Using Web Survey Panels to Estimate Population Characteristics: A Comparison of Alternative Approaches

Douglas Rivers
Stanford University and YouGov

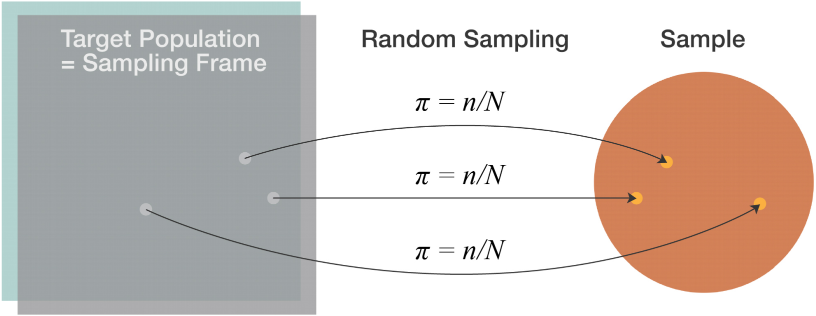
February 25, 2013

Designs for Online Panels

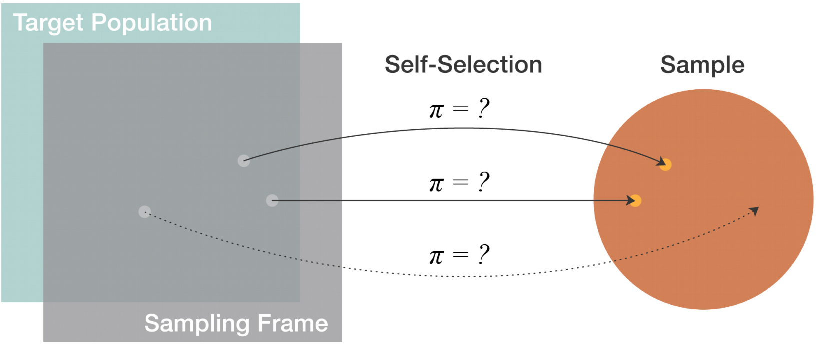
Method of Respondent Selection

	Probability	**Non-Probability**
Internet users	Nielsen TV/PC Panel *RDD recruitment*	Opt-in Web panels River samples
All adults	GfK Knowledge Panel *RDD/ABS panel* Gallup Panel *Recontacts, mixed mode*	

What is a probability sample?



The real world



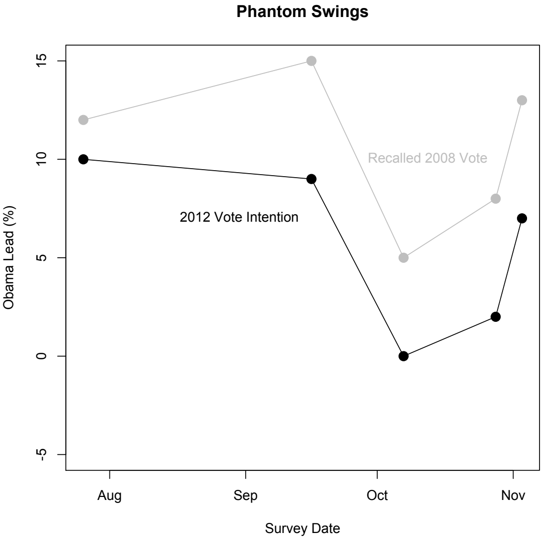
Are these probability samples?

Surveys Face Growing Difficulty Reaching, Persuading Potential Respondents

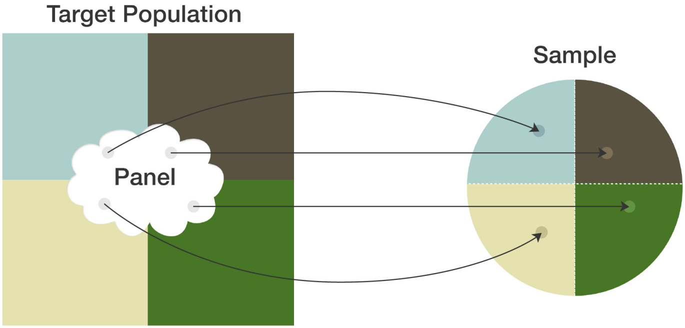
	1997	2000	2003	2006	2009	2012
	%	%	%	%	%	%
Contact rate (percent of households in which an adult was reached)	90	77	79	73	72	62
Cooperation rate (percent of households contacted that yielded an interview)	43	40	34	31	21	14
Response rate (percent of households sampled that yielded an interview)	36	28	25	21	15	9

PEW RESEARCH CENTER 2012 Methodology Study. Rates computed according to American Association for Public Opinion Research (AAPOR) standard definitions for CON2, COOP3 and RR3. Rates are typical for surveys conducted in each year.

The McCain Surge?



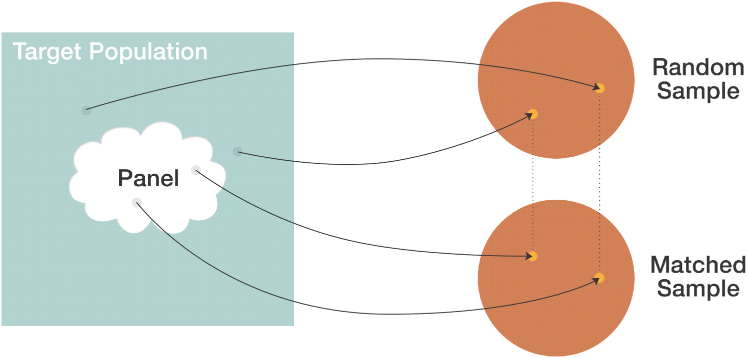
Quota sampling

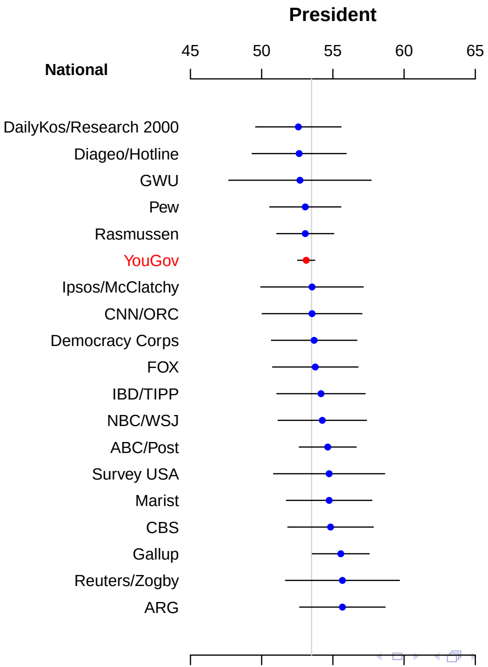


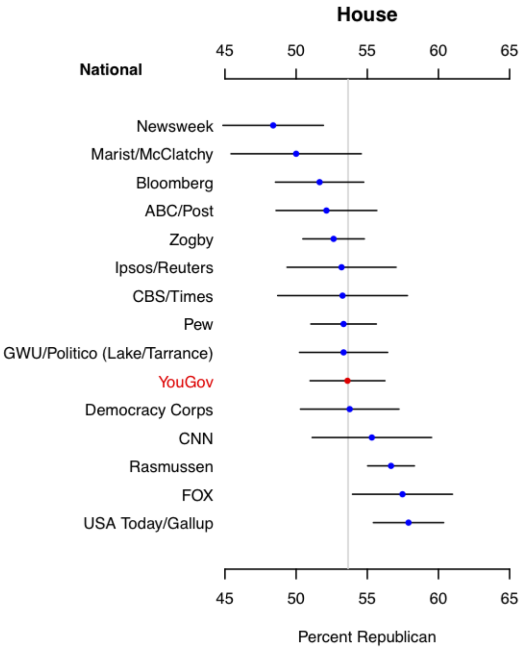
2004 Stanford Mode Study

	SRBI	KN	Pooled	SPSS/ River	Greenfield	Harris	GoZing	Survey Direct	SSI
Gender	1.4%	0.6%	0.0%	2.0%	2.9%	0.7%	3.8%	1.0%	1.0%
Age	2.6%	0.7%	2.3%	4.5%	13.3%	3.7%	35.6%	16.8%	13.8%
Race/Ethnicity	3.5%	0.5%	0.0%	10.9%	17.0%	0.3%	11.2%	13.7%	15.5%
Education	4.5%	1.5%	0.3%	20.7%	29.6%	1.1%	28.1%	25.9%	28.8%
Income	18.2%	5.4%	11.5%	13.5%	8.6%	3.0%	16.3%	12.4%	9.6%
Marital Status	9.0%	5.5%	5.3%	7.0%	7.1%	6.8%	15.4%	9.4%	8.7%
Number of Adults in HH	2.6%	7.3%	5.3%	9.2%	7.8%	3.9%	3.6%	7.1%	6.7%
Work for Pay Last Week	1.5%	2.2%	2.9%	0.5%	6.2%	0.9%	2.7%	5.1%	0.4%
Living Quarters	6.0%	2.4%	5.4%	7.5%	2.6%	3.2%	19.8%	2.9%	3.0%
Number of Bedrooms	6.4%	3.6%	2.0%	6.1%	1.7%	3.5%	5.4%	2.3%	2.7%
Number of Vehicles	4.9%	6.6%	4.6%	7.6%	6.9%	3.5%	6.7%	5.5%	5.5%

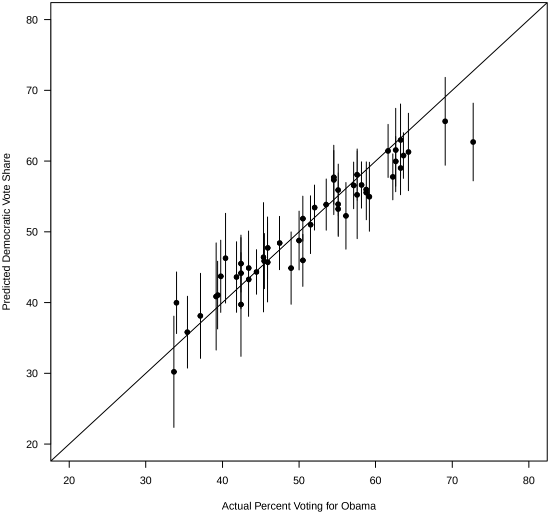
Matched sampling



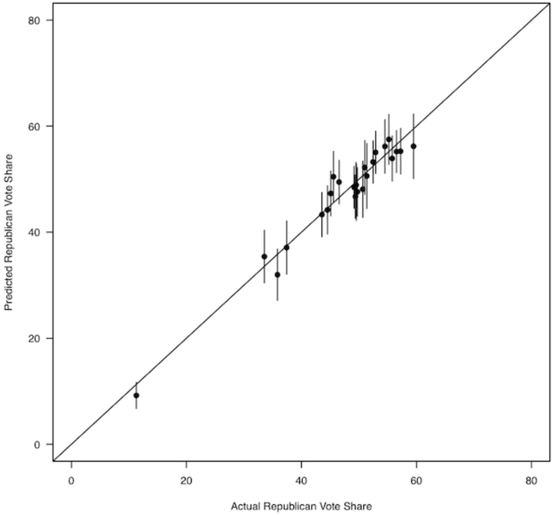

President
National
45
50
55
60
65
DailyKos/Research 2000
Diageo/Hotline
GWU
Pew
Rasmussen
YouGov
Ipsos/McClatchy
CNN/ORC
Democracy Corps
FOX
IBD/TIPP
NBC/WSJ
ABC/Post
Survey USA
Marist
CBS
Gallup
Reuters/Zogby
ARG


House
National
45
50
55
60
65
Newsweek
Marist/McClatchy
Bloomberg
ABC/Post
Zogby
Ipsos/Reuters
CBS/Times
Pew
GWU/Politico (Lake/Tarrance)
YouGov
Democracy Corps
CNN
Rasmussen
FOX
USA Today/Gallup
Percent Republican

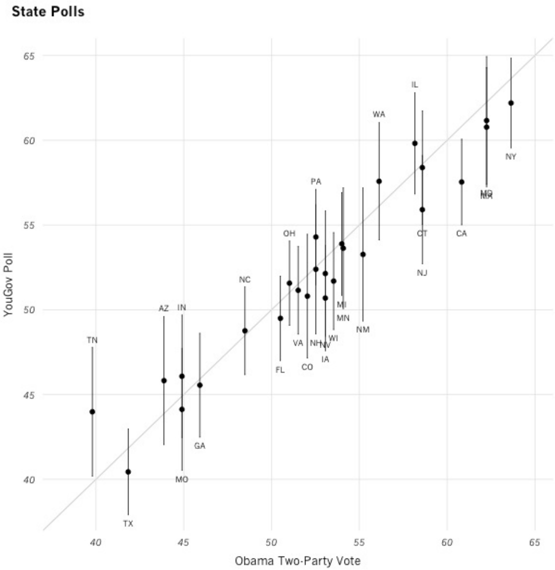
2008 State-level Results



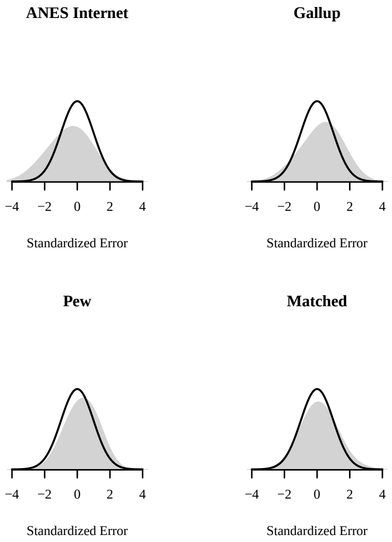
2010 State-level Results

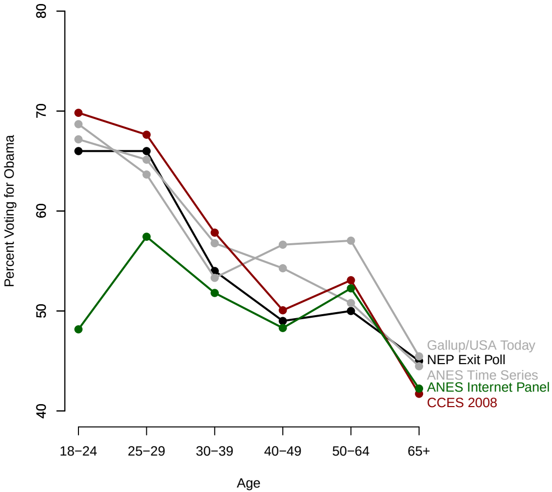


2012 State-level Results

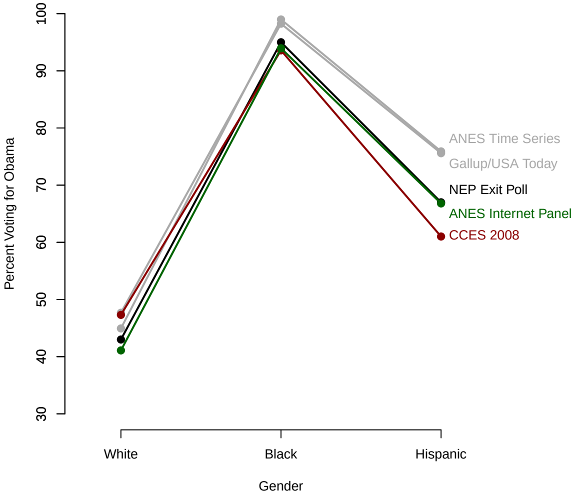


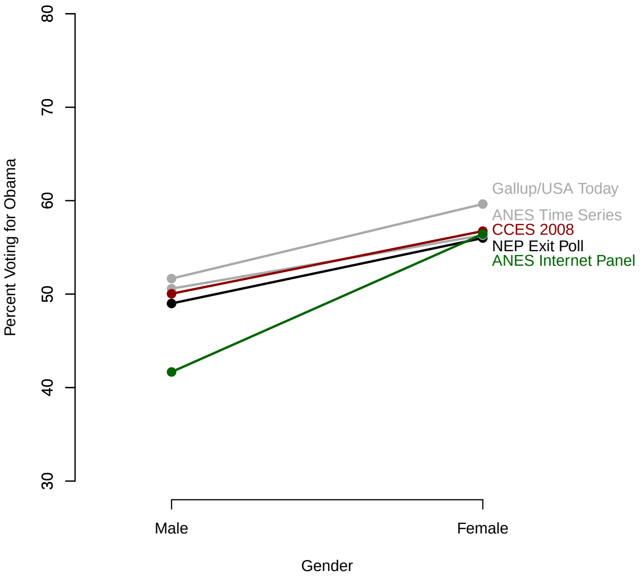
Empirical Sampling Distributions



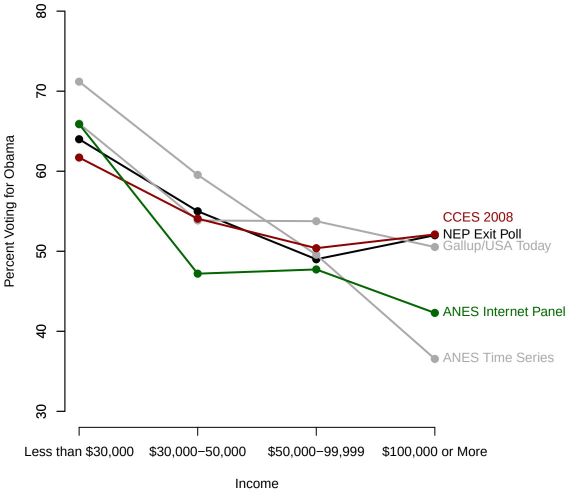


80
70
60
50
40
Percent Voting for Obama
Gallup/USA Today
NEP Exit Poll
ANES Time Series
ANES Internet Panel
CCES 2008
18−24
25−29
30−39
40−49
50−64
65+
Age


Percent Voting for Obama
100
90
80
70
60
50
40
30
ANES Time Series
Gallup/USA Today
NEP Exit Poll
ANES Internet Panel
CCES 2008
White
Black
Hispanic
Gender



Percent Voting for Obama
30
40
50
60
70
80
Gallup/USA Today
ANES Time Series
CCES 2008
NEP Exit Poll
ANES Internet Panel
Male
Female
Gender


80
70
60
50
40
30
Percent Voting for Obama
CCES 2008
NEP Exit Poll
Gallup/USA Today
ANES Internet Panel
ANES Time Series
Less than $30,000
$30,000−50,000
$50,000−99,999
$100,000 or More
Income

Comparison of 2012 YouGov and NEP Exit Poll

	YouGov		NEP	
	Obama	Romney	Obama	Romney
Gender				
Male	44%	53%	45%	52%
Female	55%	43%	55%	44%
Age				
18-29	62%	34%	60%	37%
30-44	57%	40%	52%	45%
45-64	45%	53%	47%	51%
65+	38%	61%	44%	56%
Race				
White	41%	56%	39%	59%
Black	92%	7%	93%	6%
Hispanic	61%	37%	71%	27%

Notation

We draw a sample $i = 1, \ldots, n$ using SRS, where

$$\begin{aligned} Y_i &= \text{survey measurements} \\ X_i &= \text{covariates} \\ R_i &= \text{selection indicator} \end{aligned}$$

The realized sample size is

$$n_R = \sum_{i=1}^{n} R_i$$

The mean of Y in the realized sample is

$$\bar{Y}_R = \frac{1}{n_R} \sum_{I} Y_i$$

Model

- (X_i, Y_i, R_i) are independently and identically distributed (i.i.d.).
- X is discrete with population distribution $P\{X = x\} = \mu(x)$.
- Y is dichotomous with $P\{Y = 1|X = x\} = \theta(x)$.
- The parameter of interest is $\theta = P\{Y = 1\} = \sum_x \mu(x)\theta(x)$.
- The propensity score is $\pi(x) = P\{R = 1|X = x\}$.
- In probability sampling $\pi(x)$ is known (from the design).
- In simple random sampling, $\pi(x)$ is constant.
- The Horvitz-Thompson estimator $\bar{Y}_{\mathrm{HT}} = \sum_R Y_i/\pi(X_i)$ is an unbiased estimator of θ.
- What can we do when $\pi(x)$ is unknown?

Ignorable Selection

The survey measurements are independent of sample selection conditional upon the covariates.

$$P\{Y = y | X = x, I = 1\} = P\{Y = y | X = x\}$$

Survey measurements

Covariates

Sample selection

Post-stratification

- The (unweighted) sample mean $\bar{Y} = n_R^{-1} \sum_R Y_i$ is biased if $\pi(x)$ is not constant.
- The Horvitz-Thompson estimator cannot be computed when $\pi(x)$ is unknown.
- The post-stratification estimator

$$\bar{Y}_{\mathrm{PS}} = \frac{1}{n_R} \sum_R W_i Y_i$$

  where $W_i = W(X_i)$ are the weights: $W(x) = \mu(x)/\hat{\mu}(x)$.
- By Bayes' rule,

$$P\{X = x | R = 1\} \propto P\{X = x\} P\{R = 1 | X = x\} = \mu(x)\pi(x)$$

  so W_i is approximately proportional of $1/\pi(X_i)$.
- If selection is ignorable, then $\bar{Y}_{\mathrm{PS}}$ is unbiased conditional upon the sample fractions $\hat{\boldsymbol{\mu}}$.

Model-based Variance Estimates

- It can be shown that the (conditional) variance of $\bar{Y}_{\text{PS}}$ is

$$\begin{aligned}
V(\bar{Y}_{\text{PS}}|\hat{\boldsymbol{\mu}}) &= \frac{1}{n_R^2}\sum_R W_i^2 V(Y_i|X_i) \\
&= \frac{\theta(1-\theta)}{n_R}\frac{1}{n_R}\sum_R W_i^2 \frac{\theta(X_i)[1-\theta(X_i)]}{\theta(1-\theta)} \\
&\approx V(\bar{Y}|n_R)(1+\text{CV}^2)(1-R^2_{Y\cdot X})
\end{aligned}$$

  where $\text{CV} = \text{SD}(W)/\bar{W}$ is the coefficient of variation of the weights and $R^2_{Y\cdot X}$ is the $R^2_{Y\cdot X}$ is the ratio of the variance of Y between categories of X to the total variation of Y.
- Confidence intervals: $\bar{Y}_{\text{PS}} \pm 1.96\sqrt{V(\bar{Y}_{PS}|\hat{\boldsymbol{\mu}})}$ will have approximately 95% coverage.

Non-random Selection with Ignorability

- For non-random samples, there is *some* probability $\pi(x)$ of a person with characteristics $X = x$ being included in the sample, but it is not known.
- Under the assumption of ignorability, the post-stratified estimator is unbiased.
- Ignorability is an assumption—there are no guarantees!
- The same variance estimator is valid, so we can form valid confidence intervals without knowing the selection probabilities.
- If ignobility fails, then the variance calculation is still valid, but the variance and mean square error are different.
- Ignorability can be tested when a probability sample is available with the same covariates.

The Limits of Post-stratification

- If $\hat{\mu}(x)$ is far from $\mu(x)$, then the weights will have a large coefficient of variation, and the post-stratified estimator can have a very large variance.
- Weights in excess of ten are common, and arbitrary trimming is often employed to reduce variability (at the cost of introducing bias).
- With purposive selection, we can choose $\hat{\mu}(x) \approx \mu(x)$ so the weights are nearly constant (similar to proportional allocation in stratified sampling), reducing or eliminating the need for post-stratification.
- Matching is an efficient computational algorithm for implementing purposive selection when X is high-dimensional or continuous.
- Rivers (2007) shows that matching on a single continuous covariate introduces an error $O_P(1/n)$. Abadie and Imbens (2007) have results for higher dimensions, though there is a curse of dimensionality.