BMB 510. Data Analysis and Scientific Inference

Description

An introductory course in the analysis of data and scientific inference for graduate students in Biochemistry, Molecular Biophysics, and related quantitative biomedical research areas. The course will stress fundamental principles of data analysis, best practice in presenting data, and how to draw sound scientific inferences from the data. The overall goal is to provide students the tools to carry out rigorous and reproducible scientific research.

Section I
1) Review of probability theory and the tools used for manipulating probabilities, concepts of randomness, common pitfalls and errors in data analysis.
2) Introduction to key concepts used throughout the course: probability density functions, cumulative probability distributions, and the importance of obtaining credible intervals.
3) Core elements of the Python programming language, sufficient to understand and run the software used in the later parts of the course

Section II
1) Principles of parameter estimation. Emphasis will be on robust approaches to obtaining credible intervals for parameter estimates that are valid even with small amounts of data and/or non-normal distributions, and ways to correctly incorporate results from previous experiments and other prior information.
2) Estimation of fractional and proportional parameters, population sizes, rate/time constant/decay length parameters, counting data with and without background
3) Differences between two sets of measurements, linear regression.

Section III
Higher-level aspects of analysis of data and experiments, including experimental design, quantitative comparison of models, mixture models and clustering.

After the introductory section of the course, once a week there will be a student led presentation and discussion of a paper, either a 'classic' example of data analysis (good or bad!), or a methods paper.

Intended Students

First year graduate students in BMB and other BGS graduate groups with suitable background in the mathematical and physical sciences. The course is intended as an alternative to BIOMED 611 for these more quantitative students in order to fulfill the BGS biostatistics requirement. The course will be offered spring semester, yearly.

Requirements

By permission of the instructor. Students will be required to bring a laptop to each class. The Anaconda programming and data analysis environment (https://www.anaconda.com/) will be used throughout the class. The student should install this prior to the first class. Help will provided if the student has problems. Experience with the Python programming language is not required, but as

students acquire it, it will help the students connect the lecture material to the programs they will run. Methods will be taught by example, and students are expected to run the examples themselves either on data provided by the instructor, or on suitable data from their own work.

Textbooks

Iversen, Gudmund R. 1984. *Bayesian Statistical Inference*. Sage Publications.
Sivia D, Skilling J (2006) *Data Analysis* (Oxford University Press, Oxford).
Mendenhall W, Scheaffer RL (1973) *Mathematical Statistics with Applications* (Duxbury Press,, Ma).

Evaluation

Grades will be based on homework/in class work (40%), final exam (40%), and participation in paper presentation/discussion (20%)

Homework: Students will be required to run data analysis examples either in class or as homework, and email their results to the TA to be graded. Final exam format: Each Q will involve analysis of data followed by a short text answer with interpretation/discussion of the results.

Expectations upon successful completion of the course

The students will understand the different kinds of probability: joint, conditional, marginal, and how they are used to analyze data. They will know which kind of analysis to apply depending on the type of data and what question is being asked. They will know how to obtain the usual statistical quantities – mean, variance, differences in mean and variance, etc., especially the importance of having credible intervals on every quantity they obtain and how to interpret the results of their analysis. They will recognize the confounding effects of random variation, noise, small sample size, and non-normal distributions. They will understand the principles of i) experimental design in the context of structural, physical, mechanistic experiments that form the core of modern biophysics and biochemistry research. ii) the quantitative comparison of models or hypotheses.

Instructors

Director: Kim Sharp, Kim.Sharp@pennmedicine.upenn.edu
Lecturers: BMB Faculty
TA: Martin Iwanicki,

11.00-12.00 Tues, Friday. A/C 255