Inferential Problems with Nonprobability Samples

Richard Valliant

University of Michigan & University of Maryland

9 Sep 2015

Not all nonprobability samples are created equal

- College sophomores in Psych 100
- Mall intercepts
- Volunteer samples, river samples, snowball samples
- Probability samples with low response rates

Coalitions of the willing

AAPOR task force report on non-probability samples (2013)

Not all nonprobability samples are created equal

- College sophomores in Psych 100
- Mall intercepts
- Volunteer samples, river samples, snowball samples
- Probability samples with low response rates

Coalitions of the willing

AAPOR task force report on non-probability samples (2013)

Not all nonprobability samples are created equal

- College sophomores in Psych 100
- Mall intercepts
- Volunteer samples, river samples, snowball samples
- Probability samples with low response rates

Coalitions of the willing

AAPOR task force report on non-probability samples (2013)

Declining response rates

- Pew Research response rates in typical telephone surveys dropped from 36% in 1997 to 9% in 2012 (Kohut et al. 2012)
- With such low RRs, a sample initially selected randomly can hardly be called a probability sample
- Low RRs raise the question of whether probability sampling is worthwhile, at least for some applications
  - Non-probs are faster, cheaper
  - No worse?

Polls that failed

British parliamentary election May 2015

Party	Final	Ipsos/MORI (online panel)	East Anglia/LSE/Durham U (using poll aggregation)
Conservative	51%	36%	43%
Labour	36%	35%	41%

Israeli March 2015 election (seats); online panels

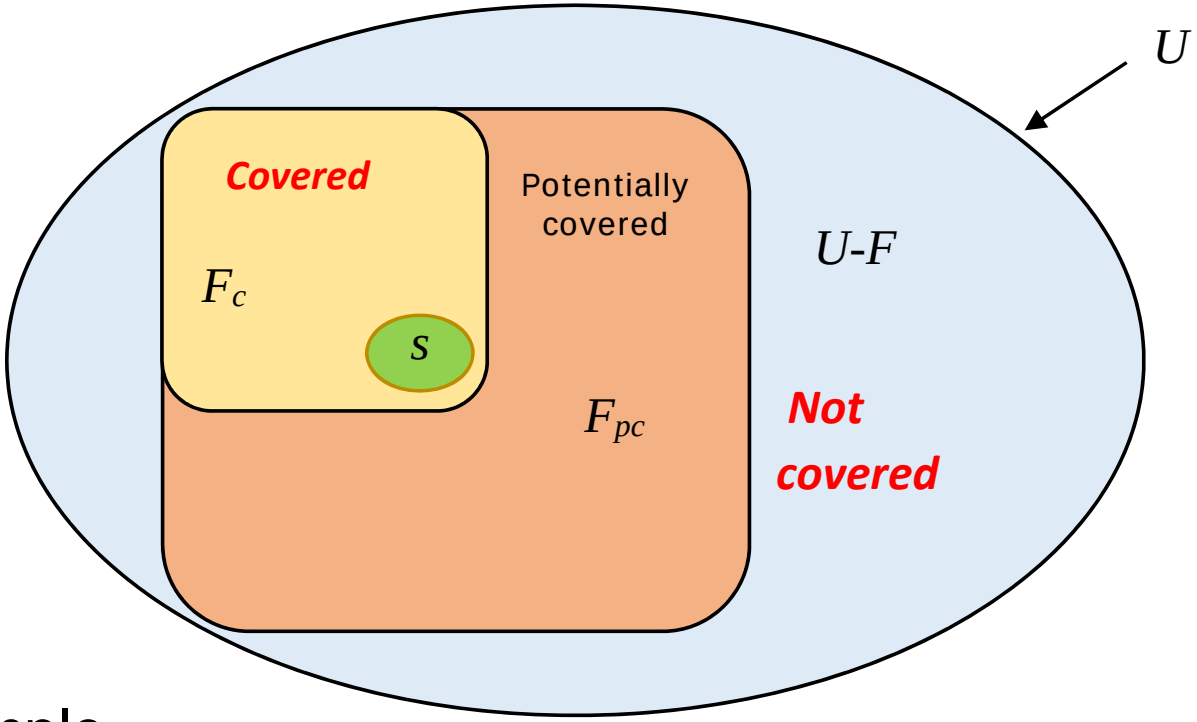
Party	Final	Smith-Reshet Bet	TNS/Walla	Maariv	Channel 1
Likud	30	21	23	21	25
Zionist Union	24	25	25	25	25

One that worked

- Xbox gamers: 345,000 people surveyed in opt-in poll for 45 days continuously before 2012 US presidential election
- Xboxers much different from overall electorate
  18- to 29-year olds were 65% of dataset, compared to 19% in national exit poll
  93% male vs. 47% in electorate
- Unadjusted data suggested landslide for Romney
- Gelman, et al. used some sort of regression and poststratification to get good estimates
- Covariates: sex, race, age, education, state, party ID, political ideology, and who voted for in the 2008 pres. election.

Wang, W., D. Rothschild, S. Goel, and A. Gelman. 2015. Forecasting Elections with Non-representative Polls. *International Journal of Forecasting*

Universe & sample



For example ...

- U = adult population
- F_{pc} = adults with internet access
- F_c = adults with internet access who visit some webpage(s)
- s = adults who volunteer for a panel

Ideas used in missing data literature

- MCAR–Every unit has same probability of appearing in sample
- MAR–Probability of appearing depends on covariates known for sample and nonsample cases
- NINR–Probability of appearing depends on covariates and y's

Table: Percentages of US households with Internet subscriptions; 2013 American Community Survey

	Percent of households with Internet subscription
Total households	74
Race and Hispanic origin of householder	
White alone, non-Hispanic	77
Black alone, non-Hispanic	61
Asian alone, non-Hispanic	87
Hispanic (of any race)	67
Household income	
Less than $25,000	48
$25,000-$49,999	69
$50,000-$99,999	85
$100,000-$149,999	93
$150,000 and more	95
Educational attainment of householder	
Less than high school graduate	44
High school graduate	63
Some college or associate's degree	79
Bachelor's degree or higher	90

Estimating a total

- Pop total $t = \sum_s y_i + \sum_{F_c - s} y_i + \sum_{F_{pc} - F_c} y_i + \sum_{U - F} y_i$
- To estimate t, predict 2nd, 3rd, and 4th sums
- What if non-covered units are much different from covered?
  - ▶ No 70+ year old Black women in a web panel
  - ▶ No 18-21 year old Hispanic males in a phone survey
- Difference from a bad probability sample with a good frame but low RR:
  ▶ No unit in $U - F$ or $F_{pc} - F_c$ had any chance of appearing in the sample

Full pop vs. Domains

- If domain is completely or mostly in the uncovered part ($U - F$, $F_{pc} - F_c$), then direct domain estimates not possible
  - Small area approach where $n_D = 0$ might be tried
- Full pop estimates may be OK if uncovered are "like" covered

Quasi-randomization

Model probability of appearing in sample

$$Pr(i \in s) = Pr(has\ Internet) \times$$
$$Pr(visits\ webpage \mid Internet) \times$$
$$Pr(volunteers\ for\ panel \mid Internet,\ visits\ webpage) \times$$
$$Pr(participates\ in\ survey \mid Internet,\ visits\ webpage,\ volunteers)$$

Reference sample

- Select a probability sample from a frame with good coverage
- Combine probability and non-probability samples together
- Estimate probability of being in non-probability sample using logistic regression (or similar)
- Use inverse probability as a Horvitz-Thompson-like weight

What does this probability mean?

In a volunteer sample, there are people who would never visit recruiting webpage or never volunteer if they did visit

The probability has no relative frequency interpretation

Reference sample

- Select a probability sample from a frame with good coverage
- Combine probability and non-probability samples together
- Estimate probability of being in non-probability sample using logistic regression (or similar)
- Use inverse probability as a Horvitz-Thompson-like weight

What does this probability mean?

> In a volunteer sample, there are people who would never visit recruiting webpage or never volunteer if they did visit

The probability has no relative frequency interpretation

Reference sample

- Select a probability sample from a frame with good coverage
- Combine probability and non-probability samples together
- Estimate probability of being in non-probability sample using logistic regression (or similar)
- Use inverse probability as a Horvitz-Thompson-like weight

What does this probability mean?

> In a volunteer sample, there are people who would never visit recruiting webpage or never volunteer if they did visit

The probability has no relative frequency interpretation

Superpopulation model

- Use a model to predict the value for each nonsample unit
- Linear model: $y_i = \mathbf{x}_i^T \beta + \epsilon_i$
- If this model holds, then

$$\begin{aligned}\hat{t} &= \sum_s y_i + \sum_{F_c - s} \hat{y}_i + \sum_{F_{pc} - F_c} \hat{y}_i + \sum_{U - F} \hat{y}_i \\ &= \sum_s y_i + \mathbf{t}_{(U-s),x}^T \hat{\beta} \\ &\doteq \mathbf{t}_{Ux}^T \hat{\beta}\end{aligned}$$

where $\hat{y}_i = \mathbf{x}_i^T \hat{\beta}$

Unit-level weights

- Prediction estimator does lead to weights

$$
\begin{aligned}
w_i &= 1 + \mathbf{t}_{(U-s),x}^T \left(\mathbf{X}_s^T \mathbf{X}_s \right)^{-1} \mathbf{x}_i \\
&\doteq \mathbf{t}_{Ux}^T \left(\mathbf{X}_s^T \mathbf{X}_s \right)^{-1} \mathbf{x}_i
\end{aligned}
$$

y's & Covariates

- If y is binary, a linear model is being used to predict a 0-1 variable
  - ▶ Done routinely in surveys without thinking explicitly about a model
- Every y may have a different model $\Rightarrow$ pick a set of x's good for many y's
  - ▶ Same thinking as done for GREG and other calibration estimators
- Undercoverage: use x's associated with coverage
  - ▶ Also done routinely in surveys

Modeling considerations

- Good modeling should consider how to predict y's and how to correct for coverage errors
- Covariates: an extensive set of covariates needed
  Dever, Rafferty, & Valliant (2008). *Svy. Rsch. Meth.*
  Valliant, Dever (2011). *Soc. Meth. Res.*
  Gelman, et al. (2015). *Intl. Jnl. Forecasting*
- Model fit for sample needs to hold for nonsample
- Proving that model estimated from sample holds for nonsample seems impossible

SE estimation

- Variance estimator must be model-based
- Replication is an option

  Jackknife or bootstrap

  Approximate jackknife
- Large sample approximation to jackknife

$$v_J = \sum_s \frac{w_i^2 r_i^2}{(1-h_{ii})^2} - n^{-1}\left[\frac{\sum_s w_i r_i}{(1-h_{ii})}\right]^2$$

  $r_i = y_i - \hat{y}_i$

  h_{ii} is a leverage
- v_J is consistent if the linear model holds with uncorrelated errors

Certification–Idea of Joe Sedransk

- AAPOR sets up certification program for panel purveyors
- Take questions from some ”legitimate survey” (CPS, HRS, NHIS, a good opinion poll)
- Panel vendor includes Q’s in its survey
  – Make estimates using vendor methods
  – Compare estimates to ones from legitimate survey
- Vendor must provide microdata and complete description of methods to be reviewed by AAPOR committee in order to be ”certified"