Indian Statistical Institute

Bachelor of Statistical Data Science (BSDS)

Course Structure and Syllabus

Contents

Curriculum **3**

1 Detailed Syllabus for Semester I **5**

Statistics I: Data Exploration . . . 5
Probability I . . . 5
Mathematics I . . . 6
Introduction to Computing . . . 7
[E] Biology I . . . 8
[E] Economics I . . . 9
[E] Earth System Sciences . . . 10

2 Detailed Syllabus for Semester II **11**

Statistics II: Introduction to Inference . . . 11
Mathematics II . . . 12
Data Analysis using R and Python . . . 13
Optimization and Numerical Methods . . . 13
[E] Biology II . . . 14
[E] Economics II . . . 14
[E] Physics . . . 16

3 Detailed Syllabus for Semester III **17**

Statistics III: Multivariate Data and Regression . . . 17
Statistical Inference . . . 17
Probability II . . . 18
Mathematics III . . . 19
Data Structures and Algorithms . . . 19

4 Detailed Syllabus for Semester IV **21**

Advanced Statistical Methods I . . . 21
Linear Statistical Models . . . 22
Stochastic Processes . . . 22
Mathematics IV . . . 23
Database Management Systems . . . 23

5 Detailed Syllabus for Semester V **25**

Regression Techniques . . . 25
Bayesian Inference . . . 26

Statistical Learning I 26
Advanced Statistical Methods II 27
Signal, Image & Text Processing 28

6 Detailed Syllabus for Semester VI **30**
Multivariate Statistical Analysis 30
Discrete Data Analysis 31
Large Sample and Resampling Methods 31
Time Series Analysis & Forecasting 32
Statistical Learning II 33

Curriculum

The course structure for the first four years is given below. The structure for the third year onwards is tentative, and subject to minor changes. The course structure for the first four years is given below. Detailed syllabi are provided for the subjects of the first three years. Syllabi for the subjects of the Fourth Year (Semesters VII and VIII) will be published in due course.

First Year

Semester I	Semester II
Statistics I: Data Exploration	Statistics II: Introduction to Inference
Probability I	Mathematics II
Mathematics I	Data Analysis using R and Python
Introduction to Computing	Optimization and Numerical Methods
Elective (1 out of 3)	*Elective (1 out of 3)*
[E] Biology I	[E] Biology II
[E] Economics I	[E] Economics II
[E] Earth System Sciences	[E] Physics

Second Year

Semester III	Semester IV
Statistics III: Multivariate Data and Regression	Advanced Statistical Methods I
Statistical Inference	Linear Statistical Models
Probability II	Stochastic Processes
Mathematics III	Mathematics IV
Data Structures and Algorithms	Database Management Systems

Third Year

Semester V	Semester VI
Regression Techniques	Multivariate Statistical Analysis
Bayesian Inference	Discrete Data Analysis
Statistical Learning I	Large Sample and Resampling Methods
Advanced Statistical Methods II	Time Series Analysis & Forecasting
Signal, Image & Text Processing	Statistical Learning II

Fourth Year

Semester VII	Semester VIII
Deep Learning I with GPU programming	Deep Learning II
Distributed and Parallel Computing	Analysis of (Algorithms for) Big Data
Elective (3 out of 5)	Data Analysis, Report writing and Presentation
[E] Genetics and Bioinformatics	*Elective (2 out of 4)*
[E] Introduction to Statistical Finance	[E] Causal Inference
[E] Clinical Trials	[E] Actuarial Statistics
[E] Sample Surveys & Design of Experiments	[E] Survival Analysis
[E] Statistical Quality Control and Operations Research	[E] Analysis of Network Data

Chapter 1

Detailed Syllabus for Semester I

Statistics I: Data Exploration

R should be introduced at the very beginning of the course and should be used to discuss, illustrate and substantiate the lectures through the analysis of suitable data sets.

1. Concepts of population and sample. Intuitive idea about a random sample.
2. Observational studies and randomized studies.
3. Types of data and methods of data collection. Primary and secondary data. Summarization and presentation of different kinds of univariate and bivariate data. Box plots, histograms, empirical c.d.f., Q-Q plots, scatter plots. Bar plots, pie charts.
4. Descriptive measures: Location, dispersion and skewness. Concept of outliers and robust measures. Comparison of various methods.
5. Analysis of bivariate data. Measures of association, correlation and simple linear regression.
6. Categorical data. Cross tabulation. Basic properties. Odds Ratio.

References

1. *Fundamentals of Descriptive Statistics* by Zealure Holcomb.
2. *Statistics* by David Freedman, Robert Pisani and Roger Purves.
3. *The Art of Statistics. How to Learn from Data* by David Spiegelhalter.

Probability I

1. Mathematical set-up for probability, equally likely outcomes, basic counting arguments.
2. Conditional probability, Bayes' rule, independent events.
3. Sampling with and without replacement, Binomial distribution, Normal and Poisson approximations to Binomial distribution.
4. Discrete random variables: Discrete Uniform, Bernoulli, Binomial, Poisson, Hyper-geometric, Geometric distributions.

5. Expectation of a random variable. Variance, standard error, higher moments and generating functions.
6. Discrete joint distributions, independent random variables, repeated trials.
7. Continuous random variables, density, cumulative distribution function, change of variable formula. Expectation, variance and higher moments for continuous random variables. Uniform, Exponential, Gamma, Beta and Normal distributions.
8. Continuous joint distributions, independence. Discrete and continuous conditional distributions and conditional expectations. Covariance and correlation of two random variables. Correlation and independence.
9. Bivariate Normal distribution, density function and basic properties. Marginal and conditional distributions. Spherical symmetry of the Bivariate Normal distribution.
10. Distribution of sums, products and quotients for continuous distributions. Student's t, χ^2 (Chi-square) and F distributions, and their elementary properties.

References

1. *Probability* by J. Pitman
2. *Introduction to Probability Theory* by R. G. Hoel, S. C. Port and C. J. Stone
3. *A First Course in Probability* by Sheldon M. Ross

Mathematics I

One variable calculus

1. Sets: Set operations. Countable and uncountable sets.
2. Functions: injective and surjective functions. Composition of functions. Inverse of a bijective function.
3. Sequences and their limits. Convergent sequences. Cauchy sequences. Series, sum of a series.
4. Catalogue of essential functions (Polynomial, Trigonometric, Exponential, Logarithmic).
5. Limit and continuity of a function. Computation of limits. Properties of continuous functions.
6. Derivative of a function. Derivatives of polynomial, exponential and trigonometric functions. Chain rule.
7. Properties of differentiable functions. Mean Value Theorem, Taylor's theorem. Maxima/minima of a function, L'Hôpital's rule.
8. Riemann integration. Some classes of integrable functions. Rules of Integration: Integration by parts, substitution rule. (Trigonometric integrals. Trigonometric substitution.) Fundamental Theorems of Calculus.
9. Improper Riemann integrals.
10. Sequence of functions: definition and examples.

Linear Algebra

1. Vector spaces, subspaces, linear independence. Basis. Dimension. Sum and intersection of subspaces.
2. Matrices. Elementary row operations. Rank of a matrix. Column space and row space.
3. Operations with partitioned matrices. Trace and determinant of a matrix.
4. Linear transformations, matrix of a linear transformation.
5. Linear equations. Homogeneous and inhomogeneous system of equations. Consistency. Solution space.

References

1. *Linear Algebra and its Applications* by Gilbert Strang
2. *Introduction to Linear Algebra* by Gilbert Strang
3. *Calculus: One-Variable Calculus with An Introduction to Linear Algebra, Vol 1* by Tom Apostol
4. *Calculus and Analytical Geometry* by George B. Thomas and Ross L. Finney

Introduction to Computing

Software Programming / General techniques

1. Introduction to number system: binary, octal, hexadecimal. Introduction to algorithms: Illustration using simple examples (Euclid's algorithm, bisection method, regula falsi). Representation of numbers and data types: Signed and unsigned integers, floating points, overflow, underflow.
2. Imperative programming: Introduction, common syntax and constructs — variables, assignment, expressions, conditionals and branching, iteration. Input / output. Functions: parameter passing, call by value, call by reference, recursion. Illustration using C / R / Python / Javascript (basic similarities and differences).
3. Arrays. Illustration using sorting.
4. Pointers, structures, dynamic allocation.
5. Data Structures: Queue, Stack, Linked lists, Trees.
6. Basic ideas of Object-oriented programming: Introduction to classes, inheritance, overloading, polymorphism.

Software / hardware / Software-hardware interface

1. Introduction to digital computers: CPU, main memory, peripherals, I/O devices, storage.
2. Random access machine model of computing. Basic concepts of formal run-time analysis.
3. Basic ideas of parallel programming.
4. Memory management strategies: C (dynamic allocation / free) vs R / Python (garbage collection)

Algorithms

1. Introduction to error analysis: round-off errors, floating point operations, error propagation, condition number and stability. Illustration using simple examples.

References

1. *C++ How to Program* by Paul Deitel and Harvey Deitel
2. *Introduction to Numerical Analysis* by Josef Stoer and Roland Bulirsch

[E] Biology I

Only if no Biology in +2

1. *Biological classification of living organisms*: Distinguishing characteristics of living and non-living things; Definition and concept of Biodiversity; Need for classification; Three domains of life; Concept of species and taxonomical hierarchy; Binomial nomenclature. Five kingdom classifications; classification of plants; classification of animals
2. *Cell and cell division*: Study of cell structure and functions; Mitosis and Meiosis, comparative account of mitosis and meiosis. Cell theory and cell as the basic unit of life, structure of prokaryotic and eukaryotic cells; Plant cell and animal cell; Cell envelope; Cell membrane, Cell wall. Chemical constituents of living cells: biomolecules, structure and function of proteins, carbohydrates, lipids, nucleic acids; Enzymes- types, properties, enzyme action; Cell cycle, mitosis, meiosis and their significance
3. *Central dogma*: Structure and function of DNA and RNA; Replication, transcription, translation.
4. *Basic biochemistry*: Metabolism of protein, carbohydrate and fat.
5. *Basic microbiology*: Microbial culture medium. Microbial morphology, growth and development. Applications of microbiology.
6. *Basic agriculture*: Plant, soil and environment interaction. Major agricultural and horticultural crops of India. Pest and diseases of crop plants.
7. *Plant Physiology*: Photosynthesis: Light harvesting complexes; mechanisms of electron transport; photoprotective mechanisms; CO2 fixation-C3, C4 and CAM pathways. Respiration and photorespiration: Citric acid cycle; plant mitochondrial electron transport and ATP synthesis; alternate oxidase; photorespiratory pathway. Nitrogen metabolism: Nitrate and ammonium assimilation; amino acid biosynthesis.
8. *Animal Physiology*: Digestive system. Cardiovascular System. Respiratory system. Nervous system. Sense organs: Vision, hearing and tactile response. Excretory system.

References

1. *Instant Notes on Biochemistry* by B. D. Hames, N. M. Hooper, J. D. Houghton, Viva publications
2. *Instant Notes on Genetics* by P. C. Winter, G. I. Hickey and H. L. Fletcher, Viva Publication
3. *Principles of Genetics* by D. P. Snustad and M. J. Simmons, John Wiley & Sons Inc
4. *Biochemistry* by U. Satyanarayana and U. Chakrapani

5. *Plant Physiology, Development and Metabolism* by Satish C. Bhatla and Manju A. Lal, Springer
6. *Textbook of Animal Physiology* by P. B. Reddy
7. *Biochemistry* by Lehninger
8. *Microbiology* by Michael Pelczar, Jr. and Noel R. Kreig
9. *Principles and Methods in Landscape Ecology* by Almo Farina
10. *Microbiology: Concepts and Applications* by Michael Pelczar, Jr., E. C. S. Chan, Noel R. Krieg, McGraw Hill Education; 5th edition
11. *General Microbiology* by R. P. Singh, Kalyani Publishers
12. *Microbiology* by Lansing Prescott, J. P. Harley, D. A. Klein, Brown (William C.) Co, U.S.; 2nd edition
13. *Microbiology Fundamentals and Applications (7th Ed.)* by S. S. Purohit, Agrobios
14. *Environmental Science* by S. C. Santra, New Central Book Agency
15. *Environmental Science: A comprehensive Treatise on Ecology and Environment* by Sovan Roy, Publishing Syndicate, Calcutta

[E] Economics I

Only if no Economics in +2

Micro Economics

1. Welfare Economics: Supply and Demand, Elasticity.
2. Consumption and Consumer behaviour.
3. Production and Theory of costs.
4. Market Organisation: Competition, Monopoly.

Macro Economics.

1. National income accounting, demand and supply.
2. Simple Keynesian model and extensions.
3. Consumption and Investment.
4. Inflation and Unemployment.
5. Fiscal policy.
6. Money, banking and finance.

References

1. *Intermediate Microeconomics* by Hal Varian.
2. *Microeconomic Theory* by Richard Layard and A.A. Walters
3. *Microeconomics in Context* by N. Goodwin, J. Harris, J. Nelson, B. Roach and M. Torras

4. *Microeconomics: behavior, institutions, and evolution* by S. Bowles
5. *Macroeconomics* by N. G. Mankiw
6. *Macroeconomics* by R. Dornbusch and S. Fisher

[E] Earth System Sciences

Prerequisite: Physics, Chemistry and Mathematics in +2

1. *The Earth System*: The Scientific Method; System Concept, Dynamic Interactions among Systems, The Energy Cycle, Introduction to Geological data.
2. *Earth as a Planet in the Solar System*: Origin of the Universe and the Solar System, Evolution of the Planets, Meteorites and Asteroids, Origin of Atmosphere, Ocean and Life.
3. *The Solid Earth*: The Earth as a Layered Planet/ Mechanical Layering of the Earth, Layers of Different Composition and Physical State, Earthquake and the Earth's Interior, Plate Tectonics, Geothermal Gradient, Magmas and Volcanoes.
4. *The Earth's Evolving Crust*: Sedimentary Strata; Sedimentary process and Sedimentary Rocks; Metamorphism and Metamorphic Rocks; Plate Tectonics, Continental Crust and Mountain Building; Understanding the past from Stratigraphic Records; Geological Time Scale.
5. *Hydrosphere, Atmosphere and Biosphere*: Water and the Hydrologic Cycle; Snow and Ice; The Oceans; The Atmosphere; Winds and the Global Air Circulation; The Earth's Climate System and the Changing Climate.
6. *Life on Earth*: A Planetary Perspective on Life; The Habitable Planet; Biogeochemical Cycles and Biological Evolution; Mass Extinctions.
7. *Resources on Earth*: Coal and Petroleum Resources; Nuclear, Wind and Hydroelectric Power Energy.
8. *Geostatistics*: Application of statistics to Geological Data; Pattern recognition of geological events; Statistics for earth resources.

References

1. *The Blue Planet – An Introduction to Earth System Science* by B.J. Skinner, S.C. Porter and D.B. Botkin
2. *The Earth Machine – The Science of a Dynamic Planet* by E.A. Mathez and J.D. Webster
3. *Understanding Earth, 5th Edition* by J. Grotzinger, T.H. Jordan, F. Press and R. Siever
4. *Planet Earth: Cosmology, Geology, and the Evolution of Life and Environment* by C. Emiliani

Chapter 2

Detailed Syllabus for Semester II

Statistics II: Introduction to Inference

Methods will be motivated through real-life examples. Performance of statistical procedures will be assessed through numerical simulations.

1. Population vs. sample. Empirical distribution. Parametric statistical models. Need for inference on a parameter.
2. Point estimation: Concept of bias, variance, mean squared error (MSE), relative efficiency.
3. Estimation of parameters by method of moments and maximum likelihood method.
4. Sampling distributions derived from Normal populations: χ^2 (Chi-square), t and F distributions. (Discussion of properties without derivation).
5. Introduction to asymptotic inference: Informal discussion of Weak Law of Large Numbers and Central Limit Theorem. Consistency of estimators. Approximating sampling distributions through CLT.
6. Numerical computation of sampling distributions based on independent samples from an arbitrary population. Comparison with CLT-based approximations.
7. Hypothesis testing: Null and alternative hypotheses, simple and composite hypotheses, Type I and Type II errors, level and power of a test, p-value. Neyman-Pearson Lemma, most powerful tests, unbiased tests.
8. Exact and large sample tests for binomial proportion and mean of a Normal distribution (one sample case). One-sided and two-sided alternatives.
9. Exact (pooled-t) and large sample tests for the equality of two Normal means. Paired t-test. Test for equality of variances of two Normal populations. Test for the equality of two binomial proportions.
10. Interval estimation: Construction using pivotal quantities and critical regions. Exact and large sample confidence intervals. Confidence intervals for Binomial, Normal and Poisson parameters (one sample case).
11. Large sample confidence intervals for the difference of two Binomial proportions. Exact and large sample confidence intervals for the difference of means of two Normal populations.

12. Sample size determination in tests of hypotheses for ensuring a specified power, and in constructing confidence intervals for ensuring a specified interval width.

References

1. *Statistics* by David Freedman, Robert Pisani and Roger Purves.
2. *The Art of Statistics. How to Learn from Data* by David Spiegelhalter.
3. *All of Statistics: A Concise Course in Statistical Inference* by L. Wasserman.
4. *Probability and Statistics* by M. H. DeGroot and M. J. Schervish.
5. *Mathematical Statistics with Applications* by D. Wackerly, W. Mendenhall, R. L. Scheaffer.

Mathematics II

Calculus

1. Infinite Series. Alternating Series. Absolute Convergence. Tests of convergence: comparison test, root test, ratio test, integral test.
2. Power Series: Radius of convergence of a power series. Differentiation/integration of power series. Revisit Trigonometric functions, Exponential Functions and Logarithms.
3. Functions of Several Variables. Limits and Continuity.
4. Partial Derivatives, directional derivatives. Differentiability. Differentiability of functions with continuous partial derivatives. Jacobian and Chain rule.
5. Matrix differentiation with examples.
6. Maximum and minimum values. Hessian matrix. Multivariate Taylor series.
7. Multiple Integrals as iterated integrals. Change of variables in multiple integrals. Jacobian formula.

Linear Algebra

1. Orthogonality and its geometric interpretation.
2. Eigenvalues and eigenvectors. Spectral decomposition. Singular value decomposition (SVD).
3. Positive semidefinite matrices. Projection matrices.
4. Matrix norms, and low rank matrix approximation using SVD.
5. Computations using matrices: QR decomposition, Gram-Schmidt orthogonalization, Cholesky decomposition.

References

1. *Linear Algebra and its Applications* by Gilbert Strang
2. *Introduction to Linear Algebra* by Gilbert Strang
3. *Calculus: One-Variable Calculus with An Introduction to Linear Algebra, Vol 1* by Tom Apostol
4. *Calculus: Multi-Variable Calculus and Linear Algebra with Applications to Differential Equations and Probability, Vol 2* by Tom Apostol

Data Analysis using R and Python

1. The REPL model for R and Python. IDEs such as RStudio and Jupyter.
2. Vectors (arrays) in R and Python (through numpy). Indexing. Other data types: Lists, data frames, dictionaries. Operators and functions, attributes. Using the help system.
3. Objects, workspace, add-on packages. Data import and export. Excel / CSV, native file formats. Common data manipulation and summary operations in R; base packages and dplyr. Scoping rules.
4. Probability calculations and simulation. Data visualization. Basic statistical modeling, formula interface.
5. Dynamic documents and notebooks. Illustration using knitr and related packages in R. Exporting as PDF / HTML reports.
6. Working simultaneously with multiple languages. Calling C / C++ / Python from R (Rcpp, reticulate).

References

1. *Introductory Statistics with R* by Peter Dalgaard
2. *R for Data Science* by Hadley Wickham and Garrett Grolemund
3. *Python Data Science Handbook* by Jake Vanderplas
4. *Python for Data Analysis* by Wes McKinney

Optimization and Numerical Methods

Focus will be on methods and applications. Relevant mathematical results will be stated without proof. Selected optimization methods will be implemented in R, Python or Julia.

Optimization

1. Role of optimization in Data Science/Statistics through motivating examples (e.g., least squares and maximum quasi-likelihood methods in linear/generalized linear models, support vector machine, penalized regression, matrix approximation).
2. Basics of convex optimization: Overview of convex sets and convex functions. Convexity preserving operations. Notions of convex hull, cone, polyhedra.
3. Optimization problems: Convex vs. non-convex optimization. Constrained vs. unconstrained optimization. Global and local optima.
4. Optimality conditions: First and second order optimality conditions. Primal and dual problems. Method of Lagrange multiplier. KKT condition.
5. Introduction to Linear Programming with examples.
6. Introduction to Quadratic Programming with examples.
7. Introduction to Gradient-based methods with examples.
8. Introduction to Second order methods with examples.

Numerical Methods

1. Numerical solution of nonlinear equations in one variable: Bisection and Newton-Raphson methods.
2. Basic concepts of interpolation. Polynomial interpolation.
3. Introduction to Numerical integration with examples.

References

1. *Convex Optimization: Algorithms and Complexity* by S. Bubeck.
2. *Convex Optimization* by S. Boyd and L. Vandenberghe.
3. *Numerical Optimization* by J. Nocedal and S. Wright.
4. *Modern Optimization with R* by P. Cortez.
5. *Julia Programming for Operations Research* by C. Kwon.
6. *Linear Algebra and Learning from Data* by G. Strang.
7. *Elementary Numerical Analysis: An Algorithmic Approach* by S. D. Conte and C. de Boor.

[E] Biology II

Prerequisite: Either Biology I or Biology in +2

1. *Predator-prey interaction*: Definitions, relationships and population dynamics, Evolution and examples.
2. *Principles of genetics*: Definition of gene and genetic code; relationship between them. Mendel's Law of genetics and application in human population. Interaction of Genes or Factor Hypothesis: Introduction; incomplete dominance or blending inheritance; lethal factor; simple interaction or two factor pairs affecting the same character; epistasis — complementary factor; supplementary factor; inhibitory factor; duplicating factor or multiple factor; polymerisms. Sex Chromosome and Sex-Linkage: Sex chromosomes; sex-linkage or sex-linked inheritance or inheritance related to sex. Cytoplasmic Inheritance or Extranuclear inheritance: Introduction; maternal effect; extranuclear inheritance.
3. *Landscape ecology*: Introduction to landscape ecology framework. Scaling patterns and processes across landscapes. Landscape heterogeneity and disturbances. Principles of landscape dynamics. Methods in landscape ecology.
4. *Environmental science*: Components of environment, Environmental pollution and management, Climate change and global warming, Environmental quality analysis.

References

See references for Biology I

[E] Economics II

Prerequisite: Either Economics 1 or Economics in +2

1. *Index numbers*: Construction of index numbers, properties, some well-known index number formulae, problem of construction of index numbers, chain indices, cost of living indices, splicing of index numbers, different types of index numbers used in India.
2. *Analysis of income and allied size distributions*: Pareto and log-normal distributions, genesis, specification and estimation, Lorenz curve, Gini coefficient.
3. *Demand analysis*: Classification of commodities, Engel curve analysis using cross-section and time series data, Engel curves incorporating household characteristics, demand projection, specific concentration curves.
4. *Production analysis*: Profit maximization, cost minimization, returns to scale, Cobb-Douglas and ACMS production functions.
5. *International Statistical system*: Overview on Sectoral statistics and other economic Statistics: Standard International classifications used for compilation of economic statistics – International Standard Industrial Classification (ISIC Rev. 4), Administrative data, Business Registers Overview of Sectoral Statistics: Agriculture, forestry, fisheries; Mining, Manufacturing, Energy & construction, Domestic Trade and Transport, Banking, insurance, financial statistics, Government finance statistics, and Services sector statistics. Employment & Labour, Prices, Merchandise trade statistics and Statistics on International Trade in Services, System of National Accounts.
6. *Measurement of vital rates*: SRS, Life table, Literacy rate, etc.
7. *Statistics of Production*: agriculture and industry, annual survey of industries, index of industrial production.
8. Price Statistics, consumer price index numbers.
9. Income and consumer expenditure distribution, poverty.
10. Employment and unemployment.

References

1. *Statistics for Economists* by P.H. Karmel and M. Polasek.
2. *Price Index Numbers* by R.G.D. Allen.
3. *Income Inequality and Poverty* by N. Kakwani.
4. *An Introduction to Econometrics* by L.R. Klein.
5. *Empirical Econometrics* by J.S. Cramer.
6. *Econometric Models, Techniques and Applications* by M.D. Intrilligator.
7. *Indian Official Statistical Systems* by M.R. Saluja.
8. *Living Standard Measurement Surveys.* World Bank
9. *National Accounts: A Practical Introduction.* United Nations, New York, 2003
10. *Business Registers - Recommendations Manual.* EUROSTAT, 2010
11. *International Standard Industrial Classification of All Economic Activities, Revision 4.* United Nations, New York, 2008.
12. *Classifications of Expenditure according to Purpose.* United Nations, New York, 2000.

13. *International Recommendations for Industrial Statistics.* United Nations, New York, 2008.

[E] Physics

Prerequisite: Physics in +2

1. Thermodynamics (Zeroth to Third Laws), Carnot Engines, Heat Equations, Various thermodynamic potentials, Relations (Requires partial derivatives, multivariable extrema).

2. Equilibrium Statistical mechanics: The microcanonical, Canonical and Grand Canonical Ensembles; Ideal Classical Gas from Statistical Mechanics; Applications of the three ensemble formulations (spin systems, gases, etc.: preferably should include a fairly detailed investigation of Ising Model in $d = 1$ and 2 with introduction of mean-field theory).

3. Classical Equilibrium Phase Transitions: Gibbs phase rule, Classification of phase transitions, Clausius-Clapeyron Equations, Rankine Cycles, etc.; Scaling ideas, Widom's scaling hypothesis, continuous phase transitions, critical exponents, calculation of critical exponents from mean-field theory in suitable systems.

4. Basics of Renormalization Group: Kadanoff's block spin transformation/coarse graining, Exposition of various concepts (dimensional analysis for scaling, relevant, marginal and irrelevant operators, fixed points and their stabilities, RG flow diagrams, corrections to critical exponents etc.) through simple problems, universality.

5. Random systems: annealed and quenched disorder, Harris criterion, introduction to spin glasses and Replica method, perturbative approach to random fixed point, introduction to percolation.

6. Computer Simulations in Classical Statistical Mechanics: Monte Carlo simulations, Molecular and particle dynamics.

References

1. *Statistical Physics of Particles* by Mehran Kardar

2. *Statistical Physics of Fields* by Mehran Kardar

3. *Statistical Mechanics* by Pathria and Beale

4. *Thermodynamics and an Introduction to Thermostatistics* by Herbert Callen

5. *Lectures on Phase Transitions and the Renormalization Group* by Nigel Goldenfeld

6. *Statistical Mechanics: Entropy, Order Parameters, and Complexity* by James P. Sethna

7. *Scaling and Renormalization in Statistical Physics* by John Cardy

8. *A Guide to Monte Carlo Simulations in Statistical Physics* by David P. Landau, Kurt Binder

Chapter 3

Detailed Syllabus for Semester III

Statistics III: Multivariate Data and Regression

1. *Review of bivariate Normal distribution*: Conditional distribution and its relation to the simple linear regression model.
2. *Multiple linear regression*: Least squares estimation. Normal equations. Properties of residuals. Multiple and partial correlation coefficients.
3. *Inference on linear regression*: Properties of least squares estimators under Gaussian noise. Hypothesis tests and confidence intervals for regression coefficients.
4. *Analysis of Variance* for one-way classified data. Interpretation of identifiability constraints. ANOVA table. Tests of hypothesis.
5. *Review of Multinomial distribution*: Conditional and marginal distributions. Large sample properties.
6. *Graphical summary of categorical response*: Stacked bar plot, tile chart, sunburst chart, parallel coordinates plot.
7. *Regression with binary response*: Odds ratio. Logistic regression and probit models. Estimation using Fisher scoring (description only).
8. *Goodness of fit for categorical response*: Tests of independence and homogeneity in contingency tables.

References

1. *Practical Regression and ANOVA with R* by Julian Faraway
2. *Applied Logistic Regression* by David W. Hosmer, Stanley Lemeshow
3. *Categorical Data Analysis* by Alan Agresti
4. *Applied Regression Analysis* by Norman N. Draper and Harry Smith

Statistical Inference

1. Review of estimation and hypothesis testing for one-parameter families of distributions.

2. Order statistics and heuristic derivation of its distribution for i.i.d. samples from continuous distributions.
3. Likelihood principle and sufficiency. Examples.
4. Unbiased estimation and Rao-Blackwell theorem. UMVUE.
5. Fisher information. Cramer-Rao lower bound for one and multi-parameter families.
6. Maximum likelihood estimation. Likelihood equations. Fisher's scoring method.
7. Most Powerful test and Neymann-Pearson lemma. Examples involving Binomial, Poisson, Normal, Uniform and Exponential Distributions.
8. Likelihood ratio test. Rao's score test. Wald's test. Examples involving standard families of distributions.
9. Connection between hypothesis testing and interval estimation. Examples.
10. Simultaneous inference: Notion of p-values and their distribution under null hypothesis. Notions of familywise error rate and false discovery rate. Methods for simultaneous testing: Bonferroni procedure, Holm procedure, Benjamini-Hochberg procedure. Discussion of advantages and disadvantages.
11. Introduction to U-statistics with examples.

References

1. *Stat Labs: Mathematical Statistics Through Applications* by Deborah Nolan, Terry Speed
2. *Mathematical Statistics* by Peter J. Bickel and Kjell A. Doksum
3. *Statistical Inference* by George Casella and Roger L. Berger
4. *U-Statistics, M_m-Estimators and Resampling* by Arup Bose and Snigdhansu Chatterjee
5. *Testing Statistical Hypothesis* (Chapter 9) by E. L. Lehmann and Joseph P. Romano
6. *Multiple Testing Procedures with Applications to Genomics* by S. Dudoit and M. van der Laan
7. *Multiple Comparisons: Theory and Methods* by Jason Hsu

Probability II

1. Distributions of functions of random vectors and Jacobian formula. Examples.
2. Sampling distribution for mean and sample variance of i.i.d. Normal observations.
3. Cauchy-Schwartz, Markov and Chebyshev inequalities. Weak Law of Large Numbers (WLLN) and Strong Law of Large Numbers (SLLN) (statements only).
4. Concept of almost sure and in probability convergence. Concept of "infinitely often", First and Second Borel-Cantelli Lemmas. Concepts of L_1, L_2 and general L_p convergence where $p > 1$.
5. Statements of the Monotone Convergence Theorem (MCT), Fatou's Lemma and the Dominated Convergence Theorem (DCT).
6. Basic notion of distributional convergence. Statement of the i.i.d. Central Limit Theorem (CLT). Slutsky's Theorem. Delta method. Multivariate CLT, Cramer-Wold device, statement and examples.

7. Basic idea of generating functions. Definition and examples of probability generating function (pgf) for discrete integer valued random variables, moment generating functions (mgf) and the characteristic function (chf). Properties of characteristic functions, uniqueness, inversion and density formula (statements only). Connection between chf and distributional convergence, Levy's Continuity Theorem (statement only).

References

1. *Probability* by J. Pitman
2. *Introduction to Probability Theory* by R. G. Hoel, S. C. Port and C. J. Stone.
3. *A First Course in Probability, 6th Edition* by Sheldon M. Ross.

Mathematics III

Graph Theory

1. Types of graphs, Simple Graph, Directed Graph, Undirected Graph, Complete Graph. Degree of a vertex in an undirected graph, Indegrees and Outdegrees of a directed graph.
2. Paths and Reachability in Graphs, Graph coloring, Vertex cover, Independent set, Matching, Representing graphs, Adjacency matrix.
3. Breadth-First Search (BFS) algorithm. Depth-First Search (DFS) algorithm. Applications.

Analysis

1. Basics of complex number system. Roots of polynomials. Power series in complex variables, complex exponential.
2. Basic properties of metric spaces. Open and closed sets, notion of convergence and continuity, compactness, completeness.
3. Normed vector spaces and Hilbert spaces (definition and examples). Basic properties of Hilbert spaces: notion of complete orthogonal basis, basis expansion.

References

1. *Topology and Modern Analysis* by George F. Simmons
2. *Introduction to Graph Theory* by Douglas B. West

Data Structures and Algorithms

1. *Introduction*: Idea of data structure design (in terms of static and dynamic data), and the basic operations needed; initial ideas of algorithms and its resource usage in terms of space and time complexity; ideas of worst case, average case and ammortized case analysis, complexity classes.
2. *Construction and manipulation of basic data structures*: Idea of abstract data types and its concrete implementation; basic data structures — list, array, stack, queue, dequeue, linked lists; binary tree and traversal algorithms, threaded tree, m-ary tree, its construction and traversals; priority queue and heap.

3. *Data Structures for searching*: Binary search trees, height-balanced binary search tree; weight-balanced binary search tree; red-black tree; binomial heap; splay tree; skip list; trie.

4. *Hashing*: separate chaining, linear probing, quadratic probing.

5. *Sorting and selection*: Finding maximum and minimum, k largest elements in order; Sorting by selection, tournament and heap sort methods, lower bound for sorting, other sorting algorithms — radix sort, quick sort, merge sort; selection of k-th largest element.

6. *Searching and set manipulation*: Searching in static table — binary search, path lengths in binary trees and applications, optimality of binary search in worst cast and average-case, binary search trees, construction of optimal weighted binary search trees; Searching in dynamic table; randomly grown binary search trees, AVL and (a, b) trees.

7. *Union-Find problem*: Tree representation of a set, weighted union and path compression analysis and applications.

8. *Graph problems*: Graph searching — BFS, DFS, shortest first search, topological sort; connected and biconnected components; minimum spanning trees, Kruskal's and Prim's algorithms, Johnson's implementation of Prim's algorithm using priority queue data structures.

9. *Algebraic problems*: Evaluation of polynomials with or without preprocessing. Winograd's and Strassen's matrix multiplication algorithms and applications to related problems, FFT, simple lower bound results.

10. *String processing*: String searching and pattern matching, Knuth-Morris-Pratt algorithm and its analysis.

References

1. *Data Structures, Algorithms and Software Principles in C* by T. A. Standish. Addison-Wesley, Reading, Mass., 1995.

2. *Data Structures Using C* by A. M. Tenenbaum, Y. Langsam and M. J. Augenstein:, Pearson, 1998.

3. *The Art of Computer Programming: Volume 1* by D. E. Knuth. Third edition, Narosa / Addison-Wesley, New Delhi / London, 1997.

4. *Fundamentals of Data Structures* by E. Horowitz and S. Sahni. Galgotia Booksource, New Delhi, 1977.

5. *The Design and Analysis of Computer Algorithms* by A. Aho, J. Hopcroft and J. Ullman. A. W. L, International Student Edition, Singapore, 1998.

6. *Introduction to Algorithms* by T. H. Cormen, C. E. Leiserson, R. L. Rivest and C. Stein. Third edition, MIT Press, 2009.

Chapter 4

Detailed Syllabus for Semester IV

Advanced Statistical Methods I

Sampling Techniques

1. *Simulating probability distributions through exact sampling*: Algorithms for simulating standard univariate distributions, multivariate Normal and Multinomial.
2. *Monte Carlo techniques*: Basic Monte Carlo, Rejection sampling, Importance sampling. Application to evaluating integrals. Comparison with numerical integration.
3. *Introduction to MCMC*: Basic introduction to discrete time Markov chains. Gibbs sampling with application to sampling from conditional distributions.

Nonparametric methods

1. *Nonparametric and distribution-free methods for testing*: Sign test and signed rank test. Mann-Whitney test, Wilcoxon rank sum statistic. Nonparametric tests for one-sample and two-sample scale problems. Kolmogorov-Smirnov test. Kruskal-Wallis test for k-sample location problem.
2. *Nonparametric density estimation*: Shifted histogram and kernel density estimation. Bandwidth selection through cross-validation.
3. *Nonparametric regression*: k-nearest neighbor regression. Kernel smoothing. Bias and variance as functions of bandwidth. Bandwidth selection through cross-validation. Splines and their applications to nonparametric regression.

Resampling techniques

1. *Introduction to Resampling techniques*: Jackknife, Bootstrap and Cross-Validation as data analytic tools. Application to inference for one parameter families and linear regression.

References

1. *MCMC from Scratch* by Masanori Hanada and So Matsuura
2. *Nonparametric Statistical Methods Using R* By John Kloke and Joseph McKean
3. *Nonparametric Statistical Inference* by Jean D. Gibbons and Subhabrata Chakraborti

4. *Resampling Methods* by Phillip I. Good
5. *Theory and Methods of Statistics* (Chapter 9) by P. K. Bhattacharya and P. Burman

Linear Statistical Models

1. Linear models from regression, ANCOVA, classification, and dimension reduction perspectives. Building linear models based on designed experiments and observational studies.
2. Gauss-Markov framework for linear models. Identifiability and estimability of parameters. Best Linear Unbiased Estimator (BLUE).
3. Projections approach to least squares model fitting. Degrees of freedom. ANOVA decomposition.
4. *Testing linear hypothesis*: test based on residual sums of squares, likelihood ratio test, Wald's test.
5. *Analysis of Variance (ANOVA)*: One-way and two-way classified data. Roles of identifiability constraints. Interaction plots for two-way classified data. Tests for main effects and interactions. Nested two-way ANOVA.
6. *Analysis of Covariance (ANCOVA)*: Model fitting and testing of hypotheses.
7. *Multiple comparisons*: Inference on contrasts. Bonferroni's method, Tukey's method, Scheffé's method, Holm's procedure.
8. *Introduction to random and mixed effects models*: Bayesian interpretation. Variance components. Marginal maximum likelihood (MML) and restricted maximum likelihood (REML) methods for estimation. Approximations to MLE. Best Linear Unbiased Predictor (BLUP). Tests of hypothesis.

References

1. *Linear Models with R* by Julian Faraway
2. *Linear Models* by S. R. Searle and M. H. J. Gruber
3. *Applied Linear Statistical Models* by M. H. Kutner, C. J. Nachtsheim, J. Neter and W. Li
4. *Multiple Comparisons: Theory and Methods* by Jason Hsu
5. *Stat Labs: Mathematical Statistics Through Applications* by Deborah Nolan, Terry Speed

Stochastic Processes

1. Introduction and examples of collection of random variables and concept of a Stochastic Process.
2. Introduction to discrete Markov chains with finite and countably infinite state space, examples including two-state chain, random walks on integer line, birth and death chains, Ehrenfest chains, etc.
3. Concept of recurrent and transient states. Absorbing states. Irreducibility, decomposition of state space into irreducible classes.
4. Concept of stationary distribution, positive and null recurrence. Periodicity, cyclic decomposition of a periodic chain. Limit theorem for a finite state, aperiodic and irreducible chain (without proof).
5. Introduction to reversible chains with examples.

6. Introduction to Markov Chain Monte Carlo (MCMC), specific examples including perfect sampling and Metropolis-Hastings algorithm.

7. Poisson process and its basic properties. Inhomogeneous and compound Poisson processes, examples. Simple birth and death processes. A brief introduction to the general continuous time but discrete state space Markov chains. Introduction to $M/M/1$, $M/M/c$ and $M/M/\infty$ queues.

References

1. *Introduction to Stochastic Processes* by R. G. Hoel, S. C. Port and C. J. Stone.
2. *Finite Markov Chains* by J. G. Kemeny, J. L. Snell and A. W. Knapp.
3. *Stochastic Processes with Applications* by R. N. Bhattacharya and E. Waymire.
4. *Markov Chain Monte Carlo in Practice* by W. R. Gilks, S. Richardson and David Spiegelhalter.
5. *Monte Carlo Statistical Methods* by C. P. Robert and G. Casella.

Mathematics IV

Calculus (7 weeks)

1. *Vectors and the Geometry of Space*: Three-Dimensional Coordinate Systems, Vectors, The Dot Product, The Cross Product, Equations of Lines and Planes, Cylinders and Quadric Surfaces.

2. *Applications of integration*: Arc Length, Area of a Surface of Revolution, Applications.

3. *Parametric equations*: Statements of Inverse Function Theorem and Implicit Function Theorem. Curves Defined by Parametric Equations, Calculus with Parametric Curves.

Differential equations (7 weeks)

1. First order differential equations, Picard's theorem, linear dependence and Wronskian.

2. Dimensionality of space of solutions, linear ODE with constant coefficients of second and higher order, Cauchy-Euler equations, Method of undetermined coefficients and method of variation of parameters.

3. Boundary Value Problems: System of linear differential equations with constant coefficients, fundamental matrix, matrix methods, Laplace Transform Method.

References

1. *Advanced Engineering Mathematics* by Erwin Kreyszig

Database Management Systems

1. *Introduction*: Purpose of database systems, data abstraction and modelling, instances and schemes, database manager, database users and their interactions, data definition and manipulation language, data dictionary, overall system structure, data integration, Database Indexing.

2. *Relational data model*: Structure of a relational database, operation on relations, relational algebra, tuple and domain relational calculus, salient features of a query language.

3. *SQL*: domain types, construction, alteration and deletion of tables, query structure and examples, natural joins and other set operations, aggregations, nested subqueries, inserting, modifying and deleting data, advanced joins, views, transactions, integrity constraints, cascading actions, autorization and roles. Hands on and practical assignments.

4. *Entity-relationship model*: Entities and entity sets, relationships and relationship sets, mapping constraints, E-R diagram, primary keys, strong and weak entities, reducing E-R diagrams to tables.

5. *Databases in application development*: cursors, database APIs, JDBC and ODBC, JDBC drivers, Connections, Statements, ResultSets, Exceptions and Warnings. Practical case studies.

6. *Normalization*: Anomalies in RDBMS, importance of normalization, functional, multi-valued and join dependencies, closures of functional dependencies and attribute sets, 1NF, 2NF, 3NF and BCNF; Discussion on tradeoff between performance and normalization. Database tuning: Index selection and clustering, tuning of conceptual schema, denormalization, tuning queries and views.

7. *Query optimization*: Importance of query processing, equivalence of queries, join ordering, cost estimation, cost estimation for complex queries and joins, optimizing nested subqueries, I/O cost models, external sort.

8. *Transaction Processing and Concurrency Control in RDBMS*: Transaction Life Cycle, Concurrent Execution of Transactions; Testing for serializability, lock based and time-stamp based protocols; Deadlock detection and Recovery.

9. *Crash recovery*: Failure classification, transactions, log maintenance, check point implementation, shadow paging, example of an actual implementation.

10. *NoSQL*: Introduction to NoSQL databases, ACID vs BASE requirements, practical exercises with one noSQL system (for example MongoDB).

References

1. *An Introduction to Database Systems* by C. J. Date, Pearson Education, Inc., 8th Edition, 2006.

2. *Database System Concepts* by A. Silberschatz, H. F. Korth and S. Sudarshan, Tata McGraw-Hill, 6th Edition, 2011.

3. *Fundamentals of Database Systems* by R. Elmasri and S. B. Navathe, Pearson Education, Inc., 4th Edition, 2004.

4. *Database Management Systems* by R. Ramakrishnan and J. Gehrke, McGraw-Hil, 3rd Edition, 2007.

5. *Database Systems: The Complete Book* by H. Garcia-Molina, J. D. Ullman and J. Widom, , Pearson Education, Inc., 2nd Edition, 2009.

6. *MySQL stored procedure programming* by G. Harrison and S. Feuerstein, O'Reilly Media, Inc., 2006.

Chapter 5

Detailed Syllabus for Semester V

Regression Techniques

1. *Multiple linear regression*: Properties of least squares estimators and residuals. (Review only)
2. *Geometric view of regression*: Orthogonal projections on subspaces associated with predictors. Decomposition of sum of squares. Degrees of freedom. Multiple and partial correlation coefficients. (Review only)
3. *Prediction in linear regression*: Point and interval prediction. Measures of prediction accuracy.
4. *Inference on linear regression*: Confidence and prediction intervals. Tests of hypotheses for linear parametric functions. Simultaneous inference.
5. *Violation of linear model assumptions*: (i) Outliers: high-leverage and influential observations, consequences, diagnostics. (ii) Heteroscedasticity: consequences, diagnostics, tests (Bruesch-Pagan and White's test). (iii) Autocorrelated errors: consequences, diagnostics, ACF and PACF plots, tests (Durbin-Watson test).
6. *Methods of addressing model violation*: Variable transformation (Box-Cox family). Weighted least squares.
7. *Multicollinearity*: Consequences, diagnostics. Notion of variable importance.
8. *Model building*: Subset selection, Forward and backward stepwise regression. Categorical predictors and dummy variables.
9. *Model selection*: Prediction and model error. Adjusted R^2, FPE, Mallows' C_p, AIC, BIC criteria. Cross-validation approach.

References

1. *Applied Regression Analysis* by Norman N. Draper and Harry Smith
2. *Modern Regression Methods* by Thomas P. Ryan
3. *Linear Models and Regression with R (an integrated approach)* by D. Sengupta and S. R. Jammalamadaka
4. *Applied Regression Analysis and Generalized Linear Models* by John Fox

Bayesian Inference

1. *Background and motivation*: What is Bayesian Inference? Why use Bayes? Bayes theorem, prior and posterior distributions. Examples.
2. *Exponential families*: One-parameter exponential families, Conjugate priors. Multi-parameter exponential families, Motivations for using exponential families. Beta-Bernoulli model, Gamma-Exponential model, Gamma-Poisson model. Normal with conjugate Normal-Gamma prior, Dirichlet-Multinomial model.
3. *Bayesian estimation*: Point estimation from posterior distribution, maximum a posteriori (MAP) estimate. Credible intervals.
4. *Bayesian hypothesis testing*: Testing hypotheses, posterior odds and Bayes factor.
5. *Choice of priors*: Conjugate priors. Non-informative prior. Jefferey's prior.
6. *Sampling from posterior distribution*: Markov chain Monte Carlo (MCMC) with Gibbs sampling, Metropolis-Hastings method. Diagnostics for convergence of MCMC schemes (emphasis on graphical summaries).
7. *Bayesian linear models*: Normal-Inverse-Gamma prior. Ridge and LASSO regression as Bayesian solutions. Spike-and-slab prior. Variable selection in linear regression.
8. *Hierarchical models*: Hierarchical Bayes and Empirical Bayes, profile likelihood.

References

1. *Bayesian Essentials with R* by J-M. Marin and C. P. Robert, Springer.
2. *Bayesian Data Analysis* by A. Gelman et al., Chapman and Hall/CRC,
3. *Likelihood and Bayesian Inference: With Applications in Biology and Medicine* by L. Held and S.-B. Daniel, Springer.
4. *An Introduction to Bayesian Analysis: Theory and Methods* by Jayanta K. Ghosh, Mohan Delampady and Tapas Samanta. Springer.

Statistical Learning I

1. *Concepts of Statistical Learning*: Prediction vs. estimation. Accuracy vs. interpretability. Supervised methods vs. unsupervised methods. Bias-variance trade-off and overfitting. Ensemble learning. Adaptive learning. Model complexity. Model selection.
2. *Classification*: Bayes rule for classification. Linear and quadratic discriminant analysis. Fisher's LDA. Multivariate logistic regression. k-nearest neighbour classifier. Decision tree classifiers including CART and Random forests. Maximum margin classifiers including Support Vector Machine (SVM) and kernel SVM.
3. *Nonparametric regression*: Kernel smoothing. Local polynomial regression. Bandwidth selection. Splines and wavelets regression with appropriate penalization schemes. Multivariate generalizations (e.g. MARS).
4. *Additive models*: Generalized additive models. Projection pursuit regression. Fitting algorithms and motivating examples.

5. *Introduction to neural networks*: Radial basis functions and single-layer neural networks. Model fitting through backpropagation.

6. *Variable selection through penalization*: Types of penalization – complexity-based, smooth, non-smooth. Soft and hard thresholding concepts. Lasso, Elastic Net and Group Lasso methods. Data-driven selection of penalty parameters.

7. *Matrix completion problem*: Motivating examples (e.g., Netflix challenge, social network recovery, linear systems identification). Compressed sensing and low-rank matrix recovery problems. Algorithms based on low rank representation and nuclear norm penalization.

8. *Tensor regression*: Motivating examples (e.g., in weather forecasting, biomedical image analysis). Tensor linear regression based on convex optimization and spectral regularization.

9. *Functional data analysis*: Curves as data points. Mean and covariance functions. Karhunen-Loève expansion. Fitting procedures for sparse and dense functional data. Brief introduction to functional linear regression.

References

1. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* by Trevor Hastie, Robert Tibshirani and Jerome Friedman

2. *An Introduction to Statistical Learning, with Applications in R* by G. James, D. Witten, T. Hastie and R. Tibshirani.

3. *An Introduction to Statistical Learning, with Applications in Python* by G. James, D. Witten, T. Hastie, R. Tibshirani and J. Taylor.

4. *Modern Multivariate Statistical Techniques* by A. J. Izenman.

5. *Linear Algebra and Optimization for Machine Learning* by C. C. Aggarwal.

6. *Linear Algebra and Learning from Data* by G. Strang.

Advanced Statistical Methods II

EM algorithm and generalizations

1. *Treatment of Missing Data*: Types of missingness and their implications. Imputation methods based on full conditional distributions. EM algorithm for computing MLE under missingness.

2. *MM algorithm for optimization*: Principles and applications to maximum likelihood estimation in generalized linear models, least squares estimation and related optimization problems.

Advanced topics in regression

1. *Introduction to Time-to-Event data*: Hazard function, failure rate, censoring. Maximum likelihood estimation for censored data. Kaplan-Meier estimator. Proportional hazard model.

2. *Penalization and regularization schemes in linear regression*: Ridge regression, Principal components regression, Partial least squares, LASSO, LARS.

3. *Robust regression*: LAD regression, quantile regression. Implementation and properties of estimators.

4. *Nonlinear regression*: Examples from compartmental model, exponential model, growth models and generalized linear models. Estimating parameters through nonlinear least squares (NLSE) methods. Optimization using Newton's and gradient descent methods. Asymptotic inference on parameters. Common issues: convergence of iterative methods, identifiability and ill-conditioning.

References

1. *The EM Algorithm and Extensions* by Geoffrey J. McLachlan and Thriyambakam Krishnan
2. *Modern Regression Methods* by Thomas P. Ryan
3. *Robust Nonlinear Regression: with Applications using R* by Hossein Riazoshams, Habshah Midi and Gebrenegus Ghilagaber
4. *Survival Analysis: A Self-Learning Text* by David G. Kleinbaum , Mitchel Klein

Signal, Image & Text Processing

Signal processing

1. Signals and their characteristics, Fourier series and Convolution, The continuous Fourier transform and its accompanying theorems, Sampling in time, aliasing, interpolation, and quantization.
2. System characteristics such as linearity, time invariance, causality and stability. Difference equations, FIR and IIR filters, basic properties of discrete-time systems.
3. The discrete-time and discrete Fourier transforms and their basic properties, elementary Z-transform, 1D correlation.

Image processing

1. From 1-D to 2-D Fourier transform. Discrete Fourier transform, Convolution and filtering, translation, scaling, derivative, rotation, and other linear operations on digital images.
2. Histogram-based image processing and thresholding.
3. Convolution kernels in the spatial and Fourier domain, low-pass, high-pass, and derivative (e.g. Sobel and Prewitt filters). Edge detection using the magnitude of the gradient.
4. Image Segmentation: detection of Discontinuities Edge linking and boundary detection Thresholding.

Text Processing

1. What is Natural Language Processing? Textual Sources and Formats 1: "What's in a Text?"
2. Tokenization, N-grams and Scriptio continua.
3. Stemming and Lemmatization, Synsets and Hypernyms.
4. Building and Tokenizing your Corpus, POS Tagging and Stopwords.
5. Text "Features" and TF-IDF Classification.
6. Named Entity Recognition (NER)
7. Topic Modeling Basics: Document Clustering and Word Vectors, Doc2vec, Word2vec.

8. Application overviews: information extraction, question answering, and machine translation.

References

1. *Digital Image Processing* by Rafael Gonzalez and Richard Woods, Pearson, 2017.
2. *Discrete-Time Signal Processing* by Alan V Oppenheim and Ronald W. Schafer, 3rd Ed.
3. *Mining Text Data* by Charu C. Aggarwal and Cheng Xiang Zhai, Springer, 2012.
4. *Natural Language Processing with Python – Analyzing Text with the Natural Language Toolkit* by Steven Bird, Ewan Klein, Edward Loper, O'Reilly 2009, website 2018.

Chapter 6

Detailed Syllabus for Semester VI

Multivariate Statistical Analysis

1. *Representation of multivariate data*: Vectors, Matrix and Tensor representation.
2. Multivariate Normal distribution and its basic properties. Distribution of linear and quadratic forms. Maximum likelihood estimation of parameters.
3. *Wishart distribution*: Definition and basic properties.
4. *Inference on mean of multivariate Normal distribution*: Two sample problem. Union-intersection test and likelihood ratio test. Hotelling's T^2 test.
5. *Multivariate regression problems and MANOVA*: Definition and properties of test statistics. Examples.
6. *Estimating covariance matrix*: Unbiased estimation. Linear and nonlinear shrinkage estimation. Estimating quadratic forms involving inverse covariance matrix. Examples.
7. *Dimension reduction methods*: Principal components analysis (PCA), Factor analysis, Canonical correlation analysis (CCA).
8. *Copula*: Specifying dependency beyond multivariate normal. Definition, basic properties and practical uses. Sklar's theorem on existence (statement only). Examples of copulas.

References

1. *Applied Multivariate Statistical Analysis* by R. Johnson and D. Wichern.
2. *Applied Multivariate Statistical Analysis* by W. Härdle and L. Simar.
3. *An Introduction to Applied Multivariate Analysis* by R. B. Everitt and T. Hothorn.
4. *Modern Multivariate Statistical Techniques* by A. J. Izenman.
5. *Multivariate Analysis* by K. V. Mardia, J. T. Kent and J. M. Bibby.
6. *An Introduction to Copulas* by Roger B. Nelsen

Discrete Data Analysis

1. *Examples of discrete data*: From observational studies (prospective, retrospective, and cross-sectional), sample surveys and designed experiments.
2. Multinomial, Poisson and Hypergeometric sampling.
3. *Contingency tables*: Joint, marginal and conditional probabilities. Independence. Relative risk. Odds ratio.
4. *Inference on Binomial and Multinomial probabilities*: Score, Wald and likelihood ratio tests.
5. *Inference for two-way contingency tables*: Pearson statistic, likelihood ratio statistic. Chi-square test of independence. Test of independence for ordinal data.
6. *Logistic regression*: Estimation procedures. Confidence intervals for effects and probabilities. Hypothesis tests. Diagnostics and model comparisons.
7. *Multicategory logit models*: Baseline-category logit model. Cumulative logit model for ordinal response.
8. *Poisson and Negative binomial regression for count data*: Specification and interpretation. Regression diagnostics and prediction. Overdispersion and zero inflated models.
9. *Log-linear models for two-way and three-way contingency tables*: Interpretation of parameters and association among categories. Model fitting and inference procedures.
10. *Bayesian approach to categorical data analysis*: Dirichlet and non-informative priors for Multinomial. Posterior distributions. Estimates of cell probabilities. Bayesian tests and credible regions for parameters in two-way contingency tables.

References

1. *An Introduction to Categorical Data Analysis* by A. Agresti.
2. *The Analysis of Cross-Classified Categorical Data* by S. A. Fienberg.
3. *Generalized Linear Models* by P. McCullagh and J. A. Nelder.
4. *Applied Linear Statistical Models* by M. H. Kutner, C. J. Nachtsheim, J. Neter and W. Li

Large Sample and Resampling Methods

Large Sample Methods

1. *Introduction*: Exact vs. approximate inference. Limitations of exact methods and the need for approximate methods. Illustrations via examples.
2. Brief review of (a) modes of convergence, (b) Laws of Large Numbers, (c) Central Limit theorem, (d) Slutsky's theorem, (e) Delta method.
3. Asymptotic distribution of functions of sample moments. The sample correlation coefficient.
4. Asymptotic distribution of Pearson's χ^2 (Chi-square) statistic.
5. *Some rank statistics*: Sign test statistic, Signed-rank test statistic, Mann-Whitney test statistic.
6. Asymptotic distribution of sample quantiles.

7. Asymptotic normality of the maximum likelihood estimate. Asymptotic distributions of likelihood ratio statistic, Rao's score statistic, Wald's test statistic.
8. Asymptotic distribution of one-sample U-statistics.

Resampling Methods

1. Importance of resampling methods. Some examples: Bias and variance estimation, estimation of distribution function, confidence interval.
2. *Jackknife in the i.i.d. case*: Bias and variance estimation, construction of large sample confidence interval. Consistency results (statements only) and discussion on inconsistency issues.
3. *Bootstrap in the i.i.d. case*: Bias and variance estimation. Estimation of distribution function and confidence interval. Parametric and nonparametric bootstrap, smooth bootstrap. Consistency results (statements only). Comparison between bootstrap approximation and Normal approximation.
4. *Bootstrap in general regression problems*: Paired bootstrap, residual bootstrap.
5. Idea of cross-validation and its use. Some examples. V-fold and leave-one-out methods.
6. Permutation tests. Demonstration through examples.

References

1. *Large Sample Techniques for Statistics* by Jiming Jiang
2. *A Course in Large Sample Theory* by Thomas S. Ferguson
3. *Bootstrap Methods and Their Application* by Anthony C. Davison and David Hinkley
4. *An Introduction to the Bootstrap* by Bradley Efron and R. J. Tibshirani
5. *U-Statistics, M_m-Estimators and Resampling* by Arup Bose and Snigdhansu Chatterjee

Time Series Analysis & Forecasting

1. *Exploratory analysis of time series*: Graphical display, classical decomposition model, concepts of trend, seasonality and cycle, estimation of trend and seasonal components.
2. *Stationary linear time series models*: Concepts of weak and strong stationarity. Autoregressive (AR), Moving Average (MA), and ARMA processes — basic properties, conditions for stationarity and invertibility. Autocorrelation function (ACF), partial auto-correlation function (PACF). Identification of time series based on ACF and PACF.
3. *Estimation of ARMA models*: AR model estimation and Yule-Walker equation. Estimation, order selection and diagnostics tests for ARMA models.
4. *Linear non-stationary model*: ARIMA model, determination of the order of integration. Trend stationarity and different stationary processes. Tests for non-stationarity, i.e., unit root tests — Dickey-Fuller test, Phillips-Perron test.
5. *Forecasting*: Exponential smoothing. Holt-Winters method. Forecasting error. Minimum MSE forecast. In-sample and out-of-sample forecast.
6. *State space models*: State space representation of ARIMA models. Kalman filtering and smoothing.

7. *Spectral analysis of weakly stationary time series*: Spectral density function (s.d.f.) and its properties. Spectral densities of AR, MA and ARMA processes. Fourier transform of time series and periodogram. Nonparametric estimation of s.d.f.

8. *Models for volatility*: ARCH and GARCH models. Identification based on PACF function. Model fitting. Forecasting volatility. Application to econometrics.

References

1. *Time Series Analysis and Its Applications with R* by R. H. Shumway and D. S. Stoffer.

2. *Introduction to Time Series and Forecasting* by P. J. Brockwell and R. A. Davis.

3. *Time Series Analysis: Forecasting and Control* by G. E. P. Box, G. M. Jenkins and G. C. Reinsel.

4. *Introductory Time Series with R* by P. S. P. Cowpertwait and A. V. Metcalfe.

5. *Forecasting: Principles and Practice* by R. J. Hyndman and G. Athanasopolous.

Statistical Learning II

1. *Concepts of unsupervised learning*: Unlabeled data – examples and challenges. Dimensionality issues and phenomena. Relevant optimization concepts.

2. *Nonlinear dimension reduction*: Concepts of data on manifolds, geodesics on manifolds. Kernel PCA. Graph-based learning methods — Local linear embedding, Laplacian eigenmaps, Isomap, Self-organizing maps. Multidimensional scaling (MDS) — motivation from visualization perspective. Metric and non-metric MDS methods.

3. *Clustering*: Measures of similarity. Traditional clustering approaches: k-means clustering, hierarchical clustering, Gaussian mixture models. Graph-based clustering: stochastic block models and community detection problems. Spectral clustering.

4. *Graphical models*: Undirected graphs and Gaussian Graphical Models (GGM). Graphical Lasso method for estimating GGM. Causal modeling and Directed Acyclic Graphs (DAG).

5. *Bayesian learning models*: Gaussian process models. Dirichlet process models. Latent Dirichlet allocation model. Examples. Computational schemes.

6. *Hidden Markov models*: Probabilistic representation. Discrete vs. continuous observation space. Viterbi and Baum-Welch algorithms for estimating states and parameters.

7. *Introduction to Reinforcement Learning*: Basic concepts. Brief introduction to Multi-arm bandit problems, Markov decision processes.

References

1. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* by Trevor Hastie, Robert Tibshirani and Jerome Friedman

2. *An Introduction to Statistical Learning, with Applications in R* by G. James, D. Witten, T. Hastie and R. Tibshirani.

3. *An Introduction to Statistical Learning, with Applications in Python* by G. James, D. Witten, T. Hastie, R. Tibshirani and J. Taylor.

4. *Modern Multivariate Statistical Techniques* by A. J. Izenman.
5. *Machine Learning: A Probabilistic Perspective* by K. P. Murphy.
6. *Linear Algebra and Optimization for Machine Learning* by C. C. Aggarwal.
7. *Linear Algebra and Learning from Data* by G. Strang.
8. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids* by Richard Durbin, Sean R. Eddy, Anders Krogh and Graeme Mitchison
9. *Reinforcement Learning: An Introduction* by R. S. Sutton and A. G. Barto.