DS-UA 201: Causal Inference

Last updated: October 18, 2023

Instructor: Professor Moy (bryant.moy@nyu.edu)
Office: 19 West 4th St., Room 223
Office Hours: Mondays 9:00am - 10:00am 19 West 4th St., Room 223 (or by appointment)

TAs:
David McGrath (dm4947@nyu.edu)
Office Hours: Tuesday 11:00am-12:00pm, 19 West 4th St., Room 416

Dias Akhmetbekov (da2669@nyu.edu)
Office Hours: Thursday 1:00pm-2:00pm, 19 West 4th St., Room 302

Jiaxu Ren (jr5674@nyu.edu)
Office Hours: Thursday 2:00-3:00pm, 60 Fifth Avenue Room 244

Lectures: Tuesdays and Thursdays 3:30pm - 4.45pm,
19 West 4th St., Room 101
Section 002: Friday 10:15am-11:05am,
60 Fifth Avenue, Room 110
Section 003: Fridays 11:15am - 12:05pm
60 Fifth Avenue, Room 110
Section 004: Fridays 12:30pm - 1:20pm,
60 Fifth Avenue, Room 110
Section 005: Fridays 9:00am - 9:50am,
60 Fifth Avenue, Room 110
Section 006: Fridays 5:55pm - 6:45pm,
60 Fifth Avenue, Room 110

Course Overview

We often want to know the relationship between cause and effect. Almost every domain has significant causal research questions that can drive decision making. Labor economists want to know whether job training programs successfully increase participants' wages. Epidemiologists want to know whether a particular medical treatment improves quality of life. Advertisers want to know whether a marketing campaign is effective at boosting sales. You've probably heard that "correlation does not imply causation." But that raises the question: What exactly is causation and how can it be determined whether an observed relationship is truly causal?

This course will teach you the fundamentals of how to reason about causality and make causal determinations using empirical data. It will begin by introducing the counterfactual framework of causal inference and then discuss a variety of approaches, starting with the most basic experimental designs to more complex observational methods, for making inferences about causal relationships from the data. For each approach, we will discuss the necessary assumptions that a researcher needs to make about the process that generated the data, how to assess whether these assumptions are reasonable, and finally how to interpret the quantity being estimated.

This course will involve combination of lectures, sections and problem sets. Lectures will focus on introducing the core theoretical concepts being taught in this course. Sections will emphasize application and discuss how to implement various causal inference techniques with real data sets. Problem sets will contain a mixture of both theoretical and applied questions. The problem sets serve as a way of reinforcing key concepts and they allow students to assess their progress throughout the course.

As a part of this course, you will be introduced to statistical programming using the R programming language. This is a free and open source language for statistical computing that is used extensively for data analysis in both academia and industry. No prior experience in programming is necessary and we recognize that students will come in with a variety of backgrounds and different levels of experience in programming. This course is designed to emphasize learning by doing and will teach statistical programming with the aim of preparing students to analyze actual data.

Prerequisites

DS-UA 111 (Data Science for Everyone) is a great introduction to probability, statistical inference and programming and is recommended for taking this course. However, because introductory statistics is taught in a variety of ways by a variety of disciplines, we are very flexible in allowing students with other backgrounds in statistics to take this course.

In general, the necessary prior knowledge of statistics required for success in this course is minimal – if you have a general familiarity with linear regression, you are more than ready for this class. The first few weeks will incorporate a review of the most important concepts (e.g. probability, random sampling, conditional averages) and we will include refreshers whenever additional concepts are introduced throughout the course. The focus of this course is on developing students' ability to reason systematically about causal relationships. Students with significant experience in data analysis and descriptive statistical inference and those with less prior background will benefit from and be able to succeed in this course.

Logistics

Lectures Lectures will introduce the main topics described in this syllabus. Attending the lecture is **strongly encouraged**. Lecture slides will be made available after the lecture.

Lab Sections Sections are designed to teach the implementation of the statistical methods we discuss in lecture in the R programming language. You will be learning how to code in R and how to generate write-ups of your analysis using RMarkdown. We will be working through sample coding exercises that aim to help you in completing the problem sets. All material (including code and data) will be made available on the course website prior to the lab section. Attendance in lab sections are mandatory and will count towards participation credit.

Online Discussion: Outside of the lecture and sections, we will be using BRIGHTSPACE as a discussion platform for the course. We encourage you to both ask and answer questions on this forum. Participation in the discussion will be taken into account when determining participation grades. The instructor and TAs will do their best to answer questions directed at them within 24 hours of posting during the work week, but please understand that this may not always be possible and some questions may require longer to be answered.

Office Hours: The instructor will be holding office hours every week unless otherwise announced. These office hours will be in person. Students who cannot attend in-person office hours for any reason are welcome to schedule an appointment (either in person or virtual) with the instructor via email. All of the TAs will also be holding office hours. The times and locations will be included at the beginning of this document.

COVID-19 Accommodations

While instruction is fully in-person, we recognize that things surrounding COVID-19 will be messy and difficult for all of us. As such, my goal is to be flexible in granting reasonable accommodations for any issues that may arise for students this year. Likewise, we hope that you will be flexible as we collectively move throughout the semester.

Textbooks

You may find the list of books useful, but you are not required to purchase your own copy. We will make use of NYU Library ebook collection and I will provide excerpts of other readings on the course website.

- Imai, Kosuke. *Quantitative Social Science: An Introduction*. Princeton University Press. 2017.
- Cunningham, Scott. *Causal Inference: The Mixtape*. Yale University Press. 2021. (Available to read online via NYU Library)

- Angrist, Joshua D., and Jorn-Steffen Pischke. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press. 2009. (Available to read online via NYU Library)
- Imbens, Guido W. and Donald B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences*. Cambridge University Press. 2010. (Available to read online via NYU Library)
- Hernán, Miguel A. and James M. Robins. *Causal Inference: What If*. Chapman & Hall/CRC. 2020. (PDF available at: https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/)

The course will follow the Imai textbook the closest. The textbook is designed to introduce students to both statistical computing and causal inference through a variety of applied examples and exercises. We hope that it will be useful to you as a reference even after the course is over.

Requirements

Students' final grades are based on three components:

- **Problem sets** (40% of the course grade). Students will complete a total of four problem sets throughout the semester. Problem sets will primarily cover topics from the lecture and section for that week and the previous weeks. Problem sets are designed to be somewhat more challenging than both the midterm and final exams and we do not expect students to perform perfectly on each problem set.
  - *Collaboration policy*: Collaboration between students on the problem sets is strongly encouraged and highly recommend that students discuss problems with each other. However, each student is expected to submit their own write-up of the answers and any relevant code. **Students may not copy each other's answers, including any R code.** Any sharing or copying of assignments is considered cheating and will result in an F in the course. A second cheating incident will, by CAS rules, result in a one-semester suspension from the College.
  - *Office hours and online discussion*: Students should feel free to discuss any questions about the problem sets with the teaching staff during sections and office hours. We also strongly encourage students to post questions about both the problem sets and the assigned readings on the course discussion board and respond to other students' questions. Responding to other students' questions will contribute to your participation grade.
  - *Submission guidelines*: Problem sets will be distributed as `PDF` and `Rmarkdown` files (`.Rmd`). You should submit your answers and any relevant R code in the same format: including an `Rmarkdown` file (`.Rmd` extension) and a corresponding compiled `.pdf` file as your submission. `Rmarkdown` combines the text formatting syntax of Markdown markup language with the ability to embed and execute chunks of `R` code directly into a text document. This allows you to present your code, graphical output,

and discussion/write-up all in the same document. We recommend that you edit the distributed `Rmarkdown` file for each problem set directly.

  - *Late submission*: All homework assignments must be submitted **on time**. If an assignment is submitted late, every day since the assignment deadline will count as a grade drop (A to A- to B to B- and so on ...). If a student cannot submit an assignment on time due to unavoidable circumstances he/she/they must submit documentation proving their circumstances, and appropriate action will be taken by the teaching team.
  - *Extra Credit*: We understand that life can get in the way of coursework, so students are allowed to replace their lowest problem set grade by completing an optional extra credit assignment. The assignment is to find a news article discussing a study that claims to find a causal effect, find the underlying paper that the article cites, read it, and write a brief 500-ish word blog post discussing the design of the study, whether the findings are persuasive, and what important caveats or information from the paper (if any) you think the news article should have mentioned. This assignment is designed to be very open-ended and to permit you to choose any topic that you are interested in. It will be due on the last lecture day.

- **Take-home midterm and final exams** (25% and 30% of the course grade respectively). *The midterm will be released after sections conclude on October 20 and must be turned in by* ***October 25th at 11:59PM.*** The midterm will consist of questions on the course material so far.

  The final will be similar in structure to the midterm, but it is designed to evaluate your knowledge of the course material. The final will be cumulative. *The final exam will be released on December 14 and will be due by* ***December 20 at 5PM.***

  Unlike the problem sets, **students are not permitted to collaborate with other students** during either of these exams. Collaboration during either of the exams will be treated as cheating. Direct any clarifying questions to the instructor. Exam times are **firm**. If a student thinks they are going to miss either exam, they will have to submit documentation proving an **unavoidable** and serious circumstances preventing them from taking the exam. Each case will be handled separately by the teaching staff and different solutions might be suggested for different cases. In absence of such documented, serious circumstances for missing either exam, students will receive a fail grade on the exam that was missed.

- **Participation** (5% of the course grade). Students are expected to take an active role in learning and engage with the course. We will take a very *broad* view of what engagement means. Asking and answering questions during lectures, labs, and posting on the discussion board will contribute to participation. *I reserve the right to administer unannounced pop quizzes in class. Grades for these pop quizzes will count towards your participation credit.*

Grading

All assignments (problem sets, exams, extra credit) will be graded on a point total, and each question's total point value will be displayed on the assignment. Point scores will be converted to percentage points. To construct final grades, percentage points will be added together and weighted

by the weight given to each final grade component. Percentage grades will then be translated to letter grades. A curve may be applied to grades at any stage of this process at the teaching team's discretion. There is no pre-determined, fixed curve, however our objective will be to have final grades that have a fair distribution that is similar to other courses of a similar level at NYU.

Grade corrections and regrading All grading decisions made by the teaching team are intended to be final. In the unusual circumstance that a student believes there has been an error in the grading for an assignment, it will be possible to submit a regrade request. To request a re-grade, send an email to both your TA *and* instructor via email. Regrade requests will be handled periodically by the teaching team. The graded exercise in question will then be re-assessed **in full** by the grader, who will make a final decision. *Note that this implies that grades may also be lowered, following a re-assessment.* Additional regrade requests after this decision will not be possible.

Computing

This course will also serve as an introduction to statistical computing using the `R` programming language. This is a free and open source programming language that is available for nearly all computing platforms. You should download and install it from `http://www.r-project.org`. Unless you have strong preferences for a specific coding environment, we recommend that you use the free RStudio Desktop Integrated Development Environment (IDE) which you can download from `https://rstudio.com/products/rstudio/download/#download`. In addition to being a great and simple to use environment for editing code, `RStudio` makes it very easy to write and compile `Rmarkdown` documents: the format in which problem sets will be distributed. In addition to base `R`, we will introduce students to data management and cleaning via the tidyverse set of packages along with basic graphics and visualization using ggplot2.

Moses Statement

Disability Disclosure Statement: Academic accommodations are available for students with disabilities. The Moses Center website is `www.nyu.edu.csd`. Please contact the Moses Center for Students with Disabilities (212-998-4980 or mosescsd@nyu.edu) for further information. Students who are requesting academic accommodations are advised to reach out to the Moses Center as early as possible in the semester for assistance.

Schedule

A **tentative** schedule of topics is provided below. This schedule is subject to change depending on time, student interest, and how the class feels about the course's pacing.

Week 1: Introduction and Math Review

Introduction (September 5)

Statistics Review: Probability, Hypothesis Testing, Estimators (September 7)

- *The following readings are for you to **skim** if you are not comfortable with probability and statistics*
- Imai, Chapter 6
- Imai, Chapter 1.3 (pp. 10-27)
- Cunningham, Probability and Regression Review p. 16-95

Week 2: Statistics, Potential Outcomes, Confounding

Ignorability and Confounding (September 12 and 14)

- Cunningham, Potential Outcomes Causal Model p. 119-148

Week 3: Experiments and Randomization Inference

Experiments (September 19)

- Imai, Chapter 2.4 (pp. 48-54)
- Angrist and Pischke, Chapter 2, "The Experimental Ideal" (pp. 11-24)

Randomization Inference (September 21)

- Cunningham Randomization Inference section pg. 148-174

Week 4: CATEs

CATEs + Treatment Effect Heterogeneity (September 26)

CATEs+ Treatment Effect Heterogeneity 2 (September 28)

Problem Set 1 Assigned on Sept. 29th, due Oct. 6th

Week 5: Blocking and Stratification

Blocking and Stratification (October 3)

Blocking and Stratification 2 (October 5)

Week 6: DAGs

LEGISLATIVE MONDAY (October 10)

DAGs (October 12)

- Cunningham, Causal Inference Mixtape, Chapter 4 DAGs

Week 7: Observational Inference

Observational Inference 1 (October 17)

Observational Inference 2 + Propensity Scores (October 19)

- Imai, Chapter 2.5.1 - 2.5.2

Midterm Released on Oct. 20 after sections conclude, Due October 25th at 11:59PM

Week 8: Midterm, Propensity Scores, Mathcing

Midterm Help Hours During Class Time (October 24)

Midterm Due Oct. 25th at 11:59PM

Propensity Scores + Matching (October 26)

- Ho et. al. 2007. "Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference." Political Analysis, Vol. 15: 199-236.

Week 9: Matching + Regression Adjustment

Matching 2 (October 31)

Regression (November 2)

- Imai, Chapter 2.5.3
- Cunningham, Probability and Regression Review

Problem Set 2 Assigned on Nov. 3rd, due Nov. 10th

Week 10: Regression Adjustment + Group Data

Regression pt. 2 (November 7)

Group Data (November 9)

- Morgan and Winship, Chapter 6
- Angrist and Pischke, Chapter 5.1

Week 11: DiD

DiD (November 14)

- Cunningham, The Causal Inference Mixtape, Chapter 10 - Differences-in-differences

DiD pt. 2 (November 16)

- Angrist and Pischke, Chapter 5.2

Problem Set 3 Assigned on Nov. 17, due Dec. 1

Week 12: IV

Instrumental Variables: assumptions and motivation (November 21)

- Cunningham, The Causal Inference Mixtape, Chapter 8 - Instrumental Variables
- Angrist and Pischke, Chapter 6
- Angrist, Imbens and Rubin (1996) "Identification of causal effects using instrumental variables." Journal of the American Statistical Association, 91:434, 444-455

NO CLASS - THANKSGIVING RECESS (November 23)

Week 13: IV + Regression Discontinuity Designs

Instrumental Variables: estimators (November 28)

RDD: Introduction and Examples (November 30)

Week 14: RDD + ML + Ethics

RDD: Estimators and Inference (December 5)

Machine Learning and Ethics (December 7)

- Jones, Jason J., et al. "Social influence and political mobilization: Further evidence from a randomized experiment in the 2012 US presidential election." PloS one 12.4 (2017).

Problem Set 4 Assigned on Dec 6, due Dec. 13th

Week 15: Bonus Content and Review

Bonus Content (December 12)

Course review and Questions about Final (December 14)

Take-Home Final Exam

Released on 12/14, Due on 12/20 at 5PM