BBC 5100. Data Analysis and Scientific Inference

An introductory course in the analysis of data and scientific inference. Intended for graduate students in biochemistry, biophysics and related quantitative biomedical research areas. The course will stress fundamental principles of data analysis: best practices in presenting data and how to draw sound scientific inferences from the data. The goal is to give students the tools to carry out rigorous and reproducible scientific research.

Section I

Introduction to Python programming, with application to data analysis: calculating basic statistical quantities, graphical presentation, numerical computation on data arrays and manipulating spreadsheets. Experienced Python programmers can test out of this section if they wish.

Section II

1) Best practices for presenting data, summary statistics, data plotting
2) Review of probability theory and the tools used for manipulating probabilities. Introduction to the key concepts of probability density distributions and cumulative probability distributions.
3) Principles of parameter estimation using Bayesian methods. 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. Examples of parameter estimation include: fraction/proportion parameters, rate/time constant parameters, lifetime/survival times, sparse data with and without background, curve fitting, and estimating differences in these parameters between two sets of measurements. Role of randomness, common pitfalls and errors in data analysis

Section III

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

Intended Students

First year graduate students in BBCB and other BGS graduate groups with suitable background in the mathematical and physical sciences. The course is intended as an alternative to BIOMED 6110 for these more quantitative students in order to fulfill the BGS biostatistics requirement. The course will be offered spring semester, yearly.

Requirements

Required for BBCB students. All others by permission of the instructor. Students are required to bring a laptop with the anaconda programming environment installed to each class. Help with software installation will be provided. No previous experience with computer programming is required. As students acquire programming skills this will help the students connect the lecture material to the programs they will run. Methods will be taught by example, and students will run the examples themselves either on data provided by the instructor, or on suitable data from their own work.

Textbook

Sharp, Kim A. (2025 v5) *Being Less Wrong: A Bayesian approach to Data Analysis and Scientific Inference* (hard copies will be provided)

Additional Texts

Introductory
Iversen, Gudmund R. 1984. *Bayesian Statistical Inference*. Sage Publications.
Sivia D, Skilling J (2006) *Data Analysis, a Bayesian Tutorial* (Oxford University Press, Oxford).
Advanced
Bayesian Data Analysis, 3rd Edition. Gelman et al.

Python source code, Python notebooks, test cases

github:kimandsharp/bbcb5100

Evaluation

Grades will be based on homework (40%), final exam (40%), participation in class discussions (20%)
Homework: Students will run data analysis examples either in class or as homework and upload results to the TA to be graded. Final exam format: Part 1. Review of basic probability theory. Part 2: Each question will involve analysis of data followed by a short text answer with interpretation/discussion of the results.

Expectations upon successful completion of the course

Understand the standard summary statistics such as mean, variance, median and quantiles, including their strengths and weaknesses.
Understand the three fundamental types of probability: joint, conditional, marginal.
Understand probability density and cumulative probability density distributions and how they are used to analyze data;
Know which kind of analysis to apply depending on the type of data and what question is being asked and how to interpret the results
Know the importance of estimating credible intervals for *every* quantity they obtain;
Recognize the confounding effects of random variation, noise, small sample size and non-normal distributions;
Understand the statistical principles of experimental design as applied to quantitative biophysics and biochemistry research.
Understand how to make quantitative comparisons of models or hypotheses.

Instructors

Director: Kim Sharp, Ph.D, sharpk@pennmedicine.upenn.edu TA: Hannah Kim hannah.kim3@pennmedicine.upenn.edu

Schedule.

Lectures: Tues, Thurs, 10.15am-11.45am, room 255 Anat/Chem. TA office hours will arranged at the first class.

Date	Topic	Reading Material
T 20 Jan	Python Notebook I: Doing math with Python, coding up statistical equations	
R 22 Jan	Python Notebook II: Data Structures: strings and lists	
T 27 Jan	Python Coding Session	
R 29 Jan	Python Notebook III: Data Structures/Control Flow: if, while, for	
T 3 Feb	Python Notebook IV: Data Input and Plotting with MatPlotLib	
R 5 Feb	Python Notebook V: Building more complex programs: Defs and Modules	
T 10 Feb	Python Notebook VI, VII: Data Manipulation: Pandas, Numerical Computing: Numpy	
R 12 Feb	Data Presentation: central tendency, spread, averaging, Simpson P'dox, plotting	p17-21 BLW text*
T 17 Feb	Data Presentation: scatter plots, linear regression. Understanding random effects Regression to the mean, small sample effects.	p17-21 p24-26 Smith, Wainer
R 19 Feb	Probability Basics.	p27-34
T 24 Feb	Probability Basics, Bayes Rule	p27-34
R 26 Feb	Parameter Estimation: Population ID	p35-39
T 3 Mar	Parameter Estimation: Proportion/Fraction	p40-44
R 5 Mar	Multi-Parameter Estimation: Difference in Proportion/fraction parameter. Thompson sampling. 2X2 Contingency tables	p40-44
T 10 Mar	No Class – Spring Break	
R 12 Mar	No Class – Spring Break	
T 17 Mar	Multiple Proportion/fraction parameters: rat tumor example	p44
R 19 Mar	Multi-Parameter Estimation: Mean and Variance	p45-51
T 24 Mar	Multi-Parameter Estimation: Difference in means, variance, multi- comparisons Hierarchical Models/Hyper-parameters: 8 schools example	p50-51
R 26 Mar	Comparison of Bayesian and Orthodox methods: The difference in two means	p52-55
T 31 Mar	Parameter Estimation: Rates of rare events. Including background counts	p56-58

R 2 Apr	Exponential Decay in time or space. Effect of windowing/censoring	p58-60
T 7 Apr	Non-exponential decay in time or space/Survival analysis	p60-64
R 9 Apr	Curve Fitting: Linear, Weibull, Polynomial, Sinusoidal	p65-70
T14 Apr	Discrete/non-parametric data. Comparing Ranks, Estimating Population Size,	p75-78
R 16 Apr	Clustering, Mixture models. Relationship to machine learning: feature identification and classification (Population ID example again)	p79-88
T 21 Apr	Review	
R 23 Apr	Final Exam	

*Textbook: Being Less Wrong, Sharp, 2025v5