STATISTICS AND DATA SCIENCE

Director of undergraduate studies: Sekhar Tatikonda (sekhar.tatikonda@yale.edu), Rm. 338, 17 Hillhouse Ave., 432-4714; statistics.yale.edu; Major FAQ and guide; undergraduate major checklist

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. This field 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.

COURSES FOR NONMAJORS AND MAJORS

S&DS 100 and S&DS 101–109 and S&DS 123 (YData) 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 nonmajors include S&DS 110 and 160.

PREREQUISITES

Multivariable calculus is 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 or a 100-level course followed by S&DS 230. Students should complete the calculus prerequisite and linear algebra requirement (MATH 222 or 225) 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 228, 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, course 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 director of undergraduate studies (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 course. Courses for research opportunities 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**

Students intending to major in Statistics and Data Science should consult the department’s guide and FAQ. 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.

**Combined B.S./M.S. degree program** Exceptionally able and well-prepared students may complete a course of study leading to the simultaneous award of the B.S. and M.S. degrees after eight terms of enrollment. See Academic Regulations, section K, Special Arrangements, "Simultaneous Award of the Bachelor’s and Master’s Degrees. Interested students should consult the DUS prior to the sixth term of enrollment for specific requirements in Statistics and Data Science.

**Roadmap** See visual roadmap of the requirements.

**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)

**Certificate in Data Science**

The Certificate in Data Science is designed for students, majoring in disciplines other than Statistics & Data Science, to acquire the knowledge to promote mature use of data analysis throughout society. Students gain the necessary knowledge base and useful skills to tackle real-world data analysis challenges. Students who complete the requirements for the certificate are prepared to engage in data analysis in the humanities, social sciences, and sciences and engineering and are able to manage and investigate quantitative data research and report on that data.

Refer to the S&DS website for more information.
PREREQUISITE
The suggested prerequisite for the certificate is an introductory course, selected from one of the following courses, S&DS 100, 101–106, 123 or 220.

REQUIREMENTS OF THE CERTIFICATE
To fulfill the requirements of the certificate, students must take five courses from four different areas of statistical data analysis. No course may be applied to satisfy the requirements of both a major and the certificate. No single course may count for two areas of study. Students are required to earn at least a B– for each course.


Statistical Methodology and Data Analysis Two from S&DS 230, 242, 312, 361, 363. ECON 136 may be substituted for S&DS 242.

Computation & Machine Learning One from S&DS 262, 235, 365, CPSC 223, 477. CPSC 323 may be substituted for CPSC 223.

Data Analysis in a Discipline Area Either two of the half-credit seminars that accompany S&DS 123; or one of the “Data Science in a Discipline Area” courses approved for the data science certificate and listed on the S&DS website.

ADVISING
More information about the certificate, including how to register, is available on the S&DS website.

REQUIREMENTS OF THE CERTIFICATE
Prerequisite: 1 term course from S&DS 100, 101–106, 123 or 220

Number of courses: 5 term courses

Distribution of courses: 1 probability and statistical theory course; 2 statistical methodology and data analysis courses; 1 computational and machine learning course; and 2 half-credit courses or 1 course in discipline area, as specified

FACULTY OF THE DEPARTMENT OF STATISTICS AND DATA SCIENCE


Associate Professors: Timothy Armstrong, Peter Aronow, Forrest Crawford, Sahand Negahban, Sekhar Tatikonda, Yihong Wu

Assistant Professors: Elisa Celis, Jessi Cisewski-Kehe, Zhou Fan, Joshua Kalla, Amin Karbasi, Roy Lederman, Vahideh Manshadi, Fredrik Savje

Senior Lecturer: Jonathan Reuning-Scherer

Lecturers: Russell Barbour, Winston Lin

†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.


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


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
they use, how they deliver them, and among which voters. In this course, we explore how data science is being used to design winning

Political campaigns have become increasingly data driven. Data science is used to inform where campaigns compete, which messages

* S&DS 172b / EP&E 209a / PLSC 453a, Introduction to Statistics: Social Sciences  Staff
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 and Russell Barbour
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

S&DS 106a, Introduction to Statistics: Data Analysis  Jonathan Reuning-Scherer and William Brinda
An introduction to probability and statistics with emphasis on data analysis.  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

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 100b, Data Science Ethics  Elisa Celis
In this course, we introduce, discuss, and analyze ethical issues, algorithmic challenges, and policy decisions that arise when addressing real-world problems via the lens of data science. We grapple with the normative questions of what constitutes bias, fairness, discrimination, or ethics when it comes to data science and machine learning in applications such as policing, health, journalism, and employment. We incorporate technical precision by introducing quantitative measures that allow us to study how algorithms codify, exacerbate and/or introduce biases of their own, and study analytic methods of correcting for or eliminating these biases. Lastly, we study the social implications of these decisions, and understand the legal, political and policy decisions that could be used to govern data-driven decision making by making them transparent and auditable. We read critical commentary by practitioners, state-of-the-art technical papers by data scientist and computer scientists, and samples of legal scholarship, moral and ethical philosophy, readings in sociology, and policy documents. We often ground our discussions around recent case studies, controversies, and current events. Prerequisites: One from S&DS 238, S&DS 241, S&DS 242, or the equivalent; and one from S&DS 230, ECON 131, or the equivalent. Suggested courses: one from: EP&E 215, PHIL 175, PHIL 177, SOCY 144, PLSC 262, PLSC 320, or equivalent.  so

* S&DS 160b / AMTH 160b / MATH 160b, The Structure of Networks  Ronald Coifman
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 171b, YData: Text Data Science: An Introduction  Derek Feng
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. Text Data Science 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 YData (S&DS 123). Prerequisite: S&DS 123, which may be taken concurrently.  QR  ½ Course cr

* S&DS 172b / EP&E 328b / PLSC 347b, 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
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. This 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 123, YData. Prerequisite: S&DS 123, which may be taken concurrently. QR ½ Course cr

S&DS 220b, Introductory Statistics, Intensive  Joseph Chang
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 required. While no particular prior experience in computing is required, strong motivation to practice and learn computing are desirable. 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 or concurrently with MATH 118 or 120. QR

S&DS 241a / MATH 241a, Probability Theory  Winston Lin
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 262b / AMTH 262b, Computational Tools for Data Science  Roy Lederman
Introduction to the core ideas and principles that arise in modern data analysis, bridging statistics and computer science and providing students the tools to grow and adapt as methods and techniques change. Topics include principle component analysis, independent component analysis, dictionary learning, neural networks and optimization, as well as scalable computing for large datasets. Assignments will include implementation, data analysis and theory. Students require background in linear algebra, multivariable calculus, probability and 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; after S&DS 100-108 or S&DS 230 or S&DS 241 or S&DS 242. QR

S&DS 312a, 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. After S&DS 242 and MATH 222 or 225. QR

* S&DS 314b, Introduction to Causal Inference  Winston Lin
Introduction to causal inference with applications to the social and health sciences. Topics include randomized experiments, matching and propensity score methods, sensitivity analysis, instrumental variables, and regression discontinuity designs. Mathematical problems, data analysis in R, and critical discussions of published applied research. Prerequisite: S&DS 242 and some programming experience in R. QR

S&DS 315a / PLSC 340a, Measuring Impact and Opinion Change  Joshua Kalla
This course introduces students to measuring impact. Political campaigns, marketers, governments, and non-profit organizations regularly try to produce opinion change through TV, radio, online ads, mail, and door-to-door canvassing. Are these efforts successful at producing opinion change? In this course, we learn how to use experiments and natural experiments to measure the impact of opinion change efforts, and how to be appropriately skeptical of findings that claim to measure impact. This course also teaches data analysis skills in R. Prerequisite: S&DS 242 and some programming experience in R. QR

S&DS 351b / EENG 434b / MATH 351b, Stochastic Processes  Amin Karbasi
Introduction to the study of random processes including linear prediction and Kalman filtering, Poisson counting process and renewal processes, Markov chains, branching processes, death-birth 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 352b / MB&B 452b / MCDB 452b, Biomedical Data Science, Mining and Modeling  Mark Gerstein and Matthew Simon
Techniques in data mining and simulation applied to bioinformatics, the computational analysis of gene sequences, macromolecular structures, and functional genomics data on a large scale. Sequence alignment, comparative genomics and phylogenetics, biological
databases, geometric analysis of protein structure, molecular-dynamics simulation, biological networks, microarray normalization, and machine-learning approaches to data integration. Prerequisites: MB&B 301 and MATH 115, or permission of instructor.

**S&DS 355a, 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 several others. Programming is central to the course, and is based on the Python programming language. Prerequisites: Two of the following courses: S&DS 230, 238, 240, 241 and 242; previous programming experience (e.g., R, Matlab, Python, C++), Python preferred. 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**  Derek Feng
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 364b / AMTH 364b, 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**  John Emerson
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**  Sekhar Tatikonda
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.