COMM 290: Intro to Data Analysis for Communication (Fall 2020)

Tuesdays and Thursdays, 15:00-16:30
Virtual Classroom: TBD

Jin Woo Kim
jin.woo.kim@asc.upenn.edu
Office Hours: By Appointment
Syllabus Updated: 8/3/2020

Course Overview

Data science is becoming an indispensable part of understanding communication behavior and effects. In this course, we will learn the basic tools of data analysis and the R programming language. We apply them to answer various questions in communication science. Can reluctant parents be convinced to vaccinate their children? Can get-out-the-vote mailings mobilize voters? Does the diffusion of political rumors affect public opinion? Are toxic comments more likely to go viral on Facebook? These are examples of the questions that we will answer using pre-existing datasets as well as a new online experiment that we will run together as part of the course.

There is no prerequisite for this class and students are not expected to have any familiarity with statistical programming. Students will be given step-by-step instructions and we will work together to analyze the datasets. For the final project, each student will write a research note based on their analysis of the new experimental data. At the end of this course, students will be able to use quantitative data to extract statistical patterns and answer empirical questions. These skills will be extremely useful in various settings, from academia to the media and tech industry and more.

Optional: Interested students are welcome to continue to work on the experimental data after the class ends and co-author a paper that will be submitted to the next annual conference of the *International Communication Association*.

Learning Objectives

At the end of this course, students should be able to

- Use R to analyze data
- Use data to answer empirical questions about communication behavior and effects

- Extract and visualize statistical patterns in data
- Understand the methods sections of academic articles

Course Structure

This class will be fully remote with an emphasis on synchronous instruction during the published time slot for the course. Real-time attendance is recommended but not mandatory. Typically, we will begin each class by learning a statistical concept in lecture. Then we will write R code together to apply the concept in practice, often drawing on actual datasets. Our goal is to go through as many in-class exercises as possible, so students can feel comfortable implementing the same concept in other contexts. I will ask you to submit the code you wrote at the end of each class, which will be due before the next class. I will use them to check if everyone is following the material and provide feedback if necessary. In-class exercise codes will *not* be graded but will count toward your participation score. Video of each class will be uploaded on Canvas. We will make our best efforts to address all of your coding-related questions during class. Contact me troubleshooting session in case you fail to solve coding errors.

Computing

We will use R in this class. R is a **free** open-source programming environment. It is a powerful and flexible tool that has become extremely popular in the data science world. It is widely used in both academia and industry. RStudio is a graphical interface to R. **You should bring your laptop to class with both R and RStudio installed.**

> To install R (v4.0.0), visit https://cloud.r-project.org/
> To install RStudio (v1.2.5042), visit https://rstudio.com/products/rstudio/download/

Learning how to use R may seem challenging for those who don't have previous experiences with statistical methods and computing. I will provide handouts with step-by-step instructions for the data analyses that we carry out in each class. They will be cover most of what you will need to know to be able to complete the homework assignments. But if you want to delve deeper into R, try:

> Getting Started with R
> R for Data Science

Required Textbook

Imai, Kosuke. *Quantitative Social Science: An Introduction*. Princeton University Press

Course Website

TBD

Evaluation

There will be no exam. Grades will be calculated based on:

- Homework (50%)
- Final Project (30%)
- Participation (20%)

Homeworks

There will be five official homeworks throughout the semester. In general, they will ask you to use the tools covered in class to analyze other datasets. You may work with other students on the homeworks, but **you should write your own code**. In case you choose to collaborate, you should write the names of the students with whom you collaborated. Each homework will be distributed two weeks before it is due. See the schedule below for specific dates. There will be smaller impromptu homeworks (e.g., taking a short survey experiment) as well. These minor assignments won't be graded. But failure to turn them in will lower your participation grade.

Final Project

You will write a short research note based on the analysis of the new online experiment that we will run together. The note will consist of (1) an abstract, (2) a very brief introduction that presents the research question, (3) a hypothesis section, (4) a method section detailing how the experiment was conducted and how the data was analyzed, (5) a result section presenting your findings, and (6) a very brief conclusion. Final paper rubric is attached at the end of this syllabus. More details will be announced later in the semester.

Schedule (Subject to Change)

Below is a tentative schedule of this course. We will adjust the pace, if necessary. Check the Canvas site for updates.

	Date	Topic	Reading (Imai)	Assignments
Class 1	9/1	Course Introduction		
Class 2	9/3	Intro to R & RStudio 1	1.3	HW1 Out
Class 3	9/8	Intro to R & RStudio 2	1.3	
Class 4	9/10	Causality 1	2.1-2.2	
Class 5	9/15	Causality 2	2.3-2.4	
Class 6	9/17	Measurement 1	2.6	HW1 Due
Class 7	9/22	Measurement 2	3.1-3.4.1	HW2 Out
Class 8	9/24	Measurement 3		
Class 9	9/29	Data Manipulation 1: Recoding (1)		
Class 10	10/1	Data Manipulation 2: Recoding (2)		
Class 11	10/6	Data Manipulation 3: Scaling (1)		HW2 Due
Class 12	10/8	Data Manipulation 4: Scaling (2)		HW3 Out
Class 13	10/13	Data Manipulation 5: Merging		
Class 14	10/15	Data Collection Exercise 1: Survey on Qualtrics		
Class 15	10/20	Prediction	4.1	
Class 16	10/22	Regression 1: Intro to Regression	4.2.1-4.2.3	HW3 Due
Class 17	10/27	Regression 2: Making Sense of Regression Outputs	4.2.4-4.6	HW4 Out
Class 18	10/29	Regression 3: Regression with Categorical Predictors	4.3.1	
Class 19	11/3	Regression 4: Regression with Binary Outcomes		Final Guideline
Class 20	11/5	Regression 5: Regression with Multiple Predictors	4.3.2	
Class 21	11/10	Regression 6: Heterogeneous Effects (1)	4.3.3	HW4 Due
Class 22	11/12	Regression 7: Heterogeneous Effects (2)	4.3.3	HW5 Out
Class 23	11/17	Regression 8: Creating Regression Tables		
Class 24	11/24	Data Visualization 1: Plotting Means		
Class 25	11/24	Data Visualization 2: Plotting Treatment Effects		
Class 26	12/1	Data Visualization 3: Plotting Heterogeneous Effects		HW5 Due
Class 27	12/3	Data Visualization 4: Plotting Trends		Final Update Due
Class 29	12/8	Final Paper Q&A		
Class 29	12/10	Course Review		
Final	12/15			Final Paper Due

Final Paper Rubric

Criteria	5	4	3	2 or lower
Abstract, Intro, and Conclusion	Clear, correct and concise description of the research; insightful discussion about this study's implications and limitations	Correct and relatively clear description of the research	Some errors or omission in description of the research	Vague or incorrect description of the research
Hypotheses	Identifies and precisely describes appropriate and *original* research hypotheses; provides strong theoretical motivation	Identifies appropriate research hypotheses	Hypotheses not described precisely; one hypothesis is not appropriate (e.g., not testable or not about treatment effects, etc.)	Hypotheses are stated vaguely or incorrectly; two or more hypotheses are not appropriate
Methods	Specifies important aspects of how study was conducted in very clear manner	Specifies important aspects of how study was conducted in relatively clear manner	Doesn't specify some important aspects of how study was conducted; methods not always well-explained	Does not provide or clearly explain most important aspects of how study was conducted
Statistical Analysis	Correct and precise statistical approach tailored for the hypotheses. Coding is very efficient.	Statistical approach is largely correct and error-free	Some errors in coding or analyses; one hypothesis is not properly tested	Numerous errors in coding or analyses; two or more hypotheses are not properly tested
Results	Figures and tables illustrate findings in an intuitive and easy-to-understand; text explains results precisely and without errors	Figures and tables illustrate findings reasonably clearly; textual explanations of results are clear;	Figures and tables unappealing or poorly constructed; some imprecision or errors in textual discussion of results	Figures and tables sloppy or hard to understand; text vague or incorrect; incomplete investigation of hypotheses
Writing	Exceptionally well-written—precise, clear, concise and elegant	Very well-written—clear and articulate	Moderately well-written—but not always clear	Unclear, awkward, or imprecise writing