S681 brief syllabus (draft, subject to change)

Brad Luen

Spring 2019

1 Instructor

Dr. Brad Luen
Office: Informatics East 210
bradluen at indiana.edu (put S681 in the subject line)

2 Description

This course is a survey of statistical methods that do not rely on parametric assumptions. Knowledge of introductory statistics at the level of S320/S520 is assumed; this course is in some ways a sequel. As such, it will review the parametric techniques learned in that and similar introductory courses, and compare them to nonparametric alternatives to see when one technique outperforms another.

The course material will be organized in six modules (names and order subject to change):

1. **EDA and basic concepts:**
   - R and RStudio: the basics
   - What's EDA?
   - Using `ggplot()`
   - EDA and models
   - Review of inference
   - Review of simple linear regression
2. **Nonparametric tests:**
   - The sign test
   - Permutation tests
   - Rank tests
   - Nonparametric confidence intervals
   - k-sample tests
   - Issues with multiple testing
3. **Empirical distributions and the bootstrap:**

- Fitting parametric distributions
- Estimating the empirical distribution
- Estimating the PDF
- Censored data
- The bootstrap principle
- Bootstrap confidence intervals

4. **Multiple linear regression:**
    - Visualizing lots of variables
    - Understanding multiple regression
    - Interactions
    - Multiway data
    - Dummy variables
    - Building models
5. **Nonparametric and penalized regression:**
    - Loess
    - Cross-validating regression models
    - Splines
    - Penalized regression
    - Fitting additive models
    - Interpreting additive models
6. **GLMs and other advanced models:**
    - Visualizing discrete data
    - Logistic regression
    - Building logistic regression models
    - Models for count data
    - Models for categorical data
    - Generalized additive models

The computation in the course will be performed in R.

3 Prerequisites

The prerequisite to this course is STAT-S 520. You should already know how to calculate probabilities using software or otherwise for the fundamental probability distributions like the binomial and the normal. You should know the forms and interpretations of t-tests, confidence intervals, and the simple linear regression line.

You should have some experience with R. If not, download R for free from `cran.r-project.org`, install it, and start playing around with it. The clearest introductory guide to actually doing statistics in R that I know of is John Verzani's *simpleR* at `http://cran.r-project.org/doc/contrib/Verzani-SimpleR.pdf`. The material up to chapter 7 (Simulations) will be immediately useful.

4 Grading

Discussion and quick checks (5%?): A small proportion of the grade will be reserved for participation in discussion and for quick check questions answered during the course of working through the modules.

Problem sets (60%?): There will be one of these for each module (six in total.) A typical problem set will consist of 8–10 problems.

Projects (35%?): There will be three major projects, each of which is expected to take about a week. These may be done in pairs. Clear communication will be important for a good grade.