

## EPSY 887 Computational Statistics: Institute (Spring 2013)

**Instructors:** R. Pruzek (rpruzek@albany.edu) & J. Bryer (jason@bryer.org)

**Meeting Time:** Thursdays 4.15 - 7.05

**Grading:** 3 credits, S/U grading

**Course Website:** <https://github.com/jbryer/CompStats>

This seminar will provide an introduction to statistical programming for data analysis with an emphasis on the analysis of large datasets. With the increased availability of large national and international datasets (e.g. PISA, TIMMS, NAEP, ECLS) there is a great opportunity and potential for researchers to focus on important substantive research questions that are difficult to address by other means. However, the analysis of large datasets requires special analytical procedures not found in commercial statistics software. Utilizing the open source statistical software R, students will be introduced to tools and procedures for analyzing large datasets with an emphasis on conducting transparent and reproducible research.

### Topics:

- Introduction to R (e.g. data input, recoding, etc.)
- Reshaping data
- Data visualization vis-à-vis a grammar of graphics and related methods
- Introduction to programming for data analysis (e.g. loops, conditional statements, functions, etc.)
- Missing data issues and methods
- Analysis of complex surveys (e.g. use of replicate weights and multiple plausible values)
- Document preparation and typesetting with LaTeX and Sweave
- R package development
- Software project management principles as applied to data analysis (e.g. source control, progress tracking, versioning, Github, R-Forge, etc.).
- Other data analytic topics as identified by students and appropriate for analysis of large datasets. Topics may include propensity score analysis, multilevel modeling, IRT, random forests, regression trees, multivariate methods, etc.

Students are encouraged to bring their own data and/or research questions as this seminar will emphasize applied statistics and analysis. As often as possible, assignments will focus on student interests. Class examples however, will utilize the Programme for International Student Assessment (PISA; OECD 2009).

**Prerequisites:** Two graduate level statistics courses (e.g. EPSY 530 and EPSY 630) and as much software knowledge as you can muster.

### Textbooks

Kabacoff, R.J. (2011). *[R in Action: Data Analysis and Graphics with R](#)*. Shelter Island, NY: Manning.

Matloff, N. (2011). *[The Art of R Programming](#)*. San Francisco, CA: No Starch Press