SS856 DATA ANALYTICS FOR BIOINFORMATICS: STATISTICS 3-0-0 3

COURSE OBJECTIVE

This course is not to 'just' introduce Statistics to the students. The primary objective of this course is to give a rigorous introduction to statistics, nonetheless, the secondary objective is to help them in understanding data and the art of collecting, analyzing and interpreting them appropriately in the context of modern bioinformatics-related data analytics. Specific learning objectives for this course are:

1. To provide basic knowledge on collection, characteristics and presentation of data.
2. To learn about the description of a particular scenario with the help of statistics.
3. To develop an understanding about the distribution of the individual data at an aggregate level.
4. To help students to learn the skill to find associations between several factors under concern.
5. To provide knowledge about the techniques to infer about a scenario based on the given numerical information.

Note: Each objective will have a bioinformatics related problem to get the students to start thinking about biology, and the rich set of challenges it provides.

LEARNING OUTCOMES

This course on Statistics is planned and designed to help the students to develop a better understanding of data, its characteristics, behaviour, analysis and interpretation. Statistics demonstrate real world phenomena efficiently and precisely, and bioinformatics problems are looked at where applicable to provide a gentle introduction to the students about open areas of research in bioinformatics. The knowledge of data sciences will help students to not only acquire analytical skills assessing the numbers, but also to take new directions in bioinformatics research and convincing decisions based on the analysis.

COURSE RELEVANCE

Every observation in biology is now documented, and most often numerically. Bioinformatics research is highly data driven, where Statistics help us to effectively analyze scenarios based on evidence, develop ideas, troubleshooting etc. Subjects related to bioinformatics are often complex in nature and varying in different setups. It is essential to assess the nature of intertwined factors, degree and direction of their associations to design appropriate policies which in turn ensure the

optimal output. This course will help students to not only learn the tools and techniques, but also to identify the appropriate techniques and their uses for a particular case.

TEACHING METHODS

This is a regular class where students will meet the instructor for each session covering the topics outlined below.

REQUIRED COURSE MATERIALS AND READINGS

The prescribed textbook for this course is

1. *Statistics for Management* by TN Srivastava and Shailaja Rego, Tata- McGraw Hill (3rd Edition)
2. Biostatistics: Third Revised Edition, Veer Bala Rastogi

OPTIONAL COURSE MATERIALS & READINGS (CASES, ARTICLES, REPORTS ETC)

Suggested books: *Statistical Methods* by NG Das.

EVALUATION CRITERIA

Students are evaluated based on their performance in personal learning paper; class participation and quiz; mid-term and end-term examination; and group project and presentations. The weightage for various components will be as follow:

First Internals	20%
Mid-term examination	20%
End-term examination	30%
Project on Bioinformatics	30%

DETAILS OF SESSION

Serial No.	TOPICS TO BE COVERED
1.	Presentation of Statistical Data and Frequency Distribution
2.	Central tendency
3.	Measures of dispersion
4.	Correlation
5.	Linear Regression
6.	Logistic Regression
7.	Introduction to Probability and Venn Diagram
8.	Conditional Probability
9.	Probability Distribution: Binominal Distribution
10.	Poisson Distribution & Normal Distribution
11.	Problem set based on literature in bioinformatics - I
12.	Hypothesis testing
13.	Test of Significance
14.	t-test
15.	Sampling Theory
16.	Chi-Square
17.	ANOVA
18.	Problem set based on literature in bioinformatics - II
19.	Interpolation and Extrapolation
20.	Forecasting
21.	Problem set based on literature in bioinformatics - III
22.	Time Series
23.	Non-parametric analysis: WSR test, Mann-Whitney Test, Kruskal-Wallis Test
24.	Prevalence rate, incidence rate, odds ratio, hazard ratio, relative risk
25.	Problem set based on literature in bioinformatics - IV