POLI SCI 403
Probability and Statistics
Fall 2024

Instructor: Gustavo Diaz (gustavo.diaz@northwestern.edu)
Teaching Assistant: Artur Baranov (artur.baranov@u.northwestern.edu)
GitHub Repository: https://github.com/gustavo-diaz/ps403
Canvas: https://canvas.northwestern.edu/courses/216757

Time and Place: Tues/Thurs 11:00am – 12:20pm, Scott Hall 212
Lab Hours: Fridays 2:00 – 3:20pm, Scott Hall 101
Student Hours: Schedule an appointment

Course Overview

This is the first course in the quantitative methods sequence for graduate students in the Department of Political Science. The course focuses on statistical inference from a social science perspective. Topics include probability, inference from random samples, linear regression, maximum likelihood estimation, identification, and causal inference.

Learning Objectives

The goal of statistical inference is to use the data we have to learn about something for which we do not have data. That connection requires making assumptions. This course aims at introducing tools and developing skills to conduct statistical inference with as few assumptions as possible.

By the end of the course, you will be able to apply statistical methods to conduct your own analyses, explain statistical tools and concepts in your own words, evaluate the credibility of applied and methodological work, and continue learning more advanced methods.

Prerequisites

There are no formal requirements to take this course other than enrollment in the Political Science PhD program or express approval.

This course does not assume prior training on statistics or quantitative methods beyond a grasp of US high school level algebra and calculus (which is covered in math camp). For example, you know that integrals have something to do with calculating the area under a curve, but you do not need to remember how to do integration by hand. I do assume you know how computer file systems work. For example, you are able to determine where a file is located in your laptop.

I expect you to participate actively, productively, and respectfully in our meetings. Some of the material addresses complicated concepts or uses math extensively. I do not expect you to understand every single equation for this course, but I do expect you to read carefully enough that you would understand every equation if you chose to revisit the material after taking this course.

Requirements

Reading

The main textbook we will follow is:

> **Textbook**
>
> Aronow, P.M. and Benjamin T. Miller. 2019. *Foundations of Agnostic Statistics*. Cambridge University Press

The rest of the syllabus refers to this book as AM. You can purchase a physical or digital copy directly from the publisher, although it is usually cheaper on online storefronts like Amazon. A digital copy is available through the library subscription at no additional cost.

The book tends to err on the side of brevity and mathematical rigor. Much of our class discussion, assignments, and additional reading will involve untangling and applying the topics in AM. Additional readings should be available through university library subscriptions or distributed promptly otherwise. You can find URLs for additional readings in the Schedule section.

A companion file to this syllabus includes recommended resources that may be useful to complement current or future learning.

Computing

We will use R and RStudio to work on assignments and classroom demonstrations. The advantage of R is that it is free and open source, meaning that you will be able to apply everything you learn in this course anywhere else. The disadvantage is a somewhat steep learning curve. I believe the investment is worthwhile for anyone working with data or in data-adjacent careers. You are welcome to use different software for statistical computing, but I cannot guarantee I will be able to help with troubleshooting.

You can install R and RStudio on your personal computer, which is the preferred use case. You can use this link for installation instructions on Windows and MacOS. See this link for installation instructions on Chromebooks, which is a bit more involved.

You can also use Posit Cloud to access RStudio from any web browser. A free account should suffice for the purposes of this course and has the advantage of letting you access your work across devices.

If you ever need more computing resources than what a personal computer or a free Posit Cloud account allow, you should consider requesting access to the Quest Analytic Nodes from Northwestern IT. I do not anticipate this to be relevant for this course, but it may be useful in the future.

Evaluation

Your final grade in this course will depend on the following:

- Participation
- Lab assignments (8 total, due Mondays 9 AM)
- Replication paper (due December 11 9AM)

Participation

This course does not formally require attendance, but I do expect the usual level of accountability required in a small graduate seminar. That means attending class regularly, doing the reading, asking questions, and working to foster a productive learning environment for everyone. Your participation inside and outside the classroom will be marked as satisfactory or unsatisfactory by the end of the quarter. If your participation leans toward an unsatisfactory mark, I will notify you by the end of Week 6 and give you feedback on how to improve.

Lab assignments

We will have weekly assignments aimed at practicing the application of course material with statistical software. These will range from coding exercises to problems that help illustrate theorems. In general, the goal of the lab assignments is to show why statistical analyses are (or should not be) conducted in a certain way.

Usually, we will start working on the lab assignment during our class meeting on Thursdays. This will give you an opportunity to clarify goals and expectations. On most weeks, you will need to work on your own time to finish the labs. You are welcome to work in groups during our meeting times and beyond, but you must submit individual reports.

Labs are due on Mondays at 9AM after they are assigned and must be submitted through Canvas in PDF format (you may use the original lab `.qmd` files as templates). Labs will be marked as *outstanding*, *satisfactory*, *unsatisfactory*, or *fail* (the last one applies if you did not submit or barely tried). I will mark late submissions as a fail unless I give you written approval to submit later. You can resubmit any labs marked as unsatisfactory at any time up until the final paper deadline. If your lab is marked as unsatisfactory, I will give you detailed feedback on what needs to be done to receive a satisfactory mark.

For most labs, I *expect* you to get stuck or be confused. Remember that the purpose of these assignments is to learn, and the quickest path to mastery is making mistakes. I find equal value on getting something right as I do on getting something wrong and doing your best to understand what went wrong and explaining how things should look like if they had gone right.

The implication is that, to assign a satisfactory or outstanding mark to your lab, I will be mostly looking for evidence that you have learned something useful.

Replication paper

You will submit a short replication paper as your final assignment. The goal of this paper is to apply what you have learned this quarter on a topic of your interest by reproducing the analysis of a previously published article, reflecting on the way it was conducted, consider what could be done differently, and possibly improve upon it.

You should think of this as the first step toward writing a publishable article in your field. You can find some guidelines on how to choose a publication to replicate here.

You are required to schedule a meeting with me to discuss the topic of your replication paper. The paper does not need to be restricted to the methods covered in our class (the only requirement is some form of statistical analysis), but this means you should be willing to work to learn the new methods on your own. I believe this is a valuable skill to practice early on.

You are allowed to work with a co-author or alone as you see fit. I am also open to discussing a different kind of paper if you know what you want to do.

Final papers are due on Wednesday, December 11 at 9AM. You should submit your paper in PDF format through Canvas. Your paper should follow the format of a research note, around 4,000 words and focusing on the main point. You are also required to submit an appendix in PDF format including all the code used to reproduce your tables, figures, and calculations. You may also include less important details in the appendix to keep the paper concise.

Like lab assignments, your final paper will be marked as *outstanding*, *satisfactory*, *unsatisfactory*, or *fail*. You are also welcome to resubmit a final paper marked as unsatisfactory at any point, but that means I will not be able to update your grade until after the final grade report deadline.

We will discuss more details about the final paper during Week 1 and throughout the term.

Grading

This course uses a labor-based grading agreement, commonly known as contract grading. In this course, instead of being given a final grade based on how "good" your submitted assignments are, your final grade will be based on the amount of labor you put into the course. The goal is to decouple grades from performance and emphasize learning and effort.

You will get a default grade if you meet the contract. It will go lower if you miss parts of the contract, it will go higher if you meet the baseline plus other criteria.

To meet the baseline grading contract (B+), you should:

- Complete Lab 0
- Be late (by a maximum of 24 hours) on no more than one lab assignment
- Submit the final paper before the deadline
- Complete 7 out of 8 lab assignments with a satisfactory or outstanding mark

- Receive a satisfactory or outstanding mark in the final paper
- Have a satisfactory participation status by the end of the semester

To get an A-, you should meet the baseline grading contract AND meet at least one of the following:

- Receive an outstanding mark in at least 3 labs
- Receive an outstanding mark in the final paper

To get an A, you should complete the baseline contract AND one of the following:

- Complete **both** requirements listed to receive an A-
- Receive and outstanding mark in 7 out of 8 labs

Your grade will go below a B+ if you miss work. Unless otherwise agreed upon in writing on a case-by-case basis, the following criteria outlines how deviating from the baseline contract will impact your grade (assuming everything else constant):

- Participation marked as unsatisfactory: B
- 2/8 labs marked as unsatisfactory or failed: B-
- 3/8 labs marked as unsatisfactory or failed: C
- 4/8 or more labs marked as unsatisfactory or failed: F
- Final paper marked as unsatisfactory or failed: F

By signing up for this course, you accept the terms of the grading contract. We will discuss potential amendments in Week 1. Amendments to the grading contract beyond this point should be agreed upon unanimously by all participants, including students and the instructional team.

Schedule

Week 1 (September 24/26): Preliminaries

Reading:

- AM Introduction
- King, Gary. 2006. "Publication, Publication." *PS: Political Science and Politics* 39 (1): 119-125
- King, Gary. "How to Write a Publishable Paper as a Class Project"
- Schwartz, Martin A. 2008. "The importance of stupidity in scientific research." *Journal of Cell Science* 121 (11): 1771

Lab 0: Project-oriented workflow in RStudio

Week 2 (October 1/3): Probability theory

Reading:

- AM Chapter 1
- Freedman, David A. ad Philip B. Stark. 2003. "What is the chance of an earthquake?" *NATO Science Series IV: Earth and Environmental Sciences*. 32: 201-213

Lab 1: Calculating and estimating probabilities in R

Week 3 (October 8/10): Summarizing distributions

Reading:

- AM Chapter 2
- Gelman, Andrew. 2023. "What is a standard error?" *Journal of Econometrics* 237 (1): 105516
- Wooldridge, Jeffrey M. 2023. "What is a standard error? (And how should we compute it?)" *Journal of Econometrics* 237 (1): 105517
- Powell, James L. 2023. "Discussion of 'What is a standard error?'" *Journal of Econometrics* 237 (1): 105518

Lab 2: Quantities of interest

Week 4 (October 15/17): Random samples

Reading:

- AM Chapter 3
- Mooney, Christopher. 1996. "Bootstrap Statistical Inference: Examples and Evaluations for Political Science." *American Journal of Political Science* 40 (2): 570-602
- Goldstein, Harvey and Michael J.R. Healy. 1995. "The Graphical Presentation of a Collection of Means." *Journal of the Royal Statistical Society* 158 (1): 175-177
- Knol, Miriam J., Wiebe R. Pestman, and Diederick E. Grobbee. 2011. "The (mis)use of overlap of confidence intervals to assess effect modification." *European Journal of Epidemiology* 26: 253-254

Lab 3: Estimation and uncertainty

Week 5 (October 22/24): Regression

Reading:

- AM Chapter 4
- Berk, Richard. 2010. "What You Can and Can't Properly do with Regression." *Journal of Quantitative Criminology* 26: 481-487
- Achen, Christopher H. 2005. "Let's Put Garbage-Can Regressions and Garbage-Can Probits Where They Belong." *Conflict Management and Peace Science* 22 (4): 327-339
- Hansen, Bruce E. 2022. "A Modern Gauss-Markov Theorem." *Econometrica* 90 (3): 1283-1294

Lab 4: Adjusting for covariates

Week 6 (October 29/31): Parametric models

Reading:

- AM Chapter 5
- Stigler, Stephen M. 2007. "The Epic Story of Maximum Likelihood." *Statistical Science* 22 (4): 598-620
- James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2021. *An Introduction to Statistical Learning with Applications in R*. Springer. Chapters 4.1-4.3
- Achen, Christopher. 2002. "Toward a New Political Methodology: Microfoundations and ART." *Annual Review of Political Science* 5: 423-450

Lab 5: Maximum likelihood estimation

Week 7 (November 5/7): Missing data

Reading:

- AM Chapter 6
- Lall, Ranjit. 2016. "How Multiple Imputation Makes a Difference." *Political Analysis* 24 (4): 414-433
- Arel-Bundock, Vincent and Krzysztof J. Pelc. 2018. "When Can Multiple Imputation Improve Regression Estimates?" *Political Analysis* 26 (2): 240-245
- Pepinsky, Thomas. 2018. "A Note on Listwise Deletion versus Multiple Imputation." *Political Analysis* 26 (4): 480-488

Lab 6: Dealing with missing data

Week 9 (November 12/14): Causal inference I

Reading:

- AM Chapter 7
- Fisher, R.A. 1935. *The design of experiments*. Oliver & Boyd. Chapters 1-2.
- Splawa-Neyman, Jerzy. 1990. "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9." *Statistical Science* 5 (4) 465 - 472

- Rubin, Donald B. 1990. "Comment: Neyman (1923) and Causal Inference in Experiments and Observational Studies.". *Statistical Science* 5 (4): 472-480
- Keele, Luke, Corrine McConnaughy, and Ismail White. 2012. "Strengthening the Experimenter's Toolbox: Statistical Estimation of Internal Validity." *American Journal of Political Science* 56 (2): 484-499

Lab 7: Randomization inference and hypothesis testing

Week 10 (November 19/21): Causal inference II

Reading:

Regression Discontinuity

- Cattaneo, Matias D., Nicolás Idrobo, and Rocío Titiunik. 2020. *A Practical Introduction to Regression Discontinuity Designs: Foundations*. Cambridge University Press. Chapters 1-4
- Cattaneo, Matias D., Nicolás Idrobo, and Rocío Titiunik. 2024. *A Practical Introduction to Regression Discontinuity Designs: Extensions*. Cambridge University Press. Chapter 2
- Stommes, Drew, P.M. Aronow, and Fredrik Sävje. 2023. "On the reliability of published findings using the regression discontinuity design in political science." *Research and Politics*

Difference-in-Differences

- Cunningham, Scott. 2021. *Causal Inference: The Mixtape*. Yale University Press. Chapter 9
- Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan. 2004. "How Much Should We Trust Differences-In-Differences Estimates?" *The Quarterly Journal of Economics* 119 (1): 249-275
- Roth, Jonathan, Pedro H.C. Sant'Anna, Alyssa Bilinski, and John Poe. 2023. "What's trending in difference-in-differences? A synthesis of the recent econometrics literature" *Journal of Econometrics* 235 (2): 2218-2244

Lab 8: Quasi-experiments

Week 11 (November 26): Flex week/future directions

We are leaving this week open to discuss topics that may be of interest to the group, I can think of five options:

- Bayesian statistics
- Statistical/machine learning for prediction
- Machine learning for statistical inference
- Generative AI/large language models
- Methods work by Northwestern faculty

Week 12 (December 3/5): Research note presentations

Northwestern University Syllabus Standards

This course follows the Northwestern University Syllabus Standards. Students are responsible for familiarizing themselves with this information.

Use of Generative AI Systems

The use of generative artificial intelligence in this course is encouraged as long as it is used to amplify humans instead of replacing them. Any form of cheating, including improper use of content generated by artificial intelligence, constitutes a violation of Northwestern's academic integrity policy.

Copilot is the University's supported artificial intelligence service. When using Copilot while actively logged in with a Northwestern account, data is stored securely in Northwestern's Microsoft tenant, and Microsoft will not use it for product improvement or to train its AI models.