Political Science 230
Introduction to Statistics for Political Scientists

Professor Jake Bowers (jwbowers@illinois.edu)
TA Aya Kachi (aya.teaching@gmail.com)
Moodle: `https://courses.las.illinois.edu/course/view.php?id=524`

Spring 2011

General Information

This class is an introduction to applied statistics as practiced in political science. It is computationally intensive, and, as such, will enable students to write quantitative social science papers. That is, by the end of the course, a successful student will be able to find social science data online, download it, analyze it, and write about it in the same way that graduate students, professors, and public policy professionals do. The only difference between a student finishing this course and social science professionals will be *experience* with the substance, theory and methods of social science, not the tools used or the general application of statistical methods.

Location and Time

The whole class meets in 219 David Kinley Hall and Wednesdays from 11:00 AM to 11:50 AM.

Sections meet in 338 Davenport Hall on Friday mornings at 9, 10, and 11.

Moodle enrollment key is: Ilovestatistics!

Office Hours

Jake's office hours are Thurs 1-3pm by appointment in 231 Computing Applications Building (CAB) or other times by appointment. If you know in advance that you want to come to office hours, please email me to reserve a 20 minute slot. Please make an appointment if you want to come to office hours or if you would like to meet at times other than the office hours.

Aya's office hours are Tuesday 2:30-4:30pm in 158 English Building.
Note: Students should bring their own laptops (if they own laptops) to Aya's office hours if they have computing-related questions.

Goals and Expectations

More than anything we assume a **willingness to engage** with mathematics, data analysis, computer programming, and the practice of social science thinking and writing. We also assume you've taken at least one class in algebra at the level taught in most high schools in the United States and have used a personal computer to read and type email and other documents and have used Google online.

We also assume that you will have read the syllabus and that you keep up to date on changes in the syllabus (which will be announced in class). You should not expect a response to emails that ask a question which is answered in the syllabus.

In-Class Work

The class itself will involve work in groups at your computers nearly every class meeting. This is not a lecture class but an experiment in hands-on learning. At the beginning of each class, we will hand out worksheets with problems that will require you to use the R statistical computing language. The problems will be designed first to introduce you to the idea of scientific computing as practiced in the social sciences and then to the basics of social science data analysis and frequentist statistical inference. You will work on the worksheets during the class-time (in groups of about 3 people). We will collect one worksheet from each of you at the end of the class. We will grade one problem from each worksheet selected at random with fixed probability.

Here is how we might choose the problem to grade in R, assuming 4 problems but not knowing much about R:

```
> set.seed(1234567) ##Ensure that the random numbers I produce are the same on each run of the program.
> ##go to http://rseek.org, search for: how can I generate a list of numbers
> problem.numbers<-seq(1,4) ##make a list of problem numbers
> problem.numbers ##print the list just to make sure we got it right

[1] 1 2 3 4

> ##go to http://rseek.org, search for: how can I draw a random sample from a list of numbers
> sample(problem.numbers,size=1) ##choose one at random

[1] 3

> ##Notice that over the course of the term, assuming four problems per session, and about 25 sessions,
> ##we'll grade each problem about the same amount of the time but not exactly the 1/4 of the time.
> problems.graded<-replicate(25,sample(problem.numbers,size=1))
> table(problems.graded)

problems.graded
1 2 3 4
5 8 4 8

> table(problems.graded)/25 ##to convert the totals into proportions

problems.graded
   1    2    3    4
0.20 0.32 0.16 0.32

> ##However, if we had 10000 class sessions, the same procedure would allow us to grade each
> ##problem nearly exactly 1/4 of the time.
> problems.graded<-replicate(10000,sample(problem.numbers,size=1))
> table(problems.graded)

problems.graded
   1    2    3    4
2497 2608 2469 2426

> table(problems.graded)/10000

problems.graded
    1     2     3     4
0.250 0.261 0.247 0.243
```

Participation Quality participation does not mean "talking a lot." It includes attending section; thinking and caring about the material; and expressing your thoughts respectfully and succinctly and thoughtfully. Participation, in this class, will mostly refer to your active involvement in your sections, but the quality of your general involvement in lectures, emails, and office hours will also be taken into account.

Final Report Each of you will write a final paper on longer than about 5 pages (but we hope closer to 2 pages) due on May 13th by 5:00 PM. This paper is an opportunity for you to use the ideas from this class to pursue some data analysis on a topic that interests you.

We will have several assignments oriented around your paper to (1) give you practice with the techniques under discussion and (2) push your paper along so that the quality of papers turned in at the end is high.

We anticipate that the report will take the following form: On the first page you will name a dataset and three variables [one outcome, one cause or explanation, one control]. You will simulate how the outcome ought to depend on the cause "controlling for" the control variable. And you will present a graph to illustrate this simulation. You will then explain, in a few sentences, why you believe this relationship ought to exist the way that the simulation shows that it does.

One the second page, you will fit a linear regression model to the data and graph the results. You will interpret the regression table (explaining what p-values and confidence intervals mean as well as the substantive meaning of the regression coefficients). If the graph based on the dataset does not match up with your expectations, you'll speculate, in a few sentences, about why.

Grades We'll calculate your grade for the course this way:

50% In-Class Work and Attendance Participation is 70% in-class work grade and 30% attendance. We'll drop the lowest 3 of the daily worksheet grades. The worksheets will tend to be multiple choice but will require you to do open-ended data analysis to arrive at the correct answer. If you answer the randomly chosen problem correctly you will receive an A on that worksheet (100%). If you answer incorrectly you will receive a B on that worksheet (86.99%). If you do not answer the question chosen for grading, you will receive 0%. [Obviously, if you do not attend the that day, you will receive a zero for your worksheet grade]. Attendance will be a simple percentage of the number of class sessions you attended. In-class work happens in-class. It may not be turned in late or made-up at a later date without official excuses [for example, if you are hospitalized in the middle of the term, but the Dean thinks that you should not drop the course, we will work with you, your doctors and the relevant Dean to enable you to complete the course.]

40% Final Reports Grades on the final reports will be based on the clarity of your writing and thinking and the correctness of your data analysis. You may turn in reports late, but you will lose $\frac{1}{3}$letter grade for each day that you are late (e.g. an "A" assignment would become a "A-" assignment after 1 day, a "B+" assignment after two days, ..., a "C" assignment after 6 days). The Final proposals are a part of the Final Report, and, as such the Final Report Grade will be max(Final Report*.80+Final Proposal*.20,Final Report).

10% Participation The Professor and TA will consult with each other to assign a letter grade reflecting the quality of participation.

Incomplete Work Assignments not turned in will be counted as zero in the calculation of the final grade.

Books

Required: Kaplan, D. (2009). *Statistical Modeling: A Fresh Approach.* `http://www.macalester.edu/~kaplan/ism/` [Called "ISM" for the rest of the syllabus]

Recommended: Gonick, L. and Smith, W. (1993). *The cartoon guide to statistics.* HarperPerennial New York, NY Nice coverage of hypothesis testing and confidence intervals as well as other topics at a very accessible level.

Verzani, J. (2005). *Using R for Introductory Statistics.* Chapman & Hall/CRC Another nice textbook combining statistics with R. (see `http://wiener.math.csi.cuny.edu/UsingR` for more materials related to this book.)

Becker, H. S. (1986). *Writing for Social Scientists: How to Start and Finish Your Thesis, Book, or Article.* University of Chicago Press A wonderful book on social science writing. We will be grading the final papers under the assumption that you write the way Becker advises us to write.

Abelson, R. (1995). *Statistics as Principled Argument.* Lawrence Erlbaum, New York Provides some very useful frameworks for how one might use statistics within the context of doing scholarly work.

Computing

In this class, we will be using the R statistical language. This means that we will be learning some computer programming skills. We will be typing sequences of commands in the R language in to a

text editor and then asking the R interpreter to execute these commands. We will not be pointing and clicking to execute statistical analyses.

Computing is an essential part of modern statistical data analysis — both for producing persuasive information from data and for conveying that information to decision makers. So we will pay attention to computing, with special emphasis on understanding what is going on behind the scenes.

The final reports must be turned in on the class Moodle either as pdf, postscript, or html. Documents in Microsoft Word format (or Wordperfect, or Pages, or OpenOffice) will not be accepted. Neither the professor nor the TA can be counted on to read any document not in pdf, postscript, plain text, or html formats. The in-class work, of course, will be completed with pencil and paper after using R to produce the computations.

Schedule

Note: This schedule is preliminary and subject to change. If you miss a class make sure you contact me or one of your colleagues to find out about changes in the lesson plans or assignments.

Wednesday, January 19 — A Taste of the Course

Task Bring laptops if you have them.

Watch: The introduction to Rstudio videos (`https://courses.las.illinois.edu/mod/resource/view.php?id=40975`)

Section Read ISM § 1.4. Bring laptops if you have them.

Monday, January 24 — What does it mean to use statistics to answer political questions and Scientific Computing using R

Read ISM Chap. 1

Wednesday, January 26 — What do we mean when we say "data"?

Read ISM Chap. 2

Monday, January 31 — Why do we want to talk about variation?

Read ISM Chap. 3

Wednesday, February 2 — Why do we want to talk about variation?

Read ISM Chap. 3

Monday, February 7 — Why do we care about models?

Read ISM Chap. 4

Wednesday, February 9 — Why do we care about models?

Read ISM Chap. 4

Monday, February 14 — What is a linear model? How are linear models useful?

Read ISM Chap. 5

Wednesday, February 16 — What is a linear model? How are linear models useful?

Read ISM Chap. 5

Monday, February 21 — How can we fit linear models to data?

Read ISM Chap. 6

Wednesday, February 23 — How can we fit linear models to data?

Read ISM Chap. 6

Monday, February 28 — Correlation . . .

Read ISM Chap. 7

Wednesday, March 2 — . . . vs. Causation

Read ISM Chap. 8

Monday, March 7 — Why "hold constant" and what does this mean anyway?

Read ISM Chap. 8

Wednesday, March 9 — The "holding constant" problem.

Read ISM Chap. 8

Monday, March 14 — How can linear models teach us about causal relationships?

Read Gelman and Hill 2007, Chapter 9; Berk 2004, Chapter 6.5

Wednesday, March 16 — Causal inference with linear models

Read Gelman and Hill 2007, Chapter 9; Berk 2004, Chapter 6.5

Monday, March 21 — Spring Break

Monday, March 28 — How can we talk about uncertainty about (or confidence in) our models? What is a "sampling distribution"?

Read ISM Chap. 14

Wednesday, March 30 — Sampling Distributions and model uncertainty

Read ISM Chap. 14

Friday, April 1 — Draft Report Proposals Due by 5pm

Monday, April 4 — Other uses for sampling distributions: Predictive Assessments, and More Intervals

Read Gelman and Hill 2007, Chapter 8

Wednesday, April 6 — Other uses for sampling distributions

Read Gelman and Hill 2007, Chapter 8

Monday, April 11 — Why test a hypothesis?

Read ISM Chap. 15

Wednesday, April 13 — Why test a hypothesis?

Read ISM Chap. 15

Friday, April 15 — Final Report Proposals Due by 5pm

Monday, April 18 — Can math can make our lives easier?

Read Freedman, Pisani and Purves 2007, Chap 16–18. The Law of Large Numbers and the Central Limit Theorem.

Wednesday, April 20 — Do we always have to resample/permute?

Read Freedman, Pisani and Purves 2007, Chap 16–18. The Law of Large Numbers and the Central Limit Theorem.

Monday, April 25 — Hypotheses about whole models? Evaluating fit. More Predictive Plotting and Checking

Read ISM Chap. 16

Wednesday, April 27 — Hypotheses about whole models? Evaluating fit.

Read ISM Chap. 16

Monday, May 2 — Putting it all together: Report Writing I

Read TBA

Wednesday, May 4 — Putting it all together: Report Writing II

Read TBA

Friday, May 13 — Final papers due by 5:00pm