	82
20	88
25	92
30	95

Calculate:

1. Regression coefficients (slope and intercept)
2. Coefficient of determination (R^2)
3. Predict accuracy for 18,000 training samples
4. Interpret R^2 in context of model performance

Q6. [20 Marks]

a. [10 Marks]

Explain the concept of two-way ANOVA with example. How does it differ from one-way ANOVA?

- Describe main effects and interaction effects
- State assumptions of two-way ANOVA
- Explain when to use Friedman's test as non-parametric alternative
- Give example from A/B testing in data science

Provide a neat diagram showing two-way ANOVA design.

b. [10 Marks]

Write comprehensive notes on (any two):

Option 1: Chi-square distribution

- Definition and properties
- Relationship with normal distribution
- Applications in hypothesis testing
- Degrees of freedom concept
- Use in goodness-of-fit tests

Option 2: Weibull distribution

- Definition and parameters
- Shape of distribution for different parameters
- Applications in reliability analysis

- Use in survival analysis and ML
- Comparison with exponential distribution

Option 3: Stem & Leaf Plot

- Construction method
- Advantages over histogram
- When to use
- Example from real data
- Interpretation

Option 4: Box Plot

- Components (Q1, Q2, Q3, IQR, whiskers, outliers)
- Construction steps
- Interpretation
- Use in comparing distributions
- Applications in anomaly detection

END OF PAPER

Notes for 2025 Examination:

High Priority Topics:

1. **Hypothesis Testing** - Core foundation, expect detailed questions
2. **Linear Regression** - With coefficient of determination
3. **Correlation vs Regression** - Conceptual understanding critical
4. **Chi-Square Test** - With complete calculations
5. **Normal Distribution** - Multiple probability calculations
6. **ANOVA** - One-way mandatory, two-way increasingly important
7. **Confidence Intervals** - With standard error calculations

Trending Topics (Likely in 2025):

1. **Kruskal-Wallis Test** - Non-parametric alternative gaining importance
2. **p-value interpretation** - Critical for data science
3. **Friedman Test** - Repeated measures scenarios
4. **Type I & II Errors** - Real-world consequences
5. **Sampling Methods** - For big data scenarios

Formula Sheet to Prepare:

- All distribution formulas (Normal, Binomial, Poisson, t, F, Chi-square)
- Regression formulas (slope, intercept, R^2)
- Correlation coefficient
- Standard error and confidence interval
- Test statistics (z, t, F, Chi-square)
- ANOVA computations

Exam Tips:

- Always state hypotheses clearly
- Show step-by-step calculations
- Draw diagrams where asked
- Interpret results in context
- Practice numerical problems extensively
- Memorize critical values tables
- Understand when to use each test

Best of luck for your 2025 exams!

Compiled by: Nitin Gupta