SYLLABUS OF

Statistics Core Courses

B. Sc APPLIED STATISTICS WITH DATA SCIENCE

{Programme sanctioned as per order No. GO (Ms) No. 150/2021 dated 25/2/21}

CBCSS-UG 2019 Regulation

Course Pattern

Core	: Applied Statistics
Complementary 1	: Mathematics for Applied Statistics (Mathematics for integrated M.Sc. Statistics programme)
Complementary 2	: Data Science

Course Structure

Semester I				
Course Code	Course	Course Title	Hours/ Week	Credits
A 01	Common Course 1	English Course -I	5	4
A 02	Common Course 2	English Course -II	4	3
A 07	Common Course 3	Language other than English-I	4	4
ASD1 B01	**Core Course 1**	**Statistics – I Descriptive Statistics & Basic R Programming**	**4**	**4**
MST1 C01	1st Complementary Course -1	Mathematics –1	4	3
BCSD S1 C01	2nd Complementary Course -1	Computer Science -1	4	3
	Audit Course			
TOTAL				**21**

Semester 2				
Course Code	Course	Course Title	Hours/ Week	Credits
A 03	Common Course 4	English Course -III	5	4
A 04	Common Course 5	English Course -IV	4	3
A 08	Common Course 6	Language other than English- II	4	4
ASD2 B02	**Core Course 2**	**Statistics – II Probability and Mathematical Expectation**	**4**	**4**
MST2 C02	1st Complementary Course -2	Mathematics -2	4	3
BCSDS2 C02	2nd Complementary Course -2	Computer Science -2	4	3
	Audit Course			
TOTAL				**21**

Semester 3				
Course Code	Course	Course Title	Hours/ Week	Credits
A05	Common Course 7	English Course –V	5	4
A09	Common Course 8	Language other than English- III	5	4
ASD3 B03	**Core Course 3**	**Statistics – III Probability Distributions**	**5**	**4**
MST3 C03	1st Complementary Course -3	Mathematics -3	5	3
BCS DS3 C03	2nd Complementary Course -3	Computer Science -3	5	3
	Audit Course			
TOTAL				**18**

Semester 4				
Course Code	**Course**	**Course Title**	**Hours /Week**	**Credits**
A06	Common Course 9	English Course -VI	5	4
A10	Common Course 10	Language other than English- IV	5	4
ASD4 B04	Core Course 4	Statistics – IV **Statistical Inference**	5	4
MST4 C04	1st Complementary Course -4	Mathematics -4	5	3
BCS DS4 C04	2nd Complementary Course -4	Computer Science -4	5	3
	Audit Course			
TOTAL				**18**

Semester 5				
Course Code	**Course**	**Course Title**	**Hours /Week**	**Credits**
ASD5 D01 ASD5 D02 ASD5 D02	ECONOMIC STATISTICS QUALITY CONTROL BASIC STATISTICS	Open Course {Offered to Other Departments of the College}	3	3
ASD5 B05	Core Course 5	Statistics –V **Sampling Theory and Design of Experiments**	6	5
ASD5 B06	Core Course 6	Statistics – VI **Multivariate Techniques for Data Science**	6	5
ASD5 B07	Core Course 7	Statistics – VII **Regression Analysis**	5	4
ASD5 B08	Core Course 8	Statistics -VIII **Introduction to Data Science and Statistical Machine Learning**	5	4
TOTAL				**21**

Semester 6				
Course Code	**Course**	**Course Title**	**Hours /Week**	**Credits**
ASD6 B09	Core Course 9	Statistics – IX **Time Series and Index Numbers**	5	4
ASD6 B10	Core Course 10	Statistics – X **Statistical Quality Control, Actuarial Statistics, Linear Programming and Demography**	5	4
ASD6 B11	Core Course 11	Statistics -XI **Data Mining Techniques**	5	5
ASD6 B12	Core Course 12	Statistics -XII **Statistical Computing using R or Python (Practical)**	5	4
ASD6 B13	Core Course 13	Statistics -XIII Project/ Internship in Industry	2	2
ASD6 B14(E)	Core 14 Elective	**Data Visualization Techniques**	3	2
ASD6 B15(E)	Core 15 Elective	**Operations Research**		
ASD6 B16(E)	Core 16 Elective	**Mathematical Economics**		
TOTAL				**21**

SEMESTER I

Statistics –1

ASD1 B01 - DESCRIPTIVE STATISTICS AND BASIC R PROGRAMMING

Teaching Hrs. 04 Credit : 04

Expected Course Outcome:-

1. ***Students will get an awareness on national and international official statistical systems and its functions***
2. ***Students will be capable of collecting and presenting data in different formats***
3. ***Students will be able to identify suitable measures of central values, dispersions and symmetry and will be able to interpret the nature of data based on it.***
4. ***Students will be able to perform these computations using R programme***

Module 1: **National and International Statistical systems**. Role, function and activities of Central Statistical organizations like MOSPI; CSO, NSSO; State Departments like DES; Organization of large-scale sample surveys. General and special data dissemination systems. Scope and Contents of population census of India. National Statistical Commission: Need, Constitution, its role, functions etc., International Statistical Organizations under UN like UNSD, UNFPA and UNDP. Careers in Statistical organizations **[10 hrs]**

Module 2: - Overview of the R Programming- Operators, Printing Values, Basic Data Types, Control Structures, Functions, Packages, Running R Code , Reading Data, Text Files, Cleaning Up Data, Identifying Data Types, **Data Visualization ,** Basic Visualizations, Scatterplots, Visualizing Aggregate Values with Bar plots and Pie charts , Common Plotting Tasks , Layered Visualizations Using ggplot2, Creating Plots Using qplot() , ggplot(): Specifying the Grammar of the Visualization , Themes., **Exploratory Data Analysis:-** Summary Statistics , Dataset Size, Summarizing the Data Ordering Data by a Variable, Group and Split Data by a Variable , Box Plots , Histograms, Measuring Data Symmetry, Skewness and Kurtosis.

[25 hrs]

Module 3: **Data Collection and Presentation:** - Origin and meaning of Statistics, Importance and Scope of Statistics, limitations and misuses of Statistics, Concepts of statistical population and sample. Census and sampling. Primary and secondary, data. Different types of data – quantitative, qualitative, geographical and chronological. Continuous and discrete data. Nominal, Ordinal Interval and Ratio Scales, Methods of collection of primary data. Designing of a questionnaire and schedule. Classification and tabulation of data. Frequency distributions. Graphical presentation, Frequency curve, Histogram and Stem and leaf chart.

[15 hrs]

Module 4: - **Measures of Central Tendency and Dispersion** – Arithmetic Mean, Geometric Mean, Harmonic Mean, Median and Mode; Partition Values – Quartiles, Deciles and Percentiles. Dispersion ; Mean deviation, Quartile deviation, Variance, Standard deviation, Coefficient of variation; Moments- raw and central moments, relation between central and raw moments; skewness, Kurtosis, Box plot. **[14 hrs]**

Books for Study.

1. Gupta S. C. and Kapoor, V. K. (2014): *Fundamentals of Mathematical Statistics*, Sultan Chand & Co.
2. Manas A. Pathak (2014) Beginning Data Science with R, Springer Cham Heidelberg NewYork Dordrecht London

Books for References

1. Goon A. M., Gupta M. K., Das Gupta. B. (1999): *Fundamentals of Statistics*, Vol. I, World Press, Calcutta.
2. B L Agrawal (2013): Basic Statistics – New Age International Publishers.
3. Official publications of MOSPI, UNDP, CSO etc.
4. Chirag Sha (2020) A Hands-On Introduction to Data Science, Cambridge University press, University Printing House, Cambridge CB2 8BS, United Kingdom
5. Roger D. Peng(2015), R Programming for Data Science, Leanpub book.

Semester II

Statistics – 2

ASD2 B02 : PROBABILITY AND MATHEMATICAL EXPECTATION

Teaching Hrs. 04 Credit : 04

Expected Course Outcome:-

1. ***Students will be able to define probability and to apply the concept in solving real life problems***
2. ***Students can define random variables and probability functions related with random variables***
3. ***Students will be able to define expected value and to define generating functions***
4. ***To explain the concept of correlation and regression and to apply these concepts.***

Module 1: **Basic Probability** :- Random experiments, Sample space, Event, Classical definition of probability, Statistical regularity, Field, Sigma field, Axiomatic definition of probability and simple properties, Addition theorem, Conditional probability of two events, Multiplication theorem, Independence of events-pair wise and mutual, Bayes theorem. **[16 hrs]**

Module 2: **Random variable** Discrete and continuous random variables, functions of random variables. Probability mass function and probability density function with illustrations. Distribution function and its properties. Transformation of random variables. **[16 hrs]**

Module 3: -Bivariate Random Variables- Bivariate distribution and statement of its properties. Joint, marginal and conditional distributions. Independence of random variables. Transformation of bivariate random variables. **[12 hrs]**

Module 4: **Mathematical Expectation and Generating Functions** -Definition and properties, mean and variance of a random variable. Addition and multiplication theorems on expectation. Raw and central moments.. Mode and median of discrete and continuous random Variables. Covariance and correlation coefficient. Cauchy- Schwartz's inequality. Conditional expectation (regression function) and conditional variance. **Generating Functions** - Probability generating function, moment generating function, cumulant generating function, characteristic function and their Properties. Methods of Computing Mean and Variance from the moment generating function and Characteristic function with examples. **[20 hrs]**

Books for Study

1. Gupta S. C. and Kapoor, V. K. (2014): *Fundamentals of Mathematical Statistics*, Sultan Chand & Co.
2. John E Freund (2013), Mathematical Statistics (8th edn), Pearson Education, NewDelhi

Books for References

1. Mood A. M., Gray bill F. A., Boes D C (2017): Introduction to the theory of statistics -
   Tata Magrow Hill.
2. Goon A. M., Gupta M. K., Das Gupta. B. (1999): *Fundamentals of Statistics*, Vol. I, World Press, Calcutta.
3. Rohatgi, VK and Saleh AK (2015) An introduction to Probability and Statistics (3rd Edition) John Wiley and Son Inc

Semester III

Statistics – 3

ASD3 B03 PROBABILITY DISTRIBUTIONS

Teaching Hrs. 05 Credit : 04

Expected Course Outcome:-

1. ***Equip the students to use discrete standard distributions to explain random phenomena***
2. ***To equip the students to interpret probabilistic nature of data using continuous probability distributions.***
3. ***Apply laws of large numbers and central limit theorems in solving problems***
4. ***Students will be aware on sampling distributions and its properties***

Module 1: **Discrete Distributions :-** Uniform, Bernoulli, Binomial, Geometric, Poisson - mean, variance, m.g.f, their properties - Fitting of Binomial and Poisson, memory less property of Geometric distribution

[20 hrs]

Module 2: **Continuous Distributions** - Uniform, Beta (two types), Exponential, Gamma - mean, variance, m.g.f, characteristic function, their properties - memory less property of exponential distribution, Normal Distribution - Properties, fitting of normal distribution, use of standard normal tables for various probability computation. Bivariate normal- marginal and conditional distributions

[24 hrs]

Module 3: **Law of Large Numbers and Limit Theorems**- Chebyshevs inequality, convergence in probability, Chebyshevs and Bernoulli's forms of weak law of large numbers, Lindberg-Levy form of Central Limit Theorem -Normal distribution as a limiting case of binomial and Poisson under suitable assumptions.

[16 hrs]

Module 4: **Sampling Distributions** - Concept of random sample and statistic, sampling distribution of a statistic, standard error, sampling distributions of the mean and variance of a random sample arising from a normal population. chi-square, t and F distributions- derivations, properties, uses and inter relationships

[20 hrs]

Books for Study

1. Gupta S. C. and Kapoor, V. K. (2014): *Fundamentals of Mathematical Statistics*, Sultan Chand & Co
2. John E Freund (2013), Mathematical Statistics (8th edn), Pearson Edn, NewDelhi

Books for References

1. Mood A. M., Gray bill F. A., Boes D C (2017): Introduction to the theory of statistics –
   Tata Magrow Hill.
2. Goon A. M., Gupta M. K., Das Gupta. B. (1999): *Fundamentals of Statistics*, Vol. I, World Press, Calcutta.
3. Rohatgi, VK and Saleh AK (2015) An introduction to Probability and Statistics (3^{rd} Edition) John Wiley and Son Inc

Semester IV

Statistics – 4

ASD4 B04 : STATISTICAL INFERENCE

Teaching Hrs. 05 Credit : 04

Expected Course Outcome:-

1. ***Enable students to find point and interval estimate the population parameters in different contexts***
2. ***Students will build an understanding on different statistical hypothesis and its formulation***
3. ***Students can apply these knowledge to test small sample and large sample problems***
4. ***Students will be able to apply non- parametric tests in different contexts***

Module 1: **Point and Interval Estimation**. Concept of estimation , Desirable properties of a good estimator, unbiasedness, consistency, sufficiency, Fisher - Neyman factorization theorem (Statement and application only), efficiency, Cramer - Rao inequality; Methods of Estimation - method of maximum likelihood, method of moments, Bayesian estimation method. **Interval Estimation**: Large sample confidence interval for mean, equality of means, equality of proportions. Derivation of exact confidence intervals for means , variance and ratio of variances based on Normal, t, chi square and F distribution **[28 hrs]**

Module 2: Testing of Hypotheses; concept of testing hypotheses, simple and composite hypotheses, null and alternative hypotheses, type I and type II errors, critical region, level of significance, power of test. Most powerful tests Uniformly most powerful test, Neyman Pearson Lemma (statement only). Sequential sampling and SPRT (Basic concepts only). **[22 hrs]**

Module 3: **Large sample and Small sample tests** :- Large sample tests concerning mean, equality of means, proportions, equality of proportions. Small sample tests based on *t* distribution for mean, equality of means and paired*t* test, one-way ANOVA. Tests based on *F* distribution. Tests based on chi square distribution – Test for thesignificance of population variance, goodness of fit and for independence of attributes. Test for correlation coefficients. **[16 hrs]**

Module 4: **Non parametric tests** – Advantages and disadvantages of non-parametric tests; Kolmogorov - Smirnov test; one sample and two sample sign tests; Wilcoxon signed rank test; Median test; Mann Whitney test; Kruskal Wallis test and test for randomness (run test). **[14 hrs]**

Books for Study

1. Gupta S. C. and Kapoor, V. K. (2014): *Fundamentals of Mathematical Statistics*, Sultan Chand & Co
2. Gupta S. C. and Kapoor, V. K.: *Fundamentals of Applied Statistics,* Sultan Chand & Co
3. John E Freund (2013), Mathematical Statistics (8th edn), Pearson Edn, New Delhi

Books for References

1. Mood A. M., Gray bill F. A., Boes D C (2017): Introduction to the theory of statistics –Tata Magrow Hill.
2. Goon A. M., Gupta M. K., Das Gupta. B. (1999): *Fundamentals of Statistics*, Vol. I, World Press, Calcutta.
3. Rohatgi, VK and Saleh AK (2015) An introduction to Probability and Statistics (3rd Edition) John Wiley and Son Inc

Semester V

Statistics – 5

ASD5 B05 : SAMPLING THEORY AND DESIGN OF EXPERIMENTS

Teaching Hrs. 06 Credit : 05

Expected Course Outcome:-

1. ***Student should be able to design and implement sample survey***
2. ***Students can estimate population parameters using sampling distributions using different sampling methods***
3. ***Students will be capable to apply ANOVA and ANCOVA***
4. ***Students will be able to apply different experimental designs in appropriate contexts***

Module 1: **Census and Sampling:** Principal steps in sample survey, sampling vs census, sampling and non-sampling errors and types of sampling. Simple random sampling: SRSWR and SRSWOR- methods of selecting a SRS, unbiased estimators of population characteristics, their variances and estimators of the variances under both SRSWR and SRSWOR, estimation of sample sizes in SRS

[24 hrs]

Module 2: **Stratified, Systematic and Cluster sampling:** Method of selecting a stratified random sample, unbiased estimators of population characteristics and their variances, allocation of sample size in stratified sampling proportional and optimum allocations, estimates of population characteristics and their variances under these allocations, comparison of stratified random sample with SRS. systematic sampling - linear systematic sampling, estimation of population characteristics and the expressions for variance of the estimator under linear systematic sampling, comparison of systematic sampling with SRS and stratified random sampling, circular systematic sampling, Cluster sampling and Multistage sampling (definition only).

[26 hrs]

Module 3: Linear estimation, estimability of parametric functions and BLUE Gauss- Markov theorem-Linear Hypothesis Analysis of variance, one way and two way classification (with single observation per cell), Post Hoc Tests - Least Significant Difference (LSD) test, Analysis of covariance with a single observation per cell (Concept and model only).
[26 hrs]

Module 4: Principles of design-randomization-replication-local control, completely randomized design; Randomized block design; Latin square design. Missing plot technique; comparison of efficiency; Greco-Latin square design (Concept only). Basic concepts of factorial experiments. **[20 hrs]**

Books for Study

1. Daroja Singh and F S Chaudhary, Theory and Analysis of Sample Survey Designs, WielyEstrn Limitted
2. Gupta S. C. and Kapoor, V. K.: *Fundamentals of Applied Statistics,* Sultan Chand & Co

Books for References

1. Montgomery, D C, Design and Analysis of Experiments, John Wiley
2. Murthy M N, Sampling theory and methods, Statistical Publishing society, Calcutta.
3. Cochran W.G, Sampling Techniques, Wiely Estern.
4. M N Das and N Giri, Design of Experiments, New Age international,
5. D.D Joshy, linear Estimation and Design of Experiments, Wiley Eastern

Statistics – 6

ASD5 B06 MULTIVARIATE TECHNIQUES FOR DATA SCIENCE

Teaching Hrs. 06 Credit : 05

Expected Course Outcome:-

1. ***Student will learn the multivariate normal distribution and its properties***
2. ***Students will be able to make inference on multivariate mean vector***
3. ***Students will learn on Principle component analysis and applications***
4. ***Students will make use of the concept of Factor analysis in real problems***

Module 1: **Multivariate Normal Distribution** - The Multivariate Normal Density, Properties of multivariate normal density, Distribution of a linear combination of the components of a normal random vector, Sampling from a Multivariate Normal Distribution and Maximum Likelihood Estimation, Maximum likelyhood estimation of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, The Sampling Distribution of X-bar and S, Assessing the Assumption of Normality, Evaluating the Normality of the Univariate Marginal Distributions, Evaluating Bivariate Normality.

[26 hrs]

Module 2: **Inference about Mean Vector :-** The Plausibility of $\boldsymbol{\mu_0}$ as a Value for a Normal Population Mean, Hotelling's T^2 and Likelihood Ratio Tests,General Likelihood Ratio Method, Confidence Regions and Simultaneous Comparisons of Component Means, Simultaneous Confidence Statements, A Comparison of Simultaneous Confidence Intervals with One-at-a- Time Intervals, The Bonferroni Method of Multiple Comparisons, Comparison of several Multivariate means, Paired Comparisons and a Repeated Measures Design Comparing Mean Vectors from Two Populations, Comparing Several Multivariate Population Means (One-Way MANOVA). **[28 hrs]**

Module 3 Principal Component Analysis - Population Principal components ; Principal Components Obtained from Standardized Variables, Principal Components for Covariance Matrices with Special Structures, Summarizing Sample Variation by Principal Components The Number of Principal Components, Interpretation of the Sample Principal Components, Standardizing the Sample Principal Components, Graphing the Principal Components **[22 hrs]**

Module 4 Factor Analysis; The Orthogonal Factor Model ,Methods of Estimation The Principal Component (and Principal Factor) Method, A Modified Approach-the Principal Factor Solution, The Maximum Likelihood Method, A Large Sample Test for the Number of Common Factors, Factor Rotation, Factor Scores. [**20 hrs]**

Books for Study

1. Gupta S. C. and Kapoor, V. K. (2014): *Fundamentals of Mathematical Statistics*, Sultan Chand & Co
2. Richard A. Johnson and Dean W. Wichern (2019), Applied Multivariate Statistical Analysis, Prentice hall India, 7th Edition, 2019.

Books for Reference

1. Brian Everitt and Torsten Hothorn (2011) An Introduction to Applied Multivariate Analysis with R Springer New York .
2. Daniel Zelterman, Applied Multivariate Statistics with R , Springer

Statistics – 7

ASD5 B07: REGRESSION ANALYSIS

Teaching Hrs. 05 Credit : 04

Expected Course Outcome:-

1. ***Student will be able to explain the context of regression methods and can estimate regression coefficients***
2. ***Students can fit multiple linear regression models in real life data sets***
3. ***Students will be capable of checking different models and to select adequate one***
4. ***Students will be able to fit logistic and polynomial regression models and to check the suitability of such models***

Module 1: **Correlation and Regression**- Curve fitting, principle of least squares, fitting of straight lines, parabolas, exponential curves. Bivariate linear correlation – Scatter diagram. Pearsons correlation coefficient, Spearman's rank correlation coefficient. Bivariate linear regression – regression lines, coefficients of regression. Multiple and partial correlation for three variables (definition only). Computations using R. **[20 hrs]**

Module 2 : Regression Model building: regressor, response, error, uses of regression. Simple linear regression: Simple liner regression model, assumptions, Least square estimation of parameters, Properties of the Least-square estimators and the fitted Regression Model. Estimation of σ^2, Hypothesis testing of slope and intercept. Interval estimation of regression parameters (Slope, intercept and σ^2). Coefficient of determination. Estimation of regression parameters by the method of Maximum likelihood.

[20 hrs]

Module 3: Multiple Linear Regression: Multiple linear regression model, assumptions, least square estimation of parameters, Properties of the Least-square estimators Hypothesis testing in Multiple linear regression (ANOVA), Test on individual regression coefficients, Interval estimation of coefficients, slope and intercept, co-efficient of determination.

[20 hrs]

Module 4: Model adequacy checking: Residual analysis, Methods of scaling residuals – standardized residuals, studentized residuals, PRESS residuals, R- Student. Residual plots – Normal probability plots, plot of residuals against fitted values, plot of residuals against the regressor, plot of residuals in time sequence. PRESS Statistic, R^2 for prediction based on PRESS. Introduction to Polynomial and logistic regression

[20 hrs]

Book for Study

1. D C. Montegomerry, E A Peak and G G Vining (2003) , Introduction to Linear Regression Analysis, Wiley

Books for Reference

1. Seber (1997) , Linear Regression Analysis, Wiley
2. D. D Joshi (1987) , Linear Estimation and Design of Experiments, Wiley
3. D N Gujarathi, D C Porter and G Sangeetha, (2003) Basic Econometrics, Mc Graw Hill

Statistics –8

ASD5 B08: INTRODUCTION TO DATA SCIENCE AND STATISTICAL MACHINE LEARNING

Teaching Hrs. 05 Credit : 05

Expected Course Outcome:-

1. ***Student will get and understanding on role and significance of data science and machine learning in present day industrial domains***
2. ***Students will be able to classify the data sets using statistical pattern recognition method***
3. ***Students will be able to make inference on data based on training data sets using different rules***
4. ***Students are expected to apply Kernal rules for classification***

Module 1 : **Introduction to Data Science and Machine Learning ,** What is Data Science Data Science in Finance , Public Policy, Politics, Healthcare, Urban Planning, Education, Libraries etc ; Data Science and Statistics, Data Science and Computer Science, Data Science and Engineering, Data Science and Business Analytics , Data Science and Computational Social Science Relationship between Data Science and Information Science, Information vs. Data Users in Information Science , Computational Thinking , Skills for Data Science, Tools for Data Science, Issues of Ethics, Bias, and Privacy in Data Science. Why Machine Learning? Some Applications, -Image Recognition, Speech Recognition, Medical Diagnosis, Statistical Arbitrage, Measurements, Features, and Feature Vectors, Supervised Learning

[25 hrs]

Module 2: The Pattern Recognition Problem , Decision Rules, Success Criterion, The Best Classifier: Bayes Decision Rule, Continuous Features and Densities, The Optimal Bayes Decision Rule **,** Bayes Theorem, Bayes Decision Rule, Optimality **20 hrs]**

Module 3:- Learning from Examples Lack of Knowledge of Distributions, Training Data, Assumptions on the Training Data, A Brute Force Approach to Learning, Curse of Dimensionality, Inductive Bias, The Nearest Neighbor Rule, Performance of the Nearest Neighbor Rule, Intuition and Proof Sketch of Performance, Using more Neighbors

[20 hrs]

Module 4:- Kernel Rules A Variation on Nearest Neighbor Rules, Kernel Rules, Universal Consistency of Kernel Rules, Potential Functions, More General Kernels, Neural Networks: Multilayer Feedforward Networks, Neural Networks for Learning and Classification, Perceptrons, Threshold, Learning Rule for Perceptrons, Representational Capabilities of Perceptrons,

[15 hrs]

Books for Study

1. **Chirag Sha (2020**) A Hands-On Introduction to Data Science, Cambridge University press, University Printing House, Cambridge CB2 8BS, United Kingdom.
2. **Sanjeev Kulkarni and Gilbert Harman**, An Elementary Introduction to Statistical Learning Theory, John Wiley and Sons. Inc. publication

Books for References

1. **Trevor Hastie, Robert Tshibirai, Jerome Friedman** The Elements of statistical learning : Data Mining, Inference and Prediction, 2nd edition, , Springer publication
2. **Gareth James, Daniela Witten,Trevor Hastie, Robert Tibshirani** (2013) An Introduction to statistical learning with applications in R, Springer,

Semester 6

Statistics – 9

ASD6 B09 : TIME SERIES ANALYSIS AND INDEX NUMBERS

Teaching Hrs. 05 Credit : 04

Expected Course Outcome:-

1. ***Students will build an understanding on time series data sets and its statistical analysis***
2. ***Students can distinguish different times series modes like MA, ARMA ARIMA etc. and will be able to analyse real life data using these models***
3. ***Students will make an understanding on different index numbers and their uses.***
4. ***Students will be able to test for best index numbers***

Module 1: Time series analysis: Examples of Time series, different components, illustrations, additive and multiplicative models, methods of measuring trends, analysis of seasonal fluctuations, construction of seasonal indices. Time-series as discrete parameter stochastic process, auto covariance and autocorrelation functions and their properties. Exploratory time Series analysis, tests for trend and seasonality, exponential and moving average smoothing. Holt and Winters smoothing, forecasting based on smoothing.

[20 hrs]

Module 2: Stationary processes, (1) moving average (MA), (2) auto regressive (AR), (3)ARMA and (4) AR integrated MA(ARIMA) models, choice of AR and MA periods. Discussion (without proof) of estimation of mean, auto covariance and autocorrelation function.

[25 hrs]

Module 3: Index numbers: Meaning and definition-uses and types-problems in the construction of index numbers-simple aggregate and weighted aggregate index numbers. Laspeyere's, Paasche's, Marshall-Edgeworth's and Fisher's, Dorbish-Bowley's andKelly's indices. Quantity Index Numbers

[20 hrs]

Module 4: Test for consistency of index numbers-factor reversal, time reversal and unit test, Chain base index numbers-Base shifting-splicing and deflating of index numbers. Consumer price index numbers-family budget enquiry-limitations of index numbers.

[15 hrs]

Book for Study

1. SC Gupta and V K Kapoor, Fundamentals of applied statistics, Sulthan chand and sons
2. Jonathan D. Cryer and Kung-Sik Chan (2008), Time Series Analysis with Applications in R, Springer

Books for Reference

1. Goon A M, Gupta M K and Das Gupta, Fundamentals of Statistics Vol II, The World press,Calcutta.
2. Box, G.E.P and Jenkins, G.M. (1970). Time Series Analysis, Forecasting and Control. Holden Day, Sanfransisco.
3. SP Gupta Statistical methods , Sulthan chand and sons

Statistics – 10

ASD6 B10 :STATISTICAL QUALITY CONTROL, ACTUARIAL STATISTICS, LINEAR PROGRAMMING AND DEMOGRAPHY

Teaching Hrs. 05 Credit : 04

Expected Course Outcome:-

1. ***Students will learn applications of statistical methods in product and process quality control***
2. ***Students will be get an idea on how statistical principles are applied in actuarial science***
3. ***Students are expected to solve constrained optimization problems using the principles of linear programming***
4. ***Students will become familiar with statistics used in population studies***

Module 1: **Statistical Quality Control:-** Process and Product Control, Control charts, Control limits, Control charts for variables- X-bar and R charts, Control charts for attributes- p-chart and d-chart, Control chart for number of defects- c-charts, Natural tolerance limit and specification limits, Acceptance sampling by variables, sampling plans for single specification limit with known and unknown and unknown variance, Sampling plans with double specification limits., comparison of sampling plans by variables and attributes, Continuous sampling

[20 hrs]

Module 2: : **Actuarial statistics -** Rate of interest: simple and compound, effective rate of interest, nominal rate of interest, relationship between these rate of interest, force of interest, present value, accumulated value, future lifetime random variable, survival function, force of mortality, probability laws of mortality, Annuities, various types of annuities, numerical problems

[20 hrs]

Module 3: **Linear Programming Problem-**Introduction, mathematical formulation of LPP, solution of LPP – graphical method, convex sets, statement of extreme value theorem, simplex method, Transportation problem: mathematical formulation, solution of a transportation problem -North-West Corner rule, Matrix minima, Vogel's Approximation Method **[15 hrs]**

Module 4: **Population Studies -** Nature and Scope of population studies, Sources of population data- Population census, Registration of vital events, World population – Growth and Distribution, Demographic transition, Population growth in India- Factors affecting population growth. **Mortality, Fertility, Migration and Life tables-** Concept of Mortality, Measures of Mortality, Sex and age pattern of mortality, Infant mortality, Levels and trends in mortality, Mortality in India, Concept of Fertility, Physiological factors affecting fertility, Social and cultural factors, Measures of fertility, Levels and trends in fertility, Factors affecting decline in fertility, age patterns of fertility of India. Concept of Migration, Internal migration, methods of measuring migration, Differential migration, Population policies, India population policy, Life tables, Types of life tables, Structure and function of life tables. Abridged life tables- Reed Merrel, Greville and Kings methods.

[25 hrs]

Books for Study

1. S.C. Gupta and V K Kapoor, Fundamentals of Applied Statistics, Sulthan Chand and Son.
2. Shylaja R Deshmukh : Actuarial statistics ,Universities Press (India) Pvt. Ltd
3. Kantiswarup, P K Gupta, Manmohan: Operations Research, Sultan Chand and Sons
4. Asha A Bhende and Tara Kanitkar (2015) Principles of Population Studies , Himalaya Publishing House.

Books for References

1. Montgomory, R.C. (1985), Introduction to Statistical Quality Control. 4th edition. Wiley, New-York.
2. C.D.Daykin, T. Pentikainen et al, Practical Risk Theory of Acturies, Chapman and Hill
3. M E Atkinson & D C M Dickson : An Introduction to Actuarial studies Second Edition, Edward Elger Publishing limited. UK , USA
4. Mukhopadhaya. P, Applied Statistics, New Central Book Agency (P)Ltd., Calcutta.
5. Goon A. M, Gupta M. K and Das Gupta B : Fundamentals of Statistics Vol II.
6. S.Kalavathy: Operations Research Vikas Publishing House Pvt Ltd.

Statistics – 11

ASD6 B11 : DATA MINING

Teaching Hrs. 05 Credit : 05

Expected Course Outcome:-

1. ***Students will be able to understand the context of data mining and text mining***
2. ***Students will be able to use different classifiers for the classification of data***
3. ***Students will be able to use computation techniques for cluster analysis***
4. ***Students will be able to apply text mining techniques in practical situations.***

Module 1: **Introduction :-** What Is Data Mining? , The Origins of Data Mining , Data Mining Tasks, Types of Data, Attributes and Measurement, Types of Data Sets, Data Quality, Measurement and Data Collection Issues , Data Preprocessing – Aggregation, Sampling, Dimensionality Reduction , Feature Subset Selection , Feature Creation , Discretization and Binarization ., Variable Transformation; Measures of Similarity and Dissimilarity , Similarity and Dissimilarity between Simple Attributes.

[20 hrs]

Module 2 Classification: Basic Concepts, General Framework for Classification, Decision Tree Classifier - A Basic Algorithm to Build a Decision Tree, Methods for Expressing Attribute Test Conditions , Measures for Selecting an Attribute Test Condition , Algorithm for Decision Tree Induction , Characteristics of Decision Tree Classifiers ; Model Overfitting ; Model Selection , Model Evaluation , Model Comparison

[20 hrs]

Module 3 Association Analysis and Cluster Analysis : Frequent Item set Generation, The *Apriori* Principle, Rule Generation , Compact Representation of Frequent Itemsets, Evaluation of Association Patterns. **Cluster Analysis:-** What Is Cluster Analysis? Different Types of Clusterings, The Basic K-means Algorithm, Bisecting K-means, Agglomerative Hierarchical Clustering , Basic Agglomerative Hierarchical Clustering Algorithm Lance-Williams Formula for Cluster Proximity, Key Issues in Hierarchical Clustering, Outliers

[20 hrs]

Module 4 Text Mining :- Reading Text Input Data, Common Text Preprocessing Tasks, Stop-Word Removal, Stemming, Term Document Matrix, Weighting Function, Text Mining Applications, Frequency Analysis, Text Classification.

[20 hrs]

Book for Study

4. **Pang-Ning Tan, Michael Steinbach, Anujkrptne and Vipin Kumar** (2019) Introduction to Data Mining (2nd Edition) Pearson Education Ltd.
5. **Manas A. Pathak** (2014) Beginning Data Science with R, Springer International Publishing

Books for Reference

1. **Paolo Giudici and Silvia Figini** (2009) Applied Data Mining for Business and Industry, Second Edition, John Wiley & Sons Ltd
2. **Nina Zumel and John Mount (2020)** Practical Data Science with R 2nd Edition, Manning Publications
3. **Daniel T Larose (2005)** Discovering Knowledge in Data : An Introduction to Data Mining, John Wiley & Sons, Inc.

Statistics – 12

ASD6 B12 : STATISTICAL COMPUTATIONS USING R AND PYTHON

Teaching Hrs. 05 Credit : 04

Expected Course Outcome:-

1. ***Students will develop skills to do statistical problems using R or Python***
2. ***Students will exhibit the capacity to undertake statistical analysis works of real life problems***
3. ***Students will have the ability to correlate theoretical knowledge with applied problems***
4. ***Students will capable to work in industries and organizations as a statistical data analyst***

Statistical Computing will be a practical course. Practical is to be done using R or Python Programming. At least five statistical data oriented/supported problems should be done from each course. Each student shall maintain practical Record and the same shall be submitted for verification at the time of external examination.

The practical is based on the following courses they studies from 1st to 6th semesters

Statistics –1 : **DESCRIPTIVE STATISTICS AND BASIC R PROGRAMMING**
Statistics – 2: **PROBABILITY AND MATHEMATICAL EXPECTATION**
Statistics – 3: **PROBABILITY DISTRIBUTIONS**
Statistics – 4 :**STATISTICAL INFERENCE**
Statistics – 5 : **SAMPLING THEORY AND DESIGN OF EXPERIMENTS**
Statistics – 6 : **MULTIVARIATE TECHNIQUES FOR DATA SCIENCE**
Statistics – 7: **REGRESSION ANALYSIS**
Statistics –8 : **INTRODUCTION TO DATA SCIENCE AND STATISTICAL MACHINE LEARNING**
Statistics – 9 : **TIME SERIES ANALYSIS AND INDEX NUMBERS**
Statistics – 10 : **STATISTICAL QUALITY CONTROL, ACTUARIAL STATISTICS, LINEAR PROGRAMMING AND DEMOGRAPHY**
Statistics – 11: **DATA MINING**

The Board of Examiners (BoE) shall decide the pattern of question paper and the duration of the external examination. The external examination shall be conducted and evaluated on the same day jointly by two examiners – one external and one internal, appointed at the centre of the examination by the University on the recommendation of the Chairman, BoE. The question paper for the external examination at the centre will be set by the external examiner in consultation with the Chairman, BoE and the HoDs of the centre. The questions are to be evenly distributed over the entire syllabus. Evaluation shall be done by assessing each candidate on the scientific and experimental skills, the efficiency of the algorithm/program implemented, the presentation and interpretation of the results.

Statistics – 13

ASD6 B13 : PROJECT WORK/ INTERNSHIP

Teaching Hrs. 02 Credit : 02

Expected Course Outcome:-

1. ***Students will capable to apply the theoretical knowledge they acquired in a practical situation***
2. ***The project work will helps the students to enhance their Research skills.***
3. ***Students will get an exposure to study the applications of Statistics in Industry or organization***
4. ***Students will improve theory professional skills to work in public or private sectors***

The following guidelines may be followed for project work.

1. The project work should be assigned to students in the 5^{th} semester and to be completed in last semester of the degree course.
2. A project may be undertaken by a group of students, the maximum number in a group shallnot exceed 5. However the project report shall be submitted by each student.
3. There shall be a teacher from the department to supervise the project and the synopsis of the project should be approved by that teacher. The head of the department shall arrange teachers for supervision of the project work.
4. As far as possible, topics for the project may be selected from the applied branches of statistics, so that there is enough scope for applying and demonstrating statistical skills learnt in the degree course.
5. Field/Industrial/Organization visit is mandatory for the data collection.

Statistics – 14 (Elective-I)

ASD6 B14(E) :DATA VISUALIZATION TECHNIQUES

Teaching Hrs. 03 Credit : 02

Expected Course Outcome:-

1. ***Students will able to interpret data plots and understand core data visualization concepts such as correlation, linear relationships, and log scales.***
2. ***Students will be able to explore the relationship between two continuous variables using scatter plots and line plots.***
3. ***Students can translate and present data and data correlations in a simple way, data analysts use a wide range of techniques — charts, diagrams, maps, etc.***

Module 1: Data Foundations and Visualization Foundations:- Types of data, Structure within and between records, Data Preprocessing, Visualization process, Semiology of graphical symbols, visual variables.

Module 2: Visualization techniques for spatial data- one-dimensional, two-dimensional and three dimensional data and dynamic data, Visualization techniques Geospatial data-Point data, line data and area data.

Module 3: Visualization techniques for time-oriented data- characterizing time oriented data, Time bench, Visualization for multivariate data- Point based techniques, line based techniques , region based techniques. Combinations of techniques.

Textbook:

1. Ward, Grinstein Keim, Interactive Data Visualization: Foundations, Techniques, and Applications. Natick: A K Peters, Ltd, 1st Edition, 2010

Reference :

1. Kieran Healy, Data Visualization: A Practical Introduction, 1st Edition, 2018
2. Andy Krik, Data Visualization: a successful design process 1st Edition, 2012
3. Corey Lanum, Visualizing Graph Data 1st Edition, 2016

Statistics – 15 (Elective- II)

ASD6 B15(E) OPERATIONS RESEARCH

Teaching Hrs. 03 Credit : 02

Expected Course Outcome:-

1. ***Students will understand the context of application of optimization techniques in various situations***
2. ***Students will get the ability to apply optimizations methods in transportations problems***
3. ***Students will develop the ability to apply assignment tools in optimization situations***
4. ***Students will be an understanding on network problems***

Module 1: **Operations Research** Origin and Development of OR, Objectives of OR, Modeling and types of models in OR Linear Programming: Mathematical formulation of LPP, graphical solutions of a L.P.P. Simplex method for solving LPP. Artificial Variables - Two phase method, Big M-method, Concept of Duality in L.P.P, Dual simplex method.

Module 2: **Transportation Problems and Assignment Problems** General transportation problem. Methods for finding initial basic feasible solutions by North West corner rule, Least cost method and Vogel's approximation method (VAM). MODI method to find the optimal solution. Unbalanced transportation problem and degeneracy (definitions and simple problems only). Assignment problem-Hungarian method to find optimal assignment.

Module 3: **Network Analysis: -**Drawing the Network Diagram – Analysis of Network, Calculation of Critical Path – PERT, Expected Completion Time and its Variance.

Book for study

1. Kanti Swarup, Gupta P.K., Man Mohan (2010): *Operations Research*, Sultan Chand and Sons, New Delhi.

References.

1. Taha, H.A. (2014).*Operations Research*, Pearson Education Publication.
2. Gupta R.K. (2010): *Operations Research*, Krishna Prakashan Media (P) Ltd., Meerut.

Statistics – 16 (Elective-III)

ASD6 B16 (E) MATHEMATICAL ECONOMICS

Teaching Hrs. 03 Credit : 02

Expected Course Outcome:-

1. ***Students can understand apply mathematical tools in economic analysis***
2. ***Students will develop an understanding on applications of calculus in demand supply Analysis***
3. ***Students will be able to apply mathematical theories in production contexts***

Module 1: Demand and Supply Analysis-Concept of demand, demand function, elasticity of demand, elasticity of substitution, relation between elasticity of demand, price, average revenue, total.

Module 2: Consumer Behaviour-Concept of utility, cardinal and ordinal utility, maximization of utility, budget constraint and equilibrium of consumer, income and substitution effects of a price change, Sluky equation.

Module 3: Production Theory-Output and input relation, total, average, marginal products in case of production with single variable input, production isoquants and economic region of production. Meaning and nature of production functions, returns to scale, linearly homogeneous production functions and its properties, Euler's theorem and its applications for various standard production functions.

Books for study

1. Allen R.G.D. (2014).*Mathematical Analysis for Economists,* Trinity Press.

References

1. Henderson, J.M. and Quandt, R.E (2003).*Micro EconomicTheory: A Mathematical Approach*, (3rded.), McGraw-Hill Education (India) Pvt.Ltd.
2. Simon, C.P.and Blume, L. (2010): *Mathematics for Economists*,(1st ed.),Viva Books.
3. Madnani, G.M.K.andMehta, B.C (2014).*Mathematics for Economists*, Sultan Chand & Sons, New Delhi.
4. Koutsoyiannis ,A.(2008).*Modern Micro Economics*,(2nded.), Macmillan publishers

OPEN COURSES

1. ASD5 D01: ECONOMIC STATISTICS

Module 1: Time series analysis: Economic time series, different components, illustrations, additive and multiplicative models, determination of trends, growth curves, analysis of seasonal fluctuations, construction of seasonal indices 24 hours

Module 2: Index numbers: Meaning and definition-uses and types-problems in the construction of index numbers-simple aggregate and weighted aggregate index numbers. Test for consistency of index numbers-factor reversal , time reversal and unit test, Chain base index numbers-Base shifting-splicing and deflating of index numbers. Consumer price index numbers-family budget enquiry-limitations of index numbers. 30 hours

Books for references

1. S C Gupta and V K Kapoor, Fundamentals of Applied Statistics, Sulthan Chands and sons
2. Goon A M, Gupta M K and Das Gupta, Fundamentals of Statistics Vol II, The World Press, Calcutta

2. ASD5 D02: QUALITY CONTROL

Module 1: General theory of control charts, causes of variations in quality, control limits, sub-grouping, summary of out-of-control criteria, charts of ttributes, np chart, p chart, c chart, Charts of variables: X bar chart, R Chart and sigma chart, Revised control charts, applications and advantages 30 hours

Module 2: Principles of acceptance sampling-problems of lot acceptance, stipulation of good and bad lots-producer' and consumer' risk, simple and double sampling plans, their OC functions, concepts of AQL, LTPD,AOQL, Average amount of inspection and ASN function 24 hours

References

1. Grant E L, Statistical quality control, McGraw Hill
2. Duncan A J, Quality Control and Industrial Statistics, Taraporewala and sons
3. Montegomery D C, Introduction to Statistical Quality Control, John Wiley and son

3. ASD5 D02: BASIC STATISTICS

Module 1: Elements of Sample Survey: Census and Sampling, advantages, principal step in sample survey-sampling and non-sampling errors. Probability sampling, judgment sampling and simple random sampling. 10 hours

Module 2: Measures of Central tendency: Mean, median and mode and their empirical relationships ; Measures of Dispersion: absolute and relative measures, standard deviation and coefficient of variation. 12 hours

Module 3: Fundamental characteristics of bivariate data: univariate and bivariate data, scatter diagram, Pearson's correlation coefficient, limit of correlation coefficient. Curve fitting, principle of least squares, fitting of straight line. 15 hours

Module 4: Basic probability: Random experiment, sample space, event, algebra of events, Statistical regularity, frequency definition, classical definition and axiomatic definition of probability-addition theorem, conditional probability, multiplication theorem and independence of events (limited to three events). 17 hours

References

1. V. K. Rohatgi, An Introduction to Probability Theory and Mathematical Statistics, Wiley Eastern.
2. S.C.Gupta and V. K. Kapoor, Fundamentals of Mathematical Statistics, Sultan Chand and Sons
3. A.M. Mood, F.A. Graybill and D C Bose, Introduction to Theory of Statistics, McGraw Hill
4. John E Freund, Mathematical Statistics (6th edn), Pearson Edn, New Delhi

MODEL QUESTION PAPERS

FIRST SEMESTER U.G. DEGREE EXAMINATION
(CBCSS—UG)
Statistics
ASD1 B01 - DESCRIPTIVE STATISTICS AND BASIC R PROGRAMMING

Time: 2 ½ Hours Max.: 80 Marks

PART A
Each question carries 2 marks.

1. What are the different statistical organizations under MOSPI
2. List out the important activities of Central Statistics Organization.
3. What are the main roles of state level statistical organizations?
4. Write about vectors in R
5. Write a note on data visualization in R
6. What is Box-Plot
7. Distinguish between primary and secondary data
8. What is the difference between census and sampling
9. The mean salary of 80 male employees in a firm is Rs. 5200 and that of 20 females in the same firm is Rs. 4200. What is the mean salary of all the employees in that firm?
10. Define Harmonic Mean and Geometric mean.
11. Compare range and quartile deviation.
12. The mean, median and standard deviation of a moderately asymmetrical distribution are 25, 23 and 4.5 respectively. Calculate Pearson's coefficient of skewness.
13. Define co-efficient of variation ? What it indicates?
14. What is Histogram
15. Explain skewness and kurtosis.

PART B
Each question carries 5 marks

16. Explain major functions of United Nation Statistics Division
17. Explain the functions of NSSO.
18. Explain different data structures in R
19. Explain different types of operators in R
20. Explain scales of measurement

21. For a group of 150 candidates, the mean and standard deviation of scores were found to be 38 and 16 respectively. Latter on it was found that the scores 45 ad 53 were misread as 54 and 35 respectively. Find the standard deviation of corrected figures.

22. Obtain the median for the following frequency distribution

X	1	2	3	4	5	6
Frequency	8	10	11	16	20	25

23. Establish the relationship between row moments and central moments

PART C

Each question carries 10 marks (Answer any **TWO** Questions)

24. Write an essay on different types of data.

25. Police records show the following numbers of daily crime reports for a sample of days during the winter months and a sample of days during the summer months. Compare the variability of the two periods.

Winter	18	20	15	16	21	20	12	16	19	20
Summer	28	18	24	32	18	29	23	38	28	18

26. Illustrate the concept of ogives with an example

27. What are partition values? Explain with an example

SECOND SEMESTER U.G. DEGREE EXAMINATION
(CBCSS—UG)
Statistics
ASD2 B02 : PROBABILITY AND MATHEMATICAL EXPECTATION

Time: 2 ½ Hours Max.: 80 Marks

PART A
Each question carries 2 marks.

1) Define random experiment. Write an example.
2) State addition theorem of probability for three events.
3) Define independence of events.
4) Write the axiomatic definition of probability.
5) Distinguish between discrete and continuous random variables.
6) Define distribution function. Write any two properties of distribution function.
7) Let $f(x) = 2x+3$, $0 < x < 1$; 0 otherwise. Verify whether $f(x)$ is a probability density function or not.
8) Define distribution function of a bivariate random variable and write any two properties.
9) When do you say that two random variables are independent?
10) Define expectation of a random variable. Write any two properties of expectation.
11) Show that $V(aX + b) = a^2V(X)$.
12) Let X_1 and X_2 be two independent random variables. Prove that $M_{X1+X2}(t) = M_{X1}(t)M_{X2}(t)$.
13) Define characteristic function and write any two properties.
14) State and prove the multiplication theorem of expectation
15) Define conditional expectation and conditional variance of a bivariate random variable

PART B
Each question carries 5 marks

16) A problem in Statistics is given to 3 students A,B and C whose chances of solvingit are ½, ¾ and ¼ respectively. What is the probability that the problem will be solved?
17) Let f(x) = 1, 0<x<1; 0 otherwise. Find the distribution of Y= -2log X
18) For a discrete r.v. X with probability distribution

x	-2	-1	0	1	2
P(X=x)	0.1	0.2	0.3	k	0.2

Find the value of (i) k (ii) $p(-1 \leq X \leq 1$

19) State and prove Baye's theorem.

20) For a discrete bivariate random variable (X, Y), prove that $E\{E(X|Y)\} = E(X)$

a. Let X be a random variable with the following probability distribution

x	-3	6	9
P(X = x)	1/6	1/2	1/3

Compute (i) V(X) and (ii) $E(2X+1)^2$

21) A random variable X has the density function $f(x) = 1/2\sqrt{x}$; $0 < x < 1$ Obtain the m.g.f.of X and hence fid its mean and variance.

22) The joint pdf of a bivariate random variable ifs $f(x\ y) = 2$; $0 < x < 1$; $0 < y < x$. Checkfor independence f X ad Y

PART C

Each question carries 10 marks (Answer any **TWO** Questions)

28. A box contains 3 blue and 2 red balls while another box contains 2 blue and 5 red balls. A ball drawn at random from one of the boxes turns out to be blue. What is the probability that it came from the first box.

29. A committee of 4 people is appointed from 3 officers of the production department, 4 officers of the purchase department, 2 officers of the sales department and 1 charted accountant. Find the probability of forming the committee in the following manner.

(1) There must be one from each category

(2) It should have at least one from the purchase department

(3) The charted accountant must be in the committee.

30. If (X, Y) is a bivariate discrete random variable with joint probability mass function

X / Y	1	2	3
1	$\frac{2}{21}$	$\frac{3}{21}$	$\frac{4}{21}$
2	$\frac{3}{21}$	$\frac{4}{21}$	$\frac{5}{21}$

31. Find the marginal distributions of X and Y (ii) Compute P(X|Y=1) and (ii)Compute P(Y|X=2)

32. If X and Y are two r.v.s having the joint pdf $f(x\ y) = 2 - x - y$;

$0 < x, y < 1$.

Find $\rho_{xy.}$

THIRD SEMESTER U.G. DEGREE EXAMINATION,
(CBCSS—UG)
Statistics
ASD3 B03 PROBABILITY DISTRIBUTION

Time:Two and half Hours **Maximum: 80 Marks**.

Use of calculator and Statistical table are permitted.

Section A
Each question carries 2 marks.
Overall Ceiling 30.

1.Define binomial distribution and explain its parameters.
2. Derive mean and variance of uniform distribution
3.The mean and variance of binomial distribution are 4 and 4/3 respectively, Find P(x≥1)
4.Obtain mean of poisson distribution
5. Define exponential distribution
6.Briefly explain parameter and statistic with examples
7. If the variance of X following Poisson distribution is 5, find P (X = 5).
8. Define negative binomial distribution.
9. State Bernoulli's law of large numbers.
11. Define a Bernoulli random variable.
12. What do you mean by the sampling distribution of a statistic
13.Define convergence in probability.
14.State weak law of large numbers.
15.What are the advantages of Chebyshev's Inequality ?

Section B
Each question carries 5 marks.
All questions can be attended. Overall Ceiling 30.

16.Establish the relation between a t variate and an F variate.
17. Obtain the sampling distribution of $\frac{ns^2}{\sigma^2}$where S^2 is the variance of a sample of size n taken from a Normal Distribution.
18. State and prove the lack of memory property of exponential distribution.
19. Find the mean and variance of X following beta distribution of first kind.
20. Obtain the mean of Normal Distribution
21.State and prove the additive property of Binomial Distribution
22. If p(x=2)=P(x=3) where X follows poissondistribution, findmgf of X
23.State and prove central limit theorem

Section C
Answer any two questions.
Each question carries 10 marks.

24.Obtain mean and variance of Binomial distribution
25. Explain about chi- square distribution and chi- square test
26. State the limiting conditions and prove that under the stated conditions binomial distribution approaches to Poisson distribution.
27. State and prove Chebycheve's inequality. Also find a lower bound to P (2<x<6) where the mean and variance of X are 4 and 3 respectively.

FORTH SEMESTER U.G DEGREE EXAMINATION
(CBCSS UG)
Applied Statistics with Data Science

ASD4 B04 : STATISTICAL INFERENCE

Time: 2 ½ Hours Max.: 80 Marks

Use of calculator and Statistical table are permitted.

PART A

Each question carries 2 marks.

1. Distinguish between parameter and statistic.
2. Describe properties of good estimator
3. Define unbiasedness and efficiency of an estimator.
4. Describe interval estimation.
5. Define confidence interval and confidence coefficient.
6. Write the confidence for the mean of normal distribution when population standard deviation is unknown.
7. Define simple and composite hypotheses.
8. Distinguish between Type I error and Type II error.
9. Distinguish between small sample tests and large sample tests.
10. Write the critical regions of large sample test.
11. Compare parametric and non-parametric tests
12. List the assumptions of one sample t test.
13. What is meant by goodness of fit?
14. Write short note on Median Test
15. Briefly discuss run test.

PART B

Each question carries 5 marks

16. Explain method of moments of estimation of parameters
17. Find the maximum likelihood estimator of λ for the Poisson distribution.
18. Establish confidence interval for the difference of proportions of two binomialpopulations
19. Explain pared t test. Give a practical situation where this test is suitable.
20. A sample of 25 boys who passed SSLC examination are found to have mean marks 50 with standard deviation 5 for English. The mean marks of 18 girls are found to be 48 with standard deviation 4 for the same subject. Does this indicate any significance difference between the marks of boys and girls assuming the population standard deviation are equal?
21. In a sample of 600 men from a certain city 400 are fund to be smokers In 900 from another city 450 are smokers Do the data indicate that the cities are significantly different as far as smoking habits of people are concerned.

22. Tests were carried out to assess the strength of single fibre yarn spun on two different machines A and B and the results are given below:

Machine A	4	4.4	3.9	3	4.2	4.4	5
Machine B	5.3	4.3	4.1	4.4	5.3	4.2	3.8

Assuming the samples have been taken from normal population, test the hypothesis that variability is same for both the machines.

23. Explain Chi square test for independence of attributes.

PART C

Each question carries 10 marks (Answer any **TWO** Questions)

24. Derive the confidence interval for the mean of a Normal population N(▪, □^{2}), when (a) □ is known (b) □ is unknown and the sample size is small

25. List basic assumptions of ANOVA and explain the procedure or performing an ANOVA test.

26. Fit a Poisson distribution for the following data and test the goodness of fit.

X	0	1	2	3	4	5	6
frequency	275	72	30	7	5	2	1

27. Explain (i) Wilcoxon signed rank test and (ii) Mann Whitney test

FIFTH SEMESTER U.G DEGREE EXAMINATION
(CBCSS UG)
Applied Statistics with Data Science

ASD5 B05 : SAMPLING THEORY AND DESIGN OF EXPERIMENTS

Time: 2 ½ Hours Max.: 80 Marks

PART A
Each question carries 2 marks.

1. Distinguish between census and sampling.
2. What is meant by judgement sampling?
3. Define secondary data. State its major sources.
4. Compare sampling and non-sampling errors.
5. Write any four properties of a good questionnaire.
6. Compute the total number of samples of size n = 2 from a population of N = 6.
7. In a systematic sampling N =40 and n = 4, find the value of k.
8. Briefly explain cluster sampling.
9. Explain Gauss Markov set up of linear model.
10. Define estimable parametric function.
11. Discuss the concept of ANCOVA.
12. Distinguish between CRD and RBD.
13. Define treatment.
14. How to compare various designs of experiments?
15. Distinguish between 2^2 and 2^3 factorial designs.

PART B
Each question carries 5 marks

16. What are the advantages of sampling over census?
17. Explain the probability sampling and non-probability sampling with the help of examples.
18. Explain the concept of stratified sampling.
19. Obtain an unbiased estimate of population mean in simple random sampling with replacement. Find the variance of the estimate.
20. Consider three independent random variables y_1 , y_2 and y_3 having common variance σ^2 and $E(Y_1)=\theta_1-\theta_2, E(Y_2)=\theta_1+\theta_2$, $E(Y_3)=2\theta_1-\theta_2$. Show that 3 $\theta_1-2\theta_2$ is an estimable parametric function

21. Find the least square estimate of the parameter vector θ in Gauss - Markov model and also find an unbiased estimator of σ^2.
22. What is meant by analysis of variance of experimental data? What are the assumptions used in it?
23. Give the analysis for completely randomized design

PART C

Each question carries 10 marks (Answer any **TWO** Questions)

24. Explain in detail the principal steps in a sample survey.
25. What is systematic sampling? Explain the estimation of population characteristics of systematic sampling
26. State and prove Gauss-Markov theorem
27. Explain the principles of design of experiments

FIFTH SEMESTER U.G DEGREE EXAMINATION
(CBCSS UG)
Applied Statistics with Data Science

ASD5 B06 MULTIVARIATE TECHNIQUES FOR DATA SCIENCE

Time : 2 $^{1}/_{2}$ Hours **Max: 80 Marks**

PART A
Each question carries 2 marks. Maximum mark =25

1. Define variance covariance matrix. Discuss any two properties.
2. Define Bivariate normal distribution.
3. X and Y are two independent identically multivariate normal variate with meanμ and variance covariance matrix Σ find the PMF of $Z = X - 2Y$.
4. Find the sampling distribution of $\overline{X}$, when samples are drawn from a multivariate normal population.
5. What are the methods to evaluate normality of bivariate data?
6. Define Hotelling's T^2 – statistics.
7. Discuss general likelihood ratio method.
8. What is Bonferroni method for multiple comparisons?
9. State the confidence region for the difference of two mean vectors.
10. What are one-at-a-time confidence intervals?
11. State the one-way MANOVA model.
12. How will you graph principal components?
13. What is principal component analysis? How it is used?
14. Explain how factor scores are used in data analysis.
15. What is orthogonal factor model?

PART B
Each question carries 5 marks. Maximum mark =35

16. Derive the characteristic function of multivariate normal distribution.
17. For Xdistributed as $N_3(\mu, \Sigma)$, find the distribution of $\begin{bmatrix} X_1 - X_2 \\ X_2 - X_3 \end{bmatrix}$
18. Let X ~ N_p (μ, Σ). If $X^{(1)}$ and $X^{(2)}$ are two sub vectors of X, obtain the conditional distribution of $X^{(1)}$ given $X^{(2)}$.
19. The data matrix for a random sample of size $n = 3$ from a bivariate normal population is $X = \begin{bmatrix} 6 & 9 \\ 10 & 6 \\ 8 & 3 \end{bmatrix}$.Evaluate the observed T^2 for$\mu_0^{'} = [9,5]$

20. How will you test the equality of covariance matrices of two multivariate normal distributions on the basis of independent samples drawn from two populations?
21. In Principal component analysis derive the first principal component.
22. Derive the likelihood ratio test for testing equality of covariance matrices.
23. Discuss the effect of an orthogonal transformation in factor analysis method.

PART C
Each question carries 10 marks. Answer any two questions

24. Derive the MLE's of the parameters μ, Σ of multivariate normal distribution. Show that these MLEs are independently distributed.
25. Derive the likelihood ratio test for assigned mean based on a sample of size N drawn from $N(\mu, \Sigma)$ assuming Σ is unknown.
26. What are sample principle components? Explain the procedure for standardising sample principal components.
27. Discuss the procedures of obtaining factor scores by least squares and regression methods.

FIFTH SEMESTER U.G DEGREE EXAMINATION
(CBCSS UG)
Applied Statistics with Data Science

STA5B07- REGRESSION ANALYSIS

Time: 2 ½ Hours Max.: 80 Marks

Part A
(Answer at least 10 questions. Each question carries 2 marks. All questions can be attended. Overall ceiling 25)

.

1. Describe scatter diagram.
2. Explain Spearmans Rank Correlation.
3. List some of the uses of regression.
4. Show that the least square estimate of multiple linear regression coefficient is unbiased.
5. Write the hypotheses and test statistic for testing the significance of slope coefficient in simple linear regression model.
6. Show that the residual mean square is an unbiased estimate of population variance of random error component in linear regression model.
7. Write the properties of least square estimates of simple linear regression coefficients.
8. What is the importance of ANOVA in multiple linear regression?
9. Define hat matrix.
10. Distinguish between R^2 and adjusted R^2.
11. What is the significance of coefficient of determination?
12. Write a situation where logistic regression is applicable.
13. List the assumptions of logistic regression.
14. What is residual analysis
15. Define logit function.

Part B
Answer at least 5 questions, *Each question carries 5 marks, All questions can be attended overall ceiling 35*

16. Derive the variance of regression coefficients in simple linear regression
17. Estimate the variance of the response variable.
18. From a study conducted by the Department of transportation on driving speed and mileage for midsize automobiles, following results are obtained:

Driving speed (x)	30	50	40	55	30	25	60	25

Mileage (y)	28	25	25	23	30	32	21	35

Fit a linear regression model for the mileage and interpret the result.

19. Consider the simple linear regression model $y=\beta_0+\beta_1 x+\varepsilon$ with $E(\varepsilon)=0, Var(\varepsilon)=\sigma^2 \square \varepsilon$ uncorrelate

20. Describe multiple linear regression model.

21. Discuss the properties of least square estimators in multiple linear regression.

22. Explain polynomial regression models and list some of its applied areas.

23. Explain the logistic regression model with a Binary response variable.

PART C

Each question carries 10 marks (Answer any **TWO** Questions)

24. Derive the least square estimates of simple linear regression coefficients and show that they are unbiased.

25. Explain the role of residual in model adequacy checking.

26. Explain test for significance of regression coefficients in multiple linear regression

27. Describe the estimation of regression coefficients in logistic regression model.

FIFTH SEMESTER U.G DEGREE EXAMINATION
(CBCSS UG)
Applied Statistics with Data Science

ASD5 B08: INTRODUCTION TO DATA SCIENCE AND STATISTICAL MACHINE LEARNING

Time: 2 ½ Hours Max.: 80 Marks

Part A

(Answer at least 10 questions. Each question carries 2 marks. All questions can be attended. Overall ceiling 25)

.

1. What is Data Science ?
2. Distinguish data and information
3. What are the goals of data science ?
4. What is a decision rule?
5. Define what is meant by the probability of error of a decision rule.
6. What is a Baye's rule ?
7. What is inductive bias
8. What is the curse of dimensionality?
9. What is classification ?
10. What is cluster Analysis?
11. Define Neural Networks?
12. Define perceptrons?
13. State Bayes Theorem.
14. What is recurrent Neural Network
15. What is feed forward neural network?

Part B

Answer at least 5 questions, *Each question carries 5 marks, All questions can be attended overall ceiling 35*

16. Explain how data science is related with Statistics ?
17. If Ω_0 and Ω_1 are the subsets for which a decision rule c decides 0 and 1 respectively . write expression for the average error rate of c?
18. Explain Bayes Decision rule
19. What is training data? What are the assumptions on training data?
20. Explain brute force approach to the learning problem

21. Explain conditions required on *kn* for the *kn*-NN rule to be universally consistent?
22. What conditions are required on the smoothing parameter *hn* for a kernel rule to be universally consistent?
23. Explain supervised and unsupervised machine Learning

PART C

Each question carries 10 marks (Answer any **TWO** Questions)

25. Identify any four areas in which data science is being used and describe how it is used?
26. Explain decision rule? Assuming that a decision rule for a pattern recognition problem is as follows:

Decide 1 (with cost Rs 10 for a wrong answer) when $0 \leq x \leq 1$, and decide 0 (with cost Rs 5 for a wrong answer) when $1 < x \leq 2$ and the feature space is [0,2], with no cost for a right answer. What is the expression (in terms of $P(0)$, $P(1)$, $p(x|0)$ and $p(x|1)$) for the average cost of that rule?

27. Discuss the nearest neighbor model in detail.
28. What is the NN rule and how does the expected error from the use of this rule compare with the Bayes error rate?) What conditions on *kn* are required for the *kn*-NN rule to have an asymptotic error rate equal to the Bayes error rate?

SIXTH SEMESTER U.G DEGREE EXAMINATION
(CBCSS UG)
Applied Statistics with Data Science
ASD6 B09 : TIME SERIES ANALYSIS AND INDEX NUMBERS

Time: 2 hours and a half **Max: 80Marks**

Part A
(Answer at least 10 questions. Each question carries 2 marks. All questions can be attended. Overall ceiling 25)

1. What is meant by stationary process?
2. Define auto correlation function. What are their properties?
3. Explain the components of a time series.
4. What are the limitations of index numbers?
5. Give the formula for Fisher's ideal index. Why it is said to be 'ideal'?
6. Explain ARIMA model.
7. What is meant by deflating of index numbers?
8. What type of bias observed in Laspeyre's and Paasche's index number?
9. What is unit test?
10. Discuss the procedure in Winters smoothing.
11. Distinguish between simple aggregate and weighted aggregate index numbers.
12. Explain the additive and multiplicative model for a time series.
13. Define consumer price index numbers.
14. Explain Dorbish-Bowley's and Kelly's indices.
15. What type of bias observed in Laspeyre's and Paasche's index number?

Part B
Answer at least 5 questions, *Each question carries 5 marks, All questions can be attended overall ceiling 35*

16. Write short on "Time series as discrete parameter stochastic process".
17. Below are given the figures of production (in thousand tons) of a sugar factory:

Year (t):	1999	2000	2001	2002	2003	2004	2005
Production (y_t):	77	88	94	85	91	98	90

Fit a straight line by the method of least squares and obtain the trend values.

18. Write short on the choice of AR and MA periods.
19. Calculate Laspeyre's , Paasche's, Fisher's index number of prices for the following data:

Commodity	Base Year		Current Year	
	Price (p_0)	Quantity (q_0)	Price (p_1)	Quantity (q_1)
A	10	12	12	15
B	7	15	5	20
C	5	24	9	20

D	16	5	14	5

20. Discuss (without proof) the estimation of mean , auto covariance and auto correlation function.
21. Define an index number. What are their uses?
22. Write short on shifting, splicing of index numbers.
23. Compute by Fisher's formula, the price and quantity index numbers from the data given below:

Commodity	Base Year		Current Year	
	Price (in Rs)	Total Value	Total Expenditure	Quantity (in Kg)
1	12	600	2400	120
2	10	1000	960	80
3	14	840	1050	70
4	16	480	900	50
5	18	720	800	40
6	22	1540	900	60
7	20	1800	1600	100
8	15	1200	1440	80

Examine whether this satisfies Factor Reversal test and Time Reversal Test.

Part C (Essay type Questions)

(Answer any 2 question)

Each question carries 10 marks.

24. Disucss Moving Average (MA), Auto Regressive (AR) and ARMA models.
25. Write an essay on "Problems in the Construction of Index Numbers".
26. The following table gives the number of workers employed in a small industry during the years 1996-2005. Calculate the four-yearly moving averages:

Years :	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005
No. of workers:	430	470	450	460	480	470	470	500	490	480

SIXTH SEMESTER U.G DEGREE EXAMINATION
(CBCSS UG)
Applied Statistics with Data Science
ASD6B10 – STATISTICAL QUALITY CONTROL, ACTUARIAL STATISTICS, LINEAR PROGRAMMING AND DEMOGRAPHY

Time: 2 hours and a half **Max: 80Marks**

Part A
(Answer at least 10 questions. Each question carries 2 marks. All questions can be attended. Overall ceiling 25)

1. What are the applications of C charts?
2. What is meant by Statistical Quality Control?.
3. State the causes of producing variation in the quality of a product.
4. What are the advantages of process control?
5. If an investor deposits Rs 40000/- in a bank that pays compound interest at the rate of 6% pa, what will be the maturity value after 5 years?
6. Define force of interest. Calculate 'i' if □ = 7%.
7. Define survival function.
8. Define slack and surplus variables.
9. What are the merits of simplex method?
10. What are the assumptions of LPP?
11. Define Basic Feasible Solution.
12. Define demography.
13. What is meant by abridged life tables?
14. Define Infant Mortality Rate.
15. Define (i) Mortality (ii) Fertility

Part B
Answer atleast 5 questions, *Each question carries 5 marks, All questions can be attended overall ceiling 35*

16. Explain the construction and working of X bar chart and R chart.
17. During an examination of equal lengths of cloth, the following numbers of defects were observed.

    2, 3, 4, 0, 5, 6, 7, 4, 3, 2. Draw appropriate control chart and comment on the process control.

18. An investor deposits Rs 1000 in a saving institution. Each payment is made at the end of the year. If the payment deposited earns 6% interest compounded annually, what amount will he receives at the end of 10 years?.

19. Find the present value at the rate of 6% pa of Rs 300/- payable 5 years hence?
20. (i) What are the major sources of demographic data?
    (ii) Discuss the measures of fertility with illustrations.
21. What are the uses of life tables? State the components of a life table.
22. Solve graphically

    Maximize $z = 3x_1 + 4x_2$
    subject to $4x_1 + 2x_2 \leq 80$
    $2x_1 + 5x_2 \leq 180$
    $x_1, x_2 > 0$
23. Determine initial Basic feasible solution to the following transportation problem using north west corner rule.

	D_1	D_2	D_3	D_4	
O_1	6	4	1	5	14
O_2	8	9	2	7	16
O_3	4	3	6	2	5
	6	10	15	4	

Part C (Essay type Questions)
(Answer any 2 question)
Each question carries 10 marks.

24. Use simplex method to solve the following LPP.

    Maximise $z = x_1 + x_2 + 3x_3$,
    subject to $3x_1 + 2x_2 + x_3 \leq 3$,
    $2x_1 + x_2 + 2x_3 \leq 2$, $x_1, x_2 > 0$.
25. Analyse the following data by means of an X bar chart and R chart and comment whether the process is in state of control. The sample size is 5 for each sample.

Set 1	X bar	R
1	5.08	0.36
2	5.28	0.98
3	4.92	0.64
4	5	0. 78
5	4.82	0.9
6	5.42	1.21
7	5.25	1.05
8	5.35	1.12
9	4.99	1.2
10	5.45	0.88
11	5.34	0.78
12	5.25	0.64
13	4.93	0.95
14	4.85	1.05
15	5.34	1.38

26. Distinguish between Census -0 Vital registration systems and sample registration system. Discuss their merits and demerits.
27. Define Annuity. Explain various types of annuities and also derive its present value expressions.

SIXTH SEMESTER U.G DEGREE EXAMINATION
(CBCSS UG)
Applied Statistics with Data Science
ASD6B11 – DATA MINING TECHNIQUES

Time: 2 hours and a half **Max: 80Marks**

Part A
(Answer at least 10 questions. Each question carries 2 marks. All questions can be attended. Overall ceiling 25)

1. What is sparse data matrix?
2. Summarize the various pre-processing activities in data mining.
3. Explain association rule mining.
4. Which are the measures of selecting best slips?
5. What is minimum description length principle?
6. Explain the attribute selection method in decision trees.
7. Compute the Gini index for the data M, F, M, M, M, M, F, F, F, M, F, M, M.
8. What is an association rule?
9. Define monotonicity property.
10. Explain confidence-based pruning.
11. What is cluster analysis? Give an application of cluster analysis.
12. Compare Hierarchical and Partitional clustering.
13. What is bisecting K-means algorithm?
14. Discuss key issues in hierarchical clustering.
15. Explain term document matrix.

Part B
(Answer at least 5 questions *Each question carries 5 marks, All questions can be attended overall ceiling 35*

16. How is data mining related to business decisions**.**
17. Illustrate different data mining tasks.
18. Let (3,2,1) and (5,6,2) are two points in R^3. Using three different distance measures, find the distance between the points.
19. Briefly explain Hunt's algorithm.
20. Illustrate model overfitting using an example.
21. What is KNN algorithm?
22. Explain the apriori principle.
23. Discuss different types of clusters.

Part C

(*Answer any 2 question,* *Each question carries 10 marks.*

24. Briefly explain different approaches of data preprocessing.
25. Discuss in details different methods for comparing classifiers.
26. State and Explain K-means Algorithm.
27. ExplainText mining. Discuss various methods for text mining.