Washington University in St. Louis

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.

The department aims to provide a foundation for ambitious and innovative digital transformation across a range of disciplinary areas, filling a vital niche in the current academic landscape that leverages the emerging opportunities of computational and data science. SDS values foundational as well as transdisciplinary scholarship and will focus on using data to offer solutions to some of the most complex global issues.

The Department of Statistics and Data Science offers two Bachelor of Arts degrees:

1. The AB in Statistics provides flexible and rigorous training in statistics for a wide range of career paths in industry or further graduate studies.
2. Jointly offered with CSE, the AB in Data Science offers students the formal foundation needed to understand the applicability and consequences of the various approaches to analyzing data with a focus on statistical modeling and machine learning.

Both programs are offered as a prime or a double major. In addition, SDS offers a Minor in Statistics and an Accelerated AB/AM in Statistics. All of our programs are flexible enough to allow a broad range of double majors or major/minor combinations. Majors are encouraged to complete additional work in other related areas.

Why choose the AB in Statistics? Data permeates every aspect of our lives. The ability to comprehend, synthesize, analyze, extract valuable information, and draw sound conclusions from data is a must-have in almost all human endeavors and activities. Statistics not only facilitates the implementation and evaluation of mathematical and computational models that represent reality but also acknowledges the inherent randomness in data, rendering it indispensable for the making of informed decisions in a wide range of domains.

Why choose the AB in Data Science? Data science arises in the midst of a new era of data revolution and the challenges faced by the standard mathematical and statistical approaches when dealing with massive datasets, high dimensionality, and extremely complex data objects. These datasets appear in modern applications ranging from medicine to climatology to social sciences, to name just a few. Students trained in data science are already in high demand across a wide spectrum of industries. Data science is by nature interdisciplinary, requiring the mastery of a variety of skills and concepts, including many traditionally associated with the fields of statistics, computer science, and mathematics. In crafting the BA in Data Science, SDS and CSE have sought to leverage courses that are already taught as much as possible, while at the same time judiciously introducing a handful of new courses that capture unique aspects at the intersection of the two disciplines. The program features a novel practicum component during which students undertake a mentored experience to apply their knowledge and skills in industry or research.

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 careful planning. For more information, visit the Statistics and Data Science page of the graduate Arts & Sciences *Bulletin.*

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

Majors

- Data Science Major
- Statistics Major

Minors

- Statistics Minor

Courses

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

L87 SDS 1011 Introducation to Statistics
Basic concepts of statistics. Data collection (sampling, designing experiments), data organization (tables, graphs, frequency distributions, numerical summarization of data), statistical inference (elementary probability and hypothesis testing). Prerequisites: 2 years of high school algebra.
Credit 3 units. A&S IQ: NSM, AN

L87 SDS 2200 Elementary Probability and Statistics
An elementary introduction to statistical concepts, reasoning and data analysis. Topics include statistical summaries and graphical presentations of data, discrete and continuous random variables, the logic of statistical inference, design of research studies, point and interval estimation, hypothesis testing, and linear regression. Students will learn a critical approach to reading statistical analyses reported in the media, and how to correctly interpret the outputs of common statistical routines for fitting models to data and testing hypotheses. A major objective of the course is to gain familiarity with basic R commands to implement common data analysis procedures. Students intending to pursue a major or minor in statistics or wishing to take 400 level or above statistics courses should instead take Math/SDS 3200 or Math/SDS 3211. Prerequisite: Math 131
Credit 3 units. A&S IQ: NSM, AN Art: NSM

L87 SDS 3200 Elementary to Intermediate Statistics and Data Analysis
An introduction to probability and statistics. Major topics include elementary probability, special distributions, experimental design, exploratory data analysis, estimation of mean and proportion, hypothesis testing and confidence, regression, and analysis of variance. Emphasis is placed on development of statistical reasoning, basic analytic skills, and critical thinking in empirical research studies. The use of the statistical software R is integrated into lectures and weekly assignments. Required for students pursuing a major or minor in statistics or wishing to take 400 level or above statistics courses. Prerequisite: Math 132. Though Math 233 is not essential, it is recommended.
Credit 3 units. A&S IQ: NSM, AN Art: NSM

L87 SDS 3211 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.
Credit 3 units. A&S IQ: NSM, AN Art: NSM

L87 SDS 322 Biostatistics
A second course in elementary statistics with applications to life sciences and medicine. Review of basic statistics using biological and medical examples. New topics include incidence and prevalence, medical diagnosis, sensitivity and specificity, Bayes' rule, decision making, maximum likelihood, logistic regression, ROC curves and survival analysis. Prerequisite: CSE 131 or 200; Math/SDS 3200, Math/SDS 3211, or a strong performance in Math/SDS (with permission of the instructor).
Credit 3 units. A&S IQ: NSM

L87 SDS 400 Undergraduate Independent Study
Approval of instructor required
Credit variable, maximum 3 units.

L87 SDS 408 Nonparametric Statistics
Statistical methods that make few or no assumptions about the data distribution. Permutation tests of different types; nonparametric confidence intervals and correlation coefficients; jackknife and bootstrap resampling; nonparametric regressions. If there is time, topics chosen from density estimation and kernel regression. Short computer programs will be written in a language like R or C. Prerequisite: CSE 131 or 200, Math 3200 and Math 493, or permission of instructor
Credit 3 units. A&S IQ: NSM

L87 SDS 420 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.
Credit 3 units. A&S IQ: NSM Art: NSM

L87 SDS 4211 Statistics for Data Science II
This builds on the foundation from the first course (SDS I) and further develops the theory of statistical hypotheses testing. It also covers advanced computer intensive statistical methods, such as the Bootstrap, that will make extensive use of R. The emphasis of the course is to expose students to modern statistical modeling tools beyond linear models that allow for flexible and tractable interaction among response variables and covariates/feature sets. Statistical modeling and analysis of real datasets is a key component of the course. Prerequisites: Math/SDS 3211, or Math/SDS 3200 and Math/SDS 493; Math/SDS 439 (Math/SDS 439 can be taken concurrently).
Credit 3 units. A&S IQ: NSM, AN Art: NSM

L87 SDS 4311 Statistics for Humanities Scholars: Data Science for the Humanities
A survey of statistical ideas and principles. The course will expose students to tools and techniques useful for quantitative research in the humanities, many of which will be addressed more extensively in other courses: tools for text-processing and information extraction, natural language processing techniques, clustering & classification, and graphics. The course will consider how to use qualitative data and media as input for modeling and will address the use of statistics and data visualization in academic and public discourse. By the end of the course students should be able to evaluate statistical arguments and visualizations in the humanities with appropriate appreciation and skepticism. Details. Core topics include: sampling, experimentation, chance phenomena, distributions, exploration of data, measures of central tendency and variability, and methods of statistical testing and inference. In the early weeks, students will develop some facility in the use of Excel; thereafter, students will learn how to use Python or R for statistical analyses.
Same as L93 IPH 431
Credit 3 units. A&S IQ: HUM, AN EN: H

L87 SDS 434 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.
Credit 3 units. A&S IQ: NSM

L87 SDS 439 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.
Credit 3 units. A&S IQ: NSM Art: NSM

L87 SDS 4392 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).
Credit 3 units. A&S IQ: NSM

L87 SDS 459 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).
Credit 3 units. A&S IQ: NSM

L87 SDS 460 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

L87 SDS 461 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.
Credit 3 units. A&S IQ: NSM Art: NSM

L87 SDS 462 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

L87 SDS 475 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.
Credit 3 units. A&S IQ: NSM Art: NSM

L87 SDS 493 Probability
Mathematical theory and application of probability at the advanced undergraduate 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. permission of the instructor. Prerequisites: Math/SDS 3200 and Math 233.
Credit 3 units. A&S IQ: NSM Art: NSM

L87 SDS 494 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.
Credit 3 units. A&S IQ: NSM Art: NSM

L87 SDS 495 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.
Credit 3 units. A&S IQ: NSM Art: NSM

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

L87 SDS 4971 Topics in Statisitcs: 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

L87 SDS 499 Study for Honors
Prereq: Senior standing, a distinguished performance in upper level statistics courses, and permission of the Chair of the Undergraduate Committee. Register for the section (listed in department header) corresponding to your honors project supervisor.
Credit 3 units.