STATISTICS AND DATA SCIENCE

Directors of undergraduate studies: Sekhar Tatikonda, Rm. 338, 17 Hillhouse Ave., 432-4714, sekhar.tatikonda@yale.edu; statistics.yale.edu

Statistics is the science and art of prediction and explanation. The mathematical foundation of statistics lies in the theory of probability, which is applied to problems of making inferences and decisions under uncertainty. Practical statistical analysis also uses a variety of computational techniques, methods of visualizing and exploring data, methods of seeking and establishing structure and trends in data, and a mode of questioning and reasoning that quantifies uncertainty. Data science expands on statistics to encompass the entire life cycle of data, from its specification, gathering, and cleaning through its management and analysis, to its use in making decisions and setting policy. It is a natural outgrowth of statistics that incorporates advances in machine learning, data mining, and high-performance computing, along with domain expertise in the social sciences, natural sciences, engineering, management, medicine, and digital humanities.

Students majoring in Statistics and Data Science take courses in both mathematical and practical foundations. They are also encouraged to take courses in the discipline areas listed below.

The B.A. in Statistics and Data Science is designed to acquaint students with fundamental techniques in the field. The B.S. prepares students to participate in research efforts or to pursue graduate school in the study of Data Science.

Requirements of the Statistics Major for the Class of 2019 With DUS approval, the following changes to the requirements of the major may be fulfilled by students who declared their major under previous requirements.

Requirements of the Statistics and Data Science Major for the Class of 2020 and subsequent classes The requirements of the new Statistics and Data Science major are indicated below.

COURSES FOR NONMAJORS AND MAJORS

S&DS 100 and S&DS 101 through 109 only assume knowledge of high-school mathematics. Students who complete one of these courses should consider taking S&DS 230. This sequence provides a solid foundation for the major. Other courses for non-majors include S&DS 110 and 160.

PREREQUISITES

Multivariable calculus and linear algebra are required and should be taken before or during the sophomore year. This requirement may be satisfied by one of MATH 120, ENAS 151, MATH 230, or the equivalent.

REQUIREMENTS OF THE MAJOR

Students who wish to major in Statistics and Data Science are encouraged to take S&DS 220. But, students may also enter the major by taking a 100-level course followed by S&DS 230. Students should complete the calculus prerequisite and linear algebra requirement as early as possible, as they provide mathematical background that is required in many courses.

B.A. degree program The B.A. degree program requires eleven courses, ten of which are from the seven discipline areas described below: MATH 222 or 225 from Mathematical Foundations and Theory; two courses from Core Probability and Statistics; two courses that provide Computational Skills; two courses on Methods of Data Science; and three courses from any of the discipline areas. The remaining course is fulfilled through the senior requirement.

B.S. degree program The B.S. degree program requires fourteen courses, including all the requirements for the B.A. degree as well as S&DS 242, which counts as one of the required courses from Core Probability and Statistics. The two remaining courses may be chosen from Core Probability and Statistics; Computational Skills; Methods of Data Science; Mathematical Foundations and Theory; or Efficient Computation and Big Data discipline areas.

Discipline Areas The seven discipline areas are listed below.

Core Probability and Statistics These are essential courses in probability and statistics. Every major should take at least two of these courses, and should probably take more. Students completing the B.S. degree must take S&DS 242. Examples of such courses include: S&DS 238, 241, 242, 312, 351

Computational Skills Every major should be able to compute with data. While the main purpose of some of these courses is not computing, students who have taken at least two of these courses will be capable of digesting and processing data. While there are other courses that require more programming, at least two courses from the following list are essential. Examples of such courses include: S&DS 220 or 230; 262, 425, CPSC 100 or 112, or ENAS 130 (substitution of CPSC 201 or 223 is permitted)

Methods of Data Science These courses teach fundamental methods for dealing with data. They range from practical to theoretical. Every major must take at least two of these courses. Examples of such courses include: S&DS 313, 361, 363, 365, 430, 468, EENG 400, CPSC 477

Mathematical Foundations and Theory All students in the major must know linear algebra as taught in MATH 222 or 225. Students who have learned linear algebra through other courses (such as MATH 230, 231) may substitute another course from this category.

Students pursuing the B.S. degree must take at least two courses from this list and those students contemplating graduate school should take additional courses from this list as electives. Examples of such courses include: S&DS 364, 400, 410, 411, CPSC 365, 366, 469, MATH 222, 225, 244, 250, 260, 300, or 301

Efficient Computation and Big Data These courses are for students focusing on programming or implementation of large-scale analyses and are not required for the major. Students who wish to work in the software industry should take at least one of these. Examples of such courses include: CPSC 223, 323, 424, 437

Data Science in Context Students are encouraged to take courses that involve the study of data in application areas. Students learn how data are obtained, how reliable they are, how they are used, and the types of inferences that can be made from them. These course selections should be approved by the DUS. Examples of such courses include: ANTH 376, EVST 362, GLBL 191, 195, LING 229, 234, 380, PLSC 454, PSYC 258

Methods in Application Areas These are methods courses in areas of applications. They help expose students to the cultures of fields that explore data. These course selections should be approved by the DUS. Examples of such courses include: CPSC 453, 470, 475, ECON 136, 420, EENG 445, S&DS 352, LING 227

Substitution Some substitution, particularly of advanced courses, may be permitted with DUS approval.

Credit/D/Fail A maximum of one course taken Credit/D/Fail may be counted toward the requirements of the major, with permission of the DUS.

SENIOR REQUIREMENT

Students in both the B.A. degree program and B.S. degree program complete the senior requirement by taking a capstone course (S&DS 425) or an individual research project. Research projects include S&DS 490, S&DS 491, or S&DS 492, and must be advised by a member of the department of Statistics and Data Science or by a faculty member in a related discipline area. Students must complete a research project to be eligible for Distinction in the Major.

ADVISING

Statistics and Data Science can be taken either as a primary major or as one of two majors, in consultation with the DUS. Appropriate majors to combine with Statistics and Data Science include programs in the social sciences, natural sciences, engineering, computer science, or mathematics. A Statistics concentration is also available within the Applied Mathematics major.

REQUIREMENTS OF THE MAJOR

Prerequisites *Both degrees*—MATH 120, ENAS 151, MATH 230, or equivalent
Number of courses *B.A.*—11 term courses beyond prereqs (incl senior req); *B.S.*—14 term courses beyond prereqs (incl senior req)
Specific courses required *B.A.*—MATH 222 or 225; *B.S.*—same, plus S&DS 242
Distribution of courses *B.A.*—2 courses from Core Probability and Statistics, 2 courses from Computational Skills, 2 courses from Methods of Data Science, and 3 electives chosen from any discipline area with DUS approval; *B.S.*—same, plus 2 additional electives from any discipline area (except Data Science in Context and Methods in Application Areas) with DUS approval
Substitution permitted With DUS approval
Senior requirement *Both degrees*—Senior Seminar (S&DS 490) or Senior Project (S&DS 491 or S&DS 492) or Statistical Case Studies (S&DS 425)

FACULTY OF THE DEPARTMENT OF STATISTICS AND DATA SCIENCE

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 (*Acting Chair*), Hemant Tagare (*Radiology & Biomedical Engineering*), Van Vu (*Mathematics*), †Heping Zhang (*Public Health & Biostatistics*), †Hongyu Zhao (*Public Health & Biostatistics*), 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, Amin Karbasi (*Electrical Engineering*), 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

†A joint appointment with primary affiliation in another department or school.

S&DS 101—106, Introduction to Statistics and Data Science

A basic introduction to statistics, including numerical and graphical summaries of data, probability, hypothesis testing, confidence intervals, and regression. Each course in this group 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 of classes are attended by all students in S&DS 101–106 together, as general concepts and methods of statistics are developed. The remaining weeks are divided into field-specific sections that 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. No prerequisites beyond high school algebra. May not be taken after S&DS 100 or 109.

Students enrolled in S&DS 101–106 who wish to change to S&DS 109, or those enrolled in S&DS 109 who wish to change to S&DS 101–106, must submit a course change notice, signed by the instructor, to their residential college dean by Monday, October 2. The approval of the Committee on Honors and Academic Standing is not required.

S&DS 101a / E&EB 210a, Introduction to Statistics: Life Sciences Jonathan Reuning-Scherer and Walter Jetz
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. QR

S&DS 102a / EP&E 203a / PLSC 452a, 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. QR

S&DS 103a / EP&E 209a / PLSC 453a, 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. QR

S&DS 105a, Introduction to Statistics: Medicine Jonathan Reuning-Scherer
Statistical methods used in medicine and medical research. Practice in reading medical literature competently and critically, as well as practical experience performing statistical analysis of medical data. QR

Courses in Statistics and Data Science

S&DS 100b, Introductory Statistics Staff
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. May not be taken after S&DS 101–106 or 109. QR

S&DS 109a, Introduction to Statistics: Fundamentals Jonathan Reuning-Scherer
General concepts and methods in statistics. Meets for the first half of the term only. May not be taken after S&DS 100 or 101–106.
½ Course cr

[S&DS 110, An Introduction to R for Statistical Computing and Data Science]

S&DS 123b / S&DS 523b, YData: An Introduction to Data Science Jessica Cisewski and Staff
Computational, programming, and statistical skills are no longer optional in our increasingly data-driven world; these skills are essential for opening doors to manifold research and career opportunities. This course aims to dramatically enhance knowledge and capabilities in fundamental ideas and skills in data science, especially computational and programming skills along with inferential thinking. YData is an introduction to Data Science that emphasizes the development of these skills while providing opportunities for hands-on experience and practice. YData is accessible to students with little or no background in computing, programming, or statistics, but is also engaging for more technically oriented students through extensive use of examples and hands-on data analysis. Python 3, a popular and widely used computing language, is the language used in this course. The computing materials will be hosted on a special purpose web server. QR

* **S&DS 160b / AMTH 160b / MATH 160b, The Structure of Networks** Staff
Network structures and network dynamics described through examples and applications ranging from marketing to epidemics and the world climate. Study of social and biological networks as well as networks in the humanities. Mathematical graphs provide a simple common language to describe the variety of networks and their properties. QR

S&DS 220b, Introductory Statistics, Intensive Xiaofei Wang
Introduction to statistical reasoning 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 taught and used extensively, as well as application of statistical concepts to analysis of real-world data science problems. MATH 115 is helpful, but not a required. QR

S&DS 230a 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. Prerequisite: a 100-level Statistics course or equivalent, or with permission of instructor. QR

S&DS 238a, Probability and Statistics Joseph Chang
Fundamental principles and techniques of probabilistic thinking, statistical modeling, and data analysis. Essentials of probability, including conditional probability, random variables, distributions, law of large numbers, central limit theorem, and 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 for calculations, simulations, and analysis of data. After MATH 118 or 120. QR

S&DS 241a / MATH 241a, Probability Theory Yihong Wu
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. After or concurrently with MATH 120 or equivalent. QR

S&DS 242b / MATH 242b, Theory of Statistics Andrew Barron
Study of the principles of statistical analysis. Topics include maximum likelihood, sampling distributions, estimation, confidence intervals, tests of significance, regression, analysis of variance, and the method of least squares. Some statistical computing. After S&DS 241 and concurrently with or after MATH 222 or 225, or equivalents. QR

S&DS 262a / AMTH 262a / CPSC 262a, Computational Tools for Data Science Staff
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, programming. Prerequisites: after or concurrently with MATH 222, 225, or 231; after or concurrently with MATH 120, 230, or ENAS 151; after or concurrently with CPSC 100, 112, or ENAS 130. QR

S&DS 312a, Linear Models Winston Lin
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. After S&DS 242 and MATH 222 or 225. QR

S&DS 351b / EENG 434b / ENAS 496b / MATH 251b, Stochastic Processes Yihong Wu
Introduction to the study of random processes including linear prediction and Kalman filtering, Poison counting process and renewal processes, Markov chains, branching processes, birth-death processes, Markov random fields, martingales, and random walks. Applications chosen from communications, networking, image reconstruction, Bayesian statistics, finance, probabilistic analysis of algorithms, and genetics and evolution. Prerequisite: S&DS 241 or equivalent. QR

S&DS 361b / AMTH 361b, Data Analysis Staff
Selected topics in statistics explored through analysis of data sets using the R statistical computing language. Topics include linear and nonlinear models, maximum likelihood, resampling methods, curve estimation, model selection, classification, and clustering. After S&DS 242 and MATH 222 or 225, or equivalents. QR

S&DS 363b, Multivariate Statistics for Social Sciences Jonathan Reuning-Scherer
Introduction to the analysis of multivariate data as applied to examples from the social sciences. Topics include principal components analysis, factor analysis, cluster analysis (hierarchical clustering, k-means), discriminant analysis, multidimensional scaling, and structural equations modeling. Extensive computer work using either SAS or SPSS programming software. Prerequisites: knowledge of basic inferential procedures and experience with linear models. QR

S&DS 364b / AMTH 364b / EENG 454b, 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 and finance. After STAT 241. QR

S&DS 365a or b, Applied Data Mining and Machine Learning Staff
Techniques for data mining and machine learning from both statistical and computational perspectives, including support vector machines, bagging, boosting, neural networks, and other nonlinear and nonparametric regression methods. Discussion includes 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 to apply them to data. After S&DS 242. QR

S&DS 400b / MATH 330b, 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 assumed. QR

S&DS 410a, 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. Prerequisite: level of S&DS 241.

* **S&DS 425b, 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 statistical software. Prerequisites: prior course work in probability and statistics, and a data analysis course at the level of STAT 361, 363, or 365 (or STAT 220, 230 if supported by other course work). QR

* **S&DS 430a / AMTH 437a / ECON 413a / EENG 437a, 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. Prerequisites: MATH 120 and 222, or equivalents. May not be taken after AMTH 237. QR

* **S&DS 480a or b, Individual Studies** Staff
Directed individual study for qualified students who wish to investigate an area of statistics not covered in regular courses. A student must be sponsored by a faculty member who sets the requirements and meets regularly with the student. Enrollment requires a written plan of study approved by the faculty adviser and the director of undergraduate studies.

[S&DS 490, Senior Seminar and Project]

S&DS 491a and S&DS 492b, Senior Project Sekhar Tatikonda
Individual research that fulfills the senior requirement. Requires a faculty adviser and DUS permission. The student must submit a written report about results of the project.

GRADUATE COURSES OF PARTICULAR INTEREST TO UNDERGRADUATES

Courses in the Graduate School are open to qualified undergraduates. Descriptions of graduate courses in Statistics & Data Science are available on the departmental website. Permission of the instructor and of the director of graduate studies is required.