Title: Thinking about data: 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 Healey, 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 A**: distribution plots vs. box plots; PIN data visualization; dual-code for image analysis

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

**Resources**:

Leek JT “[How to share data with a statistician](https://github.com/jtleek/datasharing)”

Wickham H “[Tidy Data](http://vita.had.co.nz/papers/tidy-data.pdf)” (pdf)

Allen EA, Erhardt EB, Calhoun VD (2012) “Data Visualization in the Neurosciences:

Overcoming the Curse of Dimensionality” *Neuron* **74**:603-608. (pdf)

Intro (15):

* Why this course?
* Importance of becoming a life-long statistician: e.g. subscribe to datacolada.org
* Course mechanics and expectations

Data (45)

* How to share data with a statistician. tidy data / clean data
* View your data in the rawest form that is visually comprehensible: PIN data (mapPinData.m)
* Think hard about effective ways to visualize your data: PIN data cont’d
* In publications: Hide less; show more (Allen et al. 2012):
  + distribution plots (a.k.a. ‘violin’ plots) vs. whisker plots vs. bar plots (distributionPlotDemo.m)
  + matrix of covariate plots
  + dual-code colorbars for images: dualcodeImage.m
* Sampling error vs. bias
  + class exercise: distorted ruler
  + XLM example: meng-statistics\_perils\_of\_self-selection.pdf (1 page)

Probability (60)

* using simulations to solve (difficult!) probability problems
  + class picks number ‘randomly’ [1,4]: FourChoiceProbabilityFunction.m
  + extension to birthday problem: BirthdayProblem.m
* using simulations to gain intuition
  + Central Limit Theorem demo: CLTdemo.m
* Frequentist vs. Bayes from 40k feet
  + Bayes theorem: 4 views
    - rule for inverting conditional probabilities
    - method for combining new data with existing knowledge
      * twin problem: U/S shows 2 boys; 1/3 twins identical; P(twins/US)?
    - method for updating beliefs as evidence accumulates over time
      * Bayes dice demo (Bayes factors, etc.)
    - a religion (compare and contrast w/ frequentist)

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

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

In class: introduction to bootstrap; bootstrapping standard errors and confidence intervals; failure mode of the bootstrap;

**ideas**: point estimate vs. interval estimate; std. error as most common and easily calculated confidence interval (68%); simple analytical formula; bootstrap formula gives same answer but can be applied to any point estimate we can calculate from our data

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

MATLAB built-in functions: ‘bootstrp’, ‘bootci’

**Resources**: Efron & Tibshirani, “An Introduction to the Bootstrap”

Week #3: Bootstrap II: Hypothesis testing

6. Analysis of proportions

In class: bootstrap example from E&T for proportion data (‘etASAdemo.m’); compare w/ built-in Fisher’s Exact Test (‘fishertest’)

In class: Resampling for hypothesis tests and CI hypothesis testing: permutation test (‘etMouse2SamplePermutationTest.m’) vs. bootstrap test (‘etMouse2SampleBootstrapTest.m’)

Week #4: Linear Regression I / Bootstrap III

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

In class: basics of MATLAB regression tools: ‘etHormoneRegression.m’, ‘etCellSurvivalReg.m’ (bootstrapping SEs when using other than LSE to do regression)

MATLAB built-in functions: ‘regress’

Week #5: Linear Regression II

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

In class: GLM logistic regression on microstim experiment data (‘mStimLogisticRegressionDemo.m’)

MATLAB built-in functions: ‘glmfit’,’ fitglm’

Week #6: Linear Regression III

In class: GLM to fit Poisson point process model to place cell data (‘placeCellFitEx.m’)

In class: over-fitting (‘OverFit.m’); cross-validation ; regularization

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

Video: 33. Multiple comparisons

In class: multiple comparisons; “researcher degrees of freedom”; simulation of stopping rule for data collection (‘dfSim2.m’); power calculations via simulation (‘powSimTtest2.m’)

p-hacking demo: <https://fivethirtyeight.com/features/science-isnt-broken/#part1>

(Start class with this demo!)

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

In class: distribution of p-values under H0; P-curve; ‘MultipleComparisonsExercise.m’; Stein’s paradox: ‘JSdemo.m’; false discovery rates: ‘FDRdemo.m’

General Resources:

“[Seeing Theory](http://students.brown.edu/seeing-theory/)”: a visual introduction to probability and statistics

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>

<http://jmxpearson.com/matlab-neurobio/exercises/week-1-1-exploring-tables.html>

Words of wisdom cum mantras:

“After 20 years of teaching, I’ve come to the conclusion that teaching skills works

better than teaching concepts (or, should I say, trying to teach concepts).”

--Andrew Gelman (2008)

“"When there is something to permute, it is a good idea to do so .”

--Efron & Tibshirani, “An Introduction to the Bootstrap” (1998), p. 218

Advert for Teaching Fellows:

I am looking to hire two Teaching Fellows for a statistics *quarter* course that I am teaching during the first half of the fall semester of 2017. Each TF position should be thought of as two consecutive jobs: I will require considerable help refining course materials over the summer and then class-time assistance once the course starts. Each TF will be paid $7,500.

A bit about the course: If you are familiar with the Boot Camp in Quantitative Methods (Neurobiology 306qc) that I teach with Mike Springer, you have a good idea of the pedagogical approach for the course meetings. The major addition is that, prior to classes, the 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.

Applicants for the TF position should have a basic familiarity with probability and statistics and be experienced MATLAB programmers. Please express your interest directly to Rick ([rborn@hms.harvard.edu](mailto:rborn@hms.harvard.edu)), briefly summarizing your qualifications. Further questions about course details should be directed to the same party.