DS-UA 201: Causal Inference

Last updated: November 17, 2021

Instructor:	Marco Morucci (marco.morucci@nyu.edu)
Office:	Room 631, 6th floor, 60 5th Avenue
Office Hours:	Mondays 4:00-5:00pm Room 244, 60 th Ave.
TAs:	Yutong Chen (yc4127@nyu.edu)
Office Hours:	Wednesdays 10:00-11:00am Room 244, 60 th Ave.
	Jessie Guo (jg6180@nyu.edu)
Office Hours:	Thursdays, 11:30am-12:30am Room 244, 60 th Ave.
	Surya Sama (ss14323@nyu.edu)
Office Hours:	Tuesdays, 10:00am:11:00am Room 206, 60 th Ave.
Lectures:	Mondays and Wednesdays 2pm-3:15pm,
	Silver Center for Arts and Science, 100 Washington Square East, room 206
Lab Section 1:	Thursdays 10:15am-11:05am,
	7194 Mercer Street, room 208
Lab Section 2:	Fridays 2:45pm-3:35pm,
	Bobst Library, 70 Washington Square South, room LL138
Course Website	`https://brightspace.nyu.edu/d2l/home/130191`
Course Q&A on Piazza	`https://piazza.com/nyu/fall2021/dsua201/home`

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 and serve as a way of reinforcing key concepts and allowing students to assess their progress and understanding 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. Please contact the instructor, Marco Morucci, if you are interested in enrolling but do not have DS-UA 111 as a prerequisite.

In general, the necessary prior knowledge of statistics required for success in this course is very 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. Due to the present circumstances, in-person attendances at lectures and labs will not be required, however it is **strongly recommended**. Lecture slides will be made available right after the lecture. Per NYU policy, lectures and labs in the first two weeks of class (until Sep 15) will be recorded and made available to students who cannot attend in person. After the end of add/drop we will evaluate whether to keep recoding lectures depending on student demand, need, and burden on the teaching team.

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. As with lectures, attendance is not mandatory but is strongly recommended. Lab sections will be recorded and made available on the class website until Sep. 15.

Online Discussion: We will be using PIAZZA (https://piazza.com/nyu/fall2021/dsua201/home) as the primary discussion platform for the course. We encourage you to both ask and answer questions on this forum. Participation on PIAZZA 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. These office hours will be virtual over zoom until Sep. 15th. After that date instructor office hours will switch to in person at the scheduled time. 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 the TAs will also be holding office hours, with at least one TA holding office hours remotely over zoom each week.

COVID-19 Accommodations

While instruction at NYU is fully in-person this semester's course will be taking place in an environment that is very much not normal – there is still a massive pandemic. It is absolutely okay to understand and recognize that things 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 figure out how to best re-enter in-person teaching. We encourage any and all feedback on the course design and am open to making changes as we progress through the semester.

In the event that in-person classes are no longer possible, this course will transition to a fully online format. This will not affect the meeting times for the course, but all lectures and labs will be conducted entirely via Zoom. Likewise, all assignments and exams will be made fully take-home. We will adjust the course schedule as needed to account for the likely disruption to student schedules.

Textbooks

You may find the following books useful, however, these are not required for the course:

- Imai, Kosuke. *Quantitative Social Science: An Introduction*. Princeton University Press. 2017.
- Angrist, Joshua D., and Jorn-Steffen Pischke. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press. 2009.

- Imbens, Guido W. and Donald B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences*. Cambridge University Press. 2010.
- 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. The course will primarily be focusing on chapters 1-3 with occasional excerpts from later chapters.

Requirements

Students' final grades are based on three components:

- **Problem sets** (35% 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 week. 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. Problem sets will be assigned on Wednesdays and will be due two weeks later on Tuesday (by 11:59pm).
  - *Collaboration policy*: Collaboration between students on the problem sets is strongly encouraged and highly recommend that students discuss problems with each other via video conferencing or via the course's online discussion board. 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 (Piazza) 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, especially in the present pandemic, and students may be unable to finish a problem set or do as well as they would have liked. Students are therefore 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 day of classes, **December 13**.

- **In-class midterm and take-home final exams** (25% and 35% of the course grade respectively). The midterm will be in-class during lecture time on October 27th, and will consist of questions on the course material so far. The final will be similar in structure to the problem sets and is designed to evaluate your knowledge of the course material. The final will be cumulative. Unlike the problem sets, **students are not permitted to collaborate with other students** during either of these exams. Collaboration during either of this exam will be treated as cheating. The teaching staff will answer any clarifying questions in person during the midterms on the Piazza discussion board during the final. 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. The final exam will be released on 12/15 and will be due by 11:59pm on 12/21.

- **Participation** (5% of the course grade). Students are expected to take an active role in learning and engage with the course. Because many students will be attending remotely and asynchronously, We will take a very *broad* view of what engagement means. Asking and answering questions during lectures and lab will contribute to participation, but so will interactions on the PIAZZA discussion board or asking questions during office hours.

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 through an online form that will be made available in Brightspace. 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.

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

Introduction (September 8)

Week 2: Statistics and probability review

Statistics Review (September 13)

Statistics Review Pt. 2 (September 15)

Week 3: Causation

Introduction to Causation (September 20)

Ignorability and unconfoundedness (September 22)

Week 4: Randomized Experiments

Randomized Experiments (September 27)

Randomization Inference (September 29) Problem Set 1 Assigned on Sept. 29th, due Oct 12th.

Week 5: Randomized Experiments and contextual covariates

Treatment Effect Heterogeneity (October 4)

Blocking and Stratification (October 6)

Week 6: Observational inference

Introduction to Observational Inference (October 12)

Directed acyclic graphs (October 13) Problem Set 2 Assigned on Oct. 13th, due Nov. 9th.

Week 7: Weighting Methods

The Propensity Score (October 18)

Introduction to Matching (November 1)

Week 8: Midterm

Matching Estimators (October 25)

In-class midterm (October 27)

Week 9: Matching Methods

Regression pt. 1 (November 1)

Regression pt. 2 (November 3)

Week 10: Regression Adjustment

Causal Inference with grouped data (November 8)

Difference in differences pt. 1(November 10) Problem Set 3 Assigned on Nov. 10th, due Nov. 23rd.

Week 11: Difference in Differences

Difference in differences pt. 2(November 15)

Instrumental Variables: assumptions and motivation (November 17)

Week 12: Instrumental Variables

Instrumental Variables Estimators(November 22)

RDD: Introduction and Examples (November 24) Problem Set 4 Assigned on Nov. 24th, due Dec. 7th.

Week 13: Regression Discontinuity Designs

RDD: Estimators and Inference (November 29)

Sensitivity Analysis (December 1)

Week 14: Sensitivity Analysis

Causal Inference and Machine Learning (December 6)

Causal Inference and Ethics (December 8)

Week 15: Review and Conclusions

Course review and final practice (December 13)

Take-Home Final Exam

Released on 12/15, Due on 12/21

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.