Sociology V3212
Introduction to Data Analysis (Statistics/Methods)

Fall 2006

	Office	Email address	Office hours
Instructor: Matthew Salganik	270 IAB	mjs2105@columbia.edu	Tu: 1:30-2:30, F: 11-noon
Teaching Assistant: Uri Shwed	805 IAB	us2121@columbia.edu	F: 11-noon

1 Course information

The lectures will be Monday and Wednesday from 5:40-6:55pm in 252 Engineering Terrace. There is also a required lab session which will be on Friday from 10:00-10:50am in the same room. Please make sure you have also registered for the lab section. This course is required for all sociology majors. If you have not already taken Evaluation of Evidence (SOCV1125), please see the instructor for permission to take this course.

2 Books and software

The textbook for the course will be *Intro Stats (2nd edition)* by De Veaux, Velleman, and Bock. There will also be additional readings available from the instructor. In the course we will use the computer program Stata to analyze data so you will need to have access to a computer with Stata to complete your homework assignments and exams. All computer labs on campus have Stata. You can also access Stata using cunix by following these directions: `www.columbia.edu/acis/eds/stat_pak/stata/stata-unix.html`. If you want to purchase Stata for your personal computer, you should purchase Intercooled Stata (not Small Stata). For Columbia students, the cost is $89 for a one year license or $145 for a permanent license. More information is available from: `www.columbia.edu/acis/software/licsenses/stata/`. Students who are not familiar with Stata may also want to purchase *A Gentle Introduction to Stata* by Acock. However, this book is not required because there are many places on the web to get Stata help.

3 Course overview and goals

This course will introduce students to the analysis of quantitative data in the social sciences. Since most real data analysis is done using a computer, students will also learn to use the computer program Stata. The four main goals of the course are to prepare students to:

- write senior research papers
- read sociological research as it is published in professional journals
- understand the use of statistics in policy debates
- participate in future statistics courses

In lecture, we will cover chapters 1-22 of the De Veaux book (excluding chapters 10, 16, and 17) at the rate of approximately one chapter per lecture. There will also be outside readings to illustrate ideas from the text as well as cover topics not included in the text (multivariate regression and causal inference). The lab will be used to re-enforce the material covered in lecture and introduce the students to Stata. The course will be divided into three main sections: 1) exploring and understanding data, 2) relationships between variables, and 3) sampling, randomness, and inference.

3.1 Exploring and understanding data

We will review the different types of data and how they can be presented graphically. We will also discuss ways of numerically describing the center and spread of a distribution. Some of these concepts will be developed around the debate on income inequality in the United States. Students will analyze income data from the 1965-2005 Current Population Survey (CPS) to explore these issues.

Key concepts:

- categorical and continuous data
- contingency tables/cross-tabs (row percents, column percents, conditional distribution)
- bar charts
- histograms
- mean and median
- variance, standard deviation, inter-quartile range
- re-scaling and standardizing data
- visually testing for normality (normal probability plot)

Sample of outside reading (provided by instructor): "For Richer" by Paul Krugman, *New York Times Magazine* (2002), "The Income Inequality Debate" by Herbert Stein, *The Wall Street Journal* (1996), "The Median Isn't the Message" by Stephen Jay Gould, *Discover Magazine* (1985), "Million-Dollar Murray" by Malcolm Gladwell, *The New Yorker* (2006).

3.2 Relationships between variables

In the second section of the course we will develop techniques to study the relationships between variables. This task is the core of most social research and can be quite tricky. In addition to statistical issues, we will also consider questions of research design and the limits of what we can really learn about social systems. The methods developed in this part of the course will be introduced around the the policy debate surrounding school choice and vouchers. Students will analyze student performance data from the National Longitudinal Study of Youth (NLSY) to explore these issues.

Key concepts:

- scatter plots
- correlation as a measure of association
- lurking variables (correlation is not causation)

- linear regression
- residuals
- outliers
- cautions about regression
- multivariate regression
- dummy variables
- interactions
- collinearity as a practical and technical concern
- limits of observational studies
- experiments, strengths and weaknesses
- field experiments, natural experiments, and instrumental variables

Sample of outside reading (provided by instructor): "Public schools perform near private ones in study" by Diana Jean Schemo, *The New York Times* (2006), "Schools are her business" by John Cassidy, *The New Yorker* (2000), "Economist's study on school competition stirs debate" by Jon Hilsenrath, *The Wall Street Journal* (2005).

3.3 Sampling, randomness, and inference

Most analysis of social data involves working with a sample of cases rather than the whole population. First, we will discuss how these samples can be collected. This part of research is often neglected, but it can cause real problems if not done properly. Then we will discuss how to use information from the sample to make inference about the population. The section of the course will be more mathematical and will lay the groundwork for any future coursework in statistics.

Key concepts:

- sampling and the different kinds of samples
- randomness and probability
- distribution of sample statistics
- confidence intervals for proportions
- hypothesis tests, type I and type II errors
- p-values
- comparing proportions
- problems with hypothesis tests and statistical significance

4 Homework, tests, and grading

Students will have weekly homework assignments. You can work on these assignments together, but each student must turn in a separate write-up. Also, each student must write their own Stata code. In addition to the homework, there will also be a number of short quizzes given in both lab and lecture. Finally, there will be a mid-term exam and a cumulative final exam. Both of these exams will have an in-class component which will be timed, closed-book, and without Stata and a take-home component which will be untimed, open-book, and with Stata. Your grade will be based on your homeworks and quizzes (30%), the mid-term (30%), and the final exam (40%).

5 A note on course content

There is really too much important and interesting material to cover in just one semester (some people spend their whole lives studying statistics). Given this overload of materials, I had to make some decisions about what to include and what to exclude. Compared with other courses, our treatment of hypothesis testing, probability distributions, and transformations of variables will be reduced. It's not that these topics are not important. Rather, I have reduced the treatment of them to give us more time for other topics. Specifically, we will cover both multivariate regression and causal inference. These topics are central to both social research and policy debates taking place in the media (for example, the school voucher debate). These topics may also be important in your senior papers. We will also spend time learning to work with real data which, unlike class examples, include messy things like complex sampling designs, missing data, and top-coding. Again, this is likely to be important while you are writing your senior papers. Finally, since many of the methods will be introduced around the debates on income inequality and school vouchers, we will spend some time talking about the substantive issues in these debates. Good data analysis requires substantive knowledge and cannot be done in a vacuum.