SRI DHARMASTHALA MANJUNATHESHWARA COLLEGE (AUTONOMOUS)
UJIRE – 574 240
DAKSHINA KANNADA, KARNATAKA STATE
(Re-Accredited by NAAC at 'A++' Grade)
Ph: 08256-236221, 236101(O), Fax: 236220
e-Mail:sdmcollege@rediffmail.com, sdmcollege@sdmcujire.in, Website: www.sdmcujire.in

DEPARTMENT OF PG STUDIES & RESEARCH IN STATISTICS

Syllabus of

Master's Degree in

STATISTICS

CHOICE BASED CREDIT SYSTEM (CBCS)

SEMESTER SCHEME

(2016-2017 onwards)

BOS meeting held on 18-08-2023

Academic Council meeting held on 02-09-2023

Preamble:

Revision of the Syllabus for the Two years Master Degree (Choice Based Credit System – Semester Scheme) Programme in Statistics.

The PG BOS in Statistics has prepared the revised Syllabus for M.Sc. Statistics (CBCS based) in its meeting held on 5th September 2019, as per the guidelines suggested by Mangalore University and University Grants Commission, New Delhi. It was resolved to implement this new syllabus from the academic year 2020-21.

In the present revised syllabus, the suggested course pattern includes Hard Core, Soft Core and Open Elective courses with 92 credits for the entire programme. The syllabus consists of 14 Hard Core courses (4 credits each) including 11 theory (3 in I, II, III, and 2 in IV semesters), 2 practicals (in I semester) and one Project work (in IV semester), with a total of 56 credits. It also consists of 10 Soft Core courses (3 credits each) including 5 theory (1 in I, II, III, and 2 in IV semesters) and 5 practicals (2 in II, III, and 1 in IV semesters), with a total of 30 credits. The BOS has also proposed 2 Open Elective courses (1 each in II and III semesters) with 3 credits each (with a total of 6 credits), to be offered to non-Statistics students. But the credits of Open Elective courses are not considered for CGPA. All together total credits come to 92 (including the credits for Open Elective courses), otherwise, a total of 86 credits.

Faculty of PG Studies in Statistics: PGSTAT056

Programme Specific Outcomes:

PSO1: Show the ability to use the knowledge on theoretical foundations for the development of various statistical concepts and procedures.

PSO2: Develop technical skills in probability modelling and statistical inference for the practical application of statistical methods in their future employment.

PSO3: Be able to find solutions to real world problems by applying quantitative modelling and data analysis techniques.

PSO4: Exhibit the skills in the use of computational and statistical software to develop and execute various statistical techniques and statistical computing algorithms.

PSO5: Demonstrate theoretical knowledge and applications of parametric, semi-parametric and non-parametric testing procedures.

PSO6: Design experiments and surveys with a view of providing solutions to real life problems.

PSO7: Be able to use statistical reasoning, formulate a problem in statistical terms, perform exploratory analysis of data, and carry out a variety of advanced inferential procedures.

PSO8: Be familiar with tackling emerging problems through applications of statistics.

Course Objective:

- To prepare the younger generation to tackle the emerging problems through applications of statistics.
- To develop the department as a hub for teaching, research and extension activities in the field of modern statistical theory and methods.

List of Hardcore papers:

i. Real Analysis
ii. Probability and Distributions – I
iii. Theory of Sampling
iv. Probability and Distributions – II
v. Design and Analysis of Experiments
vi. Theory of Estimation
vii. Theory of Testing of Hypothesis
viii. Regression Analysis
ix. Multivariate Analysis
x. Time Series Analysis
xi. Reliability and Survival Analysis

List of Softcore papers:

i. Linear Algebra
ii. Data Management and Statistical Computing with Python
iii. Stochastic Processes

iv. Statistical Modelling
v. Big Data Analytics
vi. Artificial Intelligence
vii. Elements of Statistical Computing
viii. Survival Analysis
ix. Stochastic Finance
x. Data Mining
xi. Bayesian Inference
xii. Statistical Methods for Reliability
xiii. Nonparametric Inference
xiv. Actuarial Methods
xv. Pattern Recognition and Image Processing
xvi. Operations Research

Value Added Courses (Certificate Courses):

i. Microsoft Excel (Basic to Advance)
ii. R and Excel for Data Science
iii. R for Data Science
iv. R for Advanced Statistical Methods & Machine Learning

Course Pattern for M.Sc. Statistics Programme

I Semester:

Course Code	Title of the Paper	Hrs/Week	Credits
STH 411	Real Analysis	4	4
STS 412	Linear Algebra	3	3
STH 413	Probability and Distributions - I	4	4
STH 414	Theory of Sampling	4	4
STP 415	Practical based on R programming and STH 414	8	4
STP 416	Practical based on Programming in Python	8	4
	Mini project**		
		Total	**23**

****Mini project will be incorporated as 'Skill Component'. II Semester:**

Course Code	Title of the Paper	Hrs/Week	Credits
STE 421 A/B/C	Introductory Statistics and Data Analysis/ Questionnaire Design and Sample Selection/ Data Visualization (Open Elective Course)	3	3*
STH 422	Probability and Distributions - II	4	4
STH 423	Design and Analysis of Experiments	4	4
STH 424	Theory of Estimation	4	4
STS 425	Data Management and Statistical Computing with Python	3	3
STP 426	Practical based on STH 423 and STH 424 using R	6	3
STP 427	Practical based on STS 425	6	3
	Mini project**		
		Total	**21+3***

***This credit is not included for CGPA.**

****Mini project will be incorporated as 'Skill Component'.**

III Semester:

Course Code	Title of the Paper	Hrs/Week	Credits
STE 531 A/B/C	Inferential Statistics and Data Analysis/ Categorical Data Analysis/ Demographic Methods and Analysis (Open Elective Course)	3	3*
STH 532	Theory of Testing of Hypothesis	4	4
STH 533	Regression Analysis	4	4
STH 534	Multivariate Analysis	4	4
STS 535	Stochastic Processes	3	3
STP 536	Practical based on STH 532 and STH 533 using R	6	3
STP 537	Practical based on Machine Learning with Python	6	3
	Mini project**		
		Total	**21+3***

***This credit is not included for CGPA.**

****Mini project will be incorporated as 'Skill Component'. IV Semester:**

Course Code	Title of the Paper	Hrs/Week	Credits
STH 541	Time Series Analysis	4	4
STH 542	Reliability and Survival Analysis	4	4
STS 543	Statistical Modelling	3	3
STS 544	Big Data Analytics	3	3
STP 545	Practical based on STH 541, STH 542 and STS 543 using R	6	3
STP 546	Project Work**	8	4
		Total	**21**

****Project Work will involve 'Skill Component'.**

Scheme of Internal Assessment Evaluation

The scheme of evaluation for internal assessment marks shall be as follows:

i.	Two Internal Tests	20 marks
ii.	Seminar/Assignments/Classroom Activities etc.	10 marks
	Total:	**30 marks**

Question Paper Pattern

The pattern of question paper in theory examinations shall be as follows:

i. There shall be totally 8 questions in which Q. No. 1 is compulsory. Students have to answer any 4 questions from the remaining 7 questions.

ii. Q. No. 1 will contain 8 questions of short answer type, each question carrying 3 marks. Students will have to answer any 6 questions. Thus Q. No. 1 carries 18 marks.

iii. Q. No. 2 to Q. No. 8 will be of long answer type, each question carrying 13 marks.

The distribution of marks will be as follows:

Q. No. 1	3×6 = 18
Any 4 questions out of remaining 7 questions	13×4 = 52
	Total = 70

I Semester

STH 411 - REAL ANALYSIS (4 Credits)

<u>Rationale/Learning Objectives</u>:

- To gain knowledge on the fundamental aspects involved in the theory of real analysis.
- To explore the concept of limits and its usage in sequences and series of real numbers.
- To learn the concept of convergence and its implementation in sequences and series of functions.
- To understand the theory involved in the concept of functions of two variables.

<u>Outcomes</u>:

- Ability to understand the fundamental properties of real numbers that lead to the formal development of real analysis.
- Understand the concept of limits and how they are used in sequences and series of real numbers.
- Understand the concept of convergence and how they are implemented in sequences and series of functions.
- Understand the implementation of theoretical aspects involved in functions of two variables.

Syllabus

- **<u>Unit 1</u>**: Elements of set theory, sets in Euclidean space of k-dimensional R^k rectangles. Metric spaces, neighbourhood, interior point and limit point, open and closed sets, Bolzano-Weierstrass theorem in R^2, compact set, real-valued functions, Heine-Borel theorem (Statement only), continuity and uniform continuity. (13hrs)
- **<u>Unit 2</u>**: Sequences and Series of real numbers - Cauchy sequence, convergence of bounded monotone sequence. Limit superior, limit inferior and limit properties. Series of positive terms - tests for convergence, divergence. Series of arbitrary terms - absolute and conditional convergence. (13hrs)
- **<u>Unit 3</u>**: Sequences of functions - uniform convergence and point wise convergence, series of functions - uniform convergence, Weierstrass' M test. Power series and radius of convergence. Reimann-Stieltjes integration continuous integrand and monotonic/differentiable integrator. (13hrs)

- **Unit 4**: Functions of two variables - partial and directional derivatives. Maxima and minima of functions, maxima-minima under constraints (Lagrange's multipliers). Improper integrals. (13hrs)

Books for Reference:

1. Apostol, T. M. (1985). *Mathematical Analysis*. Narosa India Ltd.
2. Bartle, R. G. (1975). *The Elements of Real Analysis* (2nd ed.). Wiley.
3. Courant, R. and John, F. (1965). *Introduction to Calculus and Analysis*. Wiley.
4. Goldberg, R. R. (1970). *Methods of Real Analysis*. Oxford Publishing Co.
5. Khuri, A. T. (1993). *Advanced Calculus with Applications in Statistics*. John Wiley.
6. Rudin, W. (1976). *Principles of Mathematical Analysis*. McGraw Hill.

STS 412 - LINEAR ALGEBRA (3 Credits)

Rationale/Learning Objectives:

- To gain knowledge on the theoretical foundations involved in the concepts of field and vector geometry.
- To study about the computational techniques involved in the theory of matrix algebra, linear transformations, and systems of linear equations.
- To explore the concepts of eigen values and eigen vectors and its importance in matrix theory.
- To understand the theory and applications involved in the concepts of quadratic forms, vector and matrix differentiation.

Outcomes:

- Awareness of necessary theoretical foundations on field and vector geometry, which will help them better understand linear models and multivariate analysis.
- Develop algebraic skills and knowledge on computational techniques essential for the study of matrix algebra, linear transformations, and systems of linear equations.
- Understand the concepts of eigen values and eigen vectors and their implementation in matrix theory.
- Understand the theoretical aspects of quadratic forms, vector and matrix differentiation, with their practical implementation.

Syllabus

- **Unit 1**: Fields, vector spaces, subspaces, linear dependence and independence, basis and dimension of a vector space, finite dimensional vector spaces, completion theorem. Examples of vector spaces over real fields. Vector spaces with an inner product, Gram-Schmidt orthogonalization process, orthonormal basis. (10hrs)
- **Unit 2**: Row and column spaces of a matrix. Rank and inverse of a matrix, properties of inverse. Rank of a product of matrices, partitioned submatrices, rank factorization of a matrix, rank of a sum, inverse of a partitioned matrix. General linear system of equations, generalized inverse, Moore-Penrose inverse, idempotent matrices. Solutions of matrix equations. (10hrs)
- **Unit 3**: Characteristic roots and vectors, Cayley-Hamilton theorem, minimal polynomial, similar matrices. Algebraic and geometric multiplicity of characteristic roots, spectral decomposition of a real symmetric matrix, reduction of a pair of real symmetric matrices, singular value decomposition. (10hrs)

- **Unit 4**: Real quadratic forms, reduction and classification of quadratic forms, index and signature. Extrema of quadratic forms. Vector and matrix differentiation.

(10hrs)

Books for Reference:

1. Biswas, S. (1984). *Topics in Algebra of Matrices*. Academic Publications.
2. Hadley, G. (1987). *Linear Algebra*. Narosa.
3. Graybill, F. A. (1983). *Matrices with Applications in Statistics*.
4. Rao, A. R. and Bhimasankaran, P. (1992). *Linear Algebra*. Tata McGraw Hill.
5. Rao, C. R. (1973). *Linear Statistical Inference and its Applications* (2nd ed.). Wiley.
6. Rao, C. R. and Mitra, S. K. (1971). *Generalized Inverse of Matrices and its Applications*. Wiley.
7. Searle, S. R. (1982). *Matrix Algebra Useful for Statistics*. Wiley.

Practicals based on R Programming

1. Fundamentals of R
2. Graphical analysis using R
3. Descriptive statistics using R
4. Functions in R
5. Probability distributions in R
6. Gram-Schmidt orthogonalization process
7. Fundamentals of matrix algebra
8. Rank of a matrix using echelon form
9. Generalized inverse of a matrix
10. Inverse of a matrix by the method of partitioning
11. Spectral decomposition of a real symmetric matrix
12. Principal minors of a symmetric matrix

STH 413 - PROBABILITY AND DISTRIBUTIONS - I (4 Credits)

Rationale/Learning Objectives:

- To understand the necessary theoretical foundations required for the development of probability theory.
- To gain knowledge on the fundamentals and principles involved in probability theory.
- To study about the characteristics of various discrete and continuous univariate distributions.
- To learn about various transformation techniques, order statistics, truncated and mixed distributions.

Outcomes:

- Familiarity with necessary theoretical foundations on the developments of probability theory.
- Ability to understand the fundamental aspects and principles of probability theory.
- Exhibit learning about the standard discrete and continuous univariate distributions and its characteristics.
- Show improved knowledge on various transformation techniques, order statistics, truncated and mixed distributions.

Syllabus

- **Unit 1**: Algebra of sets, sequence of sets and limits, fields and sigma-fields, minimal sigma-field. Events, sample space. Probability measure, probability space, property of probability measure, properties related to sequences of events, independent events, conditional Probability. (13hrs)
- **Unit 2**: Measurable functions, random variables, probability induced by a random variable. Definition of simple random variables. Integration of measurable functions with respect to measures. Expectation, properties of expectation, moments, inequalities. (16hrs)
- **Unit 3**: Standard discrete and continuous univariate distributions and their properties, probability generating function and moment generating function. Bivariate normal and multinomial distributions. Transformation techniques. Distribution of functions of random variables. (13hrs)

- **<u>Unit 4</u>**: Order Statistics - their distributions and properties, joint and marginal distributions. Distribution of range and median. Truncated and mixture distributions. (10hrs)

Books for Reference:

1. Ash, R. B. and Catherine Doleans-Dade (2000). *Probability and Measure Theory*. Academic Press.
2. Bhat, B. R. (1999). *Modern Probability Theory* (3rd ed.). New Age International Publishers.
3. Johnson, S. and Kotz (1972). *Distributions in Statistics*. Vols. I, II and III, Houghton and Miffin.
4. Mukhopadhyaya, P. (1996). *Mathematical Statistics*. Calcutta Publishing House.
5. Pitman, J. (1993). *Probability*. Narosa.
6. Rao, C. R. (1973). *Linear Statistical Inference and its Applications* (2nd ed.). Wiley Eastern.
7. Rohatgi, V. K. and Saleh, A. K. Md. E. (2015). *An Introduction to Probability Theory and Mathematical Statistics*. Wiley Eastern.
8. Laha, R. G. and Rohatgi, V. K. (1979). *Probability Theory*. Wiley Eastern.
9. Ross, S. M. (1993). *First Course in Probability*. Academic Press.
10. Billingsley, P. (1986). *Probability and Measure*. John Wiley and Sons.

STH 414 - THEORY OF SAMPLING (4 Credits)

Rationale/Learning Objectives:

- To explore theoretical foundations involved in the development of sampling theory and PPSWR sampling technique.
- To gain knowledge on the theoretical concepts involved in the development of PPSWOR sampling technique.
- To learn the theory and applications of single stage cluster sampling and two stage sampling.
- To study about two phase sampling and also regarding randomized response techniques and official statistics for national development.

Outcomes:

- Ability to understand the theoretical aspects involved in the development of sampling theory and PPSWR sampling technique.
- Exhibit theoretical knowledge on various concepts involved in the development of PPSWOR sampling technique.
- Ability to understand the theory of single stage cluster sampling and two stage sampling with their real life applications.
- Understand the theory involved in two phase sampling and explore about randomized response techniques and official statistics for national development.

Syllabus

- **Unit 1**: Basic Concepts - sampling design, sampling scheme, sampling strategy, interpenetrating subsampling, concept of non-random sampling. Probability proportional to size with replacement (PPSWR) sampling - selection of PPSWR sample, estimation of population mean, total and their sampling variances. Hansen-Hurwitz strategy, estimation of sampling variance. Comparison with SRSWR, estimation of gain due to PPSWR sampling. (13hrs)
- **Unit 2**: Varying probability without replacement (PPSWOR) sampling - some properties of sampling design. Horwitz-Thomson estimator, sampling variance of population total and its unbiased estimator. Sen-Midzuno sampling scheme, Des-Raj's ordered estimator (general case), Murthy's unordering principle (sample of size two). (13hrs)

- **<u>Unit 3</u>**: Single stage cluster sampling - concepts, estimation of efficiency of cluster sampling, clusters of varying sizes. Two-stage sampling - notions, estimation of population total and its variance, efficiency of two-stage sampling relative to cluster and uni-stage sampling. (13hrs)
- **<u>Unit 4</u>**: Ratio and regression estimators based on SRSWOR, method of sampling, bias and mean square errors, comparison with mean per unit estimator. Two-phase sampling - notion, double sampling for ratio estimation, double sampling for regression estimation. Randomized response techniques - Warner's model, related and unrelated questionnaire methods, non-sampling errors. Official Statistics for National Development - NSO, CSO, MOSPI, Human Development Index. Measuring inequality in income - Lorenz Curve, Gini coefficient. (13hrs)

Books for Reference:

1. Cochran, W. G. (1977). *Sampling Techniques* (3rd ed.). Wiley.
2. Des Raj and Chandok (1998). *Sampling Theory*. Narosa Publication.
3. Mukhopadhyay, P. (1998). *Theory and Methods of survey Sampling*. Prentice Hall of India.
4. Murthy, M. N. (1977). *Sampling Theory and Methods*. Calcutta: Statistical Publishing Society.
5. Sampath, S. (2001). *Sampling Theory and Methods*. Narosa Publishers.
6. Sen, A. (1997). *Poverty and Inequality*.
7. Singh, D. and Chaudhary, F. S. (1986). *Theory and Analysis of Sample Survey Designs*. New Age International Publishers.
8. Sukhatme, P. V., Sukhatme, B. V., Sukhatme, S. and Ashok (1984). *Sampling Theory of Surveys with Applications*. ICAR Publication.
9. Banett, V. (2002). *Sample Survey: Methods and Principles*. Arnold Publishers.

Practicals based on Theory of Sampling

1. Determination of sample size
2. Probability proportional to size with replacement (PPSWR) - I
3. Probability proportional to size with replacement (PPSWR) - II
4. Probability proportional to size without replacement (PPSWOR) - I
5. Probability proportional to size without replacement (PPSWOR) - II
6. Des Raj's ordered estimator and Murthy's unordered estimator

7. Cluster sampling with clusters of equal size
8. Cluster sampling with clusters of unequal size
9. Two-stage sampling
10. Ratio and regression method of estimation

Practicals based on Programming in Python

1. Program to perform arithmetic operations and display the output using placeholders.
2. Programs based on decision making statements.
3. Loop programming exercises.
4. Programs based on probability and statistics.
5. Programs to perform mathematical operations.
6. Programs for matrix operations.

II Semester

STE 421A - INTRODUCTORY STATISTICS AND DATA ANALYSIS

(Open Elective Course)

Rationale/Learning Objectives:

- To explore different techniques involved in descriptive statistics and their practical applications.
- To study about the characteristics and applications of binomial, Poisson and normal distributions.
- To understand the real life applications of various sampling techniques.

Outcomes:

- Ability to explain the various techniques involved in descriptive statistics and their implementation in real life problems.
- Show an improved knowledge about standard univariate distributions such as binomial, Poisson and normal, and understand its applications.
- Ability to apply various sampling techniques in real life.

Syllabus (3 Credits)

- **Unit 1**: Statistics - introduction, meaning, definition and scope of the subject as a science of decision making against uncertainty. Data types, methods of collection, presentation in the form of tables and graphs. Descriptive Statistics - measures of central tendency, positional averages, measures of dispersion, skewness and kurtosis. Methods of summarizing categorical data - univariate and bivariate contingency tables. Box plots - construction and interpretations. Exploratory data analysis using descriptive measures and graphical tools. (13hrs)
- **Unit 2**: The concept of random experiment, simple events, sample space, types of events, probability of an event, rules of probability, conditional probability, Baye's rule, exercises on computation of probabilities using these rules to fix the ideas. The concept of random variables - discrete and continuous type, Binomial, Poisson and Normal distributions - their use in practical applications, computing probabilities using these distributions. (13hrs)
- **Unit 3**: Sampling methods - population and sample, parameter and statistic, concept of a random sample, simple random sampling, stratified sampling, systematic, sampling,

sample size determination. The concept of sampling distribution of a statistic and standard error. (14hrs)

Books for Reference:

1. Campbell, R. C. (1974). *Statistics for Biologists*. Cambridge University Press.
2. Chatfield, C. (1981). *Statistics for Technology*. Chapman and Hall.
3. Frank, H. and Athoen, S. C. (1997). *Statistics: Concepts & Applications*. Cambridge University Press.
4. Medhi, J. (1992). *Statistical Methods: An Introductory Text*. Wiley Eastern Limited.
5. Ross, S. M. (2017). *Introductory Statistics*. Academic Press.
6. Rice, J. A. (2006). *Mathematical Statistics and Data Analysis*. Singapore: Thomson-Duxbury.

STE 421B - QUESTIONNAIRE DESIGN AND SAMPLE SELECTION

(Open Elective Course)

Rationale/Learning Objectives:

- To design questionnaires in a way that make them attractive towards target respondents.
- To explore different methods of primary data collection.
- To understand the practical implementation of various sampling techniques.

Outcomes:

- Ability to formulate a relevant questionnaire for real life problems.
- Able to understand and implement different methods of data collection through questionnaire.
- Ability to implement relevant sampling techniques in real life.

Syllabus (3 Credits):

Unit 1: Introduction, qualities of a good questionnaire, types of questionnaires: exploratory questionnaire (qualitative) and formal standardized questionnaire (quantitative). Questionnaire question types: open-ended questions, multiple choice questions, dichotomous questions, scaled questions, and pictorial questions; questions to avoid in a questionnaire. Checking the validity and reliability of questionnaire. (14hrs)

Unit 2: Steps involved in the development of a questionnaire, methods of reaching target respondents: personal interviews, group or focus interviews, mailed questionnaires, and telephone interviews. Advantages and disadvantages of questionnaires. Examples of questionnaires. Introduction to pilot surveys and its use in questionnaire development and modification. (13hrs)

Unit 3: Sampling: introduction, techniques: probability sampling - simple random, systematic, stratified, and cluster sampling; non-probability sampling - convenience, quota, judgement, and snowball sampling. Applications of sampling. Sample size determination. (13hrs)

Reference:

1. http://www.fao.org/3/w3241e/w3241e05.htm
2. https://www.kyleads.com/blog/questionnaire/
3. https://www.digitalvidya.com/blog/sampling-techniques/

STE 421C - DATA VISUALIZATION (Open Elective Course)

Rationale/Learning Objectives:

- To learn about how to develop simple summaries and exploratory graphs for univariate data.
- To explore different visualization techniques for bivariate data.
- To gain knowledge on multivariate data visualization for real life problems.

Outcomes:

- Ability to formulate and visualize univariate data.
- Able to implement various visualization techniques for bivariate data in real life.
- Capable to obtain multidimensional graphs for real life applications.

Syllabus (3 Credits):

- **Unit 1**: Tables and Univariate Graphs: Introduction, Tables, Q-Q Plot, Categorical (Bar chart, Pie chart, Tree map), Quantitative (Histogram, Frequency polygon, Frequency Curve, Ogives, Stem and leaf plot, Dot chart). Applications and examples. (12hrs)
- **Unit 2**: Bivariate Graphs: Categorical vs. Categorical (Stacked bar chart, Grouped bar chart, Segmented bar chart), Quantitative vs. Quantitative (Scatterplot, Line plot, Area chart), Categorical vs. Quantitative (Bar chart on summary statistics, Box plots). Applications and examples. (14hrs)
- **Unit 3**: Multivariate Graphs and Statistical Models}: Grouping, Faceting, Correlation plots, Linear Regression, Scatter Plot Matrix, Parallel coordinates plot, Star plot, Chernoff faces, Growth curve. Applications and examples. (14hrs)

Reference:

1. https://rkabacoff.github.io/datavis/
2. https://www.analyticsvidhya.com/blog/2015/07/guide-data-visualization-r/
3. https://towardsdatascience.com/a-guide-to-data-visualisation-in-r-for-beginners-ef6d41a34174
4. https://www.tutorialspoint.com/excel_data_analysis/excel_data_analysis_visualization.htm

STH 422 - PROBABILITY AND DISTRIBUTIONS - II (4 Credits)

Rationale/Learning Objectives:

- To learn necessary theoretical foundations of probability and measure theory.
- To explore the concept of convergence and its theoretical aspects.
- To understand the importance of weak law of large numbers.
- To study about characterization properties of the distributions, central limit theorem, and sampling distributions.

Outcomes:

- Ability to understand the conceptual basis for the probability and measure theory.
- Understand the theoretical aspects of convergence and its associated concepts.
- Ability to understand the theory of weak law of large numbers.
- Gain knowledge about characterization properties of the distributions, along with the concepts of central limit theorem and various sampling distributions.

Syllabus

- **Unit 1**: Measure, probability measure, properties of a measure and probability, Carathedory extension theorem (statement only). Lebesgue and Lebesgue-Stieltjes measure on the real line. Absolute continuity, definition of Radon-Nikodym derivative and illustrations. (13hrs)
- **Unit 2**: Monotone convergence theorem, Fatou's lemma and dominated convergence theorem. Borel-Cantelli lemma, convergence in probability, convergence almost surely, convergence in distribution, convergence in r^{th} mean, convergence theorem for expectations. Slutsky's theorem. (13hrs)
- **Unit 3**: Weak law of large numbers - Kolmogorov's generalized WLLN (proof of sufficient condition only), Khintchine's WLLN as special case, Chebyshev's WLLN. Kolmogorov's strong law of large number sequence of independent and iid random variables. Kolmogorov's inequality. (13hrs)
- **Unit 4**: Characteristic function - properties, inversion theorem (statement only and proof for density version), uniqueness theorem, continuity theorem (statement only). Central limit theorem, Lindeberg-Levy and Liapounov central limit theorems. Statement of Lindeberg-Feller form (statement only). Application of these theorems. Sampling distributions, chi-square, t, F and non-central chi-square, mgf of non-central

chi-square distribution, reproductive property. Non-central t and non-central F.

(13hrs)

Books for Reference:

1. Bhat, B. R. (1999). *Modern Probability Theory* (3rd ed.). New Age International Publishers. (To be used as Text.)
2. Mukhopadhyaya, P. (1996). *Mathematical Statistics*. Calcutta Publishing House.
3. Pitman, J. (1993). *Probability*. Narosa.
4. Billingsley, P. (1986). *Probability and Measure*. John Wiley and Sons.
5. Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. Wiley.
6. Ash, R. B. and Catherine Doleans-Dade (2000). *Probability and Measure Theory*. Academic Press.
7. Athreya, K. B. and Lahiri, S. N. (2006). *Measure Theory and Probability Theory*.
8. Rao, C. R. (1973). *Linear Statistical Inference and its Applications* (2nd ed.). Wiley Eastern.
9. Rohatgi, V. K. and Saleh, A. K. Md. E. (2015). *An Introduction to Probability Theory and Mathematical Statistics*. Wiley Eastern.

STH 423 - DESIGN AND ANALYSIS OF EXPERIMENTS (4 Credits)

Rationale/Learning Objectives:

- To study about the fundamentals and principles involved in designed experiments.
- To gain knowledge on the theory and applications of complete and incomplete block designs.
- To learn about balanced incomplete block design, fixed, mixed and random effects models.
- To explore the real life applications of factorial experiments, complete and partial confounding, nested designs, split-plot designs, and strip-plot designs.

Outcomes:

- Demonstrate necessary theoretical foundations on the fundamentals and principles involved in designed experiments.
- Exhibit theoretical knowledge and real life applications of complete and incomplete block designs.
- Ability to understand the theory and applications of balanced incomplete block design, fixed, mixed and random effects models.
- Ability to understand the importance and applications of various experimental designs such as factorial experiments, complete and partial confounding, nested designs, split-plot designs, and strip-plot designs, in analyzing real life problems.

Syllabus

- **Unit 1**: Gauss-Markov setup, normal equations and least squares estimates, estimable function and estimation space, variance and covariance of least squares estimates, estimation of error variance, estimation with correlated observations, simultaneous estimates of linear parametric functions. Tests of hypothesis for one and more than one linear parametric functions. Confidence intervals and regions, analysis of variance, power of F-test, multiple comparison tests - Tukey and Bonferroni, simultaneous confidence interval. (13hrs)
- **Unit 2**: Introduction to designed experiments, general block design - complete block design, incomplete block design and its information matrix, criteria for connectedness, balance and orthogonality. Intra-block analysis - estimability, best point estimates/interval estimates of estimable linear parametric functions and testing of linear hypotheses, estimation of parameters. (13hrs)

- **<u>Unit 3</u>**: BIBD - definition, concept of connectedness, balancing, properties, estimability, recovery of inter-block information. Analysis of covariance in a general Gauss-Markov model, applications to CRD and RCBD. Fixed, mixed and random effects models, variance components estimation, study of various methods. (13hrs)
- **<u>Unit 4</u>**: General factorial experiments, factorial effects - best estimates and testing the significance of factorial effects, study of 2^n and 3^n factorial experiments in randomized blocks. Complete and partial confounding. Nested designs. Split-plot, strip plot designs. (13hrs)

Books for Reference:

1. Bapat, R. B. (2012). *Linear Algebra and Linear Models*. Hindustan Book Agency.
2. Rao, C. R. (1973). *Linear Statistical Inference and its Applications*. Wiley Eastern.
3. Aloke Dey (1986). *Theory of Block Designs*. Wiley Eastern.
4. Dean, A. and Voss, D. (1999). *Design and Analysis of Experiments*. Springer.
5. Chakrabarti, M. C. (1962). *Mathematics of Design and Analysis of Experiments*. Asia.
6. Cochran and Cox, D. R. (1957). *Experimental Designs*. John Wiley.
7. Das, M. N. and Giri, N. (1979). *Design and Analysis of Experiments*. Wiley Eastern.
8. Giri, N. (1986). *Analysis of Variance*. South Asian Publishers.
9. John, P. W. M. (1911). *Statistical Design and Analysis of Experiments*. Macmillan.
10. Joshi, D. D. (1987). *Linear Estimation and Design of Experiments*. Wiley Eastern.
11. Montgomery, C. D. (1976). *Design and Analysis of Experiments*. New York: Wiley.
12. Mukhopadhyay, P. (1998). *Applied Statistics*. Books and Allied (P) Ltd.
13. Pearce, S. C. (1984). *Design of Experiments*. New York: Wiley.
14. Rao, C. R. and Kleffu, J. (1988). *Estimation of Variance Components and Applications*. North Holland.
15. Searle, S. R., Casella, G. and McCullugh, C. E. (1992). *Variance Components*. Wiley.

Practicals based on Design and Analysis of Experiments

1. One way analysis of variance
2. Two way analysis of variance
3. Least squares estimation
4. Analysis of incomplete block design
5. Analysis of balanced incomplete block design
6. Analysis of covariance
7. Analysis of factorial experiment
8. Analysis of 2^n factorial experiment
9. Analysis of 3^n factorial experiment
10. Confounding in 2^n factorial experiment
11. Split-plot design

STH 424 - THEORY OF ESTIMATION (4 Credits)

Rationale/Learning Objectives:

- To study about the fundamentals of estimation theory.
- To learn the theory and practical implementation of minimum variance unbiased estimation and confidence interval construction.
- To explore the concept of consistency of estimators.
- To understand the theory and applications of maximum likelihood estimation.

Outcomes:

- Develop necessary theoretical foundations on the fundamental aspects involved in estimation theory.
- Able to understand and apply the concepts associated with minimum variance unbiased estimation and also explore about confidence interval construction.
- Ability to apply the concepts associated with consistency of estimators.
- Gain knowledge about the theory and applications of maximum likelihood estimation.

Syllabus

- **Unit 1**: Parametric models, likelihood function, example from standard discrete and continuous models. Plotting likelihood functions. Sufficiency, Neyman factorization criterion, Fisher information for single and several parameters. Minimal sufficient statistic, likelihood equivalence. Exponential families and Pitman families. Completeness, Ancillary Statistics, Basu's theorem and applications. (13hrs)
- **Unit 2**: Minimum variance unbiased estimation, unbiasedness, locally unbiased estimators, minimum variance, locally minimum various, mean squared error, Cramer-Rao lower bound approach. Minimum variance unbiased estimators (MVUE), Rao-Blackwell theorem, completeness, Lehman-Scheffe's theorem, necessary and sufficient condition for MVUE. Bhattacharya bounds (without proof). Introduction to interval estimation, construction of confidence intervals using pivot. (13hrs)
- **Unit 3**: Consistency, estimation of real and vector valued parameters, invariance properties. Consistency of estimators by method of moments and method of percentiles, mean squared error criterion, asymptotic relative efficiency, consistent asymptotic normal (CAN). (13hrs)
- **Unit 4**: Method of maximum likelihood - notion, MLE in location and scale family, exponential family, Cramer family (statement only). Cramer-Huzurbazar theorem.

Solutions to likelihood equations method of scoring, Newton-Raphson and other iterative procedures. Fisher lower bound to asymptotic variance, extension to multi-parameter case (without proof). (13hrs)

Books for Reference:

1. Casella, G. and Berge, R. L. (2002). *Statistical Inference* (2nd ed.). Singapore: Thomson-Duxbury.
2. Kale, B. K. (1999). *A First Course on Parametric Inference*. Narosa Publishing House.
3. Lehman, E. L. (1986). *Theory of Point Estimation*. John Wiley.
4. Rao, C. R. (1973). *Linear Statistical Inference and its Applications*. Wiley Eastern.
5. Rohatgi, V. K. and Saleh, A. K. L. (2001). *An Introduction to Probability and Mathematical Statistics*. Wiley Eastern.
6. Rajagopalan, M. and Dhanavanthan, P. (2012). *Statistical Inference*. Phi Learning Pvt. Ltd.
7. Zacks, S. (1981). *Parametric Statistical Inference*. Pergamon Press.

Practicals based on Theory of Estimation

1. Estimation by the methods of moments and percentile
2. Uniformly minimum variance unbiased estimator
3. Unbiased estimation
4. Estimation by the method of maximum likelihood - I
5. Estimation by the method of maximum likelihood - II

STS 425 - Data Management and Statistical Computing with Python (3 Credits)

Rationale/Learning Objectives:

- To learn the fundamental aspects involved in Python programming paradigms.
- To gain knowledge on how to use pandas for data manipulation.
- To study about how to generate data visualizations using Python.

Outcomes:

- Be able to gain comprehensive knowledge on the fundamentals involved in Python programming paradigms.
- Exhibit insights on using Pandas in Python required for data manipulation.
- Be able to explore about how to generate powerful data visualizations using Python.

Syllabus

- **Unit 1**: Using Numpy - Basics of NumPy-Computation on NumPy-Aggregations-Computation on Arrays Comparisons, Masks and Boolean Arrays-Fancy Indexing-Sorting Arrays-Structured Data: NumPy's Structured Array. (14hrs)
- **Unit 2**: Data Manipulation with Pandas - Introduction to Pandas Objects-Data indexing and Selection-Operating on Data in Pandas-Handling Missing Data-Hierarchical Indexing - Combining Data Sets. (14hrs)
- **Unit 3**: Visualization and Matplotlib - Basic functions of matplotlib-Simple Line Plot, Scatter Plot-Density and Contour Plots-Histograms, Binnings and Density-Customizing Plot Legends, Colour Bars-Three-Dimensional Plotting in Matplotlib. (12hrs)

Books for Reference:

1. VanderPlas, J. (2016). *Python Data Science Handbook - Essential Tools for Working with Data*. O'Reily Media, Inc.
2. Zhang, Y. (2016). *An Introduction to Python and Computer Programming*. Springer Publications.
3. Thareja, R. *Python Programming using Problem Solving Approach*. Oxford University Press.

4. Grus, J. (2016). *Data Science from Scratch First Principles with Python*. O'Reilly Media.
5. Padmanabhan, T. R. (2016). *Programming with Python*. Springer Publications.

Practicals based on Data Management and Statistical Computing with Python

1. Programs to perform a number of mathematical operations on arrays such as trigonometric, statistical and algebraic routines.
2. Programs to perform fast vectorized array operations for data munging and cleaning, subsetting and filtering, transformation, and any other kinds of computations.
3. Exercises to handle n-dimensional arrays, broadcasting, performing operations, data generation, etc.
4. Programs to work with arrays, queries, and dataframes.
5. Programs to perform the groupby, merge, and join methods in Pandas.
6. Exercises using matplotlib to generate various plots.
    a. Plot charts, individually and in multiples.
    b. Customize the style and appearance of different plot components.
    c. Choose different chart types based on data type and requirement.
    d. Show the distribution of data.
    e. Customize objects in Matplotlib.
    f. Plot data in 2D and 3D.

III Semester

STE 531A - INFERENTIAL STATISTICS AND DATA ANALYSIS

(Open Elective Course)

<u>Rationale/Learning Objectives</u>:

- To understand the theory and applications of various parametric and nonparametric hypothesis tests.
- To study the practical implementation of one-way analysis of variance test.
- To explore the real life applications of correlation and regression techniques.

<u>Outcomes</u>:

- Able to identify the basics of hypothesis testing and perform parametric and nonparametric hypothesis tests.
- Able to conduct one-way analysis of variance hypothesis test.
- Apply correlation and regression techniques to real life problems.

Syllabus (3 Credits)

- **<u>Unit 1</u>**: The concept of hypothesis and tests of hypothesis: null hypothesis, alternate hypothesis, test statistic, level of significance, p-value, testing hypothesis about population means, and population proportions, confidence intervals. Nonparametric tests - sign test, Wilcoxon-Mann-Whitney test, Wilcoxon signed rank test. Contingency tables, chi-square test for independence of attributes. (16hrs)
- **<u>Unit 2</u>**: Testing for the equality of several population means. The concept of analysis of variance, one way analysis of variance, its utility in the analysis of survey data and data obtained from designed experiments. (10hrs)
- **<u>Unit 3</u>**: Regression and correlation - bivariate data, correlation, scatter plot, correlation coefficient and its properties, testing for correlation coefficient, rank correlation. Regression - use of simple linear regression model to study the linear relationship between two variables, fitting the simple linear regression model, testing significance of regression coefficient, coefficient of determination. (14hrs)

Books for Reference:

1. Campbell, R. C. (1974). *Statistics for Biologists*. Cambridge University Press.
2. Chatfield, C. (1981). *Statistics for Technology*. Chapman and Hall.
3. Frank, H. and Athoen, S. C. (1997). *Statistics: Concepts & Applications*. Cambridge University Press.
4. Medhi, J. (1992). *Statistical Methods: An Introductory Text*. Wiley Eastern Limited.
5. Ross, S. M. (2017). *Introductory Statistics*. Academic Press.
6. Rice, J. A. (2006). *Mathematical Statistics and Data Analysis*. Singapore: Thomson-Duxbury.

STE 531B - CATEGORICAL DATA ANALYSIS
(Open Elective Course)

Rationale/Learning Objectives:

- To gain theoretical knowledge on various statistical procedures used for categorical data analysis.
- To explore different statistical approaches used for analyzing two-way and three-way tables.
- To study about statistical models for analyzing binary, polytomous and multivariate categorical responses.

Outcomes:

- Develop conceptual understanding and application of statistical procedures for analyzing categorical data.
- Able to identify designs of contingency tables and recommend appropriate measures of association and statistical tests.
- Ability to develop models for binary, polytomous and multivariate categorical responses, interpret results regardless of model parameterization and diagnose model fits.

Syllabus (3 Credits):

- **Unit 1**: Introduction and Probability distributions: What is categorical data analysis, Scales of measurement, A brief history of categorical methods, Probability distributions for categorical variables, Frequency distribution tables for discrete variables, The hypergeometric distribution, The Bernoulli distribution, The binomial distribution, The multinomial distribution, The Poisson distribution. Maximum likelihood estimation: a single proportion, Hypothesis testing for a single proportion, Confidence intervals for a single proportion, Goodness-of-fit: comparing distributions for a single discrete variable. (12hrs)
- **Unit 2**: Analyzing Contingency Tables: Probability Structure for Contingency Tables, Comparing Proportions in 2×2 Contingency Tables, The Odds Ratio, Chi-Squared Tests of Independence, Testing Independence for Ordinal Variables, Contingency

tables for three categorical variables, Marginal and conditional independence, Inferential statistics for three-way tables. (14hrs)

- **<u>Unit 3</u>**: Generalized Linear Models: Components of a Generalized Linear Model, Generalized Linear Models for Binary Data, Generalized Linear Models for Counts and Rates, Statistical Inference and Model Checking, Fitting Generalized Linear Models. (14hrs)

Books for Reference:

1. Agresti, A. *An Introduction to Categorical Data Analysis* (3rd ed.). Hoboken, NJ: John Wiley & Sons, 2019. Series: Wiley Series in Probability and Statistics, ISBN 9781119405269.
2. Azen, R. and Walker, C. M. (1969). *Categorical Data Analysis for the Behavioral and Social Sciences*. Routledge, Taylor & Francis Group, New York, ISBN 978-1-84872-836-3.

STE 531C - DEMOGRAPHIC METHODS AND ANALYSIS
(Open Elective Course)

Rationale/Learning Objectives:

- To study about the basic demographic indicators and explore various sources of demographic data.
- To gain knowledge on the theory and applications of descriptive statistics, correlation and regression techniques.
- To understand the key measures and techniques used in studying population behavior and change.

Outcomes:

- Understand the basic demographic indicators and explore the different sources of demographic data.
- Understand the theory and applications of descriptive statistics, along with correlation and regression techniques.
- Develop conceptual understanding and applications of the key measures and techniques used in studying population behavior and change.

Syllabus (3 Credits):

- **Unit 1**: Demography: introduction, purpose, nature of demographic information: births, fertility, fecundity, deaths, mortality, life expectancy, migration. Data collection methods: census, sample surveys, registration of vital events, population registers, and administrative records. (10hrs)
- **Unit 2**: Statistical measures: measures of central tendency - arithmetic mean, median, mode; normal and skewed distributions; measures of dispersion - variance and standard deviation, quantiles; correlation and linear regression. (16hrs)
- **Unit 3**: Measurement of population, measures of fertility: crude birth rate, age-specific fertility rate, general fertility rate, and total fertility rate; measures of mortality: crude death rate, age-specific death rate, standardized death rate, infant mortality rate, neo-natal mortality rate, and maternal mortality rate. Life table and its components. (14hrs)

Books for Reference:

1. Yusuf, F., Martins, J. M., and Swanson, D. A. (2014). *Methods of Demographic Analysis*. Springer, New York, London.
2. Carmichael, G. A. (2016). *Fundamentals of Demographic Analysis: Concepts, Measures and Methods*. Springer, New York, London.

STH 532 - THEORY OF TESTING OF HYPOTHESIS (4 Credits)

Rationale/Learning Objectives:

- To gain theoretical knowledge on the fundamentals involved in testing of hypothesis theory.
- To understand the theory and practical implementation of the concepts such as uniformly most powerful, uniformly most powerful unbiased, and likelihood ratio tests.
- To study about the theoretical aspects and applications of interval estimation.
- To explore the real life applications of various nonparametric tests.

Outcomes:

- Show acquisitions of adequate foundations on the fundamentals involved in testing of hypothesis and understand its importance.
- Show learning about the theoretical aspects of uniformly most powerful, uniformly most powerful unbiased, likelihood ratio tests, and its implementation in practical problems.
- Develop knowledge about the theoretical aspects of interval estimation and its implementation in practical problems.
- Exhibit knowledge on various nonparametric tests and its applications in real life problems.

Syllabus

- **Unit 1**: Framing of null hypothesis, critical region, level of a test, randomized and non-randomized tests, two kinds of error, size of a test, p-value, power function. Most powerful tests in class of size a test, Neyman-Pearson lemma, MP test for simple null against simple alternative hypothesis. UMP tests for one sided null against one sided alternatives, monotone likelihood ratio property. Extension of these results in Pitman family when only upper or lower end points depend on the parameter. (13hrs)
- **Unit 2**: Non-existence of UMP test. Neyman-Pearson generalized lemma (statement only), concept of UMP for simple null against two sided alternatives in one parameter exponential family and UMPU tests with application to one parameter exponential family, UMP for two sided null (statement only). Likelihood ratio test (LRT),

asymptotic distribution of LRT statistic, Pearson's chi-square test for goodness of fit, Bartlett's test for homogeneity of variances, large sample tests.

(16hrs)

- **<u>Unit 3</u>**: Interval estimation, confidence level, construction of confidence intervals by inverting acceptance region. Shortest expected length confidence interval, evaluating interval estimators using size and coverage probability and test related optimality, uniformly most accurate one-sided confidence interval and its relations to UMP test for one sided null against one sided alternative hypothesis.

(10hrs)

- **<u>Unit 4</u>**: U-statistics, properties and asymptotic distributions (in one and two sample case). Nonparametric tests: One sample test - test based on total number of runs, the ordinary sign test, the Wilcoxon signed rank test, the Kolmogorov-Smirnov one sample goodness of fit test. Two sample tests - the median test, the Wilcoxon-Mann-Whitney test, Kolmogorov-Smirnov two sample test. (13hrs)

Books for Reference:

1. Casella, G. and Berger, R. L. (2002). *Statistical Inference*. Wadsworth Group.
2. Gibbons, J. D. (1971). *Nonparametic Inference*. McGraw Hill.
3. Kale, B. K. (1999). *A First Course on Parametric Inference*. Narosa Publishing House.
4. Lehmann, E. L. and Romano, J. (2008). *Testing Statistical Hypotheses*. John Wiley.
5. Pratt, T. W. and Gibbons, J. D. (1981). *Concepts of Nonparametric Theory*. Springer.
6. Rao, C. R. (1973). *Linear Statistical Inference and its Applications*. Wiley Eastern.
7. Rohatgi, V. K. and Saleh, A. K. L. (2001). *An Introduction to Probability and Mathematical Statistics*. Wiley Eastern.
8. Rajagopalan, M. and Dhanavanthan, P. (2012). *Statistical Inference*. Phi Learning Pvt. Ltd.

Practicals based on Theory of Testing of Hypothesis

1. Size and power of the test
2. Most powerful test
3. Uniformly most powerful test for discrete distributions
4. Uniformly most powerful test for continuous distributions
5. Large sample test – I
6. Large sample test – II
7. Uniformly most powerful unbiased test
8. Confidence interval using pivotal method
9. Interval estimation
10. Testing for randomness using run test
11. Sign test
12. Wilcoxon signed rank test
13. Kolmogorov-Smirnov test
14. Median test

STH 533 - REGRESSION ANALYSIS (4 Credits)

Rationale/Learning Objectives:

- To explore the theory and real life applications of linear regression techniques.
- To study the importance of regression diagnostics.
- To understand the consequences of the problem of multicollinearity in linear regression.
- To learn about the problems of heteroscedasticity and autocorrelation, and understand its consequences.
- To gain theoretical knowledge on the concepts of simultaneous equation models and identification problem.

Outcomes:

- Show an acquisition of necessary theoretical foundations on linear regression techniques and its extensive use in data analysis.
- Skilled in model adequacy checking and regression diagnostics.
- Able to understand the consequences of multicollinearity and work on its remedies.
- Able to test and understand the consequences of heteroscedasticity and autocorrelation, and work on its remedies.
- Familiarity with the theoretical aspects of simultaneous equation models and identification problem.

Syllabus

- **Unit 1**: Simple linear regression, multiple linear regression, basic assumptions, ordinary least squares (OLS) - estimation and their properties, tests of hypothesis about regression coefficients, likelihood ratio criterion. Dummy variables. Prediction - best linear unbiased predictor. (10hrs)
- **Unit 2**: Regression diagnostics and specification tests - residual analysis for identifying influential observations, recursive residuals and their applications, specification tests, subset selection of explanatory variables, Mallows C_p statistic. Use of prior information. Restricted least squares estimators and mixed regression estimator. (10hrs)
- **Unit 3**: Violation of basic ideal conditions - disturbance with non-zero mean, asymptotically uncooperative regressors. Multicollinearity - its consequences and testing. Ridge estimator and its properties, ridge regression. Stochastic regressors,

autoregressive models, instrumental variables, errors in variables. Distributed lag models. (10hrs)

- **<u>Unit 4</u>**: Heteroscedasticity - tests for heteroscedasticity. Generalized least squares (GLS) estimators and its properties, feasible generalized least squares estimators. Grouping of observations. Sets of Regression Equations. Auto correlation - its consequences and testing for autocorrelation, estimation and prediction. Autoregressive conditional heteroscedasity (ARCH) models. (10hrs)
- **<u>Unit 5</u>**: Simultaneous equation models. Identification problem, identification using linear homogeneous restrictions on structural parameters, rank and order conditions, estimation in simultaneous equation models. Indirect least squares, two stage least squares, structural equation modelling. (12hrs)

Books for Reference:

1. Cook, R. D. and Weisberg, S. (1982). *Residual and Influence in Regression*. London: Chapman and Hall.
2. Draper, N. R. and Smith, H. (1998). *Applied Regression Analysis* (3rd ed.). New York: Wiley.
3. Gunst, R. F. and Mason, R. L. (1980). *Regression Analysis and its Application - A Data Oriented Approach*. Marcel Dekker.
4. Montgomery, D. C., Peck, E. A. and Vining, G. G. (2003). *Introduction to Linear Regression Analysis*. John Wiley.
5. Ryan, T. P. (1997). *Modern Regression Methods*. New York: John Wiley.
6. Seber, G. A. F. and Lee, A. J. (2003). *Linear Regression Analysis* (2nd ed.). New York: John Wiley.
7. Fomby, T. B., Hill, C. R. and Johnson, S. R. (1988). *Advanced Econometric Methods*. Springer.
8. Greene, W. H. (2002). *Econometric Analysis* (5th ed.). New York: Prentice Hall.
9. Johnston, J. and Dinardo, J. (1996). *Econometric Methods* (4th ed.). McGraw-Hill.
10. Maddala, G. S. (1992). *Introduction to Econometrics* (2nd ed.). New York: Macmillan.
11. Gujarati, D. N. (2004). Basic Econometrics (4th ed.). McGraw-Hill.

Practicals based on Regression Analysis

1. Simple linear regression analysis
2. Multiple linear regression analysis
3. Regression diagnostics
4. Backward elimination and forward selection procedures
5. Multicollinearity diagnostics
6. Heteroscedasticity
7. Autocorrelation

STH 534 - MULTIVARIATE ANALYSIS (4 Credits)

Rationale/Learning Objectives:

- To gain knowledge on the theoretical foundations required for analyzing multivariate data.
- To explore different inferential techniques used for analyzing multivariate data.
- To understand the theory and applications of multivariate techniques such as principal component analysis, canonical correlation analysis, and factor analysis.
- To learn the theoretical aspects of discrimination and classification, and cluster analysis, and understand its applications in real life.

Outcomes:

- Show an acquisition of necessary theoretical foundations for analyzing multivariate data.
- Exhibit theoretical knowledge on various inferential techniques used for analyzing multivariate data.
- Develop theoretical knowledge about multivariate techniques such as principal component analysis, canonical correlation analysis, factor analysis, and its applications in real life problems.
- Gain conceptual understanding about multivariate techniques such as discrimination and classification, cluster analysis, and its applications in real life problems.

Syllabus

- **Unit 1**: Nature of a multivariate problem, main types of multivariate problems, objectives of multivariate analysis. Organization of multivariate data, descriptive statistics, visualization techniques. Multivariate normal distribution properties, maximum likelihood estimators of the parameters. Independence of sample mean vector and sample covariance matrix. Assessing the assumptions of normality Q-Q plot, chi-square plot, transformations to near normality. (13hrs)
- **Unit 2**: Inference problems in multivariate normal distribution, Hotellings T^2, Mahalanobis D^2 statistics, likelihood ratio tests for collinearity, q-sample problem. Roy's union and intersection test. Test for symmetry. Confidence regions, simultaneous confidence statements. Independence of subvectors, sphericity test. Wishart matrix, statement of Wishart distribution, its properties and applications. (13hrs)

- **Unit 3**: Principal component analysis (PCA) - definition and properties, graphing the principal components, sample principal components, interpretation of zero, small and repeated eigenvalues, component loadings and component correlations, the problem of scaling, tests of hypotheses. Canonical correlation analysis - canonical variates and canonical correlations, sample canonical variates, sample canonical correlations, inference problems. Factor analysis - orthogonal factor model, factor loadings, estimation of factor loadings, factor scores. (13hrs)
- **Unit 4**: Classification and discrimination problems - concepts of separation and classification, Bayes and Fisher's criteria, classification rules based on expected cost of misclassification (ECM) and total probability of misclassification (TPM), classification with two multivariate normal populations (equal and unequal covariance matrices), evaluating classification rules, classification with several populations, Fisher's linear discriminant function, tests associated with discriminant functions. Cluster Analysis - distances and similarity measures, hierarchical clustering methods, k-means method. (13hrs)

Books for Reference:

1. Anderson, T. W. (1984). *An Introduction to Multivariate Analysis* (2nd ed.). John Wiley.
2. Flury, B. (1997). *A First Course in Multivariate Statistics*. Springer Texts in Statistics.
3. Kshirasagar, A. M. (1972). *Multivariate Analysis*. Marcel Dekker.
4. Mardia, K. V., Kent, J. T. and Bibby, J. M. (1979). *Multivatiate Analysis*. Academic Press.
5. Rao, C. R. (1973). *Linear Statistical Inference and its Applications*. Wiley Eastern.
6. Johnson, R. A. and Wichern, D. W. (1986). *Applied Multivariate Statistical Analysis* (6th ed.). Prentice Hall of India.
7. Rencher, A. C. (2003). *Methods of Multivariate Analysis*. Wiley.

Practicals based on Multivariate Analysis

1. Multidimensional data visualization
2. Multivariate normal inference – I
3. Multivariate normal inference – II
4. Testing the independence of subvectors
5. Two-sample problem
6. Sphericity test
7. Principal component analysis – I
8. Principal component analysis – II
9. Canonical correlation analysis
10. Discrimination and classification

STS 535 - STOCHASTIC PROCESSES (3 Credits)

Rationale/Learning Objectives:

- To gain knowledge on the necessary theoretical foundations of stochastic process.
- To understand the theory and applications of Poisson process.
- To study about the theoretical aspects of renewal theorem.
- To explore the real life applications of branching process and understand its importance.

Outcomes:

- Acquire necessary theoretical knowledge on the foundations of stochastic processes.
- Explore the theoretical aspects associated with Poisson process and its real life applications.
- Demonstrate conceptual understanding and applications of renewal theory.
- Gain knowledge about the theoretical aspects of branching process and its applications.

Syllabus

- **Unit 1**: Introduction to stochastic processes - classification according to state space and time domain. Stationary process - weakly stationary and strongly stationary processes. Countable state Markov chains (MCs), Chapman Kolmogorov equations, calculation of n-step transition probability and its limit. Stationary distribution, classification of states, random walk and gamblers ruin problem, estimation of TPM for finite states of MC. (10hrs)
- **Unit 2**: Discrete state space continuous time MC, Kolmogorov-Feller differential equations, Poisson process, birth and death process, applications to queues. Wiener process as a limit of random walk, first passage time and other problems. (10hrs)
- **Unit 3**: Renewal theory - elementary renewal theorem and applications. Statement and uses of key renewal theorem, study of residual life time process. (10hrs)
- **Unit 4**: Branching process - Galton-Watson branching process, probability of ultimate extinction, distribution of population size. Martingale in discrete time - definition and elementary properties, convergence theorem, applications. (10hrs)

Books for Reference:

1. Basu, A. K. (2003). *Introduction to Stochastic Processes*. Narosa Publications.
2. Bhat, B. R. (2000). *Stochastic Models: Analysis and Applications*. New Age International.
3. Karlin, S. and Taylor, H. M. (1975). *A First Course in Stochastic Processes*. Academic Press.
4. Medhi, J. (1982). *Stochastic Processes*. Wiley Eastern.
5. Ross, S. M. (1983). *Stochastic Processes*. John Wiley & Sons.
6. Lawler, G. F. (2006). *Introduction to Stochastic Processes* (2nd ed.). Chapman and Hall.

Practicals based on Stochastic Processes

1. Realization of stochastic processes.
2. Calculation of n-step transition probabilities.
3. Classification of states and mean recurrence time of state.
4. Simulation of Markov chain and estimating the stationary distribution of ergodic Markov chain.
5. Simulation of Poisson processes.
6. Realization of queues and computations of typical events limiting.
7. Simulation of branching process and estimating its mean and variance.

Practicals based on Machine Learning with Python

1. Compare the performance of the OLS, ridge and the Lasso models on the movie collection test data. Establish which of these methods is giving the lowest test error.
2. Hyperparameter optimization using random search and grid search.
3. Implementing different types of encoding for categorical data.
4. Implement logistic regression model with scikit-learn.
5. Implementation of logistic regression in python - step by step.
6. Implementation of KNN classifier using scikit-learn.
7. Multiple linear regression from scratch using python.
8. Implement decision tree and random forest models, compare the performance in python using scikit-learn.
9. Build PCA from scratch in python and also using scikit-learn and fit the model for the transformed data.
10. Perform factor analysis using python.
11. Perform linear discriminant analysis (LDA) in python.
12. Perform cluster analysis using python.

IV Semester

STH 541 - TIME SERIES ANALYSIS (4 Credits)

Rationale/Learning Objectives:

- To study the fundamentals involved in the theory of time series analysis.
- To gain knowledge on the theory and applications of MA, AR, ARMA and ARIMA models.
- To learn about spectral density function and the estimation process involved in MA, AR and ARMA models.
- To explore the practical implementation of Box-Jenkins approach, exponential and Holt Winters smoothing techniques to model and forecast time series data.

Outcomes:

- Develop necessary theoretical foundations on the fundamental concepts involved in time series analysis.
- Gain theoretical knowledge on various time series models such as MA, AR, ARMA and ARIMA, and its real life applications.
- Understand the concept of spectral density function and also regarding the estimation process involved in MA, AR and ARMA models.
- Implement Box-Jenkins approach, exponential and Holt Winters smoothing techniques to model and forecast time series data empirically.

Syllabus

- **Unit 1**: Simple descriptive techniques - time series plots, trend, seasonal effect. Tests for trend and seasonality - estimation and elimination of trend and seasonal components. Exponential and moving average smoothing. Time series as discrete parameter stochastic process. Stationarity, autocovariance and autocorrelation function and their properties. Partial autocorrelation function. (13hrs)
- **Unit 2**: Probability models - White noise model, random walk, linear processes, moving average (MA), autoregressive (AR), ARMA and ARIMA models, invertibility, ACF and PACF of these processes. Spectral properties of stationary models - periodogram, spectrum. (13hrs)
- **Unit 3**: Spectral density function - estimation of spectral densities of AR, MA and ARMA models. Sample ACF and PACF for model identification. Model building -

estimation of mean, autocovariance function and autocorrelation function. Estimation in AR models, Yule-Walker equations, estimation in MA model and ARMA models. Order selection in AR and MA models. (13hrs)

- **Unit 4**: Forecasting - forecast mean square error (FMSE), least squares prediction, BLUP, innovation algorithm. Box-Jenkins forecasting for ARMA Models. Forecasting through exponential smoothing and Holt Winters smoothing. Residual analysis and diagnostic checking. Non-stationary time series models and their identification. Introduction to ARCH and GARCH models. (13hrs)

Books for Reference:

1. Box, G. E. P. and Jenkins, G. M. (1976). *Time Series Analysis: Forecasting and Control*. San Francisco: Holden Day.
2. Brockwell, P. J. and Davis, R. S. (2002). *Introduction to Time series and Forecasting* (2nd ed.). Springer.
3. Chatfield, C. (1996). *The Analysis of Time Series: An Introduction*. Chapman Hall.
4. Kendall, M. G. and Ord, J. K. (1990). *Time Series* (3rd ed.). Edward Arnold.
5. Montgomory, D. C. and Johnson, D. A. (1977). *Forecasting and Time Series Analysis*. McGraw Hill.
6. Tanaka, K. (1996). *Time Series Analysis*. Wiley Series.
7. Tsay, R. S. (2005). *Analysis of Time Series*. John Wiley & Sons.

Practicals based on Time Series Analysis

1. Time series plots and elimination of trend and seasonality.
2. Estimation of ACF and PACF.
3. Model identification and estimation of ARMA model.
4. Model identification and estimation of ARIMA model.

STH 542 - RELIABILITY AND SURVIVAL ANALYSIS (4 Credits)

Rationale/Learning Objectives:

- To understand the theoretical aspects involved in the concept of coherent structure.
- To learn the fundamental concepts of reliability theory and understand its real life applications.
- To learn the importance and applications of different censoring situations in survival analysis.
- To explore different lifetime models and understand its applications in medical studies.

Outcomes:

- Understand the concept of coherent structure, its representation and importance in industrial applications.
- Show an acquisition of necessary theoretical foundations on the fundamentals of reliability theory and its applications in real life problems.
- Understand different censoring situations and its implementation in the theoretical concepts of survival analysis.
- Show an increased learning about statistical lifetime models used in medical sciences and also regarding the estimation of survival function.

Syllabus

- **Unit 1**: Coherent structures, representation of coherent systems in terms of paths and cuts, duals systems, modules of coherent systems. Reliability of system of independent of components, association of random variables, bounds on system reliability, improved bounds on system reliability using modular decompositions, lifetime distribution of k out of n system. (13hrs)
- **Unit 2**: Measures of reliability, survival/failure rate, hazard function, cumulative hazard function, lack of memory property, graphs of the system reliability functions. Notion of aging, life distributions of coherent systems, classes of life distributions - parametric and nonparametric models, mean residual lifetime with survival function. NBU, NBUE, NWU, NWUE classes of life distributions and their implications. (13hrs)
- **Unit 3**: Complete and censored samples, type I, II and random censoring, life distributions - Exponential, Gamma, Weibull, Lognormal, Pareto family. Estimation of

parameter for exponential and gamma distribution under various censoring situations. Confidence interval for parameters of Exponential, Weibull, and Lognormal distributions. Wald, Score and LR tests for Exponential against Gamma and Weibull.

(13hrs)

- **Unit 4**: Estimation of survival function - Kaplan-Meier estimator, Nelson-Aalen estimator, Greenwoods formula. Other life table estimators. Actuarial method of estimation of survival function. Semi-parametric regression for failure rate, Cox's proportional hazards model with one and more number of covariates, log likelihood function, log linear hazards, test for regression coefficients with and without ties.

(13hrs)

Books for Reference:

1. Barlow, R. E. and Proschan, F. (1975). *Statistical Theory of Reliability and Life Testing: Probability Models*. Holt, Rinehart and Winston Inc.
2. Barlow, R. E. and Proschan, F. (1996). *Mathematical Theory of Reliability*. John Wiley.
3. Tobias, P. A. and Trindane, D. C. (1995). *Applied Reliability* (2nd ed.). CRC Press.
4. Lawless, J. R. (1982). *Statistical Models and Methods for Lifetime Data*.
5. Bain, L. J. and Engelhardt (1991). *Statistical Analysis of Reliability and Life Testing Data*.
6. Zacks, S. (1992). *Introduction to Reliability Analysis: Probability Models and Statistical Methods*. Springer.
7. Cox, D. R. Oakes, D. (1984). *Analysis of Survival Data*. New York: Chapman and Hall.
8. Kalbfeish, J. D. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data* (2nd ed.). John Wiley & Sons, Inc.
9. Deshpande, J. V. and Purohit, S. G. (2005). *Lifetime Data: Statistical Models and Methods*. World Scientific.

STS 543 - STATISTICAL MODELLING (3 Credits)

Rationale/Learning Objectives:

- To study the fundamentals of Bayesian theory and understand its real life applications.
- To gain theoretical knowledge on statistical concepts such as nonparametric density estimation, nonparametric regression and resampling techniques.
- To explore the real life applications of advanced regression techniques such as logistic, multi-logit, count data, and log linear regression.

Outcomes:

- Explain the Bayesian framework for data analysis and demonstrate when the Bayesian approach can be beneficial.
- Understand the importance of some advanced statistical concepts such as nonparametric density estimation, nonparametric regression and resampling techniques.
- Exhibit theoretical knowledge on some advanced regression techniques such as logistic, multi-logit, count data, and log linear regression and understand its applications.

Syllabus

- **Unit 1**: Introduction to Bayesian theory and philosophy - loss function and risk, foundations of optimal decision making, Bayes rule, minimax rule, admissibility. Prior and posterior distributions, conjugate families, non-informative priors - uniform and Jeffrey's prior. Bayesian estimation. Introduction to credible sets, Bayesian hypothesis testing, Bayesian prediction. (15hrs)
- **Unit 2**: Nonparametric density estimation (kernel based), nonparametric regression techniques - kernel, nearest neighbour, local polynomial (LOESS regression) and spline based methods. Concept of resampling techniques - bootstrap and jackknife methods. Bootstrap procedure - hypothesis testing and bootstrap interval estimation. (12hrs)
- **Unit 3**}: Introduction to generalized linear model (GLM), logistic regression, multi-logit regression, count data regression, log linear regression. (13hrs)

Books for Reference:

1. Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis* (2nd ed.). New York: Springer-Verlag.
2. Ghosh, J. K., Delampady, M. and Samanta, T. (2006). *An Introduction to Bayesian Analysis: Theory and Methods*. New York: Springer Texts in Statistics.
3. Box, G. E. P. and Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis*. Massachusetts: Addison-Wesley, Reading.
4. Hardle, W. (1990). *Applied Nonparametric Regression*. Cambridge University Press.
5. Hardle, W. (1991). *Smoothing Techniques*. Springer Science & Business Media.
6. Wasserman, L. (2004). *All of Statistics: A Concise Course in Statistical Inference*. Springer Science & Business Media.
7. Wasserman, L. (2005). *All of Nonparametric Statistics*. Springer Science & Business Media.
8. Dobson, A. J. (1983). *Introduction to Statistical Modelling*. Chapman and Hall.
9. Agresti, A. (1990). *Categorical Data Analysis* (3rd ed.). Wiley.
10. Myers, R. H., Montgomery, D. C., Vining, G. G. and Robinson, T. J. (2010). *Generalized Linear Models: with Applications in Engineering and the Sciences* (2nd ed.). John Wiley & Sons.
11. Davison, A. C. and Hinkley, D. V. (1991). *Bootstrap Methods and their Application*. Cambridge University Press.
12. Higgins, J. J. (2004). *An Introduction to Modern Nonparametric Statistics*. Brooks/Cole.

STS 544 - BIG DATA ANALYTICS (3 Credits)

Rationale/Learning Objectives:

- To gain theoretical knowledge on the concepts involved in big data.
- To understand how Hadoop ecosystem tools can be used in solving big data problems.
- To study about the concepts and techniques involved in business intelligence.
- To explore various data mining techniques such as neural networks, association rule mining, text mining, web mining and social network analysis.

Outcomes:

- Ability to understand the theoretical aspects involved in big data.
- Implement Hadoop ecosystem tools in solving big data problems.
- Explore the concepts and techniques involved in business intelligence used for decision making purpose.
- Show an understanding of theoretical knowledge on various data mining techniques such as neural networks, association rule mining, text mining, web mining and social network analysis.

Syllabus

- **Unit 1**: Introduction to Big Data - Classification of Digital Data, Characteristics of Data, Evolution of Big Data, Challenges with Big Data, Business Intelligence Vs Big Data, A Typical Data Warehouse Environment. Big Data Analytics - Introduction to Big Data Analytics, Classification of Analytics, Importance of Big Data Analytics, Data Science, Terminology used in Big Data Environment. (10hrs)
- **Unit 2**: Hadoop - Hadoop Distributed File System Basics, Hadoop MapReduce Framework, MapReduce Programming. Hadoop Essential Tools - Apache HIVE, Apache PIG, Sqoop, Apache Flume. (10hrs)
- **Unit 3**: Business Intelligence Concepts and Application, Data Warehousing, Data Mining, Data Visualization. (10hrs)
- **Unit 4**: Artificial Neural Network, Association Rule Mining, Text Mining, Web Mining, Social Network Analysis. (10hrs)

Books for Reference:

1. Acharya, S. and Chellappan, S. (2015). *Big Data and Analytics*. Wiley Publications.
2. Eadline, D. (2016). *Hadoop 2 Quick-Start Guide: Learn the Essentials of Big Data Computing in the Apache Hadoop 2 Ecosystem* (1st ed.). Pearson Education. ISBN-13: 978-9332570351.
3. Maheshwari, A. (2017). *Data Analytics* (1st ed.). McGraw Hill Education. ISBN-13: 978-9352604180.
4. White, T. (2015). *Hadoop: The Definitive Guide*. (4th ed.). O'Reilly Media. ISBN-13: 978-9352130672.
5. Lam, C. *Hadoop in Action*. Manning. ISBN 9781935182191.

ARTIFICIAL INTELLIGENCE (3 Credits)

Rationale/Learning Objectives:

This course enables the students to identify the problems where artificial intelligence is required and the different methods available.

- **<u>Unit 1</u>**: What is artificial intelligence? Problems, Problem Spaces and search, Heuristic search technique.

  (10hrs)

- **<u>Unit 2</u>**: Knowledge Representation Issues, Using Predicate Logic, representing knowledge using Rules, Symbolic Reasoning under Uncertainty.

  (10hrs)

- **<u>Unit 3</u>**: Statistical reasoning, Weak Slot and Filter Structures, Strong slot-and-filler structures.

  (10hrs)

- **<u>Unit 4</u>**: Game Playing, Natural Language Processing, Learning.

  (10hrs)

Books for Reference:

1. Rich, E., Knight, K. and Nair, S. B. *Artificial Intelligence* (3rd ed.). McGraw Hill.
2. Rusell, S. and Norving, P. *Artificial Intelligence: A Modern Approach* (2nd ed.). Pearson Education.
3. Patterson, D. W. *Introduction to Artificial Intelligence and Expert Systems*. Prentice Hall of India.
4. Luger, G. (2002). *Artificial Intelligence: Structures and Strategies for Complex Problem Solving* (4th ed.). Pearson Education.
5. Rolston, D. W. *Artificial Intelligence and Expert Systems Development*. McGraw Hill.
6. Padhy, N. P. (2015). *Artificial Intelligence and Intelligent Systems*. Oxford University Press.

ELEMENTS OF STATISTICAL COMPUTING (3 Credits)

Rationale/Learning Objectives:

This course provides foundations for statistical simulation and validation of models.

- **Unit 1**: Random number generation, requisites of a good random number generator, methods of random number generation such as linear congruential, mixed congruential and multiplicative congruential. Testing of random number generator, run test, Kolmogrov-Smirnov test, sign test, rank test, gap test, digit frequency test and serial correlation, selection of a random number generator. Methods of generating random observations such as inverse transforms, composition, convolution and acceptance-rejection. (10hrs)
- **Unit 2**: Simple optimization method, direct search, grid search, interpolatory search, gradient search. Newton-Raphson method, Muller's method, Aitken's extrapolation, simple problems and applications. (10hrs)
- **Unit 3**: Methods to compute integrals - quadrature formula, double integration, singularity, Gaussian integration. Monte Carlo Methods - Monte Carlo integration and simple case studies, applications of Monte Carlo methods to compute expected values of functions of random variables such as Laplace transform, fourier transform etc., some case studies. (10hrs)
- **Unit 4**: Approximating probabilities and percentage points in selected probability distribution, verification of WLLN and CLT using random number generator, simulating null distribution of various test statistics, simple applications and case studies. (10hrs)

Books for Reference:

1. Kennedy, W. J. Gentle, J. E. (1980). *Statistical Computing*. Marcel Dekker.
2. Sen, K. V. (1993). *Numerical Algorithm Computation in Science and Engineering* (2nd ed.). Affiliated East West Press.
3. Law, A. M. and Kelton, W. D. (2000). *Simulation, Modeling and Analysis* (3rd ed.). Tata McGraw Hill.
4. Rajaraman, V. (1993). *Computer Oriented Numerical Methods* (4th ed.). Prentice Hall.
5. Ripley, B. D. (1987). *Stochastic Simulation*. John Wiley.
6. Ross, S. M. (2000). *Introduction to Probability Models*. Academic Press.
7. Ross, S. M. (2013). *Simulation*. Academic Press.
8. Thisted, R. A.(1988). *Elements of Statistical Computing*. Chapman and Hall.

Practicals based on Elements of Statistical Computing

1. Generation of random numbers by acceptance rejection method.
2. Generation of random numbers by linear congruential method.
3. Solution of the equations using iterative methods.
4. Numerical integration.
5. Monte-Carlo integration.
6. Empirical distributions of the test statistics.
7. Applications of CLT.

SURVIVAL ANALYSIS (3 Credits)

Rationale/Learning Objectives:

This course provides theoretical foundations on statistical lifetime models being used in medical sciences and industries.

- **Unit 1**: Complete and censored samples, type I, II and random censoring, life distributions - Exponential, Gamma, Weibull, Lognormal, Pareto family. Estimation of parameter for exponential and gamma distribution under various censoring situations. Confidence interval for parameters of Exponential, Weibull, and Lognormal distributions. Wald, Score and LR tests for Exponential against Gamma and Weibull. (10hrs)
- **Unit 2**: Life tables - standard methods for uncensored and censored data, asymptotic properties of estimates under a random censorship model. Failure rate, mean residual life and their elementary properties. Estimation of survival function - Kaplan-Meier estimator, Greenwoods formula. Other life table estimators. Actuarial method of estimation of survival function. (10hrs)
- **Unit 3**: Fully parametric analysis of dependency accelerated life model - simple form, log logistic accelerated life model, proportional hazards model in relation with accelerated life model. Semi-parametric regression for failure rate, Cox's proportional hazards model with one and more number of covariates, log likelihood function, log linear hazards, test for regression coefficients with and without ties. (10hrs)
- **Unit 4**: Two sample problem - Gehan test, log rank test, Mantel-Haenszel test. Competing risks model - parametric and nonparametric inference for these model. (10hrs)

Books for Reference:

1. Cox, D. R. Oakes, D. (1984). *Analysis of Survival Data*. New York: Chapman and Hall.
2. Kalbfeish, J. D. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data* (2nd ed.). John Wiley \& Sons, Inc.
3. Lawless, J. F. (2002). *Statistical Models and Methods for Lifetime Data*. John Wiley & Sons, Inc.
4. Miller, R. G. (1981). *Survival Analysis*. John Wiley & Sons, Inc.
5. Hosmer, D. W., Lemeshow, S. and May, S. (2008). *Applied Survival Analysis: Regression Modeling of Time-to-Event Data* (2nd ed.). John Wiley & Sons, Inc.

6. Deshpande, J. V. and Purohit, S. G. (2005). *Lifetime Data: Statistical Models and Methods*. World Scientific.

Practicals based on Survival Analysis

1. Plotting empirical distribution and empirical survival functions
2. Confidence interval for the parameter of exponential distribution under random censoring
3. Confidence interval for the parameters of Weibull distribution
4. Nonparametric estimation of survival function using Kaplan-Meier estimator
5. Identification of distribution using graphical procedure
6. Construction of life table from medical follow-up study
7. Accelerated failure time model
8. Cox proportional hazards model
9. Testing for the equality of survival functions

STOCHASTIC FINANCE (3 Credits)

Rationale/Learning Objectives:

This course provides foundations on fundamentals of financial markets and stocks and to analyze the data on finance.

- **Unit 1**: Basic concepts of financial markets and stocks, types of traders, forward contracts and futures, call and put options, European option and American options. Interest rates, continuous compounding, present value analysis, bond pricing, risk free interest rates. Returns, gross returns, log returns. (10hrs)
- **Unit 2**: Portfolio theory, mean variance portfolio theory. One risky asset and one risk free asset, two risky assets. Sharpes ratio, tangency portfolio, optional mix of portfolio. Market portfolio, beta, security market line, and capital asset pricing model (CAPM) and their assumption. Value at risk (VAR), nonparametric and parametric estimation of VAR, VAR for a derivative and for a portfolio of assets, delta normal method, simulation of VAR models. (10hrs)
- **Unit 3**: Financial derivatives, options, pricing via arbitrage, law of one price. Risk neutral valuation, arbitrage theorem. Convexity of cost of call option, binomial model single and multiperiod binomial model. Modeling returns - lognormal model, random walk model, modeling through geometric Brownian motion process. Ito lemma (without proof). Arbitrage theorem. The Black Scholes formula and assumptions, properties of the Black Scholes option cost. (10hrs)
- **Unit 4**: Black Scholes Merton differential equations and assumptions, the delta hedging arbitrage strategy, volatility and estimating the volatility parameter, implied volatility. Pricing American options, pricing of an European option using Monte Carlo and pricing an American option using finite difference methods. Call options on dividend paying securities. (10hrs)

Books for Reference:

1. Ross, S. M. (2003). *An Elementary Introduction to Mathematical Finance*. Cambridge University Press.
2. Ruppert, D. (2004). *Statistics and Finance: An Introduction*. Springer International Edition.
3. Hull, J. C. (2008). *Options, Futures and Other Derivatives*. India: Pearson Education.
4. Cuthbertson, K. and Nitzsche, D. (2001). *Financial Engineering: Derivatives and Risk Management*. John Wiley & Sons Ltd.
5. Leuenberger, D. G. (1998). *Investment Science*. Oxford University Press.
6. Wilmott, P. (2000). *Quantitative Finance*. John Wiley & Sons.
7. Tsay, R. S. (2005). *Analysis of Time Series*. John Wiley & Sons.

DATA MINING (3 Credits)

Rationale/Learning Objectives:

This course provides foundations on various statistical methods used in data analysis, including artificial intelligence and machine learning techniques.

- **Unit 1**: Data mining - motivations and importance, knowledge discovery in databases (KDD) process search - introduction, querying, approximation and compression. Kinds of data considered for data mining, basic data mining tasks, data mining issues. Data mining models - predictive and descriptive, inter connections between statistics, data mining, artificial intelligence and machine learning, applications of data mining. (10hrs)
- **Unit 2**: Data marts, databases and data warehouses, OLTP systems, multidimensional models - data cubes, OLAP operations on data cubes, multidimensional schemes. Data Pre-processing - data cleaning, data integration, data transformation and data reduction. Visualization techniques for multidimensional data - scatter plot matrix, star plots, Andrews plots, Chernoff faces, parallel axis plots. (10hrs)
- **Unit 3**: Supervised learning - classification and prediction, statistical classification, linear discriminants, Mahalanobis linear discriminant, Fisher's linear discriminant, Bayesian classifier, regression based classification, k-NN (nearest neighbour) classifier. Tree classifiers - decision trees, ID3 algorithm, CART. Artificial neural networks (ANN) - the learning problem, perceptron, the delta rule, multilayer feed forward neural network, back propagation learning algorithm. Support vector machines - Lagrangian formulation and solution, measuring classifier accuracy. (10hrs)
- **Unit 4**: Unsupervised learning - clustering problem, similarity and distance measures, partitioning algorithms - k-means, k-medoids (PAM) algorithms. Density based clustering algorithms (DBSCAN). Association rule mining - market basket analysis, frequent item sets, support and confidence of an association rule, a priori algorithm, partition algorithm. (10hrs)

Books for Reference:

1. Han, J. and Kamber, M. (2002). *Data Mining: Concepts and Techniques*. USA: Morgan Kaufman Publishers.
2. Dunham, M. H. (2005). *Data Mining: Introductory and Advanced Topics*. Pearson Education.
3. Hastie, T., Tibshirani, R. and Friedman, J. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer.
4. Berthold, M. R. and Hand, D. J. (2003). *Intelligent Data Analysis: An Introduction* (2nd ed.). Springer.
5. J. P. Marques de Sa (2001). *Pattern Recognition: Concepts, Methods and Applications*. Springer.
6. Chattamvelli, R. (2009). *Data Mining Methods*. Narosa Publishing House.

BAYESIAN INFERENCE (3 Credits)

Rationale/Learning Objectives:

This course provides theoretical knowledge on Bayesian techniques for data analysis and inference.

- **Unit 1**: Introduction and philosophy - loss function and risk, foundations of optimal decision making, Bayes rule, minimax rule, admissibility, sufficiency and Rao-Blackwellization. (10hrs)
- **Unit 2**: Utility theory, utility and loss, personal utility function, prior and posterior, conjugate families, non-informative priors - uniform prior, Jeffrey's prior, left and right invariant prior. (10hrs)
- **Unit 3**: Bayesian analysis - the posterior distribution, Bayesian estimation, credible sets, Bayesian hypothesis testing, Bayesian prediction, empirical Bayes analysis, hierarchical Bayes analysis, Bayesian robustness. (10hrs)
- **Unit 4**: Bayesian computation - analytic approximation, the EM algorithm, Monte Carlo sampling, Markov Chain Monte Carlo methods. (10hrs)

Books for Reference:

1. Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis* (2nd ed.). New York: Springer-Verlag.
2. Ghosh, J. K., Delampady, M. and Samanta, T. (2006). *An Introduction to Bayesian Analysis: Theory and Methods*. New York: Springer Texts in Statistics.
3. Box, G. E. P. and Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis*. Massachusetts: Addison-Wesley, Reading.

Practicals based on Bayesian Inference

1. Computation of risk functions.
2. Computation of Bayes and minimax rules.
3. EM algorithm.

STATISTICAL METHODS FOR RELIABILITY (3 Credits)

Rationale/Learning Objectives:

This course provides theoretical knowledge on reliability techniques for data analysis and inference.

- **Unit 1**: Coherent structures, representation of coherent systems in terms of paths and cuts, duals systems, modules of coherent systems. Reliability of system of independent of components, association of random variables, bounds on system reliability, improved bounds on system reliability using modular decompositions. (10hrs)
- **Unit 2**: Measures of reliability, graphs of the system reliability functions. Notion of aging, life distributions of coherent systems, distributions with increasing failure rate average arising from shock models, preservation of life distribution classes under reliability operations. Reliability bounds, lifetime distribution of k out of n system. (10hrs)
- **Unit 3**: Classes of life distributions - parametric and nonparametric models, mean residual lifetime with survival function. Applicable in replacement models, NBU, NBUE, NWU, NWUE classes of life distributions and their implications. Shock models leading to NBU. Age replacement and block replacement policies. Renewal theory useful in replacement models. (10hrs)
- **Unit 4**: Replacement policy comparisons, preservation of life distribution classes under reliability operations. Reversed hazard rate, cumulative reversed hazard function, relation between hazard function and reversed hazard function. Lack of memory property. (10hrs)

Books for Reference:

1. Barlow, R. E. and Proschan, F. (1975). *Statistical Theory of Reliability and Life Testing: Probability Models*. Holt, Rinehart and Winston Inc.
2. Barlow, R. E. and Proschan, F. (1996). *Mathematical Theory of Reliability*. John Wiley.
3. Tobias, P. A. and Trindane, D. C. (1995). *Applied Reliability* (2nd ed.). CRC Press.
4. Lawless, J. R. (1982). *Statistical Models and Methods for Lifetime Data*.
5. Bain, L. J. and Engelhardt (1991). *Statistical Analysis of Reliability and Life Testing Data*.
6. Zacks, S. (1992). *Introduction to Reliability Analysis: Probability Models and Statistical Methods*. Springer.

NONPARAMETRIC INFERENCE (3 Credits)

Rationale/Learning Objectives:

This course provides theoretical knowledge on various nonparametric testing procedures for data analysis.

- **Unit 1**: Empirical distribution function, Glivenko-Cantelli theorem, Kolmogorov goodness of fit test. One sample U-statistics, kernel and symmetric kernel, two sample U-statistics, asymptotic distribution of U-statistics. UMVUE property of U-statistics, asymptotic distribution of linear function of order statistics. (10hrs)
- **Unit 2**: Rank tests, locally most powerful rank test, linear rank statistics and their distributional properties under null hypothesis, Pitman's asymptotic relative efficiency. (10hrs)
- **Unit 3**: One sample location problem, sign test and signed rank test, two sample Kolmogorov-Smirnov tests. Two sample location and scale problems. Wilcoxon-Mann-Whitney test, normal score test, ARE of various test based linear rank statistics. Kruskal-Wallis k sample test. (10hrs)
- **Unit 4**: Cox's proportional hazards model, rank test (partial likelihood) for regression coefficients, concepts of Jackknifing method of quenouille for reducing bias, bootstrap methods, confidence intervals. (10hrs)

Books for Reference:

1. Cox, D. R. and Oakes, D. (1983). *Survival Analysis*. Chapman and Hall.
2. Davison, A. C. and Hinkley, D. V. (1991). *Bootstrap Methods and their Application*. Cambridge University Press.
3. Fraser, D. A. S. (1957). *Nonparametric Methods in Statistics*. John Wiley.
4. Gibbons, J. D. (1985). *Nonparametric Statistical Inference* (2nd ed.). Marcel Dekker.
5. Hajek, J. and Sidak, Z. (1961). *Theory of Rank Tests*. Academic Press.
6. Puri, M. L. and Sen, P. K. (1971). *Nonparametric Methods in Multivariate Analysis*. Wiley.
7. Randles, R. H. and Wolfe, D. A. (1979). *Introduction to the Theory of Nonparametric Statistics*. Wiley.

ACTUARIAL METHODS (3 Credits)

Rationale/Learning Objectives:

This course provides theoretical knowledge on various models used in insurance, risk analysis and theory of credibility.

- **Unit 1**: Review of decision theory and actuarial applications. Loss distributions - modeling of individual and aggregate losses, moments, fitting distributions to claims data, deductibles and retention limits, proportional and excess of loss reinsurance, share of claim amounts, parametric estimation with incomplete information. (10hrs)
- **Unit 2**: Risk models - models for claim number and claim amount in short term contracts, moments, compound distributions, moments of insurer's and reinsurer's share of aggregate claims. (10hrs)
- **Unit 3**: Review of Bayesian statistics - estimation and application. Experience rating - rating methods in insurance and banking, claim probability calculation, stationary distribution of proportion of policyholders in various levels of discount. (10hrs)
- **Unit 4**: Delay/run-off triangle - development factor, basic and inflation-adjusted chain-ladder method, alternative methods, average cost per claim and Bornhuetter Ferguson methods for outstanding claim amounts, statistical models. (10hrs)

Books for Reference:

1. Bowers, N. L., Gerber, H. U., Hickman, J. C., Jones, D. A. and Nesbitt, C. J. (1997). *Actuarial Mathematics* (2nd ed.). Society of Actuaries.
2. Klugman, S. A., Panjer, H. H., Willmotand, G. E. and Venter, G. G. (1998). *Loss Models: From Data to Decisions*. John Wiley \& Sons.
3. Daykin, C. D., Pentikainen, T. and Pesonen, M. (1994). *Practical Risk Theory for Actuaries*. Chapman Hall.

PATTERN RECOGNITION AND IMAGE PROCESSING (3 CREDITS)

Rationale/Learning Objectives:

This course provides theoretical knowledge on various statistical techniques involved in pattern recognition and image processing for data analysis and inference.

Pattern Recognition

- **Unit 1**: Review of Bayes classification - error probability, error bounds, Bhattacharya bounds, error rates and their estimation. Parametric and nonparametric learning, density estimation. Classification trees. k-NN rule and its error rate. (8hrs)
- **Unit 2**: Neural network models for pattern recognition - learning, supervised and unsupervised classification. Unsupervised classification - split/merge techniques, hierarchical clustering algorithms, cluster validity, estimation of mixture distributions. (8hrs)
- **Unit 3**: Feature selection - optimal and suboptimal algorithms. Some of the other approaches like the syntactic, the fuzzy set theoretic, the neurofuzzy, the evolutionary (based on genetic algorithms), and applications. Some recent topics like data mining, support vector machines, etc. (8hrs)

Books for Reference:

1. Fukunaga, K. (1990). *Introduction to Statistical Pattern Recognition* (2nd ed.). New York: Academic Press.
2. Devijver, P. A. and Kittler, J. (1982). *Pattern Recognition: A Statistical Approach*. Prentice Hall.
3. Jain, A. K. and Dube, R. C. (1988). *Algorithms for Clustering Data*. Prentice Hall.
4. Everitt, B. S. (1993). *Cluster Analysis*. Halsted Press.
5. Fu, K. S. (1982). *Syntactic Pattern Recognition and Applications*. Prentice Hall.
6. Bezdek, J. C. (1981). *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum Press.
7. Hastie, T., Tibshirani, R. and Friedman, J. H. (2001). *Elements of Statistical Learning*. Springer-Verlag.
8. Ripley, B. D. (1996). *Pattern Recognition and Neural Networks*. Cambridge University Press.
9. Theodoridis, S. and Koutroumbas, K. (1999). *Pattern Recognition*. Academic Press.

Image Processing

- **Unit 4**: Introduction, image definition and its representation. Typical IP operations like enhancement, contrast stretching, smoothing and sharpening, greylevel thresholding, edge detection, medial axis transform, skeletonization/thinning, warping. (8hrs)
- **Unit 5**: Segmentation and pixel classification, object recognition, some statistical (including Bayesian) approaches for the above, like Besag's ICM algorithm, deformable templates approach of Grenander. (8hrs)

Books for Reference:

1. Young, T. Y. and Fu, K. S. (1986). *Handbook of Pattern Recognition and Image Processing*. Vols. 1 & 2, Academic Press.
2. Jain, A. (1989). *Fundamentals of Digital Image Processing*. Prentice Hall.
3. Castleman, K. R. (1996). *Digital Image Processing*. Prentice Hall.
4. Mardia, K. V. and Kanji, G. K. (1993). *Statistics and Images*. Carfax.

OPERATIONS RESEARCH (3 Credits)

Rationale/Learning Objectives:

This course provides theoretical foundations on optimization techniques for managerial decision making process.

- **Unit 1**: Linear programming problem (LPP) - definition, formulation. Simplex method - canonical form, improving non-optimal basic feasible solution (B.F.S), conditions for optimality, conditions for unboundedness. Convex sets, geometry of simplex method extreme point and B.F.S, existence of B.F.S, existence of optimal B.F.S. Two phase method, big M method. (10hrs)
- **Unit 2**: Duality theory of LPP - weak duality theorem and its properties, the fundamental duality theorem, complementary slackness theorem. Dual simplex method. Sensitivity analysis. Integer programming cutting plane technique, Gomory's algorithm for pure integer program. Dynamic Programming, multistage decision making problems, Bellman's principle of optimality, recursive nature of computation, application, applications of dynamic programming. (10hrs)
- **Unit 3**: Inventory theory - nature of inventory problem, motives for carrying inventory, deterministic inventory model with decay. Probabilistic inventory models, continuous review and periodic review systems, (s, S) policy, heuristic solution of lot size reorder point model ((Q, r) policy). (10hrs)
- **Unit 4**: Queuing theory - characteristics of queues, M/M/1 system, steady state solution, measures of effectiveness, waiting time distributions, Little's formula, M/M/1/K system, M/M/C system, machine interference problem, M/G/1 system, Pollaczek Khintchine formula. (10hrs)

Books for Reference:

1. Gross, D. and Harris, C. M. (1985). *Fundamentals of Queuing Theory* (2nd ed.). John Wiley.
2. Hadley, G. (1975). *Linear and Combinatorial Programming*. John Wiley & Sons.
3. Murty, K. G. (1976). *Linear and Combinatorial Programming*. John Wiley & Sons.
4. Kambo, N. S. (1991). *Mathematical Programming Techniques*. Affiliated East West Press.
5. Taha, H. A. (2001). *Operations Research: An Introduction* (6th ed.). India: Prentice Hall.
6. Sivazlian, B. D. and Stanfel, L. E. (1975). *Analysis of Systems in Operations Research*. Prentice Hall.
7. Daellenbach, H. G. and George, J. A. (1978). *Introduction to Operations Research Techniques*. Allyn and Bacon Inc.

CERTIFICATE COURSE ON MICROSOFT EXCEL (Basic to Advance)

This Course helps you to master in Excel, and learn the powerful features Excel has to offer to analyze the data.

Learning Objectives:

- Understand the practicality of excel.
- Knowledge of formatting, functions & formulas.
- Learn to use advanced features, graphs & presentation techniques to maximize impact.
- Perform data cleaning, processing & manipulation techniques using superpower functions & formulas.
- Build a dashboard / summary report with dynamic charts & tables.
- How macros and VBA automate your spreadsheets and increase interactivity.

Learning Outcome:

- Apply visual elements and advanced formulas to a worksheet to display data in various formats.
- Learn to use advanced functions & features of excel to improve productivity, enhance spreadsheets with templates, charts, graphics, and formulas and streamline the operational work.
- Automate common tasks & apply more advanced analysis techniques to more complex data.

Course Syllabus:

Section 1: Introduction to Excel	**Hours: 1**
a) Purpose & application of Excel, Understanding the Excel interface - Menu Options, Create & Save Spreadsheets, Save As Formats, Limitations, Insert & delete rows / columns, Printing.	
b) Navigation & Editing: Moving around the spreadsheets, Entering information into cells, Types of data, Clipboard, Transformation, Hide rows/columns.	
c) Protecting & Sharing: Protect sheet / workbook - Locking cells.	

Section 2: Data Handling	Hours: 12
a) Sorting, Filters & Advanced Filters, Remove duplicates, Text to columns, Cell reference.	
b) Presentation: Formatting - Cell (Alignment, Height & width, Wrapping, Merging), Numbers (Currency, %, Decimal, negative), Custom Format.	
c) Conditional Formatting - Changing the format of the values depends upon the cell value, conditional format formulas	
d) Data Cleaning - Extracting / Combining text, for typos & bugs – LEFT (), RIGHT (), LEN (), FIND ()	
e) Performing Math with Date & Time: TODAY (), NOW (), DATEVALUE (), YEAR (), MONTH (), DAY (), TEXT ()	
f) Lookup & Reference: VLOOKUP, HLOOKUP, INDEX, MATCH, OFFSET & NAMED RANGES, INDIRECT	
g) Logical Functions: Automatic decision making - IF ELSE, AND, OR, NOT, NESTED IF ELSE.	
h) Information Functions : ISERROR, ISBLANK, CELL, ISTEXT	
i) Text Functions: TRIM, MID, LOWER, UPPER, PROPER, REPT, TRUNC, CONCATENATE etc…	
j) Formula Evaluation: Debugging errors in formula.	

Section 3: Data Validation	Hours: 1
a) Controlling user inputs to reduce the risk of error & increase efficiency – Validation Criteria (List, Date, Time, Text Length etc…)	

Section 4: Data Analysis	Hours: 16
a) Summarizing Data - SUM () Family, COUNT () Family, AVERAGE, MEDIAN, MIN, MAX, STDEV etc…	
b) Array Formulas – Perform multiple calculations in one cell - SUMPRODUCT	
c) Pivot Tables & Pivot Charts, Adding Slicers - Value Field Settings, Filtering, Grouping, Sorting, Changing layout & format etc…	

d) Data Visualization - Charts & Sparkline's: Static & Dynamic charts, formatting

& Designing.
e) Analysis Tools - Apply various statistical methods to analyze the data. o Correlation Analysis o OLS Regression: Simple & Multiple Linear Regression Analysis. o ANOVA: Single / Two Factor o Random Number Generation o T - test (Paired / Two samples) o Understanding & Interpretation of statistical results.

Section 5: Visual Basic and Macro	**Hours: 10**
a) Recording Macro and understanding the code behind.	
b) Create macros by writing VBA scripts.	
c) Creating user forms & recording the data.	
d) User defined functions with VBA.	
e) Adding Add-Ins in Excel.	

Learning Method:

- Hands-on Training - Classroom (70%)
- Homework (10%)
- Project (20%)

Assessment:

- Practical Exam: 100 Marks **Total Hours Required: 40**

CERTIFICATE COURSE ON R AND EXCEL FOR DATA SCIENCE

This Course helps you to master in R and Excel, and learn the powerful functions and features to analyze and visualize the data effectively.

Learning Objectives:

- Learn R language fundamentals and basic syntax.
- Learn how to program in R, How to use R for effective data analysis.
- Explore R syntax, functions & packages.
- Analyze real world challenges in data management; explore general practices of data science.
- Understand the practicality of excel.
- Knowledge of formatting, functions & formulas.
- Learn to use advanced features, graphs & presentation techniques to maximize impact.
- Perform data cleaning, processing & manipulation techniques using superpower functions & formulas.
- Build a dashboard / summary report with dynamic charts & tables.

Learning Outcome:

- Become familiar with the major R data structures.
- Create your own functions & visualizations.
- Learn to write your own project syntax ranging from importing data into R to apply standard and more advanced statistical analysis methods.
- Apply visual elements and advanced formulas to a worksheet to display data in various formats.
- Learn to use advanced functions & features of excel to improve productivity, enhance spreadsheets with templates, charts, graphics, and formulas and streamline the operational work.

Course Syllabus: R Programming

Section 1: Getting Started with R **Hours: 1**
a) History of R, Installation of R & R studio, Loading Add-on packages, choosing repositories, Accessing data in packages.
b) Help & Documentation – Help for Functions/Packages/Data Sets.
c) Data Types - Vectors, Lists, Matrices, Arrays, Factors, Data Frames.
d) Variables - Variable Assignment, Finding & Deleting Variables, Data Type of a Variable.
e) Operators - Arithmetic, Relational, Logical, Assignment & Miscellaneous Operators.
Section 2: Programming Language Basics **Hours 6**
a) Simple Manipulations: Numbers & Vectors, Vectors & Assignment, Vector Arithmetic, Logical Vectors, Character Vectors.
b) Generating Sequences & Missing Values.
c) Index Vectors: Selecting & Modifying subsets of a data set.
d) Objects, their modes & attributes: Intrinsic attributes: mode and length, changing the length of an object, class of an object.
e) Ordered and unordered factors: A specific example, The function tapply () and ragged arrays.
f) Arrays and matrices: Arrays, Array indexing - Subsections of an array, Index matrices.
g) The array () function: Mixed vector and array arithmetic. The recycling rule, The outer product of two arrays, Generalized transpose of an array.
h) Matrix facilities: Matrix multiplication, Linear equations and inversion, Forming partitioned matrices cbind () and rbind (), The concatenation function c() with arrays, Frequency tables from factors.
i) Lists and Data frames: Lists, Constructing and modifying lists, Concatenating lists. Data Frames, Making data frames, working with data frames.

Section 3: Functions	Hours: 2
a) Built-in functions (Numeric/Character/Statistical/Other), User define function, Calling function, Defining new binary operators, Assignments within functions, more advanced examples, Applying functions to matrices & data frames	

Section 4: Advanced Data Management Hours: 9
a) Data Input & Output: Changing directories, Managing files & workspace, Reading data from files, writing data from R, Connection to External data sources.
b) View / Edit Data, Objects / Variable types, Converting Objects / Variables, Selecting Variables / Observations.
c) Applying Functions: lappy, sapply, tapply, apply, mapply
d) Combining Variables with c, cbind, rbind functions
e) Combining data with Vector / Matrix / Data Frame / List Function
f) Working with Date & Time
g) Finding NA / NaN & Replacing
h) Conditional Transformation / Decision Making
i) Control Flow (Repetition & Looping / Conditional Execution)
j) Variables: Renaming Variables / Observations, Creating New / Recoding Variables, Keeping & Dropping Variables
k) Generating Random Numbers
l) Data Sets : Stacking/Concatenating/Adding Datasets, Joining / Merging Data Frames
m) Summary: Creating Summarized / Aggregated Datasets (dplyr)
n) Reshaping the data (Reshape Package)
o) Removing Duplicates, Sorting the Data Frames
p) Value Labels or Formats (and Measurement Level)

Section 5: Data Visualization	Hours: 2
a) Traditional Graphics, Graphics with ggplot2 & Advanced Graph Types.	

Section 6: Data Analysis	Hours: 5
b) Basic Statistics: Descriptive Statistics, Frequency & Contingency tables, T – tests, Non-parametric tests of group differences	
c) Predictive Modeling: Correlation Analysis, Splitting data into training & validation, OLS Regression - Simple / Multiple Linear Regression.	
d) Regression diagnostics: Non-Normality, Multicolinearity, Non-linearity, Non-constant error variance.	
e) Unusual Observations & Corrective measures: Outliers, High-leverage points & Influential Observations.	
f) Choosing Best Regression Model: Comparing Models, Variable selection	
g) Model Validation: Cross-Validation	
h) Assessment of Regressors: Relative Importance.	

Course Syllabus: Excel

Section 1: Introduction to Excel	Hours: 1
a) Purpose & application of Excel, Understanding the Excel interface - Menu Options, Create & Save Spreadsheets, Save As Formats, Limitations, Insert & delete rows / columns, Printing.	
b) Navigation & Editing: Moving around the spreadsheets, Entering information into cells, Types of data, Clipboard, Transformation, Hide rows/columns.	
c) Protecting & Sharing: Protect sheet / workbook - Locking cells.	
Section 2: Data Handling	**Hours: 7**
a) Sorting, Filters & Advanced Filters, Remove duplicates, Text to columns, Cell reference.	
b) Presentation: Formatting - Cell (Alignment, Height & width, Wrapping, Merging), Numbers (Currency, %, Decimal, negative), Custom Format.	
c) Conditional Formatting - Changing the format of the values depends upon the cell value, conditional format formulas	
d) Data Cleaning - Extracting / Combining text, for typos & bugs – LEFT (), RIGHT (), LEN (), FIND ()	

e) Performing Math with Date & Time: TODAY (), NOW (), DATEVALUE (), YEAR (), MONTH (), DAY (), TEXT ()

f) Lookup & Reference: VLOOKUP, HLOOKUP, INDEX, MATCH	
g) Logical Functions: Automatic decision making - IF ELSE, AND, OR, NOT, NESTED IF ELSE.	
h) Information Functions : ISERROR, ISBLANK, CELL, ISTEXT	
i) Text Functions: TRIM, MID, LOWER, UPPER, PROPER, REPT, TRUNC, CONCATENATE etc…	
Section 3: Data Analysis	**Hours: 7**
a) Summarizing Data - SUM () Family, COUNT () Family, AVERAGE, MEDIAN, MIN, MAX, STDEV etc…	
b) Array Formulas – Perform multiple calculations in one cell - SUMPRODUCT	
c) Pivot Tables & Pivot Charts, Adding Slicers - Value Field Settings, Filtering, Grouping, Sorting, Changing layout & format etc…	
d) Data Visualization - Charts & Sparkline's: Static & Dynamic charts, formatting & Designing.	
e) Analysis Tools - Apply various statistical methods to analyze the data. • Correlation Analysis • ANOVA : Single / Two Factor • Random Number Generation • T - test (Paired / Two samples) • Understanding & Interpretation of statistical results.	

Learning Method:

- Classroom (70%)
- Homework (10%)
- Project (20%)

Assessment:

- Practical Exam: 100 Marks **Total Hours Required: 40**

CERTIFICATE COURSE ON R FOR DATA SCIENCE

This course will offer you to learn Data Science in R from scratch.

Learning Objectives:

- Learn R language fundamentals and basic syntax.
- Learn how to program in R, How to use R for effective data analysis.
- Explore R syntax, functions & packages.
- Analyze real world challenges in data management; explore general practices of data science.

Learning Outcome:

- Become familiar with the major R data structures.
- Create your own functions & visualizations.
- Learn to write your own project syntax ranging from importing data into R to apply standard and more advanced statistical analysis methods.

Course Syllabus: R Programming

Section 1: Getting Started with R **Hours: 2**
a) History of R, Installation of R & R studio, Loading Add-on packages, choosing repositories, Accessing data in packages.
b) Help & Documentation – Help for Functions/Packages/Data Sets.
c) Data Types - Vectors, Lists, Matrices, Arrays, Factors, Data Frames.
d) Variables - Variable Assignment, Finding & Deleting Variables, Data Type of a Variable.
e) Operators - Arithmetic, Relational, Logical, Assignment & Miscellaneous Operators.

Section 2: Programming Language Basics **Hours 9**
a) Simple Manipulations: Numbers & Vectors, Vectors & Assignment, Vector Arithmetic, Logical Vectors, Character Vectors.
b) Generating Sequences & Missing Values.
c) Index Vectors: Selecting & Modifying subsets of a data set.

d) Objects, their modes & attributes: Intrinsic attributes: mode and length, changing the length of an object, class of an object.
e) Ordered and unordered factors: A specific example, The function tapply () and ragged arrays.
f) Arrays and matrices: Arrays, Array indexing - Subsections of an array, Index matrices.
g) The array () function: Mixed vector and array arithmetic. The recycling rule, The outer product of two arrays, Generalized transpose of an array.
h) Matrix facilities: Matrix multiplication, Linear equations and inversion, Forming partitioned matrices cbind () and rbind (), The concatenation function c() with arrays, Frequency tables from factors.
i) Lists and Data frames: Lists, Constructing and modifying lists, Concatenating lists. Data Frames, Making data frames, working with data frames.
Section 3: Functions **Hours: 3**
a) Built-in functions (Numeric/Character/Statistical/Other), User define function, Calling function, Defining new binary operators, Assignments within functions, more advanced examples, Applying functions to matrices & data frames
Section 4: Advanced Data Management **Hours: 12**
a) Data Input & Output: Changing directories, Managing files & workspace, Reading data from files, writing data from R, Connection to External data sources.
b) View / Edit Data, Objects / Variable types, Converting Objects / Variables, Selecting Variables / Observations.
c) Applying Functions: lappy, sapply, tapply, apply, mapply
d) Combining Variables with c, cbind, rbind functions
e) Combining data with Vector / Matrix / Data Frame / List Function
f) Working with Date & Time
g) Finding NA / NaN& Replacing
h) Conditional Transformation / Decision Making
i) Control Flow (Repetition & Looping / Conditional Execution)
j) Variables: Renaming Variables / Observations, Creating New / Recoding

Variables, Keeping & Dropping Variables
k) Generating Random Numbers
l) Data Sets : Stacking/Concatenating/Adding Datasets, Joining / Merging Data Frames
m) Summary: Creating Summarized / Aggregated Datasets (dplyr)
n) Reshaping the data (Reshape Package)
o) Removing Duplicates, Sorting the Data Frames
p) Value Labels or Formats (and Measurement Level)
Section 5: Data Visualization **Hours: 5**
a) Traditional Graphics, Graphics with ggplot2 & Advanced Graph Types.
Section 6: Data Analysis **Hours: 9**
b) Basic Statistics: Descriptive Statistics, Frequency & Contingency tables, T – tests, Non-parametric tests of group differences
c) Predictive Modeling: Correlation Analysis, Splitting data into training & validation, OLS Regression - Simple / Multiple Linear Regression.
d) Regression diagnostics: Non-Normality, Multicollinearity, Non-linearity, Non-constant error variance.
e) Unusual Observations & Corrective measures: Outliers, High-leverage points & Influential Observations.
f) Choosing Best Regression Model: Comparing Models, Variable selection
g) Model Validation: Cross-Validation
h) Assessment of Regressors: Relative Importance.

Learning Method:

- Classroom (70%)
- Homework (10%)
- Project (20%)

Assessment:

- Practical Exam: 100 Marks **Total Hours Required: 40**

CERTIFICATE COURSE ON R FOR ADVANCED STATISTICAL METHODS & MACHINE LEARNING

This course will offer you to learn how to apply advanced statistical methods & machine learning algorithms.

Learning Objectives:

- Apply advanced statistical methods which include discovery & exploration of complex multivariate relationships among variables.

Learning Outcome:

- Become familiar with the major R data structures.
- Create your own functions & visualizations.
- Learn to write your own project syntax ranging from importing data into R to apply standard and more advanced statistical analysis methods.

Course Syllabus:

Section 1: Advanced Statistical Methods Using R	**Hours: 20**
a) Generalized Linear Models: Logistic Regression, Multinomial Logistic Regression, Poisson Regression	
b) Ridge Regression	
c) Forecasting: Time Series Analysis – ARIMA	
d) Cluster Analysis: Hierarchical & Portioning cluster analysis (K- Means)	
e) Classification: Decision Tree & CHAID	
f) Text Mining: Word Cloud	
g) Dimensionality Reduction: PCA & Factor Analysis	
Section 2: Machine Learning with R	**Hours: 20**
a) Classification: Support Vector Machine, Random Forest Method, Navie Bayes	
b) Gradient Boosting Model	
c) Artificial Neural Network – Single Layer & Multiple Layer	

Learning Method:

- Classroom (70%)
- Homework (10%)
- Project (20%)

Assessment:

- Practical Exam: 100 Marks **Total Hours Required: 40**