Design and Analysis of Sample Surveys

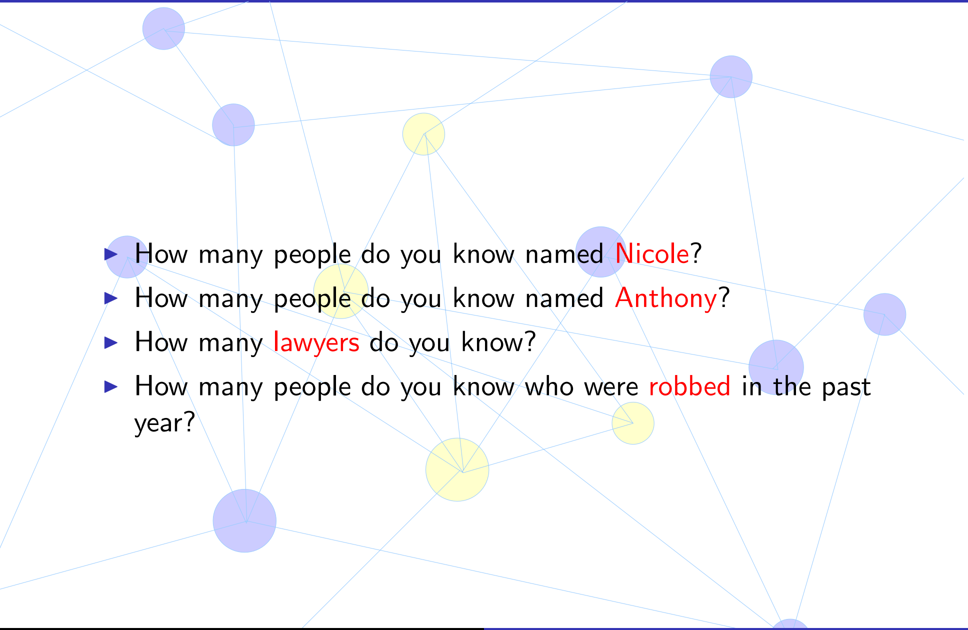
Andrew Gelman
Department of Statistics and Department of Political Science
Columbia University

Class 14a: Network sampling

Ronald Reagan was a statistician

"But the simplest way for each of you to judge recovery lies in the plain facts of your own individual situation. Are you better off than you were last year?"

Who do you know?



- How many people do you know named Nicole?
- How many people do you know named Anthony?
- How many lawyers do you know?
- How many people do you know who were robbed in the past year?

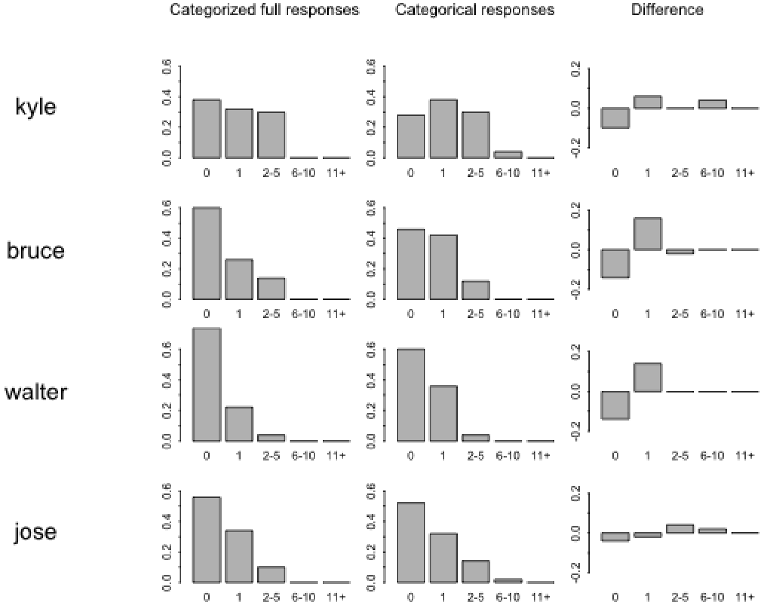
Scale-up method

- On average, you knew 0.6 Nicoles
- 0.13% of Americans are named Nicole
- Assume 0.13% of your acquaintances are Nicoles
- Estimate: on average, you know $0.6/0.0013 = 450$ people

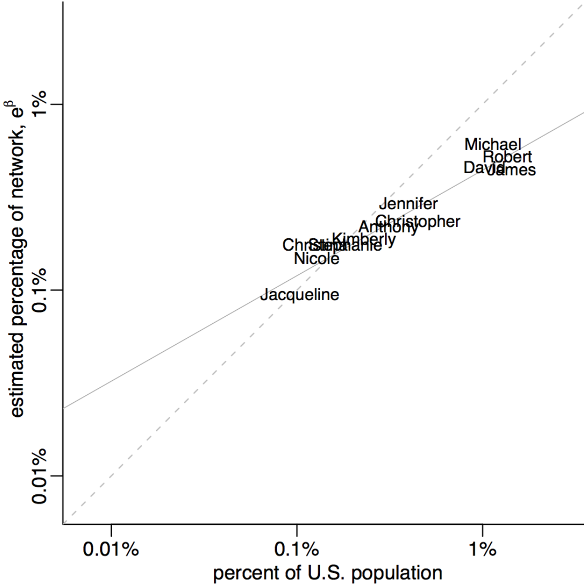
- On average, you know 0.8 Anthonys
- 0.31% of Americans are named Anthony
- Estimate: on average, you know $1.6/0.0031 = 260$ people

- Why do these differ?

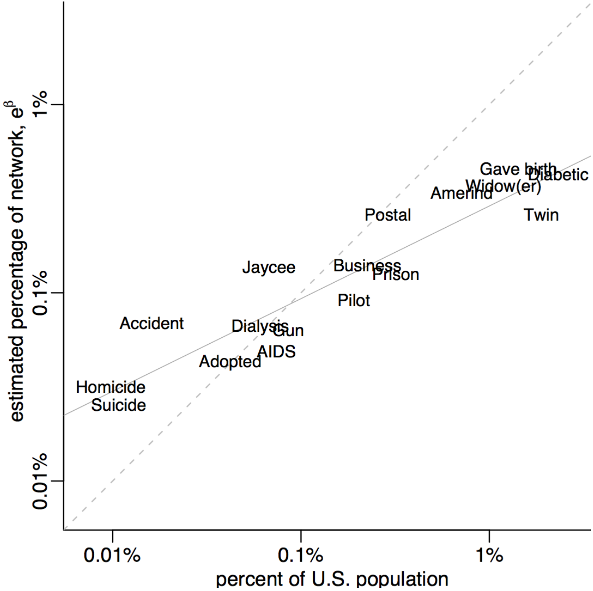
Question wording for “How many X’s” surveys



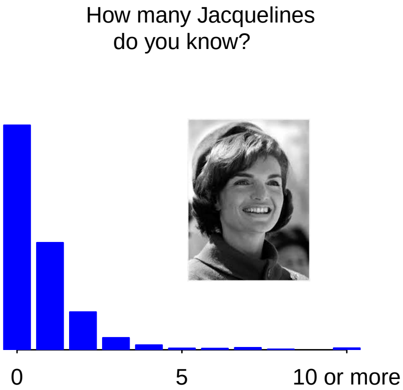
Perception vs. reality: names

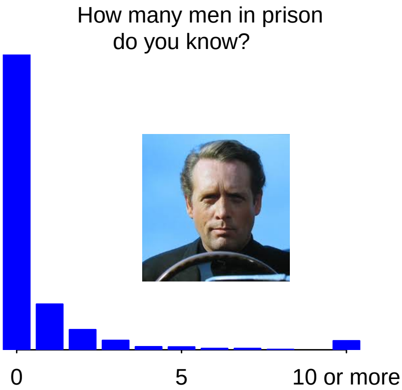


Perception vs. reality: groups

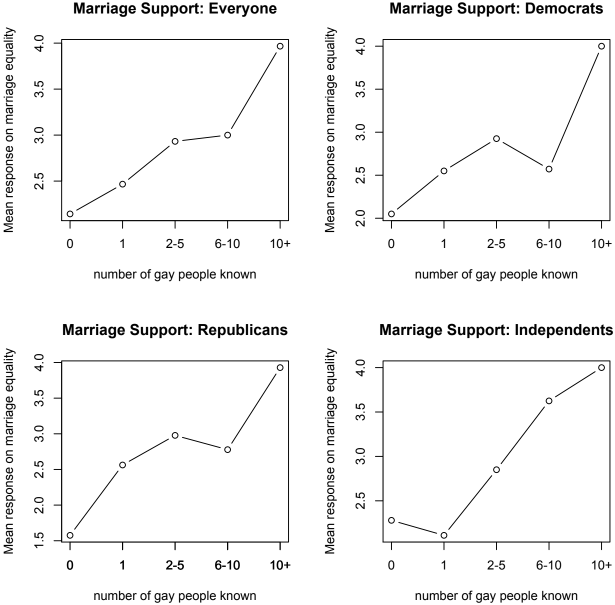


Social structure!


How many Jacquelines do you know?


How many men in prison do you know?

The gay penumbra


Marriage Support: Everyone
Marriage Support: Democrats
Marriage Support: Republicans
Marriage Support: Independents
Mean response on marriage equality
number of gay people known
0
1
2-5
6-10
10+
1.5
2.0
2.5
3.0
3.5
4.0

Penumbra sampling: goals

- Learning about small groups (getting to the "penumbra")
- Relating networks to attitudes and behavior (gay rights, foreign policy, etc.)
- Learning about people who are not in the sample

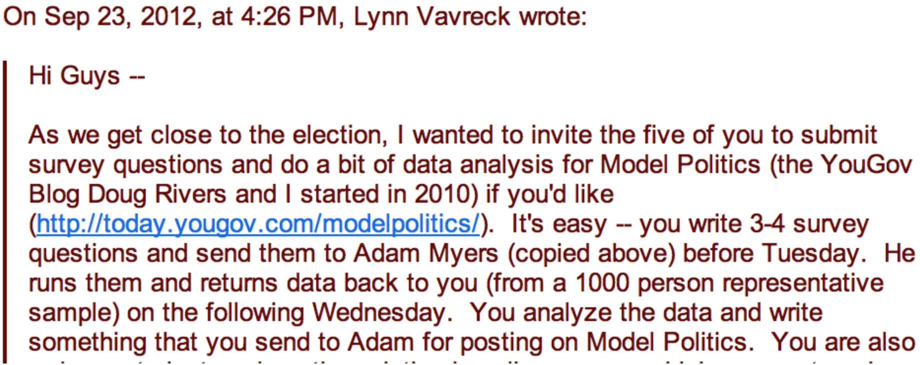
Penumbra sampling: methods

- Network sampling, snowball sampling, referral sampling, . . .
- Hypernetwork sampling (the National Congregations Study)
- Questions about social networks (to learn about the survey respondent)
- Questions about attributes of friends, neighbors, etc. (to learn about others)
- Learning about unreachable groups (children, prisoners, dead people, surly people)
- Identifying penumbras of small groups (homeschoolers, Jews, . . .)

Polarization during the 2012 election campaign

- An opportunity
- Some survey questions
- Data analysis

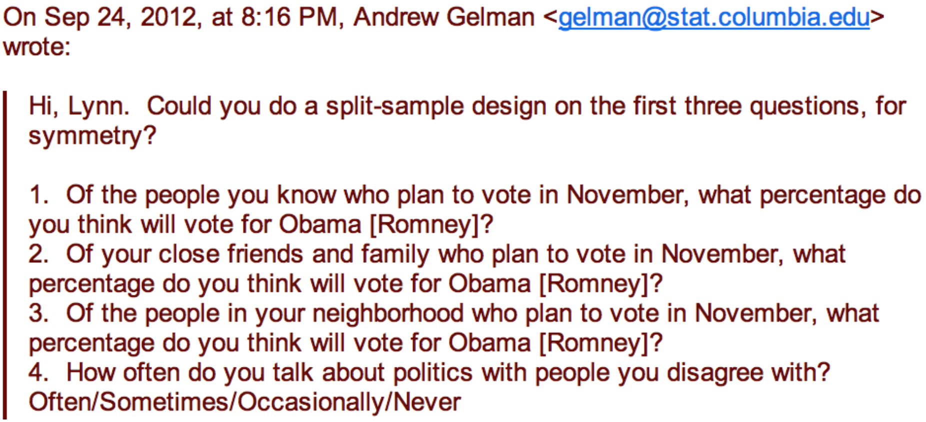
An opportunity



On Sep 23, 2012, at 4:26 PM, Lynn Vavreck wrote:

> Hi Guys --
>
> As we get close to the election, I wanted to invite the five of you to submit survey questions and do a bit of data analysis for Model Politics (the YouGov Blog Doug Rivers and I started in 2010) if you'd like (http://today.yougov.com/modelpolitics/). It's easy -- you write 3-4 survey questions and send them to Adam Myers (copied above) before Tuesday. He runs them and returns data back to you (from a 1000 person representative sample) on the following Wednesday. You analyze the data and write something that you send to Adam for posting on Model Politics. You are also

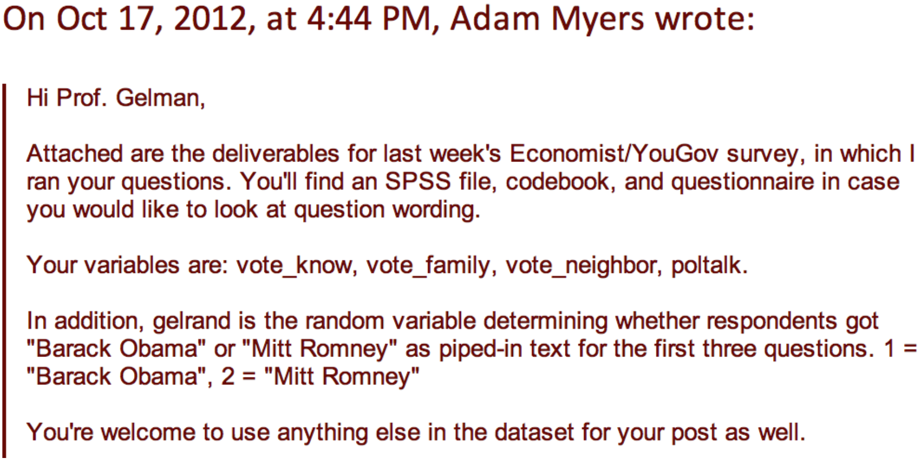
Some survey questions



On Sep 24, 2012, at 8:16 PM, Andrew Gelman <gelman@stat.columbia.edu> wrote:

> Hi, Lynn. Could you do a split-sample design on the first three questions, for symmetry?
>
> 1. Of the people you know who plan to vote in November, what percentage do you think will vote for Obama [Romney]?
> 2. Of your close friends and family who plan to vote in November, what percentage do you think will vote for Obama [Romney]?
> 3. Of the people in your neighborhood who plan to vote in November, what percentage do you think will vote for Obama [Romney]?
> 4. How often do you talk about politics with people you disagree with?
> Often/Sometimes/Occasionally/Never

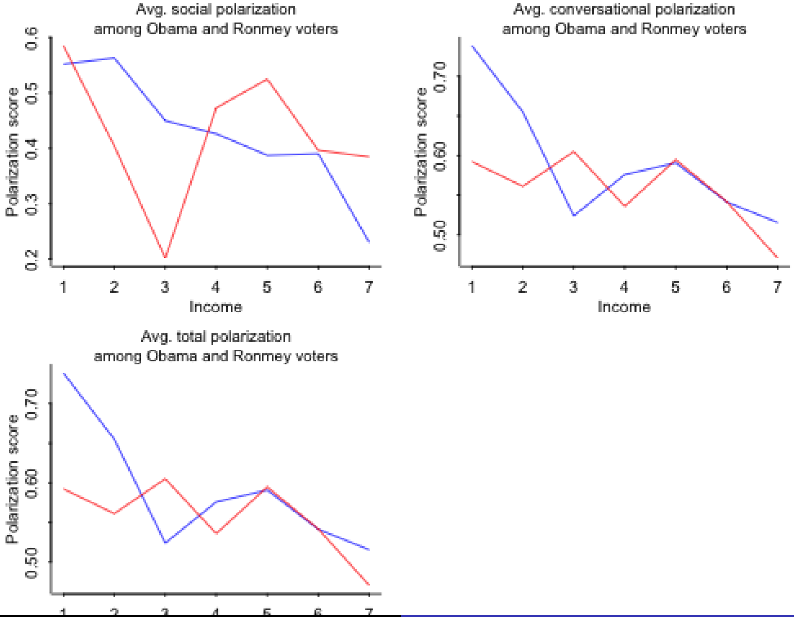
Data!



On Oct 17, 2012, at 4:44 PM, Adam Myers wrote:

> Hi Prof. Gelman,
>
> Attached are the deliverables for last week's Economist/YouGov survey, in which I ran your questions. You'll find an SPSS file, codebook, and questionnaire in case you would like to look at question wording.
>
> Your variables are: vote_know, vote_family, vote_neighbor, poltalk.
>
> In addition, gelrand is the random variable determining whether respondents got "Barack Obama" or "Mitt Romney" as piped-in text for the first three questions. 1 = "Barack Obama", 2 = "Mitt Romney"
>
> You're welcome to use anything else in the dataset for your post as well.

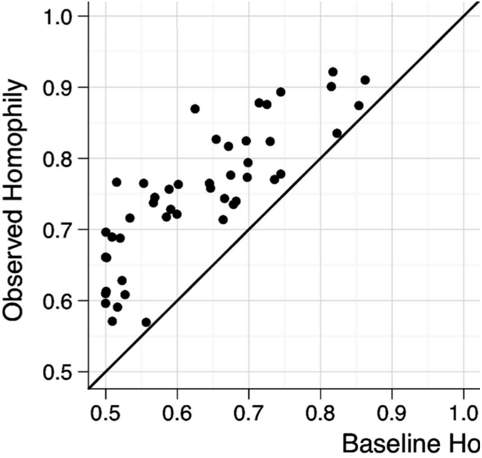
Polarization during the 2012 election campaign



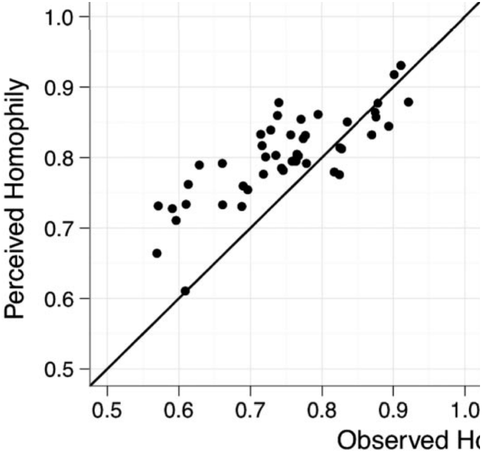
Our next penumbra project

- What groups to study?
- Descriptive findings
- Causal identification

Your friends tend to agree with you



You think your friends *really* agree with you



Polarization, and perception of polarization

- Two models for the Friend Sense results
  - Projection
  - Stereotyping
- Other ways to measure perceptions of polarization?

Fractal sampling

- Fractal sampling in time
- Fractal sampling in space
- Gathering fractal information by sampling or interviewing
- Turning a bug into a feature
- Practical challenges