2022 Fall Stat 156: Causal Inference

Peng Ding
University of California, Berkeley
Email: pengdingpku@berkeley.edu
Homepage: https://sites.google.com/site/pengdingpku/

- Units: 4
- Lectures: MW 5:00–6:30 pm; VLSB 2050
- Office hours: Tuesday 4–6pm; Evans Hall 425
- GSI: Lei Shi (Ph.D. in Biostatistics)
- GSI lab sessions: TBD
- GSI office hours: TBD

1. Description

This course focuses on approaches to causal inference using the potential outcomes framework and causal diagrams at an intuitive level. We will cover both theory and data analysis. Students are to be exposed to statistical questions relevant to decision and policy making.

Topics covered include classical randomizaed experiments, observational studies, instrumental variables, principal stratification, and mediation analysis. Applications of these methods are drawn from a variety of fields including political science, economics, public health, and medicine.

Weekly lectures and lab sections will balance theory and applications: lectures focus on concepts and methods, while lab sections address applications and implementations using the statistical language R.

Textbook

- Lecture notes written by the instructor, available on Bcourses (the lecture notes will be updated through the semester)

References

- Angrist, J. and Pischke, J. (2008) Mostly Harmless Econometrics. Princeton University Press.
- Imbens, G. W. and Rubin, D. B (2015) Causal Inference for Statistics, Social, and Biomedical Sciences. Cambridge University Press.

bCourses We will post all lecture material (lecture notes, sample R code, homework, etc.) on bCourses. Your assignment grades will also be posted on bCourses. We will also use bCourses for course-related announcements.

Students with disabilities If you need accommodations for any physical, mental, or learning disability, please get in touch with me so that we can make the necessary arrangements.

2. Prerequisites

- Stat 135; if you have taken CS 189 you can also take this course with my permission (ask me by email)
- Stat 151A strongly recommended
- Software R

3. Evaluation

Students will be graded on a curve based on the following components:

- Homework (30%): approximately 8 biweekly assignments, mix of theory and data analysis
- Midterm (20%)
- Group project with a final report (30% for the final report; 20% for the video presentation): There will be two students to a group who are randomly assigned by the GSI. The GSI will provide more information later in the semester.

  The final report should be 15–20 pages and the video presentation should be 20–30 minutes.

  If you are analyzing real data, please make sure that the data are publicly available.

Academic integrity Code of Student Conduct is available at `http://students.berkeley.edu/uga/conduct.pdf`

It is highly recommended that students form study groups in order to complete the homework assignments. Although it is recommended that people work together in order to complete the assignments, students must hand in their own individual answers. Photocopies and other reproductions of someone else's answers are not acceptable. Students should hand in the answers to the problem sets, and all computer code written to find those answers.

During exams, students are not allowed to communicate or cooperate with anyone in any way about exam. Any questions should be asked directly to me and the GSI.

Late assignments Due dates will be strictly enforced, but you can drop one homework assignment with the lowest score.

4. Course outline

4.1. Association and paradoxes: week 1

- Measures of association: risk difference, risk ratio, odds ratio, difference-in-means, regression coefficients
- Yule–Simpson Paradox and real-life examples

4.2. Neyman–Rubin potential outcomes model: week 2

- Intervention and outcome
- Experimental units and potential outcomes
- Causal effects
- Treatment assignment mechanism and the propensity score
- Inference of causal effects

4.3. Randomized experiments: weeks 2, 3

- Completely randomized experiment: difference-in-means estimator, Neyman's repeated sampling evaluation [week 2/1]
- Stratified randomized experiment, post-stratification [week 2/2]
- Rerandomization and analysis of covariance [week 3/1]
- Matched pair designs [week 3/2]

4.4. Unconfounded observational studies: weeks 4, 5, 6

- Observational studies, confounding by observed and unobserved covariates, unconfounded treatment assignment mechanism, ignorability, selection on observables [week 4/1]
- Causal inference based on outcome regression, including the stratified estimator, and its drawbacks [week 4/2]
- Causal inference based on treatment assignment: propensity score: covariate balance, stratification, weighting [week 5]
- Causal inference combining outcome regression and treatment assignment: doubly-robust estimator [week 6/1]
- Matching estimators: univariate and multivariate matching, propensity score matching, bias-corrected matching estimator [week 6/2]

4.5. Violations and difficulties of the ignorability assumption: weeks 7, 8, 9

- Unmeasured confounding and sensitivity analysis: Cornfield inequality and extensions, Rosenbaum–Rubin–Imbens sensitivity analysis [week 7]
- Negative exposure and negative outcome approach to assess the impact of unmeasured confounding [week 8/1]
- Lack of overlap in covariate distributions, trimming, target population, regression discontinuity [week 8/2]
- Pearl's critiques: causal diagram, back-door criterion, M-bias and Z-bias [week 9]

4.6. Instrumental variables: weeks 10, 11

- Encouragement design, identification and estimation under monotonicity and exclusion restriction, local average treatment effect (LATE) [week 10]

- Two-stage least squares and its connection to the potential outcomes approach [week 11]

4.7. Causal inference with intermediate variables: week 12

- Principal stratification, lack of nonparametric identifiability, mixture distribution, large-sample bounds [week 12/1]

- Mediation analysis, Pearl's nest potential outcomes and sequential ignorability, Baron–Kenny method [week 12/2]

4.8. Other topics depending on time

- Time-varying treatment and outcomes: Robins' g-computation

- Causal inference with longitudinal or panel data: fixed effects model, synthetic control