BGS Training Requirement in Statistics

All BGS students are required to have an understanding of statistical methods and their application to biomedical research. Most students take BIOM611, Statistical Methods in Experimental Design and Analysis, in the spring semester of their first year. BIOM611 is a key component of BGS’ Required Training in Rigor and Reproducibility. Due to this requirement, students cannot “place out” of a statistics course while in the graduate program.

BIOM611 is not required for GCB or GGEB students, who take more advanced courses in statistics as part of their graduate group requirements. Combined degree students may elect to postpone taking BIOM611 (or a substitute) until the spring following the candidacy exam.

Students who have taken the equivalent of BIOM611, perhaps through their undergraduate work, are required to take a more advanced statistics course, such as those courses listed below. Students with strong interests in mathematics or computing may also choose to take a more advanced statistics course. In November of each year, first year students are given access in Canvas to a pre-test. The pre-test will often be based on an examination or quiz given in BIOM611 in the previous year. The pre-test is self-scoring and intended to help students understand whether they are ready for advanced work beyond BIOM 611. A set of guidelines will be supplied to the students.

Options for students seeking alternatives to BIOM 611, presented from intermediate to highly advanced level, are listed below. Students should consult with their graduate group course advisor to select an appropriate option.

**STAT 431 Statistical Inference (Spring) – Intermediate**
This course offers an advanced undergraduate level exploration of statistical techniques for data analysis, with an emphasis on developing computational tools and an understanding of when and how to use them. The latter will require a level of mathematical maturity as we examine the theoretical underpinnings of the explored methods. Interpretation of the results and analysis of assumptions is a key part of the course. As such, the course is appropriate for mathematically inclined students who wish to learn hands-on computational techniques for data analysis. Topics include (1) collection, summary, and display of data, (2) estimation, hypothesis testing, and confidence intervals, and (3) simple and multiple linear regression. Students will experiment with these ideas on data examples using statistical software. Prerequisites: STAT 430

**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.

**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.

**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.

**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).
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.

EPIP 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 EPIP 527 or equivalent preparation in biostatistics, including generalized linear models, principles of first-year calculus and matrix algebra.

EPIP 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: EPIP 510 or equivalent and EPIP 526 or equivalent.

EPIP 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: EPIP 510 or equivalent and EPIP 526 or equivalent, and permission of instructor.

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).

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 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.

**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.

**Resources:**

Penn Libraries has a new statistical software consultant, Shruthi Arvind. She will be offering appointments to provide assistance with statistical software, including Stata, R, and SPSS. For more information about the scope of her consults and to schedule an appointment, please follow this link: [https://penntrl.wordpress.com/2016/09/27/statistical-software-consultations-now-available-at-weigle-information-commons/](https://penntrl.wordpress.com/2016/09/27/statistical-software-consultations-now-available-at-weigle-information-commons/)
A variety of online statistics training videos (along with training videos on other subjects) are available to Penn students, faculty and staff via Lynda - http://www.upenn.edu/computing/lynda