**Title**: Thinking about data: statistics for the life sciences

**Format**: quarter course, Eight 2-hour meetings

**Description**: Probability & statistics taught with an emphasis on using simulations and re-sampling methods to both analyze data and understand core statistical concepts.

**Instructors**: Richard Born, MD (in class); Brian Healy, PhD (video lectures)

Motivations:

1. Most existing statistics courses are simultaneously too superficial (e.g. flow-chart approach to test selection) and too detailed (e.g. calculations of covariance), and fail to develop intuitions that are central to statistical thinking (e.g. What is a p-value? a Bayes Factor?).
2. There is an abundance of good web-based material that can be taken advantage of if the students understand basic concepts and have basic programming skills.
3. We are not teaching our students the statistical approaches they will actually need in the world of big data.

Meetings: Eight 2-hour class sessions (quarter course)

Approach: Prior to class, students will view online lectures from Dr. Brian Healy’s biostatistics course. In class, we will focus on coding exercises to build intuition through simulations and to practice different approaches to analyzing real data sets, with an emphasis on resampling methods (i.e. bootstrap and permutation tests). All in-class programming will be done in MATLAB. Students must have either taken Neurobiology 306qc or demonstrate proficiency to the instructor.

Week #1: Data and Probability

**Video**: 1: Introduction to Biostatistics, 2: Basics of Probability

**Class**: distribution plots vs. box plots; PIN data visualization; dual-code for image analysis

**Exercises**: Bayes vs. Frequentist; Bayes dice demo; central limit theorem demo;

Week #2: Bootstrap I: Standard error and confidence intervals

**Video**: 3. Hypothesis testing/t-test; Video: 5. Nonparametric approaches Video:

**Class**: introduction to bootstrap; bootstrapping standard errors and confidence intervals; failure mode of the bootstrap

**Exercises**: ‘etMouseCI2.m’, ‘etMouse2SampleStats.m’, ‘etPCAdemo.m’; ‘etCorrCI.m’; ‘etBootstrapFailure.m’

Week #3: Bootstrap II: Hypothesis testing

**Video**: 6. Analysis of proportions

**Class**: bootstrap for proportion data; CI and hypothesis testing; permutation test

**Exercises**: ‘etASAdemo.m’, ‘etMouse2SamplePermutationTest.m’, ‘etMouse2SampleBootstrapTest.m’

Week #4: Linear Regression I

**Video**: 7. Linear regression and correlation; 12. Regression diagnostics

**Class**: basics of MATLAB regression tools; bootstrapping SEs when using other than LSE to do regression

**Exercises**: ‘etHormoneRegression.m’, ‘etCellSurvivalReg.m’

Week #5: Linear Regression II

**Video**: 11. Multiple Linear Regression I; 16/17. Logistic Regression

**Class**: GLM logistic regression on microstim experiment data

**Exercises**: ‘mStimLogisticRegressionDemo.m’

Week #6: Linear Regression III

**Video**: None

**Class**: GLM to fit Poisson point process model; over-fitting; cross-validation; regularization

**Exercises**: ‘placeCellFitEx.m’, ‘OverFit.m’

Week #7: Power, P-hacking and Reproducibility (“torturing data ethically” – XLM)

**Video**: 33. Multiple comparisons

**Class**: multiple comparisons; “researcher degrees of freedom”; power calculations via simulation

**Exercises**: ‘dfSim2.m’ , ‘powSimTtest2.m’

Week #8: Intro to Empirical Bayes: False Discovery Rates

**Video**: None

**Class**: distribution of p-values under H0; P-curve; Stein’s paradox; false discovery rates

**Exercises**: ‘MultipleComparisonsExercise.m’, ‘JSdemo.m’, ‘FDRdemo.m’

General Resources:

Datacolada: <http://datacolada.org/>

[MIT 18.05](https://ocw.mit.edu/courses/mathematics/18-05-introduction-to-probability-and-statistics-spring-2014/index.htm)

Nature’s statistics for biologists: <http://www.nature.com/collections/qghhqm/content/practical-guides>