Exploring Sampling Techniques to Reduce Respondent Burden

Wei, Yijun(Frank)
USDA-NASS
Yijun.Wei@nass.usda.gov

Bejleri, Valbona
USDA-NASS
Valbona.Bejleri@nass.usda.gov

Outline

- Purpose of research
- Brief review of sampling techniques
  - Probability sampling
  - Model-based sampling
  - Model-aided sampling
  - Model-assisted sampling
- Sampling procedures used at NASS
- Coordination function
- Simulation study
- Summary

Purpose of Research

Adopting a sampling design that will allow for

- Efficient estimators
  - Consistent
  - Efficient with respect to the variance
  - Unbiased
- Fixed sample size
- Simple implementation
- Optimal coordination of surveys
  - Small respondent burden

Review of Sampling Techniques

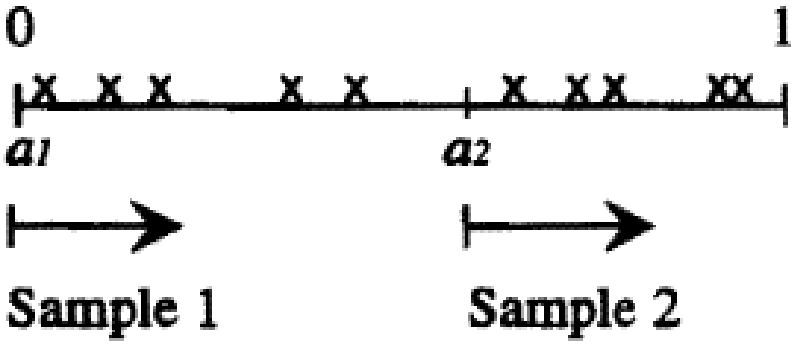
- Permanent random numbers, PRNs
- Poisson (or Bernoulli) Sampling
- Collocated random numbers, CRNs
- Births and deaths, and PRNs/CRNs
- Sample rotation
- Probability proportional to size
- Model-based, Model-aided, Model-assisted survey sampling
- Coordination function

Permanent random numbers (PRNs)

- Each unit $U_1, \ldots, U_N$ is associated with a random number X_i, $X_i \sim U(0,1)$, and X_i are i.i.d
- All the units are ordered ascendingly based on the random number $X_i s$ (i.e. order the $X_i s$ in ascending order)
- The first *n* units on the list are selected
- Applied in Jales technique (Ohlsson, 1992)
  - Associate the same random number (PRN) with the same unit
  - New businesses (births) are assigned new random numbers and closed-down (deaths) are withdrawn from the register

Positive/Negative Coordination of Two Surveys

- PRN (i.e. Permanent Random Number) is the same random number associated with the same unit



- Negative coordination starting points in each survey are far away from one another; a_1 and a_2 are far away

- Positive co-ordination is vice versa; a_1 and a_2 close to one another (Ohlsson,1992)

PRN- Equal Probability Starting Unit (EPSU)

- Sampling within an interval $(a\ b) \subset (0,1)$ instead of beyond some point a_0 is known as a Poisson sampling (Ohlsson, 1992) or Bernoulli sampling (Sarndal, Swensson, & Wretman, 1992)
- Yields fixed sampling fraction but not a fixed sample size
- Disadvantages
  - If the goal is to coordinate two or more surveys by minimizing the overlap between them, if not well distributed, PRNs may cause problems
  - If all random numbers are falling in a small range, there would be starting points that will not yield points for selection
  - Unbiased Horvitz –Thompson estimator is known to have very poor precision in connection with Poisson sampling (Ernst & Casady, 2000, and Ohlsson,1992)

Application of Permanent Random Number

- System for coordination of samples from the business register at Statistics Sweden (SAMU)
- All surveys in SAMU are stratified according to industry
  - Three hierarchical levels of stratification by industry from which to choose
  - Several other qualitative variables for stratification and selection, including but not limited to regions and variables related to ownership
  - Poisson sampling is used for sample selection
- SAMU has a built-in allocation system that can perform Neyman allocation on any of the size variables

Collocated Random Numbers

- Generate PRN $R_i \sim U(0,1)$ for each $X_i s$
- Record the rank R_i for each X_i
- Generate a single random number $\varepsilon \in U(0,1)$ and calculate

  $u_i = (R_i - \varepsilon)/N_0$ for each unit on the frame
- These u_i serve as new permanent random numbers for population units (Ernst & Casady, 2000)
- Spaces PRNs assigned to the population units equally
- Eliminates the clumping that can occur with PRNs
- For sample size of 20 or larger, PRN and CRN techniques have similar performance (Ernst & Casady, 2000)

Application of Collocated Random Numbers

- U.S. Bureau of Labor Statistics (BLS) routinely conducts several large-scale establishment surveys in a Federal/State cooperative environment
  - Current Employment Statistics (CES), Occupational Employment Statistics (OES), Occupational Safety and Health (OSH), Job Opening and Labor Turnover Statistics (JOLTS)
- Primary source of BLS is business establishment list (BEL)
  - Reports for each unemployment insurance (U.I.) account for 7 million records
- CRN used by BLS
  - to achieve the amount of sample overlap desired within a survey from one time period to another
  - to minimize the overlap in samples among different surveys

Births and Deaths from the Sample

- Birth: unit was not in the sample previously, but is rotated in at a later time
- Two ways of PRN deal with birth
  - Create separate strata for birth, apply the same strata definition, or broader strata definition
- CRN dealing with birth
  - A random number from $U(0,1)$ is drawn independently for each birth unit, the entire population is re-ordered
- Death: unit was in sample previously, but is rotated out at a later time
- Appropriate when entire sample is being rotated
- Used by several agencies, such as: Bureau of Labor Statistics (BLS) for Occupational Employment Statistics Survey (Ernst & Casady, 2000)

Probability Proportional to Size – πps Scheme

- Inclusion probabilities π_k proportional to size s_k, $k = 1,2, \ldots, N$
- Admits sample coordination over time and between surveys
- $y = y_1, y_2, \ldots, y_k, \ldots y_N$ – unknown characteristic of population
- $s = s_1, s_2, \ldots, s_k, \ldots s_N$, where $s_k > 0$, and *k = 1, 2, . . . , N* - known characteristics of population

$$\lambda_k = n s_k \Big/ \sum_{j=1}^{N} s_j$$

- Systematic πps, Random frame order, Frame ordered by size πps, Sunter πps

Traditional Probability Sampling

- For well-defined population of interest, design the sample so that selected units are in some sense representative of the whole population (Smith, 1983)
- Random selection process with known probabilities of selection (King, 1985)
- Can support inference only to the population implied by the sampling frame

Model-based Sampling Approach

- Model used to define the distribution of target population with respect to variable of interest (Stephenson, 1979)
- Not selected randomly and sampling weights not used
- Often incorporates poststratification for making descriptive inference of a specific population (Smith, 1983)
- Inference based on superpopulation model, using *a priori* sampling distribution achieved during data collection (King, 1985 and Deville, 1991)

Model-Aided Survey Sampling (MAS)

- Combining traditional probability sampling with quota sampling is a type of model-based sampling
- Units not selected randomly and sampling weights not used
- Simulation study of O*Net data (Berzofsky, 2008 and 2012) shows:
  - MAS retains probabilistic features underlying traditional sampling method
  - Keeps biasedness as traditional method, and if model assumptions hold, there is no bias in estimates produced (Deville,1991)
  - Reduces the response burden
  - Corrects skewness (avoids non-response bias) of survey
- MAS may not be effective design for initial data collection study for which little prior information exists about target population

Model assisted survey sampling

- Choosing an estimator that leads to a valid inference with respect to the sampling design, even if the model is misspecified by considering an optimal strategy rather than an optimal estimator

Model-based or model-assisted?

- When the superpopulation has fully explainable heteroscedasticity, one chooses the same sampling design for both approaches with inclusion probabilities proportional to standard deviations of errors of the model

- Horvitz-Thompson estimator is BLUE under model-assisted approach

Sampling Procedures Used at NASS

- Multivariate Probability Proportional to Size (MPPS)- employed in Crop Survey (CS)
- Sequential Interval Poisson Sampling (SIP)- employed in Agricultural Resource Management Survey (ARMS) to control overlap between ARMS from the previous year and the Crop APS sample for current year
- Stratified sampling - employed in all other surveys
- All fit into probability sampling approach

Multivariate Probability Proportional to Size (MPPS) at NASS

- Each farm i has unique probability π_i of selection,

  $$\pi_i = \min\{1, \max\{p_i^{(m)}, m = 1 \ldots M\}\},$$

  where $p_i^{(m)}$ is item m selection probability

- Probability is determined by available auxiliary data through optimal allocation consideration

- General assumption: variance is proportional to (a power of) the auxiliary data value, incorporating a desired item-level sample size

Sequential Interval Poisson Sampling (SIP) at NASS

- Employed in Agricultural Resource Management Survey (ARMS)
- Control overlap between ARMS from previous year and Crop APS sample for current year
- Sampling process where each element of population sampled is subjected to independent Bernoulli trial
- Each element of population may have different probability of being included in sample

Coordination Function and Response Burden

- Measurable and preserves *U(0,1)* probability
- Quantifies response burden, which enables to easily minimize or maximize it (Guggemos, 2012)
- Response burden is defined as: the number of times a unit is selected to participate in a survey

Steps:

- First sample S_1 is selected based on the permanent random number. Cumulative response burden is calculated for each chosen unit
- The *n*th sample S_n is selected using coordination function to update the random number for each unit and whether a unit will be selected based on its “new” random number

Coordination Function

- A simulation study is conducted, where 10 samples, consisting of 25 units each, are taken from the population of 100 simultaneously. The results showed:
  - Response burden is reduced by ~ 50%
  - It is quite robust against the volatility of parameters characterizing population and survey design
  - However, it is counterproductive when the sampling rate is too small

Simulation Results for Using Coordination Function in SRS

number of appearance	Coordination function(%)	SRS(%)	PPS(%)	SRS and PPS(%)
1	10.83	20.064	20.779	21.540
2	22.574	29.705	21.544	27.974
3	68.892	26.419	21.275	24.593
4	0.6671	15.615	17.081	15.458
5	0.0740	6.098	11.396	7.222
6	0.0039	1.713	5.375	2.483
7	0.0001	0.330	1.939	0.627
8	0	0.005	0.529	0.092
9	0	0.001	0.075	0.011
10	0	0	0.007	0

Coordination Function on Agricultural Yield Survey, Row Crop Survey, and Crop Sample Survey

number of appearance	Coordination function(%)	PPS(%)
1	79	76.3
2	21	22.8
3	0	0.9

Summary

- Sampling techniques along with their applications
- Coordination function tested on simulated data as well as on NASS survey data
- Coordination function works well if sampling rate is not too low

- Next step
  - Other sampling techniques will be tested in the future
  - Using simulated data and NASS survey data
  - Checking the efficiency of the estimator derived based on the new and current method.

Acknowledgements

This work has been done under SOSST of RRRT at USDA-NASS. Team members are:

Wendy Barboza, Franklin Duan, Jonathan Lisic, Brian Richards, Shareefah Williams, Valbona Bejleri, Yijun Wei

References

Berzofsky, M. E., Welch. B. L., Williams, R. L., and Biemer, P. 2008. *Using a model-aided sampling paradigm instead of a traditional sampling paradigm in a nationally representative establishment survey.* Research Triangle Park, NC: RTI Press.

Berzofsky, M. E., McRitchie, Brendle, M. B. (2012). Model-aided sampling: An empirical review. (2012). In proceedings of the Fourth International Conference on Establishment Surveys, 2012

Brewer, K. R. W., L. J. Early, and Muhammad Hanif. "Poisson, modified Poisson and collocated sampling." Journal of Statistical Planning and Inference 10.1 (1984): 15-30.

Deville, J.-C. (1991). A theory of quota surveys. Survey Methodology, 17, 163-181.

Ernst, L. R., and Casady, R. J.. (2000). Permanent and collocated random number sampling and the coverage of births and deaths. *Journal of Official Statistics* 16.3: 211

Guggemos, Fabien, and Olivier Sautory (2012). "Sampling coordination of business surveys conducted by Insee." In proceedings of the Fourth International Conference of Establishment Surveys.

King, B. (1985). Surveys combining probability and quota methods of sampling. Journal of the American Statistical Association, 80(392), 890-896.

Ohlsson, E. (1992). SAMU, The system for Co-ordination of Samples from the Business Register at Statistics Sweden. 18

Rosén, B. (1997). On sampling with probability proportional to size. *Journal of Statistical Planning and Inference*, *62*(2), 159-191.

Sarndal, C., Swensson, B. and Wretman, J. (1992). Model Assisted Survey Sampling. Springer-Verlag

Smith, T. M. F. (1983). On the validity of inference from non-random samples. Journal of the Royal Statistical Society: Series A., Pt. 4, 146, 394-403

Stephenson, C. B. (1979). Probability sampling with quotas: An experiment. *Public Opinion Quarterly*, *43*(4), 477-496.

Any Questions?

Thank you!