Survey research in the digital age

Bernhard Clemm von Hohenberg
Department of Computational Social Science
GESIS

Summer Institutes in Computational Social Science
July 28, 2023

Schedule

- 9.00-9.45 Introduction & total error survey framework
- 9.45-10.15 Probability and non-probability sampling
- Coffee break
- 10.30-11.00 Computer-administered interviewing
- 11.00-11.30 Linking surveys to big data
- 11:30-13:00 Intro and begin group exercise
- Lunch (or Eisbach plunge)
- 14:00-15:45 Continue group exercise

Schedule

	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

Warm-up exercise

Go to https://forms.gle/AtdDu6hS8RiuhUWB6 and indicate your height in cm (e.g., "194") and your sex. If you don't know your height, or don't want to tell, give an estimate or some plausible number.

Warm-up exercise

Let's have a look at

- … the average height of the population (all of you)
- … the average height estimated from a random sample
- … the average height estimated from a non-random (but probability) sample

Probability samples

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

Probability samples

- Probability sample (roughly): every unit from a frame population has a known and non-zero probability of inclusion
  - In the case of a simple random sample, this inclusion probability is n/N, with n being the sample size and N being size of population

Probability samples

- Probability sample (roughly): every unit from a frame population has a known and non-zero probability of inclusion
  - In the case of a simple random sample, this inclusion probability is n/N, with n being the sample size and N being size of population
- In practice, we rarely deal with a simple random sample
- However, if we know the inclusion probability, we can get an unbiased estimate of the population mean.

Probability-based estimation

Horvitz-Thompson estimator: the estimator for the population mean $\bar{y}$ is

$$\hat{\bar{y}} = \frac{1}{N} \sum_{i \in s} \frac{y_i}{\pi_i}$$

where π_i is person i's probability of inclusion. Verbally, this is a weighted sample mean where the weights are inversely related to the probability of selection.

Theory vs. practice

Inference from probability samples in theory

known information about sampling
$+$ respondents
$=$ estimates

Theory vs. practice

Inference from probability samples in theory

known information about sampling
$+$ respondents
$=$ estimates

Inference from probability samples in practice

auxiliary information
$+$ assumptions
$=$ estimated information about sampling

estimated information about sampling
$+$ respondents
$=$ estimates

Theory vs. practice

Inference from probability samples in theory

known information about sampling
$+$ respondents
$=$ estimates

Inference from probability samples in practice

auxiliary information
$+$ assumptions
$=$ estimated information about sampling

estimated information about sampling
$+$ respondents
$=$ estimates

Inference from non-probability samples in practice

auxiliary information
$+$ assumptions
$=$ estimated information about sampling

estimated information about sampling
$+$ respondents
$=$ estimates

Theory vs. practice

$$\hat{\bar{y}} = \frac{1}{N} \sum_{i \in s} \frac{y_i}{\hat{\pi}_i}$$

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

Theory vs. practice

- Not all probability samples look like miniature versions of the population—but, with appropriate weighting, probability samples can yield unbiased estimates

Theory vs. practice

- Not all probability samples look like miniature versions of the population—but, with appropriate weighting, probability samples can yield unbiased estimates
- How you collect your data impacts how you make inference

Theory vs. practice

- Not all probability samples look like miniature versions of the population—but, with appropriate weighting, probability samples can yield unbiased estimates
- How you collect your data impacts how you make inference
- Key to many adjustment methods is to use external information and make assumptions

Theory vs. practice

- Not all probability samples look like miniature versions of the population—but, with appropriate weighting, probability samples can yield unbiased estimates
- How you collect your data impacts how you make inference
- 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)

Back to warm-up

Back to the average height of your group:

- If I know the inclusion probability (e.g., 0.8 for males, 0.2 for females), it does not matter that the sample is non-probability
- If I know the frequency of males/females in the population (e.g., 0.5/0.5), I can build a weighted average (0.5*average males + 0.5*average females)
- However, I often don't know *why* more males participated than females; if I see more males than females in my sample, I have to *assume* that this imbalance is caused by *sex* and not something else
- In sum, I need assumptions to use sex as auxiliary information to estimate inclusion probability

Forecasting elections with non-representative polls

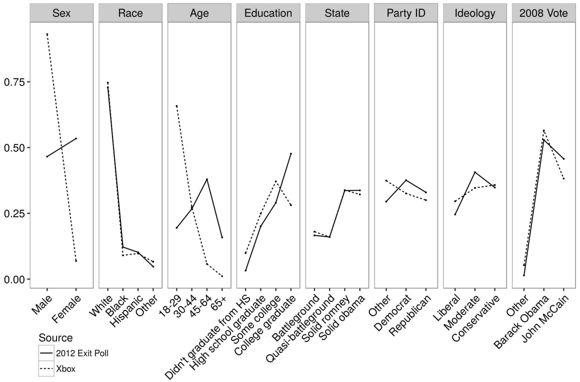
Wei Wang[a,*], David Rothschild[b], Sharad Goel[b], Andrew Gelman[a,c]

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

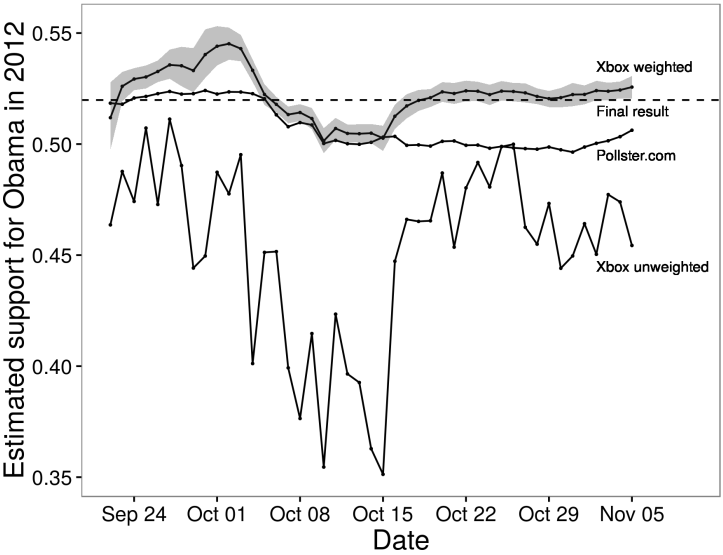
[b] *Microsoft Research, New York, NY, USA*

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

https://www.journals.elsevier.com/international-journal-of-forecasting/editors-choice-articles/forecasting-elections-with-non-representative-polls



- 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

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

Sharad Goel
Stanford University
scgoel@stanford.edu

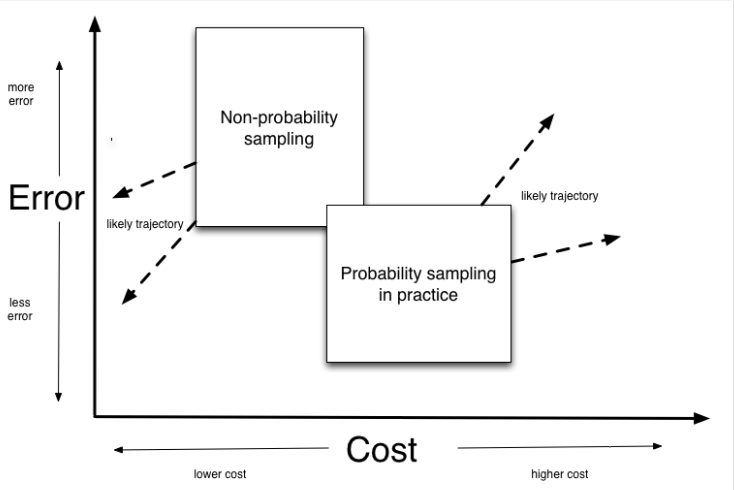
Adam Obeng
Columbia University
adam.obeng@columbia.edu

David Rothschild
Microsoft Research
davidmr@microsoft.com

https://5harad.com/papers/dirtysurveys.pdf

The future

... according to Matt Salganik



The future

RESEARCH SYNTHESIS: Are Nonprobability Surveys Fit for Purpose?

Jennifer Jerit
Professor
Department of Government
Dartmouth College

Jason Barabas
Professor, Department of Government
Director, Rockefeller Center for Public Policy and the Social Sciences
Dartmouth College

March 21, 2023

"In studies comparing the accuracy of probability and nonprobability samples in relation to government records, the former consistently outperforms the latter"

https://bpb-us-e1.wpmucdn.com/sites.dartmouth.edu/dist/d/2388/files/2023/05/JeritBarabas_NPS_Mar2023-1.pdf

The future

"Despite substantial drops in response rates since a prior comparison, the probability samples interviewed by telephone or the internet were the most accurate. Internet surveys of a probability sample combined with an opt-in sample were less accurate; least accurate were internet surveys of opt-in panel samples."

https://academic.oup.com/poq/article-abstract/82/4/707/5151369

JOURNAL ARTICLE

The Accuracy of Measurements with Probability and Nonprobability Survey Samples: Replication and Extension Get access ›

Bo MacInnis, Jon A Krosnick, Annabell S Ho, Mu-Jung Cho

Public Opinion Quarterly, Volume 82, Issue 4, Winter 2018, Pages 707–744, https://doi.org/10.1093/poq/nfy038

Published: 31 October 2018

Summary

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

Summary

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

Summary

- 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

Summary

- 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
- However, still a lot of debate about how non-probability samples really perform

Questions