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) assume knowledge of high-school mathematics only. 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, MATH 302, 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 or 226) 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 or MATH 226 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. (B.S. degree candidates must take S&DS 242 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 242), Computational Skills, Methods of Data Science, 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 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 312, 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 or MATH 226. 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, MATH 226 244, 250, MATH 255, MATH 256, 260, 300, 301, or MATH 302

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 Credit/D/Fail may not be counted toward the requirements of the major (this includes prerequisite courses).

Roadmap See visual roadmap of the requirements.

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.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 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.

REQUIREMENTS OF THE MAJOR

Prerequisites *Both degrees*—MATH 120, ENAS 151, MATH 230, MATH 302, 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 or MATH 226; *B.S.*—same, plus 1 Core Probability and Statistics course must be 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 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 Seminar (S&DS 490) or Senior Project (S&DS 491 or S&DS 492) or Statistical Case Studies (S&DS 425)

Statistics and data science is the art of answering complex questions from numerical facts, called data. The mathematical foundation of statistics lies in the theory of probability, which is applied to make 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. Knowledge of statistics is necessary for conducting research in the sciences, medicine, industry, business, and government. 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.

S&DS 100 and the 101–106 group provide an introduction to statistics and data science with no mathematics prerequisite. These courses are alternatives; they do not form a sequence. Each course in the S&DS 101–106 group emphasizes 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 (life sciences, political science, social sciences, medicine, or data analysis). The half-term, half-credit course S&DS 109 offers the same introduction to statistics as the 101–106 group, but without applications to a specific field.

S&DS 123 (YData) is an introduction to data science that emphasizes developing skills, especially computational and programming skills, along with inferential thinking. YData is designed to be accessible to students with little or no background in computing, programming, or statistics, but is also engaging for more technically oriented students through the extensive use of examples and hands-on data analysis. In addition, there are associated YData seminars, half-credit courses in a specific domain developed for extra hands-on experience motivated by real problems in a specific domain.

S&DS 230 emphasizes practical data analysis and the use of the computer and has no mathematics prerequisite.

For students with sufficient preparation in mathematics, S&DS 238 covers essential ideas of probability and statistics, together with an introduction to data analysis using modern computational tools.

The sequence S&DS 241 and S&DS 242 offers the mathematical foundation for the theory of probability and statistics, and is required for most higher-level courses. Some courses require only S&DS 241 as a prerequisite.

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:c S&DS 100, 101–109, 123 or 220, or an introductory data analysis course from another department.

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.

Probability and Statistical Theory One from S&DS 238, 240, 241, 242. Advanced students may substitute S&DS 351 or S&DS 364 or EENG 431.

Statistical Methodology and Data Analysis Two from S&DS 230, 242, 312, 361, 363, PLSC 349. ECON 136 may be substituted for S&DS 242.

Computation & Machine Learning One from S&DS 262, 265, 317, S&DS 355, 365, CPSC 223, 477, PHYS 378, PLSC 468. CPSC 323 may be substituted for CPSC 223.

Data Analysis in a Discipline Area Two half-credit courses or one full-credit 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.

REQUIREMENTS

Prerequisite 1 term course from S&DS 100, 101–109, 123 or 220 (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 2 half-credit courses or 1 course in discipline area, as specified

FACULTY OF THE DEPARTMENT OF STATISTICS AND DATA SCIENCE

Professors †Donald Andrews, Andrew Barron, †Jeffrey Brock, Joseph Chang, †Katarzyna Chawarska, †Xiaohong Chen, †Nicholas Christakis, †Ronald Coifman, †James Duncan, John Emerson (*Adjunct*), †Debra Fischer, †Alan Gerber, †Mark Gerstein, Anna Gilbert,

John Hartigan (*Emeritus*), †Edward Kaplan, †Harlan Krumholz, John Lafferty, David Pollard (*Emeritus*), †Nils Rudi, Jasjeet Sekhon, †Donna Spiegelman, Daniel Spielman, †Hemant Tagare, †Van Vu, †Heping Zhang, †Hongyu Zhao, Harrison Zhou, †Steven Zucker

Associate Professors †Peter Aronow, †Forrest Crawford, Ethan Meyers (*Visiting*), Sahand Negahban, Sekhar Tatikonda, Yihong Wu

Assistant Professors Elisa Celis, Zhou Fan, †Joshua Kalla, †Amin Karbasi, Roy Lederman, †Vahideh Manshadi, †Fredrik Savje, †Ilker Yildirim

Senior Lecturer Jonathan Reuning-Scherer

Lecturers William Brinda, Elena Khusainova

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

View Courses