Stat 27850/30850 Autumn 2023
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 group projects, both of which will include both theoretical and computational components.

Prerequisites: Stat 24400 or equivalent, and comfortable programming in R or Python or Matlab.

Course info

- Course times: Tue/Thu 12:30–1:50pm, Harper 130
- Instructor: Rina Barber (`rina@uchicago.edu`)
  Office hours: Week 1 = Thu 11:00am -12:00pm, Week 2–9 = Fri 11:00am–12:00pm (Jones 214)
- TAs: Yu Gui (`yugui@uchicago.edu`) and Wanrong Zhu (`wanrongzhu@uchicago.edu`)
  Office hours: Tue 6:00–7:00pm (Zoom) and Wed 4:30–5:30pm (Jones 226)
- All course materials & assignments will be on Canvas. Announcements and Q&A will be on Ed.
- The final course grade will be calculated as: Problem sets: 50%, group assignments: 50%

Contacting us

- For any questions about the material or for general questions about an assignment, please post a public question on Ed (you can choose to post anonymously).
- For specific questions about your work on an assignment (i.e., questions that cannot be posted publicly because it would reveal too much of the solution), please ask us via a private post on Ed.
- For any questions about a graded assignment, please use the regrade request feature on Gradescope.
- For other questions such as enrollment, prerequisites, grades, etc, please contact the instructor by email.

Handing in assignments

- Assignments are due at the start of class on Thursdays (12:30pm).
- Unless otherwise noted, for all assignments, show all your work / code / plots / etc.
- To give additional flexibility, late assignments will be accepted with a penalty of 2% per hour (late time is rounded up, i.e., one minute late counts as one hour late). There will be an optional problem set (Problem set 5) at the end of the quarter that can replace a missed problem set or a low score on a problem set. Group assignments cannot be replaced. **We cannot give extensions or exceptions to these policies.**
- Assignments are submitted and graded via Gradescope (which can be accessed from the Canvas course page).
  For each problem, Gradescope will prompt you to tag the pages containing your answer to that problem. Please be sure to do this to help the TAs grade efficiently. The timestamp on your assignment is the time it was uploaded (i.e., time spent tagging will not make your submission late).
- If you are having trouble uploading to the website and run out of time, please email your work to the instructor or TA before the time the assigment is due as proof of completion. The time of your email will count as the time of your submission. We do not accept the time stamp of the file on your computer as proof of completion.

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 group assignments (the real data analysis critique, and the two projects), students work in groups of size 2 or 3 or 4, and are expected to be fully collaborating on all aspects of the work. The grade on each project will be given to the entire group. Students must work in a group and may not hand in individual projects. Feel free to post on Ed if you are looking for team members. Please contact the instructor if any issues arise.

Computing The problem sets and the data analysis projects 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. The TAs can provide support as needed for students who are new to programming in R.

Projects Additional information for the projects:

- Project 1 will be based on a concrete data set, and you will design questions to explore and test, and compare existing methods or create a new method to analyze your questions.
- Project 2 will have two options: a data analysis option and a theory option.
- The instructor and/or TAs will offer appointments for groups to come in for feedback on their project ideas, during weeks 5 & 6 (for Project 1) and during week 9 & finals week (for Project 2)

Schedule (tentative—topics may change as needed)

Week	Topics	Due (12:30pm on Thursdays)
1 (T/Th Sep 26&28)	Intro to problems in modern inference Hypothesis testing Testing the global null Family-wise error rate & false discovery rate	
2 (T/Th Oct 3&5)	Multiple testing methods: Benjamini–Hochberg & related methods	Problem set 1
3 (T/Th Oct 10&12)	Confidence intervals & false coverage rate Permutation tests Intro to high-dimensional linear regression	Real data analysis critique (group assignment)
4 (T/Th Oct 17&19)	High-dimensional regression & variable selection: Lasso & related methods Selective inference in linear models	Problem set 2
5 (T/Th Oct 24&26)	High-dimensional regression continued: Selective inference in linear models continued Asymptotic inference via debiasing	
6 (T/Th Oct 31&Nov 2)	Selective inference for other problems: Ranking, clustering, etc	Project 1 (group assignment)
7 (T/Th Nov 7&9)	High-dimensional regression continued: Knockoff filter for linear regression Model-X knockoffs	
8 (T/Th Nov 14&16)	Distribution-free predictive inference: Holdout methods Jackknife+ and cross-validation methods Conformal prediction	Problem set 3
9 (T/Th Nov 28&30)	Distribution-free inference continued: Prediction beyond the iid setting Regression, calibration, & other topics Additional topics if time: e-values, sequential testing, etc.	Problem set 4
Finals (Dec 4–8)		Project 2 (group assignment) & optional problem set 5