

# Sociology 500: Applied Social Science Statistics

Brandon Stewart, Simone Zhang, Ian Lundberg\*

Class: 9:00am–10:30am Mondays and Wednesdays (Wallace Hall 165)

Precept: 9:00am–11:00am Thursdays (Wallace Hall 165).

## **Brandon Stewart**

bms4@princeton.edu, scholar.princeton.edu/bstewart

Phone: 609-258-5094

Office Hours: When the door is open (Wallace Hall 149)

## **Simone Zhang, Preceptor**

szz@princeton.edu

Office hours: Monday 1-3 (Wallace Hall 190)

## **Ian Lundberg, Preceptor**

ilundberg@princeton.edu

Office hours: Wednesday 6-8 (Wallace Hall 190)

## **1 The Basics**

### **1.1 Overview**

This course is the first of the two-semester sequence for Ph.D. students in Sociology. In this sequence, students will learn the statistical and computational principles necessary to perform modern, flexible, and creative analysis of quantitative social data. The overarching goal of the year is to move you from being consumers of quantitative research to producers of it. It will require a lot of hard work for all of us to achieve that goal; however, we have structured the class to provide you the maximal return on every hour of work you put in. As you read through this syllabus you will find numerous avenues for seeking help. If you are willing to put in the time- we are happy to help.

By the end of this semester, you will be able to:

- Critically read, interpret and replicate the quantitative content of many articles in the quantitative social sciences
- Conduct, interpret, and communicate results from analysis using multiple regression (including dummy variables and interactions).

---

\*Last Edited: September 15, 2016

- Explain the limitations of observational data for making causal claims, and begin to use existing strategies for attempting to make causal claims from observational data.
- Write clean, reusable, and reliable R code.
- Feel empowered working with data

More broadly by the end of the year, you will be able to:

- Conduct, interpret, and communicate results from analysis using generalized linear models.
- Understand the fundamental ideas of missing data, modern causal inference and hierarchical models.
- Teach yourself new statistical methods by reading the literature.
- Build a solid, reproducible research pipeline to go from raw data to final paper.

In terms of statistical content Soc 500 covers basic probability, univariate inference, linear regression and its applications. Soc 504 will cover maximum likelihood estimation, generalized linear models and the foundations of three special topics: missing data, causal inference and hierarchical models.

Upon finishing the course sequence, students should be able to read an original scholarly article describing a new statistical technique, implement it in computer code, estimate the model with relevant data, understand and interpret the results, and explain the results to someone unfamiliar with statistics.

As an important part of the sequence, students learn how to make novel contribution to the scholarly literature. We will do this through a replication project in the second semester of the sequence.

## 1.2 Class and Precept

Formal instruction for the course is split into two pieces: class and precept/lab. The course meets two times a week and will cover the core statistical material. The precept meets once per week and will focus on practical computational skills. Both are an essential part of the learning process.

## 1.3 Prerequisites

The most important prerequisite is a willingness to work hard on possibly unfamiliar material. Statistical methods is like a language and it will take time and dedication to master its vocabulary, its grammar, and its idioms. However like studying languages, statistics yields to daily practice and consistent effort.

Formally, the prerequisites vary for different types of students. For graduate students in the Sociology Department there are no course prerequisites. For anyone else, send Brandon an email. It is helpful but not essential if at some point you have taken calculus and have been exposed to basic matrix operations. We recommend that you are familiar with the material covered in the pre-semester methods camp.

## 1.4 Online material

We have a Blackboard site where we post detailed lecture notes, precept notes, a PDF version of this document, assignments, and other materials.

# 2 Materials

## 2.1 Computational Tools

The best way, and often the only way, to learn new statistical procedures is by doing. We will therefore make extensive use of a flexible (open-source and free) statistical software program called R as well as a number of companion packages. R is probably the most widely used statistical software. We recommend using R with RStudio. You will learn how to program in this class, if you do not know already.

## 2.2 Books

### Required

- Fox, John. 2016 *Applied Regression Analysis and Generalized Linear Models. 3rd Edition.* (We will use this in the second semester as well)
- Angrist, Joshua D. and Jörn-Steffen Pischke. 2008. *Mostly Harmless Econometrics: An Empiricist's Companion.* Princeton University Press.

**Special Access** We have generously been granted access to two manuscripts which are under contract. These are being provided to us as a favor by the author. DO NOT distribute, share, or post in a public way.

- Imai, Kosuke. 2017. *A First Course in Quantitative Social Science.*
- Aronow, Peter and Benjamin Miller. 2017. *Theory of Agnostic Statistics*

### Assigned But Free

- Matloff, Norman. 2011. *The Art of R Programming: A Tour of Statistical Software.* No Starch Press. (Available free through the library).
- Blitzstein and Hwang. 2014. *Introduction to Probability*
- Hernán, Miguel A. and James M. Robins. 2012. *Causal Inference.* Forthcoming, Cambridge University Press. (Note that this book is still being written and you can find draft PDFs on the linked page above.)
- A variety of papers, book components will be assigned as well, available on the web.

**Suggested** It is often helpful to see the same material in alternative ways. Thus here are some other texts you might consult.

- Berksekas, Dimitri P. and John N. Tsitsiklis, *Introduction to Probability*. Athena Scientific. (Also available as lecture notes online.)
- Freedman, David, Pisani, Robert, and Purves, Roger. 2007. *Statistics*. W.W. Norton & Company. 4th edition.
- Gelman, Andrew and Hill, Jennifer. 2007. *Data Analysis Using Regression and Multi-level/Hierarchical Models*. Cambridge University Press.
- Imbens and Rubin. 2016. *Causal Inference for Statistics, Social and Biomedical Sciences*.
- Morgan, Stephen L, and Christopher Winship. 2014. *Counterfactuals and Causal Inference: Second Edition*. Cambridge University Press. (although the first edition is also good).
- Ashenfelter, Orley, Levine, Philip, and Zimmerman, David. 2003. *Statistics and Econometrics: Methods and Applications*. John Wiley & Sons.
- Wooldridge, Jeffrey M. *Introductory Econometrics*. New York: South-Western. 5th edition. (earlier editions are fine)
- Wickham, Hadley. *Advanced R* (available free online)

### 3 Assignments

There are three main types of assignments:

1. **Preparing for class and precept:** For many classes and some precepts there will be some reading that you must do before class. I expect you to come 100% prepared. I will not assign an unreasonable amount of reading and thus I won't spend valuable class time summarizing readings that you should have done before class.
2. **Weekly problem sets:** Learning data analysis takes practice. The problem sets are described below.
3. **Final exam:** A cumulative take-home final exam will conclude the semester.

#### 3.1 Readings

There are readings for each topic and they generally cover the mathematical underpinnings. Reading statistical work can be challenging the first time you do it. There will be a temptation to skip over all the math - don't! The math is often where the action is in statistical work. Read carefully and go line by line making sure you understand.

To help with this process we will be using a tool called *Perusall* for many of the assigned readings. Perusall is a new ebook platform with collaborative annotation that allows you to post and answer questions directly in the text itself. This gives us the opportunity to answer questions outside of class in the text itself. So asking good questions not only helps you, it helps your

classmates. If you know the answer to a question that another student posted, please make a contribution to the class and try to answer it!

Obviously, read the required readings and any others that pique your curiosity. In addition, though, engage with the readings: take notes, write down your impressions or confusions, talk with your classmates, and post questions on Perusall or Piazza. All of your classes should be pushing your research forward and you will be more creative the more you actively read.

### 3.2 Problem Sets

Statistical methods are tools and it isn't very instructive to read a lot about hammers or watch someone else wield a hammer. You need to get your hands on a hammer or two. Thus, in this course, you will have homework on a weekly basis. The assignments will be a mix of analytic problems, computer simulations, and data analysis.

Assignments should be completed in R Markdown which allows you to show both your answers and the code you used to arrive at them. Don't worry if you don't know R Markdown, we will show you how it works. Your wonderful preceptors will provide you with more detailed instructions before the first assignment is due.

Each week's homework will be made available on Blackboard starting Wednesday at noon and is due Thursday the following week (8 days later) at the start of precept. Solutions will also be available directly after precept through Blackboard. The problem sets including looking at the solutions key is an extremely important part of the learning process, so please keep up with the work!

Problem sets will be graded from 0-5 pts. We will also have a bonus point for aesthetics including presentable graphs, clear code, nice formatting and well written answers.

You can have **one** no questions asked extension of one week on a problem set of your choosing. If you don't take an extension, we will drop your lowest grade (of any partially completed problem set). If all your problem sets are completed and with top-level grades (such that dropping the lowest wouldn't help you), we will add a comparable grade bonus to your final exam. When submitting the work on which you claim the extension please include a note indicating the original date and that you are claiming your one extension; you do not need to explain why you are taking the extension.

**Code Conventions:** Throughout the course, students will receive feedback on their code from the professor, the preceptor, and other students. Therefore, consistent code conventions are critical. We encourage you to use the Hadley Wickham's R Style Guide but any consistent style is fine. Good coding style is an important way to increase the readability of your code (even by a future you!). We will explain how to automatically check your code style using RStudio and a package called `lintr`.

**Collaboration Policy:** Unless otherwise stated, we encourage students to work together on the assignments, but you should write your own solutions (this includes code). That is, no copy-and-paste from other people's code. You would not copy-and-paste from someone else's paper, and you should treat code the same way. However, we strongly suggest that you make a solo effort at all the problems before consulting others.

There will be up to two problem sets which will not allow collaboration (sort of like take-home midterms). The final will be a more formal take-home exam on which again, there is no collaboration.

### 3.3 Help

We know that statistics can be challenging and help is available when you need it. We have made every effort to give you the tools you need to succeed in this course. Ultimately though it is your responsibility to put in the effort and seek out that help.

First, the readings provide ample sources of information and the suggested reading list contains many versions of the same material but presented from a different angle. Precept material and lecture slides will all be posted on Blackboard and can then be referenced.

For questions about the material and problem sets we will be using Piazza. You will not be required to post, but the system is designed to get you help quickly and efficiently from classmates, the preceptors, and the professor. **Unless the question is of a personal nature or completely specific to you, you should not e-mail teaching staff**; instead, you should post your questions on Piazza. The course staff will be monitoring the page, but we encourage you to help your classmates as well. I will post the link to the course page here at the start of class

The preceptors will each be holding 2 hours of office hours each week. Brandon will explain his office hour policy in class.

### 3.4 Grading

Final grades will be a weighted average of the final exam (30%) and the weekly problem sets (70%). We reserve the right to provide some bonus credit for active *participation* inside and out of class. For example a student who actively assists their classmates on Piazza or Perusall by answering questions or who engages productively in class might be entitled to a small bonus.

## 4 Course Outline

The following is a preliminary schedule of course topics. We may adjust the schedule due to comprehension, time, and interest. Please note also that readings are subject to change.

### Week 1: Introduction and Probability - Sept 14

- Course Details, Outline and Requirements
- Probability Basics
- Sample Spaces, Events, Law of Total Probability, Bayes Rule

Reading

- Aronow and Miller Chapter 1.1 (random events)
- Optional: Imai Chapters 6.1-6.2

### Week 2: Random Variables - Sept 19, 21

- Random Variables
- Marginal, joint, and conditional distributions

- Expectations, Conditional Expectations
- Covariance, correlation, and independence

Reading

- Aronow and Miller Chapters 1.2-2.3 (probability theory, summarizing distributions)
- Optional: Blitzstein and Hwang Chapters 1-1.3 (probability), 2-2.5 (conditional probability), 3-3.2 (random variables), 4-4.2 (expectation), 4.4-4.6 (indicator rv, LOTUS, variance), 7.0-7.3 (joint distributions)
- Optional: Imai Chapter 6 (probability)

**Week 3: Learning from Random Samples - Sept 26, 28**

- Populations, samples, estimation
- Point estimation
- Properties of estimators
- Interval Estimation

Reading

- Aronow and Miller Chapter 3.1-3.1.5 (estimation), 3.2.1 (confidence intervals)
- Fox Chapter 3: Examining Data

**Week 4: Testing and Regression - Oct 3, 5**

- Hypothesis testing
- Nonparametric regression
- Parametric models and linear regression
- Bias-variance tradeoff
- Regression as a predictive model

Reading

- Aronow and Miller Chapter 3.2.2 (testing)
- Fox Chapter 2: What Is Regression Analysis?
- Fox Chapter 5.1 Simple Regression
- Aronow and Miller 4.1.1 (bivariate regression)
- Momentous Sprint at the 2156 Olympics, by Andrew J. Tatem, Carlos A. Guerra, Peter M. Atkinson, and Simon I. Hay, Nature 2004.
- Optional: Imai Ch 2

## **Week 5: Simple Linear Regression in Scalar Form - Oct. 10, 12**

- Mechanics of Ordinary Least Squares
- Assumptions of the linear model
- Properties of least squares
- Inference with regression

### Reading

- Fox Chapter 6.1 Simple Regression
- Optional: Imai 4.2

## **Week 6: Linear Regression with Two Regressors - Oct. 17, 19**

- Mechanics of regression with two regressors
- Simpson's Paradox
- Omitted variables and multicollinearity
- Dummy variables, interactions, and polynomials

### Reading

- Fox Chapter 5.2.1. (Least Squares with Two Variables)
- Fox Chapter 7.1-7.3 (Dummy-variable Regression, Interactions)

## **Week 7: Multiple Linear Regression- Oct. 24, 26**

- Matrix algebra and mechanics of multiple linear regression
- Inference in a multiple linear regression model
- Concerns about p-values and multiple testing
- Bootstrap

### Reading

- Fox Chapter 9.1-9.4 skip 9.1.1-9.1.2 (Linear Models in Matrix Form)
- Aronow and Miller 4.1.2-4.1.4 (Regression with Matrix Algebra)
- Optional Fox Chapter 10 (Geometry of Regression)
- Optional: Imai Chapter 4.3-4.3.3
- Optional: Angrist and Pischke Chapter 3.1 (Regression Fundamentals)

- Optional: Nunzo, R. (2014) Scientific method: Statistical errors *Nature*.
- Optional: Leek and Peng (2015) Statistics: P values are just the tip of the iceberg *Nature*
- Optional: Cohen, J. (1994). The earth is round ( $p < .05$ ) *American Psychologist*.
- Optional: Simmons, J. et al. (2014) False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science*.

## Fall Break

### Week 8: Regression in Social Science- Nov 7, 9

- Presenting Results and Making a Case
- Frameworks for Causal Inference (Potential Outcomes and Causal Graphs)
- Visualization and Quantities of Interest

#### Reading

- Healy and Moody (2014). Data Visualization in Sociology. *Annual Review of Sociology*
- Morgan and Winship (2015) Chapter 1: Causality and Empirical Research in the Social Sciences
- Morgan and Winship (2015) Chapter 13.1: Objections to Adoption of the Counterfactual Approach
- Optional: Aronow, Peter M. and Cyrus Samii. 2016. “Does regression produce representative estimates of causal effects?” *American Journal of Political Science*.
- Optional: Morgan and Winship (2015) Chapter 2 (Potential Outcomes), Chapter 3 (Causal Graphs)
- Optional: Hernan and Robins (2016) Chapter 1: A definition of a causal effect

### Week 9: What Can Go Wrong and How To Fix It- Nov 14, 16, 21

- Diagnostics with Residuals
- Unusual and Influential Data → Robust Estimation (Day 1)
- Nonlinearity → Generalized Additive Models (Day 2)
- Unusual Errors → Sandwich Standard Errors (Day 3)

#### Reading

- Fox Chapters 11-13

- Optional: Fox Chapter 19 Robust Regression
- Optional: King and Roberts “How Robust Standard Errors Expose Methodological Problems They Do Not Fix, and What to Do About It.” *Political Analysis*, 2, 23: 159179.
- Optional: Aronow and Miller Chapters 4.2-4.4 (Inference, Clustering, Nonlinearity)
- Optional: Angrist and Pishke Chapter 8 (Nonstandard Standard Error Issues)

## Thanksgiving

### Week 10: Causality With Measured Confounding- Nov 28, 30

- Review of Causal Framework
- The Experimental ideal
- Graphical Models of Causal Effects
- The Assumption of No Unmeasured Confounding
- Choosing Conditioning Variables
- Colliders and Back-Door Criterion

#### Reading

- Angrist and Pishke Chapter 2 (The Experimental Ideal) Chapter 3.2 (Regression and Causality)
- Selection from Morgan and Winship Chapters 3-4 TBD (Causal Graphs and Conditioning Estimators)
- Optional: Elwert and Winship (2014) “Endogenous selection bias: The problem of conditioning on a collider variable” *Annual Review of Sociology*
- Optional: Morgan and Winship Chapter 11 Repeated Observations and the Estimation of Causal Effects

### Week 11: Unmeasured Confounding and Instrumental Variables- Dec 5, 7

- The Assumption of No Unmeasured Confounding
- Classical Approach to Instrumental Variables
- Modern Approach to Instrumental Variables

#### Reading

- Fox Chapter 9.8 Instrumental Variables and TSLS
- Angrist and Pishke Chapter 4 Instrumental Variables

- Morgan and Winship Chapter 9 Instrumental Variable Estimators of Causal Effects
- Optional: Hernan and Robins Chapter 16 Instrumental Variable Estimation
- Optional: Sovey, Allison J. and Green, Donald P. 2011. “Instrumental Variables Estimation in Political Science: A Readers’ Guide.” *American Journal of Political Science*

## **Week 12: Repeated Observations and Panel Data- Dec 12, 14**

- Fixed effects
- Panel Data and Causal Inference
- Difference-in-Differences
- Regression Discontinuity

### *Reading*

- Angrist and Pishke Chapter 5 Parallel Worlds: Fixed Effects, Differences-in-Differences and Panel Data
- Optional: Imai and Kim “When Should We Use Linear Fixed Effects Regression Models for Causal Inference with Longitudinal Data”
- Optional: Angrist and Pishke Chapter 6 Regression Discontinuity Designs
- Optional: Morgan and Winship Chapter 11 Repeated Observations and the Estimation of Causal Effects

## **5 Inspirations**

The development of this course has been influenced by a number of people particularly: Matt Blackwell, Dalton Conley, Adam Glynn, Justin Grimmer, Jens Hainmueller, Chad Hazlett, Gary King, Kosuke Imai, Kevin Quinn and Matt Salganik. Teppei Yamamoto. Thanks to all of these excellent teachers for sharing their ideas, slides and/or syllabi with me.