POLI SCI 403: Probability and Statistics

Fall 2024

Instructor: Gustavo Diaz
Time and Place: Tues/Thurs 11:00am – 12:20pm, room TBD
Office Hours: TBD
GitHub Repository: TBD

Teaching Assistant: TBD

Course Overview

This is the first course in the required two-quarter 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 do so 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.

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's usually cheaper in storefronts like Amazon. TBD whether this will be available at the library.

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.

The final section of 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 lab assignments and the final research note. 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 you troubleshoot problems.

You can install R and RStudio on your personal computer, which is the preferred workflow. 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

be sufficient 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 allows, 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 dates TBD)
- Replication paper (due date TBD)

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 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 tricky puzzles that you may try to solve. In general, the goal of the lab assignments to show why statistical analyses are (or should not be done) in a certain way.

Our usual workflow will be to start working at the lab assignment during our class and TA section meetings 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 TBD in the week after they are assigned and they must be submitted in PDF format (you may use the original lab `.qmd` files as templates). Labs will be marked as *satisfactory*, *unsatisfactory*, or *fail* if you did not submitted or barely tried. I will mark late submissions as a fail unless I give you written approval submit later. However, you can resubmit any labs marked 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 trying to understand what went wrong and how things should look like if it had worked right.

The implication is that, to assign a satisfactory mark to your lab, I will be mostly looking at 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 previously published work, 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: https://gking.harvard.edu/papers

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 topics 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 have a solid idea of something you would prefer work on rather than a replication.

Final papers are due on TBD. Your should submit your paper in PDF format (I will show you how to write a PDF that integrates data analysis and writing in the R workshop). Your paper should follow the format of a research note, around 4,000 words. 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.

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

Grading

Schedule

Week 1 (September 24/26): Preliminaries

Reading:

Lab 0:

Week 2 (October 1/3): Probability theory

Reading:

Lab 1:

Week 3 (October 8/10): Summarizing distributions

Reading:

Lab 2:

Week 4 (October 15/17): Random samples

Reading:

Lab 3:

Week 5 (October 22/24): Regression

Reading:

Lab 4:

Week 6 (October 29/31): Parametric models

Reading:

Lab 5:

Week 7 (November 5/7): Missing data

Reading:

Lab 6:

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

Reading:

Lab 7:

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

Reading:

Lab 8:

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

Reading:

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

Academic Integrity

Accessibility

Generative AI

Resources

Statistics

Computing