Katz
Katz School
of Science and Health

M.S. in Applied Statistics
Course Descriptions

Computational Statistics and Probability
Arguably, most of data science is statistical learning, which requires strong foundational knowledge in probability and statistics. And applying computational methods such as direct simulation, shuffling, bootstrapping, and cross-validation to statistical problems is often more intuitive, and intuitive and can provide solutions where analytical methods would prove computationally intractable. This course introduces students to the statistical analysis of data using modern computational methods and software. Probability, descriptive statistics, inferential statistics and computation methods such as simulations sample distributions, shuffling, bootstrapping, and cross-validation will be covered.

Multivariate Analysis
This is a course focusing on multivariate statistical methods. High dimensional statistics is one of the most important advanced field of statistics that is used extensively by practitioners and is not covered by the undergraduate curriculum. The course topics include advanced linear algebra, multivariate normal distribution and derived distributions, ANOVA and MANOVA, Principal Component Analysis, Factor Analysis and canonical correlations.

Non-parametric Statistical Learning
This course is an introduction to two important modern fields of statistics: (1) nonparametric statistics and (2) modern statistical learning theory – the theoretical foundation of machine learning. Topics include (A) the nonparametric estimation of the cumulative distribution function (Glivenko-Cantelli, Kolmogorov-Smirnov and Dvoretzky-Kiefer Wolfowitz theorems), (B) nonparametric estimation of probability density functions (kernel methods), (C) nonparametric regression, (D) Bias-variance tradeoff and double descent phenomena, (E) VC-dimension and Rademacher complexity.

Data Acquisition and Management
Data Acquisition and Management focuses on the data structures, data design patterns, algorithms, methods, and best practices for the pre-modeling phases of data science workflows, including problem formulation, gather, analyze, explore, model, and communicate, analytics programming focuses on the gather, analyze, and explore workflow steps. This comprises the “data wrangling” work which is where most data scientists spend the majority of their time. Because data science is iterative, this preparatory work informs the modeling phase. Often, the creation and validation of new models requires going back for additional data, different data transformations, and exploration of data distributions. In short, every effective data scientist needs to master analytics programming. Course topics include reading from or writing to databases, text files, and the web; shaping data into “tidy” data frames, exploratory data analysis, data imputations, feature engineering, and feature scaling.

Capstone in Applied Statistics
Students will demonstrate a holistic mastery of theory and practices through in two ways: 1) an applied project, thesis, or equivalent graduate-level activity approved by the program director and supervised by a faculty member. Examples include developing a statistical application or methodology, publishing a research paper at a peer-reviewed conference, or creating a startup company through YU’s Innovation lab—though students may propose other related work based on their interests. 2) a portfolio of work from three previous courses that can be shown to potential employers. The Capstone will include four components: a brief proposal and project schedule; the main deliverable; a portfolio of previous work; and a final presentation. Faculty will provide students with mentorship and feedback at each stage of their project or thesis.

Mathematics of Finance
This course provides a rigorous introduction to the mathematical foundations of finance. The following fundamental topics will be covered: risk, arbitrage, mathematical models for asset price movements (based on trees, PDEs, and martingales), pricing of financial derivatives, and hedging. It will also provide an elementary introduction to stochastic calculus, and to the Black-Scholes model. Computer simulations are an integral part of the course.

Time Series Analysis
This course provides a rigorous introduction to modeling and prediction of financial time series. Course topics include basic concepts of linear time series analysis such as stationarity and autocorrelation function, regression models with time series errors, seasonality, unit-root non-stationarity, and long-memory processes; analysis in the presence of conditional heteroscedasticity and serial correlations of asset returns; heavy-tailed distributions, and their application to financial risk management, in

particular, modern valuations of credit risk; multivariate time series analysis; co-integration and arbitrage opportunity in pairs trading. The course places great emphasis on empirical data analysis. It uses real examples and software applications. The course aims to broaden the horizons of students in applied mathematics and to provide conceptual background to students who are interested in a career in financial industry.

Stochastic Processes
This is an introductory course to stochastic processes. The course serves two purposes: (1) to discussing the most fundamental topics in the theory of stochastic processes, (2) to prepare students for quantitative job interviews by solving a big number of brain teasers and applied interview questions in stochastic processes.

Stochastic Calculus
This in an introductory course to stochastic calculus: a prime example of how probability and statistics are applied in financial mathematics. Topics include the rigorous mathematical treatment of Brownian motion, stochastic integrals, stochastic differential equations, with the most important examples, such as the Langevin equation and Black-Scholes SDE together with its application in pricing European options.

Introduction to Biostatistics
Biostatistics is one of the most important and quickly growing areas of applied statistics. The Introduction to Biostatistics course introduces the basic concepts of biostatistics with a special emphasis in practical applications while substantially expanding the statistical methods from the class AS X02. Topics include simple and multivariateregression, Analysis of variances, likelihood ration test, logistic regression, mediation analysis, missing data handling.

Advanced Biostatistics
Advanced biostatistics is key for medical and epidemiologic research. In this class, students will be exposed more indepth to key procedures of biostatistical methods that are essential for the daily life of a biostatistician. Topics range from pure asymptotic statistical subjects such as survival analysis, Kaplan-Meier estimators through Cox hazard model to practical hands-on experience involving Poisson regression, longitudinal methods and random effects model.

Statistics in Trials
Designing and conducting clinical trials requires a multidisciplinary skillset, ranging from biostatistical theory and practices to a good understanding of the ethics of trials, and the legal jargon of the regulations. This class offers comprehensive overview to this multidisciplinary landscape. Students will have also the opportunity to liaise with faculty and students in the Katz School’s Biotechnological Management and Entrepreneurship MS program.

Bioinformatics

Modern Biology is data driven and bioinformatics is the combination of Biology and computing. Bioinformatics uses computational tools to analyze and interpret large amounts of data. This course will introduce you to some of the powerful tools currently used to study problems that include large scale genetic sequencing, analysis of query sequences to find similarities, protein structure prediction, and machine learning methods. This course exposes you to the interdisciplinary nature of Bioinformatics and serves as an introduction to the computational tools frequently used for the practice of modern biology and medicine. The laboratory component of this course will introduce students to basic web-based tools (BLAST for searching for similar sequences in proteins and nucleic acids, access databases for protein structures and genes, application of visualization programs like PyMOL to display molecular structures, to name a few), clustering algorithms, machine learning, logistic regression, decision trees, a basic introduction to computer programming languages like Python, and R to analyze and visualize data, write simple Python scripts as well as the Unix shell scripts to drive the bioinformatics tools. The course will also train students to read and interpret primary literature articles, identify a research question, write research reports, and make oral presentations.

Machine Learning

In classical programming, answers are obtained from rules and data. In machine learning, rules are obtained from data and answers. The widespread availability and sharing of data, and improvements in computing capacity, processing methods, and algorithms have given machine learning the power to deliver game-changing systems and technologies to organizations that compete on predictive, prescriptive, and/or autonomous analytics. In this course, we'll look at methods for using, tuning, and comparing machine learning algorithms, based on measures of performance, accuracy, and explainability. We'll also look at recent advances and trends in automated machine learning.

Predictive Models

Predictive modeling answers the question, “What will happen next?” Linear regression and logistic regression are foundational predictive modeling methods, used to predict continuous and categorical output respectively. The main topics covered in this course include simple and multiple linear regression, variable selection and shrinkage methods, binary logistic regression, count regression, weighted least squares, robust regression, generalized least squares, multinomial logistic regression, generalized linear models, panel regression, and nonparametric regression.

Bayesian Methods
Bayesian inference provides powerful tools to model random variables. While Bayesian methods often yield the most accurate theoretical results, historically analytical complexity made it challenging to apply Bayesian methods against less trivial problems. Now, the confluence of more powerful computing resources and improved computational algorithms make Bayesian methods the best choice for tackling some of the most complex data science problems. Bayesian analysis is increasingly important in academic research, and research and is the preferred standard statistical analysis tool in data science practice. In this course, we'll build from Bayes' probability foundations to first applying Bayesian methods to infer binomial probabilities, then hierarchical models, and finally generalized linear models. We'll provide comparisons between frequentist approaches and Bayesian approaches. We'll build basic algorithms from scratch, as well as using high-performance Markov Chain Monte Carlo (MCMC) methods.

Internship in Applied Statistics
This course allows students to participate in an off-campus internship supervised by 1.) a staff person at the internship site and 2.) overseen by a faculty advisor. The internship site must be approved by the program director and the overall duration of student work must be no less than 150 hours. At the start of the internship, the student and faculty advisor will jointly develop specific learning objectives tailored to the nature of the internship. Over the course of the internship, students will be required to submit weekly reflections, and at the end of the internship, students write a final paper that represents the culmination of the work.

Special Topics in Applied Statistics
This course provides the opportunity to offer special interest courses on emerging theory, phenomena, and technologies in the field of applied statistics, stochastic processes, and/or in modern machine learning, data science, and big data generally. This will be an advanced class, whether seminar style or project based. Students are required to complete an appropriate project or other deliverable in line with the number of credits awarded for the course.

Independent Study in Applied Statistics
The course provides the student with the flexibility to learn more about a topic of interest outside of the formal course setting. The subject should be chosen in consultation with a faculty advisor who acts as the student's supervisor, and with the permission of the program director. The student is required to submit a course contract describing the course of study and its specific learning objectives. Course credit is determined in advance by the instructor with the approval of the program director.