SET A

Name:

Enrolment No: Tutorial Batch: Signature:

QUESTION BOOKLET
INDIAN INSTITUTE OF TECHNOLOGY ROORKEE
MEHTA FAMILY SCHOOL OF DATA SCIENCE AND ARTIFICIAL INTELLIGENCE
DAI-101: Introduction to Data Science
Autumn Semester 2025

Total Marks: 45 **MTE** **Time: 90 Minutes**

Note: There are 48 questions in this Question Booklet. You need to answer any 45 Questions. There is ONE mark for each correct answer. A negative marking of 1/3 is applicable for an incorrect answer.

USEFUL DATA:

Critical Z-values for Common Confidence Levels $(1-\alpha)$. The value given is for $Z_{\alpha/2}$	Probability Values for Some Z-Scores
$z_{0.05} = 1.645$	P(Z<1)≈0.8413
$z_{0.025} = 1.96$	P(Z<2)≈0.9772
$z_{0.005} = 2.576$	P(Z<3)≈0.9987

1. Consider the following two statements for a fixed sample size n and standard deviation σ:
   S1: Higher the confidence level, longer the resulting Confidence Interval
   S2: Lower the confidence level, lower the resulting Confidence Interval

   Then,

   (a) Only S1 is TRUE.
   (b) Only S2 is TRUE.
   (c) Both S1 and S2 are TRUE.
   (d) Neither S1 nor S2 is TRUE.

2. If the length of the Confidence Interval (CI) at $\alpha = 0.01$ and $\alpha = 0.05$ are l_1 and l_2 respectively, then l_1/l_2 is

   (a) 1.08 (b) 1.32 (c) 1.44 (d) 1.58

3. What is the confidence level for the interval

$$\mu - 2.14\frac{\sigma}{\sqrt{n}} \leq \bar{x} \leq \mu + 2.14\frac{\sigma}{\sqrt{n}}$$

   (Given CDF of standard normal distribution for 2.14 is 0.9838)?

(a) 96.76 (b) 98.38 (c) 95.0 (d) 99.0

4. A random sample has been taken from a normal distribution. Output from a software package is as follows. Values at ? are to be calculated by you.

Variable	N	Mean	SE Mean estimate	Sample Std. Dev.	Variance	Sum $\sum X_i$
X	?	?	1.5	6.0	?	752.0

What is the Confidence Interval on the population mean?

Given $t_{0.025,14} = 2.145$, $t_{0.025,15} = 2.131$ and $t_{0.025,16} = 2.120$.

(a) (43.68, 50.32)
(b) (43.80, 50.20)
(c) (43.92, 50.08)
(d) (44.02, 49.99)

5. Which of the following statements is **NOT** correct related to decisions in hypothesis testing?
   (a) No error: Fail to reject H_o when H_o is true
   (b) Type I error: Fail to reject H_o when H_o is true
   (c) Type I error: Reject H_o when H_o is true
   (d) No error: Reject H_o when H_o is false

6. The number of people visiting a garden to watch flower show in 2024 is approximately normally distributed. The population mean is thought to be 100, and the population standard deviation is 40. You wish to test the $H_o: \mu = 100 \; versus \; H_1: \mu \neq 100$ with a sample of $n = 25$ people. Find the Type I error probability α given the acceptance region is $98 \leq \bar{x} \leq 102$. Given CDF for standard normal variable at 0.25 is 0.5987.
   (a) 0.4013
   (b) 0.5987
   (c) 0.8026
   (d) 0.9011

Answer Q. No. 7 to 10 based on the data of the following confusion matrix:

Predicted Values ⟹ / True Values ⇓	**Class 1**	**Class 2**	**Class 3**
Class 1	**100**	**50**	**10**
Class 2	**10**	**150**	**20**
Class 3	**10**	**20**	**50**

7. What is F1-Score for Class 1?
   (a) 10/12
   (b) 10/16
   (c) 5/14
   (d) 5/7

8. What is Precision for Class 2?
   (a) 15/22
   (b) 5/6
   (c) 3/4
   (d) 3/8

9. What is Recall for Class 3?
   (a) 5/8
   (b) 5/16
   (c) 2/5
   (d) 1/2

10. What is the classification accuracy of the model?
   (a) 5/8
   (b) 5/7
   (c) 4/7
   (d) 2/7

11. Consider the following two statements about Precision and Recall:
   S1: Precision measures how accurate the positive predictions are by a trained model
   S2: Recall measures the fraction of the positives the trained model has identified

   Then,

   (a) Only S1 is TRUE.
   (b) Only S2 is TRUE.
   (c) Both S1 and S2 are TRUE.
   (d) Neither S1 nor S2 is TRUE.

12. Which of the following statement is correct about p-value?
   (a) The p-value is the smallest level of significance that would lead to rejection of the null hypothesis.
   (b) The p-value is the probability that null hypothesis is false.
   (c) The p-value is the probability that null hypothesis is true.
   (d) The p-value is the probability of wrongly rejecting the null hypothesis when it is true.

13. A biased coin with probability of heads = 0.9 is tossed 10000 times. What is the probability of observing more than 9030 heads? [You may have to use central limit theorem]
   (a) 0.0013
   (b) 0.178
   (c) 0.0228
   (d) 0.1587

14. Let X be a Poisson random variable with mean as 10. Let Y be a uniform random variable with minimum value as 5 and maximum value as 15. Let Z = X + 2Y. What is the expectation of random variable Z?
    (a) 20
    (b) 30
    (c) 40
    (d) Can't be answered without knowing if X and Y are independent or not

15. Let X be a random variable following normal distribution with mean = 10 and standard deviation = 5. Let Y be a binomial random variable with n = 100 and p=0.4. Let Z = X + 3Y. Assume X and Y are independent. What is the variance of random variable Z?
    (a) 241
    (b) 77
    (c) 97
    (d) 261

16. A dataset has two variables X and Y. For each point in the dataset, it is observed that X = 2Y + 5. What is the correlation between X and Y?
    (a) 0.5
    (b) 0.25
    (c) 0
    (d) 1

17. Consider two random variables X_1 and X_2 for which $E(X_1X_2) = E(X_1)E(X_2)$. Which of the following statements is correct?
    (a) X1 and X2 are independent
    (b) X1 and X2 are independent only if E(X1) = E(X2)
    (c) X1 and X2 are independent only if $E(X1) \neq E(X2)$
    (d) None of the above

18. Consider an examination consisting of 30 multiple-choice questions. Each question has 4 options, with exactly one correct answer. A student randomly selects an answer for every question. The scoring scheme awards +5 points for a correct answer and -2 points for an incorrect answer. Let X be the total score obtained by the student. What is the expected value E(X) of the student's score.
    (a) 0
    (b) -7.5
    (c) -10
    (d) -15

19. Consider an examination consisting of 30 multiple-choice questions. Each question has 4 options, with exactly one correct answer. A student randomly selects an answer for every question. The scoring scheme awards +5 points for a correct answer and -2 points for an incorrect answer. Let X be the total score obtained by the student. What is the variance Var(X) of the student's score.
    (a) 2205/8
    (b) 2225/8
    (c) 2245/8
    (d) 2425/8

20. A simple random sample with replacement is collected to estimate the average time spent on exercise by the youth in the country. The observed data points are: 1, 2, 3, 4, 5, 9. What is the unbiased estimate for variance of the time spent in exercise by youth of the country?
    (a) 4
    (b) 40/6
    (c) 8
    (d) 10

21. The random variable X follows a uniform distribution with minimum value as 10 and maximum value as 15. The random variable Y follows a uniform distribution with minimum value as 20 and maximum value as 30. Assuming X and Y to be independent, what is the expectation of the random variable $X - Y$?
    (a) 37.5
    (b) -7.5
    (c) 20
    (d) -12.5

22. The random variable X follows a uniform distribution with minimum value as 10 and maximum value as 15. The random variable Y follows a uniform distribution with minimum value as 20 and maximum value as 30. Assuming X and Y to be independent, which among the following best describes the shape of the Probability density function of X+Y ?
    (a) Rectangle
    (b) Triangle
    (c) Trapezium
    (d) Quadratic

23. Let X be a random variable following Poisson distribution with mean $\lambda = 40$. Let Y be another random variable following standard normal distribution. If Covariance of X and Y is 0.5, what is the variance of the random variable X + Y?
    (a) 40
    (b) 41
    (c) 42
    (d) 39

24. If random variables X and Y follow normal distributions independently, which of the following random variables also follow normal distribution:
    (a) $X^2 + Y^2$
    (b) Minimum $\{X, Y\}$
    (c) $3X - 4Y$
    (d) All of the above

Answer questions 25, 26 and 27 using the following data:

X	Y
2	**0**
3	**2**
4	**10**

Suppose the regression line (OLS) for the above data is given by $Y = b_0 + b_1 X$.

Then,

25. The value of b_0 is
    (a) 11
    (b) -11
    (c) -10
    (d) 10

26. The value of b_1 is
    (a) 4
    (b) 5
    (c) 6
    (d) 7

27. The MSE in the fitting is
    (a) 0
    (b) 1
    (c) 2
    (d) 2.5

28. A fitted line for some data is $Y = 10 - 0.5X$. If the sample means are $\overline{X} = 4$ and $\overline{Y}$, and this regression pass the check that the fitted line goes through $(\overline{X}, \overline{Y})$, then $\overline{Y}$
    (a) 8
    (b) 10
    (c) 6
    (d) Cannot be computed from given data

29. Suppose a regression output shows $R^2 = 0.9$ and Adjusted $R^2 = 0.4$. What does this suggest?
    (a) The model fits the data well
    (b) Too many predictors may have been added unnecessarily
    (c) All predictors are highly significant
    (d) The slope coefficients must be zero

30. Suppose exam marks (Y) increase with study hours (X), but after 8 hours of study, students become tired and marks decrease in the next 8 hours. Which regression model is best for this data?
    (a) Simple linear regression
    (b) Multiple regression
    (c) 2nd degree polynomial regression
    (d) Exponential regression

31. In multiple regression, the error term must satisfy which condition for OLS to be unbiased?
    (a) Errors have zero variance
    (b) Errors are uncorrelated with the predictors
    (c) Errors are perfectly correlated with the predictors
    (d) Errors must not be normally distributed

32. Which of the following is a drawback of fitting polynomial regression $y = a_0 + a_1x + a_2x^2 + \cdots + a_nx^n$, for m data points.
    (a) It is not linear in parameters
    (b) It cannot model nonlinear relationships
    (c) It may lead to overfitting if $n \gg m$
    (d) It cannot be estimated with least squares

33. Suppose in a regression the Total Sum of Squares (SST) = 200, and the Sum of Squares of Error (SSE) = 50. Then R^2 is
    (a) 0.25
    (b) 0.75
    (c) 0.50
    (d) 0.40

34. In a simple linear regression, the correlation between X and Y is −0.6. What is R^2?
    (a) -0.6
    (b) 0.36
    (c) 0.60
    (d) None of the above

35. Given three observations:

X1	X2	Y
1	0	6
2	1	12
3	0	14

The fitted regression model is $\hat{Y} = 4 + 3X_1 + 2X_2$ using above data. Then the residuals (errors) for each observation are respectively:

(a) $-1, 0, 1$
(b) $7, 12, 13$
(c) $-1, 12, 14$
(d) $1, 0, 1$

36. Consider the three data points $(1, 1), (2, 2)$ and $(3, 3)$ in 2-D space. If we want to project this 2-D data points into 1-D space in the direction of the principal component, then the variance of the projected data is
(a) 1
(b) $4/3$
(c) $1/2$
(d) 2

37. Consider V_1, V_2 and V_3 be the three unit eigenvectors of the covariance matrix corresponding to three distinct eigenvalues of a 3-D dataset having six observations. Let $M = [V_1 \quad V_2 \quad V_3]$ be a matrix having V_1, V_2 and V_3 as its columns. Then, which of the following is always TRUE:
(a) $\det(\mathrm{M}) = 0$
(b) $\mathrm{M}^{\mathrm{T}} = -M$
(c) $|\mathrm{Mw}| = 1$ for a unit vector $w \in \mathbb{R}^3$
(d) $(\mathrm{Mw}_1) \cdot (Mw_2) \neq 0$ for two orthogonal vectors w_1 and w_2 in $\mathbb{R}^3$

38. Consider the following two statements about Principal Component Analysis:
S1: If we select only one direction (a one-dimensional subspace) to represent the data, the sample variance of the projected points is zero if and only if the original sample points are all identical.
S2: If we select only one direction (a one-dimensional subspace) to represent the data, PCA chooses the eigenvector of the sample covariance matrix that corresponds to the least eigenvalue.
Then,
(a) Only S1 is TRUE.
(b) Only S2 is TRUE.
(c) Both S1 and S2 are TRUE.
(d) Neither S1 nor S2 is TRUE.

39. Consider a point $P(3, -3, 6)$ in 3-D space. If Q is the projection of the point P on the line having direction given by the vector $w = \left[\frac{1}{\sqrt{3}}, -\frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}\right]$, then the distance of the point Q from the origin $(0, 0, 0)$ is
(a) 3
(b) $4\sqrt{3}$
(c) 3
(d) $3\sqrt{3}$

40. Consider a 2D dataset with sample covariance matrix $\Sigma = \begin{pmatrix} 4 & 2 \\ 2 & 1 \end{pmatrix}$. Then the percentage of total variance explained by the first principal component of the data is
(a) 80%
(b) 88%
(c) 92%
(d) 100%

41. Let Σ be the sample covariance matrix of a n-dimensional dataset having n observations. Let $\Sigma = USV^T$ be the singular value decomposition of the matrix Σ. Then, which of the following is TRUE?
(a) The first principal component is given by the first row of the matrix U.
(b) The first principal component is given by the first column of the matrix U.
(c) The first principal component is given by the third row of the matrix V.
(d) The first principal component is given by the third column of the matrix V.

Consider a matrix $M = \begin{pmatrix} 1 & \alpha \\ \alpha & 1 \\ \beta & \beta \end{pmatrix}$, where $\alpha < 2$ and $\beta \leq 0$ are real constants. If the nonzero singular values (square root of the eigenvalues of $M^T M$) of the matrix M are $\sqrt{3}$ and 1. Answer Q.No. 42 to 45 based on this information:

42. The value of the constant α is
(a) 1
(b) 0
(c) -1
(d) -2

43. The value of the constant β is
(a) $-\frac{1}{2}$
(b) 0
(c) -1
(d) -2

44. If $M = USV^T$ is the singular value decomposition of the matrix M, then the matrix V is given by

(a) $\frac{1}{\sqrt{2}}\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}$

(b) $\frac{1}{\sqrt{2}}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$

(c) $\frac{1}{\sqrt{2}}\begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}$

(d) $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$

45. If $M = USV^T$ is the singular value decomposition of the matrix M, where $U = \{u_{ij}\},\ i,j = 1,2,3$, then $12(u_{11} + u_{12})^2$ is equal to

(a) 8

(b) 4

(c) 12

(d) None of these

46. Consider the following two statements about the singular value decomposition (SVD) consists of a matrix M:

(S1) SVD can be done for any matrix M.

(S2) The SVD consists of two rotations and a scaling transformation.

Then,

(a) Only S1 is TRUE.

(b) Only S2 is TRUE.

(c) Both S1 and S2 are TRUE.

(d) Neither S1 nor S2 is TRUE.

47. Consider the data set having observations $[-1, 0, 2, 3, 6]$. Suppose we apply a linear transformation to normalize (scale) the data in the interval $[0,\ 1]$. Then the mean of the original observations transforms to

(a) 1/2

(b) 3/7

(c) 2/5

(d) 4/9

48. Which of the following statements about the support (range of possible values) of Binomial and Poisson random variables is correct?

(a) Both Binomial and Poisson random variables take values only in a bounded range.

(b) Binomial random variables take values in a bounded range, while Poisson random variables take values in an unbounded range.

(c) Binomial random variables take values in an unbounded range, while Poisson random variables take values in a bounded range.

(d) Both Binomial and Poisson random variables take values in an unbounded range.