Statistics and Data Science

Statistics — as a core discipline focusing on data-driven discovery, understanding, and decision-making — is rapidly evolving and advancing in the data science era. The new Department of Statistics and Data Science (SDS) strives to be a world-class department with outstanding scholars who will transform the university's intellectual community not only through their own activities and achievements but also through synergistic collaborations with existing faculty and departments across Arts & Sciences, the McKelvey School of Engineering, and all of the other schools at the university.

SDS offers two master's degree programs in statistics and one doctoral degree in statistics. Our master's and PhD graduates regularly secure competitive positions at universities, at government institutions, and as statisticians and data scientists in industry. One of the most attractive features of our program is the friendly and supportive atmosphere that develops among our graduate students.

PhD in Statistics

Students ordinarily complete the PhD program in five years, and those students may expect up to five years of support. Continuation of support each year is dependent upon normal progress toward the degree and the satisfactory performance of duties. Students typically spend their first two years (four semesters) taking graduate courses. At the end of this time, they will have completed requirements for the master's degree. Students ordinarily start the process of finding a dissertation advisor and start their research during their second year.

A student who comes to Washington University with advanced preparation may finish in less time. Alternatively, some students find that it is advisable for them to take some preparatory courses before attempting the qualifying courses. In special cases, the time schedule may be lengthened accordingly.

Washington University's graduate student stipends are in the top 25% of stipends at similar universities, and St. Louis has a low cost of living.

Master of Arts in Statistics

SDS offers two standalone Master of Arts (AM) programs. The Accelerated AB/AM in Statistics is available only to qualified Washington University undergraduates. The AM in Statistics is available to all qualified students from any field. These programs provide students with the analytical background needed to prepare them for diverse careers, from positions in government and business to further PhD studies.

The regular AM in Statistics consists of 36 units of course work to be completed in three to four semesters. After completing a well-designed and judiciously chosen set of core courses (currently five), AM students can choose from a wide range of electives that includes several courses from other departments. AM students can also challenge themselves to take the more advanced Statistics PhD qualifier courses to prepare them for further PhD programs upon graduation. High-achieving students may also choose to complete a master's thesis.

Another distinctive feature of our program is a practical training experience, which will typically involve an internship off campus or working in a research group on campus.

The Accelerated AB/AM in Statistics allows highly qualified undergraduate majors to earn both the AB and AM degrees with two additional semesters of work (i.e., usually a total of five years). Participants can count up to 15 units of 400-/500-level course work earned during the four years of undergraduate study (with grades of B or better) toward the AM course requirements. Counting these 15 units makes it possible to finish the master's requirements in one additional year, but the program is still fast-paced and requires a lot of intense work and some careful planning. SDS expects applicants to have backgrounds comparable to the students admitted to the regular AM program.

Overview of Faculty Research

The interdisciplinary interests of our faculty span a broad range of areas including the application of statistics and data science to medicine, finance, environmental sciences, and technology. Research interests of our faculty include the following:

1. Bioinformatics
2. Bootstrap methodology
3. Environmental statistics
4. Functional data analysis
5. High-dimensional statistics
6. Statistical computing for massive data
7. Mathematical and statistical finance
8. Model selection and post-selection inference
9. Network analysis
10. Objective Bayes
11. Robust statistics
12. Statistical and machine learning
13. Time series and spatial statistics

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

Faculty

Chair

Xuming He
Kotzubei-Beckmann Distinguished Professor
PhD, University of Illinois at Urbana-Champaign
Robust statistics; quantile regression; Bayesian inference; post-selection inference

Director of Graduate Studies, PhD Program

José Figueroa-López
Professor
PhD, Georgia Institute of Technology
Inference methods for stochastic processes based on high-frequency sampling data; nonparametric estimation and model selection methods; time series analysis; high-frequency algorithmic trading, limit order book modeling, and asset price formation

Director of Graduate Studies, AM Program

Nan Lin
Professor
PhD, University of Illinois at Urbana–Champaign
Statistical computing in massive data, bioinformatics, Bayesian quantile regression, longitudinal data and functional data analysis, and statistical applications in anesthesiology

Department Faculty

Nilanjan Chakraborty
William Chauvenet Postdoctoral Lecturer
PhD, Michigan State University
High dimensional inference; time series; bootstrap

Likai Chen
Assistant Professor
PhD, University of Chicago
Time series; high dimensional data analysis; statistical learning theory

Jimin Ding
Associate Professor
PhD, University of California, Davis
Survival analysis; longitudinal data analysis; joint modeling of longitudinal and survival data; functional data analysis; nonparametric smoothing methods; systems of differential equations; dynamical systems; profile likelihood; asymptotic theories

Abigail Jager
Senior Lecturer
PhD, University of Chicago
Statistics; causal inference

Chetkar Jha
Postdoctoral Lecturer
PhD, University of Missouri–Columbia
Hierarchical Bayesian methods; high-dimensional data analysis; network analysis with applications to biomedical datasets such as single-cell RNA sequencing datasets; SNP genotyping datasets

Todd Kuffner
Associate Professor
PhD, Imperial College London
Statistics; econometrics; Bayesian asymptotics; applications of differential geometry to statistics; empirical likelihood; variable and model selection methods

Soumendra Lahiri
Stanley A. Sawyer Professor
PhD, Michigan State University
Asymptotic expansions, astrostatistics, inference for high dimensional and massive data sets, machine learning and predictive modeling, resampling and computer intensive methods, spatial statistics, time series and econometrics

Robert Lunde
Assistant Professor
PhD, Carnegie Mellon University
Statistical network analysis; time series; resampling methods; high-dimensional statistics

Debashis Mondal
Associate Professor
PhD, University of Washington
Spatial statistics; computational science; machine learning; applications in ecology (including microbial ecology); environmental sciences

Debjoy Thakur
Postdoctoral Lecturer
PhD, Indian Institute of Technology
Spatial statistics, resampling method, copula, spatial extreme, statistical neural network

Bowen Xie
Postdoctoral Lecturer
PhD, Iowa State University
Queuing theory; stochastic control problems; mathematical finance

Degree Requirements

- Statistics, Accelerated AB/AM
- Statistics, AM
- Statistics, PhD

Courses

Visit online course listings to view semester offerings for L87 SDS.

L87 SDS 500 Independent Work
Prerequisites: graduate standing (or, for advanced undergraduates permission of the Department's Director of Undergraduate Studies)
Credit variable, maximum 6 units.

L87 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

L87 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. Prerequisite: CSE 131 or 200, Math/SDS 3200 and Math/SDS 493.
Same as L87 SDS 494
Credit 3 units. A&S IQ: NSM Art: NSM

L87 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. Prerequisite: Math/SDS 493 or the equivalent. Some knowledge of basic ideas from analysis (e.g. Math 4111) will be helpful: consult with instructor.
Credit 3 units.

L87 SDS 5062 Theory of Statistics II
Continuation of Math/SDS 5061. Prerequisite: Math/SDS 5061 or permission of instructor.
Credit 3 units.

L87 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. Prerequisites: Math 309; Math/SDS 493 or Math/SDS 3211.
Same as L87 SDS 495
Credit 3 units. A&S IQ: NSM Art: NSM

L87 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. Prerequisites: an introductory statistics course at the level of Math/SDS 3200; a course in linear algebra at the level of Math 309 or 429; some acquaintance with fundamentals of computer programming (CSE 131); Math/SDS 493.
Credit 3 units.

L87 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. Prerequisites: Math/SDS 5071 and a course on mathematical statistics at the level of Math/SDS 494 (can be taken concurrently).
Credit 3 units.

L87 SDS 507M 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. Prerequisite: Current graduate enrollment in a program in DBBS, medicine or public health, or permission of instructor.
Credit 3 units.

L87 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. Prerequisite: CSE 131 or 200; Math/SDS 3200, or Math/SDS 3211.
Same as L87 SDS 420
Credit 3 units. A&S IQ: NSM Art: NSM

L87 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. Prerequisites: CSE 131 or 200; Math 309 or 429; Math/SDS 3200 or Math/SDS 3211.
Same as L87 SDS 434
Credit 3 units. A&S IQ: NSM

L87 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. Prerequisite: CSE 131 or 200; a course in linear algebra (such as Math 309 or 429); Math/SDS 3211 or Math/SDS 3200 and Math/SDS 493 (493 can be taken concurrently). If Math/SDS 3211 is taken, Math/SDS 493 is not required.
Same as L87 SDS 439
Credit 3 units. A&S IQ: NSM Art: NSM

L87 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. Prerequisites: Math/SDS 439 and a course in linear algebra (such as Math 309 or 429).
Same as L87 SDS 4392
Credit 3 units. A&S IQ: NSM

L87 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. Prerequisite: Math/SDS 493 or Math/SDS 3211; Math/SDS 3200, Math/SDS 494 or Math/SDS 4211.
Same as L87 SDS 461
Credit 3 units. A&S IQ: NSM Art: NSM

L87 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. Prerequisite: Math 233; Math 309 or 429; multivariable-calculus-based probability and mathematical statistics (Math/SDS 493-494 or Math/SDS 3211/4211), not taken concurrently; acquaintance with fundamentals of programming in R.
Same as L87 SDS 475
Credit 3 units. A&S IQ: NSM Art: NSM

L87 SDS 5211 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. Prerequisite: Multivariable Calculus (Math 233). 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.
Same as L87 SDS 3211
Credit 3 units. A&S IQ: NSM, AN Art: NSM

L87 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. Prerequisite: CSE 131; Math 309; multivariable-calculus-based probability and mathematical statistics (Math/SDS 493-494 or Math/SDS 3211/4211).
Same as L87 SDS 459
Credit 3 units. A&S IQ: NSM

L87 SDS 533 Mathematical Statistics I
Credit 3 units.

L87 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.
Same as L87 SDS 460
Credit 3 units. A&S IQ: NSM

L87 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.
Same as L87 SDS 462
Credit 3 units. A&S IQ: NSM

L87 SDS 5480 Topics in Statistics
Topic varies with each offering.
Same as L87 SDS 496
Credit 3 units. A&S IQ: NSM Art: NSM

L87 SDS 551 Advanced Probability I
Credit 3 units.

L87 SDS 552 Advanced Probability II
Credit 3 units.

L87 SDS 553 Topics in Advanced Probability
Credit 3 units.

L87 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.

L87 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.

L87 SDS 554 Topics in Advanced Probability II
Credit 3 units.

L87 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.
Same as L87 SDS 4971
Credit 3 units. A&S IQ: NSM Art: NSM

L87 SDS 579 Topics in Statistics
Credit 3 units.

L87 SDS 584C Multilevel Models in Quantative Research
This course covers statistical model development with explicitly defined hierarchies. Such multilevel specifications allow researchers to account for different structures in the data and provide for the modeling of variation between defined groups. The course begins with simple nested linear models and proceeds on to non-nested models, multilevel models with dichotomous outcomes, and multilevel generalized linear models. In each case, a Bayesian perspective on inference and computation is featured. The focus on the course will be practical steps for specifying, fitting, and checking multilevel models with much time spent on the details of computation in the R and Bugs environments. PREREQ: Math 2200, Math 3200, Poli Sci 581, or equivalent.
Same as L32 Pol Sci 584
Credit 3 units.

L87 SDS 586 Topics in Statistics
Credit 3 units.

L87 SDS 590 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 variable, maximum 3 units.

L87 SDS 591 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.