STATISTICS AND DATA SCIENCE (S&DS)

S&DS 100b, Introductory Statistics  Ethan Meyers
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 101a / E&EB 210a, Introduction to Statistics: Life Sciences  Walter Jetz and 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. QR

S&DS 102a / EPRE 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 / EPRE 209a / PLSC 453a, Introduction to Statistics: Social Sciences  Jonathan Reuning-Scherer
Descriptive and inferential analysis 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  Ethan Meyers and 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

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

Introductory statistical concepts beyond those covered in high school AP statistics. Includes additional concepts in regression, an introduction to multiple regression, ANOVA, and logistic regression. This course is intended as a bridge between AP statistics and courses such as S&DS 230, Data Exploration and Analysis. Meets for the second half of the term only. Prerequisites: A previous statistics course in high school. May not be taken after S&DS 100, S&DS 101–106, PSYC 100, or any other full semester Yale introductory statistics courses. Students should consider S&DS 103 or both S&DS 108, 109. ½ Course cr

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 123b / CPSC 123b / PLSC 351b / S&DS 523b, YData: An Introduction to Data Science  Ethan Meyers
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 175b / SOCY 163b, YData: Measuring Culture  Daniel Karell
Culture is increasingly digital. Cultural objects, such as songs and artwork, are frequently digitized. Creating culture objects often involves digital tools and takes place in digital domains. The effects of culture on our social lives are now typically mediated by digital platforms and devices. In this introductory course, we explore how data science is being used to measure the cultural landscape, the consumption and production of culture, and the impact of culture on society. To do so, we review current theories and methodologies, as well as conduct our own analyses of popular culture, the rhetoric and social connections underlying online extremist communities, and other topics. The course provides opportunities to practice the data science skills presented in S&DS 123 with applications to the social scientific study of culture. This course can be taken concurrently with S&DS 123 or after successfully completing it. QR, SO

* S&DS 176b, YData: Humanities Data Mining  Catherine DeRose
What new modes of inquiry become available when we transform novels into bags of words and images into pixels? What is lost in the process? This course explores how we can use computational methods to pursue questions in the humanities, while also looking at how humanistic methods can inform the work of algorithms in research and society at large. We begin this course with a series of questions at the intersections of the humanities and quantitative analysis: What is data? How can we turn texts into data? To explore these questions from both theoretical and technical perspectives, each course week is divided into discussion and lab sessions. Discussion sessions introduce concepts and humanities-based case studies that ground the hands-on technical work done in the labs. We survey
some of the most popular methods in modern data science—classification, vectorization, and visualization—to see what kinds of questions we can ask and answer. We conclude the semester with open lab sessions during which students leverage the skills covered in this course to create their own data science projects with cultural heritage data.

S&DS 200b, Introductory Statistics, Intensive  Staff
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 is desirable.   QR

S&DS 230a or b / S&DS S230E, 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 240a, An Introduction to Probability Theory  Elisa Celis
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 counts towards the Data Science certificate but not the Statistics and Data Science major. Prerequisite: MATH 115.  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  William Brinda and 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  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 principal component analysis, independent component analysis, dictionary learning, neural networks and optimization, as well as scalable computing for large datasets. Assignments 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 240-248 or S&DS 230 or S&DS 241 or S&DS 242. Enrollment is limited; requires permission of the instructor.  QR

S&DS 265a, 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 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 317b, Applied Machine Learning and Causal Inference  Jas Sekhon
We cover approaches to causal inference using machine learning. Machine learning methods include bagging, boosting, random forests, and neural networks. Causal topics include randomized experiments with and without noncompliance, observational studies with and without ignorable treatment assignment, instrumental variables, and regression discontinuity. Assignments provide students with hands-on experience with the methods. Applications are drawn from a variety of fields including political science, economics, public health, and medicine. Programming is central to the course and is based on the R programming language. Prerequisites: The equivalent of at
least two of the following courses: S&DS 230, 238, 241 and 242; previous programming experience (e.g., R, Matlab, Python, C++), R preferred. Strong knowledge of OLS is assumed.  

**S&DS 351b / EENG 434b / MATH 251b, Stochastic Processes**  
Staff  
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.  

**S&DS 352b / MB&B 452b / MCD 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 361b / AMTH 361b, Data Analysis**  
Elena Khusainova  
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.  

**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.  

**S&DS 364b / AMTH 364b / EENG 454b, Information Theory**  
Andrew Barron  

**S&DS 365b, Intermediate Machine Learning**  
John Lafferty  
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.  

**S&DS 400b / MATH 330a, 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.  

**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 425a or b, 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 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).  

**S&DS 431a / AMTH 431a, Optimization and Computation**  
Anna Gilbert  
This course is designed for students in Statistics & Data Science who need to know about optimization and the essentials of numerical algorithm design and analysis. It is an introduction to more advanced courses in optimization. The overarching goal of the course is teach students how to design algorithms for Machine Learning and Data Analysis (in their own research). This course is not open to students who have taken S&DS 430. Prerequisites: Knowledge of linear algebra, multivariate calculus, and probability. Linear Algebra, by MATH 222, 223 or 230 or 231; Graph Theory, by MATH 244 or CPSC 365 or 366; and comfort with proof-based exposition and problem sets, such as is gained from MATH 230 and 231, or CPSC 366.  

**S&DS 432b, Advanced Optimization Techniques**  
Sekhar Tatikonda  
This course covers fundamental theory and algorithms in optimization, emphasizing convex optimization. Topics covered include convex analysis; duality and KKT conditions; subgradient methods; interior point methods; semidefinite programming; distributed methods; stochastic gradient methods; robust optimization; and an introduction to nonconvex optimization. Applications accepted from statistics & data science, economics, engineering, and the sciences. Prerequisites: Knowledge of linear algebra, such as MATH 222, 225; multivariate calculus, such as MATH 120; probability, such as S&DS 241/541; optimization, such as S&DS 431/631; and, comfort with proof-based exposition and problem sets.
* 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 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.