STATISTICS AND DATA SCIENCE

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

Chair
Joseph Chang

Directors of Graduate Studies
Andrew Barron (24 Hlh, andrew.barron@yale.edu)
John Emerson (24 Hlh, john.emerson@yale.edu)

Professors Donald Andrews (Economics), Andrew Barron, Jeffrey Brock (Mathematics),
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),
Anna Gilbert, John Hartigan (Emeritus), Edward Kaplan (School of Management/
Operations Research), Harlan Krumholz (Internal Medicine), John Lafferty, David Pollard
(Emeritus), Nils Rudi (School of Management), Jasjeet Sekhon, Donna Spiegelman
(Biostatistics), Daniel Spielman, Hemant Tagare (Radiology & Biomedical Engineering)
, Van Vu (Mathematics), Heping Zhang (Biostatistics), Hongyu Zhao (Biostatistics),
Harrison Zhou, Steven Zucker (Computer Science)

Associate Professors Peter Aronow (Political Science), Forrest Crawford (Biostatistics),
Amin Karbasi (Electrical Engineering), Ethan Meyers (Visiting), Sahand Negahban,
Sekhar Tatikonda, Yihong Wu

Assistant Professors Elisa Celis, Zhou Fan, Joshua Kalla (Political Science), Roy
Lederman, Vahideh Manshadi (School of Management/Operations), Fredrik Savje
(Political Science), Zhuoran Yang, Ilker Yildirim (Psychology)

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 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), S&DS 631 (Optimization and Computation), S&DS 632 (Advanced
Optimization Techniques), 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. With the permission of the DGS and under special circumstances, appropriate courses may be taken at the undergraduate level in departments outside of Statistics and Data Science to fulfill these elective requirements.

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 for several terms to acquire professional training. All students are required to teach, usually for two terms, regardless of the nature of their funding. This teaching is typically completed in the first two years of study, although the actual timing is at the discretion of the DGS. Students who require additional support from the Graduate School after their second year will be required to teach additional terms, if needed.

COMBINED PH.D. PROGRAM

The Department of Statistics and Data Science also offers, in conjunction with the Department of Political Science, a combined Ph.D. in Statistics and Data Science and Political Science. For further details, see Political Science.

MASTER’S DEGREES

M.A. in Statistics (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. With the permission of the DGS and under special circumstances, appropriate courses may be taken at the undergraduate level in departments outside of Statistics and Data Science to fulfill elective requirements.

M.A. in Statistics (en route to the Ph.D. in other areas of study) Pursuit of this degree requires an application process managed by the DGS of Statistics and Data Science followed by approval from the DGSs from both programs and the cognizant Graduate School dean. This degree is awarded upon completion of eight term courses in Statistics, chosen in consultation with the DGSs, with all grades HP or higher. Most of these courses should be in addition to the requirements of the primary Ph.D. program. With the permission of the DGS and under special circumstances, appropriate courses may be taken at the undergraduate level in departments outside of Statistics and Data Science to fulfill elective requirements. This degree also has an academic teaching fellow requirement, to be determined by the DGSs from both programs and the cognizant Graduate School dean.
Terminal M.A. in Statistics Students are also admitted directly to a terminal master of arts program in Statistics. To qualify for the M.A., the student must successfully complete an approved program of eight term courses with an average grade of HP or higher and receive at least one grade of Honors, chosen in consultation with the DGS. With the permission of the DGS and under special circumstances, appropriate courses may be taken at the undergraduate level in departments outside of Statistics and Data Science to fulfill elective requirements. Full-time students must take a minimum of four courses per term. Part-time students are also accepted into the program. All students are expected to complete two terms of full-time tuition and residence, or the equivalent, at Yale. See Degree Requirements: Terminal M.A./M.S. Degrees, under Policies and Regulations.

Terminal M.S. in Statistics and Data Science Students are also admitted directly to a terminal master of science program in Statistics and Data Science. To qualify for the M.S., the student must successfully complete an approved program of twelve term courses with an average grade of HP or higher and receive at least two grades of Honors, chosen in consultation with the DGS. With the permission of the DGS and under special circumstances, appropriate courses may be taken at the undergraduate level in departments outside of Statistics and Data Science to fulfill elective requirements. Full-time students must take a minimum of four courses per term. Part-time students are also accepted into the program. All students are expected to complete three terms of full-time tuition and residence, or the equivalent, at Yale. See Degree Requirements: 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 Ethan Meyers
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.

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**  Ethan Meyers 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**  Robert Wooster 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**  Robert Wooster

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 523b, YData: An Introduction to Data Science**  Ethan Meyers
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 and skills 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. Enrollment is limited.

**S&DS 530a or b / PLSC 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: after or concurrently with MATH 118 or MATH 120.

**S&DS 540a, An Introduction to Probability Theory**  Robert Wooster
Introduction to probability theory. Topics include probability spaces, random variables, expectations and probabilities, conditional probability, independence, discrete and continuous distributions, central limit theorem, Markov chains, and probabilistic modeling. This course may be appropriate for non-S&DS graduate students. Prerequisite: MATH 115 or equivalent.

**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**  Robert Wooster

**S&DS 551b / ENAS 502b, Stochastic Processes**  Amin Karbasi
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 562b, Computational Tools for Data Science**  Roy Lederman
An introduction to computational tools for data science. The analysis of data using regression, classification, clustering, principal component analysis, independent component analysis, dictionary learning, topic modeling, dimension reduction, and network analysis. Optimization by gradient methods and alternating minimization. The application of high-performance computing and streaming algorithms to the analysis of large data sets. Prerequisites: linear algebra, multivariable calculus, and programming.

**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, Introductory Machine Learning**  John Lafferty
This course covers the key ideas and techniques in machine learning without the use of advanced mathematics. Basic methodology and relevant concepts are presented in lectures, including the intuition behind the methods. Assignments give students hands-on experience with the methods on different types of data. Topics include linear regression and classification, tree-based methods, clustering, topic models, word embeddings, recurrent neural networks, dictionary learning, and deep learning. Examples come from a variety of sources including political speeches, archives of scientific articles, real estate listings, natural images, and others. Programming is central to the course and is based on the Python programming language.

**S&DS 572a, 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.

**S&DS 600a, 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**  Harrison Zhou
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 617b / PLSC 511b, Applied Machine Learning and Causal Inference Research Seminar**  Jas Sekhon
In this seminar we discuss recent advances in machine learning and causal inference. Emphasis is placed on research designs and methods that have succeeded. We carefully examine successful examples to see why they work. The seminar is also a forum for students to discuss the research designs and methods needed in their own work. It should be particularly helpful for students writing their prospectus or designing a major research project. Applications are drawn from a variety of substantive domains including political science, economics, medicine, and public health. It is assumed that students come with diverse backgrounds. A good background would be provided by S&DS 542, ECON 551, or equivalent, plus some experience with applications and statistical computing. More important than the precise course background are research maturity and familiarity with modern statistical and machine-learning methods.

**S&DS 625a or b, Statistical Case Studies**  Staff
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. Enrollment limited; requires permission of the instructor.

**S&DS 627a and S&DS 628b, Statistical Consulting**  Jay Emerson
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. Enrollment limited; requires permission of the instructor. ½ Course cr per term

**S&DS 631a / AMTH 631a, Optimization and Computation**  Yang Zhuoran
An introduction to optimization and computation motivated by the needs of computational statistics, data analysis, and machine learning. This course provides foundations essential for research at the intersections of these areas, including the asymptotic analysis of algorithms, an understanding of condition numbers, conditions for optimality, convex optimization, gradient descent, linear and conic programming, and NP hardness. Model problems come from numerical linear algebra and constrained least squares problems. Other useful topics include data structures used to represent graphs and matrices, hashing, automatic differentiation, and randomized algorithms. Prerequisites: multivariate calculus, linear algebra, probability, and permission of the instructor. Enrollment is limited, with preference given to graduate students in Statistics and Data Science.
S&D 632b, Advanced Optimization Techniques Sekhar Tatikonda
This course covers fundamental theory and algorithms in optimization, emphasizing convex optimization. Topics covered include convex analysis; duality and KKT conditions; subgradient methods; interior point methods; semidefinite programming; distributed methods; stochastic gradient methods; robust optimization; and an introduction to nonconvex optimization. Applications from statistics and data science, economics, engineering, and the sciences. Prerequisites: knowledge of linear algebra, such as MATH 222 or MATH 225; multivariate calculus, such as MATH 120; probability, such as S&D 541; optimization, such as S&D 631; and comfort with proof-based exposition and problem sets.

S&D 661b, Data Analysis Brian Macdonald
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&D 542.

S&D 662a, Statistical Computing Jay Emerson
Topics in the practice of data analysis and statistical computing, with particular attention to problems involving massive data sets or large, complex simulations and computations. Programming with R, C/C++, and Perl/Python, computational efficiency, memory management, interactive and dynamic graphics, and parallel computing.

S&D 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&D 665a, Intermediate Machine Learning John Lafferty
S&D 365 is a second course in machine learning at the advanced undergraduate or beginning graduate level. The course assumes familiarity with the basic ideas and techniques in machine learning, for example as covered in S&D 265. The course treats methods together with mathematical frameworks that provide intuition and justifications for how and when the methods work. Assignments give students hands-on experience with machine learning techniques, to build the skills needed to adapt approaches to new problems. Topics include nonparametric regression and classification, kernel methods, risk bounds, nonparametric Bayesian approaches, graphical models, attention and language models, generative models, sparsity and manifolds, and reinforcement learning. Programming is central to the course, and is based on the Python programming language and Jupyter notebooks.

S&D 690a or b, Independent Study Jay Emerson
By arrangement with faculty. Approval of DGS required.

S&D 695b, Summer Internship in Statistics and Data Science Jay Emerson
The purpose of this course is to provide students with the opportunity to gain practical experience in statistics and data science. Students who identify a suitable summer internship consult with the DGS and prepare a one-page description of the plan. The internship must be full-time: 35–40 hours per week for 10–12 weeks during the
summer. Upon completion of the internship, the student must submit a written report of the work to the instructor no later than October 1. Prerequisites: completion of at least one term of the M.S. program (or the M.A. program if transferring into the M.S. program) and permission of the DGS.

S&DS 700a or b, Departmental Seminar  Staff
Presentations of recent breakthroughs in statistics and data science.  o Course cr