King Fahd University of Petroleum and Minerals
Department of Mathematics
STAT-503: Probability and Statistics for Data Science (Term 242)

Instructor: Dr. Muhammad Riaz
Office: 5-332
Phone: 013-860-7622
E-mail: riazm@kfupm.edu.sa
Office Hours: will be announced later

Course Description:

STAT 503 Probability and Statistics for Data Science (3-0-3)

Description: Selected topics from Probability theory, Statistical Inference, and Information Theory for Data Science with an emphasis on the implementation using statistical software, toolboxes, and libraries like R, NumPy, SciPy, Pandas, and Statsmodels. Topics include Probability; Conditional Probability; Bayes' Theorem; Random variables; Discrete and Continuous Distributions; Central Limit Theorem; Point Estimation MLE and MAP; Confidence Interval Estimation; Hypothesis Testing; Non-parametric Statistics; Synthetic Data; Entropy, Mutual Information; Information Gain.

Prerequisite*: Graduate Standing*

Course Main Objectives:

- probability handling
- Dealing with Discrete and Continuous Distributions
- Point Estimation and Confidence Interval Estimation
- statistical inference
- Using the Non-parametric techniques
- Learning about Synthetic Data and mutual Information

Textbook: Matloff, N. Probability and Statistics for Data Science Math + R + Data, CRC Press 2019
Software Packages: R Language + RStudio

Assessment*

Activity	Weight
Classwork (quizzes, home works, assignments, attendance, bonuses, etc.)	15%
Mid Term Exam	30%
Project	15%
Final Exam (Comprehensive)	40%

Letter Grades: The letter grades will follow a relative grading scheme, which depends on the average of all students enrolled in the course.

R Language and RStudio: All R commands, procedures and packages will be explained in the class and the student are expected to practice them during and after the class.

Project Description
The project should be based on a real problem (with complete description) and a detailed analysis using the skills developed in the course. All results of the project should be made available numerically with the software/packages used in class. There should be some concluding remarks that refer to the real implications of your chosen problem. You may use online sources in your project with proper citation/reference.
Project Requirements:

- Each group should contain 5 students.
- Each group should submit the following:
  - a formal report (pdf)
  - a power point presentation
- Deadline: The end of semester (before the last day of classes)

Weekly Schedule

Week	Topics
Week 1 Jan. 12-16	**Descriptive Statistics** • Numerical Description of Data (mean, median, mode, variance, Standard deviation, quartiles. percentiles & IQR etc.) • Graphical Description of Data (Stem and Leaf, Histogram, Box Plots (Shapes of distributions and the outlier) etc.)
Week 2 Jan. 19-23	**Probability** • Sample Space and Events • Addition and Multiplication Rules • Conditional Probability and Bayes' Theorem
Week 3 Jan. 26-30	**Discrete Probability Distributions** • Probability Distributions, Probability Mass Functions, Cumulative Distribution Functions • Mean and Variance of a Discrete Random Variable • Some Selective Discrete Distributions (Binomial, Hypergeometric, Poisson etc.)
Wek 4 Feb. 02-06	**Continuous Probability Distributions** • Probability Distributions, Probability Density Functions, Cumulative Distribution Functions • Mean and Variance of a Continuous Random Variable • Some Selective Continuous Distributions (Normal, Exponential, Weibull, etc.)
Week 5 Feb.09-13	**Sampling Distribution** • Point Estimation • Sampling Distributions • Central Limit Theorem
Week 6 Feb.16-20	**Statistical Intervals for a Single Sample** • Confidence Interval for the Population Mean • Confidence Interval for the Population Proportion • Confidence Interval for the Population Variance
Week 7 Feb. 23-27	**Hypothesis Testing for a Single Sample** • Testing of the Population Mean • Testing of the Population Proportion • Testing of the Population Variance
Week 8 Mar.02-06	**Statistical Intervals for two Samples** • Confidence Interval for two Population Means • Confidence Interval for two Population Proportions • Confidence Interval for two Population Variances
Week 9 Mar. 09-13	**Hypothesis Testing for Two Samples** • Testing of two Population Means • Testing of two Population Proportions • Testing of two Population Variances
Week 10 Mar. 16-20	**Hypothesis Testing for More than Two Samples** • Testing of more than two Population Means • Developing Analysis of variance (ANOVA) technique
Week 11 Apr. 06- 10	**Some Variations of ANOVA technique** • Extending ANOVA technique • One-way and Two-way ANOVA
Week 12 Apr.13-17	**Non-parametric Statistics for Location** • Introduction to Non-Parametric tests • Some Selective Non-Parametric Tests for Location Parameter(s) (Sign, Wilcoxon, Mann-Whitney, Kruskal-Wallis etc.)
Week 13 Apr. 20-24	**Non-parametric Statistics for Variability** • Extending Non-Parametric tests for other parameters • Some Selective Non-Parametric Tests for Variability Parameter(s) (Moses, Mood, Ansari-Bradley, Fligner-Killeen etc.)
Week 14 Apr.27- May 01	**Information Theory and Data Science** • Introduction to data synthesis • Information-theoretic concepts for realistic data generation, privacy preservation, and utility maximization. • Synthetic data for training datasets
Week 15 May. 04-08	**Information Theory and Data Science** • Entropy • Mutual Information and Information Gain.
Week 16 May 11	Review / Catch up

Important Notes:

Blackboard: All contacts or announcements between the instructor and the students are supposed to be through Blackboard, so the student must check Blackboard at least once a day.

Academic Integrity: All KFUPM policies regarding ethics and academic honesty apply to this course.

Important Rules

1- Student is not allowed to enter the exam hall without either KFUPM ID cards or Saudi ID/ Iqama ID cards.
2- Students are not allowed to carry mobile phones and smart watches to the exam halls.
3- Students need to strictly adhere to the attendance policy of the university.
4- DN-Grade will be assigned to the eligible students after their instructors have warned them twice.

Cheating in Exams

Cheating or any attempt of cheating by use of illegal activities, techniques and forms of fraud will result in a grade of **DN** in the course along with reporting the incident to the higher university administration for further action. Cheating in exams includes (but is not restricted to):

➢ looking at the papers of other students

➢ talking to other students

➢ using mobiles or any other electronic devices.

Missing an Exam

In case a student misses an exam for a legitimate reason (such as medical emergencies), he/she must bring an official excuse from Students Affairs. Otherwise, he/she will get zero in the missed exam.

Attendance

- Students are expected to attend all lecture classes.
- If a student misses a class, he is responsible for any announcement made in that class.
- Attendance on time is very important. Mostly, attendance will be checked within the first five minutes of the class. Entering the class after that, is considered as one late, and every two times late equals to one absence.
- A DN grade will be awarded to any student who accumulates more than 20% unexcused absences or 33.3% excused and unexcused absences.

The usage of mobile phones and apple watches

- Students are not allowed to use mobiles for any purpose during class time unless given permission.
- Violations of these rules will result in a penalty decided by the instructor.
- Academic Integrity: All KFUPM policies regarding ethics apply to this course. See the Undergraduate Bulletin.