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 approach throughout is Bayesian. The overall goal is to provide students the tools to carry out rigorous and reproducible scientific research.

Section I
Presenting and summarizing data: Using Python/Jupyter Notebooks to do mathematical manipulations of data, calculate averages, standard deviations, etc, make histograms, box plots, scatter plots and line plots, perform numerical computations on data arrays.

Section II
1) Role of randomness, common pitfalls and errors in data analysis.
2) Review of probability theory and the tools used for manipulating probabilities. Introduction to key concepts of probability density functions, cumulative probability distributions.
3) 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.
3) Examples of Parameter Estimation will include: fraction/proportion parameters, population sizes, rate/time constant/decay length parameters, counting data with and without background, differences in parameters 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.

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

Required for BMB students. All others by permission of the instructor. Students are required to bring a laptop to each class. The miniconda programming environment https://conda.io/miniconda.html equipped with the matplotlib, numpy and jupyter notebook modules will be used throughout the class. Students should install this prior to the first class. Help with software installation will be provided if necessary. 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.

Textbook

Sharp, Kim A. (2018) *Being Less Wrong: A Bayesian approach to data analysis and Scientific Inference* (hard copies will be provided first day of class)

Additional Texts

Iversen, Gudmund R. 1984. *Bayesian Statistical Inference*. Sage Publications.
Sivia D, Skilling J (2006) *Data Analysis, a Bayesian Tutorial* (Oxford University Press, Oxford).

Evaluation

Grades will be based on homework (40%), final exam (40%), participation in class discussions (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, Ph.D, sharpk@pennmedicine.upenn.edu
TA: Michael Cory, mcory@pennmedicine.upenn.edu

Schedule

11.00-12.30 Tues, Friday. rm 255 Anat/Chem Bldg.

Date	Topic	pages in BLW text[1]
F 17 Jan	Python Notebook I: Doing math with Python	
T 21	Python Notebook II: Programming up statistical equations	
F 24	Python Notebook III: Data Structures 1.	
T 28	Python Notebook IV: Data Structures 2	
F 31	BMB Interview Day 1.	
T 4 Feb	Python Notebook V: Data Plotting with MatPlotLib	
F 7	Python Notebook VI: Tables and Data Manipulation using Pandas	
T 11	Data Presentation: central tendency, spread, Averaging, Simpson Pdox	
F 14	BMB Interview Day 2.	pp29-34
T 18	Data Presentation: contingency tables, Monty Hall, Ioannides paper	pp29-34
F 21	Data Presentation: linear regression, correlation coefficients	pp58-61
T 25	Probability Basics	pp15-21
F 28	Probability Basics, Bayes Rule	pp15-21
T 3 Mar	Parameter Estimation: Population ID	
F 6	Parameter Estimation: Proportion/Fraction	
T 10	Spring Break	pp36-37
F 13	Spring Break	pp38-39
T 17	Parameter Estimation: Rates	pp42-46
F 20	Parameter Estimation: Rate+Bkgnd, Decay length, windowing	pp42-46
T 24	Multi-Parameter Estimation: Difference in Proportion/fraction parameter	pp51-52
F 27	Multi-Parameter Estimation: Mean and Variance	pp53-55
T 31	Multi-Parameter Estimation: Difference in means, variance multi- comparisons	pp55-57
F 3 Apr	Prior information	pp47-50
T 7	Estimating Population Size, Rank Tests	pp40-41
F 10	Clustering	pp63-67
T 14	Mixture models, hierarchical models.	pp68-72
F 17	Review	
T 21	Final Exam	

[1]"Being Less Wrong", Sharp, 2018