Statistics and Data Science

Contact:	Department of Statistics and Data Science
Email:	sds@wustl.edu
Website:	https://sds.wustl.edu

Courses

SDS 1600 Introduction 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).
Credit 3 units. A&S IQ: NSM, AN
Typical periods offered: Spring

SDS 1998 Statistics and Data Science Elective: 100-Level
Credit 3 units.

SDS 2020 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
Typical periods offered: Fall, Spring

SDS 2211 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.
Credit 3 units.
Typical periods offered: Spring

SDS 3020 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
Typical periods offered: Fall, Spring

SDS 3030 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
Typical periods offered: Fall, Spring

SDS 3110 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
Typical periods offered: Spring

SDS 3996 Statistics and Data Science Elective: 300-Level
Credit 3 units.

SDS 4000 Undergraduate Independent Study
This course is for independent study. Approval of instructor required.
Credit 3 units.
Typical periods offered: Fall

SDS 4010 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.
Credit 3 units. A&S IQ: NSM Art: NSM
Typical periods offered: Fall, Spring

SDS 4020 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: Fall

SDS 4030 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
Typical periods offered: Fall, Spring

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

SDS 4110 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 4120 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 4130 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
Typical periods offered: Fall, Spring

SDS 4140 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
Typical periods offered: Spring

SDS 4155 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 4210 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. Acquaintance with fundamentals of programming in R is helpful.
Credit 3 units. A&S IQ: NSM Art: NSM
Typical periods offered: Fall

SDS 4310 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 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.
Credit 3 units. A&S IQ: HUM, AN BU: HUM EN: H

SDS 4430 Statistical Learning
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.
Credit 3 units. A&S IQ: NSM

Typical periods offered: Spring

SDS 4440 Mathematical Foundations of Data Science
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. A willingness to learn new mathematics as needed is essential.
Credit 3 units. A&S IQ: NSM

Typical periods offered: Spring

SDS 4480 Topics in Statistics: Machine Learning Methods in Biological Sciences
Novel scientific discoveries are made nowadays by analyzing increasingly large and noisy biological datasets thanks to next-generation high-throughput technology. Machine learning methods which have been developed to extract complex patterns from image, text and speech datasets are now regularly being utilized to investigate conjectures in biology and medicine. The goal of this course will be to review some key concepts and methods in statistical learning and apply these to biological datasets. The course's focus is on the methods and applications rather than the theory and is intended for a broad audience. We will first explore predictive algorithms which perform classification and regression based on training datasets, such as logistic regression, decision trees, random forests, boosting, naive Bayes classifiers, Gaussian process regression, linear discriminant analysis and support vector machines. As much as time allows it, we will then review clustering algorithms and dimensionality reduction techniques used to identify patterns in large-scale biological datasets, such as hierarchical clustering, mixture models and principal component analysis.
Credit 3 units. A&S IQ: NSM Art: NSM

Typical periods offered: Fall

SDS 4481 Special Topics in Statistics and Data Science: An Introduction in Python
At the end of the course, students will have a solid grasp of Python programming basics and have been exposed to the entire data science workflow. This includes interacting with SQL databases to query and retrieve data, through to data wrangling, reshaping, summarizing, analyzing and ultimately reporting results.The course will introduce and use popular Python libraries such as pandas, numpy, seaborn and matplotlib and use the Jupyter notebooks framework for coding. "Special Topics in Statistics and Data Science" is a variable-credit course that covers computational/practical methods or tools of broad interests in Statistics and Data Science. The title of the course will change from semester to semester.
Credit 1.5 units.

Typical periods offered: Spring

SDS 4720 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 4971 Topics in Statistics: Data Mining
Data mining is the process of uncovering meaningful patterns and making predictions from data, often in large and complex datasets. This course provides both a practical and theoretical foundation in data mining, emphasizing computational techniques and statistical reasoning. Students will explore supervised and unsupervised learning methods, applying them to real-world datasets across different domains. The course balances conceptual understanding with hands-on programming, ensuring students grasp both the principles behind data mining algorithms and their implementation. Topics covered include essential data mining techniques, such as information retrieval and similarity measures, dimensionality reduction (PCA, MDS, Isomap, UMAP), clustering (k-means, hierarchical clustering), regression (linear regression, ridge regression, lasso), and classification (KNN, LDA, QDA, decision trees). Additional topics, including advanced clustering methods, ensemble learning techniques (bagging, boosting, random forests), and artificial neural networks, may be covered as time permits. All programming throughout the course will be conducted in Python.
Credit 3 units. A&S IQ: NSM Art: NSM

Typical periods offered: Fall

SDS 4990 Study for Honors
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.

Typical periods offered: Fall, Spring

SDS 4996 Statistics and Data Science Elective: 400-Level
Credit 3 units.