PARVATHANENI BRAHMAYYA
SIDDHARTHA COLLEGE OF ARTS & SCIENCE
Autonomous
Siddhartha Nagar, Vijayawada–520010
Re-accredited at 'A+' by the NAAC

Course Code				**23STMIL232**			
Title of the Course				**Random Variables and Statistical Methods**			
Offered to:				**B.Sc. Honours Data Analytics**			
L	**4**	**T**	**0**	**P**	**0**	**C**	**3**
Year of Introduction:	**2024-25**			**Semester:**		**3**	
Course Category:	**MINOR**			**Course Relates to:**	**Local, Regional, National, Global**		
Year of Revision	**NA**			**Percentage:**	**NA**		
Type of the Course:				**SKILL DEVELOPMENT**			
Crosscutting Issues of the Course :				**NA**			
Pre-requisites, if any				**23STMIL121**			

Course Description: Course Description:
This course provides an in-depth exploration of statistical methods essential for analysing the data. Key topics include random variables, mathematical expectations, generating functions, and correlation and regression analysis. Students will learn to apply both theoretical and practical statistical tools, with a focus on problem-solving using software like Excel. By the end of the course, students will be equipped to handle complex data sets, perform in-depth statistical analyses, and interpret results to support decision-making in cognitive computing environments.

S. No	COURSE OBJECTIVES
1	Understand the distinction between discrete and continuous random variables
2	Calculate joint probability mass functions (PMFs) and joint probability density functions (PDFs).
3	Understand the concept of variance and its relationship to expected value.
4	Understand the concept of generating functions: Students should be able to define and interpret generating functions, recognizing their role in representing sequences and distributions.
5	Interpret correlation coefficients to assess the strength and direction of relationships and fit the simple linear regression model to real-time data.

Course Outcomes

At the end of the course, the student will be able to…

NO	COURSE OUTCOME	BTL	PO	PSO
CO1	Understand and apply concepts of univariate and bivariate random variables, including probability functions and distribution properties.	K2	1	1
CO2	Analyze the relationships between variables through joint, marginal, and conditional distributions, using Excel for practical exercises.	K4	1	1
CO3	Apply mathematical expectations, variance, and covariance in solving statistical problems, utilizing Excel for visualization and computations.	K3	1	1
CO4	Demonstrate the use of generating functions in deriving properties of distributions and apply Chebyshev's inequality to solve probability problems.	K2	1	1
CO5	Interpret and calculate correlation and regression coefficients, using Excel to visualize and analyze real-world data sets.	K5	1	1

For BTL: K1: Remember; K2: Understand; K3: Apply; K4: Analyze; K5: Evaluate; K6: Create

CO-PO-PSO MATRIX									
CO NO	**PO1**	**PO2**	**PO3**	**PO4**	**PO5**	**PO6**	**PO7**	**PSO1**	**PSO2**
CO1	**2**							**2**	
CO2	**3**							**3**	
CO3	**3**							**3**	
CO4	**2**							**2**	
CO5	**3**							**3**	

Use the codes 3, 2, 1 for High, Moderate and Low correlation Between CO-PO-PSO respectively.

Syllabus

Course Details

Unit-I Random Variables Univariate (12hours)

Introduction – Discrete random variable – probability mass function- Problems. Continuous random variables - Probability density function – problems. Distribution function – Properties of distribution function- discrete & continuous distribution function - Problems. Transformation of one-dimensional random variable- problems.

Case Studies

Customer Arrival Times at a Restaurant

1. Problem: A restaurant wants to estimate the number of customers that will arrive during a particular hour.
2. Random Variable: The number of customers arriving in an hour.
3. Distribution: Poisson distribution, as the events (customer arrivals) are independent and occur at a constant rate.
4. Application: The Poisson distribution can help the restaurant estimate the average number of customers per hour and the probability of different numbers of customers arriving. This information can be used for staffing purposes, inventory management, and other operational decisions.

Exercises/Projects:

1. Exercise1: Coin Tossing
2. Problem: A fair coin is tossed 5 times. Let X be the number of heads obtained.
3. Find the probability mass function of X.
4. Calculate the expected value of X.
5. Calculate the variance of X.

Unit-II Bivariate Random variables (12-hours)

Two-dimensional random variables-Definition-Discrete and Continuous bi- variate random variables- Joint, Marginal and Conditional distributions and their properties- Problems. Two-dimensional distribution function- Joint, Marginal and Conditional distribution functions. Independence of random variables.

Case studies:

1. Case Study: Joint Distribution of Height and Weight
2. Problem: A researcher is interested in studying the relationship between height and weight in a population.
3. Bivariate Random Variables: X = height, Y = weight.
4. Joint Probability Distribution: The researcher can use a joint probability distribution to model the probability of observing a particular height and weight combination. This could be a bivariate normal distribution, for example, if the heights and weights are assumed to be normally distributed and correlated.
5. Applications: The joint distribution can be used to calculate the correlation between height and weight, predict one variable based on the other, and analyze the joint distribution of other related variables (e.g., BMI).

Exercises/Projects:

1. Project1: Simulating and Visualizing Bivariate Normal Distributions
2. Goal: To explore the effects of mean, variance, and correlation on bivariate normal
3. Distributions through simulation and visualization.
4. Steps: Generate random samples: Use a random number generator to generate samples from a bivariate normal distribution with specified mean vector and covariance matrix.
5. Visualize the data: Create scatter plots, histograms, and density plots to visualize the joint distribution of the two variables.

6. Vary parameters: Experiment with different values for the mean vector and covariance matrix to observe how the distribution changes.
7. Analyze the effects: Analyse the impact of changes in mean, variance, and correlation on the shape, location, and spread of the distribution.

Unit-III Mathematical Expectations (12hours)

Introduction- Mathematical expectation of a random variable-expected of function of a random variable. Properties of expectations. Addition and Multiplication theorems on expectation. Variance – Properties of variance and covariance- properties of covariance . Inequalities involving expectation- Cauchy – Schwartz.

Case studies:

1. Expected Return on Investment
2. Problem: An investor is considering investing in a stock. They want to estimate the expected return on their investment.
3. Mathematical Expectation: The expected return is calculated as the weighted average of the possible returns, where the weights are the probabilities of each return occurring.
4. Application: The investor can use historical data or market analysis to estimate the probability distribution of the stock's returns. By calculating the expected return, they can assess whether the investment aligns with their risk tolerance and expected returns.

Exercises/Projects:

1. Predicting Insurance Claims: A Project on Mathematical Expectations

   Project Objective:
2. To estimate the expected cost of insurance claims based on historical data, using statistical and machine learning techniques.

   Data Requirements:
3. Historical claim data: This should include information on the number of claims, claim severity, and relevant risk factors (e.g., age, location, policy type).
4. Policyholder information: Data on policyholders' characteristics (e.g., age, gender, occupation) can be used to identify risk factors.

   Methodology:

   i. Data cleaning and pre-processing: Clean and prepare the data by handling missing values, outliers, and inconsistencies.
   ii. Exploratory data analysis: Explore the data to identify patterns, correlations, and relationships between variables.
   iii. Feature engineering: Create new features or transform existing features to improve predictive power.
   iv. Model selection: Choose appropriate statistical or machine learning models based on the nature of the data and the desired outcome. Some potential models include:
   v. Linear regression: Suitable for predicting continuous variables like claim costs.
   vi. Generalized linear models (GLMs): Can handle different types of response variables (e.g., count data, binary outcomes).
   vii. Decision trees and random forests: Can handle both continuous and categorical variables and can capture non-linear relationships.
   viii. Gradient boosting machines (GBMs): Powerful ensemble models that can handle complex relationships.

ix. Model training and evaluation: Train the selected models on the training data and evaluate their performance using appropriate metrics (e.g., mean squared error, mean absolute error, R-squared).
x. Model deployment: Deploy the best-performing model to predict future claim costs.

Unit-IV Generating functions. (12hours)

Definition of Moment Generating Function (m.g.f.)- properties- problems, Definition of Cumulate Generating Function (c.g.f.)- properties, Definition of Probability Generating Function (p.g.f.)- Properties , Definition of Characteristic Function (c.f.) – properties- and Chebychev's Inequality- Problems

Examples

1. Finding moments: As mentioned above, the MGF can be used to find the mean, variance, and higher-order moments of a distribution.
2. Identifying distributions: If you know the MGF of a distribution, you can often identify the distribution itself.
3. Deriving new distributions: MGFs can be used to derive new distributions from existing ones.
4. Calculating probabilities: In some cases, probabilities can be calculated using the MGF.

Exercises/Projects:

Exercise 1: Geometric Distribution
Problem: Find the moment generating function of a geometric distribution with parameter p.
Exercise 2: Binomial Distribution
Problem: Find the moment generating function of a binomial distribution with parameters n and p.
Exercise 3: Negative Binomial Distribution
Problem: Find the moment generating function of a negative binomial distribution with Parameters r and p.
Exercise 4: Poisson distribution
Problem: Find the moment generating function of a Poisson distribution with parameter λ. (You've already solved this one in a previous question.)
Exercise 5: Sum of Independent Random Variables
Problem: If X and Y are independent random variables with moment generating functions M_X(t) and M_Y(t), respectively, find the moment generating function of X + Y.

Unit -V Correlation &Regression (12hours)

Introduction- Meaning correlation- scatter diagram, Karl Pearson's coefficient of correlation, limits for correlation coefficient- Calculation of the Correlation Coefficient for a Bivariate Frequency Distribution and spearman rank correlation. Definition, types of regression lines and properties of regression coefficients.

Case Study

Impact of Advertising on Sales

i. Problem: A marketing team wants to assess the impact of advertising spending on sales. Correlation and Regression:
ii. Correlation: The team can calculate the correlation between advertising spending and sales to determine if there is a relationship.
iii. Regression: A linear regression model can be used to predict sales based on advertising spending.

Exercises/Projects:

i. A Project on Correlation and Regression
ii. Project Objective: To analyze real-world datasets to identify relationships between variables using correlation and regression techniques.

Data Selection:

i. Choose relevant datasets: Consider datasets from fields like economics, finance, social sciences, or environmental science.
ii. Ensure data quality: Verify the accuracy, completeness, and consistency of the data.
Data Exploration:
iii. Descriptive statistics: Calculate summary statistics (mean, median, mode, standard deviation) for each variable.
iv. Data visualization: Create visualizations (e.g., histograms, scatter plots, box plots) to explore the distribution of variables and identify potential relationships.

Correlation Analysis:

i. Calculate correlation coefficients: Compute Pearson, Spearman, or Kendall correlation coefficients to measure the strength and direction of the linear or monotonic relationship between variables.
ii. Interpret the results: Analyze the magnitude and sign of the correlation coefficients to understand the strength and direction of the relationship.

Regression Analysis:

i. Build regression models: Create linear or multiple regression models to predict one variable based on others.
ii. Evaluate model performance: Assess the accuracy of the models using metrics like R-squared, mean squared error, and mean absolute error.
iii. Interpret the results: Analyze the coefficients of the regression models to understand the impact of each predictor variable on the response variable.

Textbook:

Fundamentals of Mathematical Statistics, 11th Edition, 2010, S. C. Gupta and V. K. Kapoor, Sultan Chand & Sons, New Delhi

Reference Books:

1. Mathematical Statistics with Applications, 2009, K.M.Ramachandran and Chris P.Tsokos Academic Press(Elsevier), Haryana .
2. Probability and Statistics, Volume I & II, D. Biswas, New central book Agency (P) Ltd, New Delhi.

PARVATHANENI BRAHMAYYA
SIDDHARTHA COLLEGE OF ARTS & SCIENCE
Autonomous
Siddhartha Nagar, Vijayawada–520010
Re-accredited at 'A+' by the NAAC

23STMIL232: Random Variables and Statistical Methods.

Minor-2 B.Sc. Honours Data Analytics Semester III
Time: 3 hours Maximum Marks: 70

Section - A

Answer the following questions. **5 X 4M = 20M**

1. a. Define the Random variables and state its types. (CO-1, K-1)
(OR)
b. Define distribution function and state its properties (CO-1, K -1)

2. a. Explain the concepts marginal and conditional probability distribution(CO-2, K-2)
(OR)
b. Explain the concept of joint probability function and state its types.(CO-2, K-2)

3. a. Show that the mathematical expectation of the sum of two random variables is equal to the sum of their individual expectations. (CO-3, K-2)
(OR)
b. State and prove multiplication theorem of two events. (CO-3, K-2)

4. a. Define moment generating function. (CO-4, K-1)
(OR)
b. Define cumulate generating function. (CO-4, K-1)

5. a. Explain correlation and its types. (CO-5, K-2)
(OR)
b. Explain scatter diagram method.(CO-5, K-2)

Section - B

Answer the following questions. **5 X 10M = 50M**

6. a. A random variable has the following probability distribution

x	0	1	2	3	4	5	6	7	8
P(X= x)	a	3a	5a	7a	9a	11a	13a	15a	17a

(i) Determine 'a' (ii) Find $P(X<3)$, $P(X \geq 3)$ and $P(0<X<5)$
(iii) Find the distribution function of X. (CO-1, K-3)
(OR)

b. The diameter of an electric cable, say X, is assumed to be a continuous random variable with p.d.f. $f(x) = 6x(1-x),\ 0 \leq x \leq 1$.
i) Check that f(x) is p.d.f.,
ii) Determine a number **b** such that P(X <**b**) = P(X >**b**).(CO-1, K-3)

7. a. Two discrete random variables X and Y have the joint probability density function:

$P_{XY}(x, y) = \dfrac{\lambda^x e^{-\lambda} p^y (1-p)^{x-y}}{y!(x-y)!}$, y = 0, 1, 2,…,x ; x = 0, 1, 2,…, where λ, p are constants with $\lambda > 0$ and 0 < p < 1.The conditional distribution of Y given X and X given Y. (CO-2, K-3)

(OR)

b. If X and Y are two random variable having joint probability function:

$$f(x, y) = \begin{cases} \frac{1}{8}(6 - x - y);\ 0 \le x < 2,\ 2 \le y < 4 \\ 0 \quad\quad otherwise \end{cases}$$

Find (i) $P(X < 1 \cap Y < 3)$ (ii) $P(X + y < 3)$ (iii) $P(X < 1 | Y < 3)$ (CO-2, K-3)

8. a. State and prove the Linear combination of Random variables (CO-3, K-5)

(OR)

b. State and prove Cauchy Schwartz inequality. (CO-3, K-5)

9. a. Explain moment generating function and characteristic function (CO-4, K-3)

(OR)

b. Use Chebyshev's inequality to determine how many times a fair coin must be tossed in Order that the probability will be at least 0.90 that the ratio of the observed number of Heads to the number of tosses will lie between 0.4 and 0.6. (CO-4, K-3)

10a. Calculate the correlation coefficient for the following data (CO-5, K-3)

X	65	66	67	67	68	69	70	72
Y	67	68	65	68	72	72	69	71

(OR)

b. Derive the two regression lines X on Y and Yon X.

(Co-5, K-3)

***