Data Analysis & Statistics
POLS W3704
Columbia University, Spring 2013
MW 2:40-3:55
Professor Donald Green

Office hours: Tuesdays 9-12
Office location: IAB 814
Email: dpg2110@columbia.edu
Web: https://sites.google.com/site/donaldpgreen

Teaching Assistants: Noah Buckley <nmb2137@columbia.edu> and Ryan Luby <rpl2126@columbia.edu>

Course overview: This course provides a hands-on introduction to social science data analysis. The aims of the course are threefold: (1) to familiarize students with core concepts in statistics, such as sampling distributions and hypothesis testing, (2) to introduce students to software that can be used to analyze data and simulate statistical processes, and (3) to apply statistical techniques to political phenomena. Special emphasis will be placed on the interpretation and presentation of statistical results. Approximately half of our meetings will be devoted to lectures; the other half will be lab sessions involving hands-on data analysis.

Prerequisites: This course is intended for political science majors, but non-majors are welcome. A solid grasp of high school algebra is required. Familiarity with statistics and statistical software is helpful but not required.

Readings: Students are expected to keep up with each week's reading. We will be using two course books.

The first is a (free) resource *Online Statistics Education: A Multimedia Course of Study* by David M. Lane, et al., Version 2.0. A PDF of the book and accompanying lectures, videos, quizzes, and animated applets may be found at http://onlinestatbook.com/2/index.html.

The second is the 2012 release of the textbook *Statistical Modeling: A Fresh Approach* (2nd edition) by Daniel T. Kaplan. This book provides R examples for each of the statistical concepts it discusses will help guide our lab sessions.

Statistical software: Our lab sessions and assignments will use the free software package R, which may be downloaded at http://www.r-project.org. Although R is more challenging to use than other point-and-click statistical programs, it is worth the extra effort. In order to minimize any anxiety students may have about programming in R, we will use a "template-driven" approach throughout the term. Through Courseworks, I will distribute the programs for each lab session, and students will download, adapt, and run them. Students will not be

required to develop their own programs (although they are certainly encouraged to do so if they wish). The Kaplan book provides a basic introduction to R, and the Web is replete with free on-line guides and tutorials.

Assignments: We will have a take-home assignment for each of the four modules (descriptive statistics, probability, sampling distributions and hypothesis testing, regression). Each assignment will count for 10% of the overall grade. Assignments are due on the day of the exam for each module. There will also be an in-class exam at the end of each module. Each exam will count for 10% of the overall grade. The in-class final exam (which will also function as the test for the fourth module of the class) will count for 30%. Full attendance and participation are assumed; failure to show up for class will count against the final grade.

The course schedule is as follows:

January 23: Introduction and Overview

January 28 and 30: Descriptive statistics: types of variables; categorical distributions; location and dispersion of continuous variables; communicating descriptive information through graphs and tables; an introduction to R

Reading: Kaplan chapters 1-2; Lane chapters 1-3

February 4 and 6: Bivariate statistics: measures of association and graphical representations; how transformations (logs, ranks) change the numerical and visual assessment of relationships; the corrosive effects of measurement error; what correlations do and don't tell us about linear/non-linear relationships and about causal influence.

Reading: Kaplan chapter 3 and 9.2-9.3; Lane chapter 4

February 11: Time-series relationships and the idea of leads, lags, and placebo tests; out-of-sample tests to avoid spurious correlations

February 13: First midterm

February 18 and 20: Introduction to probability: probability axioms, independence, conditional probability

Readings: Lane chapter 5.1 – 5.7

February 25 and 27: Bayes rule, prior beliefs, and posterior beliefs

Readings: Lane chapter 5.13 – 5.16

March 4 and 6: working with probability statements based on the normal distribution; using probability models (normal, Poisson) to characterize actual social processes; using

simulations in order to come up with approximate solutions to difficult problems

Readings: Kaplan chapter 11; Lane chapter 5.8 – 5.11, 7

March 11: Second midterm

March 13: Sampling distribution of a mean calculated based on a random sample from a population, the Central Limit Theorem, and hypothesis testing

Reading: Lane chapter 9

March 25 and 27: Confidence intervals for the population mean and difference-in-means

Reading: Kaplan chapter 5; Lane chapter 11

April 1: Review of hypothesis testing and confidence intervals

April 3: Third midterm

April 8 and 10: Introduction to regression

Reading: Kaplan chapters 6, 7, 8.1; Lane chapter 14.1 – 14.9

April 15 and 17: Using regression to predict and forecast

Reading: Kaplan chapter 12

April 22 and 24: Multiple regression, indicator variables, and interactions

Reading: Kaplan chapter 10; Lane chapter 14.10

April 29 and May 1: Experiments and causal inference

Reading: Kaplan chapter 17, 18; Lane chapters 6.6, 6.7

TBD: In-class final exam