Advanced Quantitative Research Methodology, Gov2001, Gov1002, and E-2001

Gary King, Jennifer Pan, and Molly Roberts

Class: 2–4pm Mondays (CGIS South 010, Tsai Auditorium); Section: 6–7:30pm Thursdays (CGIS K354).

Gary King
King@Harvard.edu, http://GKing.Harvard.edu
Phone: 617-500-7570, Assistant: 617-495-9271
Office: 1737 Cambridge Street, N313

Jennifer Pan, Teaching Fellow
jjpan@fas.harvard.edu
Office hours: 2–4pm Wednesdays (CGIS HMDC Basement Lab), or by appointment

Molly Roberts, Teaching Fellow
roberts8@fas.harvard.edu
Office hours: 4–6pm Mondays (CGIS HMDC Basement Lab), or by appointment

Online material At http://bit.ly/gov2001 you will find a link to the full class web site, which includes detailed lecture notes, a PDF version of this document, links to electronic copies of the readings (access to one must be purchased; see below), a video annotation tool, a text annotation tool, assignments, and other materials.

Who Takes This Course? Do I have to take it for a grade? Can I sit in? Following Gov 2000 or the equivalent course, Gov 2001 is the second in the methods sequence for Government Department graduate and undergraduate students. While not required, most Government graduate students doing empirical work take the course. Graduate students in other departments and schools at Harvard (and in the area) also take the course. Undergraduates preparing to write quantitative theses are especially welcome to take Gov 1002, which is taught along with this class. Non-Harvard students and others may also take this course by registering through the Harvard extension school, for which course credit is available if desired (see course number E-2001).

If there are seats in the room you're welcome to attend even if you're not formally registered, but if possible we would appreciate if you would sign up formally (as our teaching fellows get paid more!). If you are not a Harvard student, you can easily do this via Harvard extension school course E-2001.

If you need cross-registration papers signed, please bring them to the first class. We observe that students who take the course for a grade participate more and get far more out of the experience (even among many of those who think or say it will be otherwise), but pass/fail and formal auditing are okay with us too.

Overview Building on the analytical and theoretical background of Gov 2000, this course gives you the tools to build statistical models useful in real social science research. The course covers how to develop new approaches to research methods, data analysis, and statistical theory. More advanced statistical theory is not required when data and variables fit standard assumptions. Since this is not usually the case in political science and related disciplines, we often cannot use ready-made statistical procedures developed elsewhere and for other purposes. Once an underlying theory of inference is understood, it is easy to "reinvent" known statistical solutions to accommodate social science data, or to conceive original approaches and new statistical estimators when required.

Upon finishing the course, students should be able to read an original scholarly article describing a new statistical technique, implement it in computer code, estimate the model with relevant data, understand and interpret the results, and explain the results to someone unfamiliar with statistics.

As an important part of the course, students learn how to make novel contribution to the scholarly literature. As a result, a substantial portion of those who complete the course publish a revised version of their class paper in a scholarly journal. For most students, this is their first professional publication.

Prerequisites Gov 2000, a course in linear regression (with matrices), or the equivalent.

What to Expect in Class This class is designed as a *collective* experience. This means that other students will be counting on you (and you on them), and so please come to class highly prepared. If you don't understand something, that's perfectly fine; we'll figure it out together and make sure no one is left behind. But if you don't put in the effort, it will hurt what everyone gets out of the class.

We have redesigned the course again this year, this time including several new teaching and learning technologies. We expect you to make a genuine effort to participate in the following activities prior to the class for which they are assigned:

- Watch an approximately two-hour video of Gary King giving a class lecture from a previous year.
- Complete the assigned readings
- For both the videos and the readings, our web site offers separate but integrated collaborative annotation tools. This means that if you find a portion of the lecture, or a passage in the reading, difficult or confusing, you should post a question about it. You can pause the video or stop reading and do this right there as you watch or read. Similarly, if you think you may know an answer to a query another student posted, or have a suggestion, please make a contribution to the class and try to answer it. Similarly, if you merely have an interesting idea related to the video or text, please contribute that as well.

Being prepared by having watched the videos and done the reading enables us to devote class time to difficult, confusing, or interesting ideas that arise. We will also be able to make more detailed connections to student projects and interests.

Computational Tools The best way, and often the only way, to learn new statistical procedures is by doing. We will therefore make extensive use of a flexible (open-source and free) statistical software program called R and a companion package called Zelig. R is probably the most widely used statistical software, and Zelig is one of the most widely used packages in R. You will learn how to program in this class, if you do not know already.

For hardware, you are welcome to use your own computers. To install R and Zelig on your computer, see `http://gking.harvard.edu/zelig`. You are also welcome to use the HMDC computer labs, which have computers with R already installed on them. Harvard affiliates also have the option of registering for a Research Computing Environment (RCE) account through http://hmdc.harvard.edu. Having an RCE account allows you access to HMDC's servers, which are fast and well-equipped to handle large data sets or time-intensive procedures. In addition, these servers supply a persistent desktop environment that is accessible from any computer with an Internet connection.

Most of the probability and statistical theory in this class will be taught in the context of "Monte Carlo simulation" (which we do not expect you to know prior to the course). We will write computer programs to verify, or substitute for, more difficult or impossible formal mathematical proofs. This intuitive technique will make it much easier to understand and to implement new statistical methods.

Problem Sets In addition to the final paper, you will have to turn in weekly. We strongly encourage students to work together on the problem sets and quizzes, but the work that you turn in must ultimately be your own. Problem sets are due each week at the beginning of section in the dropbox on the course website. Immediately after section a key will be posted including just the answers (not the full worked out solutions). You will then have 24 hours (until Friday at 6PM) to check your answers and rewrite *one* problem that you got incorrect on your problem set and resubmit the *entire* problem set to the online dropbox. You may only rewrite a problem that you previously attempted. At 6PM on Friday, the full solution key will be posted so you can review your answers. Because we will be posting answer/solution keys immediately after deadlines, late work will not receive any credit. You can still turn in late work for feedback and help learning the material. The problem sets including looking at the solutions key is an extremely important part of the learning process, so please keep up with the work!

Replication Paper The main assignment is to write a research paper that replicates an existing piece of scholarship. The goal of the paper is to apply some advanced method to, or develop one for, a substantive problem in your field of study. You should aim to produce a publishable article, and, in fact, most students do publish their final paper in a scholarly journal. (I know it sounds hard, but that's only because you haven't learned some of the material we go over in class!) More information about the paper can be found at `http://gking.harvard.edu/papers/`.

You must choose a co-author and a paper to replicate by Thursday, March 1, at 5pm, by which point you should email us a PDF copy of the paper along with a brief paragraph explaining your choice. On Thursday, March 22, you must turn in a draft of the paper with little text but with figures and tables, and a proposed table of contents for your paper, in a relatively polished form. You should also arrange to hand over all of the data and information necessary to replicate the results of your analysis and reproduce your tables and figures. (Many students email their files; students with larger datasets often set up shared Dropboxes.) On that day, you will hand over your paper and materials to another student, and, in exchange, you will receive another student's paper. Your task for the following week is to replicate the other student's analysis and write a memo to this student (with a copy to us), pointing out ways to make the paper and the analysis better. You will be evaluated based on how helpful, not how destructive, you are.

The final version of the paper is due the first day of Reading Period, *Thursday, April 26, at 5pm*. You must turn in a hard copy of the paper and arrange to hand over all data and code (either by email or by Dropbox). You must also follow standard academic practice and create a permanent replication archive by uploading all your data and code to the Gov2001 Dataverse (`http://dvn.iq.harvard.edu/dvn/dv/gov2001`).

If you need an extension with the replication paper, you do not need to ask permission: We will accept papers until *Monday, April 30, at 5pm*, but since you will have had more time, papers turned in after the *April 26* deadline will be graded according to proportionately higher standards. The number of incompletes we plan to give is governed by a Poisson distribution with $\lambda = 0.01$, so please plan accordingly.

Once all papers are turned in, we will turn over your replication paper to another student, and assign you a replication paper to evaluate. Your last assignment for the class will then be to read and comment on a fellow student's work and to grade this student according to certain guidelines we will provide. Your main objective is to give the student feedback on what changes and improvements need to happen in order for the paper to be published. As always, you will be evaluated based on how helpful, not how destructive, you are. Your comments on your fellow student's paper are due *May 7, at 5pm.*

Special Rules for Extension and Distance Learning Students This course is being offered as part of the Harvard Extension School's Distance Education Program. The recorded class meetings that you will view are from the Harvard FAS course, Government 2001, and this meets once per week throughout the term. Even though your participation will take place online, you are responsible for homework, readings, quizzes, and all other work. This means you should participate in the text and video annotation like the rest of the class. You should also take the online quizzes that follow the lecture videos. There will also be weekly on-campus section meetings and office hours for students who are able to attend. The sections will also be videotaped. Please see the Harvard Extension School distance education web site for more information.

Students taking the class through the extension school will complete a final exam instead of the replication paper. They will, however, participate in the replication assignment by replicating others' work. An extension school student who prefers to do the replication paper instead of the final exam must get permission from us by the third week of the semester and make satisfactory progress with weekly assignments during the semester.

All students will need to have access to the course webpage, which is operated by FAS. If you do not already have a Harvard ID, please make arrangements to get one or to set up an XID.

Readings All readings are available on the text annotation web site. All are freely accessible to members of this class, except for the required text: Gary King, *Unifying Political Methodology: The Likelihood Theory of Statistical Inference.* (Ann Arbor: University of Michigan Press, 1998). We made special arrangements with the publisher so that you may obtain access to this book on the text annotation web site by either buying physical copy of this book (available at the Coop) or buying or renting electronic access (for as short as a 30 day rental available for $10) at `http://bit.ly/gov2001upm`. When you have obtained access through one of these means, sign up on the class web site and we will give you access to the text annotation tool.

In addition to the required text, we will assign a wide variety of scholarly papers. We will announce at the end of every class meeting what the reading assignment will be.

Help Help is available when you need it. If you have any questions about the homework, your paper, or anything else related to the course, please email the class list at gov2001-l@lists.fas.harvard.edu. Since all three of us and all students will be reachable via this list, it's a very efficient way to get answers to questions that do not fit as comments on the video and text annotation tool sites. Please also respond to inquiries if you happen to know the answer. (If you don't want to receive all the mailings, use a mail filter to put them in a separate folder, although we do not recommend this for this mailing list, and we don't do it ourselves.)

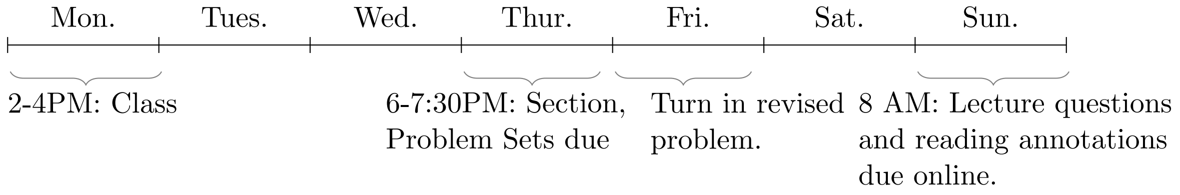
We will also use the class list to email regarding course logistics, including readings, video assignments, and problem sets. Please sign up for the class email list at `http://lists.fas.harvard.edu/mailman/listinfo/gov2001-l`.

We are also making available the course list from previous years. You will have access to all the questions asked by previous generations of students, many who are now professors at top universities.

Grading Final grades will be a weighted average of the replication paper, weekly problem sets and quiz questions outside of class, and *participation.* (There will be no final exam.)

"Participation" includes preparing for and joining in the discussion in class, making a serious effort to contribute to collaborative text and video annotation, answering email list queries if you have a suggestion, and other ways of helping your classmates learn more. Finally, since everyone learns more when more connections exist among students, finding ways to help build class camaraderie can also count as part of participation.

Weekly Schedule The timeline below gives the outline of the weekly schedule. Students are expected to (1) Watch the 2-hour lecture video, (2) Do the assigned readings, (3) Answer the lecture questions and perform annotations in the readings online (Sun 8AM), (4) Attend Class (Mon. 2-4PM), (5) Complete the problem set (Thu. 6PM), (6) Attend section (Thu. 6-7:30PM), (7) Turn in a revised problem set (Fri. 6PM).


Mon.
Tues.
Wed.
Thur.
Fri.
Sat.
Sun.
2-4PM: Class
6-7:30PM: Section, Problem Sets due
Turn in revised problem.
8 AM: Lecture questions and reading annotations due online.

Keeping up with the weekly schedule is extremely important not only for your learning but for the rest of the class as well.

Outline and Lecture Notes After the foundational material is presented (roughly the first third of the class), I will introduce a large variety of statistical models and methods. I will choose these based on what makes sense from a pedagogical perspective at first, but as the semester goes on I will choose more and more material based on students interest and class projects.

For more information on the content of the class, see the detailed lecture notes online, which gives a general outline. Here's another version of some of the material:

Foundations

1. What is statistics?
2. What is political methodology?
3. Models and a language of inference
4. The role of simulation
   (a) To solve probability problems
   (b) to evaluate estimators
   (c) to compute features of probability distributions
   (d) to transform statistical results into quantities of interest
5. Stochastic components (normal, log-normal, Bernoulli, Poisson, etc)
6. The relationship between stochastic and systematic components and data generation processes
7. Systematic components (linear, logit, etc.)
8. Uncertainty and Inference
   (a) Probability as a model of uncertainty
   (b) Probability distributions, theory, discrete, continuous, examples
9. Inference
   (a) Inverse probability problems
   (b) The likelihood theory of inference
   (c) The Bayesian theory of inference
   (d) Detailed example: Forecasting presidential elections
10. Properties of maximum likelihood estimation (finite sample, asymptotic, etc.)
11. Precision of likelihood estimates

Specific Topics We will not get to all these topics, and the list of topics we do cover will likely include others than those listed here, depending on student interest.

1. Discrete regression models
   (a) Binary variables
   (b) Interpreting functional forms
   (c) Ordinal variables

   (d) Grouped uncorrelated binary variables

   (e) Event count models — Correlated and uncorrelated events; over and under dispersion.

2. Basic time series models
3. Basic multiple equation models, including identification
4. Multinomial choice models
5. Models for selection bias, censoring, and truncation
6. Models for duration
7. Hurdle models
8. Case-control designs
9. Model dependence
10. Matching as nonparametric preprocessing
11. Rare events
12. Compositional data
13. Missing data (item and unit nonresponse) problems
14. Text Analysis

References

Required

King, Gary. 1989. *Unifying Political Methodology: The Likelihood Theory of Statistical Inference.* Ann Arbor: University of Michigan Press.

A variety of papers will be assigned as well, available on the web.

Recommended

It is also helpful to have access to a book on R/S programming such as

Fox, John. 2002. *An R and S-Plus Companion to Applied Regression.* Sage Publications.

Imai, Kosuke, Gary King, and Olivia Lau. 2011. *Zelig: Everyone's Statistical Software*, Manuscript.

Ripley, Brian D. and Venables, William N. 2002. *Modern Applied Statistics with S*, Springer.

Suggested

Pawitan, Yudi. 2001. *In All Likelihood: Statistical Modelling and Inference Using Likelihood.* Oxford University Press

Barnett, Vic. 1982. *Comparative Statistical Inference.* 2nd edition. Wiley.

Chiang, Alpha. 1984. *Fundamental Methods of Mathematical Economics.* McGraw-Hill.

DeGroot, Morris H. 1986. *Probability and Statistics* Addison-Wesley. or Mendenhall, William and Robert J. Beaver. 1994. *Mathematical Statistics with Applications.* Duxbury.

Edwards, A.W.F. 1984. *Likelihood.* Cambridge University Press.

Gelman, Andrew et al. 2004. *Bayesian Data Analysis.* Chapman and Hall.

Gill, Jeff. 2008. *Bayesian Methods: A Social and Behavioral Sciences Approach*, 2nd ed, Chapman and Hall.

Harvey, Andrew C. 1990. *The Econometric Analysis of Time Series.* MIT Press.

Joreskog, Karl G. and Dag Sorbom, edited by Jay Magidson. 1979. *Advances in Factor Analysis and Structural Equation Models.* University Press of America.

King, Gary. 1997. *A Solution to the Ecological Inference Problem: Reconstructing Individual Behavior from Aggregate Data.* Princeton: Princeton University Press.

Kleppner, Daniel and Norman Ramsey. *Quick Calculus.* Wiley.

Lee J. Bain and Max Engelhardt. 1987. *Introduction to Probability and Mathematical Statistics.* Duxbury.

McCullagh, Peter and J. A. Nelder. 1993. *Generalized Linear Models* Chapman-Hall.

Mills, Terence C. 1990. *Time Series Techniques for Economists.* New York: Cambridge University Press.

Norman J. Johnson and Samuel Kotz. *Distributions in Statistics*, four volumes. John Wiley and Sons.

Rice, John A. 1995. *Mathematical Statistics and Data Analysis, 2nd Ed.* Belmont, CA: Duxbury Press.

Rubinsten, Reuven Y. 1981. *Simulation and the Monte Carlo Method*, New York: John Wiley.

Schafer, Joseph L. 1997. *Analysis of Incomplete Multivariate Data.* New York: Chapman-Hall.

Tanner, Martin A. 1996. *Tools for statistical inference: observed data and data augmentation methods*, 3rd edition. New York: Springer.