Non-probability Sampling for Finite Population Inference

Jill A Dever & Richard Valliant

RTI International & Universities of Michigan and Maryland

AAPOR Webinar (October 18, 2016)

Webinar Goals

- Understand the different types of non-probability samples currently in use
- Understand how non-probability samples can be affected by errors such as coverage and nonresponse
- Understand what methods of estimation can be used for non-probability samples and the arguments used to justify them
Motivation for Non-probability Sampling

- Low response rates for many probability samples (Kohut et al. 2012)
- Ever increasing costs with ever decreasing funds
- Nonsampling errors
- The need for speed
- Data are everywhere just waiting to be analyzed!!!

Examples of “New-ish” Sources of Data

- Twitter
- Facebook
- Snapchat
- Mechanical Turk
- SurveyMonkey
- Web-scraping
- Pop-up Surveys
- Data warehouses
- Probabilistic matching of multiple sources
New Sources of Data: Example Studies

- Analysis of medical records including text to predict heart disease (Giles & Wilcox 2011)
- Correlates of local climate & temperature with spread of infectious disease (Global Pandemic Initiative)
- MIT's Billion Prices Project–Price indexes for 22 countries from web-scraped data
- Marketing of e-cigarettes (Kim et al. 2015)
- Political polls and political issues (e.g., Clement 2016; Conway et al. 2015; Dropp & Nyhan 2016)
- Prediction of social stability (e.g., Kleinman 2014)
- Public health events, outbreaks (Harris et al. 2014; Kim et al. 2012)
- Research on subscribers to PatientsLikeMe.com
- Ad-hoc surveys via Amazon’s Mechanical Turk
- Google flu and dengue fever trends (defunct)

Probability vs. Non-probability Samples

Probability sampling:
- Presence of a sampling frame linked to population
- Every unit has a known probability of being selected
- Design-based theory focuses on random selection mechanism
- Probability samples became touchstone in surveys after Neyman (JRSS 1934) article

Non-probability sampling:
- Investigator does not randomly pick units with KNOWN probabilities
- No population sampling frame available/desired
- Underlying population model is important
- Differing opinions on reporting estimates of error
Probability vs. Non-probability Samples

- We focus on surveys with the goal to use sample to make estimates for *entire finite population*—external validity.
- Many applications of big data analysis use non-probability samples. Population may not be well defined.
- Many probability surveys have such low RRs they basically are non-probability samples.
  - Pew Research response rates in typical telephone surveys dropped from 36% in 1997 to 9% in 2012 (Kohut et al. 2012)
- Recommendations for using non-probability samples:
  - AAPOR task force reports on non-probability samples (2013) & online samples (2010)
  - Perils and potentials of self-selected entry (Keiding & Louis 2016)

Three Categories of Non-probability Samples

- **Convenience**—units at hand selected; notion overlaps with accidental, availability, opportunity, haphazard or unrestricted sampling.
- **Matched**—units are drawn into study (panel) based on characteristics, i.e., controlled selection.
- **Network**—a set of units form starting seeds, which sequentially lead to additional units selected (aka snowball, respondent driven sampling).
  (Note: Sirken network sampling is an exception)
Types of Convenience Samples

- **Volunteer sampling**—recruitment at events (e.g. sports, music, etc.) and other locations (e.g. mall intercept, street recruitment), limited (if any) refusal conversion
- **River sampling**—general or study-specific invitation through banner/pop-up web ads, etc.
- **Mail-in surveys**—type of volunteer sampling with paper-and-pencil questionnaires, distributed as leaflets at public locations (e.g. hotels, restaurants) or enclosed in magazines, journals, newspapers, etc.
- **Tele-voting** (text message)—type of volunteer sampling where people are invited to express their vote by calling-in or by sending a text (TV shows, contests)
- **Observational**—“you get what you see”

Types of Matched Samples

- **Purposive sampling**—selection follows some judgment or arbitrary ideas of looking for a “representative” sample
- **Expert selection**—subject experts pick the units, e.g., two most typical settlements selected from a region
- **Quota sampling**—sample “improved” by obtaining targeted socio-demographic quotas (e.g. region, gender, age) to reflect population distribution
- **Balanced sampling**
**Comments on Balanced Sampling**

- Samples selected until means or other quantities match the population (Särndal et al. 2003)
- Estimates are either unweighted (e.g., *average*) or via a model
- Quota sampling is a subset and focuses only on observable characteristics
- Shown to protect against misspecified inferential models (Royall & Herson 1973; Valliant et al. 2000)
- For probability-based balanced sampling
  - Survey weights are required (e.g., Horvitz-Thompson estimation)
  - Cube method randomly chooses from a set of balanced samples (Deville & Tillé 2004)

**Survey Errors**

- Coverage
- Selection bias
  - Coverage and/or selection bias is a problem if the seen (sample) part of the population differs from the unseen (nonsample) in such a way that the sample cannot be projected to the full population
- Nonresponse
  - (some unknown nonresponse for non-probability surveys)
- Attrition
- Measurement error (e.g., satisficing—provide an acceptable answer instead of the correct one)
Non-probability Electoral Polls: Many Failures

Early failure of a non-probability sample
- 1936 Literary Digest mail survey
- 2.3 million subscribers plus automobile and telephone owners
- Predicted landslide win by Alf Landon over FDR
- Excluded core lower-income supporters of FDR

More recent failures
- British parliamentary election May 2015
  Sturgis et al. (2016) is a post mortem
- Israeli Knesset election March 2015
- Scottish independence referendum, Sep 2014
- US state of Michigan democratic primary, 2016

Non-probability Electoral Polls: 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
- Wang et al. (2015) used multilevel regression and poststratification to get good estimates with covariates
  - sex, race, age, education, state, party ID, political ideology, and who they voted for in the 2008 presidential election
### Comparing Probability and Non-probability Samples

**Mixed results**

- Kennedy et al. (2016)—compared 9 non-probability and 1 probability sample
- Dutwin & Buskirk (2016)—some techniques show benefits (e.g., sample matching) but ....
- Tourangeau et al. (2013)—examined wt adjustments for 8 opt-in web panels using weight with mixed results
- Yeager et al. (2011)—compared RDD and non-probability internet survey with results varying by type of variable
- Valliant & Dever (2011)—effective propensity scores are possible with weighted reference survey cases

---

### Universe & Sample

![Diagram of Universe & Sample](image)

For example ...

- **$U$** = adult population
- **$U - F$** = adults without internet access
- **$F_{pc}$** = adults with internet access
- **$F_{c}$** = adults with internet access who visit some webpage(s)
- **$s$** = adults who volunteer for a panel
Illustration of a Coverage Problem

- Volunteer web panel surveyed about voting intentions
- Support for 2 candidates differs by age group
- Suppose the panel has no one in older groups

<table>
<thead>
<tr>
<th>Age group</th>
<th>Percentage of voters favoring (fictional)</th>
<th>Proportion of total Presidential vote in 2012</th>
</tr>
</thead>
<tbody>
<tr>
<td>18 to 24</td>
<td>47</td>
<td>0.09</td>
</tr>
<tr>
<td>25 to 44</td>
<td>46</td>
<td>0.30</td>
</tr>
<tr>
<td>45 to 64</td>
<td>40</td>
<td>0.39</td>
</tr>
<tr>
<td>65 to 74</td>
<td>30</td>
<td>0.13</td>
</tr>
<tr>
<td>75+</td>
<td>23</td>
<td>0.09</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>1.00</td>
</tr>
<tr>
<td>Total excl. 65 and older</td>
<td></td>
<td>40</td>
</tr>
<tr>
<td>Total excl. 75 and older</td>
<td></td>
<td>41</td>
</tr>
</tbody>
</table>

Correcting for Sample Imbalance

- Quota sampling or other type of controlled recruiting (YouGov/Polymetrix); no weights needed
- Weights to correct imbalance of sample compared to pop

- Two approaches to weighting
  - Quasi-randomization weighting
  - Superpopulation modeling of $y$'s

*Both involve modeling*
Flavors of Missing Data

- **MCAR** (Missing completely at random)
  — Every unit has the same probability of appearing in the sample

- **MAR** (Missing at random)
  — Probability of appearing depends on covariates known for sample and nonsample cases

- **NMAR** (Not missing at random)
  — Probability of appearing depends on covariates and $y$’s

Population Inference: Estimating a Total

- Pop total $t = \sum_{s} y_i + \sum_{F, s} y_i + \sum_{F, F_c} y_i + \sum_{U - F} y_i$

- To estimate $t$, predict 2nd, 3rd, and 4th sums

- What if non-covered units are much different from covered?

  - 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
Population Inference: Quasi-randomization Approach

Model probability of appearing in sample

\[ Pr(i \in s) = Pr(\text{has Internet}) \times \]
\[ Pr(\text{visits webpage} \mid \text{Internet}) \times \]
\[ Pr(\text{volunteers for panel} \mid \text{Internet, visits webpage}) \times \]
\[ Pr(\text{participates in survey} \mid \text{Internet, visits webpage, volunteers}) \]

Probabilities are sometimes estimated with special Reference (probability) sample or an existing sample (ACS, NHIS, etc.)

Propensity score method:
- Put \( s \) and reference sample together
- Estimate pseudo-inclusion probability \( \pi_i = Pr(i \in s) \) via binary regression
- Use \( 1/\pi_i \) as a weight
- Model covariates:
  - demographic items
  - webographic (attitudinal) items
    - mixed results (Schonlau et al. 2007; Lee et al. 2009)
  - covariates highly correlated with \( y \)'s (Lee 2006; Dever et al. 2015)
Population Inference: Quasi-randomization Approach

Binary regression to estimate propensity scores:
- Code non-probability cases = 1; reference cases = 0
- \( w_{\text{non-prob}} = 1 \) for non-probability sample cases
- \( w_{\text{Ref}} \) = survey weight for reference survey cases
- Propensities estimate probability of being in non-prob sample within whatever pop the reference weights to. Cases:
  - \( w_{\text{Ref}} \uparrow \) adult pop with internet access
  - \( w_{\text{Ref}} \uparrow \) adult pop regardless of internet access
- Caveats—reference survey weighting must correct for any coverage and nonresponse error
- Poststratification, raking, or other calibration often applied after getting pseudo-inclusion probabilities

Assumptions important for propensity score methods (Valliant & Dever 2011):
- Surveys are disjoint (no respondent overlap)
- Nonparticipants in non-probability survey are MAR
- Large reference survey from target population
- Identical key items on covariates in both questionnaires
- Propensity scores:
  - have common support in reference and non-probability (distributions overlap)
  - estimated with reference survey weights
Population Inference: Superpopulation “Prediction” Approach

- Use a model to predict the value for each nonsample unit (Valliant et al. 2000)
- Linear model: \( y_i = x_i^T \beta + \epsilon_i \)
- If this model holds, then

\[
\hat{t} = \sum_s y_i + \sum_{F_c-F_p} \hat{y}_i + \sum_{U-F_p} \hat{y}_i \\
= \sum_s y_i + t^T_{(U-s)x} \hat{\beta} \\
\hat{y}_i = x_i^T \hat{\beta}
\]

*Note*: Nonlinear models require individual \( x \)'s for nonsample units

Population Inference: Superpopulation (Prediction) Approach

\[
\hat{t} = \sum_s y_i + t^T_{(U-s)x} \hat{\beta}
\]

- \( \hat{\beta} = A_s^{-1}X_s^Ty_s \), where \( A_s = X_s^TX_s \)
- \( X_s \) is \( n \times p \) matrix of covariates for the sample units
- \( y_s \) is the \( n \)-vector of sample \( y \)'s

Resulting weight:

\[
w_i = 1 + t^T_{(U-s)x} A_s^{-1}x_i
\]

where \( t_{(U-s)x} \) = vector of \( x \) totals for nonsample units

*Note*: With this \( \hat{\beta} \), weights do not depend on \( y \)'s

Similar structure to generalized regression estimation (GREG)
Methods of Inference  
Model for $y$

## $y$’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
- Covariate selection: LASSO, CART, random forest, boosting, other machine learning methods
- Covariates: an extensive set of covariates needed
  (Dever, Rafferty & Valliant 2008; Valliant & Dever 2011; Wang et al. 2015)
- Model fit with sample needs to hold for nonsample (difficult [impossible?] to prove)
Pros and Cons with Quasi-Randomization and Superpopulation

Quasi-randomization
- Pro = general weights for estimating any $y$
- Con = possible bias with respect to the superpopulation model for $\pi_i$

Superpopulation
- Pro = model-specific estimators with lower variance than quasi-randomization
- Con = possible bias with respect to the superpopulation model for $y_i$

Notes: Model misspecification a worry for both
Bayesian variations available for each
See review paper by Elliott & Valliant (forthcoming)

Software

Quasi-randomization
- Propensity classes: pclass function in R PracTools package (Valliant et al. 2015)
- WTADJUST and WTADJX in SUDAAN 11 (Kott 2016; RTI 2012)
- Custom-written software in SAS, Stata, R, etc.

Superpopulation modeling
- calibrate function in R survey package (Lumley 2014)
- ReGenesees in R (Zardetto 2015)
- WTADJUST and WTADJX in SUDAAN 11 (Kott 2016; RTI 2012)
- ipfraking in Stata (Kolenikov 2014)
- sreweight in Stata (Pacifico 2014)
- svycal in future version of Stata

Set weights to 1 in design-based calibration routines
Simulation Study: Set-up (Valliant & Dever 2011)

- Data: 2003 Michigan Behavioral Risk Factor Surveillance Survey (MI BRFSS)
- 2,845 sample persons bootstrapped to \( N = 50,000 \) study population
- \( R = 10,000 \) simulation runs with two samples:
  - Volunteer sample
    - Volunteers selected by Poisson sampling; \( n = 500 \) (expected)
    - Logistic regression for volunteering; probabilities based on having internet access
    - Volunteering probabilities generated with logistic regression with covariates: age, race, gender, wireless phone, education, income
  - Reference sample—srswor of \( n = 500 \) from non-volunteers

Simulation Study: 4 Estimators Evaluated

- individual propensity weights (1: propensity wts)
- average propensity weights in each of five subclasses (2: avg propensity wts)
- propensity-poststratified estimator (3: propensity PS)
- calibration to population totals of covariates (no propensity adjustment) using a regression estimator (4: calibration to X); example of a prediction estimator

10,000 simulations with 500 in each volunteer & reference samples
Simulation Study: Statistical Results

Unweighted propensity parameter estimates

Weighted propensity parameter estimates

Simulation Study: Key Findings

- Reference survey weights need to be used to estimate propensities of volunteering
- Estimates with individual propensities or average propensity weights within classes are biased with unweighted propensity estimates, but less so with weighted
- Propensity-poststratification poor with unweighted or weighted propensity estimates
- GREG and estimate with individual propensity weights generally have smallest biases
- If probability of volunteering depends on $y$ analysis variables, all estimators are biased
Other Research

- Desire to compare estimates from non-probability against “the truth” leads researchers to contrast probability and non-probability surveys
- Quasi-randomization techniques do not always work (e.g., Dever & Brown 2016; Willis et al. 2015; Rothschild & Goel 2014; Valliant & Dever 2011; Yeager et al. 2011; Lee & Valliant 2009; Schonlau et al. 2009; Rivers 2007; Duffy et al. 2005)
- Limited comparisons with model-based estimation
- Lingering concerns
  - Were right covariates available?
  - Were they used correctly—multiway interactions?
  - Poor modeling leads to biased estimators

Variance estimation

Quasi-randomization
- Treat pseudo-inclusion probabilities in same way as designed-based selection probabilities
- Design-based variance estimators apply. Justification is consistency under quasi-randomization distribution
- Linearization or replication can be used
  Replication shows most promise (Lee & Valliant 2009)
- Need to decide whether strata and clusters are appropriate

Superpopulation modeling
- Compute variance under model used for point estimates with variance based on squared residuals
- Replication estimators also justified (Valliant et al. 2000)
- Bayesian models, e.g., credibility interval (Santos, Buskirk & Gelman 2012) with(out) applying survey design effects
- Justification is consistency under superpopulation model
Takeaway

- Non-probability samples do not have the (false?) assurance of complete population coverage that probability samples do
- Inference to finite populations is possible but only with either correct modeling of
  - Chance of being in sample, or
  - Dependence of analysis variables on covariates
- Convincing users that a non-probability sample represents nonsample part of population will always be an issue (true for low RR probability samples, too)

Diagnostics

- Work needed on diagnostics for "representativity"
- Are non-probability estimates aiming at desired target population?
- Distance measure
  \[ D^2 = (\hat{\theta} - \theta)^T [\text{cov}(\hat{\theta} - \theta)]^{-1} (\hat{\theta} - \theta) \]
  
  Compare to a chi-square distribution or \( F \)
- Validation items in \( \hat{\theta} \) are not used in non-probability weight calculation; may not be of direct interest in the survey
The Future

- Quasi-randomization—model pseudo-inclusion probabilities
- Superpopulation models—model the \( y \)'s
- Combination

Which is better???
References