Name: **Enrolment No:**	UPES UNIVERSITY OF TOMORROW

UPES
End Semester Examination, December 2023

Course: Probability and Statistics
Semester: III
Program: B. Tech. CSE — **Time:** 3 hrs.
Course Code: CSEG 2036P — **Max. Marks:** 100

SECTION A
(5Qx4M=20Marks)

S. No.		**Marks**	**CO**
Q 1	**Discuss** covariance of random variables. **Illustrate** that for random variable X and Y, $$\text{Cov}(aX+b, cY+d) = ac \times \text{Cov}(X,Y)$$ for constants a, b, c and d.	**4**	**CO2**
Q 2	**Define** Marginal Probability Distributions. **Apply** your understanding of marginal probability mass functions to evaluate that for random variables X and Y, given the joint probability mass function, <table><tr><td></td><td>$Y=0$</td><td>$Y=1$</td><td>$Y=2$</td></tr><tr><td>$X=0$</td><td>$\frac{1}{9}$</td><td>$\frac{1}{6}$</td><td>$\frac{1}{9}$</td></tr><tr><td>$X=1$</td><td>$\frac{1}{6}$</td><td>$\frac{1}{9}$</td><td>$\frac{1}{3}$</td></tr></table> **Identify** if X and Y statistically independent.	**4**	**CO1**
Q 3	**Outline** what is meant by random variables. **Identify** c and d if we have a random variable X with the associated probability density function, $$f(x) = cx^{d-1}, 0 \le x \le 1$$ and if the second central moment $E[X^2]$ is 0.6.	**4**	**CO1, CO2**

Q 4

Define sample spaces. **Identify** the set expression as well as Venn diagram representation for the following cases:

1. At least one of the events A, B, or C occurs
2. At most two of the events A, B, or C occur.

for a sample space S and three events A, B and C.

4 | **CO1**

Q 5

Discuss correlation coefficient. **Identify** $Var(X')$, $Var(Y')$ and $r_{X'Y'}$ in terms of $Var(X)$, $Var(Y)$ and r_{XY} respectively, if X is the height of students in a class in centimeters and Y is the weight of the students in kilograms, and we undertake a transformation to height in inches (X') and weight in pounds (Y'):

$$X \to X' = 0.3937 \times X$$
$$Y \to Y' = 2.2046 \times Y$$

4 | **CO2**

SECTION B

(4Qx10M= 40 Marks)

Q 6

Choice 1: **Define** the Kruskal Wallis H Test, its null and alternate hypothesis as well as its relevant test statistic. **Describe** any one assumption relevant to this statistical test. **Highlight** how it is better than one-way ANOVA.

Apply your understanding of the Kruskal Wallis H Test for analyzing the scores of three groups of students (Group A, Group B and Group C) with

Group A	73	76	87	91
Group B	66	72	81	83
Group C	62	64	71	74

Given: The critical value for the H test for 2 degrees of freedom and $n_1 = 4$, $n_2 = 4$ and $n_3 = 4$ at $\alpha = 0.05$ is 5.692.

Choice 2: **Define** regression, principle of least squares and residuals**.** **Describe** what is meant by multiple regression model.

Apply your understanding of nonlinear regression to fit a least-square curve of the form $y = \frac{b}{x(x-a)}$ to the following data:

x	3.6	4.8	6.0	7.2	8.4	9.6	10.8
y	0.83	0.31	0.17	0.10	0.07	0.05	0.04

10 | **CO4**

Q 7	**Define** and **illustrate** the concepts of hypothesis, null hypothesis, alternative hypothesis, significance level and Type I/II errors. **Remember** and **explain** the generalized scheme for Hypothesis Testing. **Apply** and **discuss** Hypothesis Testing for Population Standard Deviations. **Illustrate** if the difference between standard deviations of samples A and B drawn from a normal distribution with population standard deviation $\sigma = 15$ is significant (for $\alpha = 0.1$) if $s_A = 8$, $s_B = 10$, $n_A = 150$ and $n_B = 200$, given z: -2, -1.9, -1.8, -1.7, -1.6, -1.5, -1.4 p: 0.04, 0.06, 0.07, 0.09, 0.11, 0.13, 0.16	**10**	**CO4**
Q 8	**Define** ANOVA and the motivation behind formulating such a statistical characterization. **Discuss** the different kinds of sums of squares, degrees of freedom and mean squares. **Identify** the relevant test statistic (ANOVA coefficient) in terms of mean squares, for ANOVA. **Apply** ANOVA ($\alpha = 0.05$) for an experiment testing three types of food on separate groups of rats over a 5-week period. The main goal is to determine if there were any significant differences in the average weight (measured in grams) of the rats each week across the three food groups. The data is Food 1 / Food 2 / Food 3: 8 / 4 / 11 12 / 5 / 8 19 / 4 / 7 8 / 6 / 13 6 / 9 / 7 11 / 7 / 9 <u>Given</u>: F-statistic for degrees of freedom (2,15) at $\alpha = 0.05$ is 3.68.	**10**	**CO4**
Q 9	**Define** a Decision Tree. **Describe** what is node purity and **highlight** one advantage and one disadvantage of using Decision Trees.	**10**	**CO5**

Table for Q 7:

z	-2	-1.9	-1.8	-1.7	-1.6	-1.5	-1.4
p	0.04	0.06	0.07	0.09	0.11	0.13	0.16

Table for Q 8:

Food 1	Food 2	Food 3
8	4	11
12	5	8
19	4	7
8	6	13
6	9	7
11	7	9

Expand on the two ways in which Decision Trees can have variable selection criterion for node allocation.

Apply your understanding of the *Gini index approach* for Decision Trees to analyze 15 students' performance in an online exam. The predictors for this data-set encompass details such as whether the student is enrolled in other online courses, their academic background and whether they are currently employed or not.

S.No.	Target Variable	Predictor Variables		
	Result	Other Online Courses	Student Background	Working Status
1.	Pass	Yes	Mathematics	Not Working
2.	Fail	No	Mathematics	Working
3.	Fail	Yes	Mathematics	Working
4.	Pass	Yes	CS	Not Working
5.	Fail	No	Other	Working
6.	Fail	Yes	Other	Working
7.	Pass	Yes	Mathematics	Not Working
8.	Pass	Yes	CS	Not Working
9.	Pass	No	Mathematics	Working
10.	Pass	No	CS	Working
11.	Pass	Yes	CS	Working
12.	Pass	No	Mathematics	Not Working
13.	Fail	Yes	Other	Working
14.	Fail	No	Other	Not Working
15.	Fail	No	Mathematics	Working

SECTION-C
(2Qx20M=40 Marks)

Q 10			
Q 10	**Define** a Poisson random variable $X \sim \mathrm{Po}(\lambda)$ and **highlight** the expression for its probability distribution. **Derive** the mean and variance of the Poisson distribution, considering $e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}$. **Show** that this probability distribution satisfies the properties of probabilities. **Define** a Gamma Function and **highlight** any two properties of the Gamma Function. **Expand** on your understanding of the Gamma Distribution $Y \sim \mathrm{Gamma}(\alpha, \beta)$, with the expression for its probability distribution. **Derive** the mean and variance of the Gamma Distribution.	**20**	**CO3**

	Identify the values of $\Gamma(4), \Gamma\left(\frac{7}{2}\right)$ and $\Gamma(-3)$. **Apply** your understanding of cumulative distribution functions to show that $P(Y \leq \lambda) = P(X \geq \alpha)$ for $X \sim \text{Po}(\lambda)$ and $Y \sim \text{Gamma}(\alpha, \beta)$, given that the cumulative distribution function for the Poisson distribution is $$F(x, \lambda) = \sum_{k=0}^{x} \frac{e^{-\lambda}\lambda^{x}}{k!}$$ and we take $\alpha = 2, \beta = 1$.		
Q 11	**Define** a normal distribution and a standard normal table. **Derive** the points of inflection of a normal distribution. **Calculate** the probability that a randomly selected student from UPES has IQ lesser than 70, given that the IQ scores of the students of UPES follow a normal distribution with a mean (μ) of 100 and a standard deviation (σ) of 15. Given: The following segment of the standard normal table <table><tr><td>z</td><td>-3</td><td>-2</td><td>-1</td><td>0</td><td>1</td><td>2</td><td>3</td></tr><tr><td>Value</td><td>0.001</td><td>0.023</td><td>0.159</td><td>0.5</td><td>0.841</td><td>0.977</td><td>0.999</td></tr></table> **Determine** all IQ scores that comprise the top 10% of the class, given that the z-score corresponding to $z \approx 1.3$ is 0.9. **Discuss** sample statistics and **describe** the Method of Moments (MoM). **Highlight** any two properties of a good estimator in sample statistics. *Choice 1:* **Identify** the MoM estimator of the population parameters for n independent and identically distributed samples taken from a Gamma distribution. *Choice 2:* **Calculate** the probability that the sample mean height of these students (for a sample of 25 students taken from the distribution mentioned above) is greater than 106. Given: $p_{\alpha=0.05}(z = 2) = 0.977$.	**20**	**CO3**