KANNUR UNIVERSITY

(Abstract)

M.Sc Statistics with Data Analytics programme- Scheme, Syllabus and Model Question Papers - Core/Elective Courses under Credit Based Semester System (CBSS) - Implemented with effect from 2022 Admission - Orders issued.

ACADEMIC C SECTION

AcadC/16764/MSc/ Stati- Data Analysis/2022 Dated: 07.02.2023

Read:-1. U.O No. Acad C1/11460/2013 dated 12.03.2014.

2. U Os No. Acad C1/11460/2013 dated 05.12.2015, 22.02.2016, 29.04.2017 & 01.08.2018.

3. U.O No. ACAD A/ASO A2/14768/2022 Dated 09.12.2022.

4. Letter No Nil Dated 19.09.2022.

5. Letter No. AcadC/16764/MSc/ Stati- Data Analysis/2022 Dated: 22.12.2022.

6. E mail from HoD, Dept. of Statistical Sciences, Dated 06.01.2023 forwarding the remarks

7. Letter No. AcadC/16764/MSc/ Stati- Data Analysis/2022 Dated: 21.01.2023.

8. E-mail dated 23.01.2023, received from the Principal, Don Bosco Arts and Science College, Angadikadavu

ORDER

1. As per the paper read (1) above, the Regulations for PG Programmes under Credit Based Semester System (CBSS) in Affiliated Colleges were implemented in the University w.e.f 2014 admission and certain modifications were effected to the same, as per the paper read (2) above.

2. Provisional Affiliation was sanctioned to M.Sc Statistics with Data Analytics Programme at Don Bosco Arts and Science College, Angadikadavu for the academic year 2022-23, as per the paper read (3) above.

3. As Kannur University is not having the Syllabus of aforementioned subject and all activities of the reconstituted Board of Studies are kept in abeyance in the light of the judgement in WA No.1530/2021 Dated 22.03.2022, the Principal, Don Bosco Arts & Science College submitted a draft Scheme, Syllabus and Model Question Papers of M.Sc Statistics with Data Analytics Programme, prepared by Dr.A.P Kuttykrishanan (statistics), former Pro Vice Chancellor of this University in consultation with four other subject experts, as per the paper read (4) above.

4. As per the paper read (5) above, the HoD, Dept. of Statistical Sciences, Kannur University was requested to scrutinize the Scheme, Syllabus and Model Question Papers of aforesaid programme and he forwarded the remarks/recommendations as per paper read (6) above.

5. As per the paper read (7) above, the Principal, Don Bosco Arts and Science College, Angadikadavu was requested to incorporate the recommendations of the HoD, Dept. of Statistical Sciences and to modify the draft scheme, Syllabus and Model Question Papers of MSc Statistics with Data Analytics Programme for implementation w.e.f 2022

7. The Principal, Don Bosco Arts and Science College, Angadikadavu submitted the modified Scheme Syllabus and Model Question Papers of MSc Statistics with Data Analytics Programme

after incorporating the suggestions of the the HoD, Dept. of Statistical Sciences, as per the paper read (8) above.

8. The Vice Chancellor, after considering the matter in detail, and in exercise of the powers of the Academic Council conferred under section 11(1) Chapter III of the Kannur university Act 1996, has accorded sanction to implement the Scheme, Syllabus and Model Question papers of MSc Statistics with Data Analytics Programme (CBSS) at Don Bosco Arts and Science College, Angadikadavu, w.e.f 2022 admission onwards and to report the same to the Academic Council.

9. The Scheme, Syllabus and Model Question papers of MSc Statistics with Data Analytics Programme (CBSS) are uploaded on the website of the University.

10. Orders are issued accordingly

Sd/-
Narayanadas K
DEPUTY REGISTRAR (ACAD)
For REGISTRAR

To:
1. The Principal, Don Bosco Arts and Science College, Angadikadavu
2. The Examination Branch (Through PA to CE)

Copy To:
1. PS to VC/PA to PVC/PA to R
2. DR/AR I/AR II (Academic)
3. The Computer Programmer
4. The Web manager (for uploading on the University Website)
5. EG 1/EX CI (Exam)
6. SF/DF/FC

Forwarded / By Order

SECTION OFFICER

KANNUR UNIVERSITY
CIVIL STATION P.O.
KANNUR
Pin-670 002
THAVAKKARA

Scheme and Syllabus

M. Sc Statistics with Data Analytics Programme

(UNDER CREDIT BASED SEMESTER SYSTEM (CBSS)

FROM 2022 ADMISSION ONWARDS

1. **Programme Specific Outcomes (PSO)**

- Expertise in the field of statistical theory and its applications.
- Expertise on data analysis using statistical techniques.
- Expertise to use statistical software for data analysis.
- Enables to apply data analysis tools using computer programming.
- Expertise to take up responsibilities as efficient Statistician/Data Analysis Expert/Research Officers in various fields.
- Enable to lead the team to conduct Statistical survey and preparing tools for the conduct survey.

2. **Eligibility for Admission**:

Candidates who have successfully completed any of the following three degree programmes are eligible for admission to M. Sc Statistics with Data Analysis Programme, as per the existing University/ Government orders.

- B.Sc. Degree with Statistics or Applied Statistics as the Core Course (Main) with not less than 50% marks or equivalent grade excluding subsidiaries/ complementary.
- B.Sc. Degree with Statistics & Mathematics double main and B.Sc .Degree with Mathematics as Core Course (Main) and Statistics as one of the complimentary Courses (Subsidiary) with not less than 55% marks or equivalent grade.
- B.Sc. degree with Computer Science as main and Statistics/Mathematics as a complimentary subject with not less than 55% marks or equivalent grade.
- B.Tech/ B.E Degree in any Branch with Mathematics paper as one of the subjects in each semester for the first four semesters of the Programme.
- The index score for preparing the rank list shall be calculated on the basis of the marks/ grade points of Main (Core Courses) and Subsidiaries (Complimentary courses) scored by the candidates in the B.Sc Degree programme. 50% of the total seats in each category shall be reserved for the B.Sc degree holders in Statistics single main or Statistics and Mathematics double main if there are claimants. For B.Tech/B.E candidates index score is calculated on the basis of total marks/grade points of the B.Tech/B.E programme and marks/grade points scored in Mathematics courses of the programme.

3. **Duration of the Programme**: The duration of M.Sc Statistics with Data Analysis

Programme shall be a minimum of 2 years consisting of 4 Semesters. Each Semester consists of a minimum of 450 contact hours distributed over 90 working days.

4. **Structure of the Programme:** The total credits for the Programme is 80. Core courses have a total credit of 68 and elective have 12 credits. Core Course is a Course that every student admitted to the Programme must successfully complete to receive the degree and cannot be substituted by any other Course. An Elective Course is a Course which can be substituted by an equivalent Course from the Subject.

5. **Evaluation Scheme:** Evaluation scheme of each course shall contain two parts:(i) Continuous Evaluation (CE) and (ii) End Semester Evaluation (ESE) 20% weightage shall be given for CE and 80% weightage shall be given for ESE.

6. **End Semester Examination for Theory papers:**

   For each paper the duration of the examination is 3 hours and maximum mark is 80. The question paper will have 3 parts: Part A, Part B and Part C.

   Part A will consist of 8 short answer each carrying 2marks and a candidate has to answer all of them. (16 marks)

   Part B will consist of 6 questions each carrying 4 marks and the candidate has to answer 4 questions from this part (16 marks)

   Part C will consist of 6 questions each carrying 12 marks and the candidate has to answer4 questions from this part. (48 marks)

7. **End Semester Examination for Practical Papers:**

   The practical paper will be conducted using R and Python Programme.

   For each practical paper, a record of work done by the student should be prepared and submitted for internal evaluation. The components of CA mark for the practical paper are Attendance (2 marks), Record book (8 marks) and class test (10 marks). The Board of examiners will prepare the question paper for the practical examination (ESE) covering the papers specified in the syllabus. An external examiner along with an internal examiner, appointed by the University will conduct the practical examination and its evaluation. For each practical paper, the duration of the examination is 3 hours and the maximum mark is 80.

   End-semester Evaluation in practical courses shall be conducted and evaluated by two examiners-one internal and one external.

   A candidate shall be permitted to appear for the ESE of a Practical Course only if she/ he has submitted the Record certified by the concerned Head of the Department.

8. **Continuous Evaluation (CE):** CE of a course shall be based on a transparent system consisting of periodic written tests, assignments, seminars and attendance in respect of theory courses and lab skill, record, tests and attendance in respect of practical courses. The weightage assigned to various components for CE for theory and practical are as follows.

Components of CE (Theory)

	Component	% of internal marks
a	Two test papers	40
b	Assignment	20
c	Seminar	20
d	Attendance	20

Components of CE (Practical)

	Component	% of internal marks
a	Two test papers	40
b	Lab skill	20
c	Record	20
d	Attendance	20

9. **Project work**

A project work has to be done using primary or secondary data and to be submitted at the end of the fourth semester. For field survey project, students can do the survey on a topic and the data analysis and reporting can be done based on the data collected. The project report should consist of literature review, methodology, data analysis and summary.

Evaluation of Project Work: There shall be Continuous Evaluation and End Semester Evaluation in the case of Project Work. Each candidate has to submit a project report/ dissertation by the end of the IV semester. The ESE of the Project Work shall be conducted by two External Examiners, at the end of the Programme only.

10. **Pass Condition**

A candidate with 40% of aggregate marks and 40% separately for ESE for each course shall be declared to have passed in that course. Those who secure not less than 40 % marks (both ESE and CA put together) for all the courses of a semesters hall be declared to have successfully completed the semester. The marks obtained by the candidates for CE in the first appearance shall be retained (irrespective of pass or fail).

The candidates who fail in theory unit shall reappear for theory unit only, and the marks secured by them in practical unit, if passed in practical, will be retained. A candidate who fails to secure a minimum for a pass in a course will be permitted to write the same

examination along with the next batch. For the successful completion of a semester, a candidate should pass all courses.

For the **successful completion of a Programme** and award of the degree, a student must pass all Courses satisfying the minimum credit requirement (80) and must complete all Semester successfully.

Course Structure

Semester I

Course Code	Course Name	Credit	Duration of End Semester Exam	Max. Marks (CE)	Max. Marks(ESE)
MST1C01	Mathematical Methods for Statistics	4	3 hrs	20	80
MST1C02	Probability Theory	4	3hrs	20	80
MST1C03	Distribution Theory	4	3hrs	20	80
MST1C04	Statistical Programming Using R	4	3 hrs	20	80
MST1P01	Statistical Computing-I (Lab based on R Programming)	3	3 hrs	20	80

Semester II

Course Code	Course Name	Credit	Duration of End Semester Exam	Max. Marks (CE)	Max. Marks (ESE)
MST2C05	Data Base Management System with SQL/PL-SQL	4	3 hrs	20	80
MST2C06	Statistical Inference	4	3hrs	20	80
MST2C07	Sampling and Design of Experiments	4	3hrs	20	80
MST2C08	Statistics using Python Programming	4	3 hrs	20	80
MST2P02	Statistical Computing- II (Lab based on Python Programming)	2	3 hrs	20	80

Semester III

Course Code	Course Name	Credit	Duration of End Semester Exam.	Max. Marks (CE)	Max. Marks (ESE)
MST3C09	Regression Analysis	4	3 hrs	20	80
MST3C10	Stochastic Processes and Time Series Analysis	4	3hrs	20	80
MST3C11	Big Data Analytics	4	3hrs	20	80
MST3E	Elective -I	4	3 hrs	20	80
MST3P03	Statistical Computing III (Lab based on R and Python)	3	3 hrs	20	80

Semester IV

Course Code	Course Name	Credit	Duration of End Semester Exam.	Max Marks CE	Max. Marks (ESE)
MST4C12	Multi Variate Analysis	4	3 hrs	20	80
MST4E	Elective – II	4	3hrs	20	80
MST4OE	Open Elective	4	3hrs	20	80
MST4Pr	Project/Internship	8		20	80
MST4C13	Viva Voce	4		20	80

ELECTIVES

Group I (for Third Semester)

MST3E01 Survival Analysis

MST3E02 Queuing Theory

MST3E03	Reliability Modeling
MST3E04	Data Mining

Group II (for Fourth Semester)

MST4E01	Bio Statistics
MST4E02	Analysis of Clinical Trials
MST4E03	Demography
MST4E04	Machine Learning

OPEN ELECTIVE: (Fourth Semester)

MST4OE01	Neural Networks and Deep Learning
MST4OE02	Statistics with SAS

Detailed Syllabus

SEMESTER I

MST1C01: Mathematical Methods for Statistics

Course Outcome:

After successful completion of this course, student will be able to:

1. Understand the concepts of Eigen values and Eigen vectors of matrix.

2. Understand the vector space, matrices and its properties.

3. Solve systems of linear equations using multiple methods.

4. Understand the properties of quadratic forms and generalized inverse.

5. Understand the concept of Metric space and convergence of sequences.

6. Understand Reimann – Stieltjes integral and its properties.

Module 1:

The Geometry of Linear Equations, Example of Gaussian Elimination, Matrix Notation and Matrix Multiplication, Triangular Factors and Row Exchanges, Inverses and Transposes.
Vector Spaces and Subspaces, Solving Ax=0 and Ax=b, Linear dependence and Independence, Basis and Dimension, The Four Fundamental Subspaces, symmetric, orthogonal, and idempotent matrices, Rank of a matrix, Linear Transformations.

Module 2:

Introduction-Properties of the Determinant, Formulas for the Determinant, Applications of Determinants. Eigenvalues and Eigenvectors, Introduction, Diagonalization of a Matrix, Complex Matrices, Similarity Transformations. Quadratic forms, classification of quadratic forms, rank and signature, positive definite and non-negative definite matrices.

Module 3:

Metric spaces, compact set, perfect set, connected set, limit of functions, continuous function, continuity and compactness, continuity and connectedness, discontinuities, monotone functions, derivative of a real valued function, mean value theorem. Reimann- Stieltjes integral and properties.

Module 4:

Sequence of Functions and Functions of Several variables: Sequences and series of functions, uniform convergence. Uniform convergence and continuity, uniform convergence and integration, uniform convergence and differentiation, Weirstrass theorem, improper integrals, the Beta and Gamma functions. Functions of several variables, limits and continuity. Taylor's theorem and its applications. Conditions for the optima of multivariate functions.

REFERENCES

1. Rudin. W. (2013). Principles of Real Analysis (3rd Ed.), McGraw Hill.

2. Apostol T. M. (1974): Mathematical Analysis, Narosa Publishing House, New Delhi.

3. Ramachandra Rao and Bhimasankaran (1992). Linear Algebra. Tata McGraw Hill, NewDelhi.

4. Malik, S.C & Arora, S. (2006). Mathematical Analysis, Second Edition, New-age International Publishers.

5. Mathai, A. M. and Haubold, H. J. (2017). Linear Algebra – A course for Physicists and Engineers, De Gruyter, Germany.

6.Gilbert Strang (2006). Linear Algebra and Its Application, Fourth Edition, AcademicPress.

MST1C02: Probability Theory

Course Outcome:

After successful completion of this course, student will be able to:

1. Understand concepts of measure and probability, sequence of sets, sequence of measurable functions and sequence of integrals.

2. Understand distribution function and its properties.

3. Expectation of random variables and its properties.

4. Understand inequalities involving moments.

5. Understand various laws of large numbers and different central limit theorem, their mutual implications and applications.

Module 1

Sequences and limits of sets, field, Sigma field, measurable space, minimal sigma field, Borel field; Random variable, vector random variable, properties of random variables, Independence of two events, Independence of random variables; Probability space, Monotone and continuity property of probability measure, conditional probability and Bayes' Theorem for a finite number of events, Borel 0-1 law.

Module 2:

Distribution function and its properties, Jordan decomposition theorem (statement only), Correspondence theorem (statement only); Expectation and moments – definitions and simple properties , Moment inequalities – Basic, Markov, Jensen; Characteristic function of a random variable, properties, continuity and inversion theorems of characteristic functions(without proof).

Module 3:

Convergence of random variables, convergence in probability, almost sure convergence, convergence in distribution, and convergence in rth mean, properties and relations among them; complete convergence of distributions, Helly-Bray lemma (statement only), Helly-Bray theorem (statement only).

Module 4:

Central limit theorem-Demoivre-Laplace CLT, Lindberg –Levy CLT,LiaponovCLT(statement only), Lindberg- Feller CLT (statement only); Law of Large Numbers-Weak Law of Large numbers of Bernoulli, Chebychev, Poisson and Khinchine; Kolmogorov strong law of large numbers for independent random variables.

References:

1. Bhat B.R. (2014) Modern Probability theory (An introductory text book), Fourth edition, New Age International.

2. Rohatgi V.K. and SalehM. (2015) An introduction to probability and statistics, Third edition, Wiley.

3. Laha R.G. and Rohatgi V.K. (1979) Probability theory, John Wiley.

4. Basu A.K. (2012). Measure Theory and Probability, Second Edition, PHI Learning Pvt. Ltd, New Delhi.

5.Kingman,J.F.C.andTaylor,S.J.(1977).AtextbookofIntroductiontoMeasureTheoryandProbability ,3rdEdn., CambridgeUniversity Press, London.

6.Feller,W. (1968)IntroductiontoProbabilityTheoryandItsApplicationsVol.1and2,JohnWiley,New York.

MST1C03: Distribution Theory

Course Outcome:

After successful completion of this course, student will be able to:

1. Understand the concepts of discrete and continuous distributions.

2. Understand the normal distribution and various non-normal distributions, their properties and applications for scientific research.

3. Understand the Pearson family of distributions.

3. Understand the concept of multivariate distributions and their marginal and conditional distributions

4. Understand the idea of sampling and sampling distributions from infinite population

Module 1:

Univariate Discrete Distributions: Moments and moment generating functions, probability generating functions, characteristic function. Discrete uniform, binomial, Poisson, geometric, negative binomial, hyper geometric and power series distributions

Module 2:

Univariate Continuous Distributions: Uniform, normal, exponential, Weibull, Pareto, beta, Gamma, Laplace, logistic, Cauchy and log-normal distributions; Pearson family of distributions.

Module 3

Bivariate and Multivariate Distributions: Joint, marginal and conditional distributions, independence, covariance and correlations, functions of random variables and their distributions. Jacobin of transformations, bivariate normal distribution, multinomial distribution and their marginal and conditional distributions.

Module 4

Sampling Distributions: Basic concepts of sampling distributions from infinite populations, sampling from normal distributions, properties of sample mean and sample variance. Chi-square, t-distribution and F distributions, properties and applications. Non-central Chi-square, t and F-distributions. Basic concepts of order statistics and their distributions. Distribution of rth order statistics, distribution of sample median and range (for Uniform (0, 1) distribution only).

REFERENCES

1. Rohatgi, V.K. (2001). An Introduction to Probability and Statistics, 2nd Edition. John Wiley and Sons.

2. Krishnamurthy, K. (2006). Handbook of Statistical Distributions with Applications. Chapman & Hall/CRC, New-York.

3. Johnson, N.L., Kotz, S. and Balakrishnan, N. (1995). Continuous Univariate Distributions, Vol. I &Vol. II, John Wiley and Sons, New-York

4. Johnson, N.L., Kotz. S. and Kemp. A.W. (1992). Univariate Discrete Distributions, John Wiley and Sons, New York.

5. Stuart, A. Ord, A. (1994). Kendall's Advanced Theory of Statistics, Distribution Theory,6thEdition, Wiley-Blackwell.

6. Gupta, S.C. and Kapoor, V.K. (2000). Fundamentals of Mathematical Statistics, 10 th Revised Edition. Sultan Chand & Sons, New Delhi.

MST1C04: Statistical Programming Using R

Course Outcome:

After successful completion of this course, student will be able to:

1. Understand various built-in functions in R programming for statistical data analysis.
2. Understand different functions in R programming for writing computer programmes and develop computer programmes for different problems.
3. Plot cdf and pdf of standard distributions using R.
3. Test of significance of means, ANOVA, non-parametric tests, simple correlation and regression procedures and apply for real data sets.

Module 1:

Basic Concepts of R Programming: Introduction to R- Objects and their classes, operators, vectors and matrices, list and data frames, indexing and accessing data, importing and exporting data. Common built-in functions. Simple applications - Descriptive statistics. R-Graphics-Histogram, Box-plot, Stem and leaf plot, Scatter plot, Q-Q plot. Looping-For loop, repeat loop, while loop, if command, if else command.

Module 2:

Matrices and Standard Probability Distributions: Matrices, rank, determinants and inverse. Eigen values and vectors, power of matrices, g-inverse, system of linear equations, roots of algebraic and transcendental equations. Plotting of cdf and pdf for different values of the parameters of standard distributions. Generations of random samples from standard distributions, demonstrations of the sampling distributions of the standard statistics and functions of random variables-distribution of sample mean and sample variance, illustration of laws of large numbers, central limit theorems.

Module 3:

Sampling Methods: Random sample selections, estimation of mean, proportion, variance, confidence interval and efficiency under SRS, stratified random sampling, Various kind of allocation, stratification, estimators based on ratio and regression methods,pps sampling, two stage cluster sampling, and systematic sampling.

Module 4:

Testing of Hypothesis &Re-sampling Methods: Power function and OC function, parametric and non-parametric tests, single sample tests, two sample tests, test for independence, one way and two-way ANOVA, Sequential Probability Ratio Test Bootstrap methods, bias and standard errors, bootstrapping for estimation of sampling distribution, confidence intervals, variance stabilizing transformation, Bootstrapping in regression. Jackknife and cross validation: jackknife in sample surveys.

REFERENCES:

1. Maria D.U., Ana F.M. and Alan T.A. (2008*): Probability and Statistics with R.* CRC Press.
2. Dalgaard, P. (2008): *Introductory Statistics with R, (Second Edition),* Springer.
3. Purohit, S.G, Ghore,S.D and Deshmukh, S.R.(2004): *Statistics Using R.* Narosa.
4. Maria L. Rizzo (2008): *Statistical Computing with R*, Chapman & Hall/CRC.
5. Maria D. U., Ana F. M. and Alan T. A. (2008*). Probability and Statistics with R.* CRC Press.
6. Peter Dalgaard (2008). *Introductory Statistics with R, Second Edition, Springer.*
7. Draper, N. R. and Smith, H. (1998): *Applied Regression Analysis, (3rdEdition).* John Wiley, New York.
8. Casella, G. and Berger, R.L. (2002). *Statistical Inference, 2nd Edition,* Duxbury, Australia.

MST1P01: Statistical Computing-I (Lab using R Programming)

Course Outcome:

After successful completion of this course, student will be able to:

1. Acquire practical knowledge of different theoretical methods.

2. Improve the basic concepts of statistical theories using practical data.

3. Develop their ability to handle real world problems with large scale data.

Practical based on data with respect to problems discussed in module 1 and module 2 of 1st semester paper- Programming Using R.

**

SEMESTER II

MST2C05: Database Management Systems with SQL/PL-SQL

Course Outcome:

After successful completion of this course, student will be able to:

1. Understand important terms related to DBMS.

2. Analyze various normal forms.

3. Able toapply structured query language for Data Analytics.

4. Demonstrate the use of PL/SQL for Data Analytics.

Module 1:

Introductory concepts of DBMS, Introduction and applications of DBMS, Purpose of data base, Data, Independence, Database System architecture- levels, Mappings, Database, users and DBA. Relational Model: Structure of relational databases, Domains, Relations, Entity-Relationship model Basic concepts, Design process, constraints, Keys, Design issues, ER diagrams, weak entity sets, Extended E-R features – generalization, specialization, aggregation, reduction to E-R database schema

Module 2:

Relational Database design Functional Dependency – definition, trivial and non-trivial FD, closure of FD set, closure of attributes, irreducible set of FD, Normalization – 1Nf, 2NF, 3NF, Decomposition using FD, dependency preservation, BCNF, Multivalued dependency, 4NF, Join dependency and 5NF

Module 3:

SQL Concepts Basics of SQL, DDL, DML, DCL, structure – creation, alteration, defining constraints – Primary key, foreign key, unique, not null, check, IN operator, Functions - aggregate functions, Built-in functions – numeric, date, string functions, set operations, sub queries, correlated sub-queries, Use of group by, having, order by, join and its types, Exist, Any, All , view and its types. transaction control commands – Commit, Rollback, Save point

Module 4:

PL/SQL Introduction to PL/SQL, PL/SQL Identifiers, Control Structures, Composite Data Types, Explicit Cursors, Stored Procedures and Functions, Triggers, Compound, DDL, and Event Database Triggers.

REFERENCES

1. Raghu Ramakrishnan and Johannes Gehrke, Third Edition, McGraw Hill, 2003

2. Database Systems: Design,Implementation and Management, Peter Rob, Thomson Learning, 7thEdn. Concept of Database Management, Pratt, Thomson Learning, 5Edn.

3. Database System Concepts – Silberchatz, Korth and Sudarsan, Fifth Edition, McGraw Hill, 2006 4. The Complete Reference SQL – James R Groff and Paul N Weinberg

MST2C06: Statistical Inference

Course Outcomes:

After successful completion of this course, student will be able to:

1. Apply various parametric techniques with real life examples.

2. Understand the concepts of Sufficiency, Completeness and Minimum Variance Unbiased Estimation and various estimation methods and applications in real life problems

3. Apply various parametric, non-parametric and sequential testing procedures to deal with real life problems.

4. Understand various non-parametric tests used for different problems and Sequential Probability Ratio Test and developing SPRT for different situations.

Module 1:

Properties of Estimators and Minimum Variance Unbiased Estimation Sufficient statistics and minimum variance unbiased estimators, factorization theorem for sufficiency (proof for discrete distributions only), joint sufficient statistics, exponential family, minimal sufficient statistics, criteria to find the minimal sufficient statistics, ancillary statistics, complete statistics, Basu's theorem (proof for discrete distributions only), Unbiased estimator, Best Linear Unbiased Estimator(BLUE), Minimum Variance Unbiased Estimator (MVUE), Fisher information, Cramer-Rao inequality and its applications, Rao-Blackwell theorem, Lehmann - Scheffe theorem. Consistent estimators and consistent asymptotically normal estimators. Invariance property of estimators and consistent asymptotically normal (CAN) estimators.

Module 2:

Methods of Estimation: Method of moments, Method of maximum likelihood (MLE), MLE in exponential family, one parameter Cramer family, Cramer- Huzurbazar theorem. Interval estimation, shortest expected length confidence interval, large sample confidence intervals.

Module 3:

Tests of Hypotheses, Most Powerful Tests, UMP Tests and Similar Tests;Tests of hypotheses and most powerful Tests -- Simple versus simple hypothesis testing problem - Error probabilities, p-value and choice of level of significance – Most powerful tests – Neyman Pearson Lemma, generalized Neyman - Pearson lemma, MLR property, One-sided UMP tests, two sided UMP

tests and UMP unbiased tests, α –similar tests and similar tests with Neyman structure. Principle of invariance in testing of hypotheses, locally most powerful tests. Likelihood ratio tests, asymptotic distribution of likelihood ratio.

Module 4:

Non-parametric Tests and Sequential Tests: Single sample tests - testing goodness of fit, chi-square tests- Kolmogorov - Smirnov test - sign test - Wilcoxon signed rank test. Two sample tests - the chi-square test for homogeneity - Kolmogorov Smirnov test; the median test - Mann-Whitney - Wilcoxon test - Test for independence, Kendall's tau, Spearman's rank correlation coefficient, some fundamental ideas of sequential sampling - Sequential Probability Ratio Test (SPRT) - important properties, termination of SPRT - Operating Characteristic (OC) function and AverageSample Number (ASN) of SPRT - Developing SPRT for different problems.

REFERENCES

1. Kale,B.K.(2005). A First Course in Parametric Inference, Second Edition, Narosa Publishing House, New Delhi.

2.Vijay K. Rohatgi, A. K. Md. Ehsanes Saleh (2015). An Introduction to Probability and Statistics,3rd Edition, John Wiley and Sons, New York.

3. Casella, G. and Berger, R.L. (2002). Statistical Inference, Second Edition, Duxbury, Australia.

4. Lehmann, E.L (1983). Theory of Point Estimation, John Wiley and Sons, New York.

5. Rohatgi, V.K (2003). Statistical Inference, Dover Publications.

6. Rao, C.R (2002). Linear Statistical Inference and Its Applications, Second Edi- tion, John Wiley and Sons, NewYork.

7. Fraser, D.A. S. (1957): Non - parametric Methods in Statistics, Wiley, New York.

8. Ferguson, T.S. (1967): Mathematical Statistics: A Decision - Theoretic Approach. Academic Press,New York.

9. Srivastava, M. and Srivstava, N. (2009): Statistical Inference: Testing of Hypothesis, Eastern Economy Edition, PHI Learning Pvt. Ltd., New Delhi.

MST2C07: Sampling and Design of Experiments

Course Outcome:

After successful completion of this course, student will be able to:

1 .know different census and sample survey methods

2. Plan and implement sample surveys, consumer satisfaction surveys, public opinion surveys etc.

3. Aware of different designs in experimentation like CRD, RBD, LSD, BIBD, Factorial Designs, etc. 4. Apply ANOVA technique to analyse the data using Python or R.

Module 1:

Census and sampling methods, probability sampling and non-probability sampling, principal steps in sample surveys, sampling errors and non-sampling errors, bias, variance and mean square error of an estimator, simple random sampling with and without replacement- estimation of the population mean,totaland proportions, properties of the estimators, variance and standard error of the estimators, confidence intervals, determination of the sample size. Stratified random sampling- estimation of the population means, total and proportion, properties of estimators, various methods of allocation of a sample, comparison of the precisions of estimators under proportional allocation, optimum allocation and SRS. Systematic sampling.

Module 2:

Ratio method of estimation, estimation of population ratio, mean and total, Bias and relative bias of ratio estimator, comparison with SRS estimation. Regression method of estimation. Comparison of ratio and regression estimators with mean per unit method, Cluster sampling, single stage cluster sampling with equal and unequal cluster sizes, estimation of the population mean and its standard error. Multistage and Multiphase sampling (Basic Concepts),

Module 3:

Standard Gauss Markoff set up, estimability of parameters, method of least squares, best linear unbiased Estimators, Gauss – Mark off Theorem, tests of linear hypotheses. Planning ofexperiments, Basic principles of experimental design, uniformity trails, analysis of variance, oneway, two-way and three-way classification models, completely randomized design (CRD),randomized block design (RBD) Latin square design (LSD). Analysis of covariance (ANCOVA)

Module 4:

Balanced incomplete block design (BIBD); incidence Matrix, parametric relation; intrablock analysis of BIBD, basic ideas of partially balanced incomplete block design (PBIBD). Factorial experiments, 2n and 3n factorial experiments, Yates procedure, confounding in factorial experiments, basic ideas of response surface designs.

REFERENCES

1. Cochran W. G. (1999) Sampling Techniques, 3rd edition, John Wiley and Sons.
2. Mukhopadyay P. (2009) Theory and Methods of Survey Sampling, 2nd edition, PHL, New Delhi.
3. Aloke Dey (1986) Theory of Block Designs, Wiley Eastern, New Delhi.
4. Das M.N. and Giri N.C. (1994) Design and analysis of experiments, Wiley Eastern Ltd.
5. Arnab, R. (2017). Survey Sampling: Theory and Applications. Academic Press.
6. Montgomery, C.D. (2012) Design and Analysis of Experiments, John Wiley, New York.
7. Sampath S. C. (2001) Sampling Theory and Methods, Alpha Science International Ltd.,
8. Thomas Lumley (1996) Complex Surveys. A Guide to Analysis Using R, Wiley eastern Ltd.
9. Des Raj (1967) Sampling Theory. Tata McGraw Hill ,NewDelhi
10. Dean, A. and Voss, D. (1999) Design and Analysis of Experiments, Springer Texts in Statistics

MST2C08: Statistics using Python programming

Course Outcome:

After successful completion of this course, student will be able to:

1. Understand the basics of Python programming.

2. Analyze various object oriented concepts.

3. Apply Python tools for statistical analysis.

4. Demonstrate the use of graphical representations for data analytics.

Module 1:

Installing Python - basic syntax - interactive shell - editing, saving and running a script. The concept of data types - variables- assignments - mutable type - immutable types - arithmetic operators and expressions - comments in the program - understanding error messages - Control statements - operators.

Module 2:

Introduction to functions - inbuilt and user defined functions - functions with arguments and return values - formal vs actual arguments - named arguments - Recursive functions - Lambda function - OOP Concepts - classes - objects - attributes and methods - defining classes - inheritance - polymorphism

Module 3:

Introduction to Pandas - Pandas data series - Pandas data frames - data handling - grouping - Descriptive statistical analysis and Graphical representation.

Hypothesis testing - data modelling - linear regression models - logistic regression model.

Module 4:

Line graph - Bar chart - Pie chart - Heat map - Histogram - Density plot - Cumulative frequencies - Error bars - Scatter plot - 3D plot.

REFERENCES

1. Lambert, K. A. (2018). Fundamentals of Python: first programs. Cengage Learning.

2. Haslwanter, T. (2016). An Introduction to Statistics with Python. Springer International Publishing.

MST2P02: Statistical Computing -II (Lab based on Python Programming)

Course outcomes:

After successful completion of this course, student will be able to:

1. Know the basics of Python Programming Language.

2. Familiarize OOPS concepts using Python.

3. Acquaint statistical analysis using Python

4. Learn graphical representation of data for analysis using Python.

Module 1:Practical Assignments:

1. Lab exercise on data types
2. Lab exercise on arithmetic operators and expressions
3. Lab exercise on Control statements

Module 2:Practical Assignments:

4. Lab exercise on inbuilt and user-defined functions
5. Lab exercise on Recursive and Lambda function
6. Lab exercise on OOP Concepts.

Module 3: Practical Assignments:

7. Lab exercise on Pandas data series, frame, handling and grouping
8. Lab exercise on statistical analysis
9. Lab exercise on Hypothesis testing
10. Lab exercise on regression modelling

Module 4: Practical Assignments:

11. Lab exercise on graphical and diagrammatic representation.
12. Lab exercise on the density plot
13. Lab exercise on scatter and 3D plot.

**

SEMESTER III

MST3C09: Regression Analysis

Course Outcome:

After successful completion of this course, student will be able to:

1. Understand simple linear regression

2. Understand multiple regression, residual analysis for fitting a suitable model to a given data and to check the suitability.

3. Study necessary transformations and modifications to be made when model assumptions are violated.

4. Fit logistic and Poisson, non-linear and polynomial models.

Module 1:

Introduction to regression analysis: overview and applications of regression analysis, major steps in regression analysis. Simple linear regression (Two variables): assumptions, estimation and properties of regression coefficients, significance and confidence intervals of regression coefficients, measuring the quality of the fit. Effect of outliers.

Module 2:

Multiple linear regression model: assumptions, ordinary least square estimation of regression coefficients, interpretation and properties of regression coefficient, significance and confidence intervals of regression coefficients. Mean Square error criteria, coefficient of determination, Residual analysis, various types of residuals, Departures from underlying assumptions, Departures from normality. Multi-collinearity, sources, effects, diagnostics.

Module 3:

Need for transformation of variables; Box-Cox transformation, removal of heteroscedasticity and serial correlation, Leverage, and influence (concept only). Generalized least squares and weighted least squares (without derivation). Polynomial regression models, subset regression, Forward, Backward and Stepwise procedures, indicator variables, stepwise regression.

Module 4:

Introduction to nonlinear regression, linearity transformations, Least squares in the nonlinear case and estimation of parameters, Logistic regression, estimation and interpretation of parameters in a logistic regression model, Poisson regression, Generalized linear models (GLM), Prediction and estimation with the GLM, residual analysis in the GLM.

REFERENCES

1. D. C Montgomery, E.A Peck and G.G Vining (2003). Introduction to Linear Regression Analysis, John Wiley and Sons, Inc.NY.

2. S. Chatterjee and A. Hadi (2013) Regression Analysis by Example, 5th Ed., John Wiley and Sons.

3. Seber, A.F. and Lee, A.J. (2003) Linear Regression Analysis, John Wiley.

4. Lain Pardoe (2012) Applied Regression Modelling, John Wiley and Sons, Inc,.

5. P. McCullagh, J.A. Nelder,(1989) Generalized Linear Models, Chapman & Hall,. John O.Rawlings.

6. Sastry G. Pantula, David A. Dickey (1998) Applied Regression Analysis, Second Edition, Springer.

7. Draper, N. and Smith, H. (2012) Applied Regression Analysis – John Wiley & Sons

MST3C10: Stochastic Processes and Time Series Analysis

Course Outcome:

After successful completion of this course, student will be able to:

1. Know various stochastic models.

2. UnderstandMarkov chain and its properties

3 .Understand the concept of Poisson process and important queuing models.

of time series data.

4. Understand Time series models and able to predict future values to make appropriate planning and decision making.

5. Understand autoregressive /moving average models.

Module 1:

Introduction to stochastic processes: - classification of stochastic processes, wide sense and strict sense stationary processes, processes with stationary independent increments, Markov process, Markov chains- transition probability matrices, Chapman-Kolmogorov equation, first passage probabilities, recurrent and transient states, mean recurrence time, stationary distributions, limiting probabilities, Random walk.

Module 2:

Continuous time Markov chains, Poisson processes, properties, inter-arrival time distribution pure birth processes and the Yule processes, birth and death processes linear growth process with immigration, steady-state solutions of Markovian queues - M/M/1, M/M/s, M/MM/∞ models, Renewal processes – concepts only, examples, Poisson process viewed as a renewal process.

Module 3:

Time series data, examples, Time series as stochastic process, Additive and multiplicative models, stationary time series- covariance stationarity, Modelling Time Series Data, Exponential Smoothing Methods - First-Order Exponential Smoothing, Second Order Exponential Smoothing, Forecasting, Exponential Smoothing for Seasonal Data, Exponential Smoothers.

Module 4:

Time series modelling, Autocorrelation function (ACF), partial auto correlation function (PACF), correlogram, AR, MA, ARMA, ARIMA Models, Yule- Walker equations, Forecasting future values.

REFERENCES

1. Medhi J. (2017) Stochastic Processes, Second Edition, Wiley Eastern, New Delhi

2. Ross S.M. (2007) Stochastic Processes. Second Edition, Wiley Eastern, New Delhi

3. Montgomery D. C., Cheryl L. J., and Murat K. (2015) Introduction to Time Series Analysis and Forecasting. John Wiley & Sons.

4. Brockwell P.J and Davis R.A. (2002) Introduction to Time Series and Forecasting Second edition, Springer-Verlag.

5. Abraham, B., &Ledolter, J. (2009). Statistical methods for forecasting (Vol. 234). John Wiley & Sons.

6. Chatfield, C.(2004).The Analysis of Time Series - An Introduction (Sixth edition), Chapman and Hall.

MST3C11: Big Data Analytics

Course Outcome:

After successful completion of this course, student will be able to:

1. Understand the basic concepts of Data Analytics.

2. Analyze various storage techniques for data.

3. Know various methods for representing data.

4. Understand various operations on stored data.

Module 1:

Concepts of Data Analytics: Descriptive, Diagnostic, Predictive, Prescriptive analytics -Big Data characteristics: Volume, Velocity, Variety, Veracity of data - Types of data: Structured, Unstructured, Semi-Structured, Metadata - Big data sources: Human-Human communication, Human-Machine Communication, Machine-Machine Communication - Data Ownership - Data Privacy.

Module 2:

Standard Big data architecture - Big data application - Hadoop framework - HDFS Design goal - Master-Slave architecture - Block System - Read-write Process for data - Installing HDFS - Executing in HDFS: Reading and writing Local files and Data streams into HDFS - Types of files in HDFS - Strengths and alternatives of HDFS - Concept of YARN.

Module 3:

Stream processing Models and Tools - Apache Spark - Spark Architecture: Resilient Distributed Datasets, Directed Acyclic Graph - Spark Ecosystem - Spark for Big Data Processing: MLlib, Spark GraphX, SparkR, SparkSQL, Spark Streaming - Spark versus Hadoop

Module 4:

Hive Architecture - Components - Data Definition - Partitioning - Data Manipulation - Joins, Views and Indexes - Hive Execution - Pig Architecture - Pig Latin Data Model - Latin Operators - Loading Data - Diagnostic Operators - Group Operators - Pig Joins - Row Level Operators - Pig Built-in function - User-defined functions - Pig Scripts.

REFERENCES

1.Anil Maheshwari (2020). Big Data. 2nd Edition. McGraw Hill Education Pvt Ltd.

Essential Reading / Recommended Reading

2.Thomas Erl, Wajid Khattak and Paul Buhler (2016). Big Data Fundamentals: Concepts, Drivers and Techniques. Service Tech Press.

3.Julián Luengo, Diego García-Gil, Sergio Ramírez-Gallego, Salvador García, Francisco Herrera (2020). Big Data Preprocessing: Enabling Smart Data. Springer Nature Publishing.

4.Seema Acharya, SubhasiniChellappan (2019), Big Data and Analytics. 2nd Edition, Wiley India Pvt Ltd.

MST3P03: Statistical Computing III (Lab based on R and Python)

Course Outcomes:

After successful completion of this course, student will be able to:

1. Get practical knowledge of different theoretical methods using real data.

2. Improve the basic concepts of statistical theories using real world data.

Practical based on data with respect to module 3 and module 4 of 1st semester paper-Programming Using R.

**

SEMESTER IV

MST4C12: Multivariate Analysis

Course Outcome:

After successful completion of this course, student will be able to:

1. Understand basic concepts on multivariate analysis.

2. Apply multivariate techniques such as discriminant function and classification rules, principal components, canonical correlations, factor analysis, MANOVA etc.

3. Apply Hotelling's T2 and Mahalanobis D2 etc for testing hypotheses in the case of multivariate data.

Module 1:

Basic concepts on multivariate variable. Concept of random vector: Its expectation and dispersion (Variance-Covariance) matrix. Marginal and joint distributions. Conditional distributions and Independence of random vectors. Multivariate normal distribution, Marginal and conditional distribution. additive property, MLEs of mean and dispersion matrix.

Module 2:

Sample mean vector and its distribution, Hotelling's T2 and Mahalanobis' D2 statistics and applications. Tests of hypotheses about the mean vectors and covariance matrices for multivariate normal populations. Wishart distribution, Independence of sub vectors and sphericity test.

Module 3:

Fisher's criteria for discrimination and classification for two populations, Sample discriminant function. Expected cost of misclassification, classification with two multivariate normal population, classification with several multivariate normal populations. Multivariate analysis of variance (MANOVA) of one and two- way classified data. Multivariate analysis of covariance(concept only).

Module 4:

Principal components, sample principal components asymptotic properties. Canonical variables and canonical correlations: definition and estimation. Factor analysis: Orthogonal factor model, factor loadings, estimation of factor loadings, factor scores, cluster analysis-agglomerative and divisive techniques.

REFERENCES

1. Chatfield, C. (2018). Introduction to multivariate analysis. Routledge.

2. Rencher, A. C. (2012) Methods of Multivariate Analysis.(3rd ed.) John Wiley.

3. Johnson R.A. and Wichern D.W. (2008) Applied Multivariate Statistical Analysis. 6th Edition, Pearson Education.

4. Anderson, T.W. (2009). An Introduction to Multivariate Statistical Analysis, 3rd Edition, John Wiley.

5. Everitt B, HothornT,(2011). An Introduction to Applied Multivariate Analysis with R, Springer.

6. Barry J. Babin, Hair, Rolph E Anderson, & William C. Blac, (2013), Multivariate Data Analysis, Pearson New International Edition,

ELECTIVES

Group I (for Third Semester)

MST3E01: Survival Analysis

Course Outcome:

After successful completion of this course, student will be able to:

1. Understand lifetime models and life time characteristics.

2. Estimateparameters of life time characteristics.

3. Test parametric and non-parametrictests of life time characteristics.

4. Understand Parametric regression models of lifetime

5. Understand to apply of Bayesian Inference.

Module 1:

Life time models and life time characteristics: continuous and discrete models, mixture models ,hazard functions, survival function and mean residual life function. Censuring and truncation of distributions. Different censoring schemes, Ageing classes –IFR, IFRA, NBU, NBUE, HNBUE and their duals, Bathtub Failure rate.

Module 2:

Estimation and Testing of the life time characteristics parametric and nonparametric methods. Life tables, Kaplan – Meier Estimator, Estimation under the assumption of IFR/DFR.Tests of exponentiality against non-parametric classes-Total time on test, Deshpande test, two sample problem-Gehan Test, Log rank test,Mantel-Haenszel Test,Tarone-Ware Tests.

Module 3:

Parametric regression models of lifetime. Semi-parametric regression for failure rate- Cox's proportional hazards model with one and several covariates. Survival analysis and competing risk, multy variable survival models.

Module 4:

Bayesian Inference: randomized and non-randomized decision rules, risk and loss function, optimality and decision rules. Estimation, testing hypothesis, confidence interval and prediction under Bayesian approach.

REFERENCES

1. Alaen O O, Borgen O, Gjessing H K(2008):Survival& Inventory analysis,Springer.

2. Balakrishnan N, C R Rao(2001).Hand Book of Statistics Vol 20,Advances in Reliability, North Holland.

3.Bensal A K (2008): Bayesian Parametric Inference, New Age International.

4.Cox,D.R. and Oakes,D.(1984):Analysis of Survival Data,Chapman and Hall ,New York.

5.Elandt – Johnson, R.E. Johnson N.L(1980): Survival models and Data Analysis, John Wiley and sons

6.Ghosh J K,DelampadyM,Samanta T(2006):An Introduction to Bayesian Analysis,Springer.

7.Gross A J and Clark ,V.A.(1975):Survival Distributions:Reliability Applications in the Biomedical Sciences,John Wiley and Sons.

8. Hosmer D .W,Lemeshow S, May S(2008):Applied Survival Analysis ,Wiely.

9. LawkessJ.F.(2003):Statistical Models and Methods for Life Time , J Wiley

10.Sinha S K(1986) reliability and Life testing,Wiley.

MST3E02: Queueing Theory

Course Outcome:

After successful completion of this course, student will be able to:

1. Understand various Markovian queueing models and their analysis.

2. Understand transient behavior of queueing models and analysis of advanced Markovian models with bulk arrival and bulk service.

3. Understand various queueing networks and their extensions.

4. Understand various non Markovian queueing models and their analysis.

Module 1:

Markovian Queueing Models: Introduction to queueing theory, Characteristics of queueing processes, Measures of effectiveness, Markovian queueing models, steady state solutions of the

M/M/I model, waiting time distributions, Little's formula, queues with unlimited service, finite source queues.

Module 2:

Advanced Markovian Models: Transient behavior of M/M/1 queues, transient behavior of M/M/1.Busy period analysis for M/M/1 and M/M/c models. Advanced Markovian models. Bulk input M[X] /M/1 model, Bulk service M/M[Y]/1 model, Erlangian models, M/Ek/1and Ek/M/1. A brief discussion of priority queues.

Module 3:

Queueing Networks: Series queues, open Jackson networks, closed Jackson network, Cyclic queues, Extension of Jackson networks. Non-Jackson networks.

Module 4:

Non-Markovian Queueing Models: Models with general service pattern, The M/G/1 queueing model, The Pollaczek-Khintchine formula, Departure point steady state systems size probabilities, ergodic theory, Special cases M/Ek/1 and M/D/1, waiting times, busy period analysis, general input and exponential service models, arrival point steady state system size probabilities.

REFERENCES

1. Gross, D. and Harris, C.M. (1985): Fundamentals of Queueing Theory, 2nd Edition, John Wiley and Sons, New York.

2. Ross, S.M (2010). Introduction to Probability Models. 10th Edition. Academic Press, New York.

3. Bose, S.K. (2002). An Introduction to Queueing Systems, Kluwer Academic / Plenum Publishers, New York.

MST3E03: Reliability Modeling

Course Outcome:

After successful completion of this course, student will be able to:

1. Understand reliability concepts and measures.

2. Understand various lifetime Probability distributions and their structural properties.

3. Understand univariate and bivariate shock models and reliability estimation based on failure times.

4. Understand Maintenance and Replacement Policies.

Module 1:

Basic Reliability Concepts: Reliability concepts and measures; components and systems; coherent systems; reliability of coherent systems; cuts and paths; modular decomposition; bounds on reliability; structural and reliability importance of components.

Module 2:

Life Distributions and Properties: Life distributions; reliability function; hazard rate; common life distributions-exponential, Weibull, Gamma etc. Estimation of parameters and tests in these models. Notions of ageing; IFR, IFRA, NBU, DMRL, and NBUE Classes and their duals; closures or these classes under formation of coherent systems, convolutions and mixtures.

Module 3:

Shock Models: Univariate shock models and life distributions arising out of them; bi-variate shock models; common bivariate exponential distributions and their properties. Reliability estimation based on failure times in variously censored life tests and in tests with replacement of failed items; stress-strength reliability and its estimation.

Module 4:

Maintenance and Replacement Policies: Repairable systems, replacement policies, modeling of arepairable system by a non-homogeneous Poisson process. Reliability growth models; probability plottingtechniques; Hollander-Proschan and Deshpande tests for exponentiality; tests for HPP vs. NHPP with repairablesystems. Basic ideas of accelerated life testing.

REFERENCES

1. Barlow R.E. and Proschan F. (1985). Statistical Theory of Reliability and Life Testing; Holt,Rinehart and Winston.

2. Bain L.J. and Engelhardt (1991). Statistical Analysis of Reliability and Life Testing Models; Marcel Dekker.

3. Aven, T. and Jensen, U. (1999). Stochastic Models in Reliability, Springer Verlag, New York, Inc.

4. Nelson, W (1982). Applied Life Data Analysis; John Wiley.

5. Zacks, S. (1992). Introduction to Reliability Analysis: Probability Models and Statistics Methods. New York: Springer-Verlag

MST3E04: Data Mining

Course Outcome:

After successful completion of this course, student will be able to:

1. Know various data mining concepts.
2. Analyze statistical techniques for data mining.
3. Understand various data classification techniques.
4. Apply various techniques for cluster analysis.

Module 1

Introduction Data Warehousing, Multidimensional Data Model, OLAP Operations, Introduction to KDD process, Data mining, Data mining -On What kinds of Data, Data mining Functionalities, Classification of Data Mining Systems. Data Preprocessing Data Cleaning, Data Integration and Transformation, Data Reduction, Data discretization and concept hierarchy generation.

Module 2:

Exploring Data and Visualization Techniques General Concepts, Techniques, Visualizing Higher Dimensional Data, Tools Association Analysis Basic Concepts, Efficient and Scalable Frequent Item set Mining Methods:Apriori Algorithm, generating association Rules from Frequent Item sets, Improving the Efficiency of Apriori. Mining Frequent item-sets without Candidate Generation, Evaluation of Association Patterns, Visualization.

Module 3:

Classification Introduction to Classification and Prediction, Classification by Decision Tree Induction: Decision Tree induction, Attribute Selection Measures, Tree Pruning, Bayesian Classification: Bayes' theorem, Naïve Bayesian Classification, Rule Based Algorithms: Using If - Then rules of Classification, Rule Extraction from a Decision Tree, Rule Induction Using a Sequential Covering algorithm, K- Nearest Neighbour Classifiers, Support Vector Machine. Evaluating the performance of a Classifier, Methods for comparing classifiers, Visualization.

Module 4:

Prediction Linear Regression, Nonlinear Regression, Other Regression-Based Methods Cluster Analysis I: Basic Concepts and Algorithms Cluster Analysis, Requirements of Cluster Analysis' Types of Data in Cluster Analysis, Categorization of Major Clustering Methods, Partitioning Methods: k-Means and k- Medoids, From K-Medoids to CLARANS. Cluster Analysis: Hierarchical Method: Agglomerative and Divisive Hierarchical Clustering. Density-based Clustering - DBSCAN, Grid based clustering-STING Evaluation of Clustering Method.

REFERENCES

1. Pang-Ning Tan, Michael Steinbach, Vipin Kumar, 'Introduction to Data Mining'

2.Data Mining Concepts and Techniques – Jiawei Han and Micheline Kamber, Second Edition, Elsevier, 2006

3. G. K. Gupta, "Introduction to Data Mining with Case Studies", Easter Economy Edition, Prentice Hall of India, 2006.

4.Making sense of Data: A practical guide to exploratory Data Analysis and Data Mining-Glenn J Myatt.

Group II (for Fourth Semester)

MST4E01: Bio-Statistics

Course Outcome:

After successful completion of this course, student will be able to:

1. Understand survival distributions and their applications.

2. Estimate survival functions using non parametric methods.

3. Estimate probabilities of death under competing risks by maximum likelihood and modified minimum chi-square methods.

4. Understand basic biological concepts in genetics.

Module 1

Functions of survival time, survival distributions and their applications viz.exponential, gamma, weibull, Rayleigh, lognormal, death density function for a distributionhaving bath-tub shape hazard function. Tests of goodness of fit for survival distributions (WEtest for exponential distribution, W-test for lognormal distribution, chi-square test foruncensored observations). Parametric methods for comparing two survival distributions viz.L.R. test, Cox's F-test.

Module 2:

Type I, Type II and progressive or random censoring with biological examples,estimation of mean survival time and variance of the estimator for type I and type II censoreddata with numerical examples. Non-parametric methods for estimating survival function andvariance of the estimator viz. Actuarial and Kaplan-Meier methods.

Module 3

Competing risk theory, indices for measurement of probability of death undercompeting risks and their inter-relations. Estimation of probabilities of death undercompeting risks by maximum likelihood and modified minimum chi-square methods. Theoryof independent and dependent risks. Bivariate normal dependent risk model. Conditionaldeath density functions.

Module 4

Basic biological concepts in genetics, Mendel's law, Hardy-Weinberg equilibrium,random mating, distribution of allele frequency(dominant/co-dominant cases), Approach toequilibriumfor X-linked genes, natural selection, mutation, genetic drift, equilibrium whenboth natural selection and mutation are operative, detection and estimation of linkage inheredity.

REFERENCES

1. Biswas, S. (1995). Applied Stochastic Processes. A Biostatistical and Population Oriented Approach, Wiley Eastern Ltd.
2. Cox, D. R. and Oakes, D. (1984). Analysis of Survival Data, Chapman and Hall.
3. Elandt, R. C. and Johnson(1975). Probability Models and Statistical Methods in Genetics, John Wiley & Sons.
4. Ewens, W. J.(1979). Mathematics of Population Genetics, Springer Verlag.
5. Ewens, W. J. and Grant, G. R.(2001). Statistical Methods in Bio informatics: AnIntroduction, Springer.
6. Friedman, L. M., Furburg, C. and DeMets, D. L.(1998). Fundamentals of Clinical Trials,Springer Verlag.
7. Gross, A. J. and Clark, V. A.(1975). Survival Distribution; Reliability Applications inBiomedical Sciences, John Wiley & Sons.
8. Lee, Elisa, T.(1992). Statistical Methods for Survival Data Analysis, John Wiley & Sons.
9. Li, C. C.(1976). First Course of Population Genetics, Boxwood Press.
10. Miller, R. G.(1981). Survival Analysis, John Wiley & Sons.

MST4E02: Analysis of Clinical Trials

Course Outcome:

After successful completion of this course, student will be able to:

1. Understand Basics of Clinical Trails
2. Understand design of clinical trials
3. Determine Sample size in clinical trials
4. Understand the concept of meta-analysis in clinical trials.

Module 1:

Basics of Clinical Trails: Introduction to clinical trials, the need and ethics of clinical trials, bias and random error in clinical studies, Protocols, conduct of clinical trials over view of Phase I-IV trials, Data management-data definitions, standard operating procedure informed consent form, case report forms, database design, data collection systems for good clinical practice.

Module 2:

Design of Clinical Trials: Design of clinical trials- Different phases, Comparative and controlled trials, Random allocation, Randomization, response adaptive methods and restricted randomization. Methods of Blinding, Parallel group designs, Crossover designs, Symmetric designs, Adaptive designs, Group sequential designs, Zelen's designs, design of bioequivalence trials. Outcome measures.

Module 3:

Sample Size Determination and Testing: Sample size determination in one and two sample cases, comparative trials, activity studies, testing and other purposes, unequal sample sizes and case of anova. Surrogate endpoints-selection and design of trials with surrogate endpoints, analysis of surrogate end point data. Reporting and Analysis-Interpretation of result, multi-center trials.

Module 4:

Meta-Analysis: Meta-analysis in clinical trials-concept and goals, fixed and random effect approaches. Bioassay: Direct and indirect assays, Quantal and quantitative assays, Parallel line and slope ratio assays, Design of bioassays.

REFERENCES

1. Friedman, L. M., Furburg, C. D. Demets, L. (1998): Fundamentals of Clinical Trials, Springer Verlag.

2. Jennison and B. W. Turnbull (1999): Group Sequential Methods with Applications to Clinical Trails, CRC Press.

3. Kulinskaya E, Morgeathaler S, Staudte R G (2008), Meta-analysis, Wiley.

4. Das, M. N. and Giri (2008). Design of Experiments, New Age, India

5. Fleiss, J. L. (1989): The Design and Analysis of Clinical Experiments, Wiley.

6. Marubeni, E. and M. G. Valsecchi (1994): Analyzing Survival Data from Clinical Trials and Observational Studies, Wiley and Sons.

7. Piantadosi S. (1997): Clinical Trials: A Methodological Perspective. Wiley.

8 . W Rosenberger, J MLachin (2002): Randomization in Clinical Trials Theory and Practice, Wiley

MST4E03: Demography

Course Outcome:

After successful completion of this course, student will be able to:

1. Understand various aspects related to population census of India.
2. Understand different measures of Fertility.
3. Understand different measures of Mortality.
4. Understand the method of population projection.

Module-1:

Scope and content of population census of India. Population, Composition, Dependency ratio. Brief Coverage and content errors in demographic data. Adjustment of age data – use of Whipple, Myer and UN indices. Chandrasekhar – Deming formula to check completeness of registration data.

Module-2:

Measures of fertility: Stochastic models for reproduction, (Dandekar`s Modified Binomial and Poisson distributions, William Brass Model), distributions of time to first birth, inter-live birth intervals and number of births.

Module-3:

Measures of Mortality: Construction of abridged life tables (l_x–linear, exponential, Reed and Merrell`s, Grevill's) Relations between functions of Life Tables. Distributions of life table functions.

Module-4:

Stable and quasi-stable populations, intrinsic growth rate. Methods for population projection. Use of Leslie matrix.

Module-5:

Models for population growth and their fitting to population data. Linear, Exponential, logarithmic, modified logarithmic, Gompertz and Logistic Curves. Stochastic models for population growth (Pure Birth Model, Simple Birth & Death Model, B rth, death and migration model).

References:

1.Sudhendra Biswas (1995): Applied Stochastic Processes, New Age International Publishers Ltd.
2.Pathak, K.B. & Ram, F. (1998): Techniques of Demographic Analysis, Himalays Publishers
3.K.Srinivisan (1998): Basic Demographic Techniques and Applications Sage publications.
4.Asha A Bhande, Tara Kanitkar (2004): Principles of Population Studies; Himalayas publishing House.
5.Saxena H.C and Surrendran P.U: Statistical Inference.
6.Bartholomew, D.J. (1982): Stochastic Models for Social Processes, John Wiley.
7.Benjamin, B. (1969): Demographic Analysis, Ge0rge. Allen and Unwin.
8.Chain, C.L (1968): Introduction to Stochastic Processes in Biostatistics; John Wiley.
9.Cox, P.R. (1970): Applied Mathematical Demography, Springer Verlag.
10.Spiegelman, M. (1969): Introduction to Demographic Analysis; Harvard University Press.

MST4E04: Machine Learning

Course Outcome:

After successful completion of this course, student will be able to:

1. Understand machine learning techniques.

2. Apply probability techniques for machine learning.

3. Apply the concept of neural networks.

4. Understand techniques for cluster analysis.

Module 1:

Machine Learning-Examples of Machine Applications – Learning Associations – Classification Regression- Unsupervised Learning- Reinforcement Learning. Supervised Learning: Learning class from examples- Probably Approximately Correct (PAC) Learning- Noise-Learning Multiple classes. Regression – Model Selection and Generalization. Introduction to Parametric

methods- Maximum Likelihood Estimation: Bernoulli Density Multinomial Density-Gaussian Density, Nonparametric Density Estimation: Histogram EstimatorKernel Estimator-K-Nearest Neighbor Estimator.

Module 2:

Dimensionality Reduction: Introduction- Subset Selection- Principal Component Analysis, Feature Embedding-Factor Analysis- Singular Value Decomposition- Multidimensional Scaling-Linear Discriminant Analysis- Bayesian Decision Theory. Linear Discrimination: Introduction-Generalizing the Linear Model- Geometry of the Linear Discriminant- Pairwise Separation-Gradient Descent Logistic Discrimination. Optical separating hyper plane – v-SVM, kernel tricks – vectorian kernel- - defining kernel- multiclass kernel machines- one-class kernel machines.

Module 3:

Multilayer perceptron, Introduction, training a perceptron- learning Boolean functions-multilayer perceptron- back propagation algorithm- training procedures. Combining Multiple Learners, Rationale-Generating diverse learners- Model combination schemes- voting, Bagging-Boosting- fine tuning an Ensemble.

Module 4:

Cluster Analysis, Introduction-Mixture Densities, K-Means Clustering- Expectation-Maximization algorithm- Mixtures of Latent Variable Models-Supervised Learning after Clustering-Spectral Clustering- Hierarchical Clustering- Divisive Clustering- Choosing the number of Clusters.

REFERENCES

1. E. Alpaydin (2014) Introduction to Machine Learning, 3rd Edition, MIT Press.

2. Frank Kane (2012) Data Science and Machine Learning. Manning Publications.

3. C.M.Bishop, Pattern Recognition and Machine Learning, Springer.

4. T. Hastie, R. Tibshirani and J. Friedman (2016) The Elements of Statistical Learning: Data Mining, Inference and Prediction, Springer, 2nd Edition,2009.

5. Data Mining Techniques: A.K. Pujari, Universities Press, 2001

OPEN ELECTIVES:

MST4OE01: Neural Networks and Deep Learning

Course Outcome:

After successful completion of this course, student will be able to:

1. Understand concepts of artificial neural networks.

2. Apply neural network techniques.

3. Apply statistical techniques in neural networks.

4. Familiarize with training of neural networks.

Module 1:

Fundamental concepts of Artificial Neural Networks (ANN) - Biological neural networks - Comparison between biological neuron and artificial neuron - Evolution of neural networks - Scope and limitations of ANN - Basic models of ANN - Learning methods - Activation functions - Important terminologies of ANN: Weights - Bias - Threshold - Learning Rate - Momentum factor - Vigilance parameters.

Module 2:

Concept of supervised learning algorithms - Perceptron networks - Adaptive linear neuron (Adaline) - Multiple adaptive linear neuron - Back-Propagation network: Learning factors - Initial weights - Learning rate α - Momentum factor - Generalization - Training and testing of the data.

Module 3:

Concept of unsupervised learning algorithms - Fixed weight competitive net: Maxnet - Mexican Hat net - Hamming networks - Kohonen self-organizing feature maps - Learning vector quantization.

Module 4:

Introduction - Components of Convolution Neural Networks (CNN) architecture: Padding - Strides - Rectified linear unit layer - Exponential linear unit - Pooling - Fully connected layers -

Local response normalization - Hierarchical feature engineering - Training CNN using Backpropagation through convolutions - Case studies: AlexNet - GoogLeNet.

Module 5:

Stateless algorithms: Naive algorithms - Upper bounding methods - Simple reinforcement learning for Tic-Tac-Toe - Straw-Man algorithms - Bootstrapping for value function learning - One policy versus off policy methods: SARSA - Policy gradient methods: Finite difference method - Likelihood ratio method - Monte Carlo tree search.

REFERENCES

1.Charu C. Aggarwal (2018) Neural Networks and Deep Learning A Textbook, Springer International Publishing, Switzerland.

2.S.N Sivanandam, S.N Deepa (2018). Principles of soft computing. Wiley India.

3.S Lovelyn Rose, L Ashok Kumar, Karthika Renuka (2019). Deep Learning using Python. Wiley India.

4.Francois Chollet (2017). Deep Learning with Python. Manning Publishing.

5.Andreas C. Muller & Sarah Guido (2017). Introduction to Machine Learning with Python. O'Reilly Media, Inc.

MST4OE02: Statistics with SAS

Course Outcome:

After successful completion of this course, student will be able to:

1. Understand basic concepts in data analysis using SAS

2. Do programming using PROC MEANS,PROC FREQ,PROC PRINT etc

3. Understand various tests and get the knowledge on how to write,I nterpret and summarizing results.

4. Perform ANOVA using SAS.

Module 1

Basic Concepts in Data Analysis – variables, values, quantitative variables versus classification variables, observational units, scale of measurements, basic approach for research, descriptive

versus inferential statistical analysis, hypothesis testing Introduction to SAS programs – What is SAS?, three types of SAS files; Data input –Inputting questionnaire data versus other types of data, inputting data using the DATALINES statement, inputting a correlation or covariance matrix. Working with Variables and Observations in SAS study – Manipulating, subsetting, concatenating and merging data.

Module 2

Simple Descriptive Data Analysis – Introduction, PROC MEANS, creating frequency table with PROC FREQ, PROC PRINT, PROC UNIVARIATE, Test for normality, Stem-and-Leaf plot, skewness. Analysis of Bivariate Data – Significance tests versus measures of association, levels of measurement, Appropriate statistics, scattergrams with PROC GPLOT, Pearson correlation with PROC CORR, options used with PROC CORR, Spearman correlations with PROC CORR, Chi-square test of independence, Two way classification table, tabular versus raw data, assumptions underlying Pearson correlation coefficient, Spearman correlation coefficient and chi square test of independence

Module 3:

t-tests – two types of t-test, independent samples t test, independent variable and dependent variable, writing the SAS program, interpreting and summarizing the results. The paired samples t test, paired versus independent samples, problems with the paired samples approach, difference score variable, interpreting and summarizing the results, assumptions underlying the t tests.

Module 4:

One Way ANOVA with One between Subjects Factor – Basics of one way ANOVA, between subjects design, multiple comparison procedures, statistical significance versus the magnitude of the treatment effect, writing the SAS program, interpreting and summarizing the results. Factorial ANOVA with Two Between Subject Factors – Introduction, Factorial Design Matrix, significant main effects and significant interaction effects, writing the SAS program, interpreting and summarizing the results

REFERENCES

1. Norm O'Rourke, Larry Hatcher and Edward J. Stepanski (2005): Using SAS for Univariate and Multivariate Statistics, SAS Institute Inc. and Willey

2. Der, G. and Everitt, B.S.(2006). A Handbook of Statistical Analysis Using SAS, CRC Press.

PATTERN OF QUESTIONS

Questions shall be set to assess knowledge acquired, standard application of knowledge, application of knowledge in new situations, critical evaluation of knowledge and the ability to synthesize knowledge. The question setter shall ensure that questions covering all skills are set. He/she shall also submit a detailed scheme of evaluation along with the question paper.

For each paper the duration of the examination is 3 hours and maximum mark is 80. The question paper will have 3 parts: Part A, Part B and Part C. Part A will consist of 8 short answer each carrying 2 marks and a candidate has to answer all of them (16 marks). Part B will consist of 6 questions each carrying 4 marks and the candidate has to answer 4 questions from this part (16 marks). Part C will consist of 6 questions each carrying 12 marks and the candidate has to answer4 questions from this part (48 marks).

MODELQUESTIONPAPER

FIRST SEMESTER M.Sc DEGREE EXAMINATION

Branch: Statistics with Data Analytics

MST1C03: Distribution Theory

Time : 3 Hours **Maximum Marks : 80**

Part A

(Answer ALL questions. Each question carries 2 marks).

1. Define pgf and obtain the pgf of Negative Binomial distribution
2. Define generalized power series distribution and obtain its mgf.
3. Define log normal distribution and obtain its mean.
4. A truncated Poisson distribution is given by the mass function, $P(X = x) = e^{-\lambda}\lambda^x/\{(1 - e^{-\lambda}).x!\}$, x = 1, 2, 3, …, find the mean of the distribution.
5. State Fisher- Cochran theorem.
6. Define non central t distribution.
7. Define (i) r^{th} order statistic and (ii) standard error.
8. Derive the standard error of the sample mean.

(8 X 2 = 16 marks)

Part B

(Answer any FOUR questions. Each question carries 4 marks)

9. Let $X_1, X_2, X_3, \ldots, X_{k-1}$ have a multinomial distribution. Find the mgf of the distributionand hence obtain the marginal distribution of X_1
10. If X and Y are independent binomial variates with parameters (n_1, p) and (n_2, p) respectively,show that the conditional distribution of X given X + Y = n is hyper geometric.
11. Define compound Poisson distribution and obtain its probability mass function
12. Let $X_1, X_2, X_3, \ldots, X_n$ be independent normal variates having same variance and let Q be aquadratic function in these variates. Obtain the characteristic function of Q.
13. Establish the relationship between Chi-square, t and F distributions
14. Derive the standard error of the sample correlation coefficient.

(4 X 4 = 16 marks)

Part C

(Answer any FOUR questions. Each question carries 12 marks)

15. (i) Define hyper geometric distribution
    (ii) Find the factorial moments of the hyper geometric distribution and hence or otherwise derive the mean and variance of the distribution.
16. Derive the recurrence relation of cumulants of power series distribution. Show that binomial distribution and negative binomial distribution are special cases of power series distributions
17. (i) Define Standard Weibull distribution and obtain its rth raw moment.
    (ii) If Xi, i = 1,2,3,…,n are i.i.d.r.v's having Weibull distribution with three parameters, show that the variable Y = min(X1, X2, …, Xn) , also has Weibull distribution and identify the parameters.
18. Define non non central F variate, derive its pdf and obtain the mean. Also deduce the pdf of central F distribution from the pdf of the non central F distribution.
19. (i) Define order statistics. Derive the distribution of rth order statistic based on a random sample of size n from a population.
    (ii) Derive the distribution of largest sample in case of uniform (0, θ) population.
20. (i) Derive the asymptotic distribution of sample median
    (ii) Derive the standard error of sample variance

……

(4 X 12 = 48 marks)

MODELQUESTIONPAPER

FIRST SEMESTER M.Sc DEGREE EXAMINATION

Branch: Statistics with Data Analytics

MST1C01: - Mathematical Methods for Statistics

Time : 3 Hours **Maximum Marks : 80**

Part A

(Answer ALL questions. Each question carries 2marks).

1. Define closed set. Give an example.
2. What is the kind of discontinuity of the function $f(x) = (\sin 2x)/x$; $x \neq 0$; $= 0$; if $x = 0$ at the origin?
3. Using Lagrange's mean value theorem prove that if $f'(x) = 0$ for all $x \varepsilon [a, b]$, then $f(x)$ is a constant on $[a, b]$.
4. A function f is defined on R by $f(x) = x$; $0 \leq x < 1$; $= 1$; $x \geq 1$. Examine whether the derivative of $f(x)$ at $x = 1$ exists.
5. Explain dimension of a vector space
6. Explain idempotent matrix.
7. Distinguish between ordinary inverse and generalized inverse.
8. Define index and signature of the quadratic form

(8 X 2 = 16 marks)

Part B

(Answer any FOUR questions. Each question carries 4 marks)

9. Establish Cauchy's principle of convergence of sequence of real numbers.
10. Explain classification of quadratic forms with an example.
11. State Alembert's Ratio Test. Test the convergence of the infinite series.
12. State and prove Rolle's theorem of differential calculus.
13. Define extreme values. Show that the function $f(x, y) = (y-x)^4 + (x-2)^4$ has a minimum at (2, 2).
14. Describe the method of finding inverse of a square matrix with an example

(4 X 4 = 16 marks)

Part C

(Answer any FOUR questions. Each question carries 12 marks)

15. Define limit point of a set. Give an example. Also state and prove Bolzano-Weierstrass theorem.
16. Explain linear dependence and independence. Prove that linear dependence and independence in a system of vectors is not changed by scalar multiplication of the Vectorsbynon-zeroscalar.
17. State and prove a necessary and sufficient condition for Riemann-Stieltje's integrability.
18. (i) State and prove fundamental theorem of integral calculus.
    (ii) Show that every continuous function is integrable.
19. Explain basis and orthogonal basis. Also explain the Gram Schmidt orthogonalization process.
20. Explain diagonalization of a quadratic form. Prove that if A is a real symmetric matrix, then $P'AP = \Lambda$, where P is an orthogonal matrix and Λ is a diagonal matrix.

(4 X 12 = 48 marks)

MODELQUESTIONPAPER

SECOND SEMESTER M.Sc. DEGREE EXAMINATION

Branch: Statistics with Data Analytics

MST2C08: Statistics using Python programming

Time: 3 Hours **Maximum Marks: 80**

Part A

(Answer All questions. Each question carries 2 marks).

1. What are keywords? Give examples
2. How do you create Python file?
3. Briefly explain encapsulation with example.
4. Differentiate formal argument and actual argument.
5. Discuss elif statement in python with example.
6. Difference between histogram and bar chart.
7. Give applications of matplotlib in Python.
8. Define data series in pandas

(8 X 2 = 16 marks)

Part B

(Answer any FOUR questions. Each question carries 4 marks)

9. Explain the concept of function in python.
10. How is *Hypothesis Testing* using in *Linear Regression*?
11. Explain the usage of different data types in python.
12. How to use lambda function in python. Give an example.
13. Briefly explain recursive function with example.
14. What are the different ways a data frame can be created in pandas ?

(4 X 4 = 16 marks)

Part C

(Answer any FOUR questions. Each question carries 12 marks)

15. What is a function and explain its advantages. Explain its advantages. Explain with syntax how to create a user-defined functions and how to call the user-defined function from the main function with an example.
16. Differentiate the syntax of if…else and if…elif…else with an example.
17. What are the different operators used in Python?
18. Explain the OOP concept in python with example.
19. What are the uses of Pandas library? Give details with examples.
20. Write a Python class named Circle constructed by a radius and two methods which will compute the area and the perimeter of a circle.

(4 X 12 = 48 marks)

MODELQUESTIONPAPER

THIRD SEMESTER M. Sc DEGREE EXAMINATION

Branch: Statistics with Data Analytics

MST3E01: Data Mining

Time : 3 Hours **Maximum Marks : 80**

Part A

(Answer ALL questions. Each question carries 2 marks).

1. What is data mining?
2. Write a short note on OLAP operations.
3. What is data visualization?
4. How to improve the efficiency of Apriori algorithm?
5. Define tree pruning.
6. What are Bayesian classifiers?
7. What is cluster analysis?
8. Differentiate between Agglomerative and Divisive hierarchical clustering.

(8 X 2 = 16 marks)

Part B

(Answer any FOUR questions. Each question carries 4 marks)

9. Explain briefly about the classification of data mining systems.
10. What are the various data mining functionalities? Explain each with suitable example.
11. Write a note on generating association rules from frequent item sets.
12. Explain briefly about K-Nearest Neighbour classifiers.
13. Write a note on various types of data in cluster analysis.
14. Explain briefly about DBSCAN density based clustering method. **(4 X 4 = 16 marks)**

Part C

(Answer any FOUR questions. Each question carries 12 marks)

15. Explain in detail about the various steps involved in the KDD process with suitable figure.
16. What is data preprocessing and why it is important? Explain in detail about the major tasks in data preprocessing.
17. Explain about the Apriori algorithm for finding frequent item sets with an example.
18. Explain in detail about various Rule Based algorithms.
19. Explain Naïve Bayesian classification in detail with example.
20. Explain the hierarchical methods of classification in detail. **(4 X 12 = 48 marks)**