PLSC500: Foundations of Statistical Inference (Fall 2019)

Instructor: Peter Aronow (peter.aronow@yale.edu).

TFs: Angèle Delevoye (angele.delevoye@yale.edu) and Trevor Incerti (trevor.incerti@yale.edu).

Class time/location: Monday and Wednesday, 2:30-3:45, RKZ 05. Office hours and section hours TBD.

Course Overview and Prerequisites

This course provides an intensive introduction to statistical theory for quantitative social inquiry. Topics include foundations of probability theory, statistical inference from random samples, estimation theory, linear regression, maximum likelihood estimation, and nonparametric identification.

Mathematics training at the level of the political science math camp (including calculus of a single variable) is required. PLSC 529 (Mathematics for Political Science) is a corequisite for Political Science PhD students. Students from other programs are welcome, but should contact the instructor beforehand about preparation. More generally, if you are uncertain about whether or not you are prepared for the course, please feel free to contact the instructor.

Readings

The textbook for this course is Aronow and Miller (A&M):

Aronow, P. M. and Miller, B. T. 2019. *Foundations of Agnostic Statistics*. Cambridge University Press.

It is mandatory that students (attempt to) engage with the textbook readings before class. A PDF of the introductory material and Chapter 1 will be placed on Canvas. Supplementary readings will be assigned as the semester proceeds.

Assignments, Exams and Grading

Seven problem sets will be assigned on Wednesdays (dates TBD). They will be due the following week on Wednesday at 2:30-3:45pm. Although students can work in groups, we strongly encourage students to attempt the problem sets individually first. Submitted problem sets must indicate the names of all students who collaborated. Answer keys will be posted within 48 hours of the problem set due date. Accordingly, problem sets *cannot be accepted late*, and will be awarded zero credit. The lowest score (including a missed assignment) among the problem sets will be dropped.

There will be a midterm and a final exam, with dates on the calendar below. Both exams contain in-class and take-home components. Students will have the opportunity to submit corrections to their in-class components along with their take-home examinations, and partial credit will be awarded as appropriate. Similarly, we cannot accept exams late.

The final grade will be based on problem sets (30%), the midterm (20%), the final exam (40%), and class participation (10%).

Statistical Software

We will be using the R programming language, available at `http://cran.r-project.org/`. As we will be focusing on statistical programming concepts (and *not data analysis*), we will be restricting the use of packages outside of the base R functionality. The goal of your programming work will be to understand how to translate statistical ideas into computational implementation from scratch.

Plagiarism and Academic Integrity

As the Yale Graduate School notes: "Academic integrity is a core institutional value at Yale. It means, among other things, truth in presentation, diligence and precision in citing works and ideas we have used, and acknowledging our collaborations with others. In view of our commitment to maintaining the highest standards of academic integrity, the Graduate School Code of Conduct specifically prohibits the following forms of behavior: cheating on examinations, problem sets and all other forms of assessment; falsification and/or fabrication of data; plagiarism, that is, the failure in a dissertation, essay or other written exercise to acknowledge ideas, research, or language taken from others; and multiple submission of the same work without obtaining explicit written permission from both instructors before the material is submitted. Students found guilty of violations of academic integrity are subject to one or more of the following penalties: written reprimand, probation, suspension (noted on a student's transcript) or dismissal (noted on a student's transcript)."

(Tentative) Calendar

8/28. Introduction. Reading: A&M, Introduction.

8/30. Probability. Reading: A&M, Ch. 1.1.1-1.1.2.

9/4. Joint and conditional probability. Reading: A&M, Ch. 1.1.3.

9/9. Applications of conditional probability. Independence. Reading: A&M, Ch. 1.1.4.

9/11. Discrete random variables; the Cumulative Distribution Function. Reading: A&M, Ch. 1.2.1-1.2.3.

9/16. Continuous random variables. Support. Reading: A&M, Ch. 1.2.4-1.2.5.

9/18. Random vectors. Reading: A&M, Ch. 1.3-1.4.

9/23. Summarizing random variables. Reading: A&M, Ch. 2.1.

9/25. Summarizing joint distributions; the Conditional Expectation Function. Reading: A&M, Ch. 2.2.0-2.2.3.

9/30. The Best Linear Predictor. Reading: A&M, Ch. 2.2.4-2.2.5.

10/2. Multivariate generalizations of the CEF and BLP. Reading: A&M, Ch. 2.3.

10/7. Random sampling model. Reading: A&M, Ch. 3.0-3.1.

10/9. Estimation theory. Reading: A&M, Ch. 3.2-3.3.

10/14. Statistical inference. Reading: A&M, Ch. 3.4. Take-home midterm assigned.

10/21. Clustering. Midterm review. Reading: A&M, Ch. 3.5.

10/23. In-class midterm.

10/28. Linear regression; Ordinary Least Squares. Reading: A&M, Ch. 4.1.

10/30. Inference with OLS. Reading: A&M, Ch. 4.2. Take-home midterm due.

11/4. Relaxing linearity with OLS; interactions and polynomials. Reading: A&M, Ch. 4.3.0-4.3.4.

11/6. Applications of linear regression; complexity and regularization. Reading: A&M, Ch. 4.3.5-4.4.

11/11. Parametric models. Reading: A&M, Ch. 5.0-5.1.

11/13. Maximum likelihood estimation. Ch. 5.2-5.3.

11/18. Nonparametric identification; missing data. Reading: A&M, Ch. 6.0-6.1.

11/20. Causal inference. Reading: A&M, Ch. 7.0-7.1.

12/2. Estimation with missing data. Reading: A&M, Ch. 6.2-6.3.

12/4. Estimation of causal effects. Reading: A&M, Ch. 7.2-7.5.

12/9. Final review; roadmap for further study. Take-home final exam assigned.

12/11. In-class final exam.

12/18 (No Class). Take-home final exam due.