Non-probability sampling (03-03)

Matthew J. Salganik
Department of Sociology
Princeton University

Soc 596: Computational Social Science
Fall 2016

	Sampling	Interviews	Data environment
1st era	Area probability	Face-to-face	Stand-alone
2nd era	Random digital dial probability	Telephone	Stand-alone
3rd era	Non-probability	Computer-administered	Linked

Probability Samples

$$P(u_i) = \frac{p_i}{(N-1)\cdots(N-n+1)}\binom{N-1}{n-1}(n-1)! + \sum_{j\neq i}^{N}\frac{p_j}{(N-1)\cdots(N-n+1)}\binom{N-1}{n-1}(n-1)!\frac{n-1}{N-1},$$

which upon simplification becomes

$$P(u_i) = \frac{N-n}{N-1}p_i + \frac{n-1}{N-1}, \qquad (i = 1, 2, \cdots, N). \tag{19}$$

Similarly, it may be shown that for this case

$$P(u_i u_j) = \frac{n-1}{N-1}\left[\frac{N-n}{N-2}(p_i + p_j) + \frac{n-2}{N-2}\right], \qquad (i \neq j;\ i, j = 1, 2, \cdots, N). \tag{20}$$

Non-Probability Samples

http://www.chicagotribune.com/news/nationworld/politics/chi-chicagodays-deweydefeats-story-story.html

Probability Samples

unknown sampling process
weighting based on unverifiable assumptions

Non-Probability Samples

unknown sampling process
weighting based on unverifiable assumptions

- Probability sample (roughly): every unit from a defined population (a sampling frame) has a known and non-zero probability of inclusion

- Probability sample (roughly): every unit from a defined population (a sampling frame) has a known and non-zero probability of inclusion
- Not all probability samples are directly representative of the population

- Probability sample (roughly): every unit from a defined population (a sampling frame) has a known and non-zero probability of inclusion
- Not all probability samples are directly representative of the population
- But, with appropriate weighting, probability samples can yield unbiased estimates of the frame population

- Key to many adjustment methods is to use external information

- Key to many adjustment methods is to use external information
- If external information is incorrect or used improperly then you can make things worse

Imagine that you want to estimate the average height of Princeton students.

- Assume 50% are male and 50% are female
- You stand outside Frist and recruit 60 people
- Males (n= 20): Average height: 180cm
- Females (n=40): Average heigh: 170cm

What is your estimate of the average height? (think-pair-share at board)

- sample mean $= 173.3$cm $(\frac{180*20+170*40}{20+40})$

- sample mean $= 173.3\text{cm}$ $(\frac{180*20+170*40}{20+40})$
- weighted estimate $= 175\text{cm}$ $(180 * 0.5 + 170 * 0.5)$

- sample mean $= 173.3$cm $(\frac{180*20+170*40}{20+40})$
- weighted estimate $= 175$cm $(180 * 0.5 + 170 * 0.5)$

How could this go wrong?

Imagine that you want to estimate the average height of Princeton students.

- Assume 50% male and 50% female; assume 25% first-year; 25% sophomore; 25% junior; 25% senior; assume gender and class year are independent
- Your (relatively) sample does not include any female seniors. How could you use the same trick?

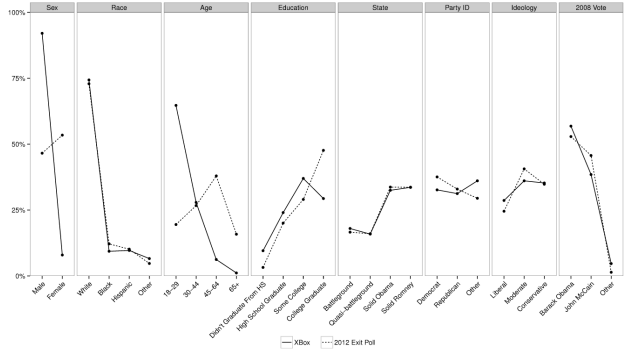
Forecasting elections with non-representative polls

Wei Wang [a,*], David Rothschild [b], Sharad Goel [b], Andrew Gelman [a,c]

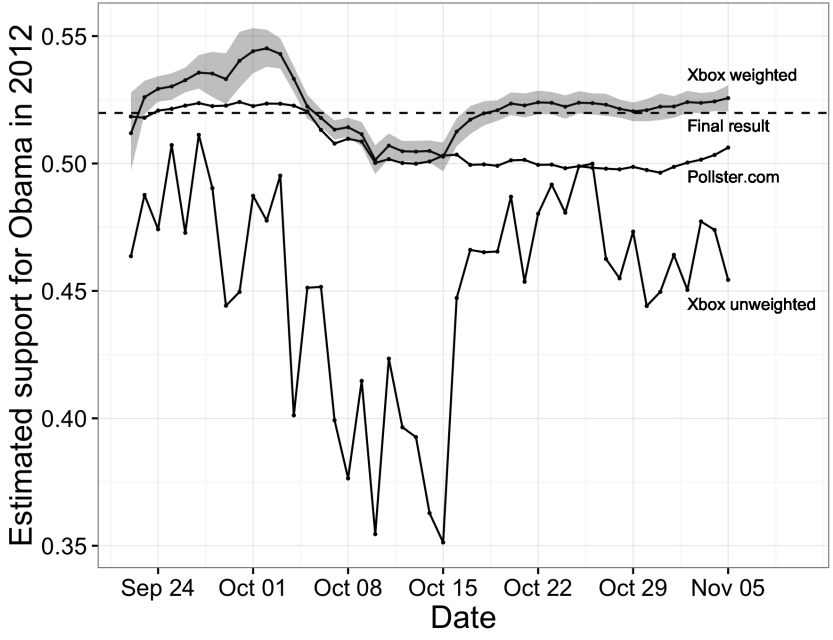
[a] *Department of Statistics, Columbia University, New York, NY, USA*

[b] *Microsoft Research, New York, NY, USA*

[c] *Department of Political Science, Columbia University, New York, NY, USA*



- about 750,000 interviews
- about 350,000 unique respondents


Estimated support for Obama in 2012
0.55
0.50
0.45
0.40
0.35
Sep 24
Oct 01
Oct 08
Oct 15
Oct 22
Oct 29
Nov 05
Date
Xbox weighted
Final result
Pollster.com
Xbox unweighted

« Scientific communication by press release

Nate Silver's website »

President of American Association of Buggy-Whip Manufacturers takes a strong stand against internal combustion engine, argues that the so-called "automobile" has "little grounding in theory" and that "results can vary widely based on the particular fuel that is used"

Posted by Andrew on 6 August 2014, 2:45 pm

http://andrewgelman.com/2014/08/06/
president-american-association-buggy-whip-manufacturers-takes-strong-stand-internal-combustion-engine-argues-called-automobile-little-grounding-theory/

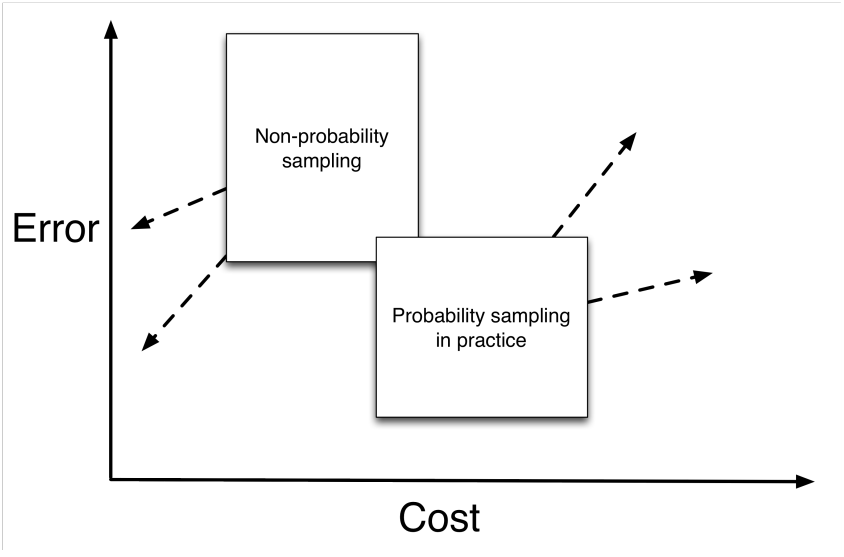
- Mr. P is just one of the many ways to post-stratifiy non-probability samples

- Mr. P is just one of the many ways to post-stratifiy non-probability samples
- the performance of Mr. P (and related methods) is an empirical question

- Mr. P is just one of the many ways to post-stratifiy non-probability samples
- the performance of Mr. P (and related methods) is an empirical question
- these methods can be applied to big data and experiments

- Mr. P is just one of the many ways to post-stratifiy non-probability samples
- the performance of Mr. P (and related methods) is an empirical question
- these methods can be applied to big data and experiments
- there are also methods that focus on sampling rather than weighting (e.g., sample matching)

- Mr. P is just one of the many ways to post-stratifiy non-probability samples
- the performance of Mr. P (and related methods) is an empirical question
- these methods can be applied to big data and experiments
- there are also methods that focus on sampling rather than weighting (e.g., sample matching)
- we should not let what happened in 1948 prevent us from trying new things today


Non-probability sampling
Probability sampling in practice
Error
Cost