

Sociology 210A

Univariate Statistics

Gabriel Rossman
rossman@soc.ucla.edu

November 5, 2009

It is no great wonder if in a long process of time, while fortune takes her course hither and thither, numerous coincidences should spontaneously occur. If the number and variety of subjects to be wrought upon be infinite, it is all the more easy for fortune, with such an abundance of material to effect this similarity of results. Or if, on the other hand, events are limited to the combinations of some finite number, then of necessity the same must often recur, and in the same sequence. There are people who take a pleasure in making collections of all such fortuitous occurrences that they have heard or read of, as looks like works of a rational power and design ...

–Plutarch

This course serves as the first part of a three quarter sequence in statistics for sociology graduate students. The 210 sequence is narrowly focused on statistical analysis and does not cover other issues in quantitative methodology such as sampling, data collection, and write-up. These issues are addressed in 212AB. Thus students interested in *creating* quantitative research should consider this the beginning of a five quarter sequence whereas those who only need to be able to *understand* quantitative research can consider 210AB a terminal sequence. 210A covers the basics of data, distributions, and central tendencies. 210B covers basic regression methods like OLS and logit. The optional course, 210C, covers advanced regression methods such as event history and random effects for clustered data. All of these courses ignore the kind of proofs built up from probability theory you would encounter in a statistics course taught in the math department and instead focus on practical considerations of how to interpret statistics, and even more important how to understand their limitations and assumptions. There are two goals for the sequence:

1. All students who have taken 210AB should be able to read and get the gist of almost any article published in *ASR* or *AJS*. Although many, if not most, quantitative articles now use the types of models taught in 210C, these models are analogous enough to the 210B models that you should

be able to basically understand them even if you'll have to take it on faith that the authors got the details right.

2. Those students who are interested in pursuing advanced quantitative methods will be well-prepared to do so.

Since the first goal applies to all students, issues that apply to it are mandatory whereas issues that apply to the second goal are optional. The main implication of this is that learning to use Stata is optional. (Stata is a very flexible, powerful, and reasonably easy statistics and database program that is very popular with social scientists.) There will be two versions of most assignments, the *output* version in which I provide tables and graphs for you to interpret and the *coding* version in which I provide you raw data and you generate the tables and graphs yourself. Students planning to pursue 210C and 212AB should do the latter version of the assignments as these classes require Stata.

Readings

Although most graduate sociology courses assign primary texts and journal articles, statistics is sufficiently normal science-y that a textbook is much more appropriate. There is one required textbook and three optional textbooks. Each week I will assign readings from the core text and optional readings from the other texts. The mandatory text is

Agresti, Alan and Barbara Finlay. *Statistical Methods for the Social Sciences* [Fourth Edition]. Upper Saddle River, NJ: Prentice Hall.

Agresti and Finlay is an introductory statistics textbook that covers most of the material in 210A and 210B while giving a brief introduction to the issues in 210C. The book's philosophy is similar to that of the class in that it emphasizes intuition and assumptions rather than proofs. Although the price is a bit steep you should expect to keep a statistics textbook to serve as a reference work. I still have and occasionally refer to all my statistics textbooks from both undergrad and grad school. If you buy your copy online search for ISBN #0130272957 to avoid getting the wrong edition.

There are also several optional texts.

Keller, Dana K. 2005. *The Tao of Statistics: A Path to Understanding (With no Math)*. Thousand Oaks, CA: Sage.

Keller consists of a series of 47 short essays that very clearly explain the intuition behind most of the concepts that come up in 210ABC. If you find you are having any trouble understanding the core text, the lecture, or the exercises then I recommend reading *Tao*.

Acock, Alan C. 2008. *A Gentle Introduction to Stata* [2nd edition]. College Station, TX: Stata Press.

Acock gives a very gradual introduction to Stata suitable for people who have never used statistical software or programming before. It covers most of the material from 210AB. This book is recommended for people who are interested in doing the coding version of the assignments but are not yet familiar with Stata

or similar packages. Don't expect a tutorial from the regular Stata manuals as they are really reference works, not textbooks.

Hamilton, Lawrence C. 2009. *Statistics with Stata (Updated for Version 10)*. Belmont, CA: Brooks and Cole.

Hamilton is similar to Acock so there's no reason to get both. The basics like loading data and creating a do file are a little bit more abrupt than in Acock. On the other hand, Hamilton covers issues through 210C. This book is recommended for people who are interested in going through 210C and who already have a basic familiarity with Stata or comparable statistics packages like SPSS. (Note that older editions will probably work about as well and be available much cheaper.)

If you are interested in software and practicing quantitative research you should also be aware of some really excellent programming tutorials through CCPR (California Center for Population Research) and ATS (Academic Technology Services). The CCPR site is excellent at clearly explaining the big picture of good programming that are essential for any kind of complex dataset construction but often get lost if you just concentrate on learning analysis commands. The ATS website and consulting has a lot on the basics but really shines for the sort of exotic syntax and software used for techniques explored in 210C.

http://ccpr.ucla.edu/Computing_Services/Tutorial/index.asp

<http://www.ats.ucla.edu/stat/stata>

Long, J. Scott. 2009. *The Workflow of Data Analysis Using Stata*. College Station, TX: Stata Press.

The *Workflow* book provides very solid advice on more advanced Stata usage related to data management. The content is similar to the CCPR and ATS sites but more thorough and systematic.

1 Introduction

- Tuesday
 - Basic Concepts
 - Categorical, Ordinal, Continuous, and Count Data
 - Sampling
- Thursday
 - Introduction to Stata

Reading

Agresti and Finlay. Chapters 1 + 2.

Acock. Chapters 1-4.

Keller. Chapters 1-6.

2 Descriptive Statistics and Bayes Theorem

- Tuesday
 - Histograms, Box Plots, and Scatterplots
 - Mean, Median, and Mode
 - Range, Standard Deviation, Quartiles
- Thursday
 - Probability and Bayes Theorem

Reading

Agresti and Finlay. Chapter 3.
Acock. Chapter 5.
Keller. Chapter 12 + 26.

3 Probability Distributions (and scholarly word processing)

- Tuesday
 - Normal Distribution (or Bell Curve)
 - Z-Scores (or Standardizing)
- Thursday
 - Text Editors vs. Word Processing
 - Generating Tables
 - Styles
 - Citations

Reading

Agresti and Finlay. Chapter 4.
Keller. Chapters 7-11.

4 Statistical Inference: Bootstrapping

- Tuesday
 - Assumption underlying standard errors
 - Resampling
- Thursday

- The Gordian knot
- QAP and other shuffling algorithms

Reading
TBA

5 Statistical Inference: Estimation

- Tuesday
 - Estimate \pm Standard Error = Confidence Interval
 - Sample Size (or n)
 - t Distribution
- Thursday
 - review session

Reading
Agresti and Finlay. Chapter 5.
Keller. Chapters 13-22.

6 Statistical Inference: Significance Tests

- Tuesday
 - Null (H_a) vs Alternative Hypothesis (H_a)
 - p -value
 - One and Two-Sided (or -Tailed) Tests
- Thursday
 - False Positives and False Negatives: The Scylla and Charybdis of Significance Tests
 - Publication Bias

Reading
Agresti and Finlay. Chapter 6.
Keller. Chapters 27-31, 34-35.

7 Comparison of Two Groups (and Stata programming)

- Tuesday
 - t -test of means
- Stata programming
 - macros
 - loops and programs
 - pipes

Reading

Agresti and Finlay. Chapter 7.
Acock. Chapter 7.

8 Association Between Categorical Variables (and Philosophy of Science)

- Tuesday
 - Contingency tables and marginal distributions
 - Expected frequencies (as null hypothesis)
 - Interpreting the χ^2 distribution
 - Odds-ratios
- Thursday
 - Popper and positivism
 - Quine and holism
 - Kuhn and scientific realism

Reading

Agresti and Finlay. Chapter 8.
Acock. Chapter 6.
Keller. Chapter 32.

9 Pathologies of Statistics

- Tuesday
 - Sampling on the Dependent Variable
 - Censorship
- Thursday (holiday)

Reading

Keller. Chapter 33.

TBA

10 More Pathologies of Statistics

- Tuesday
 - Statistical versus Substantive Significance
 - Reifying Data
 - Overcontrolling
 - Assymmetric Causation
- Thursday
 - Final

Reading

TBA