Syllabus

SOC3050 Statistics for Sociology

Fall 2019

Instructor: John Kuk

Time & Room
Class: M, W 2:30–4:00PM
Room: Seigel 306
Course website: Piazza

TA: Brian Tung (brian.tung@wustl.edu)
Office Hours: TBA

Office
John Kuk
Room: Seigel 181
Email: jskuk@wustl.edu
Office Hours: Tuesday 2:00-4:00 PM

Course Description

How can we measure racial discrimination in job hiring? What is the best way to predict election outcomes? Would universal health insurance improve the health of the poor? Do patterns of arrests in US cities show evidence of racial profiling? These are just a few of the numerous question that social scientists are tackling with quantitative data. Beyond academia, companies and non-profits have invested heavily in data science techniques to learn about their users, platforms, and programs. Data scientists at these institutions are essentially applied social scientists and employ many of the same techniques you will learn in this course.

What will you learn in this course? Our goal is to give you the ability to understand, explain, and perform social science research, with a special focus on data analysis and causal reasoning. You will be able to read and understand the methodology of most academic articles in the social sciences, but more importantly you will have a foot in the door of the data science world. This course will also introduce and familiarize students to the use of computing software for data analysis, primarily the widely-used statistical package `R`. The ability to collect and analyze data in a sophisticated manner is becoming a crucial skill set for the modern job market across industries. Finally, you will obtain data literacy that will help you be a critical consumer of evidence for the rest of your life.

This course is designed primarily for undergraduate majors and minors in sociology, but students from other academic disciplines are welcome. No prior knowledge of statistics is assumed or required.

Prerequisites

A willingness to work hard on possibly unfamiliar material is key.

Course Objectives

1. Read and process data in multiple formats and conduct basic descriptive analysis.
2. Create effective visual depictions of statistical patterns in data using visualization techniques.
3. Understand the assumptions and limitations of statistical procedures.
4. Learn the terminology and techniques of basic statistical analysis.
5. Gain an understanding of how statistics can be used to address research questions in the social sciences.
6. Become knowledgeable and critical consumers of statistical information encountered in everyday life.

Evaluation

- **Datacamp computer assignments (10%)**
  To help with learning the programming skills necessary to analyze data, I will be giving assignments in an online interactive learning system called Datacamp (www.datacamp.com). Learning to program is a lot like learning a language, it is best to have consistent, daily practice. The exercises will be interactive and guided lessons through computing topics covered that week, and they will be graded based on completion.
- **Coding challenge (10%)**
  At the end of each class, the instructor will share a short coding challenge based on the content of that class during the weeks without Datacamp assignments
- **Five homework assignments (25%)**
  You need to show that you made a good faith effort to work each question. The problem sets will be graded using a check system:

  ✓+: Problem set is 100% completed. Every question was attempted and answered, and most answers are correct. Document is clean and easy to follow. Work is exceptional. I will not assign these often.

  ✓: Problem set is 70–99% complete and most answers are correct. This is the expected level of performance.

  ✓- : Problem set is less than 70% complete and/or most answers are incorrect. This indicates that you need to improve next time. I will hopefully not assign these often.
- **In-class midterm exam (20%)**
- **Take-home final exam (25%)**
  The final exam can be substituted with a final project
- **Participation (10%)**
  Students should actively participate in all aspects of the course. Class participation will be judged based on questions asked/answered during the lectures, the precepts, and on `Piazza`.

Learning and Programming

Learning: The course will follow a "learning-by-doing" approach and will place emphasis on gaining experience in analyzing data. Students are expected to do the required readings for each week and run the code before each session. The lectures will build upon the content of the readings with a series of short in-class assignments that will introduce new statistical and programming concepts, which will then be applied to the analysis of data from published research papers or common tasks in data science. Most of the applications will be related to social science questions.

Programming Exercises: Throughout the semester we will learn the statistical programming language `R`. We will use the open-source statistical software environment *R Studio*, which makes it much easier and more intuitive to work with data using `R`. There is a steep learning curve with `R`, and you will discover that learning to program is fun and exciting, but it can also be frustrating at times. To facilitate learning of `R`, we will be using DataCamp. DataCamp will enable you to work through the programming exercises at your own pace, while accessing accessing various types of support, both within DataCamp and the broader class community. The system will teach you all you need to know to use `R` for your own analyses, and you will have access to several supplementary courses that you can use to extend your knowledge beyond what is covered in the course. With the resources and exercises provided by DataCamp, activities and instruction during class and experience working through the problem sets, we are confident that all of you will learn the language of `R` during the semester, but students should expect to spend additional time learning and practicing.

R and RStudio: We'll use `R` in this class to conduct data analysis. `R` is free, open source, and available on all major platforms. `RStudio` (also free) is a graphical interface to `R` that is widely used to work with the `R` language. You can install `R` and `RStudio` on your own computer (and you should!), but we will also provide you with an account on a cloud-based version of `RStudio` that you can run through your web browser (Chrome, Safari, Internet Explorer, Firefox, etc). This will minimize some of the pain of setting up RStudio and allow us to easily provide you with script templates for homework assignments. You can find a virtually endless set of resources for R on the internet. For beginners, there are several web-based tutorials including one from DataCamp. In these, you will be able to learn the basic syntax of `R`.

Logistics

Emails and Piazza: Email should be used for personal issues, such as to schedule an appointment outside of office hours, to request an excused absence, or for feedback about grades. I I will respond to all course-related e-mails within 24 hours. All other questions should be asked on *Piazza*. If you have a question about course content assignments, or logistics, please check *Piazza* first to see whether it has been asked already. If you email us with a question that is relevant to other members of the class, we will respond by directing you to post your comment to *Piazza*. Note that you can ask questions anonymously on *Piazza*. *Piazza* is designed so that students can answer each other's questions. I encourage you to use this feature. *Piazza* will be checked at least every 48 hours, and more frequently around the end of the term when papers are due. The link to *Piazza* will be announced later.

Collaboration Policy: We encourage students to work together on the homework assignments,

but you must write your own solutions (this includes computer code), and you must write the names of your collaborators on your assignment. I also **strongly suggest** that you make a solo effort at all the problems before consulting others. The exams will be very difficult if you have no experience working on your own. There is no collaboration allowed on the exams.

Late Submissions: For the class to work, all students must keep up with the course load throughout the semester. To encourage students to keep up with the material and to allow the instructors to provide timely feedback, assignments must be turned in on time. All assignments turned in after the deadline will be docked 20%, and an additional 10% for every 24 hours the assignment is delayed. Any submission one week after the deadline will not be accepted.

Reading and Textbooks

- Imai, Kosuke, *Quantitative Social Science*
- Diez, David M., Christopher D. Barr, and Mine Çetinkaya-Rundel. 2015. *Open-Intro Statistics*. 3rd edition. https://www.openintro.org/

Quantitative Social Science will be our primary textbook. I will occasionally assign readings from the free *Open-Intro Statistics* book.

Course Schedule

Date	Topic	Reading
8/26/2019	Introduction	
8/28/2019	No class (instructor conference attendance)	
9/2/2019	Labor day	
9/4/2019	Intro to `R` and RStudio	QSS Ch. 1
9/9/2019	Measurement I: Descriptive Statistics	QSS Ch. 2.6, 3.1 – 3.2
9/11/2019	Measurement I: Visualization	QSS Ch. 3.3
9/16/2019	Data manipulation	QSS Ch. 2.1 – 2.2
9/18/2019	Data manipulation	
9/23/2019	Causality: Introduction	QSS Ch. 2.3 – 2.7
9/25/2019	Causality: Randomized Control Trial	
9/30/2019	Measurement II: Relating Variables to Each Other	QSS Ch. 3.4 and 4.1
10/2/2019	From Bivariate Analysis to Prediction	
10/7/2019	Prediction: Regression I	QSS Ch. 4.2 –4.3
10/9/2019	Prediction: Regression II	
10/14/2019	Fall break	
10/16/2019	Midterm Exam	
10/21/2019	Prediction: Regression III	QSS Ch. 4.3
10/23/2019	Probability: Introduction	QSS Ch. 6.1 – 6.2
10/28/2019	Probability: Random Variable	QSS Ch. 6.3
10/30/2019	Uncertainty: Standard Error and Confidence Intervals	QSS Ch.7
11/4/2019	Uncertainty: Hypothesis Testing I	
11/6/2019	Uncertainty: Hypothesis Testing II	
11/11/2019	Regression and Uncertainty	
11/13/2019	Regression and Uncertainty	
11/18/2019	Putting it All Together	
11/20/2019	An Intro to Text Data and Network Data	QSS Ch. 5
11/25/2019	An Intro to Text Data and Network Data	
11/27/2019	No Class	
12/2/2019	Review I	
12/4/2019	Review II	

Acknowledgment

I learned tremulously from my teachers and friends. This course is built upon, or borrows from, course materials prepared Molly Roberts (UCSD POLI170A), Pablo Barberá (USC IR312), Matthew Blackwell (Harvard GOV50), and Andrew Heiss (BYU MPA630).

August, 2019