Semester: 1				
Programme: M.Sc. Data Science				
Course: Statistical Foundations of Data Science				
Paper code: MDTS4111			**Credits: 6**	
Hours/week: 7 (4Th + 3Pr)				
Category: Core/MDC/SEC/VAC: Core				
Theory / Practical / Composite: Composite				
No of Module: 2				
Course Outcome:				
1. **Remember** the fundamental definitions of population and sample, various scales of measurement, and the basic terminology of probability and sampling distributions.				
2. **Understand** the conceptual logic of the Theorem of Total Probability and Bayes Theorem, the structural implications of the Central Limit Theorem, and the notion of Big Data.				
3. **Apply** measures of central tendency, dispersion, skewness, and kurtosis, along with probability laws, to calculate parameters and summarize univariate data.				
4. **Analyze** bivariate data relationships using scatter plots, correlation, and regression techniques while identifying patterns of association through contingency tables and odds ratios.				
5. **Evaluate** the fit of empirical datasets to theoretical probability distributions using QQ plots and assess the significance of sampling fluctuations using Chi-square, t, and F distributions.				
6. **Create** robust statistical models and data simulations by synthesizing probability distribution properties, moment generating functions, and Monte-Carlo simulation techniques to predict data behaviour.				
Prerequisites: ***Basic knowledge about any prior course***				
SYLLABUS				
Module/Unit	**CONTENT**	**HOURS or NUMBER OF CLASSES**	**CO Mapping**	**COGNITIVE LEVEL**
Module I / Unit 1	***Analysis of Univariate Data:*** Concept of population and sample. Classification of data. Scales of measurement. Concepts of moments and quantiles. Concepts, measures & applications of central tendency, dispersion, skewness and Kurtosis. Summary tables and graphs. Notion of Big Data.	14	CO1, CO2, CO3	K1, K2, K3
Module I / Unit 2	***Analysis of Bivariate Data:*** Scatter plot. Concept of correlation and Regression. Contingency table. Notion of independence and association. Odds ratio	12	CO4	K4

	and relative risk. Chi-square and Kendall's measures.			
Module II / Unit 1	***Probability:*** Random experiment, outcomes, sample space, events. Classical, statistical and axiomatic definitions of probability. Subjective probability. Poincaré's theorem. Boole's and Bonferroni's inequality. Conditional Probability, multiplication law of probability, and independent events. Theorem of total probability. Bayes theorem. (Statement and applications only).	8	CO1, CO2, CO3	K1, K2, K3
Module II / Unit 2	***Probability Distributions:*** Random variables. PMF, PDF, CDF (graphs and properties). Empirical distribution function. Moments and quantiles. Moment generating functions. Statement of properties & applications of Discrete Uniform, Hypergeometric, Binomial, Poisson, Geometric, Rectangular, Normal, Exponential, Beta and Gamma distributions. Bivariate normal distribution. QQ Plots.	8	CO5, CO6	K5, K6
Module II / Unit 3	***Sampling Distributions:*** Concept of i.i.d random variables, parameters & statistics. Sampling fluctuations and sampling distributions of statistics. Chi-square, t, and F distributions (with applications in data analysis). Sampling distribution of the sample mean and variance & their independence under the normality assumption. Central Limit Theorem and its implications in data science. Concept of Monte-Carlo Simulation and Applications	10	CO1, CO2, CO5, CO6	K1, K2, K5, K6
Practical Using Excel	**Lab Using Spreadsheet: Some suggested topics** • Data entry, cleaning, and pre-processing in spreadsheets • Descriptive statistics using built-in functions • Data visualization techniques for exploratory analysis • Working with formulas and functions (IF, VLOOKUP/XLOOKUP, COUNTIF, SUMIF) for data manipulation and insights • Basic probability and statistical analysis (correlation, simple regression using Excel tools) Introduction to data-driven decision-making using pivot tables, filters, and summary reports			**39**

Text Books

1. Goon, A. M., Gupta, M. K., and Dasgupta, B. (2002). *Fundamentals of Statistics*, Vols. I & II (8th ed.). Kolkata: The World Press.
2. Ross, S. M. (2014). *A First Course in Probability* (9th ed.). Pearson Education.

3. Rohatgi, V. K., and Saleh, A. K. Md. E. (2009). *An Introduction to Probability and Statistics* (2nd ed., reprint). John Wiley & Sons.
4. Hogg, R. V., Tanis, E. A., and Rao, J. M. (2009). *Probability and Statistical Inference* (7th ed.). New Delhi: Pearson Education.
5. Stanton, J. M. (2013). *Introduction to Data Science*. Syracuse University.
6. Peng, R. D., & Matsui, E. (2015). *The Art of Data Science: A Guide for Anyone Who Works with Data*. Leanpub.
7. Chung, K. L. (2001). *Elementary Probability Theory with Stochastic Processes* (3rd ed.). Springer.
8. Moore, D. S., McCabe, G. P., & Craig, B. A. (2014). *Introduction to the Practice of Statistics* (7th ed.). Macmillan Higher Education.

Marks	**Theory** **CIA:** 10 **End Sem Exam:** 25+25 **Total:** 60	**Practical** **Continuous Assessment:** 40
Paper Structure for Theory Semester Exam	Short questions: 5 marks each	Long Questions: 10 Marks each
Module I	1 out of 2	2 out of 3
Module II	1 out of 2	2 out of 3

Course outcomes (COs) and Cognitive Level Mapping

COs	**CO Description**	**Cognitive levels**
CO1	**Remember** the fundamental definitions of population and sample, various scales of measurement, and the basic terminology of probability and sampling distributions.	K1
CO2	**Understand** the conceptual logic of the Theorem of Total Probability and Bayes Theorem, the structural implications of the Central Limit Theorem, and the notion of Big Data.	K2
CO3	**Apply** measures of central tendency, dispersion, skewness, and kurtosis, along with probability laws, to calculate parameters and summarize univariate data.	K3
CO4	**Analyze** bivariate data relationships using scatter plots, correlation, and regression techniques while identifying patterns of association through contingency tables and odds ratios.	K4
CO5	**Evaluate** the fit of empirical datasets to theoretical probability distributions using QQ plots and assess the significance of sampling fluctuations using Chi-square, t, and F distributions.	K5
CO6	**Create** robust statistical models and data simulations by synthesizing probability distribution properties, moment generating functions, and Monte-Carlo simulation techniques to predict data behaviour.	K6