CMU ONLINE
Carnegie Mellon University

Syllabus Overview Sample: Probability & Statistics for Data Science

Program: Applied Data Science

Faculty: Dr. Weijing Tang, Assistant Professor, Department of Statistics & Data Science

Course Number: 36640 **Units:** 6

Zoom Class: Weekday Evening, ET

Course Description

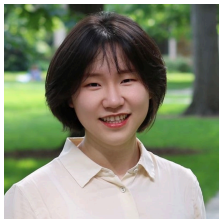
Data science is where mathematics, computation, and topic knowledge come together. It encompasses the methods and tools that inform how one collects, processes, and analyzes data and uses them to generate actionable results. This course is designed to teach you how to use probability laws and statistical methods to unlock the inferential power of data. You will learn the laws of probability and how we represent the probabilities of experimental outcomes, or random variables, using mathematical functions. You will then gain hands-on experience in using functions of observed data, or statistics, along the mathematical functions of probability, to uncover truths about data-generating processes.

Course Learning Objectives

By the end of this course, students will be able to:

- Describe fundamental concepts of probability as it is applied within data science–sample spaces, random variables, and probability distributions.
- Explain to others how probability concepts are combined with statistics to make inferences about populations from which data are sampled.
- Make statistical inferences via point estimation, confidence intervals, and hypothesis testing.
- Evaluate analytical choices made by others, taking into account the underlying assumptions and the limitations of various statistical modeling techniques.
- Adapt and use R code to perform data analyses.

Faculty Introduction



Dr. Weijing Tang is an Assistant Professor in the Department of Statistics and Data Science at Carnegie Mellon University. She earned her bachelor's degree in mathematics from Tsinghua University and her Ph.D. in statistics from the University of Michigan. Before joining Carnegie Mellon, she was a postdoctoral research fellow in biostatistics at Harvard University. Her research focuses on developing statistical methods and theory for data valuation in data-centric AI, multi-source statistical learning, network analysis, and survival analysis, with applications in the health and social sciences. In this course, she aims to help students integrate and apply the skills they have developed throughout the program to a real-world capstone project and communicate their methods, findings, and conclusions clearly to diverse audiences.

Course Topics

Topic	Learning Objectives
Probability	• Explain how probability and statistical inference fit within the overall context of data science. • Describe the outcomes of experiments via the definition of sample spaces. • Describe how we use probability axioms and relationships to quantify the relative frequencies of events in a sample space.
Conditional Probability and Random Variables	• Explain how the application of conditions impacts probability calculations. • Apply, e.g., the additive law and Bayes' rule when solving problems involving probabilities. • Explain how the concept of random variables fits within the overall context of probability.
Probability Distributions	• Represent the relative frequencies of experimental outcomes using probability distributions. • Explain how parameters affect the shape and/or domain of probability distributions and what their roles are in statistical inference. • Derive distribution properties such as the mean, the variance, and the standard deviation and explain what they represent. • Derive cumulative distribution functions and explain what they represent and how they are used in probability calculations.
Parameter Estimation	• Explain how statistics help us learn the properties of populations. • Show how we use likelihood functions to estimate distribution parameters, given observed data. • Explain how we characterize the performance of parameter

	estimators.
Confidence Intervals and Hypothesis Tests	● Explain the relationship between statistics and their sampling distributions, and confidence interval estimation and hypothesis testing. ● Derive and interpret confidence intervals for distribution parameters. ● Test preconceived hypotheses about distribution parameters and interpret the results. ● Explain the role of that the Central Limit Theorem plays in statistical inference.
Multivariate Distributions	● Explain how multivariate probability distributions describe the concurrent sampling of groups of random variables. ● Evaluate the level of covariance and correlation between any pair of random variables. ● Derive, and describe the differences between, marginal and conditional distributions.
Regression	● Explain how one uses statistical models to represent data-generating processes. ● Recognize how different data settings affect model selection. ● Explain in detail four commonly used models: linear regression, analysis of variance, the generalized linear model, and the Naive Bayes classifier.

NOTE: This is a sample designed to provide an overview of topics covered in the course based on previous course syllabi. All students will be provided an updated and final syllabus at the start of the course. Faculty, content, and schedule are all subject to change.