STATISTICS AND DATA SCIENCE

24 Hillhouse Avenue, 203.432.0666
http://statistics.yale.edu
M.A., Ph.D.

Chair
Harrison Zhou

Acting Chair (2018–2019)
Daniel Spielman

Directors of Graduate Studies
Andrew Barron (24 Hlh, andrew.barron@yale.edu)
David Pollard (24 Hlh, david.pollard@yale.edu)

Professors Donald Andrews (Economics), Andrew Barron, Joseph Chang, Katarzyna Chawarska (Child Study Center), Xiaohong Chen (Economics), Nicholas Christakis (Sociology), Ronald Coifman (Mathematics), James Duncan (Radiology & Biomedical Imaging), John Emerson (Adjunct), Debra Fischer (Astronomy), Alan Gerber (Political Science), Mark Gerstein (Molecular Biophysics & Biochemistry), John Hartigan (Emeritus), Theodore Holford (Public Health/Biostatistics), Edward Kaplan (School of Management/Operations Research), Harlan Krumholz (Internal Medicine), John Lafferty, Peter Phillips (Economics), David Pollard, Daniel Spielman, Hemant Tagare (Radiology & Biomedical Engineering), Van Vu (Mathematics), Heping Zhang (Public Health/Biostatistics), Hongyu Zhao (Public Health/Biostatistics), Harrison Zhou, Steven Zucker (Computer Science)

Associate Professors Peter Aronow (Political Science), Donald Lee (School of Management; Operations), Sekhar Tatikonda

Assistant Professors Timothy Armstrong (Economics), Jessi Cisewski, Zhou Fan, Amin Karbasi (Electrical Engineering), Roy Lederman, Vahideh Manshadi (School of Management/Operations), Sahand Negahban, Fredrik Savje (Political Science), Yihong Wu

Senior Lecturer Jonathan Reuning-Scherer

Lecturers Russell Barbour, William Brinda, Derek Feng, Winston Lin, Susan Wang

FIELDS OF STUDY
Fields of study include the main areas of statistical theory (with emphasis on foundations, Bayes theory, decision theory, nonparametric statistics), probability theory (stochastic processes, asymptotics, weak convergence), information theory, bioinformatics and genetics, classification, data mining and machine learning, neural nets, network science, optimization, statistical computing, and graphical models and methods.

SPECIAL ADMISSIONS REQUIREMENTS
GRE scores for the General Test are required. A GRE Subject Test in the area closest to the undergraduate major is recommended for the Ph.D. program and encouraged for the M.A. program. All applicants should have a strong mathematical background, including advanced calculus, linear algebra, elementary probability theory, and at least one course providing an introduction to mathematical statistics. An undergraduate major may be in statistics, mathematics, computer science, or in a subject in which significant statistical problems may arise. For those whose native language is not English, the Test of English as a Foreign Language (TOEFL) scores are required. This requirement is waived only for applicants who, prior to matriculation at Yale, will have received a baccalaureate degree or its international equivalent with three years of residency from a college or university where English is the primary language of instruction.

SPECIAL REQUIREMENTS FOR THE PH.D. DEGREE IN STATISTICS AND DATA SCIENCE
There is no foreign language requirement. Students take at least twelve courses, usually during the first two years. The department strongly recommends that students take S&DS 551 (Stochastic Processes), S&DS 600 (Advanced Probability), S&DS 610 (Statistical Inference), S&DS 612 (Linear Models), S&DS 625 (Statistical Case Studies), and S&DS 661 (Data Analysis), and requires that students take S&DS 626 (Practical Work). Substitutions are possible with the permission of the director of graduate studies (DGS); courses from other complementary departments such as Mathematics and Computer Science are encouraged.

The qualifying examination consists of three parts: a written report on an analysis of a data set, one or more written examination(s), and an oral examination. The examinations are taken as scheduled by the department. All parts of the qualifying examination must be completed before the beginning of the third year. A prospectus for the dissertation should be submitted no later than the first week of March in the third year. The prospectus must be accepted by the department before the end of the third year if the student is to register for a fourth year. Upon successful completion of the qualifying examination and the prospectus (and meeting of Graduate School requirements), the student is admitted to candidacy. Students are expected to attend weekly departmental seminars.
Students normally serve as teaching fellows (at level 20 or the equivalent) during four terms to acquire professional training. Although this may be completed during the third and fourth years, most students satisfy part of this requirement in the earlier years of study, with approval of the DGS and their adviser, in areas contributing to their professional development.

**MÁSTER’S DEGREES**

**M.A. (en route to the Ph.D. in Statistics and Data Science)** This degree may be awarded upon completion of eight term courses in Statistics with an average grade of HP or higher, and two terms of residence.

**Terminal Master’s Degree Program in Statistics** Students are also admitted directly to a terminal master’s degree program in Statistics. To qualify for the M.A., the student must successfully complete an approved program of eight term courses in Statistics with an average grade of HP or higher, chosen in consultation with the DGS. Full-time students must take a minimum of four courses per term. Part-time students are also accepted into the master’s degree program. See Terminal M.A./M.S. Degrees, under Policies and Regulations.

Program information is available online at http://statistics.yale.edu.

**COURSES**

**S&DS 500b, Introductory Statistics** William Brinda
An introduction to statistical reasoning. Topics include numerical and graphical summaries of data, data acquisition and experimental design, probability, hypothesis testing, confidence intervals, correlation and regression. Application of statistical concepts to data; analysis of real-world problems.

**S&DS 501a, Introduction to Statistics: Life Sciences** Walter Jetz and Jonathan Reuning-Scherer
Statistical and probabilistic analysis of biological problems, presented with a unified foundation in basic statistical theory. Problems are drawn from genetics, ecology, epidemiology, and bioinformatics.

**S&DS 502a, Introduction to Statistics: Political Science** Jonathan Reuning-Scherer
Statistical analysis of politics, elections, and political psychology. Problems presented with reference to a wide array of examples: public opinion, campaign finance, racially motivated crime, and public policy. Note: S&DS 501–506 offer a basic introduction to statistics, including numerical and graphical summaries of data, probability, hypothesis testing, confidence intervals, and regression. Each course focuses on applications to a particular field of study and is taught jointly by two instructors, one specializing in statistics and the other in the relevant area of application. The first seven weeks are attended by all students in S&DS 501–506 together as general concepts and methods of statistics are developed. The course separates for the last six and a half weeks, which develop the concepts with examples and applications. Computers are used for data analysis. These courses are alternatives; they do not form a sequence, and only one may be taken for credit.

**S&DS 503a, Introduction to Statistics: Social Sciences** Jonathan Reuning-Scherer
Descriptive and inferential statistics applied to analysis of data from the social sciences. Introduction of concepts and skills for understanding and conducting quantitative research. Note: S&DS 501–506 offer a basic introduction to statistics, including numerical and graphical summaries of data, probability, hypothesis testing, confidence intervals, and regression. Each course focuses on applications to a particular field of study and is taught jointly by two instructors, one specializing in statistics and the other in the relevant area of application. The first seven weeks are attended by all students in S&DS 501–506 together as general concepts and methods of statistics are developed. The course separates for the last six and a half weeks, which develop the concepts with examples and applications. Computers are used for data analysis. These courses are alternatives; they do not form a sequence, and only one may be taken for credit.

**S&DS 505a, Introduction to Statistics: Medicine** Russell Barbour and Jonathan Reuning-Scherer
Statistical methods relied upon in medicine and medical research. Practice in reading medical literature competently and critically, as well as practical experience performing statistical analysis of medical data. Note: S&DS 501–506 offer a basic introduction to statistics, including numerical and graphical summaries of data, probability, hypothesis testing, confidence intervals, and regression. Each course focuses on applications to a particular field of study and is taught jointly by two instructors, one specializing in statistics and the other in the relevant area of application. The first seven weeks are attended by all students in S&DS 501–506 together as general concepts and methods of statistics are developed. The course separates for the last six and a half weeks, which develop the concepts with examples and applications. Computers are used for data analysis. These courses are alternatives; they do not form a sequence, and only one may be taken for credit.

**S&DS 506a, Introduction to Statistics: Data Analysis** William Brinda and Jonathan Reuning-Scherer
An introduction to probability and statistics with emphasis on data analysis. Note: S&DS 501–506 offer a basic introduction to statistics, including numerical and graphical summaries of data, probability, hypothesis testing, confidence intervals, and regression. Each course focuses on applications to a particular field of study and is taught jointly by two instructors, one specializing in statistics and the other in the relevant area of application. The first seven weeks are attended by all students in S&DS 501–506 together as general concepts and methods of statistics are developed. The course separates for the last six and a half weeks, which develop the concepts with examples and applications. Computers are used for data analysis. These courses are alternatives; they do not form a sequence, and only one may be taken for credit.

**S&DS 520b, Intensive Introductory Statistics** Xiaofei Wang
An introduction to statistical reasoning designed for students with particular interest in data science and computing. Using the R language, topics include exploratory data analysis, probability, hypothesis testing, confidence intervals, regression, statistical modeling,
and simulation. Computing is taught and used extensively throughout the course. Application of statistical concepts to the analysis of real-world data science problems.

**S&DS 52b, YData: An Introduction to Data Science**  Jessica Cisewski and John Lafferty
Computational, programming, and statistical skills are no longer optional in our increasingly data-driven world; they are essential for opening doors to manifold research and career opportunities. This course aims to dramatically enhance students’ knowledge and capabilities in fundamental ideas in data science, especially computational and programming skills and inferential thinking. It emphasizes the development of these skills while providing opportunities for hands-on experience and practice. The course is designed to be accessible to students with little or no background in computing, programming, or statistics, but also engaging for more technically oriented students through extensive use of examples and hands-on data analysis. Python 3 is the computing language used.

**S&DS 530a or b, Data Exploration and Analysis**  Staff
Survey of statistical methods: plots, transformations, regression, analysis of variance, clustering, principal components, contingency tables, and time series analysis. The R computing language and Web data sources are used.

**S&DS 538a, Probability and Statistics**  Joseph Chang
Fundamental principles and techniques of probabilistic thinking, statistical modeling, and data analysis. Essentials of probability: conditional probability, random variables, distributions, law of large numbers, central limit theorem, Markov chains. Statistical inference with emphasis on the Bayesian approach: parameter estimation, likelihood, prior and posterior distributions, Bayesian inference using Markov chain Monte Carlo. Introduction to regression and linear models. Computers are used throughout for calculations, simulations, and analysis of data. Prerequisite: differential calculus of several variables; some acquaintance with matrix algebra and computing is assumed.

**S&DS 541a, Probability Theory**  Yihong Wu
A first course in probability theory: probability spaces, random variables, expectations and probabilities, conditional probability, independence, some discrete and continuous distributions, central limit theorem, Markov chains, probabilistic modeling. Prerequisite: calculus of functions of several variables.

**S&DS 542b, Theory of Statistics**  Andrew Barron

**S&DS 551b, Stochastic Processes**  Yihong Wu and Sahand Negahban
Introduction to the study of random processes, including Markov chains, Markov random fields, martingales, random walks, Brownian motion, and diffusions. Techniques in probability such as coupling and large deviations. Applications chosen from image reconstruction, Bayesian statistics, finance, probabilistic analysis of algorithms, genetics, and evolution.

**S&DS 563b, Multivariate Statistical Methods for the Social Sciences**  Jonathan Reuning-Scherer
An introduction to the analysis of multivariate data. Topics include principal components analysis, factor analysis, cluster analysis (hierarchical clustering, k-means), discriminant analysis, multidimensional scaling, and structural equations modeling. Emphasis on practical application of multivariate techniques to a variety of examples in the social sciences. Students complete extensive computer work using either SAS or SPSS. Prerequisites: knowledge of basic inferential procedures, experience with linear models (regression and ANOVA). Experience with some statistical package and/or familiarity with matrix notation is helpful but not required.

**S&DS 565a or b, Applied Data Mining and Machine Learning**  Staff
Techniques for data mining and machine learning are covered from both a statistical and a computational perspective, including support vector machines, bagging, boosting, neural networks, and other nonlinear and nonparametric regression methods. The course gives the basic ideas and intuition behind these methods, a more formal understanding of how and why they work, and opportunities to experiment with machine-learning algorithms and apply them to data. Prerequisite: after or concurrent with S&DS 542.

**S&DS 570b / ASTR 545b, YData: ExoStatistics: Exploring Extrasolar Planets with Data Science**  Jessica Cisewski
Extrasolar planets, or exoplanets, are planets orbiting stars outside our solar system. The past decade has led to a proliferation of exoplanet discoveries using various detection methods. Through the lens of data science, we investigate exoplanet datasets to learn how to find exoplanets, examine the population properties of observed exoplanets, estimate probabilities of another Earth-like exoplanet in our universe, and probe other questions about exoplanets. This course provides an introduction to exoplanet astronomy, an introduction to data science tools necessary for studying exoplanets, and opportunities to practice the data science skills presented in S&DS 523. This course can be taken concurrently with, or after successful completion of, S&DS 523. ½ Course cr

**S&DS 571b, YData: Text Data Science: An Introduction**  John Lafferty
Written language is the primary means by which humans document their observations of the world, including scientific discoveries, interpretations of history and art, health diagnoses, analyses of political events and economic trends, social interactions, and many others. Increasingly, this rapidly growing transcript is readily available in electronic form and is being used in commercial applications and to advance scientific knowledge. This course is an introduction to computational and inferential methods that use text. The focus is on simple but often powerful text-processing techniques that do not require linguistic analyses, to gain familiarity with working with text data. Sources used in the seminar include political speeches, Twitter feeds, scientific journals, online FAQ and discussion boards, Wikipedia, news articles, and consumer product reviews. Methodologies include scraping, wrangling, hashing, sorting, regressing, embedding, and probabilistic modeling. The course is based on the Python programming language within a cloud computing platform.
and is paced to be accessible to students who have previously taken or are currently enrolled in S&DS 523. Prerequisite: S&DS 523; may be taken concurrently. ½ Course cr

S&DS 572b / PLSC 524b, YData: Data Science for Political Campaigns  Joshua Kalla
Political campaigns have become increasingly data driven. Data science is used to inform where campaigns compete, which messages they use, how they deliver them, and among which voters. In this course, we explore how data science is being used to design winning campaigns. Students gain an understanding of what data is available to campaigns, how campaigns use this data to identify supporters, and the use of experiments in campaigns. The course provides students with an introduction to political campaigns, an introduction to data science tools necessary for studying politics, and opportunities to practice the data science skills presented in S&DS 523. Can be taken concurrently with, or after successful completion of, S&DS 523. ½ Course cr

S&DS 600b, Advanced Probability  Sekhar Tatikonda
Measure theoretic probability, conditioning, laws of large numbers, convergence in distribution, characteristic functions, central limit theorems, martingales. Some knowledge of real analysis is assumed.

S&DS 610a, Statistical Inference  Zhou Fan
A systematic development of the mathematical theory of statistical inference covering methods of estimation, hypothesis testing, and confidence intervals. An introduction to statistical decision theory. Knowledge of probability theory at the level of S&DS 541 is assumed.

S&DS 612a, Linear Models  William Brinda
The geometry of least squares; distribution theory for normal errors; regression, analysis of variance, and designed experiments; numerical algorithms (with particular reference to the R statistical language); alternatives to least squares. Prerequisites: linear algebra and some acquaintance with statistics.

S&DS 615b, Introduction to Random Matrix Theory and Applications  Zhou Fan
A graduate-level introduction to random matrix theory. Wigner matrices, sample covariance matrices, spiked models. Applications to statistical principal component analysis, random graphs and networks, and landscape analysis of nonconvex statistical optimization problems. Methods applicable to non-invariant models that commonly arise in statistical applications: moment method, resolvents and Stieltjes transforms, free probability, concentration of measure, Lindeberg exchange. Prerequisite: real analysis and measure-theoretic probability.

S&DS 625a, Statistical Case Studies  Xiaofei Wang
Statistical analysis of a variety of statistical problems using real data. Emphasis on methods of choosing data, acquiring data, assessing data quality, and the issues posed by extremely large data sets. Extensive computations using R.

S&DS 626a or b, Practical Work  Staff
Individual one-term projects, with students working on studies outside the department, under the guidance of a statistician.

S&DS 627a and S&DS 628b, Statistical Consulting  Derek Feng
Statistical consulting and collaborative research projects often require statisticians to explore new topics outside their area of expertise. This course exposes students to real problems, requiring them to draw on their expertise in probability, statistics, and data analysis. Students complete the course with individual projects supervised jointly by faculty outside the department and by one of the instructors. Students enroll for both terms (S&DS 627 and 628) and receive one credit at the end of the year. ½ Course cr per term

S&DS 630a, Optimization Techniques  Sekhar Tatikonda
Fundamental theory and algorithms of optimization, emphasizing convex optimization. The geometry of convex sets, basic convex analysis, the principle of optimality, duality. Numerical algorithms: steepest descent, Newton's method, interior point methods, dynamic programming, unimodal search. Applications from engineering and the sciences.

S&DS 645b / CB&B 645b, Statistical Methods in Computational Biology  Hongyu Zhao
Introduction to problems, algorithms, and data analysis approaches in computational biology and bioinformatics. We discuss statistical issues arising in analyzing population genetics data, gene expression microarray data, next-generation sequencing data, microbiome data, and network data. Statistical methods include maximum likelihood, EM, Bayesian inference, Markov chain Monte Carlo, and methods of classification and clustering; models include hidden Markov models, Bayesian networks, and graphical models. Prerequisite: S&DS 538, S&DS 542, or S&DS 661. Prior knowledge of biology is not required, but some interest in the subject and a willingness to carry out calculations using R is assumed.

S&DS 661b, Data Analysis  William Brinda
By analyzing data sets using the R statistical computing language, a selection of statistical topics are studied: linear and nonlinear models, maximum likelihood, resampling methods, curve estimation, model selection, classification, and clustering. Prerequisite: after or concurrent with S&DS 542.

S&DS 663a, Computational Mathematics for Data Science  Roy Lederman
The course explores the mechanics of the interface between mathematics, computation, and statistics in data analysis. We discuss topics in numerical computation, complexity, programming, and prototyping. Assignments include theory, programming, data analysis, individual work, collaborative work, and making mistakes. Prerequisites: linear algebra and some experience with programming (any language).
S&DS 664b, Information Theory  Andrew Barron
Foundations of information theory in communications, statistical inference, statistical mechanics, probability, and algorithmic complexity. Quantities of information and their properties: entropy, conditional entropy, divergence, redundancy, mutual information, channel capacity. Basic theorems of data compression, data summarization, and channel coding. Applications in statistics.

S&DS 671b, Selected Topics in Neural Nets  Harrison Zhou
Recent theoretical developments in neural nets. Topics include nonconvex optimization; generalization theory; overparameterization; GAN and VAE; mean field view; implicit regularization; geometry; and statistical theory.

S&DS 672a, Information Theory Tools in Probability and Statistics  Andrew Barron
Information theory techniques valuable in probability, statistics, and machine-learning research. Example topics include information inequalities in central limit analysis, in adaptive estimation, in minimax risk determination, in metric entropy calculation, in stochastic search, and in the exploration of the accuracy of deep learning networks. Prerequisite: co-enrollment in S&DS 610, or completion of S&DS 664 and S&DS 542.

S&DS 674b, Applied Spatial Statistics  Timothy Gregoire
An introduction to spatial statistical techniques with computer applications. Topics include modeling spatially correlated data, quantifying spatial association and autocorrelation, interpolation methods, variograms, kriging, and spatial point patterns. Examples are drawn from ecology, sociology, public health, and subjects proposed by students. Four to five lab/homework assignments and a final project. The class makes extensive use of the R programming language as well as ArcGIS.

S&DS 679a, High-Dimensional Statistical Estimation  Sahand Negahban
In this course we review the recent advances in high-dimensional statistics, covering concepts in empirical process theory, concentration of measure, and random matrix theory in the context of understanding the statistical properties of high-dimensional estimation methods. We also cover the computational constraints that are involved with solving high-dimensional problems and touch upon concepts in convex optimization and online learning.

S&DS 683a, Statistical Methods in Neuroimaging  Joseph Chang and Dustin Scheinost
Introduction to common statistical methods used in neuroimaging. Topics include introduction to different imaging modalities and experimental designs; modeling tasks using linear models; functional connectivity analysis; mixed effects, repeated measures, longitudinal models, power; multiple comparisons, random fields; effective connectivity, dynamic causal modeling, and variational Bayesian methods; machine-learning approaches to multi-voxel pattern analysis.

S&DS 684a, Statistical Inference on Graphs  Yihong Wu
An emerging research thread in statistics and machine learning deals with finding latent structures from data represented in graphs or matrices. This course provides an introduction to mathematical and algorithmic tools for studying such problems. We discuss information-theoretic methods for determining the fundamental limits, as well as methodologies for attaining these limits, including spectral methods, optimization techniques such as semidefinite programming relaxations, message-passing algorithms, etc. Specific topics include spectral clustering, planted clique and partition problems, sparse PCA, community detection on stochastic block models, statistical-computational trade-offs. Prerequisites: maturity with probability theory; familiarity with statistical theory at the level of S&DS 610.

S&DS 690a or b, Independent Study  Staff
By arrangement with faculty. Approval of DGS required.