BGS Training Requirement in Statistics

All BGS students are required to have an understanding of statistical methods and their application to biomedical research. Biostatistics training is a key component of BGS' Required Training in Rigor and Reproducibility.

Most BGS students in CAMB, IGG, NGG and PGG will take *either* BIOM611 Statistical Methods for the Design and Analysis of Experiments *or* BIOM612 Biostatistics, Bioinformatics, and Experimental Design in the spring semester of their first year. These courses have the primary goals of teaching students concepts in statistics and how to apply these concepts to experimental design and analysis. In order to decide which course to take, students and their advisors should review the syllabi of both courses, along with the BIOM611 table of contents to compare the course's topics with previous courses taken. Additional guidance is provided below under Comparison of BIOM611 and BIOM612.

GCB and GGEB students are exempt from taking BIOM611 or BIOM612 because they take more advanced courses in statistics as part of their graduate group's requirements. Similarly, most BMB students take BMB 510 Data Analysis and Scientific Inference instead of BIOM611 or BIOM612. Students in CAMB, IGG, NGG, and PGG who have taken the equivalent of BIOM611 or BIOM612, perhaps through their undergraduate work, are required to take a more advanced statistics course. Intermediate and advanced course alternatives are listed below.

Combined degree students, with approval of their graduate group course advisors, may elect to postpone fulfilling their statistics training requirement until the spring following the candidacy exam.

Comparison of BIOM611 and BIOM612

> BIOM611 provides a foundation in statistics to students who may not have had significant undergraduate coursework in statistics or experience in programming. BIOM612 provides similar statistical concepts, but is more suited to students with a stronger background in quantitative and computational coursework or training. BIOM611 focuses on teaching statistics with minimal computer programming and uses Rcmdr. BIOM 612 includes teaching statistics in R Studio and R programming, although experience using R or R studio is not a requirement.
>
> BIOM611 will move more slowly than 612 and will give more time to specific methods using Rcmdr as the basic software. Rcmdr is menu driven. Using Rcmdr, students will learn to create an Rmarkdown file, and thus integrate the results of statistical analysis with text-based annotation and descriptions. By the last third of the course, students will be introduced to some simple coding. Support will be given to students who do not have previous experience with programming.
>
> BIOM612 covers similar topics as 611, but moves more quickly in terms of concepts and statistical programming, and moves into some advanced concepts in data analysis. It also presupposes comfort with basic statistics and understanding of basic programming concepts such as data types, arrays and matrices, functions, loops, and other similar concepts. Thus students who wish to take BIOM612 should have at least two of the following: comfort with content found in undergraduate statistics courses, some programming experience such as with

Python or another language, and experience with independent programing projects in or outside of an undergraduate course.

Intermediate and Advanced Course Options

Options for students seeking alternatives to BIOM611 or BIOM612, presented from intermediate to highly advanced level, are listed below. Students should consult with their graduate group course advisor to select an appropriate option.

BIOL 446 Statistics for Biologists (Fall) *– Intermediate*
Introductory probability theory, principles of statistical methods, problems of estimation and hypothesis testing in biology and related areas. (Note that this course does not make use of any statistical software packages.) Prerequisites: MATH 104 or equivalent; or permission of the instructor.

BMB 510 Data Analysis and Scientific Inference (Spring) *– Intermediate*
An introductory course in the analysis of quantitative data produced, for example, by Biochemistry and Molecular Biophysics experiments. The course will stress fundamental principles of data analysis, best practice in presenting data, and how to draw scientific inferences from the data. The overall goal is to provide students the tools to carry out rigorous and reproducible scientific research. Prerequisites: For non BMB students, permission of the instructor is needed.

EDUC 767 Regression and Analysis of Variance (Fall) *– Intermediate*
This course covers design of controlled randomized experiments, analysis of survey data and controlled field experiments, including statistics models, regression, hypothesis testing, relevant data analysis, and reporting. Prerequisites: EDUC 667 or equivalent.

EPID 527 Biostatistics for Epidemiologic Methods II (Mid-Fall to Mid-Spring) *– Intermediate*
The first half of this covers concepts in biostatistics as applied to epidemiology, primarily categorical data analysis, analysis of case-control, cross-sectional, cohort studies, and clinical trials. Topics include simple analysis of epidemiologic measures of effect; stratified analysis; confounding; interaction, the use of matching, and sample size determination. The second half of this course covers concepts in biostatistics as applied to epidemiology, primarily multivariable models in epidemiology for analyzing case-control, cross-sectional, cohort studies, and clinical trials. Topics include logistic, conditional logistics, and Poisson regression methods; simple survival analyses including Cox regression. Emphasis is placed on understanding the proper application and underlying assumptions of the methods presented. Laboratory sessions focus on the use of the STATA statistical package and applications to clinical data. Prerequisites: Permission of the instructor.

STAT 431/STAT 511 Statistical Inference (Fall) *- Intermediate*
Graphical displays; one- and two-sample confidence intervals; one- and two-sample hypothesis tests; one- and two-way ANOVA; simple and multiple linear least-squares regression; nonlinear regression; variable selection; logistic regression; categorical data analysis; goodness-of-fit tests. A methodology course. This course does not have business applications but has significant overlap with STAT 101 and 102. software. Prerequisites: STAT 430

***STAT 474/974 Modern Regression for the Social, Behavioral and Biological Sciences (Spring)** - Intermediate*
Function estimation and data exploration using extensions of regression analysis: smoothers, semiparametric and nonparametric regression, and supervised machine learning. Conceptual foundations are addressed as well as hands-on use for data analysis. Prerequisites: STAT 102 or 112 or equivalent

***STAT 500 - Applied Regression and Analysis of Variance (Fall) –** Intermediate*
This is an applied graduate level course in multiple regression and analysis of variance for students who have completed an undergraduate course in basic statistical methods. Emphasis is on practical methods of data analysis and their interpretation. Covers model building, general linear hypothesis, residual analysis, leverage and influence, one-way anova, two-way anova, factorial anova. The course is primarily for doctoral students in the managerial, behavioral, social, and health sciences. Prerequisites: STAT 102 or 112 or equivalent.

***BSTA 630 Statistical Methods for Categorical and Survival Data (Fall)** – Intermediate-Advanced*
This first course in statistical methods for data analysis is aimed at first-year Biostatistics students. It focuses on the analysis of continuous data. Topics include descriptive statistics (measures of central tendency and dispersion, shapes of distributions, graphical representations of distributions, transformations, and testing for goodness of fit); populations and sampling (hypotheses of differences and equivalence, statistical errors); one- and two-sample t tests; analysis of variance; correlation; nonparametric tests on means and correlations; estimation (confidence intervals and robust methods); categorical data analysis (proportions; statistics and test for comparing proportions; test for matched samples; study design); and regression modeling (simple linear regression, multiple regression, model fitting and testing, partial correlation, residuals, multicollinearity). Examples of medical and biologic data will be used throughout the course, and use of computer software demonstrated.
Prerequisite: Multivariable calculus and linear algebra, BSTA 620 (may be taken concurrently); permission of the instructor.

***BSTA 631 Statistics Methods and Data Analysis II (Spring)** – Advanced*
This is the second half of the methods sequence and focuses on categorical data and survival data. Topics in categorical data to be covered include defining rates, incidence and prevalence, the chi-squared test, Fisher's exact test and its extension, relative risk and odds-ratio, sensitivity, specificity, predictive values, logistic regression with goodness of fit tests, ROC curves, Mantel-Haenszel test, McNemar's test, the Poisson model, and the Kappa statistic. Survival analysis will include defining the survival curve, censoring, and the hazard function, the Kaplan-Meier estimate, Greenwood's formula and confidence bands, the log rank test, and Cox's proportional hazards regression models. Examples of medical and biologic data will be used throughout the course, and use of computer software demonstrated. Prerequisites: linear algebra, calculus, BSTA 630, BSTA 620, BSTA 621 (may be taken concurrently).

***EPID 621 Longitudinal and Clustered Data (Fall)** – Advanced*
An introduction to the principles of and methods for longitudinal and clustered data analysis with special emphasis on clinical, epidemiologic, and public health applications; marginal and conditional methods for continuous and binary outcomes; mixed effects and hierarchical models; and simulations for power calculations. Each student will be required to participate in 8 labs and complete associated problem sets. Knowledge of Stata and SAS. Prerequisites: Completion of EPID 527 or equivalent

preparation in biostatistics, including generalized linear models, principles of first-year calculus and matrix algebra.

EPID 622 Applied Regression Models for Categorical Data (Fall) .5 cu *– Advanced*
This course will provide in-depth treatment of several topics in categorical data analysis. After a brief review of methods for contingency tables, we will introduce the idea of generalized linear models, and focus on two special cases – multiple logistic regression and log linear models. Each topic will be presented in detail by stating the model and covering parameter estimation and interpretation, inference, model building, regression diagnostics and assessment of model fit. Finally, we will cover extensions to both models, including models for multinomial data, analysis of matched-pair data, and random effects models. Topics will be illustrated in class with examples, and we will discuss the use of Stata to conduct the analyses. Prerequisites: EPID 510 or equivalent and EPID 526 or equivalent.

EPID 623 Survival Data Analysis (Fall) .5 cu *– Advanced*
This course will focus on the specialized issues related to the analysis of survival or time-to-event data. The course begins by closely examining the features unique to survival data which distinguishes these data from other more familiar types. Topics include non-parametric survival analysis methods, common survival functions, parametric survival models, the proportional hazards model, and common model checking methods. All methods will be illustrated by in class examples and homework sets. Prerequisites: EPID 510 or equivalent and EPID 526 or equivalent, and permission of instructor.

NGG 594/PHYS 585/BE 530/PSYC 539 Theoretical Neuroscience (Spring) *– Advanced*
This course will develop theoretical and computational approaches to structural and functional organization in the brain. The course will cover: (i) the basic biophysics of neural responses, (ii) neural coding and decoding with an emphasis on sensory systems, (iii) approaches to the study of networks of neurons, (iv) models of adaptation, learning and memory, (v) models of decision making, and (vi) ideas that address why the brain is organized the way that it is. The course will be appropriate for advanced undergraduates and beginning graduate students. Prerequisites: Mathematics: knowledge of multi-variable calculus, some linear algebra and differential equations is necessary for this course. Neuroscience: basic knowledge of the architecture of the brain and of the mechanisms of neural signaling will be very useful.

STAT 471/571/701 Modern Data Mining (Fall) *–Advanced*
Modern Data Mining: Statistics or Data Science has been evolving rapidly to keep up with the modern world. While classical multiple regression and logistic regression technique continue to be the major tools we go beyond to include methods built on top of linear models such as LASSO and Ridge regression. Contemporary methods such as KNN (K nearest neighbor), Random Forest, Support Vector Machines, Principal Component Analyses (PCA), the bootstrap and others are also covered. Text mining especially through PCA is another topic of the course. While learning all the techniques, we keep in mind that our goal is to tackle real problems. Not only do we go through a large collection of interesting, challenging real-life data sets but we also learn how to use the free, powerful software "R" in connection with each of the methods exposed in the class. Prerequisites: STAT 102 or 112 or 431

STAT 501 Introduction to Nonparametric Methods and Log-linear Models (Spring) *– Advanced*
This is an applied graduate level course for students who have completed an undergraduate course in basic statistical methods. It covers two unrelated topics: log linear and logit models for discrete data and nonparametric methods for non-normal data. Emphasis is on practical methods of data analysis and their interpretation. Course is primarily for doctoral students in the managerial, behavioral, social and health sciences and may be taken before STAT 500 with permission of instructor. Prerequisites: STAT 102 or 112 or equivalent.

STAT 503 Data Analytics and Statistical Computing (Spring) *– Advanced*
This course will introduce a high-level programming language, called R, that is widely used for statistical data analysis. Using R, we will study and practice the following methodologies: data cleaning, feature extraction; web scrubbing, text analysis; data visualization; fitting statistical models; simulation of probability distributions and statistical models; statistical inference methods that use simulations (bootstrap, permutation tests). Prerequisites: Two courses at the statistics 400 or 500 level.

STAT 550 Mathematical Statistics (Fall) *– Advanced*
Decision theory and statistical optimality criteria, sufficiency, point estimation and hypothesis testing methods and theory. Prerequisites: STAT 431 or 520 or equivalent; comfort with mathematical proofs (e.g., MATH 360).

BSTA 751 Statistical Methods for Neuroimaging (Spring) – *Advanced*
This course is intended for students interested in both statistical methodology, and the process of developing this methodology, for the field of neuroimaging. This will include quantitative techniques that allow for inference and prediction from ultra-high dimensional and complex images. In this course, basics of imaging neuroscience and preprocessing will be covered to provide students with requisite knowledge to develop the next generation of statistical approaches for imaging studies. High-performance computational neuroscience tools and approaches for voxel- and region-level analyses will be studied. The multiple testing problem will be discussed, and the state-of-the art in the area will be examined. Finally, the course will end with a detailed study of multivariate pattern analysis, which aims to harness patterns in images to identify disease effects and provide sensitive and specific biomarkers. The student will be evaluated based on 3 homework assignments and a final in-class presentation. Prerequisites: BSTA 621, BSTA 651; permission of instructor.

BSTA 787 Methods for Statistical Genetics and Genomics in Complex Human Diseases (Spring) *– Highly Advanced*
This is an advanced elective course for graduate students in Biostatistics, Statistics, Epidemiology, Bioinformatics, Computational Biology, and other BGS disciplines. This course will cover statistical methods for the analysis of genetics and genomics data. Topics covered will include genetic linkage and association analysis, analysis of next-generation sequencing data, including those generated from DNA sequencing and RNA sequencing experiments. Students will be exposed to the latest statistical methodology and computer tools on genetic and genomic data analysis. They will also read and evaluate current statistical genetics and genomics literature. Prerequisites: Introductory graduate-level courses in statistics (such as BSTA 630-632 or EPID 520-521) are required; or permission of the instructor.

BSTA 771 Applied Bayesian Analysis (Spring) *– Highly Advanced*
This course compares and contrasts Bayesian, empirical Bayes, and frequentist approaches to statistical inference. Core topics include Bayes's theorem, the likelihood principle, selection of prior distributions

(both informative and non-informative), and simulation techniques for obtaining estimates of posterior distributions. Key statistical techniques including linear models, generalized linear models, and survival models are presented from a Bayesian perspective, along with methods for model checking and model choice such as posterior predictive distributions and Bayes factors. The course emphasizes the development and estimation of hierarchical models as a means of modeling complicated real-world problems. Bayesian methods in the design and analysis of clinical trials are also considered, with emphasis on better incorporating uncertainty and the effects of missing data and non-compliance into inference. (1.0 course unit/spring.) Prerequisites: permission of instructor.

STAT 541 Statistical Methodology (Fall) *– Highly Advanced*
This is a course that prepares 1st year PhD students in statistics for a research career. This is not an applied statistics course. Topics covered include: linear models and their high-dimensional geometry, statistical inference illustrated with linear models, diagnostics for linear models, bootstrap and permutation inference, principal component analysis, smoothing and cross-validation. Prerequisites: STAT 431 or 520 or equivalent; a solid course in linear algebra and a programming language.

STAT 542 Bayesian Methods and Computation (Spring) *– Highly Advanced*
Sophisticated tools for probability modeling and data analysis from the Bayesian perspective. Hierarchical models, mixture models and Monte Carlo simulation techniques. Prerequisites: STAT 430 or 510 or equivalent or permission of instructor.

Updated 11/30/17