Design and Analysis of Sample Surveys

Andrew Gelman
Department of Statistics and Department of Political Science
Columbia University

Class 14b: Review

Class 1a: Introduction

- Goal: learning about the population
- Intermediate steps:
  - Sample to population (sampling)
  - Survey response to question of interest (measurement)

Happiness and the Tea Party movement

- A Brooks *New York Times* op-ed:

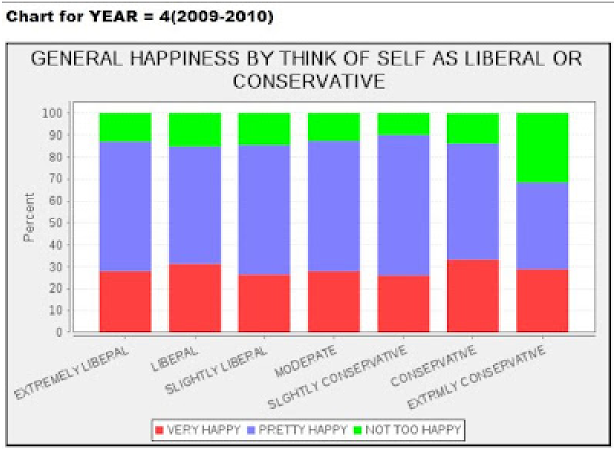
  > *People at the extremes are happier than political moderates . . . . none, it seems, are happier than the Tea Partiers . . .*

- But sociologist Jay Livingston writes:

  > *The GSS does not offer "bitter" or "Tea Party" as choices, but extreme conservatives are nearly three times as likely as others to be "not too happy."*

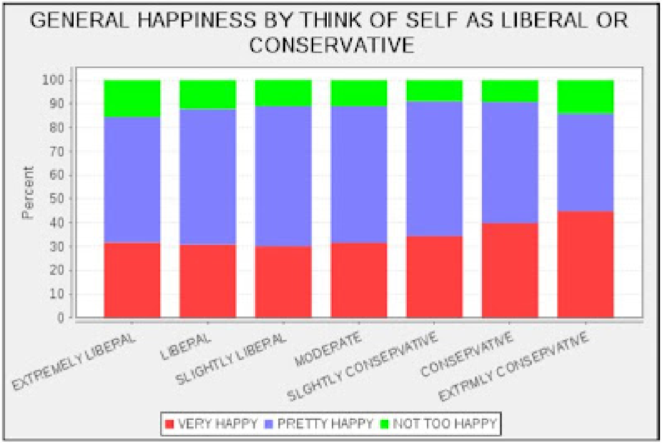
- Let's look at the data!

Data from General Social Survey



- Is this just sampling variation?
  - Sample size for “Extremely Conservative” here is 80
  - Thus the standard error for that green bar on the right is approx $\sqrt{0.3 \cdot 0.7/80} = 0.05$

How did Brooks get this wrong?



- Averaging over all the years, conservatives seem pretty happy!
- The importance of descriptive inference
    - Be careful about explaining patterns that aren't real!

Class 1b: Statistical inference and linear regression

- $\sqrt{p(1-p)/n}$ or $\sigma/\sqrt{n}$
- $\sqrt{\sigma_1^2/n_1+\sigma_2^2/n_2}$
- Sample size calculations
- $(y+2)/(n+4)$
- Weighted averages
- Living with uncertainty

Linear regression

- Interpreting coefficients
- Building models
- The role of statistical significance

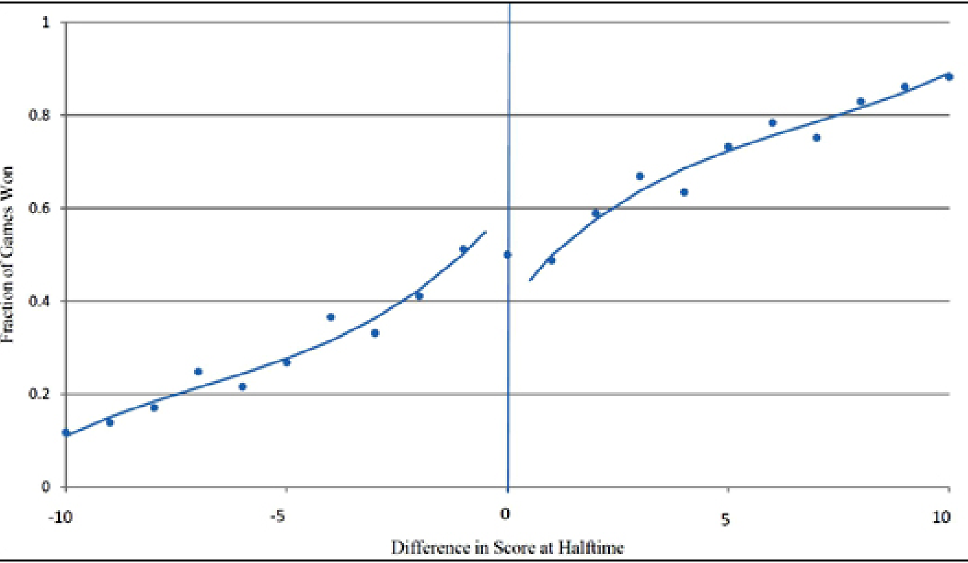
Class 2a: Logistic regression

- Logistic curve
- Divide-by-4 rule
- Latent continuous variable
- Choice models

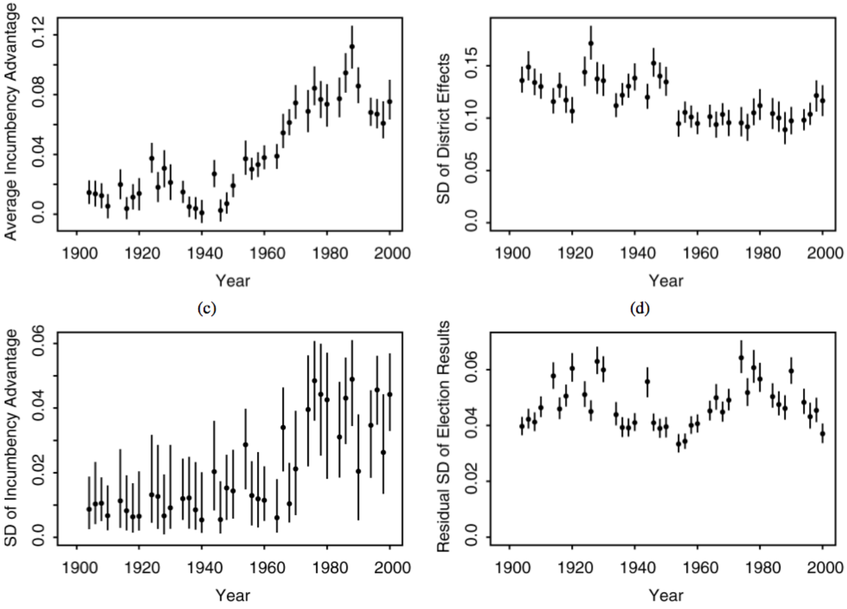
Class 2b: The challenge of estimating small effects

- The statistical significance filter
- When possible, study large effects
- Study effects in context

Business-relevant examples . . .



Example: incumbency advantage over time



(c)

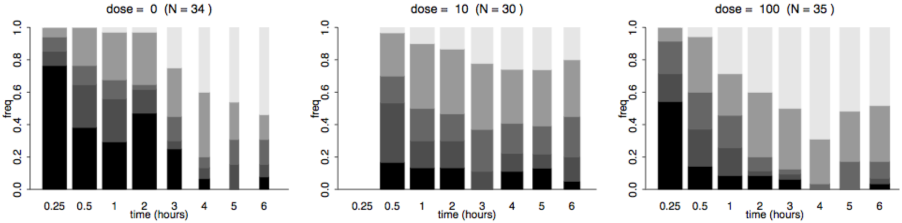
(d)

Class 3a: Nonresponse and survey adjustment

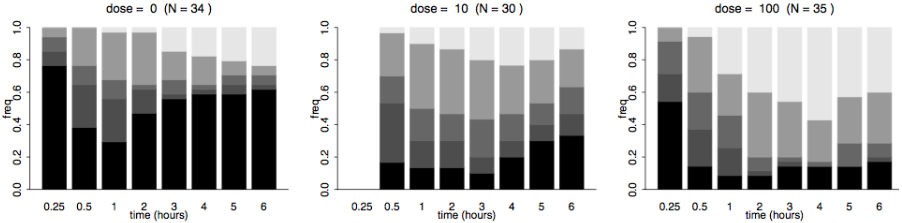
- Unit nonresponse
- Item nonresponse
- Intentional missing data
- Structural missing data

Observed and completed data

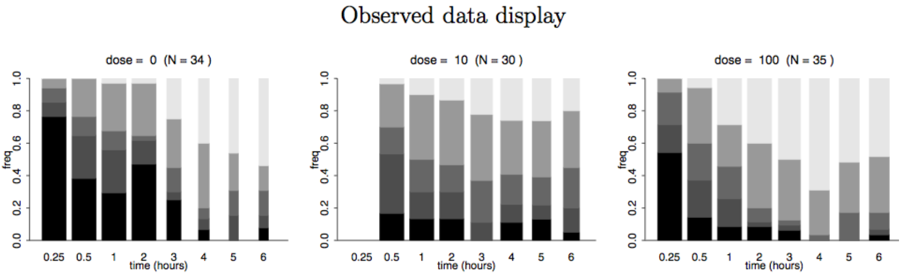
Observed data display

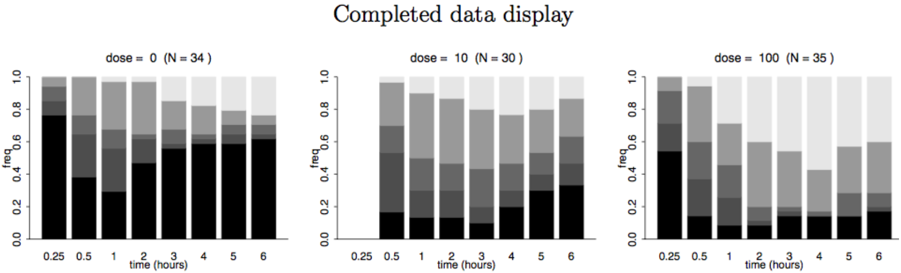


Completed data display



Observed and completed data





Class 3b: Adjusting for nonresponse

- Subsetting and imputation for item nonresponse
- Poststratification and weighting for unit nonresponse

Class 4a: Weighting and poststratification

- Where do weights come from?
- Poststratification identity: $\theta = \frac{\sum_j N_j \theta_j}{\sum_j N_j}$
- Challenges in weighting
- Challenges in poststratification
- Weighted regression using the "survey" package in R

Example: CBS/New York Times pre-election polls

```
  id     org  y state edu age adults weight
6140  cbsnyt NA     7   3   1      2    923
6141  cbsnyt  1    39   4   2      2    558
6142  cbsnyt  0    31   2   4      1    448
6143  cbsnyt  0     7   3   1      2    923
6144  cbsnyt  1    33   2   2      1    403
```

- The weight is listed as just another survey variable
- But they are actually constructed *after* the survey
- Weights $w_i = g(X_i, \theta)$

Class 4b: Ratio and regression estimation

- Ratio estimation of a ratio
- Ratio estimation of a population average
- Regression estimation
- Robustness through model and design

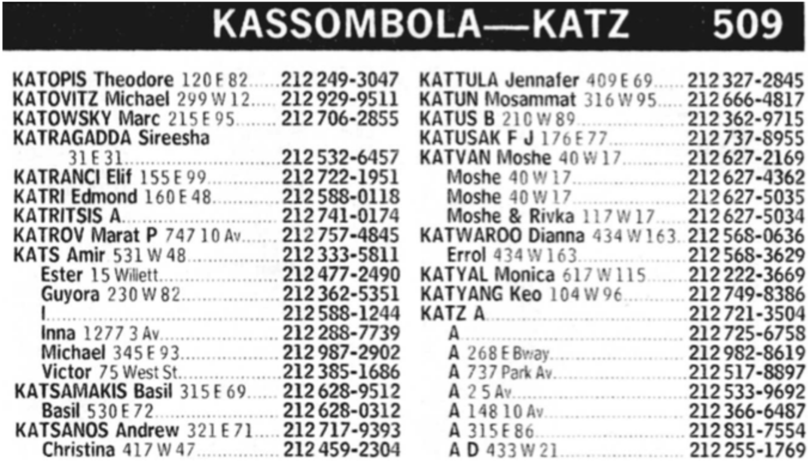
Regression estimation as a general framework

- Fit a regression, $y_i = a + bx_i + \text{error}$
- Regression estimate of $\overline{Y}$ is $\overline{y} + b\,(\overline{X} - \overline{x})$
- Special cases:
  - $b = 0$: unadjusted sample average
  - $b = \frac{\overline{y}}{\overline{x}}$: ratio estimation
  - $b = 1$: simple adjustment
- Regression estimation is valid for any b
  - Optimal for $b =$ least-squares estimate

Class 5a: Simple and stratified random sampling

- Sampling from a list
- Systematic sampling
- Stratified sampling
- Design and analysis

(A bit of) the population



KASSOMBOLA—KATZ 509

KATOPIS Theodore 120 E 82	212 249-3047
KATOVITZ Michael 299 W 12	212 929-9511
KATOWSKY Marc 215 E 95	212 706-2855
KATRAGADDA Sireesha 31 E 31	212 532-6457
KATRANCI Elif 155 E 99	212 722-1951
KATRI Edmond 160 E 48	212 588-0118
KATRITSIS A	212 741-0174
KATROV Marat P 747 10 Av	212 757-4845
KATS Amir 531 W 48	212 333-5811
Ester 15 Willett	212 477-2490
Guyora 230 W 82	212 362-5351
I	212 588-1244
Inna 1277 3 Av	212 288-7739
Michael 345 E 93	212 987-2902
Victor 75 West St	212 385-1686
KATSAMAKIS Basil 315 E 69	212 628-9512
Basil 530 E 72	212 628-0312
KATSANOS Andrew 321 E 71	212 717-9393
Christina 417 W 47	212 459-2304
KATTULA Jennafer 409 E 69	212 327-2845
KATUN Mosammat 316 W 95	212 666-4817
KATUS B 210 W 89	212 362-9715
KATUSAK F J 176 E 77	212 737-8955
KATVAN Moshe 40 W 17	212 627-2169
Moshe 40 W 17	212 627-4362
Moshe 40 W 17	212 627-5035
Moshe & Rivka 117 W 17	212 627-5034
KATWAROO Dianna 434 W 163	212 568-0636
Errol 434 W 163	212 568-3629
KATYAL Monica 617 W 115	212 222-3669
KATYANG Keo 104 W 96	212 749-8386
KATZ A	212 721-3504
A	212 725-6758
A 268 E Bway	212 982-8619
A 737 Park Av	212 517-8897
A 2 5 Av	212 533-9692
A 148 10 Av	212 366-6487
A 315 E 86	212 831-7554
A D 433 W 21	212 255-1769

Class 5b: Cluster sampling with equal cluster sizes

- Why do cluster sampling?
- Goal of equal-probability sampling
- Analysis of cluster data
- Design effects

Class 6a: Cluster sampling with unequal cluster sizes

- Sampling with equal probability at both stages
- Sampling clusters with probability proportional to size
- Adjusting for unequal sampling probabilities
- Design effects

Class 6b: Inference for regression coefficients

- Option 1: weighted regression
- Option 2: unweighted regression, including in the model all variables that affect the probability of inclusion in the survey
- Discuss
- Population mean as a special case of a regression coefficient
- Population difference as a special case of a regression coefficient
- Practical limitations of weighting
- Practical limitations of modeling

Putting it all together

- Our ideal procedure:
  - As easy to use as hierarchical regression
  - Population info included using poststratification
- Smooth transition from classical weighting
  - Equivalent weights
  - When different methods give different results, we can track it back to an interaction

Class 7a: Survey interviewing

- Questions and answers in surveys
- Evaluating survey questions
- Survey interviewing
- Surveys vs. other sources of information (administrative data, economic activity, . . .)
- Conceptual or specification errors
- Sampling and nonsampling errors
- Errors of measurement, interviewers, question wording, . . .
- Errors in reporting

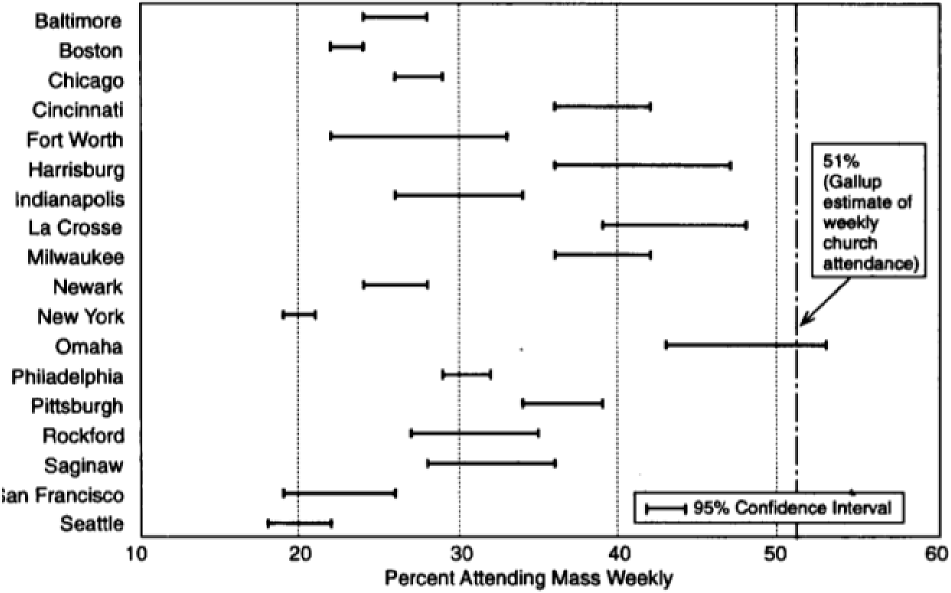
Class 7b: Challenges in survey measurement

- Difficulty of estimating small percentages
- Framing and question wording

Estimating measurement effects

- You are conducting a survey and are concerned about the possible effects of the wording of one particular question. You decide to do one of two experiments:
  - (a) Within-subject design: Put the two different wordings on the same survey form (randomizing the order of the two questions) and compare responses to the two wordings.
  - (b) Between-subject design: Randomly give one wording to half the respondents and the other wording to the other half. Compare the average responses under the two wordings.
- Give a reason why you might prefer design (a).
- Give a reason why you might prefer design (b).
- Computing standard errors for different designs

Measurement: Church attendance

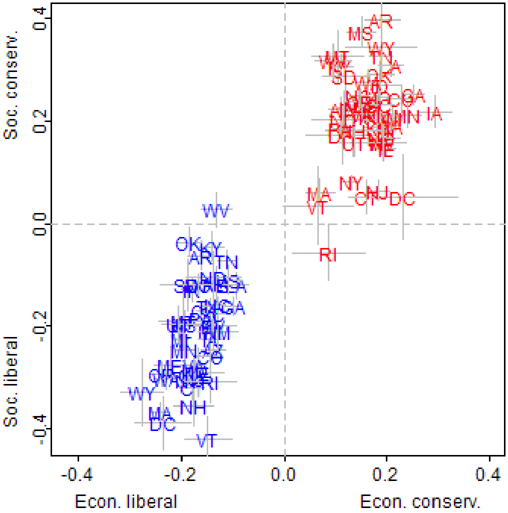


Class 8a: Using surveys to answer questions in political science

- Political attitudes and behavior
- Comparing different groups

Democrats and Republicans separately in 2000

Average economic and social ideology scores
among Bush voters (red) and Gore voters (blue) in each state



Class 8b: Conducting a survey in the real world

- Research goals
- Population, frame, and sampling design
- Constructing and testing the survey instrument
- Sampling and data collection
- Collection of auxiliary data
- Data cleaning and manipulation
- Data analysis

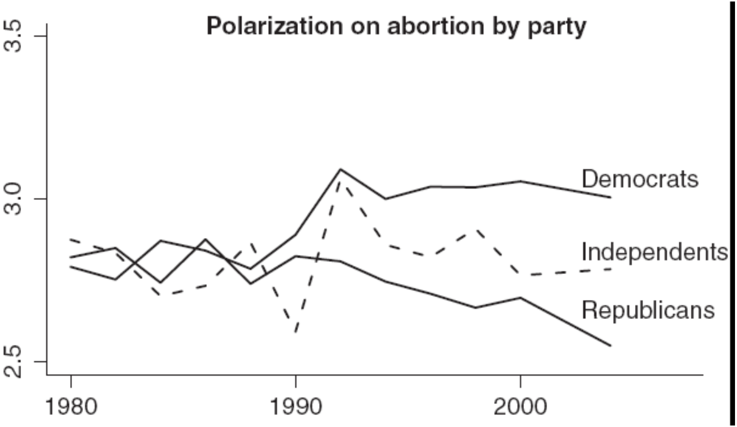
Class 9a: Voting

- Survey questions on demographics
- Political affiliation
- Issue attitudes

Class 9b: Public opinion

- Partisan polarization
- Elite and mass attitudes
- Uniform partisan swing

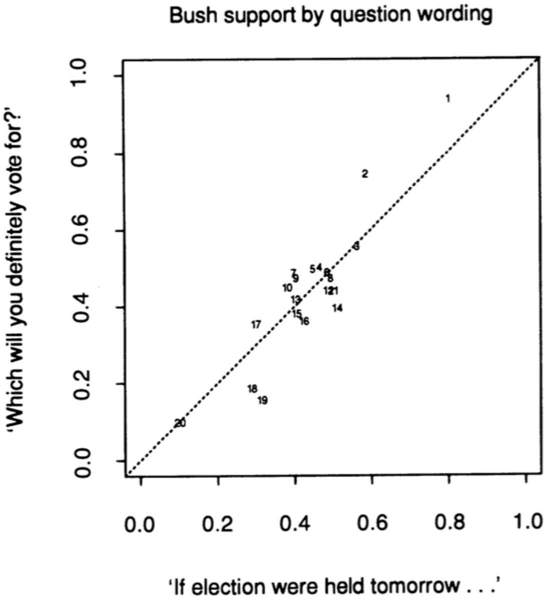
Political polarization since 1990



Class 10a: Political participation

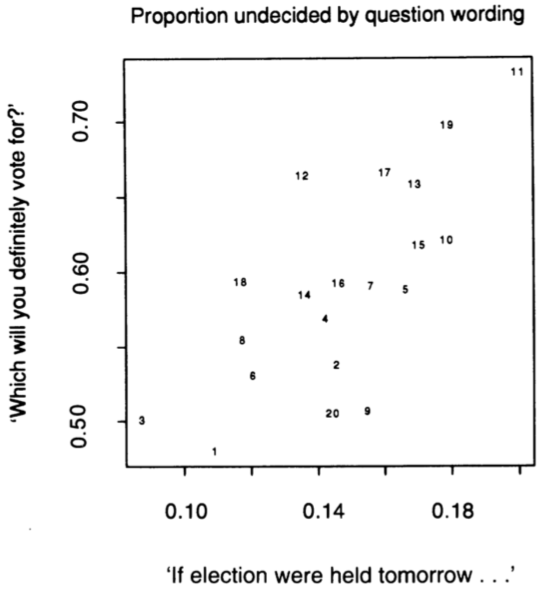
- Asking about voting
- Other forms of political participation
- Rationality of voting and responding to surveys
- Data sources

Question wording and vote intention

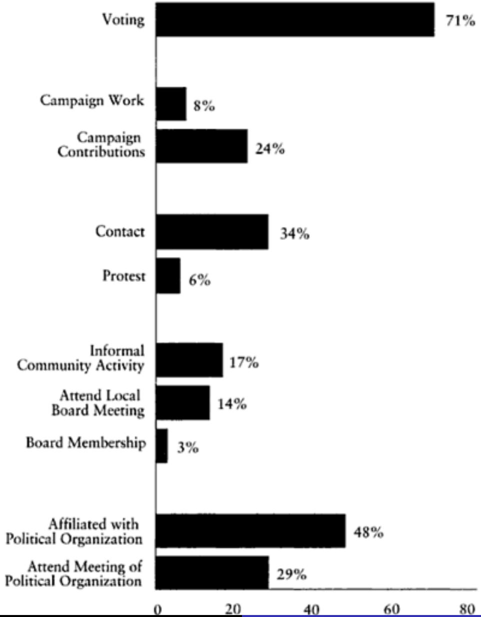


Bush support by question wording

Question wording and nonresponse



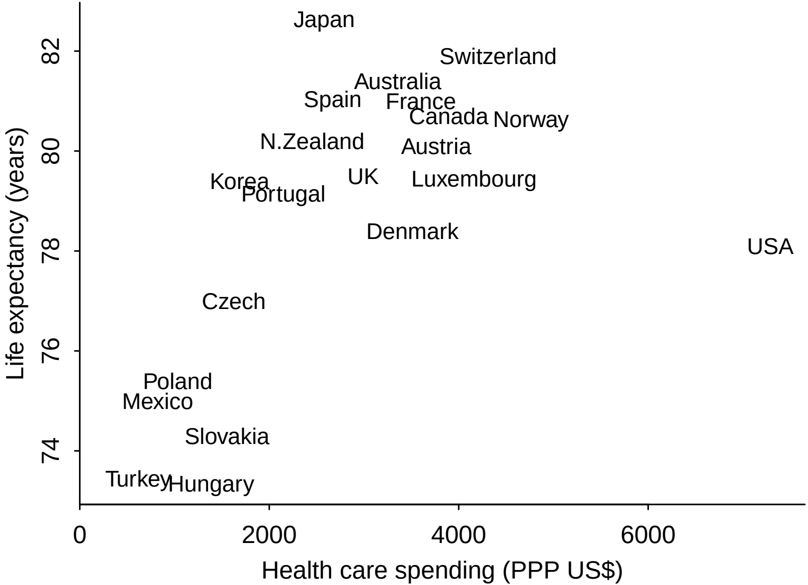
From Verba, Schlozman, Brady, *Voice and Equality* (1995)



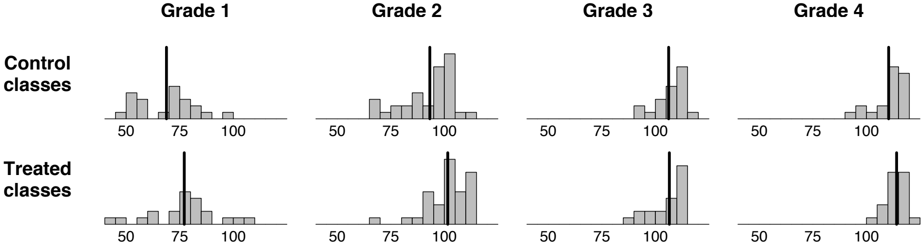
Class 10b: Understanding and displaying data

- Manipulating data in R
- Looking carefully at the data
- Effective graphing

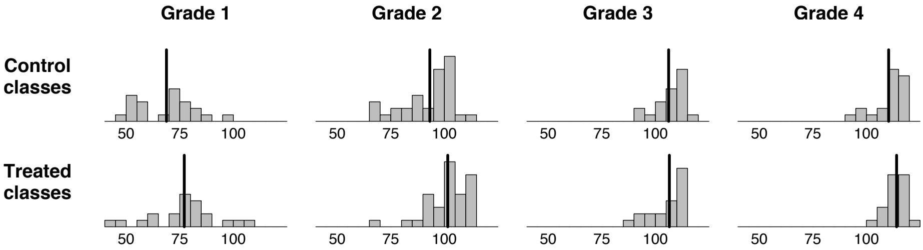
Real simple


Japan
Switzerland
Australia
Spain
France
Canada
Norway
N.Zealand
Austria
Korea
UK
Luxembourg
Portugal
Denmark
USA
Czech
Poland
Mexico
Slovakia
Turkey
Hungary
82
80
78
76
74
Life expectancy (years)
0
2000
4000
6000
Health care spending (PPP US$)

Re-expression saves space and adds clarity



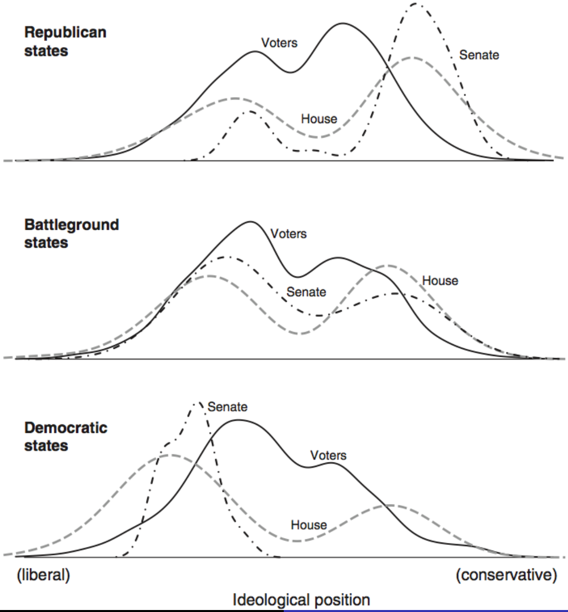
Re-expression saves space and adds clarity



Class 11a: Bayesian inference

- Calibration of probabilities
- Combining prior and data information
- Example: forecasting
- MRP

Aligning voters with Congress



Class 11b: Ideal-point modeling

- Applications: ability testing, ranking
- Compare to ideal-point modeling
- Implications for education and for data collection more generally

Class 12a: Multilevel regression and poststratification

- State-level opinions
- Comparisons to state policies
- Demographic breakdowns
- Displaying inferences

Class 12b: Challenges in multilevel regression and poststratification

- Many factors, deep interactions
- Fitting and understanding models
- Adjusting for non-census variables
- Differential nonresponse within cells

Class 13a: Low response rates in U.S. surveys

- When does it matter?
- Validation of survey data
- Methods for increasing response rates

Class 13b: Surveys in less-developed countries

- Iraq mortality survey

Class 14a: Network sampling

- Fractal sampling
- Penumbra sampling
- Learning about networks
- Averaging over networks