

# Probability and non-probability sampling

Matthew J. Salganik  
Department of Sociology  
Princeton University

Summer Institutes in Computational Social Science  
2020

The Summer Institutes in Computational Social Science is supported by grants from the Russell Sage Foundation and the Alfred P. Sloan Foundation.



	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

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

Similarly, it may be shown that for this case

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

# Non-Probability Samples



## 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 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 insight from probability samples:

- ▶ 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:

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

where  $\pi_i$  is person  $i$ 's probability of inclusion

Sometimes called:

- ▶ Horvitz-Thompson estimator
- ▶  $\pi$  estimator

# Inference from probability samples in theory

respondents } estimates  
known information about sampling }

# Inference from probability samples in theory

respondents } estimates  
known information about sampling }

---

# Inference from probability samples in practice

respondents } estimates  
estimated information about sampling }  
auxiliary information + assumptions }

## Inference from probability samples in theory

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

---

## Inference from probability samples in practice

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

---

## Inference from non-probability samples

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

$$\hat{y} = \frac{\sum_{i \in s} y_i / \hat{\pi}_i}{N}$$

where  $\hat{\pi}_i = \frac{n_g}{N_g} \quad \forall \quad i \in g$  (estimated probability of inclusion)

Requires:

- ▶ auxiliary information ( $N_g$ )
- ▶ ability to place respondents in groups
- ▶ assumptions

- ▶ Key to many adjustment methods is to use external information and make assumptions

- ▶ Key to many adjustment methods is to use external information and make assumptions
- ▶ If external information is incorrect or assumptions are wrong, then you can make things worse (but it usually seems to make things better)

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 60 Princeton students
- ▶ Males ( $n=20$ ): Average height: 180cm
- ▶ Females ( $n=40$ ): Average height: 170cm

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



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

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

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

How could this go wrong?

# Forecasting elections with non-representative polls

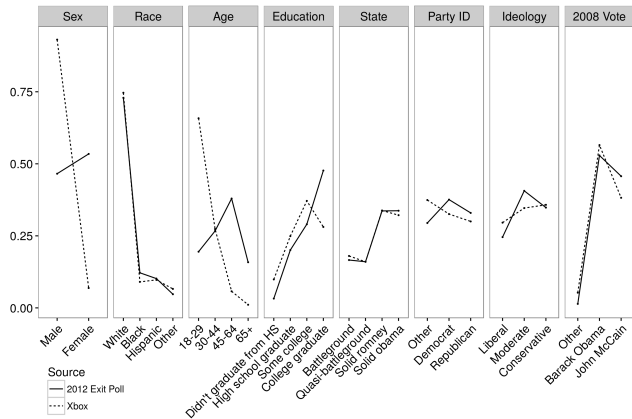
Wei Wang<sup>a,\*</sup>, David Rothschild<sup>b</sup>, Sharad Goel<sup>b</sup>, Andrew Gelman<sup>a,c</sup>

<sup>a</sup> *Department of Statistics, Columbia University, New York, NY, USA*

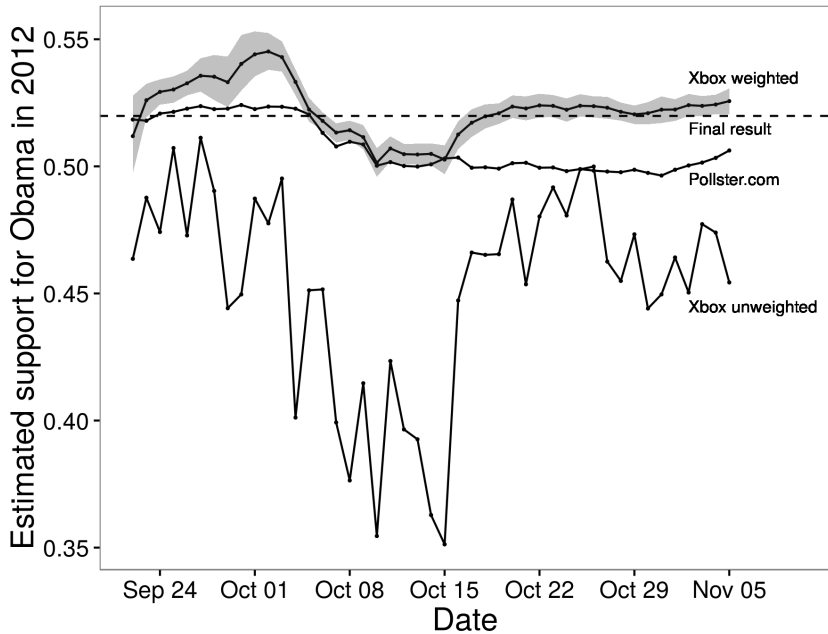
<sup>b</sup> *Microsoft Research, New York, NY, USA*

<sup>c</sup> *Department of Political Science, Columbia University, New York, NY, USA*





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

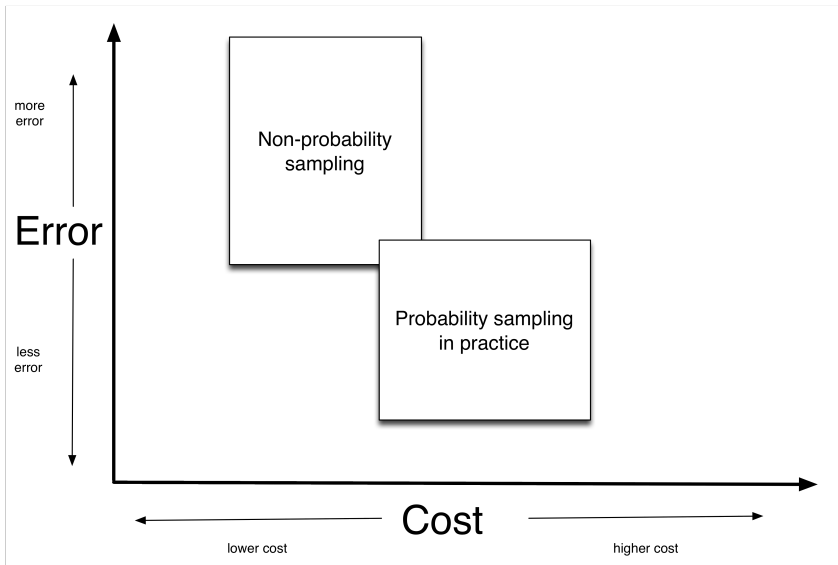


# Online, Opt-in Surveys: Fast and Cheap, but are they Accurate?

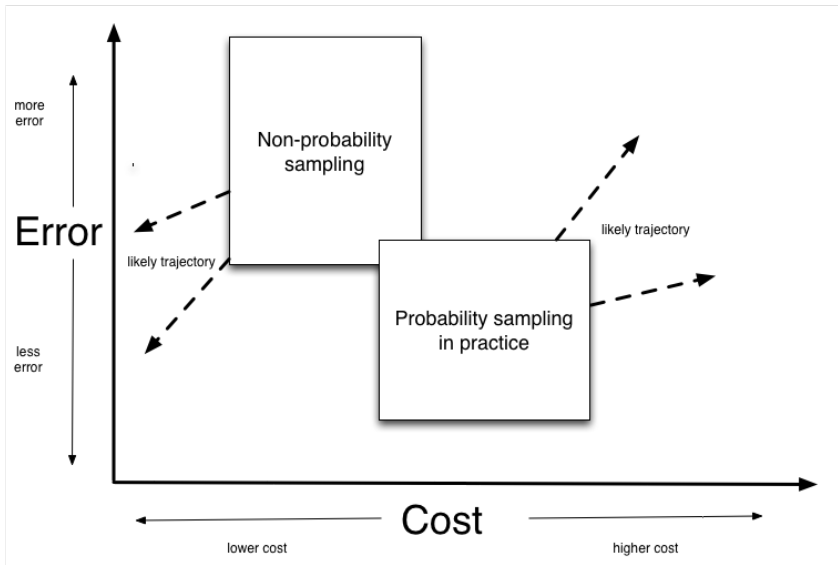
Sharad Goel  
Stanford University  
scgoel@stanford.edu

Adam Obeng  
Columbia University  
adam.obeng@columbia.edu

David Rothschild  
Microsoft Research  
davidmr@microsoft.com







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)