[Survey research in the digital age], [Probability and non-probability sampling], [Computer-administered interviews], [Combining surveys and big data], [Additions and extensions]

Matthew J. Salganik
Department of Sociology
Princeton University

1) Introduction
2) Observing behavior
3) Asking questions
4) Running experiments
5) Mass collaboration
6) Ethics
7) The future

	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

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

Similarly, it may be shown that for this case

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

Non-Probability Samples

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

Probability Samples	Non-Probability Samples
unknown sampling process weighting based on unverifiable assumptions	unknown sampling process weighting based on unverifiable assumptions

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

- Probability sample (roughly): every unit from a frame population has a known and non-zero probability of inclusion
- Not all probability samples look like miniature versions of the population

- Probability sample (roughly): every unit from a frame population has a known and non-zero probability of inclusion
- Not all probability samples look like miniature versions of the population
- But, with appropriate weighting, probability samples can yield unbiased estimates of the frame population

Main insights from probability sampling:

- How you collect your data impacts how you make inference

Main insights from probability sampling:

- How you collect your data impacts how you make inference
- Focus on properties of estimators not properties samples

Main idea and equation in sampling and estimation:

$$\hat{\bar{y}} = \frac{\sum_{i \in s} y_i / \pi_i}{N}$$

where π_i is person i's probability of inclusion

Sometimes called:

- Horvitz-Thompson estimator
- π estimator

Inference from probability samples in theory

$$\left.\begin{array}{r}\text{respondents}\\ \text{known information about sampling}\end{array}\right\}\ \text{estimates}$$

Inference from probability samples in theory

$$\left.\begin{array}{r}\text{respondents}\\ \text{known information about sampling}\end{array}\right\}\ \text{estimates}$$

Inference from probability samples in practice

$$\left.\begin{array}{r}\text{respondents}\\ \underbrace{\text{estimated information about sampling}}_{\text{auxiliary information + assumptions}}\end{array}\right\}\ \text{estimates}$$

Inference from probability samples in theory

$$\left.\begin{array}{r}\text{respondents}\\ \text{known information about sampling}\end{array}\right\}\ \text{estimates}$$

Inference from probability samples in practice

$$\left.\begin{array}{r}\text{respondents}\\ \underbrace{\text{estimated information about sampling}}_{\text{auxiliary information } + \text{ assumptions}}\end{array}\right\}\ \text{estimates}$$

Inference from non-probability samples

$$\left.\begin{array}{r}\text{respondents}\\ \underbrace{\text{estimated information about sampling}}_{\text{auxiliary information } + \text{ assumptions}}\end{array}\right\}\ \text{estimates}$$

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

- Assume 50% are male and 50% are female
- You stand outside Lewis Library and recruit a non-random sample of 60 Princeton students
- Males (n= 20): Average height: 180cm
- Females (n=40): Average heigh: 170cm

What is your estimate of the average height?

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

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

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

- weighted estimate uses auxiliary information and assumptions

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

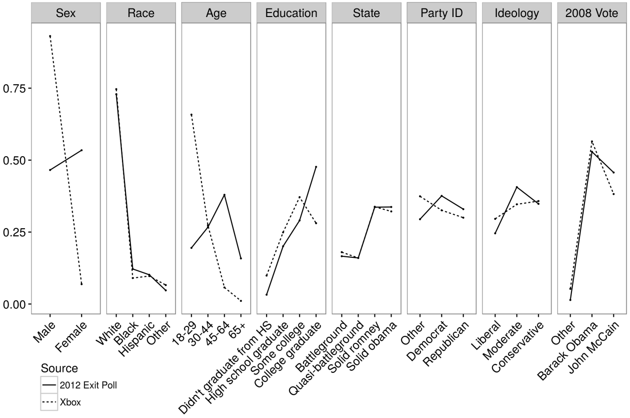
- weighted estimate uses auxiliary information and assumptions
- How could this go wrong?

Forecasting elections with non-representative polls

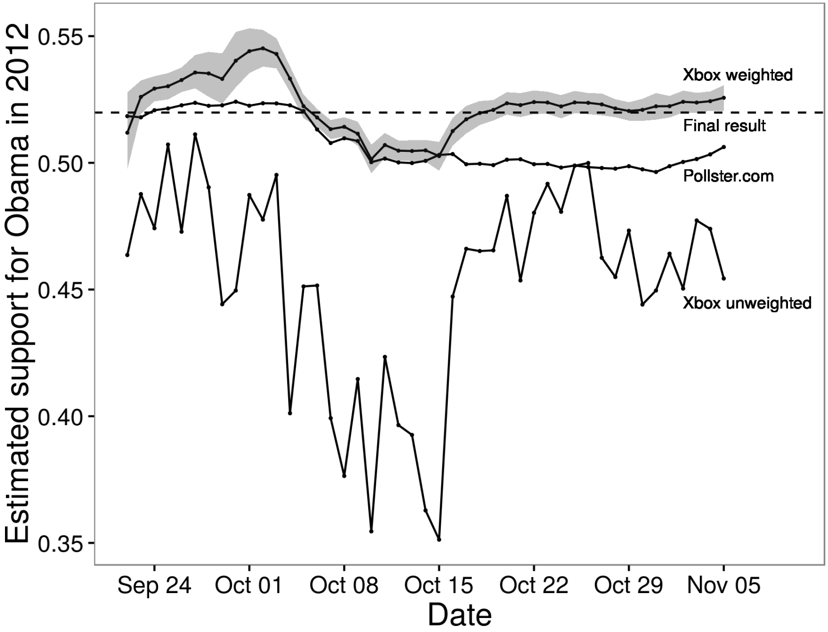
Wei Wang[a,*], David Rothschild[b], Sharad Goel[b], Andrew Gelman[a,c]

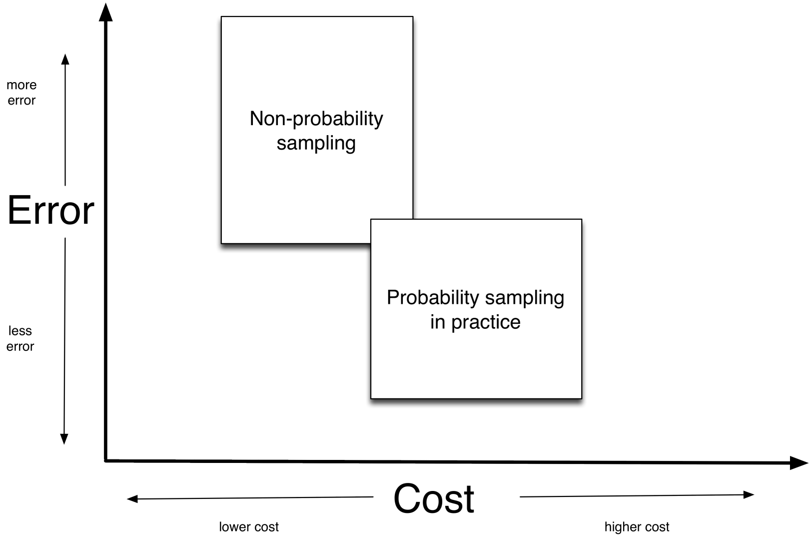
[a] *Department of Statistics, Columbia University, New York, NY, USA*
[b] *Microsoft Research, New York, NY, USA*
[c] *Department of Political Science, Columbia University, New York, NY, USA*

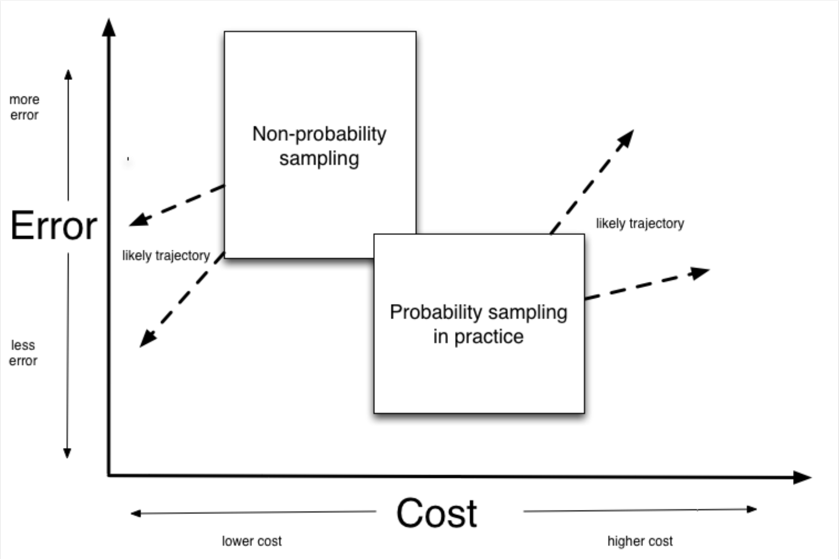
Wang et al (2015)



- 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
Xbox weighted
Final result
Pollster.com
Xbox unweighted
Sep 24
Oct 01
Oct 08
Oct 15
Oct 22
Oct 29
Nov 05
Date


more error
Error
less error
Non-probability sampling
Probability sampling in practice
Cost
lower cost
higher cost


more error
Error
less error
Non-probability sampling
likely trajectory
likely trajectory
Probability sampling in practice
Cost
lower cost
higher cost

Wrap-up:

- Samples don't need to look like mini-populations

Wrap-up:

- Samples don't need to look like mini-populations
- Key to making good estimates is for estimation process to account for the sampling process

Wrap-up:

- Samples don't need to look like mini-populations
- Key to making good estimates is for estimation process to account for the sampling process
- There is not a bright-line difference between probability sampling in practice and non-probability sampling

Wrap-up:

- Samples don't need to look like mini-populations
- Key to making good estimates is for estimation process to account for the sampling process
- There is not a bright-line difference between probability sampling in practice and non-probability sampling
- To learn more: Lohr (2009) or Sandal et al (2013)

[Survey research in the digital age], [Probability and non-probability sampling], [Computer-administered interviews], [Combining surveys and big data], [Additions and extensions]

Matthew J. Salganik
Department of Sociology
Princeton University