Course syllabus for 140.754
Ph.D. Level Applied Statistics and Methods

Instructor: Jeff Leek (jtleek@gmail.com)
TA: Parichoy Pal choudhury (ppalchou@jhsph.edu, office hour TBA)
Webpage: http://jtleek.github.com/jhsph753/
Old 754 Webpages: http://biostat.jhsph.edu/~jleek/teaching/2011/754/ (2011)
http://biostat.jhsph.edu/~jleek/teaching/appliedstat/ (2012)
Room: Genome Cafe (may change)
Time: TuTh 9:00AM-10:20AM

This course is the third term of the intensive introduction to methods for applied statistics. The goal of this sequence is to develop Ph.D. level biostatisticians who are capable of both applied data analysis and developing the next generation of statistical methods. Both data analysis and methods development require substantial hands on experience, so the focus of this class will be on hands on data analysis.

Learning Objectives:

Upon completion of this course students will be able to:

1. Obtain, clean, transform and process raw data into usable formats
2. Formulate quantitative models to address scientific questions
3. Organize and perform a complete data analysis, from exploration, to analysis, to synthesis, to communication.
4. Understand and apply a range of statistical methods for inference and prediction.
5. Develop ideas for new statistical methods, tools, and analyses

Students will also be encouraged to independently read and apply statistical methods from texts and the scientific literature that are not covered in the course. They will also be encouraged to think of improvements or variations on existing methods to address specific scientific questions.

Evaluation

1. 35% = Bi-weekly data analysis
2. 20% = Bi-weekly problems
3. 10% = Bi-weekly data analysis review
4. 10% = In class participation
5. 25% = Final Project

Data Analyses
(For more on my project philosophy see: http://bit.ly/wQT5uI)

Each student will be required to perform two data analysis projects during the course of the

class. Students will be given 2 weeks to perform each analysis. The project assignments will consist of a scientific description of the problem. Students are responsible for all stages of each data analysis from obtaining the data to the final report. At the conclusion of each analysis each student must turn in: (1) A write-up of their data analysis in a synthesized format, with numbered figures and references. (You may also include supplementary material for detailed additional calculations/analyses) and (2) a reproducible Rmd file that produces all of the numbers, figures and results in your write-up. All documents should be submitted electronically.

Grading for the data analyses will be on the following criteria:

1. Did you answer the scientific question? (30%)
2. Did you use appropriate statistical methods? (40%)
3. Was your write-up simple, clear, and precise? (20%)
4. Was your code reproducible? (10%)
5. *Did you do something incredible?* (possibility for extra points)

You may speak to your fellow students about specific statistical questions related to the projects, but the overall idea, analysis, and write-up should be your own individual work. You should cite any help you get from fellow students/TAs in your report in standard citation format.

Data Analysis Reviews

After each data analysis is turned in, they will be randomly assigned to another student for review. Your review will be due one week after it is assigned. Your comments should have the format of a typical peer review. You should include a summary of the analyses and conclusions in the project you are reviewing, any major revisions, and any minor revisions. I will also evaluate each data analysis independently to assign a grade. Synthesized comments will be made available for each project.

Homework

Every two weeks you will be assigned one more mathematical question focused on a statistical method we have covered. These problems may have multiple parts. The solutions should be typed in Latex.

Final Project

The final project will have the same format as the data analyses. It will be slightly longer than the weekly projects in terms of space and more in depth in terms of analysis.

The choice of your final project is up to you. The project should involve data/code that you can obtain, process, analyze, and synthesize yourself. You may use any of the methods you learn during the course, or any other methods you know/look up etc.

After the first week of class you will submit a project proposal (one paragraph) to the instructor who will help you determine the feasibility and appropriateness of your project. Grading for the final exam will be weighted by the difficulty of the project you undertake. The more difficult the project you take on, the greater the multiplier of your final score. The maximum possible score will still be 100%.

Students will rotate through presentations of their current results over the course of the class, so you can obtain feedback and get “unstuck”.

Tentative Material to Be Covered (will bleed over into 754)

- Obtaining data and data processing
    - Data sources
    - Data formatting
    - Data evaluation
    - Regular expressions
- Exploratory data analysis
    - Data visualization and summaries
    - Clustering
    - Principal components analysis/SVD
    - Imputation
- Simulation studies
    - Generation of data
    - Structure of studies
    - Comparing methods
- Error measures
    - Inferential error measures
    - Bootstrap
    - Predictive error measures
    - Cross validation
- Regression
    - Model agnostic errors
    - Generalized linear models
    - Generalized estimating equations
    - Interpretations marginal/conditional
    - Generalized additive models
    - Smoothing (kernel, etc.)
- High dimensional data
    - Multiple testing
    - Regularization
    - Overfitting

- Prediction Techniques
    - Linear discriminant analysis
    - Decision trees
    - Boosting
    - Bagging