

Sociology G4074: Introduction to Social Data Analysis I

December 20, 2005

Tuesday and Thursday, 10:35am-11:50am
311 Fayerweather

Lab: Wednesday 9-10:50am
Mathematics 407

	Instructor	TA
	Aaron Gullickson	Eric Johnson
office:	412 Fayerweather Hall	402 Fayerweather
email:	ag2319@columbia.edu	ebj2001@columbia.edu
office hours:	Wednesday, 2-4pm	Fridays, 10am-12pm

Course Objectives

This course will teach the fundamentals of analyzing numerical data in a social science context. Students will learn effective ways of presenting informational summaries, the use of statistical inference from samples to populations, and the linear model which forms the basis of much social science research. Emphasis will be on an intuitive understanding of statistical formulae and models, and on their practical application. In the continuation of this course next semester, students will learn some important variations of the linear model and other advanced topics.

Readings

There is only one required textbook for the class:

- Moore, David S. and George P. McCabe. 2004. Introduction to the Practice of Statistics, 5th edition. New York: W.H. Freeman and Company.

This book should be available at the Columbia Bookstore.

In addition, the following book is recommended for those who have no familiarity with the STATA software program.

- Sophia Rabe-Hesketh and Brian Everitt. *A Handbook of Statistical Analyses using Stata*. New York: Chapman and Hall.

This book should also be available from the Columbia Bookstore.

Additional material will be distributed in class or available as a PDF file on the courseworks website.

Course Organization

This course will consist of biweekly lectures (Tuesdays and Thursdays), and a lab section on Wednesdays. There will be weekly homework assignments, available from the course website, and due the Tuesday after they are assigned. The lab section is designed to introduce students to statistical programming languages, particularly STATA, and to provide a collaborative environment for working on homework problems.

There will be a midterm and a final examination for this class. The final exam will be cumulative, but disproportionately focused on material presented after the midterm exam.

Grading

Homework will count for 50% of the grade, and each of the tests will count for 25% of the grade. The two lowest homework grades will be dropped.

Course Outline

1. Describing Data
 - (a) Administrative and philosophical overview
 - (b) The idea of a distribution
 - (c) Measures of center and spread
 - (d) Measuring relationships with categorical variables
 - (e) Measuring relationships between quantitative variables
 - (f) Interpreting OLS regression and transformations
 - (g) Multivariate regression
 - (h) A research example

- (i) Regression diagnostics: outliers and influential points
- (j) Regression diagnostics: aggregate data, collinearity, and causality
- (k) The matrix approach
- (l) Review

2. Midterm

3. Data Collection and Statistical Inference

- (a) Data collection (experiments and surveys)
- (b) Probability rules
- (c) Random variables
- (d) The sampling distribution
- (e) The binomial case of the sampling distribution
- (f) Confidence intervals and hypothesis tests
- (g) Errors in hypothesis testing: statistical significance and power
- (h) t-tests and the t-distribution
 - (i) Inference for contingency tables
 - (j) Inference for bivariate regression
 - (k) Inference for multivariate regression
 - (l) Non-parametric tests
- (m) Review