Nonprobability Samples: Problems & Approaches to Inference

Richard Valliant

University of Michigan & University of Maryland

Washington Statistical Society
25 Sep 2017

1. Probability vs. nonprobability sampling
2. Inference problem
3. Methods of Inference
   - Quasi-randomization
   - Superpopulation Models for y's
4. Numerical example
5. Conclusion
6. References

Two Classes of Sampling

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]

Nonprobability sampling:

- Investigator does not randomly pick sample units with KNOWN probabilities
- No population sampling frame available (desired)
- Underlying population model is important

Review paper: [Elliott & Valliant, StatSci 2017]

[Vehover Toepoel & Steinmetz, 2016]

Types of Nonprobability Samples

AAPOR panel on nonprob samples defined three types [Baker. et al., AAPOR 2013]:

- Convenience sampling—mall intercepts, volunteer samples, river samples, observational studies, snowball samples
- Sample matching—members of nonprobability sample selected to match set of important population characteristics
- Network sampling—members of some population asked to identify other members of pop with whom they are somehow connected

Examples of Data Sources

- Twitter
- Facebook
- Snapchat
- Mechanical Turk
- SurveyMonkey
- Web-scraping
  - Billion Prices Project @ MIT, `http://bpp.mit.edu/`
  - Price indexes for 22 countries based on web-scraped data
  - Google flu and dengue fever trends
- Pop-up surveys
- Data warehouses
- Probabilistic matching of multiple sources

---

see, e.g., [Couper, SRM 2013]

Probability vs. Nonprobability samples

- Many applications of big data analysis use non-probability samples. Population may not be well defined.
- Goal in surveys is to use sample to make estimates for *entire finite population*—external validity
- Many surveys have such low RRs they 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], [Baker. et al., AAPOR 2013], [Keiding & Louis, JRSS-A 2016]

Electoral Poll Failures

- Early failure of a nonprobability sample
  - 1936 Literary Digest; 2.3 million mail surveys to subscribers plus automobile and telephone owners
  - Predicted landslide win by Alf Landon over FDR
  - Out-of-balance sample, no weighting to correct

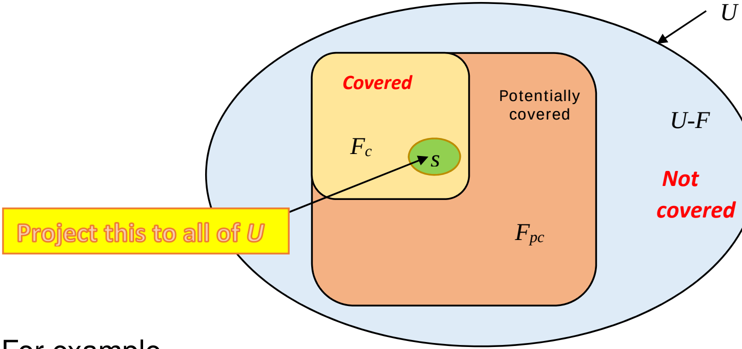
- More recent failures
  - British parliamentary election May 2015
  - Israeli Knesset election March 2015
  - Scottish independence referendum, Sep 2014
  - State polls in 2016 US presidential election
  - Out-of-balance samples, weighting did not correct, last minute decisions by voters

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 Bayesian regression and poststratification (MRP) to get good estimates
- Covariates: sex, race, age, education, state, party ID, political ideology, and who voted for in the 2008 pres. election.

[Wang, Rothschild, Goel, and Gelman, IJF 2015]

Universe & sample



For example ...

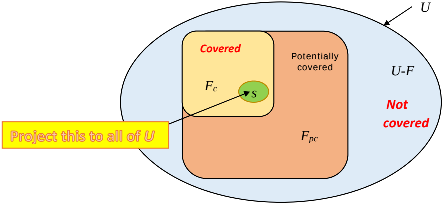
- U = adult population
- F_{pc} = adults with internet access
- F_c = adults with internet access who visit some webpage(s)
- s = adults who volunteer for a panel

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
- NMAR–Probability of appearing depends on covariates and y's

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?
- 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

Quasi-randomization

Model probability of appearing in sample

$$
\begin{aligned}
&Pr(i \in s) = Pr(has\ Internet)\times \\
&\quad Pr(visits\ webpage \mid Internet)\times \\
&\quad Pr(volunteers\ for\ panel \mid Internet,\ visits\ webpage)\times \\
&\quad Pr(participates\ in\ survey \mid Internet,\ visits\ webpage,\ volunteers)
\end{aligned}
$$

Sometimes done with *Reference* (probability) sample

Reference samples

- Reference sample is probability sample (or a census) from target pop
- Reference should cover *entire* target pop—no coverage errors
- Combine nonprobability and reference samples
- Code nonprob=1 and give weights=1; ref=0 with weights=survey weight
- Fit weighted binary regression and predict probability that a nonprob case is observed
  $p(\mathrm{x}_i; \theta)$ a function of covariates
- Weight for unit i is $1/p(\mathrm{x}_i; \theta)$

Estimation requirements

- **Common support**: for each value of $\mathbf{x}$, the probabilities of being in nonprobability sample and in reference sample are both positive
- **Common covariates**: the nonprobability and reference samples need to collect the same covariates in the same way

Common support is probably violated in many applications since some persons have zero probability of volunteering

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

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

- $\hat{\beta} = \mathbf{A}_s^{-1} \mathbf{X}_s^T \mathbf{y}_s$, where $\mathbf{A}_s = \mathbf{X}_s^T \mathbf{X}_s$
- $\mathbf{t}_{(U-s),x}$ = vector of x totals for nonsample units

Weights from superpopulation model

$$
\begin{aligned}
w_i &= 1 + \mathbf{t}^T_{(U-s),x} \mathbf{A}_s^{-1} \mathbf{x}_i \\
&\doteq \mathbf{t}^T_{Ux} \mathbf{A}_s^{-1} \mathbf{x}_i
\end{aligned}
$$

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

Similar structure to generalized regression estimation (GREG)

Prediction theory is covered in [Valliant, Dorfman, & Royall, 2000]

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, SRM 2008]
  [Valliant & Dever, SMR 2011]
  [Wang, Rothschild, Goel, and Gelman, IJF 2015]
- Model fit for sample needs to hold for nonsample
- Proving that model estimated from sample holds for nonsample seems difficult (impossible?)

Comments on Balanced Sampling

- Units selected until sample means or other quantities match the population [Sarndal & Lundquist, JSSAM 2014]
- Estimates are either unweighted (e.g., *average*) or via a model
- Quota sampling is a subset and focuses only on observable characteristics
- Some types of balance protect against misspecified inferential models [Valliant, Dorfman, & Royall, 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, BMKA 2004]

Two ways to compute weights I

- Two methods of estimation:
  (1) Quasi-randomization weights using nonprobability sample + a reference sample
  (2) Superpopulation model
- Dataset derived from the 2003 Behavioral Risk Factor Surveillance Survey (BRFSS) (Valliant & Dever, SMR 2011)
- 2,645 `mibrfss` cases are bootstrapped out to a reference "population" of 20,000.
- About 60% of persons have Internet at home
- Sample 200 persons who had access to Internet at home; stratified with older persons being less likely to volunteer for the sample and younger ones being more likely.

Sample distribution I

Age group	18-24	25-34	35-44	45-54	55-64	65+
Proportion in pop	0.056	0.134	0.197	0.226	0.170	0.217
Proportion in sample	0.120	0.310	0.185	0.205	0.135	0.045

- Sample is far out-of-balance
- Assign volunteers an initial weight of 1
- Select *srswor* reference sample from the full population. (De-duplicate if necessary)
- Weights in reference sample: N/n
- Reference sample and volunteer sample are combined
- Weighted logistic regression fitted to predict probability of being in volunteer sample using as covariates age, race, education level, and income level.

- Quasi-randomization
  - Predicted probabilities estimated with `svyglm` in R `survey`
  - Weights = 1/(pseudo-probs)
  - Sum is 19553, compared to pop size of 20,000.
  - Pseudo-weights range: 21.96 to 662.63
- Superpopulation model
  - Weights computed with `calibrate` in R `survey`
  - Bounded calibration used to avoid negative weights
  - Sum is 20,000, exactly pop size of 20,000.
  - Model-based weights range: 31.68 to 540.72

Other algorithms are available for bounding weights:
[Folsom & Singh, Proc SRM 2000], [Kott, Surv Meth 2006],
[Chang & Kott, BMKA 2008], [Kott & Chang, JASA 2010]

Proportion	Population value	Quasi-rand	Model-based	Unweighted volunteers
Smoked 100 cigarettes	0.530	0.561	0.548	0.480
		(0.048)	(0.050)	
Excellent health	0.179	0.216	0.212	0.285
		(0.036)	(0.037)	
Good or better health	0.843	0.896	0.870	0.940
		(0.036)	(0.037)	

$\Rightarrow$ Both options perform about the same here

Conclusion—Two general approaches to inference

- Quasi-randomization
  - $\approx$ Design-based (DB) inference—existing randomization theory applies
  - Pseudo-probabilities of selection apply to unit not a particular y
    $\Rightarrow$ Approach has generality of DB inference
- Superpopulation modeling
  - “Standard” model-based inference
  - Model can be different for every y
    • But, search for general set of covariates and use linear model weights to give standard set of weights
  - Modeling can be frequentist or Bayesian
  - Can allow use of more covariates than quasi-randomization as long as pop totals are available

References I

- Baker R, Brick JM, Bates NA, Battaglia M, Couper MP, Dever JA, Gile K, & Tourangeau, R (2013). *Report of the AAPOR Task Force on Non-Probability Sampling. The American Association for Public Opinion Research.*

  `http://www.aapor.org/AAPORKentico/AAPOR_Main/media/MainSiteFiles/NPS_TF_Report_Final_7_revised_FNL_6_22_13.pdf`

- *Valliant R & Dever JA (2017).*

  Survey Weights: A Step-by-step Guide to Calculation. *College Station: StataPress.*

- *Valliant R, Dever, JA, & Kreuter, F (2013).* Practical Tools for Designing and Weighting Sample Surveys. New York: Springer.

References II

- Valliant R, Dorfman A & Royall RM (2000). *Finite Population Sampling and Inference: A Prediction Approach. New York: Wiley.*

- *Chang, T & Kott PS (2008). Using calibration weighting to adjust for nonresponse under a plausible model.*

  Biometrika, 95, 3, 555-571.

- *Couper MP (2013). Is the Sky Falling? New Technology, Changing Media, and the Future of Surveys.*

  Survey Research Methods, 7(3): 145-156.
  http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.685.6246&rep=rep1&type=pdf

- Dever, J,Rafferty, A and Valliant, R (2008). Internet surveys: Can statistical adjustments eliminate coverage bias?

  Survey Research Methods, 2, 47Ű62.

References III

Deville, JC & Tillé, Y (2004). Efficient Balanced Sampling: The Cube Method.

Biometrika, 91, 893-912.

Elliott, MR & Valliant, R (2017). Inference for Nonprobability Samples

Statistical Science, 32(2), 249Ű264.

Folsom, RE & Singh, AC (2000). The generalized exponential model for sampling weight calibration for extreme values, nonresponse, and poststratification.

Proc. Survey Res. Meth. Sect., 598-603. Washington, DC: American Statistical Association.

Keiding N & Louis TA (2016). Perils and potentials of self-selected entry to epidemiological studies and surveys.

Journal of the Royal Statistical Society A, 179, 319-376.

References IV

Kott, PS (2006). Using Calibration Weighting to Adjust for Nonresponse and Coverage Errors.
Survey Methodology, 32(2): 133-142.

Kott, PS & Chang, T (2010). Using Calibration Weighting to Adjust for Nonignorable Unit Nonresponse.
Journal of the American Statistical Association, 105(491), 1265-1275.

Neyman J (1934). On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection.
Journal of the Royal Statistical Society, 97: 558-625.

References V

- Särndal, CE & Lundquist, P (2014). Accuracy in Estimation with Nonresponse: A Function of Degree of Imbalance ond Degree of Explanation.

  Journal of Survey Statistics and Methodology, 2, 361-387.

- *Valliant R & Dever JA (2011). Estimating propensity adjustments for volunteer web surveys.*

  Sociological Methods and Research, 40: 105-137.

- Vehovar V, Toepoel V, & Steinmetz S (2016). Non-probability sampling.

  in *The SAGE Handbook of Survey Methodology, chap. 22. London: Sage.*

- *Wang W, Rothschild D, Goel S, & Gelman A (2015). Forecasting Elections with Non-representative Polls.*

  International Journal of Forecasting, 31, 980-991.

References VI

Kohut A, Keeter S, Doherty C, Dimock M, & Christian L (2012).
Assessing the representativeness of public opinion surveys.
15 May 2012, http://www.people-press.org/2012/05/15/assessing-the-representativeness-of-public-opinion-surveys