STATISTICS AND DATA SCIENCE

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 and Biomedical Imaging*), John Emerson (*Adjunct*), Alan Gerber (*Political Science*), Mark Gerstein (*Molecular Biophysics and Biochemistry*), Anna Gilbert, John Hartigan (*Emeritus*), Edward Kaplan (*School of Management/Operations Research*), Harlan Krumholz (*Internal Medicine*), John Lafferty, Zongming Ma, David Pollard (*Emeritus*), Nils Rudi (*School of Management*), Jasjeet Sekhon, Donna Spiegelman (*Biostatistics*), Daniel Spielman, Hemant Tagare (*Radiology and Biomedical Engineering*), Van Vu (*Mathematics*), Yihong Wu, Heping Zhang (*Biostatistics*), Hongyu Zhao (*Biostatistics*), Harrison Zhou, Steven Zucker (*Computer Science*)

Associate Professors P.M. Aronow (*Political Science*), Forrest Crawford (*Biostatistics*), Amin Karbasi (*Electrical Engineering*), Vahideh Manshadi (*School of Management/ Operations*), Ethan Meyers (*Visiting*), Sekhar Tatikonda

Assistant Professors Elisa Celis, Zhou Fan, Joshua Kalla (*Political Science*), Roy Lederman, Lu Lu, Fredrik Savje (*Political Science*), Dustin Scheinost (*Radiology and Biomedical Imaging*), Andre Wibisono (*Computer Science*), Zhuoran Yang, Ilker Yildirim (*Psychology*), Ilias Zadik

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 be teaching fellows for a minimum of two terms, regardless of the nature of their funding. The timing of this teaching is at the discretion of the DGS.

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

Three different M.A. in Statistics are offered. All require completion of eight term courses approved by the DGS; of which one must be in probability, one must be in statistical theory, and one must be in data analysis. The remaining five elective courses may include courses from other departments and, 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.

M.A. in Statistics (en route to the Ph.D. in Statistics and Data Science) This degree requires an average grade of HP or higher, and two terms of residence.

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. All eight courses for this degree must earn grades of HP or higher. Most of the courses for the M.A. in Statistics should be in addition to the requirements of the primary Ph.D. program. 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. Students must earn an average grade of HP or higher and receive at least one grade of Honors. 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 501a / E&EB 510a, Introduction to Statistics: Life Sciences 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 Jay Emerson 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 530a / PLSC 530a, Data Exploration and Analysis Ethan Meyers
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 542a, Theory of Statistics Andrew Barron
Principles of statistical analysis: maximum likelihood, sampling distributions, estimation, confidence intervals, tests of significance, regression, analysis of variance, and the method of least squares. Prerequisite: S&DS 541.

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 580a, Neural Data Analysis Ethan Meyers
We discuss data analysis methods that are used in the neuroscience community. Methods include classical descriptive and inferential statistics, point process models, mutual information measures, machine learning (neural decoding) analyses, dimensionality reduction methods, and representational similarity analyses. Each week we read a research paper that uses one of these methods, and we replicate these analyses using the R or Python programming language. Emphasis is on analyzing neural spiking data, although we also discuss other imaging modalities such as magneto/electro-encephalography (EEG/MEG), two-photon imaging, and possibility functional magnetic resonance imaging data (fMRI). Data we analyze includes smaller datasets, such as single neuron recordings from songbird vocal motor system, as well as larger

data sets, such as the Allen Brain observatory's simultaneous recordings from the mouse visual system.

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 Harrison Zhou
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 Zongming Ma
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 625a, Statistical Case Studies Brian Macdonald
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, 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

S&DS 631a / AMTH 631a, Optimization and Computation Zhuoran Yang
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&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. Offered every other year. 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 665a, Intermediate Machine Learning John Lafferty
S&DS 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&DS 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&DS 688a, Computational and Statistical Trade-offs in High Dimensional Statistics Ilias Zadik
Modern statistical tasks require the use of both computationally efficient and statistically accurate methods. But, can we always find a computationally efficient method that achieves the information-theoretic optimal statistical guarantees? If not, is this an artifact of our techniques, or a potentially fundamental source of computational hardness? This course surveys a new and growing research area studying such questions on the intersection of high dimensional statistics and theoretical computer science. We discuss various tools to explain the presence of such "computational-to-statistical gaps" for several high dimensional inference models. These tools include the "low-degree polynomials" method, statistical query lower bounds, and more. We also discuss connections with other fields such as statistical physics and cryptography. Prerequisites: maturity with probability theory (equivalent of 241/541) and linear algebra and a familiarity with basic algorithms and mathematical statistics.

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

S&DS 700a, Departmental Seminar Staff
Presentations of recent breakthroughs in statistics and data science. 0 Course cr