Political Science 230
Introduction to Statistics for Political Scientists

Professor Jake Bowers (jwbowers@illinois.edu)
TA Jason Renn (duurenn2@illinois.edu)
Moodle: https://learn.illinois.edu/course/view.php?id=8782

Fall 2014

General Information

This class is an introduction to applied statistics as practiced in political science. It is computing intensive, and, as such, will enable students to execute basic quantitative analyses of social science data using the linear model with statistical inference arising from resampling and permutation based techniques as applied in the R statistical computing language. 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 how the analyses bear on focused social science or policy questions.

Where/When The whole class meets in 1000 Lincoln Hall on Tues and Thurs from 2:00pm to 2:50pm.

Sections meet as specified online on Fridays.

Moodle enrollment key is: Ilovestatistics!

Office Hours Jake's office hours are Monday 2–3:30pm by appointment in 432 David Kinley Hall (DKH) or other times by appointment. I am very happy to meet with you. If you know in advance that you want to come to office hours, please email me to reserve a 20 minute slot. I have found that making appointments for office hours leaves fewer students sitting in the hall waiting to talk with me. 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.

Jason's office hours are Tuesday 12:30pm–2pm and Friday 11:00am–12:00pm or by appointment in 433 David Kinley Hall. He would also prefer that the students make appointments in twenty minute blocks.

Note: Students should bring their own laptops (if they own laptops) to Jason'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 some experience with the Internet.

We also assume that you will 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 already answered in the syllabus.

This is an experimental class so you should expect that the syllabus will change throughout the term. Make sure you have the syllabus with the latest date stamp. We will announce syllabus changes via the emails sent from the Moodle.

If you have any special needs (for example any disability that you'd like us to know about) please contact us during the first two weeks of class. We are happy to help and to work with the folks from http://www.disability.illinois.edu/.

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. We anticipate that 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.

As we specify below, we will grade one problem chosen at random from each day's worksheet. Here is how we might use R to choose the problem to grade from a particular worksheet, 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))
## what does replicate() do? Type help(replicate) ?replicate in R.
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
0.25 0.26 0.25 0.24
```

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 no longer than about 10 pages. 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 and class sessions 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 will also require you to turn in two partial drafts of your paper — thus ensuring that you do not have to scramble at the end of the term to complete your paper and that you have some input on your work.

We will be working through the format of the final paper as the class proceeds. Roughly speaking, in the final paper, we will expect you to put together what you've learned in class with your own interests to execute a simple bit of statistical data analysis (including fitting a linear regression model, and graphing and interpreting the results). For example, you will produce a regression table and be able to explain what p-values and confidence intervals mean as well as the substantive meaning of the regression coefficients.

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

50% In-Class Work and Attendance This part of the grade consists of 70% in-class work grade and 30% attendance. We'll drop the lowest 3 of the daily worksheet grades. The worksheets will will require you to do open-ended data analysis to arrive at the correct answer although the answers will tend to require very little writing. 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 class 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)`.[1] The final proposal grade itself will be calculated in the same way as `max(Final Proposal*.80+Draft Proposal*.20, Final Proposal)`.

As you can see, we aim to reward improvement.

10% Participation The Professor and TA will consult with each other to assign a letter grade reflecting the quality of participation. In the past, this part of the grade has reflected both attendance and quality participation at sections and also useful conversations during office hours or email as well as a our sense about whether you did the readings and have a constructive and engaged attitude toward the course during the main twice-weekly class meetings.

Successful participation in the political science department subject pool will give you 1 percentage point per experiment that you do towards your participation grade.

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

[1] Type `help(max)` in R to find out what this command does.

Computers in class Please bring your laptops if you have them. If you do not own a laptop, you can still work in a group of other people who have laptops and will be able to complete the in-class worksheets without a problem. In fact, it is ideal if each group of 2–4 people works with one laptop and then shares the work among themselves. Of course, feel free to work on your own. We just think it will be more fun and fulfilling if you work with your colleagues in the class.

Books

Required: Kaplan, D. (2012). *Statistical Modeling A Fresh Approach*. Daniel Kaplan, Macalester College, St. Paul, MN, second edition [Called "ISM" for the rest of the syllabus. The first few chapters are available online for free at http://www.mosaic-web.org/go/StatisticalModeling/index.html.]

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.]

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 (http://www.r-project.org) and the RStudio integrated development environment for R (http://rstudio.org). This means that we will be learning some computer programming skills. We will be typing sequences of commands in the R language in a text editor (http://en.wikipedia.org/wiki/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 Jake or Jason or one of your colleagues to find out about changes in the lesson plans or assignments.

Tuesday, August 26 — A Taste of the Course

Task Bring laptops if you have them.

Watch: The Rstudio screen cast: http://rstudio.org/ or other Rstudio videos.

Check out: Some resources about R http://rstudio.org/docs/help_with_r

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

Thursday, August 28 — No Class

Tuesday, September 2 — What do we mean when we say "data"?

Task From now on bring laptops if you have them. Ensure that you either have an RStudio account (see the Moodle for instructions) or have an installed and working copy of R (using either the desktop version of RStudio or some other GUI).

Read ISM Chap. 1 and 2

Thursday, September 4 — Why do we want to talk about variation?

Read ISM Chap. 3

Tuesday, September 9 — Why do we want to talk about variation?

Read ISM Chap. 3

Thursday, September 11 — Why do we care about models?

Read ISM Chap. 6 (ISM Chap 4 as Background)

Tuesday, September 16 — Why do we care about models?

Read ISM Chap. 6 (ISM Chap 4 as Background)

Thursday, September 18 — What is a linear model? How are linear models useful?

Read ISM Chap. 7

Tuesday, September 23 — What is a linear model? How are linear models useful?

Read ISM Chap. 7

Thursday, September 25 — How can we fit linear models to data?

Read ISM Chap. 8

Tuesday, September 30 — How can we fit linear models to data?

Read ISM Chap. 8

Thursday, October 2 — Model Fit and R^2 ...

Read ISM Chap. 9

Tuesday, October 7 — Statistical Adjustment: Holding Constant

Read ISM Chap. 10

Tasks Interlude on final papers.

Thursday, October 9 — Holding Constant by Subsetting or Stratifying

Read ISM Chap. 10

Tuesday, October 14 — Holding Constant by Residualization, Covariance Adjustment, "Controlling For"

Read ISM Chap. 10

Thursday, October 16 — When is holding constant meaningful?

Read Berk 2004, Chapter 6.5 and 7.2

Tuesday, October 21 — Report proposal practice

Tasks Come to class ready to write and talk about your outcome, explanatory, and control variables, and possible data.

Thursday, October 23 — How can we talk about uncertainty about (or confidence in) our models? What is a "sampling distribution"? Confidence intervals.

Read ISM Chap. 12 (and ISM Chap 5 as background)

Tuesday, October 28 — Sampling Distributions and model uncertainty

Read ISM Chap. 12 (and ISM Chap 5 as background)

Thursday, October 30 — Report Proposal Workshop

Task Come to class prepared to work on your report proposal (i.e. your plan for writing your report).

Friday, October 31 — Report Proposals Due by 5pm

The report proposal is a proposal: it tells us what question you have and what dataset and variables you plan to use. The point of this assignment is to get you thinking seriously about your final paper and doing the hard work of finding data, grappling with codebooks, and thinking about how outcome variables, explanatory variables, and control variables relate to your questions. Specifics about the form of this proposal (and draft report and final report itself) will be handed out in class as the term progresses.

Tuesday, November 4 — What is a hypothesis test?

Read ISM Chap. 13, 14, 15

Thursday, November 6 — Why test a hypothesis about a parameter in a model?

Read ISM Chap. 13, 14, 15

Tuesday, November 11 — Why test a hypothesis about a parameter in a model?

Read ISM Chap. 13, 14, 15

Thursday, November 13 — Report Draft Workshop

Task: Come to class ready to work on the draft of the first half of your report.

Friday, November 14 — Report Drafts Due by 5pm

Tuesday, November 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.

Thursday, November 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.

Tuesday, November 25 — Fall Break

Thursday, November 27 — Fall Break

Tuesday, December 2 — Hypotheses about whole models? Evaluating fit. More Predictive Plotting and Checking

Read ISM Chap. 14

Thursday, December 4 — Hypotheses about whole models? Evaluating fit.

Read ISM Chap. 14

Tuesday, December 9 — Putting it all together: Report Writing

Task Come to class ready to work on your final reports.

Monday, Dec 15 — Final Reports due by 5:00pm