Pew Research Center

Selection bias and models in nonprobability sampling

Andrew Mercer

Senior Research Methodologist
PhD Candidate, JPSM

A BRIEF HISTORY OF BIAS IN NONPROBABILITY SAMPLES

In 2004-2005 samples varied in accuracy

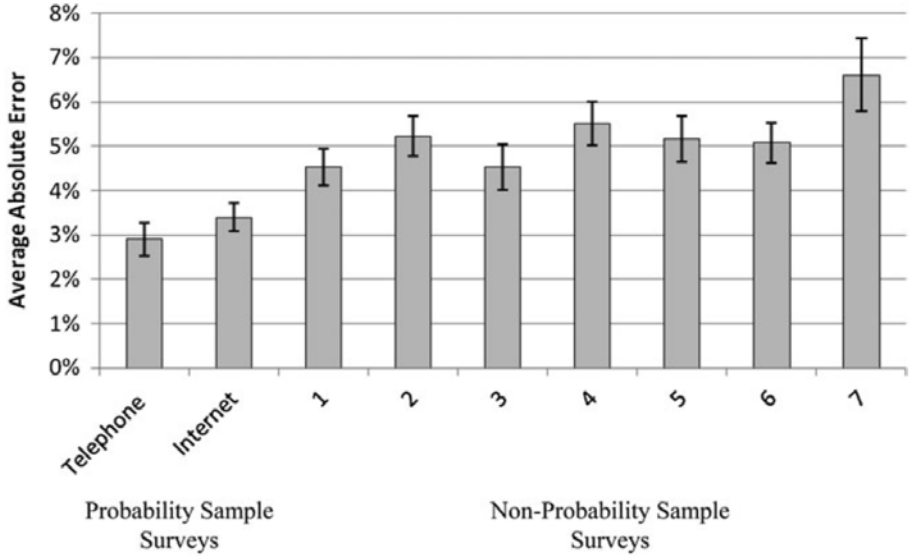


Figure 1. Average Percentage Point Absolute Errors for Commissioned Probability and Non-Probability Sample Surveys across Thirteen Secondary Demographics and Non-Demographics, with Post-Stratification.

Source: Yeager, David S., Jon a. Krosnick, Linchiat Chang, Harold S. Javitz, Matthew S. Levendusky, Alberto Simpser, and Rui Wang. 2011. "Comparing the Accuracy of RDD Telephone Surveys and Internet Surveys Conducted with Probability and Non-Probability Samples." Public Opinion Quarterly 75 (4): 709–47.

Still true in 2013

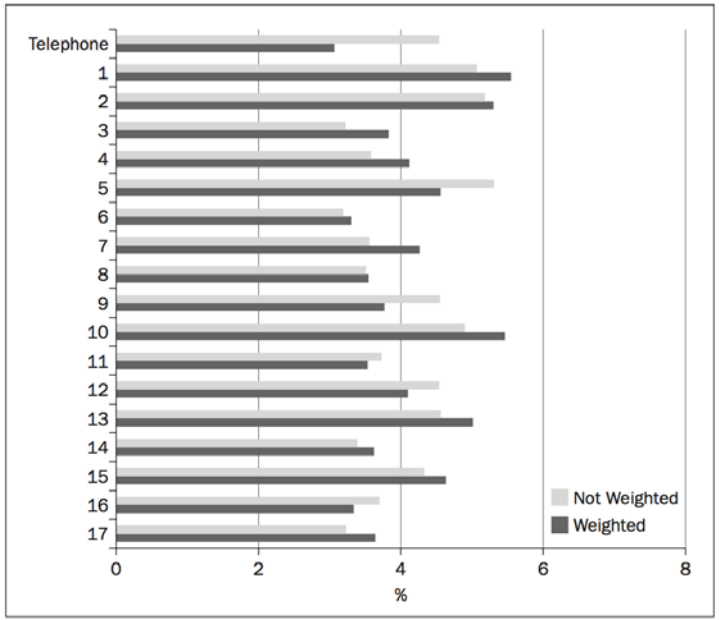
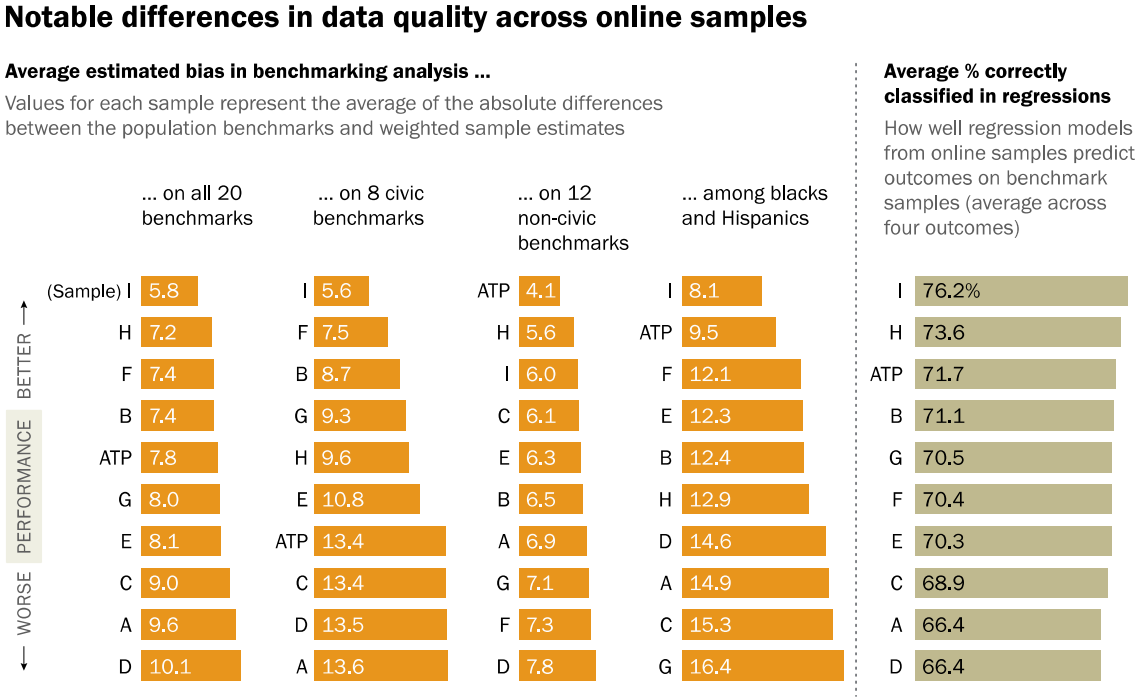


Figure 5 Average Absolute Benchmark Difference for Telephone and by Online Provider

Source: Gittelman, Steven H., Randall K. Thomas, Paul J. Lavrakas, and Victor Lange. 2015. "Quota Controls in Survey Research: A Test of Accuracy and Intersource Reliability in Online Samples." Journal of Advertising Research 55 (4): 368–79.

And in 2015



Source: Kennedy, Courtney, Andrew Mercer, Scott Keeter, Nick Hatley, Kyley Mcgeeney, and Alejandra Gimenez. 2016. "Evaluating Online Nonprobability Surveys." Pew Research Center.

What's missing?

- Quality still varies considerably between sources
- Don't know why some are better than others
- No guidance on how estimates can be improved
- Implicit assumption that standard procedures used for probability samples should work for nonprobability samples

BIAS IN SURVEY ESTIMATES

The ideal: perfect randomization

Random selection…

from the entire population…

with known probabilities of selection…

and 100% response…

..implies that for any variable we measure (and those that we don't measure), the sample distribution will match the population on average

Depends only on what know about the process. Not what we know about the population

No need for models or strong assumptions

None of these things apply to nonprobability surveys

- Selection is not random by definition
- You do not have access to the whole population
- There are no known inclusion probabilities
- You probably don't have complete response

The sample is only sure to match the population on those dimensions where it is fixed by design (e.g. quotas)

Nothing intrinsic to the process that guarantees anything

Entirely dependent on models and strong assumptions

TSE formulation for probability surveys doesn't fit well

Main sources of bias for probability-based surveys:

- Undercoverage
- Nonresponse

Not well defined for many nonprobability surveys

- No sampling frame
- Often not even a finite sample (e.g. routers, river samples)

Not very informative without random selection

- Describe how a the process deviates from randomization
- Not how a sample differs from the population or how a model is misspecified

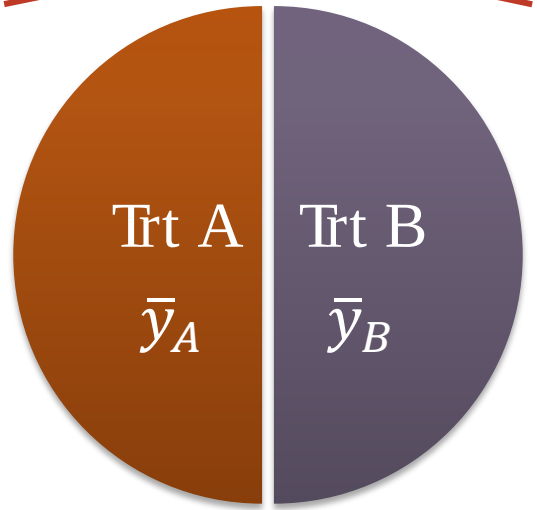
If a survey had perfect coverage and 100%nonresponse but nonrandom selection, how would you describe the error?

WE ARE NOT THE FIRST PEOPLE TO HAVE THIS PROBLEM

Parallels Between Experiments and Surveys

Observational Study

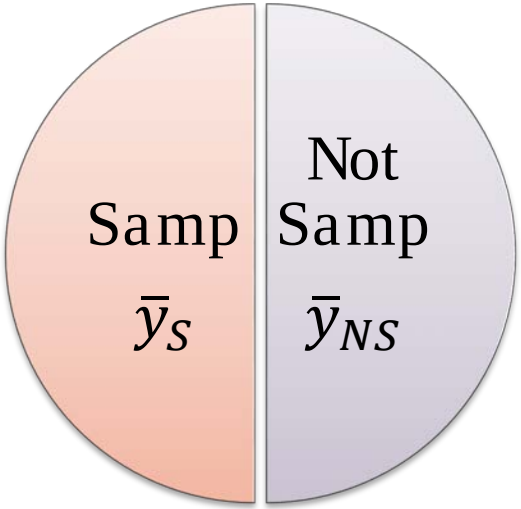
~~Randomized experiment~~



$$\bar{y}_A - \bar{y}_B = \theta$$

Nonprobability survey

~~Probability-based survey~~



$$\bar{y}_{Pop} = \bar{y}_S = \bar{y}_{NS}$$

Causal Inference and Survey Inference

Causal inference and survey inference both recognize the utility of randomization

- Probability-based surveys
- Randomized experiments

As a field, causal inference has recognized the validity of learning from observational data while recognizing the limitations

- Political science, economics, epidemiology, sociology, and most other fields all use observational data
- Decades of research developing methods for observational data

Tends to describe error in terms of failure to meet necessary conditions or violated assumptions

What conditions are required to avoid selection bias?

Exchangeability (aka ignorability, unconfoundedness):

We know and have measured all of the confounding variables that are correlated with both inclusion in the sample and the outcome of interest

Positivity (aka common support):

There are no kinds of unit with distinct values of the outcome variable that are systematically missing from the sample

Composition:

The distribution of potentially confounding variables in the sample matches the distribution in the population

A Vocabulary for Describing Specific Problems

Selection bias implies that:

$$E\left(Y \middle| S = \text{Population}\right) \neq E\left(Y \middle| S = \text{Sample}\right)$$

$$E\left(Y \middle| S = s\right) = \underbrace{\sum_{x \in X}}_{\text{(Positivity)}} \overbrace{E\left(Y \mid X = x, S = s\right)}^{\text{(Exchangeability)}} \times \overbrace{\Pr\left(X = x \mid S = s\right)}^{\text{(Composition)}}$$

Which parts of this equation don't match the between sample and population?

Quantifying Sources of Bias

With ”gold-standard” microdata representing the population, we can estimate the contribution of each of these problems to the total selection bias

Combine reference and nonprobability samples:

1. Estimate a propensity model predicting sample membership:

$$\pi(x) = \Pr(S = \text{Sample} \mid X = x)$$

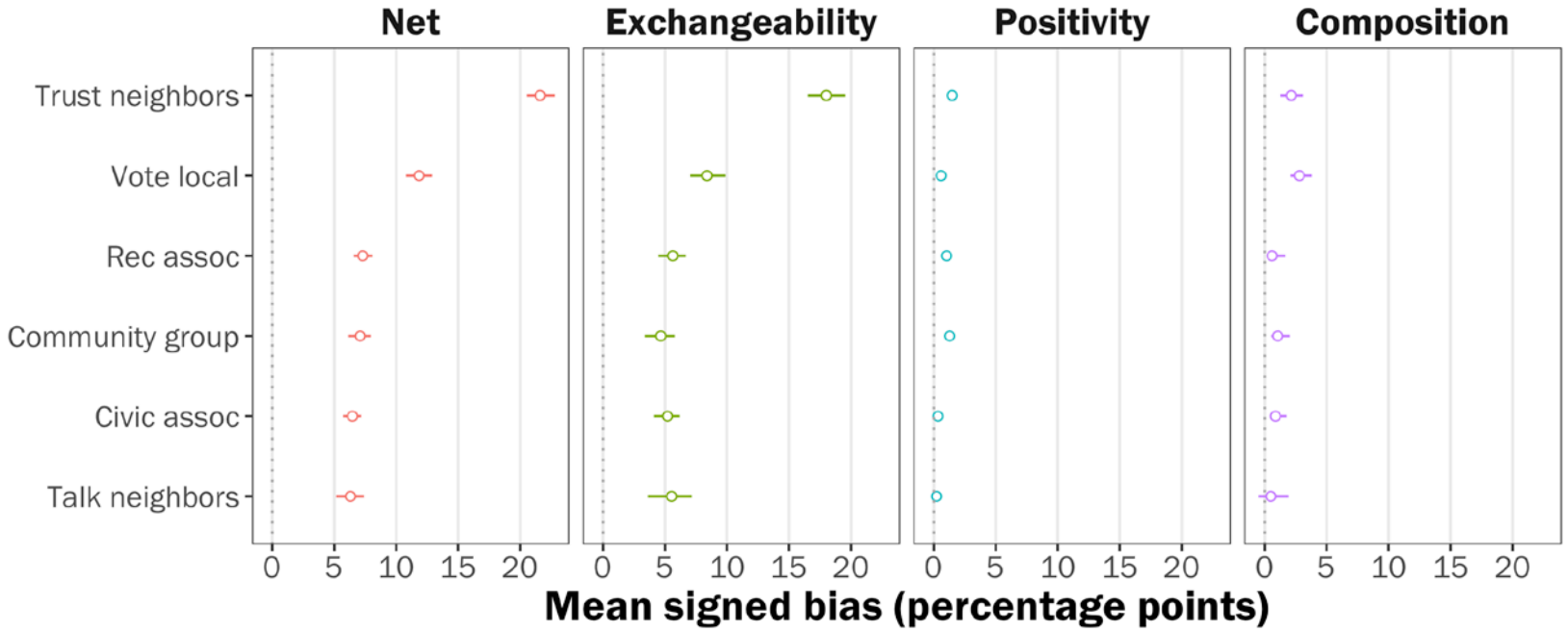
2. Estimate a response surface model for the outcome:

$$f(x, s) = E(Y \mid X = x, S = s)$$

3. Calculate differences between estimated counterfactual means

2013 CPS Civic Engagement vs. Pew Nonprob Samples

Average bias by type for 10 nonprobability samples on 6 measures of civic engagement before weighting. Estimated using BART models conditional on age, sex, race/ethnicity, education, and region.



Source: 9 nonprobability samples from Kennedy et al 2016 plus 1 sample from Mechanical Turk. Estimates produced using BART models conditioning on age, sex, race and Hispanic ethnicity, education, and region.

WHY DO SOME METHODS SEEM TO WORK BETTER?

The Xbox Study and MRP
Wang et al (2014)

Panel survey of Xbox users in the days leading up to the 2012 presidential election.

Closer to the national vote than the 2012 Pollster.com average, and very accurate on state level estimates or Obama vote share (mean error 2.5%).

93% Male and 65% 18-29 years old. 1% of 65+

Exchangeability

- Powerful covariates for predicting vote preference (i.e. party id).

Positivity

- Very large sample size. 750k interviews, 345k respondents.
- 1% is still ~3400 observations.

Composition:

- 2008 exit poll for poststratification gives composition of party. Not usually available.
- MRP regularization allows more dimensions and more granularity.

Revisiting Sample I

For each sample, we requested weights from vendors and also created our own. Used the weights that gave the lowest average error.

Sample I

- Only vendor provided weights that were used for analysis.
- Consistently lower error than other samples.
- What's different?

Exchangeability

- Selection and adjustment using more than just demographic variables (party, religious attitudes, etc…).

Positivity

- Matching to a synthetic population based on ACS may do a better job reproducing the joint distribution of selection variables than marginal quotas.

Composition:

- Combination of matching for selection, propensity score weighting and calibration.

GOING FORWARD

Some principles

- Accept necessity of assumptions
- Be explicit about assumptions and work to make sure they are justifiable
- Design with modeling assumptions in mind
- Design with model testing in mind
- Covariate selection is probably more important than anything else
- Demographics are probably not enough

Achieving the requirements may be difficult or impossible for some areas of research.

This does not make them any less required.

Research agenda

- Variable selection methods for selection and adjustment
- Combining microdata from different sources for use in modeling
- Sensitivity analyses and other methods for testing assumptions
- Measuring quality when benchmarks are unavailable

Pew Research Center

Andrew Mercer
Senior Research Methodologist

amercer@pewresearch.org

Twitter: @awmercer