Stat 30850/27850: Multiple Testing, Modern Inference, and Replicability

Course description This course examines the problems of multiple testing and statistical inference from a modern point of view. High-dimensional data is now common in many applications across the biological, physical, and social sciences. With this increased capacity to generate and analyze data, classical statistical methods may no longer ensure the reliability or replicability of scientific discoveries. We will examine a range of modern methods that provide statistical inference tools in the context of modern large-scale data analysis. The course will have weekly assignments as well as a final project, both of which will include both theoretical and computational components.

Prerequisites: Stat 24400 or equivalent. Undergraduates may enroll with permission of the instructor.

Course info

- Course times: MWF 1:30-2:20, Stuart 105.
- Instructor: Rina Foygel Barber, `rina@uchicago.edu`
  Office: Eckhart 113.
  Office hours: W 3-4:30.
- TA: Yuancheng Zhu, `yuancheng@galton.uchicago.edu`
  Office: Eckhart 8.
  Office hours: Th 4:30-6, Eckhart 117.

Grading

- Problem sets & other assignments: 50%
- Project: 50% ($\approx$ 15% project proposal + progress report, 35% presentation + final report)

Handing in assignments All assignments are due on Friday of that week. Handing in assignments on paper, or by email to the TA, are both acceptable. It's fine to hand it in by the end of the day on Friday. Past that, no late HWs will be accepted for any reason, however, the lowest HW score will be dropped (this does not include the project proposal / progress report / final report, which are all required).

Collaboration policy For problem sets, students are free to discuss the problems and collaborate on strategies for solving the problems, but all writing, code, etc, should be done completely on your own. (For example, working out a solution on the board in a group, then transferring it to the page, is not acceptable.)

For the final project, students who form a group are expected to be fully collaborating on all aspects of the work. The grade on the final project will be given to the entire group.

Computing The problem sets and the final project will all involve some amount of simulations or computation on real data. These may be carried out in R, Matlab, or Python. We will sometimes provide R code as part of a problem set, therefore all students should be comfortable using R if needed. However, this is not a prerequisite for the course; if any students are new to programming in R, we will provide a tutorial early in the quarter which should be sufficient for this course.

Final project For the final project, students may work on their own or form a group of size 2–4. (Larger groups will be expected to produce a more extensive project.) Your project can either:

- Read one or more recent papers in the field and explain the method and key ideas, along with implementing the method, testing it on real data and/or exploring its power and limitations on simulated data, trying out extensions of the method, etc;

- Or, develop and test new methods to address an interesting problem. This type of project would examine an existing problem (its importance, and why existing methods are not sufficient to handle the problem), develop a method for addressing the problem along with intuition for why this method would work and would perform well (this can optionally include theoretical results), and would test the method empirically on simulated and/or real data.

Concrete suggestions for possible projects will be discussed later in the quarter.

Working on the project will span the second half of the quarter. Each group will:

- Form a group and meet with the instructor to brainstorm possible ideas in weeks 5/6
- Turn in a 1-2 page project proposal outlining a tentative plan, in week 7
- Turn in a 2-4 page progress report describing what you've found so far (problems or questions addressed, early empirical results, etc), in week 9
- Turn in a full report (around 10 pages, but this will vary a lot depending on if your work is more theoretical or empirical), due in finals week
- Give a short final presentation, during finals week

Schedule (very tentative)

Week	Topics	Due (on Friday)
1 (Jan 4–8)	Replicability and reproducibility News coverage on replicability in medicine, psychology, political science, & other fields Classical statistical inference & testing Intro to multiple testing: testing the global null hypothesis	
2 (Jan 11–15)	Family-wise error rate, false discovery rate Methods for multiple testing to control FWER & FDR	Problem set 1
3 (Jan 18–22)	The empirical Bayes perspective on FDR Local FDR FWER and FDR without independence	P-hacking challenge
4 (Jan 25–29)	Structured testing: hierarchies / groups of hypotheses, ordered / sequential / online testing	Problem set 2
5 (Feb 1–5)	Sign error, magnitude error Confidence intervals, false coverage rate	Real data analysis critique Form project group & meet with Rina during week 5 or 6
6 (Feb 8–12)	Review: linear regression & the Lasso Asymptotic inference for linear models Intro to post-selection inference	Problem set 3
7 (Feb 15–19)	Post-selection inference for linear models	Project proposal
8 (Feb 22–26)	Post-selection inference for sequential selection procedures Post-selection inference for other problems: ranking, & others	Problem set 4
9 (Feb 29–Mar 4)	Other topics: adaptivity and privacy, & other topics TBA	Project progress report
10 (Mar 7–11)	Guest lectures	
Finals week (Mar 14–18)	Project presentations	Project final report