Gov 2000: 1. Introduction

Matthew Blackwell

Fall 2016

1/ Welcome and Motivation

Political methodology

- Political science: the systematic study of politics.
- Political methodology: the tools, techniques, and methods needed to make statistical or quantitative insights into politics.
  - Encompasses a wide variety of data types and approaches
  - Closely related to cognate fields: econometrics, sociological methods, psychometrics, biostatistics, etc.
  - Laid the groundwork for growth of data science (see Facebook/Google/OkCupid hiring)
  - A great community here at Harvard (IQSS) and beyond (Polmeth)

Why take this class?

1. Quantitative skills will make your research better.
    - Your research is judged on how convincing it is.
    - Statistics helps ensure and formalize credibility.
    - Overwhelming majority of top journal articles are quantitative.
    - You should never have to abandon a project because "you don't know how to do it."
2. Quantitative skills can get you a better job.
    - Quant literacy no longer optional.
    - Ceteris paribus, being cutting edge is a huge plus.
    - Hiring committees see potential for teaching, advising, and leadership.
3. Quantitative skills can answer big, substantive questions.

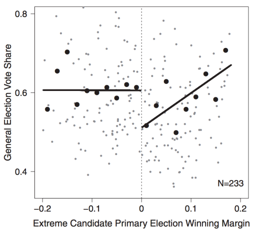
What is research?

1. Substance motivates a causal hypothesis:
    - H1: X causes Y
2. Substance and statistical theory motivate a research design:
    - How best to measure X and Y?
    - Where will variation in X and Y come from?
3. Design and statistical theory motivate analysis:
    - How best to estimate the relationship?
    - How best to assess the uncertainty of that relationship?
    - How best to present the results?

- Statistics guides us on all but the first question.
- Number 3 will be the focus of this class.

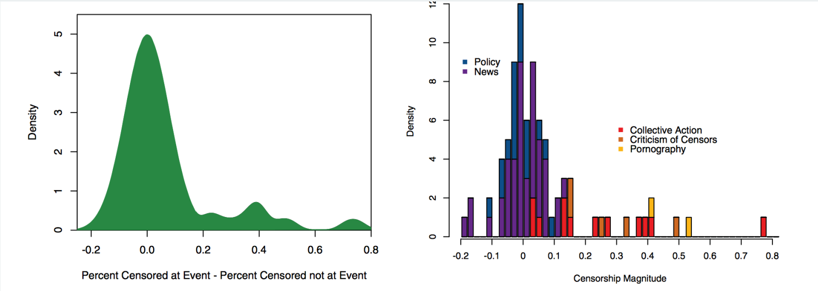
Methods tour: American

FIGURE 2. General-Election Vote Share After Close Primary Elections Between Moderates and Extremists: U.S. House, 1980–2010



- Andy Hall APSR paper
  - (Gov 2000 TF → Stanford)
- Do extremist candidates do better or worse in general election?
- Need to:
  1. measure extremism
  2. estimate the relationship
  3. determine if this is a causal.
- All of these are challenging!

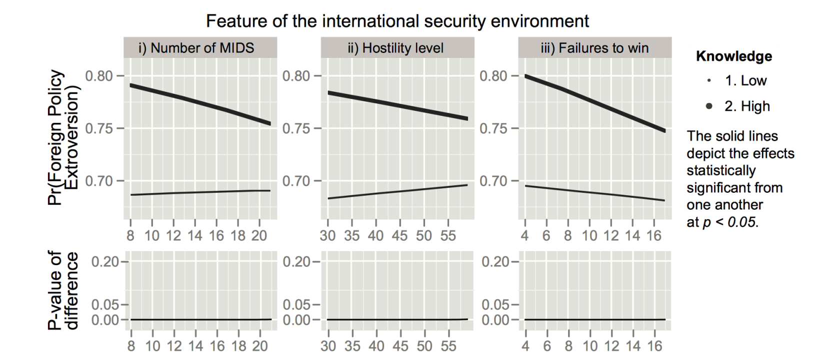
Methods tour: Comparative



(a) Distribution of Censorship Magnitude (b) Censorship Magnitude by Event Type

- Gary King, Molly Roberts, and Jen Pan APSR paper.
  - Roberts (Gov 2001 TF → UCSD)
  - Pan (Gov 2001 TF → Stanford)
- What types of messages do an authoritarian government try to censor?
- Use statistics to classify social media posts into topics.
- Use statistics to determine which topics were censored the most.

Methods tour: IR



- Josh Kertzer JoP paper.
- What are the determinants of foreign policy mood?
- Does political knowledge or the true security environment matter?
- Use statistics to see if we can determine such a relationship.

2/ Course Details

Staff

- Me: Matthew Blackwell
  - Office: CGIS K305
  - Email: `mblackwell@gov.harvard.edu`
  - Office Hours: W, 2-4pm or stop by whenever I'm in and the door is open.
  - Google chat: `mblackwell@gmail.com`
- Your TFs: they are your sage guides for everything in this class.
  - Mayya Komisarchik (`mkomisarchik@fas.harvard.edu`), G4 in the Gov Department
  - David Romney (`dromney@fas.harvard.edu`), G4 in the Gov Department

Course numbers

- Gov 2000: main course number for Gov PhD students
- Gov 2000e: alternative course number for Gov PhD students who never plan to read any empirical political science.
- Gov 1000: main course number for undergraduates.
- Stat E-190: course number for extension school students
- All course numbers will use some R.
- Some course material will be tailored to Gov 1000, Gov 2000e, and Stat E-190 undergrad credit.

Prerequisites

- All course numbers requires:
  - Knowledge of basic algebra and some exposure to basic statistics.
- Graduate-level credit requires some exposure to:
  - Calculus (limits, derivatives, integrals)
  - Linear algebra (vectors, matrices, etc)
  - Basic probability (probability axioms, joint/conditional probability, etc)
  - Basically what's covered in Gov Math Prefresher (see syllabus for link)
- Talk to us if you want resources!

Why so much math?

- Methods popular since I started grad school:
  - Text-as-data, machine learning, Bayesian nonparametrics, design-based inference, network analysis, and so many more.
- I wouldn't be able to learn or use any of those methods without a strong foundation in rigorous statistics.
- You will be using methods for the rest of your career $\rightsquigarrow$ you best invest!
  - Understanding your tools will make you better at your craft.

How much time?

- The first year of grad school is a marathon:
  - Past students spent 5–20 hours per week on the HWs alone.
  - This can be painful, but it is completely normal
- Everyone starting at a Top-10 PhD program is doing that and probably more.
- Success in academia is a combination of creativity and consistent hard work
  - Working hard on methods will give you the ability to be as creative as possible.

Computation

- We'll use R for statistical computing.
  - It's free
  - It's becoming the de facto standard in many applied statistical fields
  - It's extremely powerful, but relatively simple to do basic stats
  - Compared to other options (Stata, SPSS, etc) you'll be more free to implement what you need (as opposed to what Stata thinks is best)
- Will use it in lectures, much more help with it in sections

Teaching resources

- Lecture (where we will cover the broad topics)
- Sections (where you will get more specific, targeted help on assignments)
- Canvas site (where you'll find the syllabus, upload your assignments, and where you can ask questions and discuss topics with us and your classmates)
- Office hours (where you can ask even more questions)

Textbook

- Wooldridge, Introductory Econometrics: A Modern Approach, 5th edition.
- Any edition is fine, though you might want to check the reading list more carefully.
- Lecture notes will be other main text.

Grading

- Weekly homework assignments (50%)
- Take-home midterm exam (10%)
- Cumulative take-home final (30%)
- Participation (10%)
- PhD students: grades don't matter.

Outline of topics

- The basic outline of our semester, in backwards order:
  - **Regression**: how to determine the relationship between variables.
  - **Inference**: how to learn about things we don't know (the relationship b/w two variables) from the things we do know (the observed data).
  - **Probability**: what data we would expect if we did know the truth.
- Probability $\rightarrow$ Inference $\rightarrow$ Regression

3/ Overview of Probability and Statistics

What is statistics?

- It is branch of mathematics that studies the collection and analysis of data.
- The name statistic comes from the word state.
- Assume events are stochastic rather than deterministic.
- Model these stochastic events using probability.

Deterministic versus stochastic

- One idea that unites all of these questions in statistics is variation and uncertainty. What do we mean by this?
- Imagine someone comes to us and says, "what is the relationship between voter turnout and campaign spending?"
- Deterministic account of voter turnout in a district:

$$\text{turnout}_i = f(\text{spending}_i).$$

- What's the problem with this? Omits all other determinants:
  - open seat, challenger quality, weather on election day, having the local college football team win the previous weekend, whether or not Jimmy had to stay home sick from school

Stochastic models

- Measure everything and then add it to our model:

$$\text{turnout}_i = f(\text{spending}_i) + g(\text{stuff}_i).$$

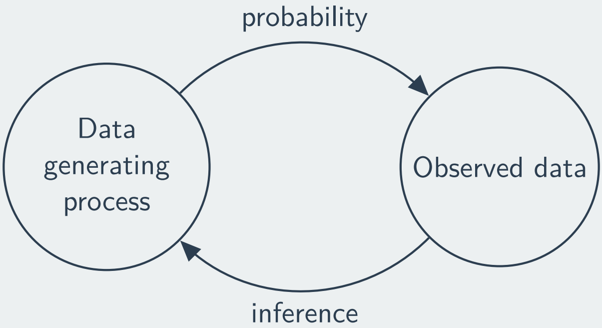
- Treat other factors as direct interest as stochastic:
    - They affect the outcome, but are not of direct interest.
    - We think of them as part of the natural variation in turnout.
- The word “stochastic” comes from the Greek word for the target that archers are supposed to shoot at.
- We know roughly where the arrows are going to fall, but not exactly where any particular arrow will be.
- Stochastic = chance variation

The error term

- When we do this, we often write this as:

$$\text{turnout}_i = f(\text{spending}_i) + u_i.$$

- Here, u_i is the error or disturbance term.
- Stochastic term represents all factors that affect turnout.
- Need some way of quantifying stochastic outcomes: probability.



Why probability?

- Next few weeks: probability.
  - Not a punishment.
  - Probability helps us study stochastic events.
  - Important for all of statistics.
- Statistical inference is a thought experiment.
- Probability is the logic of these though experiments.
- Suppose men and women were paid the same on average, but there was chance variation from person to person.
  - How likely is the observed wage gap in this hypothetical world?
  - What kinds of wage gaps would we expect to observe in this hypothetical world?
- Probability to the rescue!

The lady tasting tea

- Thought experiment posed by statistician R.A. Fisher.
  - "a genius who almost single-handedly created the foundations for modern statistical science"
- Setup of thought experiment:

  Your advisor asks you to grab a tea with milk for him before your meeting and he says that he prefers tea poured before the milk. You stop by Darwin's and ask for a tea with milk. When you bring it to your advisor, he complains that it was prepared milk-first.

- You are skeptical that he can really tell the difference, so you devise a test:
  - Prepare 8 cups of tea, 4 milk-first, 4 tea-first
  - Present cups to advisor in a random order
  - Ask advisor to pick which 4 of the 8 were milk-first.

Assuming we know the truth

- Advisor picks out all 4 milk-first cups correctly!
- Statistical thought experiment: how often would she get all 4 correct if she were guessing randomly?
  - Only one way to choose all 4 correct cups.
  - But 70 ways of choosing 4 cups among 8.
  - Choosing at random ≈ picking each of these 70 with equal probability.
- Chances of guessing all 4 correct is $\frac{1}{70} \approx 0.014$ or 1.4%.
- $\rightsquigarrow$ the guessing hypothesis might be implausible.
- You've done your first hypothesis test and calculated your first p-value!

Sample spaces and events

- A sample space Ω is the set of possible outcomes.
- $\omega \in \Omega$ is one particular outcome.
- A subset of Ω is an event and we write this as $A \subset \Omega$.
- Lady tasting tea:
    - Sample space is all the unordered ways that the advisor could choose 4 cups from the 8 available:

$$\Omega = \{1234, 1235, 1236, \dots, 5678\}.$$

The axioms of probability

- Probability quantifies how likely or unlikely events are.
- We'll define the probability $\mathbb{P}(A)$ with three axioms:

1. Nonnegativity: $\mathbb{P}(A) \geq 0$ for every event A
2. Normalization: $\mathbb{P}(\Omega) = 1$
3. Additivity: If a series of events, $A_1, A_2, \ldots, A_N$, are mutually exclusive, then the probability of any of the events is the sum of the probabilities of each event:

$$\mathbb{P}(A_1 \cup A_2 \cup \ldots \cup A_N) = \sum_{i=1}^{N} \mathbb{P}(A_i).$$

    - $\mathbb{P}$(1 or 2 cups correct) = $\mathbb{P}$(exactly 1 correct) + $\mathbb{P}$(exactly 2 correct)
    - With additivity, we can let N go to infinity.

Data generating process

- Probabilities often derived from assumptions about how the data came to be.
  - Often refer to this as the data generating process or DGP.
- DGP: flipping a fair coin $\rightsquigarrow \mathbb{P}(H) = 0.5$.
- Random sampling/selection: each outcome has equal probability.
- Example: randomly select a card from a standard deck of playing cards
  - Each of the 52 card has equal probability

$$\mathbb{P}(4\clubsuit) = \mathbb{P}(4\heartsuit) = 1/52$$

- Goal: learn about the DGP

4/ Basic Descriptive Statistics

Let's play with some data

- Data from Fulton County, GA with all registered voters.

```
## load file of all registered voters
load("../data/fulton.RData")
```

```
## size of the dataset
nrow(fulton)
```

```
## [1] 339186
```

```
## how many democrats are there
table(fulton$dem)
```

```
##
##      0      1
## 242178  97008
```

Peeking at the data

- What does the data look like?

```
## print the first few rows
fulton[1:5, ]
```

```
##   turnout black sex age dem rep urban percblk lvbdist
## 1       0     0   1  19   0   0     0  0.0523  3.4836
## 2       0     0   0  35   0   0     0  0.0288  3.2913
## 3       0     1   0  36   0   0     1  0.9924  2.8767
## 4       1     0   0  27   0   0     1  0.1112  2.5618
## 5       1     1   1  79   1   0     1  0.9923  2.7935
##   school firest church
## 1      0      0      1
## 2      1      0      0
## 3      1      0      0
## 4      0      0      0
## 5      1      0      0
```

Sample mean

- Let X_i be the age of the ith person in the data.
- Let n is the number of people in the data.
- Sample mean (or sample average): $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$
  - Sum of the values divided by the number of values.
- Describes the center of the data—what is a typical value in this sample.

Sample mean in R

- First, useful to see the ages of the first few observations:

```
fulton[1:5, "age"]
```

```
## [1] 19 35 36 27 79
```

- Now we can calculate the mean "by hand":

```
sum(fulton[, "age"])/nrow(fulton)
```

```
## [1] 42.3608
```

- Or we can use a handy R function:

```
mean(fulton[, "age"])
```

```
## [1] 42.3608
```

Sample variance

- Also want to get a sense of the spread around the center.
- Sample variance: $S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$
  - Measures how far, on average, people are from the sample mean.
  - Divide by $n - 1$ instead of n to ensure S^2 is unbiased (we'll see what this means)
- In R:

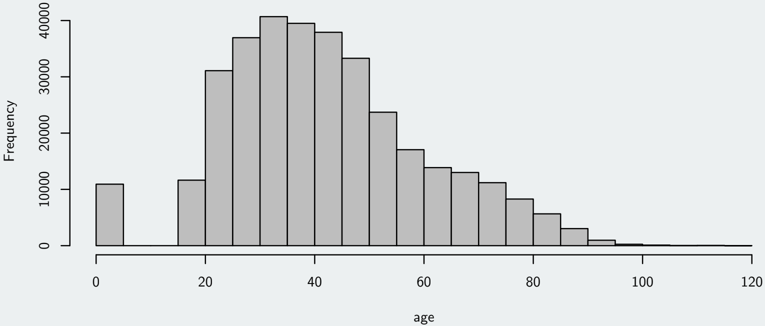
```
## sample variance of age
var(fulton[, "age"])
```

```
## [1] 331.1574
```

Visualizing the distribution

- How can we look at the distribution ages in the data?
- Histogram: height of bar = frequency of bin:

```r
hist(fulton[,"age"], col = "grey", xlab = "age", main = "")
```



Why means and variances?

- The sample mean and the sample variance help describe the data we have.
  - This is called descriptive inference.
- But they can also tell us about the data we don't have—those people not in the sample.
  - This is called statistical inference.
- If we have a sample from some population, how can we learn about the population?
- What can we learn about the average age in the population from the sample mean?
- Need to learn probability before we can answer these questions!

Next few weeks

- Random variables
- How to conceptualize observed data as probabilistic quantities.
- Probability distributions, means, variances, etc.
- Some calculus next week so brush up on integrals.