

Statistics and Data Science

Directors of undergraduate studies: Sekhar Tatikonda
(sekhar.tatikonda@yale.edu) and Brian MacDonald; (brian.macdonald@yale.edu)
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 1000, 1090, and 1230 (YData) assume knowledge of high-school mathematics only. S&DS 1080 requires some previous coursework in statistics such as high school AP Statistics. Students who complete one of these courses should consider taking S&DS 2300. This sequence provides a solid foundation for the major. Other courses for nonmajors include S&DS 1100 and S&DS 1600.

PREREQUISITES

Multivariable calculus is required and should be taken before or during the sophomore year. This requirement may be satisfied by one of MATH 1200, ENAS 1510, MATH 3020, or the equivalent.

REQUIREMENTS OF THE MAJOR

See Links to the attributes indicating courses approved for the Statistics and Data Science major requirements.

Students who wish to major in Statistics and Data Science are encouraged to take S&DS 2200 or a 1000-level course followed by S&DS 2300. Students should complete the calculus prerequisite and linear algebra requirement (MATH 2220 or MATH 2250 or MATH 2260) 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 2220 or MATH 2250 or

MATH 2260 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 subject to DUS approval. 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. Specifically, B.S. degree candidates must take S&DS 2420 and S&DS 3650 to fulfill the B.A. requirements. The three remaining courses include one course chosen from the Mathematical Foundations and Theory discipline and two courses chosen from Core Probability and Statistics (not including S&DS 2420), Computational Skills, Methods of Data Science (not including S&DS 3650), Mathematical Foundations and Theory, or Efficient Computation and Big Data discipline areas subject to DUS approval.

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 2420.

Examples of such courses include: S&DS 2380, S&DS 2410, S&DS 2420, S&DS 3120, S&DS 3510

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 2200 or S&DS 2300, S&DS 2620, S&DS 2650, S&DS 4250, CPSC 1001, or CPSC 2010 or ENAS 1300

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. Students completing the B.S. degree must take S&DS 3650.

Examples of such courses include: S&DS 3120, S&DS 3170, S&DS 3610, S&DS 3630, S&DS 3650, S&DS 4300, S&DS 4310, S&DS 4680, ECE 4000, CPSC 4460, CPSC 4520, CPSC 4770

Mathematical Foundations and Theory All students in the major must know linear algebra as taught in MATH 2220 or MATH 2250 or MATH 2260. Students who have learned linear algebra through other courses 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 3640, S&DS 4000, S&DS 4100, S&DS 4110, CPSC 3650, CPSC 3660, CPSC 4690, MATH 2220, MATH 2250, MATH 2260, MATH 2440, MATH 2550, MATH 2560, MATH 2600, or MATH 3020

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 2230, CPSC 3230, CPSC 4240, CPSC 4381

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 3476, EVST 3620, GLBL 3191, 3195, LING 3290, 2340, 3800, PSYC 2658

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 4530, CPSC 4700, CPSC 4750, ECON 2136, ECON 4420, EENG 445, S&DS 3520, LING 2270

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

Credit/D/Fail No course taken Credit/D/Fail may be applied toward the requirements of the major, including prerequisites.

Outside credit Courses taken at another institution or during an approved summer or term-time study abroad program may count toward the major requirements with DUS approval.

SENIOR REQUIREMENT

Students in both the B.A. degree program and B.S. degree program complete the senior requirement by completing an individual research project. Courses for research opportunities include S&DS 4910 or S&DS 4920 (but not both), 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.

ADVISING

Students intending to major in Statistics and Data Science should consult the department 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.A. degree program Exceptionally able and well-prepared students may complete a course of study leading to the simultaneous award of the B.S. in S&DS and M.A. in Statistics after eight terms of enrollment. See Academic Regulations, section L, Special Academic Arrangements, “Simultaneous Award of the Bachelor’s and Master’s Degrees.” Interested students should consult the DUS at the beginning of their fifth term of enrollment for specific requirements in Statistics and Data Science.

SUMMARY OF MAJOR REQUIREMENTS

Prerequisites Both degrees— one of MATH 1200, ENAS 1510, MATH 3020, 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 2220 or MATH 2250 or MATH 2260; B.S.— same as B.A. degree, although 1 Core Probability and Statistics course must be S&DS 2420 and 1 Methods of Data Science course must be S&DS 3650

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 1 Mathematical Foundations and Theory course and 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 Project (S&DS 4910 or S&DS 4920)

CERTIFICATE IN DATA SCIENCE

The Certificate in Data Science is designed for students majoring in disciplines other than Statistics and 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. Students must declare their intent to earn a certificate by the last day of add/drop period in their final term of enrollment. This is done on the *Declare Major, Concentration within the Major, Certificate* page on Yale Hub. Once declared, Degree Audit will track students' progress toward completion of the certificate.

PREREQUISITE

The suggested prerequisite for the certificate is an introductory course, selected from one of the following courses: S&DS 1000, 1080, 1090, or 1230, or an introductory data analysis course from another department.

REQUIREMENTS OF THE CERTIFICATE

See Links to courses approved for the statistical data analysis requirements.

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.

Probability and Statistical Theory One from S&DS 2380, S&DS 2400, S&DS 2410, S&DS 2420. Advanced students may substitute S&DS 3510 or S&DS 3640 or ECE 4310.

Students are held to the Statistical Methodology and Data Analysis requirements that were in place when they declared their intent to earn the S&DS Certificate. However, with approval from the director of undergraduate studies (DUS), the following requirements, updated for the academic year 2024–2025, may be fulfilled by students who declared their intent to earn the certificate in a prior term.

Statistical Methodology and Data Analysis Two from S&DS 2200 or S&DS 2300 (but not both), S&DS 2420, S&DS 3120, S&DS 3610, S&DS 3630, PLSC 2501. ECON 2136 may be substituted for S&DS 2420.

Computation & Machine Learning One from S&DS 2620, S&DS 2650, S&DS 3170, S&DS 3650, CPSC 2230, CPSC 3810, CPSC 4770, PHYS 3780, PLSC 5060, CPSC 3230 may be substituted for CPSC 2230.

Data Analysis in a Discipline Area One course from those approved for this requirement and listed on the S&DS website.

ADVISING

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

SUMMARY OF REQUIREMENTS

Prerequisite 1 term course from S&DS 1000, 1080, 1090, or 1230 (or an introductory data analysis course in another department)

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 1 course in discipline area, as specified

FACULTY OF THE DEPARTMENT OF STATISTICS AND DATA SCIENCE

Professors †Donald Andrews, †P. M. Aronow, Andrew Barron, †Jeffrey Brock, Joseph Chang, †Katarzyna Chawarska, †Xiaohong Chen, Yuejie Chi, †Nicholas Christakis, †Ronald Coifman, †James Duncan, John Emerson (*Adjunct*), †Alan Gerber, †Mark Gerstein, Anna Gilbert, John Hartigan (*Emeritus*), †Edward Kaplan, †Harlan Krumholz, John Lafferty, Zongming Ma, David Pollard (*Emeritus*), †Nils Rudi, Jasjeet Sekhon, †Donna Spiegelman, Daniel Spielman, †Hemant Tagare, †Van Vu, Yihong Wu, †Heping Zhang, †Hongyu Zhao, Harrison Zhou, †Steven Zucker

Associate Professors †Forrest Crawford, Zhou Fan, †Joshua Kalla, †Amin Karbasi, †Vahideh Manshadi, Sekhar Tatikonda

Assistant Professors Elisa Celis, Sinho Chewi, †Melody Huang, Roy Lederman, Shuangping Li, Lu Lu, Theodor Misiakiewicz, Omar Montasser, †Dustin Scheinost, †Ramina Sotoudeh, †Andre Wibisono, Zhuoran Yang, †Ilker Yildirim, Ilias Zadik

Senior Lecturers †William Casey King, Brian Macdonald, Ethan Meyers, Jonathan Reuning-Scherer

Lecturer Robert Wooster

Preceptors Lynda Aouar, Addison McGhee, Shivam Sharma, Alberto Stefanelli

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