Statistics And Data Science (M.S)

About The Program:

The Department of Statistical Science offers the Statistics and Data Science M.S. degree program to allow students to acquire in-depth knowledge of statistics. The program prepares “super analysts” who have stronger and deeper predictive analytics skills. Students learn to:

1. Understand descriptive, predictive, and prescriptive analytics; recognize when each is appropriate; and understand the fundamental methods and techniques used in each.
2. Understand the applications, assumptions, and limitations of data science models and techniques.
3. Develop expertise in the use and application of cutting-edge analytical tools and software.
4. Obtain information and insight from highly complex, large datasets.
5. Effectively communicate insights and results of analyses in managerially relevant terms.

Career Options: Graduates of the master's program are trained to serve as professional statisticians in industry, research organizations, or government.

Prerequisites for Admission: Students must have completed a minimum of two semesters of university-level Calculus or its equivalent. Adequate preparation in Linear Algebra and in Programming are also required. Prerequisites include:

CIS 5011	Programming and Data Structure (or equivalent)
MATH 5001	Linear Algebra (or equivalent)

Requirements of Program:

Total Credit Hours: 30
Culminating Event: Satisfactory completion of coursework is required to earn the degree.

Core Courses
Visualization: The Art of Numbers and the Psychology of Persuasion - Organizations are collecting an unprecedented volume of data, and analysts are producing information from data using analytics and models. None of the information that is

extracted from the data is usable unless it can be effectively communicated. In this course, we will begin with the fundamental questions of communication: Who is the audience? What is the information? What is the goal? Using these questions to focus our thoughts, we will explore the techniques that allow you to select appropriate information and to craft a narrative that clearly and effectively communicates this information using visual elements. Producing good visual displays is a combination of art and science and compromise between function and form. We will discuss how humans process and encode visual and textual information in relation to selecting an appropriate visual display, and we will cover topics including: exploratory data analyses, charts, tables, graphics, static and dynamic displays, effective presentations, multimedia content, animation, and dashboard design. Examples and cases will be used from a variety of industries.

Statistical Methods I - Introduction to applied statistics. Topics include data management, probability distributions, parameter estimation, hypothesis testing, sampling methodologies, graphical display, analysis of variance, and simple and multiple regression. Use of R, S-Plus and SAS statistical software.

Statistical Methods III

Applied Statistics and Data Science - PART I: Elements of a scientific problem, including estimands, the role of statistical models, the language of statistical modeling, notions of likelihood, finite vs infinite populations, and types of analysis. PART II: Elements of statistical modeling, including transformation theorems, sufficiency, 1-parameter and multi-parameter models, multivariate Normal models, Dirichlet-multinomial models, hierarchical models, generalized linear models, mixture models, text analysis, social network analysis. PART III: Concepts and algorithms for estimation and inference, including information, statistical efficiency, asymptotic approximations, maximum likelihood estimators, method of moments estimators, Bayesian estimators, empirical Bayes vs full Bayes estimation strategies, expectation-maximization algorithm, Monte Carlo approximations, Gibbs samplers, Metropolis-Hastings samplers, prior and posterior predictive checks, and Bayesian vs. frequentist coverage. Data Science visitors: The course will feature a series of short talks and Q&A sessions with prominent data scientists spanning academia, government, and the Tech industry.

Statistical Computing - Use of computers in the solution of statistical problems. Topics include: floating point architecture, random number generation, design of

statistical software, computational linear algebra, numerical integration, optimization methods.

Design and Analysis of Complex Experiments - Experimentation has become a strategic foundation for modern corporations. This course focuses on strategies for designing, executing, and analyzing experiments on large populations, as well as dealing with large collections of concurrent experiments. We will consider settings where units of analysis may be assigned treatment in multiple experiments that are running concurrently. The course is largely based on critical reading of recent articles and technical reports. The emphasis will be on the design of complex field experiments that produce actionable results, and in working with experimentation platforms at large IT companies.

Data Wrangling and Curation - This course will explore advanced tools and techniques for cleaning "raw" data. Real data is ubiquitous, but it is almost always "dirty". Analysis of "dirty" data can significantly distort results, which can reduce or eliminate the benefits from an analytic solution. The first step in extracting actionable information from data is to "clean" the data, and this process frequently occupies the majority of the analysis time. In this course, we will provide an in-depth look at the techniques that can be used to identify and deal with problematic data. Even after data have been "scrubbed", datasets are frequently not in the correct configuration for analysis, and we will explore techniques for merging and summarizing datasets, extracting subsets of data, and transforming variables. We will also discuss procedures for ensuring data quality and reliability in data collection.

OR

Principles of Data Management - This course covers fundamental concepts in constructing database management systems, including relational query languages, such as SQL and relational algebra, file organizations, storage management, system architectures, query processing, query optimization, transaction management, recovery, and concurrency control. Additional topics may include distributed databases, NoSQL databases and data integration.

Electives

Select two from following

Actuarial Modeling I - The heart of the traditional actuarial science curriculum, this

course examines the pricing of life insurance products by integrating concepts from probability and interest theory. It introduces random variables measuring the future lifetime of a person and the present values of life insurance and life annuity products, in both discrete-time and continuous-time settings. Students learn to calculate and interpret the mean, variance and probability functions for these random variables. In addition, students learn to determine actuarially fair premiums for life insurance and life annuity products as well as other life-contingent payments.

Actuarial Modeling II - This is a continuation of AS 5102. Students learn to calculate reserves for life insurance and life annuity products. The course then introduces Markov Chains in order to extend the pricing and reserving concepts to multiple lives (e.g. life insurance for a married couple or business partners) and multiple decrements (e.g. modeling different health statuses). The course also includes actuarial applications to pension valuation and profit testing.

Actuarial Modeling III - Focusing on short-term actuarial modeling, the course introduces a variety of frequency, severity, and aggregate loss models. Students learn to select suitable models for a given data set, to parameterize the models to the data, to assess the predictive quality of the models through various measures of confidence, and to estimate losses using credibility theory.

Actuarial Economics - This course develops the conceptual framework of microeconomics and macroeconomics with some applications in actuarial science. Topics in microeconomics include interaction between supply and demand, consumer behavior, production choices, different types of competition, factor markets, and market failure. Topics in macroeconomics include business cycles, inflation, unemployment, monetary and fiscal policy, balance of payments, international economics, and economic growth.

Knowledge Discovery and Data Mining - Basic concepts and techniques for the automated extraction of interesting patterns in large databases. Topics covered include: association-rule mining, sequence mining, web and text mining, data warehousing, information filtering, classification and clustering analysis, Bayesian and neural networks, classification and regression trees, hypotheses evaluation, feature extraction, dimensionality reduction, singular value decomposition, data compression and reconstruction, visualization of large data sets, fractals in databases, and indexing methods that support efficient data mining and queries by content. Special emphasis is given in multimedia, business, scientific, and medical

databases. Note: Students may not receive credit for both CIS 5523 and CIS 4523.

Neural Computation - Neural networks provide powerful techniques to model and control nonlinear and complex systems. The course is designed to provide an introduction to this interdisciplinary topic. The course is structured such that students from computer science, engineering, physics, mathematics, statistics, cognitive sciences, and other disciplines learn the main principles of this area as well as have an opportunity to explore promising research topics through hands-on experience with neural network simulators applied to classification and prediction problems ranging from biomedical sciences to finance and business.

Machine Learning - The goal of the field of machine learning is to build computer systems that learn from experience and are able to adapt to their environments. This introductory machine learning course will present modern machine learning algorithms for supervised and unsupervised learning. It will provide the basic intuition behind the algorithms as well as a more formal understanding of how and why they work. Students will learn how to apply machine learning algorithms on a range of real-life problems and how to evaluate their performance.

Text Mining and Language Processing - This course will cover a broad overview of problems and techniques in text mining and natural language processing. It will also provide in-depth coverage of the latest natural language processing research in selected topics. The in-depth part of the course will focus on the latest research in unsupervised information extraction. This part of the course will cover such techniques as pointwise mutual information, pattern-matching, bootstrapping, Hidden Markov Models, Conditional Random Fields, and language modeling techniques, among others.

Advanced Statistics for Urban Applications - This course teaches advanced statistical methods to examine urban processes and patterns. The course covers spatial point pattern analysis, multivariate regression, logit and probit regression, spatial econometrics, Geographically Weighted Regression (GWR), and hierarchical linear modeling.

Application Development for Geographic Information Systems - This course introduces students to geospatial software development through the creation of standalone software applications and plugins that add new functionality to major GIS software products. The emphasis will be on geospatial algorithms and object-

oriented programming. Other topics in software design will be addressed including documentation, version control, user interface design, software testing, and software project management. The course will be taught using Python, JavaScript, or another major programming language with strong geospatial support. There will be a heavy emphasis on Free and Open Source Software, and active participation in the developer community outside of the classroom. At the end of the course students will have produced fully functioning geospatial software, shared their code to a public online repository, generated documentation, and promoted their work publicly.

Statistical Methods II - Design of experiments, analysis of discrete data, introduction to nonparametric methods, logistic regression, ARIMA time series analysis, bootstrapping, jackknife, robustness, and selected topics in multivariate analysis. Use of R, S-Plus and SAS statistical software.

Survey Techniques for Business Applications - Application oriented. A course dealing with statistical and nonstatistical aspects of organizing a sample survey. Included are discussions of objectives, measurement, sample selection, pilot testing, data collection, data editing, summarization and interpretation of results in addition to describing the various sampling schemes. Students may be required to plan and execute a survey.

Nonparametric Methods - A thorough course in nonparametric statistics. Estimation and testing of hypothesis when the function form of the population distribution function is not completely specified.

Categorical Data Analysis - Sampling models and analyses for discrete data: Fisher's exact test; Logistic regression; ROC analysis; Log-linear models and Poisson regression; Conditional logistic regression; Cochran-Mantel-Haenszel test; Measures of agreement between observers; Quasi-independence; Multinomial logit models; Proportional odds model; Association models; generalized estimating equations (GEE); generalized linear mixed model (GLIMMIX); GSK models; Composite link functions. The students will need to work with R and SAS throughout the semester.

Time Series Analysis and Forecasting

Independent Study - Special study in statistics theory and methods under the supervision of a graduate faculty member.

Capstone Course

Capstone in Analytics - The capstone in analytics is the culmination of analytics-focused coursework. You will work with real data from "live" clients. Some of you will work on projects at companies for which you are interning. Others will work with MBA teams as part of our Fox Management Consulting program, providing analytics support for a live client. Others will work on primarily analytics focused projects.

Courses:

Click HERE for more information on the courses below.

Quantitative Methods for Business
Introduction to Biostatistics
Special Topics
Independent Study
Visualization: The Art of Numbers and the Psychology of Persuasion
Statistical Learning and Data Mining
Experiments: Knowledge by Design
Decision Models: From Data to Decisions
Data: Care, Feeding, and Cleaning
Advanced Business Analytics
Business Analytics II
Capstone in Analytics
Statistical Analysis for Management
Quantitative Techniques for Management
Probability and Statistics Theory I
Probability and Statistics Theory II
Statistical Methods I
Statistical Methods II
Probability and Large Sample Theory
Stochastic Processes
Statistical Methods III
Sampling Theory
Mathematics for Stat
Univariate Time Series Analysis
Linear Models I
Design of Experiments I
Applied Multivariate Analysis I
Applied Statistics and Data Science

Survey Techniques for Business Applications
Statistical Methods for Business Research I
Statistical Methods for Business Research II
Survival Analysis I
Nonparametric Methods
Categorical Data Analysis
Clinical Trials
Statistical Computing
Advanced SAS Programming
Time Series Analysis and Forecasting
Design and Analysis of Complex Experiments
Data Wrangling and Curation
Independent Study
Advanced Statistical Inference I
Advanced Statistical Inference II
Multivariate Time Series Analysis
Stat Lrng & Data Mining
Linear Models II
Design of Experiments II
Multivariate Analysis II
Survival Analysis II
Statistical Genetics: An Advanced Graduate Course
Seminar in New Topics in Statistics
Directed Study in Statistics
Preliminary Examination Preparation
Pre-Dissertation Research
Dissertation Research