National University of Computer & Emerging Sciences, Lahore
School of Computing
DS2003 - Advanced Statistics
Spring-2024

Instructor: Dr. Muhammad Ahmad Raza	**TA Name:**
Email: muhammad.ahmadraza@lhr.nu.edu.pk	**Email:**
Office Location: Office 37, 1st Floor, F Block	
Office Timings: Friday 2:00pm to 4:00pm	

Course Information

Program: BS (DS)	**Type:** Core
Credit hours: 3	**Course website:** Google Classroom
Pre-requisites: Probability and Statistics	**Class Venue:** As per timetable
Class meeting time: As per timetable	

Course Objectives

This course covers a broad range of topics related to statistics. Beginning with a recap of basic probability concepts, including the important probability distributions (Bernoulli, Binomial, Poisson, Uniform and Normal). We then hope to move forward towards the central limit theorem and then a deeper understanding of the need for confidence intervals and hypothesis testing. We will then spend the rest of the course describing different methods for hypothesis testing, including Chi-squared tests, ANOVA, linear regression models, and interpreting the correlation coefficient (R^2). Towards the end of the course, we aim to study multiple linear regression, and more advanced topics such as multicollinearity, model selection and logistic regression.

Course Learning Outcomes (CLOs)

At the end of the course students will be able to:

- Learn about basic statistical methods
- Understand the reasons for developing confidence intervals and performing hypothesis tests
- Analyze the results of different forms of hypothesis tests, and make an informed decision
- based on sample statistics.
- Build a greater understanding, theoretical foundation, and tools for applying the linear regression model. With a practical focus, he/she explores the workings of multiple regression and problems that arise in applying it, as well as going deeper into the theory of inference underlying regression and most other statistical methods. Moreover, he/she will able to understand the new types of models for binary data, emphasizing the need to fit appropriate models to the underlying processes generating the data being explained.
- Use tools such as Microsoft Excel, R or Python for basic statistical inference tasks.

Course Textbook

- OpenIntro Statistics, https://www.openintro.org/book/os/

Reference books

- TBD

Week	Topics to be covered	Assessment
1	Introduction to the course; recap of probability. Review of basic probability distributions: Bernoulli, Binomial, Negative Binomial, Geometric, Poisson, Uniform & the Standard Normal)	
2	Concept of Sampling from Population. Concept of independent and identically distributed (IID) random variables. The Law of Large Number with examples.	Assignment 01
3	Central Limit Theorem	Quiz 01
4	Confidence Intervals and Hypothesis Testing.	
5	Calculating Z-statistic, P-value & Standard Error.	Quiz 02
	Midterm 1	
6	Chi-Square Tests (Goodness of Fit, Test of Independence). 1-Sample t-test (z-test) – Hypothesis Testing and Confidence Intervals Paired t-test	
7	Difference in two means (z-test or t-test). Difference in more than two means (ANOVA test, F-stat)	Assignment 02
8	Multiple comparisons and the Bonferroni correction; pairwise comparisons	Quiz 03
9	Linear Regression Model: Line fitting, correlation, slope and intercept Conditions for the least squares line, interpretation of R^2	
10	Calculating R (correlation coefficient), the types of outliers	Quiz 04
	Midterm 2	
11	Multiple Linear Regression: calculating the intercept and slopes for various explanatory variables. Interpreting these values	
12	Multicollinearity: types, can we fix, must we fix, VIF and its interpretation. Calculating VIF, correlation matrix. Different ways to calculate R^2, Adjusted R^2.	Assignment 03
13	Model Selection (backward elimination, forward selection, using p-values, adjusted R^2)	
14	Conditions for modeling (residuals are normal, constant variability, independent, etc.) Logistic Regression (Logit function, log odds and log odds ratio, intercept and slope, interpretation).	Quiz 05
15	Logistic Regression (p-values of coefficients, confidence intervals for slope coefficients). Difference between Odds ratio and Relative Risk, Sensitivity, Specificity and Precision, Type I and Type II Errors, picking thresholds)	

Note: This weekly schedule is tentative. Any modification will be communicated.

Evaluations (Subject to change)

- Assignments (10%)
- Quizzes (15%)
- Midterm-1 and Midterm-2 Exam(s) (15+15 = 30%)
- Final Exam (45%)

Grading scheme

- The grading scheme followed will be **ABSOLUTE** in accordance with the university standards.

Passing Criteria:

- Students need to score a minimum of 50% to pass the course.

Course Policies

- Students are expected to attend all sessions. However, they might avail 20% leaves in emergency situations. Beyond this the student will not be allowed to appear in the final exam.
- Quizzes are generally unannounced; therefore, students should come prepared to class with revision of the content from the previous 2 or 3 lectures.
- Plagiarism is not tolerable in any of its form. Minimum penalty would be an 'F' grade in the course. Automated tools may be deployed to detect pirated copies. Students bear all the responsibility for protecting their assignments. In case of cheating, both parties will be considered equally responsible.
- Assignments must be submitted in time. Late submissions will face some penalty within a stipulated time period, after which they will not be accepted. REMEMBER that the overall submission time allowed includes the extra time given during which Google Classroom doesn't work. Therefore, deadlines are firm.
- Rechecking of quizzes/assignments must be done within one week of it being uploaded on FLEX. In case they are shown to you during the class, the week starts thereon.