DISCIPLINE SPECIFIC CORE COURSE – DSC - 18: STATISTICAL ANALYSIS IN PHYSICS

Course Title & Code	Credits	Credit distribution of the course			Eligibility Criteria	Pre-requisite of the course
		Lecture	Tutorial	Practical		
Statistical Analysis in Physics DSC – 18	**4**	**2**	**0**	**2**	Class XII pass with Physics and Mathematics as main subjects	Basic understanding of statistics and probability

LEARNING OBJECTIVES

This course provides an elementary introduction to the principles of Bayesian statistics and working knowledge of some of the data analysis techniques. The objective is to equip the students with certain techniques so that they may successfully apply these to the real world problems, in their research areas as well as in industry.

LEARNING OUTCOMES

After completing this course, students will be able to,

- Understand the fundamental concepts in statistical data analysis.
- Define in a Bayesian context, the likelihood, prior and posterior distributions and their role in Bayesian inference and hypothesis testing.
- Estimate the parameters of a distribution from sample.
- Perform hypothesis testing and validate a model.
- Apply multi-linear and logistic models to real life situation.

In the practical component, students will be able to

- Learn basic data analysis techniques such as linear and non-linear fittings
- Apply hypothesis testing techniques in physics
- Perform multi-linear and logistic regression analysis for a given data
- Understand the concept of gradient descent and use it for the regression analysis
- Understand the stochastic processes, Markov chains and transition probability matrix.

SYLLABUS OF DSC - 18

THEORY COMPONENT

Unit – I **(8 Hours)**

Random variables, Discrete and Continuous Probability Distributions. Bivariate and multivariate random variables, Joint Distribution Functions (with examples from Binomial, Poisson and Normal). Mean, variance and moments of a random vector, covariance and correlation matrix, eigendecomposition of the covariance matrix (bivariate problem). Cumulative Distribution Function and Quantiles. Point Estimation, Interval estimation, Central Limit Theorem (statement, consequences and limitations).

Unit – II **(11 Hours)**

Bayesian Statistics: Conditional probability and Bayes Theorem, Prior and Posterior

probability distributions, examples of Bayes theorem in everyday life. Bayesian parameter estimation. Normal, Poisson and Binomial distributions, their conjugate priors and properties. Bayes factors and model selection.

Unit – III **(11 Hours)**

Bayesian Regression: Introduction to Bayesian Linear Regression. Bayesian logistic regression and its applications. Bayesian parameter estimation for regression models. Posterior distribution of model parameters and the posterior predictive distributions.

References:

Essential Readings:

1) Schaum's Outline Series of Probability and Statistics, M. R. Spiegel, J. J. Schiler and R. A. Srinivasan, 2012, McGraw Hill Education
2) Schaum's Outline Series of Theory and Problems of Probability, Random Variables, and Random Processes, H. Hsu, 2019, McGraw Hill Education
3) Bayesian Logical Data Analysis for the Physical Sciences: A Comparative Approach with Mathematica Support, P. Gregory, 2010, Cambridge University Press
4) Linear Regression: An Introduction to Statistical Models, P. Martin, 2021, Sage Publications Ltd.
5) Data Analysis: A Bayesian Tutorial, D. S. Sivia and J. Skilling, 2006, Oxford University Press
6) Data Reduction and Error analysis for the Physical Sciences, P. R. Bevington and D. K. Robinson, 2002, McGraw-Hill Education

Additional Readings:

1) A Guide to the Use of Statistical Methods in the Physical Sciences, R. J. Barlow, 1993, Wiley Publication
2) An Introduction to Error Analysis, J. R. Taylor, 1996, Univ. Sci. Books
3) Applied Multivariate Data Analysis, Volume I: Regression and Experimental Design, J. D. Jobson, 2012, Springer-Verlag
4) Statistical Rethinking A Bayesian Course with Examples in R and STAN, Richard McElreath, 2020, CRC Press
5) Introduction to Bayesian Statistics, W. Bolstad, 2007, John Wiley

PRACTICAL COMPONENT

(15 Weeks with 4 hours of laboratory session per week)

The objective of this lab is to familiarise the students with the techniques of data analysis. The instructors are required to discuss the concepts and the pseudo-codes of the recommended programs in the practical sessions before their implementation. The implementation can be in any programming language. Inbuilt libraries can be used wherever applicable. **All units are mandatory.**

Unit 1 (12 Hours)

Probability Distributions

1) Generate sequences of N random numbers M (at least 10000) number of times from different distributions (e.g. Binomial, Poisson, Normal). Use the arithmetic mean of each random vector (of size N) and plot the distribution of the arithmetic means. Verify the Central Limit Theorem (CLT) for each distribution. Show that CLT is violated for the

Cauchy-Lorentz distribution.

2) Given a data for two independent variables (x_i, y_i). Write a code to compute the joint probability in a given sample space. Verify the same for the data generated by random number generator based on a given probability distribution of pair of independent variables (both discrete and continuous).

Unit 2 (16 Hours)

1) Hypothesis testing
   Make a random number generator to simulate the tossing of a coin *n* times with the probability for the head being *q*. Write a code for a Binomial test with the Null hypothesis H_0 ($q = 0.5$) against the alternative hypothesis H_1 ($q \neq 0.5$).

2) Bayesian Inference
   a) In an experiment of flipping a coin *N* times, *M* heads showed up (fraction of heads $f = M/N$). Write a code to determine the posterior probability, given the following prior for the probability of *f*:
      i. Beta Distribution $B(a, b)$ with given values of *a* and *b* .
      ii. Gaussian Distribution with a given mean and variance.
   b) Using the Likelihood of Binomial distribution, determine the value of *f* (fraction of heads) that maximizes the probability of the data.
   c) Plot the Likelihood (normalised), Prior and Posterior Distributions.

Unit 3 (20 hours)

Regression Analysis and Gradient Descent:

1) Given a dataset (*Xi, Yi*). Write a code to obtain the parameters of linear regression equation using the method of least squares with both constant and variable errors in the dependent variable (*Y*). The data obtained in a physics lab may be used for this purpose. Also obtain the correlation coefficient and the 90% confidence interval for the regression line. Make a scatter plot along with error bars. Also, overlay the regression line and show the confidence interval.

2) Write a code to minimize the cost function (mean squared error) in the linear regression using gradient descent (an iterative optimization algorithm, which finds the minimum of a differentiable function) with at least two independent variables. Determine the correlation matrix for the regression parameters.

3) Write a code to map a random variable *X* that can take a wide range of values to another variable *Y* with values lying in limited interval say [0, 1] using a sigmoid function (logistic function). Considering the Log Loss as the cost function of logistic regression, compute its minimum with gradient descent method and estimate the parameters.

Unit 4 (12 Hours)

Markov Chain (Any one)

1) Write a code to generate a Markov chain by defining (a finite number of) *M* (say 2) states. Encode states using a number and assign their probabilities for changing from state *i* to state *j*. Compute the transition matrix for *1, 2,..., N* steps. Following the rule, write a code for Markovian Brownian motion of a particle.

2) Given that a particle may exist in one of the given energy states (E_i, $i = 1, \ldots 4$) and the

transition probability matrix T, so that T_{ij} gives the probability for the particle to make transition from energy state E_i to state E_j. Determine the long-term probability of a particle to be in state in the state E_f if the particle was initially in state E_i.

References for laboratory work:

1) Data Science from Scratch – First Principles with Python, J. Grus, O'Reilly, 2019, Media Inc.
2) Bayes' Rule with Python: A tutorial introduction to Bayesian Analysis, J. V. Stone, 2016, Sebtel Press
3) Practical Bayesian Inference, B. Jones, 2017, Cambridge University Press
4) Modeling and Simulation in Scilab/Scicos with Scicos Lab 4.4, S. L. Campbell, Jean-P. Chancelier and R. Nikoukhah, Springer.
5) Scilab Textbook Companion for Probability And Statistics For Engineers And Scientists, S. M. Ross, 2005, Elsevier
6) Numerical Recipes: The art of scientific computing, W. H. Press, S. A. Teukolsky and W. Vetterling, 2007, Cambridge University Press