Fall 2012 | Mon. & Wed., 11:00–11:50am | Classroom: Gavett Hall 310

PSC 200: Applied Data Analysis

Matthew Blackwell

Office: Harkness 307
Office Hours: W 1:00-3:00pm (or by appointment, or just drop by)
m.blackwell@rochester.edu
http://www.mattblackwell.org

TA: Hyesung Kim
Office: Harkness 305
Office Hours: M 1:00–2:00pm, T 1:00-2:00pm

General Information

This course is about making arguments with numbers and data. Data analysis for its own sake is often quite boring, but becomes crucial when it supports claims about the social world. Thus, this course will teach you how to analyze data to investigate social science hypotheses and, further, how to incorporate those skills into a convincing argument. The goal will be to convey your data-backed arguments to any audience, regardless of their statistical knowledge. This skill is rapidly becoming vital to many fields—social science, public policy, and business.

This class involves a little math, but this is **not a math class**. I will assume zero mathematical background beyond high-school algebra and zero statistical computing experience. The philosophy of this course is the best way to learn data analysis is to actually analyze data. We will be learning largely through applications and we will see datasets at every turn—lecture, computer lab, and assignments. Remember, while we will be learning formulas and computer functions and other technical material, these are just tools to help us better understand the data. They are a poor replacement for our brains and our own reasoning is a crucial component to any data analysis.

Who should and should not take this class?

This course assumes no prior statistical or mathematical experience beyond high-school algebra. In principle, *anyone* can be successful in this class. While this is true, the course will require a good amount of work and dedication to learning the craft of data analysis. Many, many people before you (your humble instructor included) have found themselves lost when trying to learn statistics and data analysis. This feeling is completely normal and there will be many opportunities for you to get help from us. The key to remember is that you *can* do it, but it might take some extra work to get there. If you have taken a statistics class before, you may find the class to be on the slow side.

Books

The following texts are **required** for this course:

- Alan Agresti and Barbara Finlay. 2009. *Statistical Methods for the Social Sciences. Fourth Edition.* Upper Saddle River, New Jersey: Pearson Prentice Hall.

  This will be the main textbook for the course. Note that this is fourth edition, but the third edition is widely available used for considerably cheaper. Either edition is acceptable for this class.

- Larry Gonick and Woollcott Smith. 1993. *The Cartoon Guide to Statistics.* Harper-Perennial (Amazon). Does what it says on the can. Good intuitive explanations of some of the key concepts in the course.

- John Verzani, SimpleR: Using R for Introductory Statistics. This is a free ebook about R, which we will use for computation.

Note that we may circulate additional (mostly optional) readings during the term.

Computing

Many data analysis problems require computation and we will be using a free statistical software package called R and a frontend to that package called RStudio. Using a free package allows you to work on your own computers as opposed to being shackled to the labs. You should attend all classes and recitations to learn how to use R for each assignment and budget time to trial and error as you work. Over the course of the term, we will also produce notes that will help you complete specific tasks in R. This class, though, is not a test of your R ability and you should always feel free to ask the professor or teaching assistant for help.

Grading

- **30% Problem sets** - To start learning about data analysis, we will first rely on problem sets, but over time we will transition to mostly data essays (below).
- **30% Data Essays** - These are short (two to four page) essays that use the statistical techniques we have learned to answer problems in political science using real political science datasets drawn from a range of topics.
- **20% Midterm Exam** - Will take place on October 24th, with a review session during the lecture on October 22nd.
- **20% Final Exam** - The final will not directly cover the first half of the course, but concepts in the first half of the course are needed to use techniques in the second half of the course. It will be on December 20th at 7:15 PM.

Late Policy

If you turn in an assignment late, you will receive a 10% deduction in the grade on that assignment for every day late. If you submit the homework more than three days after the due date, you will receive no credit.

Attendance

Each class meeting is important; you will have a hard time keeping up with the material if you miss lectures or recitations. There will be material covered in lecture that is not in the readings. And recitation and lecture will provide you the tools (mathematical and computational) necessary to complete your assignments. If you miss class, you should contact the instructors or your fellow students to get caught up.

Collaboration

Students may discuss homeworks and assignments in pairs or small groups, however, all work must be individually written and all results and figures individually generated. You may give each other advice or help point out coding errors, but in the end you must carry out the work yourself. Occasionally, a student will email their work to friends to show how they completed a problem. If, as sometimes happens, a friend simply copies text or graphs into his or her own paper, both students will be cited for academic honesty violations. Note that in cases of academic honesty, the instructor is required to report cases to the Board of Academic Honesty, without exception.

Missed Exams

The midterm and final exam dates are firm. Missed exams may only be re-taken under the following circumstances: (1) death in the family within two weeks before the exam, (2) participation in a University-sponsored academic or sporting event, (3) unforeseen medical emergency. In the case of (1) and (2), you must inform me within 24 hours of the exam that you will miss it. In some cases, I may require supporting documentation out of fairness to other students.

Schedule

We will meet for lectures in Gavett Hall 310 on Mondays and Wednesdays from 11:00–11:50 AM. The lab sessions will meet Fridays, 11:00–11:50 AM in Gavett 244. I will orient Monday sessions to lecture and Wednesday sessions to data, with a mix of lecture, demonstrations, and hands-on problem solving by students. The schedule is subject to change, but I will always notify you in class and by email of any changes and distribute an updated syllabus.

Week 1 (September 1–5)

- **September 3 (M):** Labor Day, no class
- **September 5 (W):** Introduction, syllabus review. What is R? Why are we using it?
  - HW #1 distributed.

Week 2 (September 10–15)

- **Describing the data we do have:** histograms and scatterplots, the "center" of the data, measuring spread.
  - Agresti and Finlay, 3.1-3.2
  - *Cartoon Guide* pp. 7–18
- **September 12 (W):** HW #1 due, HW #2 distributed.

Week 3 (September 17–21)

- **All's Normal:** distributions, Z-scores, Normal tables, samples, populations, Central Limit Theorem.
  - Agresti and Finlay, 4.1–4.3

  - *Cartoon Guide*, Ch. 3

- **September 19 (W):** HW #2 due, HW #3 distributed.

Week 4 (September 24–28)

- **Samples from known populations:** Repeated samples, standard errors, opinion polls.
  - Agresti and Finlay, 4.4–4.7
  - *Cartoon Guide*, Ch. 6
- **September 26 (W):** HW #3 due, HW #4 distributed.

Week 5 (October 1–5)

- **Learning about populations:** Inference from samples, confidence intervals, election forecasting.
  - Agresti and Finlay, Chapter 5
  - *Cartoon Guide* Ch. 7
- **October 3 (W):** HW #4 due, HW #5 distributed.

Week 6 (October 8–12)

- **Learning about populations (continued):** Hypothesis testing.
  - Agresti and Finlay, 6.1–6.5
  - *Cartoon Guide* Ch. 8
- **October 8 (M):** Fall break, no class.
- **October 10 (W):** HW #5 due, HW # 6 distributed.

Week 7 (October 15–19)

- **Comparing groups:** The population difference in means, binary variables, causal effects.
  - Agresti and Finlay, 7.1–7.4
  - *Cartoon Guide*, Chapter 9.

Week 8 (October 22–26)

- **October 22** (M): Review for Midterm Exam, HW #6 due.
- **October 24** (W)**:** Midterm Exam

Week 9 (October 29–31)

- **Relationships between variables:** correlation, scatterplots, bivariate regression, ordinary least squares.
  - Agresti and Finlay, 3.5, 9.1–9.4
  - *Cartoon Guide* Ch 11
- **October 31** (W)**:** Essay #1 distributed.

Week 10 (November 5–9)

- **Bivariate regression:** interpreting slopes, residuals, adding a binary variable.
  - Agresti and Finlay, 9.3–9.7, 13.1–13.2

Week 11 (November 12–16)

- **Holding other factors constant:** multiple regression, interpreting regression coefficients.
  - Agresti and Finlay Ch. 10, 11.1–11.4
  - Alan I Abramowitz. 2008. "Forecasting the 2008 Presidential Election with the Time-for-Change Model." *PS: Political Science & Politics* 41, no. 04 (October)
- **November 14** (W)**:** Essay #1 due, Essay #2 distributed.

Week 12 (November 19–23)

- **Research design:** Causal inference, confounders, mediators.
  - Gelman and Hill, Ch. 9
- **November 21** (W)**:** No class, happy Thanksgiving.

Week 13 (November 26–30)

- **Research design (continued):** Randomized experiments, observational studies.
- **November 28 (W):** Essay #2 due, Essay #3 distributed.

Week 14 (December 3–7)

- **How effects can vary:** Interaction effects, non-linear relationships between variables.
  - Agresti and Finlay, 13.3–13.4, 14.5–14.6

Week 15 (December 10–12)

- **How effects can vary (continued):** Interaction effects, non-linear relationships between variables.
- **December 12 (W):** Essay #3 due, Final Exam review session.