

Sociology 504: Advanced Social Statistics

Sam Trejo & Alessandra Rister Portinari Maranca
Class: Monday/Wednesday 10:30 a.m.-12:00 p.m.
Lab: Friday 9:00 a.m.-11:00 a.m.
Wallace Hall 165

Professor Sam Trejo

Office: 187 Wallace Hall
Office Hours: Mondays 1:00 - 2:00pm
Email: samtrejo@princeton.edu
www.samtrejo.com

Preceptor Alessandra Rister Portinari Maranca

Office: 127 Wallace Hall (Office Hours will be held in Sociology lounge)
Office Hours: Thursdays 12:30-1:30pm
Email: arpm@princeton.edu

Note: this course is designed for first-year doctoral students in the social sciences. The prerequisite is Sociology 500 – which covers introductory probability, multivariate linear regression, and the foundations of causal inference – or a similar statistics course. Please speak with the instructor if you have not taken Sociology 500 but are considering enrolling in Sociology 504.

“Sociology is not like physics. Nothing but physics is like physics, because any understanding of the world that is like the physicist’s understanding becomes part of physics...”

– OTIS DUDLEY DUNCAN

NOTES ON SOCIAL MEASUREMENT, 1984

Course Description

This course is the second class in the required first-year statistics sequence for doctoral students in the Department of Sociology. The overarching goal of the two-course sequence is to help students grow from consumers to producers of quantitative social research. Students learn the statistical and computational principles necessary to perform modern and innovative analysis of quantitative social data. The sequence's capstone is a replication project in which students choose a published work of social science, reproduce (and extend) the paper's key results, and then present their findings via a poster session at the Department of Sociology's annual graduate research day on Friday, May 1st.

In terms of statistical content, Sociology 504 is divided into two modules. The first module – **Applied Causal Inference** – focuses on experimental and quasi-experimental methods utilizing the potential outcomes framework (also known as the Neyman–Rubin causal model). We cover randomized experiments, instrumental variables, regression discontinuity designs, matching, difference-in-differences, and synthetic control methods. The second module – **Modeling Discrete Outcomes** – concerns generalized linear models and their estimation using maximum likelihood approaches. We cover a variety of strategies for modeling dichotomous, ordinal, categorical, and count outcomes.

Formal instruction for the course is split into lecture (Monday/Wednesday) and lab (Friday). Both lecture and lab are essential parts of the learning process. The lecture covers the core statistical material, whereas the lab focuses on computational skills and applied research practices. By the end of the Spring semester, students should be able to read an original scholarly article describing a new statistical technique, implement the model with relevant data using statistical software, interpret the results, and explain the findings to someone unfamiliar with statistics. This course requires a lot of hard work, but the payoff is well worth it; if a student is willing to put in the time, we are always happy to help.

Assignments & Grading

Grades in the course will be assigned according to the following breakdown:

Participation & Attendance	10%
Weekly Memos & Problem Sets	50%
Replication & Extension Project	40%

Participation

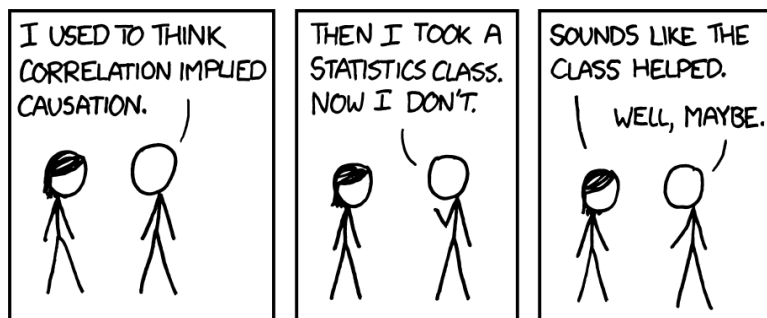
Students are expected to attend and participate in all class sessions. Please email both the instructor and the preceptor in advance explaining your situation in the event that you need an excused absence. Note, two weeks of cumulative absences, regardless of the reason a student misses class, may constitute grounds for a failing grade in a course.

Weekly Memos & Problem Sets

Each week, a homework assignment will be due by 11:59 p.m. on Sunday night. Homework assignments will be submitted and graded via Gradescope. There are two different types of homework assignments, which assigned on alternating weeks: **memos** and **problem sets**. Both memos and problem sets are to be completed individually. Memos are written responses that include reflections and updates regarding the Replication & Extension project as well as conceptual questions related to the course's required readings. Problem sets are assignments which involve mathematical derivations and/or the analysis of empirical data. Homework assignments that display a concerted effort to fully address each question will receive a grade of at least 80%, with the final 20% awarded for thoroughness, clarity, and accuracy. Late assignments sets will be penalized 10% off per day late (except in the case of documented emergencies); if you turn in a problem set late, you are on your honor to **avoid looking at the solution key** before submitting your work.

Replication & Extension Project

The primary assignment in this course is a research paper – written individually or in pairs – that applies an advanced statistical method to a substantive problem in your field of study. This assignment is structured via the replication and extension of an already published paper. Ultimately, the goal is for each student to produce a publishable article. This assignment serves in place of a final exam. There will be a number of interim deadlines, which are listed below and described further on the replication and extension project handout.



A Note on Artificial Intelligence

Students are expected to follow Princeton University's standards for academic integrity (see these [guidelines](#)). AI tools (e.g., ChatGPT, Claude, Copilot) may be used to support learning—for example, for brainstorming, clarifying concepts, improving writing clarity, debugging code, or assisting with LaTeX/Markdown—but they may not replace your own analysis or original work. If you use AI, you must submit a record of your interaction with the tool, and you must acknowledge any direct contribution to your written work in a brief footnote or author's note. This means that using AI to generate complete or partial answers to problem sets is strictly prohibited. All submitted work must reflect your own reasoning, critical engagement with the material, and independent understanding.

Laptop Use

A growing body of evidence suggests that the use of laptops, tablets, and phones in classrooms tends to be detrimental to learning. In general, we discourage their use on lecture days. However, if you want to use a device during class, we ask that you contact us outside of class to make this request. Those who choose to use their laptops will be asked to sit in the back of the room so as to provide the least distraction to other students. For more context on this policy, see [this video](#).

Due Dates

Week	Date	Assignment Due
1	Sunday, January 25	N/A
2	Sunday, February 1	Paper Selection Memo
3	Sunday, February 8	Problem Set #1
4	Sunday, February 15	Extension Conceptualization Memo
5	Sunday, February 22	Problem Set #2
6	Sunday, March 1	Data Acquisition Memo
Spring Break		
7	Sunday, March 15	Problem Set #3
8	Sunday, March 22	Replication Memo
9	Sunday, March 29	Problem Set #4
10	Sunday, April 5	Peer Feedback Memo
11	Sunday, April 12	Problem Set #5
12	Sunday, April 19	R&E Poster Draft
End of Spring Classes		
	Friday, May 1	Grad Research Symposium
	Tuesday, May 5	Final R&E Paper
	Friday May 8	Peer Review Report

Course Calendar

MONDAY		WEDNESDAY	
Jan 26th Introduction	1	28th Counterfactuals & Causality •	2
Feb 2nd Counterfactuals & Causality •	3	4th Randomized Experiments •	4
9th Instrumental Variables •	5	11th Discussion of Problem Set #1 •	6
16th Instrumental Variables •	7	18th Regression Discontinuity Designs •	8
23rd Discussion of Problem Set #2 •	9	25th Interrupted Times Series •	10
Mar 2nd Difference-in-Differences •	11	4th Synthetic Control Methods •	12
9th <i>-Spring Recess-</i>		11th <i>-Spring Recess-</i>	
16th Discussion of Problem Set #3 •	13	18th Binary Outcome Models •	14
23rd Maximum Likelihood Estimation I •	15	25th Maximum Likelihood Estimation II •	16
30th Discussion of Problem Set #4 •	17	Apr 1st Generalized Linear Models I •	18
6th Generalized Linear Models II •	19	8th Generalized Linear Models III •	20
13th Discussion of Problem Set #5 •	21	15th Regularization & Pooling •	22
20th <i>~Practice Poster Presentations~</i>	23	22nd <i>~Practice Poster Presentations~</i>	24

Course Readings

Part I: Applied Causal Inference •

- *Mostly Harmless Econometrics*. Joshua D. Angrist and Jörn-Steffen Pischke. (Princeton University Press, 2009).
- *Mastering 'Metrics*. Joshua D. Angrist and Jörn-Steffen Pischke. (Princeton University Press, 2014).
- *Counterfactuals and Causal Inference (2nd Edition)*. Stephen L. Morgan and Christopher Winship. (Cambridge University Press, 2014).

Part II: Modeling Discrete Outcomes •

- *Unifying Political Methodology: The Likelihood Theory of Statistical Inference*. King, Gary. (Cambridge University Press, 1989).
- *Statistical Methods for Categorical Data Analysis (2nd Edition)*. Daniel Powers and Yu Xie (Emerald Publishing, 2008).
- *Applied Regression Analysis and Generalized Linear Models (3rd Edition)*. John Fox. (SAGE Publications, 2015).

REQUIRED

OPTIONAL

PRINCETON

Counterfactuals & Causality

- *Statistics and Causal Inference*. Paul W. Holland. (Journal of the American Statistical Association, 1986).
- *Mostly Harmless Econometrics* Chapter 1
- *Endogenous Selection Bias: The Problem of Conditioning on a Collider Variable*. Felix Elwert and Christopher Winship. (Annual Review of Sociology, 2014).
- *What We Inherit: How New Technologies and Old Myths Are Shaping Our Genomic Future*. Sam Trejo & Daphne Martschenko. (Princeton University Press, 2026; Chapter 2).

Randomized Experiments

- *Mastering Metrics* Chapter 1 or *Mostly Harmless Econometrics* Chapter 2
- *Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market*. Matthew J. Salganik, Peter S. Dodds, and Duncan J. Watts. (Science, 2006).
- *Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects*. David S. Lee. (Review of Economic Studies, 2009).
- *Double Jeopardy: Teacher Biases, Racialized Organizations, and the Production of Racial/Ethnic Disparities in School Discipline*. Jayanti Owens. (American Sociological Review, 2022).
- *Policing the Boundaries of Blackness: How Black and White Americans Evaluate Racial Self-identifications*. Marissa Thompson, Sam Trejo, AJ Alvero, and Daphne Martschenko. (American Journal Sociology, 2026).

Instrumental Variables

- *Handle With Care: a Sociologist’s Guide to Causal Inference With Instrumental Variables*. Chris Felton and Brandon M. Stewart. (Sociological Methods & Research, 2022).
- *Mastering Metrics* Chapter 3 or *Mostly Harmless Econometrics* Chapter 4
- *Identification and Estimation of Local Average Treatment Effects*. Guido W. Imbens and Joshua D. Angrist. (Econometrica, 1994).
- *Pounds That Kill: The External Costs of Vehicle Weight*. Michael L. Anderson and Maximilian Auffhammer. (Review of Economic Studies, 2014).
- *Community and the Crime Decline: the Causal Effect of Local Nonprofits on Violent Crime*. Patrick Sharkey, Gerard Torrats-Espinosa, and Delaram Takyar. (American Sociological Review, 2017).
- *The Effect of Violent Crime on Economic Mobility*. Patrick Sharkey and Gerard Torrats-Espinosa. (Journal of Urban Economics, 2017).
- *The Effects of Active and Passive Leisure on Cognition in Children: Evidence From Exogenous Variation in Weather*. Thomas Laidley and Dalton Conley. (Social Forces, 2018).

Regression Discontinuity Designs

- *Mastering Metrics* Chapter 4 or *Mostly Harmless Econometrics* Chapter 6
- *Do Better Schools Matter? Parental Valuation of Elementary Education*. Sandra E. Black. (Quarterly Journal of Economics, 1999).
- *High Stakes in the Classroom, High Stakes on the Street: The Effects of Community Violence on Student’s Standardized Test Performance*. Patrick Sharkey, Amy Ellen Schwartz, Ingrid Gould Ellen, and Johanna Lacoe. (Sociological Science, 2014).
- *Comparing Inference Approaches for RD Designs: a Reexamination of the Effect of Head Start on Child Mortality*. Matias D. Cattaneo, Rocio Titiunik, and Gonzalo Vazquez-Bare. (Journal of Policy Analysis and Management, 2017).
- *Regression Discontinuity Designs*. Matias D. Cattaneo and Rocio Titiunik. (Annual Review of Economics, 2022).

Difference-in-Differences

- *Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania*. David Card and Alan Krueger. (American Economic Review, 1994).
- *Mastering Metrics* Chapter 5 or *Mostly Harmless Econometrics* Chapter 5
- *Simple Approaches to Nonlinear Difference-in-Differences with Panel Data*. Jeffrey M. Wooldridge. (The Econometrics Journal, 2023).
- *Prenatal Exposure to an Acute Stressor and Children’s Cognitive Outcomes*. Florencia Torche. (Demography, 2018).
- *Local Exposure to School Shootings and Youth Antidepressant Use*. Maya Rossin-Slater, Molly Schnell, Hannes Schwandt, Sam Trejo, and Lindsey Uniat. (Proceedings of the National Academy of Sciences, 2020).

Synthetic Control Methods

- *Using Synthetic Controls: Feasibility, Data Requirements, and Methodological Aspects*. Alberto Abadie. (Journal of Economic Literature, 2021).
- *The Effects of the Flint Water Crisis on the Educational Outcomes of School-Age Children*. Sam Trejo, Gloria Yeomans-Maldonado, and Brian Jacob. (Science Advances, 2024).

Binary Outcome Models

- *Applied Regression Analysis and Generalized Linear Models* Chapter 14.1
- *Behind the Curve: Clarifying the Best Approach to Calculating Predicted Probabilities and Marginal Effects From Limited Dependent Variable Models*. Michael J. Hanmer and Kerem O. Kalkan. (American Journal of Political Science, 2013).

Maximum Likelihood Estimation

- *The Epic Story of Maximum Likelihood*. Stephen M. Stigler. (Statistical Science, 2007).
- *Unifying Political Methodology* Chapters 2 & 4

Generalized Linear Models

- *Applied Regression Analysis and Generalized Linear Models* Chapters 14.2 & 15.1
- *Evicting Children*. Matthew Desmond, Weihua An, Richelle Winkler, and Thomas Ferriss. (Social Forces, 2013).
- *A Model of Text for Experimentation in the Social Sciences*. Margaret E. Roberts, Brandon M. Stewart, and Edoardo M. Airoidi. (Journal of the American Statistical Association, 2016).

Regularization & Pooling

- *Machine Learning for Sociology*. Mario Molina and Filiz Garip. (Annual Review of Sociology, 2019).
- *Estimation in Parallel Randomized Experiments*. Donald B. Rubin. (Journal of Educational Statistics, 1981).