# BMB 510. Data Analysis and Scientific Inference

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

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

## 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. 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 , 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://docs.conda.io/en/latest/miniconda.html equipped with the matplotlib, numpy, jupyter notebook, pandas and openpyxl modules will be needed for the class. Help with software installation will be provided. 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 will run the examples themselves either on data provided by the instructor, or on suitable data from their own work.

# Textbook

Sharp, Kim A. (2022) *Being Less Wrong: A Bayesian approach to Data Analysis and Scientific Inference* (hard copies will be provided)

# Python source Code

github:kimandsharp/bmb510

# 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.

# 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; know which kind of analysis to apply depending on the type of data and what question is being asked; 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; recognize the confounding effects of random variation, noise, small sample size, and non-normal distributions; 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: Saira Montermoso sairamon@pennmedicine.upenn.edu

# Time, Place

Tues, Friday, 12.00pm-1.15pm. In person, room 255 Anat/Chem.

# Schedule

| Date | Topic | pages in BLW text[1] | |
|---|---|---|---|
| **F 13 Jan** | Python Notebook I: Doing math with Python, coding up statistical equations | | |
| T 17 | Python Notebook II: Data Structures: strings and lists | | |
| F 20 | Python Coding Session | | |
| T 24 | Python Notebook III: Data Structures/Control Flow: if, while, for | | |
| F 27 | No Class- BMB recruitment | | |
| T 31 | Python Notebook IV: Data Input and Plotting with MatPlotLib | | |
| **F 3 Feb** | Python Notebook V: Building more complex programs: Defs and Modules | | |
| T 7 | Python Notebook VI: Tables and Data Manipulation using Pandas | | |
| F 10 | No Class- BMB recruitment | | |
| T 14 | Python Notebook VII: Numerical Computing using Numpy | | |
| F 17 | Data Presentation: central tendency, spread, averaging, Simpson P'dox, plotting | p29 | |
| T 21 | Data Presentation: linear regression, correlation coefficients. Gary Smith, what the luck.<br>Hotelling Review. Regression to the mean<br>Contingency tables, Ioannides paper | p29<br>p69<br>p25 | |
| F 24 | Probability Basics. Read Howard Wainer | p15 | |
| T 28 | Probability Basics, Bayes Rule | p15 | |
| **F 3 Mar** | Parameter Estimation: Population ID | p36 | |
| T 7 | No Class – Spring Break | | |
| F 10 | No Class – Spring Break | | |
| T 14 | Parameter Estimation: Proportion/Fraction | p41 | |
| F 17 | Multi-Parameter Estimation: Difference in Proportion/fraction parameter. Thompson sampling. 2X2 Contingency tables | p43 | |
| T 21 | Multiple Proportion/fraction parameters: rat tumor example | | |
| F 24 | Multi-Parameter Estimation: Mean and Variance | p53 | |
| T 28 | Multi-Parameter Estimation: Difference in means, variance, multi- comparisons | p56 | |

| | | | |
|---|---|---|---|
| | Hierarchical Models/Hyper-parameters: 8 schools example | | |
| F 31 | Using Prior information, Estimating differences in mean, proportion using statistical summaries | p47 | |
| **T 4 Apr** | Parameter Estimation: Rates of rare events. Including background counts | p47 | |
| F 7 | Exponential Decay in time or space. Effect of windowing/censoring | p63 | |
| T 11 | Non-exponential decay in time or space/Survival analysis | p64 | |
| F 14 | Discrete/non-parametric data. Estimating Population Size, Rank Tests | p35 | |
| T 18 | Curve Fitting: Linear, Weibull, Polynomial, Sinusoidal | p69 | |
| F 21 | Clustering, Mixture models. Relationship to machine learning: feature identification and classification (cf. Population ID example) | p75 | |
| T 25 | Review | | |
| F 28 | Final Exam | | |
| | | | |

[1]"Being Less Wrong", Sharp, 2022