Statistical Foundation for Data Science – Sem 3

Unit I: Basics of Data Science & Role of Statistics

1. Define **Data Science** and explain the role of statistical foundations in solving data-driven problems.

2. What is structured thinking in data science? Why are statistics, probability, and optimization essential?

3. Explain the **typology of problems in data science**—classification, regression, clustering etc.

Unit II: Probability & Distributions

1. State the axioms of probability and differentiate between **discrete** and **continuous** random variables.

2. Explain key probability distributions (binomial, Poisson, normal)—their parameters, properties, and applications in analytics.

3. What do expectation, variance, covariance, and correlation measure? Why are they important?

4. Define PMF, PDF, and CDF. How are each of them used?

5. Discuss the **Central Limit Theorem** and its significance in sampling.

Unit III: Statistical Inference & Hypothesis Testing

1. Explain the concept of **sampling distributions** and its applications.

2. Describe hypothesis testing: state null and alternative hypotheses, errors, significance level, p-value interpretation.

3. How do you construct **confidence intervals** for means, proportions, variances, and correlation?

4. Explain hypothesis tests for means, proportions, variances, and correlation.

5. What is **A/B testing**? When can it be used, and what are its limitations? ([Scribd][1], [Reddit][2], [Reddit][3], [Towards Data Science][4])

Unit IV: Regression, Regularization & Model Selection

1. Discuss the formulation and solution of **linear regression**, including interpretation of coefficients.

2. Explain **regularization techniques** like **Ridge** and **Lasso regression**, and how they mitigate overfitting. ([KDnuggets][5])

3. What is empirical risk minimization and cross-validation? How do they help in model selection?

4. Discuss feature selection methods and overfitting vs underfitting trade-off.

Unit V: Advanced Topics & High-Dimensional Methods

1. Explain **dimensionality reduction** techniques like PCA and their use in high-dimensional data.

2. Define **Rademacher complexity**, uniform convergence, and concentration inequalities in context of learning theory. ([Scribd][1])

3. What are perceptron algorithms and linear threshold functions?

4. Discuss **Lasso vs Ridge regression**, and applications of minimax strategies in classification or portfolio optimization.

5. Overview of **stochastic gradient descent (SGD)** optimization in neural network learning.

📝 Short Answer / 2–5 Mark Questions

* Define **random variable** and list key properties.

* What is **standard deviation**? How is it different from variance?

* What is **p-value**?

* Define **confidence interval**.

* What are type I and type II errors?

* Differences between **Binomial and Poisson distributions**. ([Scribd][1], [guvi.io][6])

* Explain **descriptive vs inferential statistics**. ([GUVI][7])