Statistics and Data Science

Contact:	José E. Figueroa-López
Email:	sdsadvising@wustl.edu
Website:	https://sds.wustl.edu/

Courses

SDS 5000 Independent Work
Independent Work for Credit. Graduate standing (or, for advanced undergraduates permission of the Department's Director of Undergraduate Studies)
Credit 6 units.
Typical periods offered: Fall, Spring

SDS 5007 Statistics for Medical and Public Health Researchers
This course is an introduction to basic statistical analysis for graduate students in medicine, biology, and public health. Students will be introduced to core statistical tools used to study human health outcomes. Topics include: measurement, descriptive analysis, correlation, graphical analysis, hypothesis testing, confidence intervals, analysis of variance, and regression analysis. Major components of the course include learning how to collect, manage, and analyze data using computer software, and how to effectively communicate to others results from statistical analyses. The second aspect of the course is focused on the statistical package R, which is is the most powerful, extensively featured, and capable statistical computing tool available. Course may not be used for credit in undergraduate math major/minor programs, nor in any Mathematics or Statistics graduate programs.
Credit 3 units.
Typical periods offered: Fall

SDS 5010 Probability
Mathematical theory and application of classical probability at the advanced level; a calculus based introduction to probability theory. Topics include the computational basics of probability theory, combinatorial methods, conditional probability including Bayes' theorem, random variables and distributions, expectations and moments, the classical distributions, and the central limit theorem. Prerequisites: Mutivariate Calculus (Math 233); a course in linear algebra at the level of Math 309 or Math 429. Some knowledge of basic ideas from analysis (e.g. Math 4111) will be helpful: consult with instructor.
Credit 3 units. Art: NSM
Typical periods offered: Fall

SDS 5020 Mathematical Statistics
Theory of estimation, minimum variance and unbiased estimators, maximum likelihood theory, Bayesian estimation, prior and posterior distributions, confidence intervals for general estimators, standard estimators and distributions such as the Student-t and F-distribution from a more advanced viewpoint, hypothesis testing, the Neymann-Pearson Lemma (about best possible tests), linear models, and other topics as time permits.
Credit 3 units. A&S IQ: NSM Art: NSM
Typical periods offered: Spring

SDS 5061 Theory of Statistics I
An introductory graduate level course. Probability spaces; derivation and transformation of probability distributions; generating functions and characteristic functions; law of large numbers, central limit theorem; exponential family; sufficiency, uniformly minimum variance unbiased estimators, Rao-Blackwell theorem, information inequality; maximum likelihood estimation; estimating equation; Bayesian estimation; minimax estimation; basics of decision theory.
Credit 3 units.
Typical periods offered: Fall

SDS 5062 Theory of Statistics II
Continuation of Math/SDS 5061.
Credit 3 units.
Typical periods offered: Spring

SDS 5070 Stochastic Processes
Content varies with each offering of the course. Past offerings have included such topics as random walks, Markov chains, Gaussian processes, empirical processes, Markov jump processes, and a short introduction to martingales, Brownian motion and stochastic integrals.
Credit 3 units. A&S IQ: NSM Art: NSM
Typical periods offered: Spring

SDS 5071 Advanced Linear Models I
Theory and practice of linear regression, analysis of variance (ANOVA) and their extensions, including testing, estimation, confidence interval procedures, modeling, regression diagnostics and plots, polynomial regression, collinearity and confounding, and model selection. The theory will be approached mainly from the frequentist perspective and use of statistical software (mostly R) to analyze data will emphasized.
Credit 3 units.
Typical periods offered: Fall

SDS 5072 Advanced Linear Models II
Generalized linear models including logistic and Poisson regression (hetrogeneous vairance structure, quasi-likelihood), linear mixed-effects models (estimation of variance components, maximum likelihood estimation, restricted maximum likelihood, generalized estimating equations), generalized linear mixed-effects models for discrete data, models for longitudinal data, and optional multivariate models as time permits. The computer software R will be used for examples and homework problems. Implementation in SAS will be mentioned for several specialized models.
Credit 3 units.
Typical periods offered: Spring

SDS 5111 Experimental Design
A first course in the design and analysis of experiments, from the point of view of regression. Factorial, randomized block, split-plot, Latin square, and similar design.
Credit 3 units. A&S IQ: NSM Art: NSM
Typical periods offered: Fall

SDS 5120 Survival Analysis
Life table analysis and testing, mortality and failure rates, Kaplan-Meier or product-limit estimators, hypothesis testing and estimation in the presence of random arrivals and departures, and the Cox proportional hazards model. Techniques of survival analysis are used in medical research, industrial planning and the insurance industry.
Credit 3 units. A&S IQ: NSM
Typical periods offered: Fall

SDS 5130 Linear Statistical Models
Theory and practice of linear regression, analysis of variance (ANOVA) and their extensions, including testing, estimation, confidence interval procedures, modeling, regression diagnostics and plots, polynomial regression, colinearity and confounding, model selection, geometry of least squares, etc. The theory will be approached mainly from the frequentist perspective and use of the computer (mostly R) to analyze data will be emphasized.
Credit 3 units. A&S IQ: NSM Art: NSM
Typical periods offered: Spring

SDS 5140 Advanced Linear Statistical Models
Review of basic linear models relevant for the course; generalized linear models including logistic and Poisson regression (heterogeneous variance structure, quasilikelihood); linear mixed-effects models (estimation of variance components, maximum likelihood estimation, restricted maximum likelihood, generalized estimating equations), generalized linear mixed-effects models for discrete data, models for longitudinal data, optional multivariate models as time permits. The computer software R will be used for examples and homework problems. Implementation in SAS will be mentioned for several specialized models.
Credit 3 units. A&S IQ: NSM
Typical periods offered: Spring

SDS 5155 Time Series Analysis
Time series data types; autocorrelation; stationarity and nonstationarity; autoregressive moving average models; model selection methods; bootstrap condence intervals; trend and seasonality; forecasting; nonlinear time series; filtering and smoothing; autoregressive conditional heteroscedasticity models; multivariate time series; vector autoregression; frequency domain; spectral density; state-space models; Kalman filter. Emphasis on real-world applications and data analysis using statistical software.
Credit 3 units. A&S IQ: NSM Art: NSM
Typical periods offered: Fall

SDS 5210 Statistical Computation
Introduction to modern computational statistics. Pseudo-random number generators; inverse transform and rejection sampling. Monte Carlo approximation. Nonparametric bootstrap procedures for bias and variance estimation; bootstrap confidence intervals. Markov chain Monte Carlo methods; Gibbs and Metropolis-Hastings sampling; tuning and convergence diagnostics. Cross-validation. Time permitting, optional topics include numerical analysis in R, density estimation, permutation tests, subsampling, and graphical models. Prior knowledge of R at the level used in Math 494 is required.
Credit 3 units. A&S IQ: NSM Art: NSM
Typical periods offered: Fall

SDS 5212 Statistics for Data Science I
This course starts with an introduction to R that will be used to study and explore various features of data sets and summarize important features using R graphical tools. It also aims to provide theoretical tools to understand randomness through elementary probability and probability laws governing random variables and their interactions. It integrates analytical and computational tools to investigate statistical distributional properties of complex functions of data. The course lays the foundation for statistical inference and covers important estimation techniques and their properties. It also provides an introduction to more complex statistical inference concepts involving testing of hypotheses and interval estimation. Required for students pursuing a major in Data Science. No prior knowledge of Statistics is required. NOTE: Math/SDS 3211 and Math/SDS 3200 can not both count towards any major or minor in the Statistics and Data Science Department.
Credit 3 units. A&S IQ: NSM, AN Art: NSM
Typical periods offered: Fall, Spring

SDS 5310 Bayesian Statistics
Introduces the Bayesian approach to statistical inference for data analysis in a variety of applications. Topics include: comparison of Bayesian and frequentist methods, Bayesian model specification, choice of priors, computational methods such as rejection sampling, and stochastic simulation (Markov chain Monte Carlo), empirical Bayes method, hands-on Bayesian data analysis using appropriate software.
Credit 3 units. A&S IQ: NSM
Typical periods offered: Spring

SDS 5430 Multivariate Statistical Analysis
A modern course in multivariate statistics. Elements of classical multivariate analysis as needed, including multivariate normal and Wishart distributions. Clustering; principal component analysis. Model selection and evaluation; prediction error; variable selection; stepwise regression; regularized regression. Cross-validation. Classification; linear discriminant analysis. Tree-based methods. Time permitting, optional topics may include nonparametric density estimation, multivariate regression, support vector machines, and random forests. Prerequisite: CSE 131; Math 233; Math 309 or Math 429; multivariable-calculus-based probability and mathematical statistics (Math/SDS 493-494 or Math/SDS 3211/4211); Math/SDS 439. Prior knowledge of R at the level introduced in Math/SDS 439 is assumed.
Credit 3 units. A&S IQ: NSM
Typical periods offered: Spring

SDS 5440 Mathematical Foundations of Big Data
Mathematical foundations of data science. Core topics include: Probability in high dimensions; curses and blessings of dimensionality; concentration of measure; matrix concentration inequalities. Essentials of random matrix theory. Randomized numerical linear algebra. Data clustering. Depending on time and interests, additional topics will be chosen from: Compressive sensing; efficient acquisition of data; sparsity; low-rank matrix recovery. Divide, conquer and combine methods. Elements of topological data analysis; point cloud; Cech complex; persistent homology. Selected aspects of high- dimensional computational geometry and dimension reduction; embeddings; Johnson-Lindenstrauss; sketching; random projections. Diffusion maps; manifold learning; intrinsic geometry of massive data sets. Optimization and stochastic gradient descent. Random graphs and complex networks. Combinatorial group testing. Prerequisite: Multivariable calculus (Math 233), linear or matrix algebra (Math 429 or 309), and multivariable-calculus-based probability and mathematical statistics (Math/SDS 493-494 or Math/SDS 3211/4211). Prior familiarity with analysis, topology, and geometry is strongly recommended. A willingness to learn new mathematics as needed is essential.
Credit 3 units. A&S IQ: NSM
Typical periods offered: Spring

SDS 5480 Topics in Statistics
Topic varies with each offering.
Credit 3 units. A&S IQ: NSM Art: NSM
Typical periods offered: Fall

SDS 5481 Special Topics in Statistics and Data Science: An Introduction in Python
Credit 1.5 units.

SDS 5501 Teaching Seminar
Principles and pedagogical strategies for teaching Statistics and Data Science at the college level.
Credit 1.5 units.
Typical periods offered: Spring

SDS 5502 Professional Development
This course includes topics on professional development, technical writing, and responsible conduct of research.
Credit 1.5 units.
Typical periods offered: Spring

SDS 5503 Statistics and Data Science Seminar
This weekly seminar covers recent advances in the field of Statistics and Data Science. Leaders in the field will present their recent work. The purpose is for Ph.D. students to gain knowledge about the breakthroughs in the field, the most critical problems facing the different subareas, and how researchers approach solutions to those. This exposure in turn helps them to develop professionally and may inspire advances in their own research and future research directions. The students also get familiarized with the leaders in the field. Participation also creates a sense of Data Science community within the department and WashU.
Credit 1 unit.
Typical periods offered: Spring

SDS 5510 Advanced Probability I
Advanced Probability I course description is TBD.
Credit 3 units.
Typical periods offered: Fall

SDS 5511 Advanced Probability II
Advanced Probability II course description is TBD.
Credit 3 units.
Typical periods offered: Spring

SDS 5531 Advanced Statistical Computing I
This course is the first of a sequence of two courses on advanced methods and tools for Statistical Computing. The course sequence provides opportunities to develop programming skills, algorithmic thinking, and computing strategies for statistical research. Key topics in SDS 5531 include EM algorithms, dynamic programming, random number generation, Monte Carlo methods, Markov Chain Monte Carlo (MCMC) and other advanced variants. Prereq: Math 233; a course in linear algebra at level of Math 309 or Math 429; multivariable-calculus-based probability and mathematical statistics (Math/SDS 493-494 or Math/SDS 3211/4211); Experience with a high-level programming language like R, Python, C++, etc.
Credit 3 units.
Typical periods offered: Fall

SDS 5532 Advanced Statistical Computing II
This is the second course on advanced methods and tools for Statistical Computing. This course will introduce classical methods, including the EM algorithm and its variants. It also will cover basic convex optimization theory and advanced computing tools and techniques for big data and learning algorithms. Prereq: Math 233; a course in linear algebra at level of Math 309 or Math 429; multivariable-calculus-based probability and mathematical statistics (Math/SDS 493-494 or Math/SDS 3211/4211); Experience with a high-level programming language like R, Python, C++, etc.
Credit 3 units.
Typical periods offered: Spring

SDS 5571 Topics in Advanced Probability
Credit 3 units.

SDS 5572 Topics in Advanced Probability II
Various topics related to advanced probability will be explored.
Credit 3 units.

SDS 5595 Topics in Statistics: Spatial Statistics
The course covers all three main branches of spatial statistics, namely, (1) the continuum spatial variations, (2) the discrete spatial variations and, (3) the spatial point patterns. Topics include positive definite functions, geostatistics, variograms, kriging, conditional simulations, Markov random fields, conditional and intrinsic autoregressions, Ising and Potts models, pseudolikelihood, MCMC, Inference for spatial generalized linear and mixed models, Spatial Poisson, and other point processes. The computer software R is used for examples and homework problems. Prerequisites: CSE 131; Math 233; Math 309 or Math 429; multivariable-calculus-based probability and mathematical statistics (Math/SDS 493-494 or Math/SDS 3211/4211); Math/SDS 439. Prior knowledge of R at the level introduced in Math/SDS 439 is assumed.
Credit 3 units. A&S IQ: NSM Art: NSM
Typical periods offered: Fall

SDS 5800 Topics in Statistics: Optimization Methods For Machine Learning
This is a graduate-level course designed to equip students with both fundamental and advanced optimization techniques and methods relevant to large-scale statistics and machine learning. We will begin with a concise review of the fundamentals of convex optimization, then progress to explore significant developments in first-order optimization methods across convex, nonconvex, stochastic, and distributed settings. Upon completing the course, students are expected to be capable of handling optimization-related challenges they encounter in statistics and machine learning research. This includes appropriately formulating an optimization problem, selecting or developing an efficient optimization algorithm for it, and analyzing the algorithm, based on structural properties such as convexity, smoothness, and sparsity, as well as specific settings such as online, distributed, and memory-limited contexts. Prerequisites: Fluency with reasoning and analysis using linear algebra and probability is required (MATH 309 and Math/SDS 493 or Math/SDS 3211). Students are expected to be familiar with the basics of at least one computing platform/ programming language, such as Matlab, Julia, Python, and R. Students should learn by themselves the basics of the (very user-friendly) convex optimization interpreter cvx (http://cvxr.com/cvx/) in the Matlab environment. cvx is also available in the Julia and Python environments.
Credit 3 units.
Typical periods offered: Fall

SDS 5801 Advanced Topics in Statistics
This is a variable credit-hour course in Advanced Topics in Statistics. The proposed course in Fall 2025 is a 1.5 credit-hour advanced topic course on time series analysis and highdimensional statistics. It will provide a systematic introduction to two research topics: selfnormalization (SN) for time series inference and nonlinear dependence metrics and their statistical applications. For self-normalization, we plan to cover its use for both confidence interval construction and hypothesis testing in the setting of stationary multivariate time series, functional time series, and high-dimensional time series. Change-point testing and estimation based on self-

normalization will be introduced in detail for both low and high-dimensional data. Some recent work which combines sample splitting and self-normalization will also be presented. The course assumes that the student has the basic background of time series analysis and some research experience in time series analysis is desired but not a prerequisite. For nonlinear dependence metrics, the emphasis will be placed on distance covariance, energy distance and their variants, including Hilbert-Schmidt Independence Criterion, maximum mean discrepancy, and martingale di?erence divergence, among others. The usefulness of these metrics will be demonstrated in some contemporary problems in statistics, such as dependence testing and variable screening/selection for high-dimensional data, as well as dimension reduction and diagnostic checking for multivariate time series. Some recent work on their applications to the inference of non-Euclidean data will also be discussed. The presentations are based on the research results my collaborators and I have obtained in the past and will cover methodology, theory and practical data examples.

Credit 1.5 units.

Typical periods offered: Spring

SDS 5802 Advanced Topics in Statistics:

This advanced topic course delves into five critical ideas and insights that have significantly impacted the fields of statistics and data science. From foundational concepts to influential advancements, each topic is selected for its historical significance and enduring impact. Students will develop a deep appreciation for these key ideas, associated insights, and their role in shaping contemporary statistical theory and practice. For each topic, we will discuss motivations, innovations, and impacts through presentations and discussions. Students will be required to read important papers and share their perspectives. Although many topics fit this criterion, the list includes, but is not limited to, influence functions, resampling, adaptive designs, dimension reduction, data augmentation, regularization, identifiability, and propensity scores. Depending partly on the interests of the students, we will select and focus on five topics over the course of one semester. Prerequisites: SDS 5020 and 5071.

Credit 1.5 units.

Typical periods offered: Spring

SDS 5805 Topics in Statistics

Topics in Statistics

Credit 3 units.

Typical periods offered: Fall

SDS 5901 Research

See the beginning of the mathematics listings and register for the section corresponding to supervising instructor. Prerequisite: Graduate standing and permission of the instructor.

Credit 3 units.

Typical periods offered: Fall

SDS 5910 Practical Training in Statistics

The Master of Arts in Statistics program at Department of Statistics and Data Science, Washington University in St. Louis, requires students to participate in extensive practical training as an essential component of the degree program. The program requires all full-time students to participate in practical training at least for one semester or summer session during their degree study. This requirement should be completed prior to the last semester in the degree program. The requirement does not require registration for additional credit but does require registration by ALL students, regardless of citizenship or visa status, for the zero-credit practical training course MATH 591 for one semester or summer session in which a student participates in an internship or co-op. Practical training can be fulfilled by any one of the following three methods: 1. An off-campus Internship or Co-op position with an employer in the data science industry or data science related department of a company is STRONGLY RECOMMENDED as the most preferred component of the Practical Training. The position should be related to the Statistics curriculum and span at least four weeks in duration. The student is required to submit a written report after the internship ends. 2. On-campus research, or research project participation, where the research or project is related to data science under the sponsorship of one or more of a data science institution, industry practitioner or faculty member of Washington University in St. Louis. A detailed written report on the research or project participation should be submitted and approved by a faculty member in the Department of Mathematics and Statistics. 3. Participation in the colloquium or statistics seminar in Department of Mathematics and Statistics, or other data science related research colloquium and seminar talks at Washington University in St. Louis. Students must attend talks regularly. A written report should be submitted to summarize the problems, ideas, approaches and results learned from at least four talks, and provide additional information from further reading and research of the topic.

Credit 0 units.

Typical periods offered: Fall