Yale School of Public Health
Ph.D. in Biostatistics
Curriculum (2020-2021 Matriculation)

The Ph.D. degree requires a total of 16 course units.
If a course is waived, a substitute course must be identified, approved by the student's advisor and the DGS.

Course	Title	Units	Term Offered	Term Taken	Notes
PhD Required Courses (10 course units)					
BIS 525	Seminar in Biostatistics and Journal Club	0	Fall	1st year	
BIS 526	Seminar in Biostatistics and Journal Club	0	Spring	1st year	
BIS 610	Applied Area Readings for Qualifying Exams	1	Spring	2nd year	
BIS 623	Advanced Regression Analysis or S&DS 612, Linear Models	1	Fall		
BIS 628	Longitudinal and Multilevel Data Analysis	1	Spring		
BIS 643	Theory of Survival Analysis	1	Spring		
BIS 678	Statistical Practice I	1	Fall	2nd year	
BIS 681	Statistical Practice II	1	Spring	2nd year	
BIS 691	Theory of Generalized Linear Models	1	Spring		
BIS 695	Summer Internship in Biostatistical Research	0	Summer		
BIS 508	Foundations of Epidemiology and Public Health	1	Fall	1st year	
EPH 600	Research Ethics and Responsibilities	0	Fall	1st year	
EPH 608	Frontiers of Public Health*	1	Fall and Spring	1st year preferred	One term only is requierd
S&DS 610	Statistical Inference	1	Fall		
PhD Elective Courses (6 course units) List of approved electives attached to this document (will be updated closer to the fall term)					
Other Courses					

* Students entering the program with an MPH or relevant graduate degree may be exempt from this requirement.

Updated: 4/3/20

Possible additional electives for BIS

CDE 566a Causal Inference Methods in Public Health Research
Instructor: Zeyan Liew
This course will introduce the theory and applications of causal inference methods for public health research, tailored for analyses of epidemiologic data. The practice of epidemiological and public health research requires causal thinking. The rapid development of both the theoretical frameworks and applications of causal inference methods in recent years provide opportunities to improve the rigor of epidemiological research. The course will start by introducing the principles of causal logic including counterfactuals and probability logic. Next, the course will review epidemiological study designs and sources of biases, including misinterpretations of statistics. A significant amount of time will be dedicated to the applications of causal diagrams in epidemiology, a powerful tool that can provide a basis for identifying variables that must be measured and controlled to obtain unbiased effect estimates. Students will learn to apply specific common causal modeling techniques in epidemiological research using real-world and simulated data. Novel study designs that could strengthen causal inference in epidemiological research will be discussed. Students will leave the course with a basic knowledge of causal inference methods to apply in their own research and the ability to further explore the causal inference literature.

E&EB 678 Mathematical Models and Quantitative Methods in Evolution and Ecology
Instructor: Alvaro Sanchez De Andres
In this course, we focus on how quantitative approaches are used to allow scientific inference. We discuss general principles for generating hypotheses that are testable (i.e., quantifiable). The course also examines a variety of approaches used to model population-level processes in evolution and ecology, including an overview of population genetics, quantitative genetics, optimality models, game theory, and population dynamic equations. We also discuss experimental design, statistical analyses, inference, and other quantitative methods. The course assumes a basic background in algebra, calculus, probability theory, and statistics.

PLSC 504a Advanced Quantitative Methods
Instructor: Fredrik Sävje
The aim of this course is to provide students with the understanding and tools to critically consume and conduct statistical research. The theme is the challenge of drawing reliable causal inference. We will learn: how to use graphical methods to transparently analyze and present data; how to discipline our analyses against multiple-comparisons bias; how to use nonparametric methods to avoid implausible assumptions; how strong research design is essential to causal inference; how Bayesian inference provides the mathematical vocabulary for thinking about scientific inference; how causal graphs allow us to express and analyze causal assumptions, choose control variables, and think about selection bias; how placebo tests allow us to test assumptions; how to build and understand Likelihood and Bayesian models including Logistic and Probit models; how to think about and analyze time-series cross-sectional data. We will review instrumental variables methods and regression-discontinuity designs, though it is assumed that you have already covered these in PLSC 503. The course assumes students have command of the material covered in PLSC 500 and PLSC 503, including basic probability theory, matrix algebra, and the linear regression model.

PLSC 508b Causal inference and research design
Instructor: Winston Lin
This seminar exposes students to cutting-edge empirical and statistical research across the social and health sciences, with a focus on topics relevant to causal questions in the domain of political science. Readings and discussions focus on selected methodological topics, such as experimental design, partial identification, design-based inference, network analysis, semiparametric efficiency theory, and qualitative/mixed-methods research. Topics vary from year to year. Statistical training at the level of PLSC 504 is expected, though training in probability theory at the level of S&DS 541 or ECON 550 is suggested.

CB&B 555a (CPSC 553) Machine Learning for Biology
Instructor: Smita Krishnaswamy
This course introduces biology as a systems and data science through open computational problems in biology, the types of high-throughput data that are being produced by modern biological technologies, and computational approaches that may be used to tackle such problems. We cover applications of machine-learning methods in the analysis of high-throughput biological data, especially focusing on genomic and proteomic data, including denoising data; nonlinear dimensionality reduction for visualization and progression analysis; unsupervised clustering; and information theoretic analysis of gene regulatory and signaling networks. Students' grades are based on programming assignments, a midterm, a paper presentation, and a final project.

CB&B 740a Clinical and Translational Informatics
Instructors: Richard Shiffman, Michael Krauthammer
The course provides an introduction to clinical and translational informatics. Topics include (1) overview of biomedical informatics, (2) design, function, and evaluation of clinical information systems, (3) clinical decision making and practice guidelines, (4) clinical decision support systems, (5) informatics support of clinical research, (6) privacy and confidentiality of clinical data, (7) standards, and (8) topics in translational bioinformatics.
Permission of the instructor required.

CB&B 745 Advanced Topics in Machine Learning
An overview of advances in the past decade in machine learning and automatic data-mining approaches for dealing with the broad scope of modern data-analysis challenges, including deep learning, kernel methods, dictionary learning, and bag of words/features. This year, the focus is on a broad scope of biomedical data-analysis tasks, such as single-cell RNA sequencing, single-cell signaling and proteomic analysis, health care assessment, and medical diagnosis and treatment recommendations. The seminar is based on student presentations and discussions of recent prominent publications from leading journals and conferences in the field. Prerequisite: basic concepts in data analysis (e.g., CPSC 545 or 563) or permission of the instructor.

CB&B 752b Biomedical Data Science: Mining and Modeling
Instructor: Mark Gerstein
Biomedical data science encompasses the analysis of gene sequences, macromolecular structures, and functional genomics data on a large scale. It represents a major practical application for modern techniques in data mining and simulation. Specific topics to be covered include sequence alignment, large-scale processing, next-generation sequencing data, comparative genomics, phylogenetics, biological database design, geometric analysis of protein structure, molecular-dynamics simulation, biological networks, normalization of microarray data, mining of functional genomics data sets, and machine-learning approaches to data integration. Prerequisites: biochemistry and calculus, or permission of the instructor.

CPSC 546a or b Data and Information Visualization
Instructor: Holly Rushmeier
Visualization is a powerful tool for understanding data and concepts. This course provides an introduction to the concepts needed to build new visualization systems, rather than to use existing visualization software. Major topics are abstracting visualization tasks, using visual channels, spatial arrangements of data, navigation in visualization systems, using multiple views, and filtering and aggregating data. Case studies to be considered include a wide range of visualization types and applications in humanities, engineering, science, and social science. Prerequisite: CPSC 223.

S&DS 674b Applied Spatial Statistics
Instructor: Timothy Gregoire
An introduction to spatial statistical techniques with computer applications. Topics include modeling spatially correlated data, quantifying spatial association and autocorrelation, interpolation methods, variograms, kriging, and spatial point patterns. Examples are drawn from ecology, sociology, public health, and subjects proposed by students. Four to five lab/homework assignments and a final project. The class makes extensive use of the R programming language as well as ArcGIS.

S&DS 683 Statistical Methods in Neuroimaging
Instructors: Joseph Chang, Dustin Scheinost
Introduction to common statistical methods used in neuroimaging. Topics include introduction to different imaging modalities and experimental designs; modeling tasks using linear models; functional connectivity analysis; mixed effects, repeated measures, longitudinal models, power; multiple comparisons, random fields; effective connectivity, dynamic causal modeling, and variational Bayesian methods; machine-learning approaches to multi-voxel pattern analysis.

INP 599b Statistics and Data Analysis in Neuroscience
Instructors: John Murray, Daeyeol Lee, Hyojung Seo
This course focuses on practical applications of various statistical models and tests commonly used in neuroscience research. It covers basic probability theory, hypothesis testing, and maximum likelihood estimation, as well as model comparison. The specific models and tests covered include ANOVA, regression, time series analyses, and dimension reduction techniques (e.g., PCA). Examples and homework will be given in MATLAB, which will be introduced at the beginning of the course. Previous experience in programming and basic statistics is desirable but not required.

MGT 510b Data Analysis & Causal Inference
Instructor: Robert Jensen
This course will examine how and when data can be used specifically to infer whether there is a causal relationship between two variables. We will emphasize (a) the role of an underlying theory of behavior in interpreting data and guiding analysis, as well as (b) a range of advanced techniques for inferring causality from experimental and non-experimental data, such as: regression discontinuity designs; matching estimators; instrumental variables; synthetic controls; event studies; difference-in-differences; heterogeneity modeling; audit studies; randomized controlled trials; natural experiments (and unnatural experiments); and A/B testing. The issue of causality, and the relevance of thinking about models and methods for inferring causality, is just as central and important for "Big Data" as it is when working with traditional data sets in business and public policy. Although each lecture will contain some proofs and derivations (only calculus is required), the emphasis will be on understanding the underlying concepts, the practical use, implications and limitations of techniques. Students will work intensively with data, drawing from examples across many sectors and topics, to develop the skills to use data analysis to make better decisions. All analysis will be conducted using Stata. The goals of the course are for students to become expert consumers, able to interpret and evaluate empirical studies, as well as expert producers of convincing empirical analysis themselves.

MGT 556b Big Data & Customer Analytics
Instructor: Kosuke Uetake
In the age of Big Data, companies have access to data about markets, products, customers, and much more. When deciding on strategic issues such as pricing, advertising or targeting these data can be very valuable to companies if used correctly. This course will provide you with the tools and methods that will allow you to leverage such a rich data to help shape a marketing strategy from a quantitative perspective. While students will employ quantitative methods in the course, the goal is not to produce experts in statistics; rather, students will gain the competency to interact with and manage a data scientist team. In other words, we aim to develop working knowledge of data analysis, which is necessary for managers in the age of Big Data. The course uses a combination of lectures, cases, and many data analytic exercises to learn the material. In the second half of the course, we will work on the final project as well. This is your grand opus and will allow getting your hands dirty on real data. The idea is to gain some experience working with a real company data and to apply what you learn in the class to answer real business/research questions. So, you are welcome to locate customer data from a company you may know of or, if you have difficulty obtaining such a data, we can discuss alternatives. There is another course taught by the same instructor, titled “Advanced Customer Analytic,” which focuses more on advanced methodologies and targets to those who are more data scientists rather than managers.

MGT 803b Decision Making with Data
Instructor: Peter Schott
This course is designed as a follow-up to the SOM MBA core statistics class. Its goal is to give you lots of practice analyzing data and presenting the results of your analysis relying solely on the tools introduced in the core. It will revolve around weekly, open-ended "consulting assignments" of the form: "Your client is X. They give you dataset Y. They would like to know Z. Please prepare a short presentation succinctly summarizing and justifying your answer." Class time will be split among three activities: lab time devoted to discussing issues/problems that arise in the consulting assignments; mini-lectures covering issues related to data analysis and visualization, as needed; and discussion of consulting assignment presentations.

F&ES 775b Modeling Geographic Space
Instructor: Charles Tomlin
An introduction to the conventions and capabilities of image-based (raster) geographic information systems (GIS) for the analysis and synthesis of spatial patterns and processes. In contrast to F&ES 756, the course is oriented more toward the qualities of geographic space itself (e.g., proximity, density, or interspersion) than the discrete objects that may occupy such space (e.g., water bodies, land parcels, or structures). Three hours lecture, problem sets. No previous experience is required.

Updated 8/1/19