An Empirical Comparison of Efficiency between Optimization and Non-optimization Probability Sampling of Two Units from a Stratum

Sun-Woong Kim
Steven G. Heeringa
Peter W. Solenberger

Dongguk University
&
Survey Research Center
University of Michigan

Overview

- Unequal Probability Sampling Procedures with or without replacement
- Principle of Inclusion Probability Proportional to Size (IPPS) Sampling
- Desirable Requirements in IPPS Sampling
- Nonlinear Programming (NLP) Approaches
- Criteria to Improve Efficiency Relative to Existing Sampling Strategies
- Feasibility of NLP Approaches
- Comparison of Efficiency
- Discussion

Unequal Probability Sampling Procedures with or without replacement

There exist a variety of procedures useful to select 2 primary sampling units from each stratum in multistage sampling

We may be interested in

- Probability proportional to size with replacement (PPSWR) sampling procedure
- 50 probability proportional to size without replacement (PPSWOR) sampling procedures reviewed by Brewer and Hanif (1983)

Some inclusion probability proportional to size (IPPS) sampling procedures are widely used among them

Principle of IPPS sampling

Horvitz and Thompson (1952) produced a general theory of PPSWOR sampling based on the use of the following estimator of the population total Y

$$\hat{Y}_{HT} = \sum_{i=1}^{n} \frac{y_i}{\pi_i}$$

where y_i is the value of the characteristic of the ith unit and π_i is called the first-order inclusion probability .

Also, Sen (1953) and Yates and Grundy (1953) derived the variance and variance estimator of $\hat{Y}_{HT}$ respectively:

$$Var_{SYG}\left(\hat{Y}_{HT}\right)=\sum_{i=1}^{N}\sum_{j>i}^{N}\left(\pi_i\pi_j-\pi_{ij}\right)\left(y_i/\pi_i-y_j/\pi_j\right)^2$$

$$\widehat{Var}_{SYG}\left(\hat{Y}_{HT}\right)=\sum_{i=1}^{n}\sum_{j>i}^{n}\frac{\pi_i\pi_j-\pi_{ij}}{\pi_{ij}}\left(y_i/\pi_i-y_j/\pi_j\right)^2$$

where π_{ij} is called the second-order inclusion probability .

If the π_i are approximately proportional to the y_i, the variance (or the variance estimator) can be made close to zero.

The y_i are usually unknown in practice, but the auxiliary variable x_i correlated with the y_i may be available.

By setting the π_i proportional to the x_i, a substantial reduction in the variance (or variance estimator) can be achieved. In this case rather than the squared terms $\left(y_i/\pi_i - y_j/\pi_j\right)^2$ we may focus on the following terms.

$$\pi_i\pi_j - \pi_{ij}$$

Hence IPPS sampling strategy for the smaller values of these non-squared terms would be essential.

Desirable Requirements in IPPS Sampling

IPPS sampling designs satisfying the following desirable requirements are usually preferred:

(i) The π_i are strictly proportional to the x_i

(ii) The π_{ij} are larger than zero

(iii) The non-squared terms must be larger than zero, that is, $\pi_i\pi_j - \pi_{ij} > 0$

(iv) $\pi_{ij} \big/ \pi_i\pi_j > c$, where the value of c is positive and as far from zero as possible

For the details of (iv) for the stability of the variance estimator see Hanurav (1967), Nigam, Kumar and Gupta (1984) and Kim, Heeringa and Solenberger (2004).

Although it seems that those requirements are simple, the construction of IPPS sampling satisfying all of them is unlikely to be easy.

But nonlinear programming (NLP) approaches suggested by Kim, Heeringa and Solenberger (2003, 2004), which assures IPPS sampling designs to possess those requirements, seems to be easy to implement and useful to select two primary units per stratum.

NLP Approaches (2003, 2004)

Approach I:

$$Minimize \sum_{i=1}^{N}\sum_{j>i}^{N}\left(\pi_i\pi_j - \pi_{ij}\right)^2$$

Subject to i) the bounded linear constraints

$$c_{NLP}\pi_i\pi_j < \pi_{ij} \le \pi_i\pi_j, \quad 0 < c_{NLP} \le 1$$

ii) IPPS linear constraints

$$\sum_{j\neq i}^{N}\pi_{ij} = \pi_i$$

Note. c_{NLP} is the maximum value of c that allows a solution to the NLP problem

Approach II:

$$Maximize \sum_{i=1}^{N}\sum_{j>i}^{N}\pi_{ij}$$

under the same linear constraints

Approach III:

$$Minimize \sum \widehat{Var}_{SYG}\left(\hat{Y}_{HT}\right),$$

where the summation is over all possible samples, which is equivalent to

$$Minimize \sum_{i}^{N}\sum_{j>i}^{N}\frac{\pi_i\pi_j}{\pi_{ij}}$$

under the same linear constraints

Note. $E\left[\widehat{Var}_{SYG}\left(\hat{Y}_{HT}\right)\right]=Var_{SYG}\left(\hat{Y}_{HT}\right)$

Note. Let $\pi_{ij,NLP}$ denote the second-order inclusion probabilities obtained from these NLP approaches

Criteria to Improve Efficiency Relative to Existing Sampling Strategies

The bounded linear constraints in NLP approaches have some relationships with the variances as well as variance estimators for PPSWR sampling, Brewer's (1963) method, Hanurav's (1967) method and Murthy's (1957) method. See the followings for $n = 2$:

(1) NLP Approaches vs. PPSWR Sampling

1) Variance:

$$Var_{SYG}\left(\hat{Y}_{HT}\right)=\sum_{i=1}^{N}\sum_{j>i}^{N}p_i p_j\left(1-\frac{\pi_{ij,NLP}}{4p_i p_j}\right)\left(\frac{y_i}{p_i}-\frac{y_j}{p_j}\right)^2$$

where $p_i = x_i \Big/ \sum x_i$

$$Var\left(\hat{Y}_{PPS}\right)=\frac{1}{2}\sum_{i=1}^{N}\sum_{j>i}^{N}p_i p_j\left(\frac{y_i}{p_i}-\frac{y_j}{p_j}\right)^2$$

Note that if $1-\frac{\pi_{ij,NLP}}{4p_i p_j}<\frac{1}{2}$ for all i, j, which reduces to $\frac{1}{2}\pi_i\pi_j<\pi_{ij,NLP}$, then $Var_{SYG}\left(\hat{Y}_{HT}\right)<Var\left(\hat{Y}_{PPS}\right)$.

2) Variance Estimator:

$$\widehat{Var}_{SYG}\left(\hat{Y}_{HT}\right)=\left(\frac{p_i p_j}{\pi_{ij,NLP}}-\frac{1}{4}\right)\left(\frac{y_i}{p_i}-\frac{y_j}{p_j}\right)^2$$

$$\widehat{Var}\left(\hat{Y}_{PPS}\right)=\frac{1}{4}\left(\frac{y_i}{p_i}-\frac{y_j}{p_j}\right)^2$$

Note that if $\frac{p_i p_j}{\pi_{ij,NLP}}-\frac{1}{4}<\frac{1}{4}$ for all i, j, which reduces to $\frac{1}{2}\pi_i\pi_j<\pi_{ij,NLP}$, then $\widehat{Var}_{SYG}\left(\hat{Y}_{HT}\right)<\widehat{Var}\left(\hat{Y}_{PPS}\right)$.

Similarly, for Brewer's (1963) method and Hanurav's (1967) method that are IPPS sampling procedures and Murthy's (1957) method we find the following relationships.

(2) NLP Approaches vs. Brewer's method

1) Variance and 2) Variances Estimator

If the following is achieved for all i, j, then the smaller variance or variance estimator are obtained than in Brewer's method

$$\pi_{ij,B} < \pi_{ij,NLP} < \pi_i \pi_j,$$

where $\pi_{ij,B}$ indicates the second-order inclusion probabilities obtained from Brewer's method

and

$$\pi_{ij,B} = \frac{2p_i p_j}{D} \frac{1 - p_i - p_j}{(1-2p_i)(1-2p_j)}$$

$$D = \frac{1}{2} + \left(1 + \sum_{i=1}^{N} \frac{p_i}{1-2p_i}\right).$$

Note that since $\frac{1}{2}\pi_i\pi_j > \pi_{ij,B}$ or $\frac{1}{2}\pi_i\pi_j < \pi_{ij,B}$, we would prefer $\mathbf{\frac{1}{2}}\pi_i\pi_j < \pi_{ij,NLP}$.

(3) NLP Approaches vs. Hanurav's method

1) Variance and 2) Variances Estimator

For the smaller variance or variance estimator than in Hanurav's method the following is needed all i, j,

$$\pi_{ij,H} < \pi_{ij,NLP} < \pi_i \pi_j ,$$

where $\pi_{ij,H}$ indicates the second-order inclusion probabilities obtained from Hanurav's method

and

$$\pi_{ij,H} > \frac{1}{2}\pi_i\pi_j\alpha,\ \alpha = \frac{\left[1-\frac{1}{2}\beta\Big/\left(1-p_{(N)}\right)\right]^2}{(1-\beta)}$$

$$\beta = \frac{2(1-p_{(N)})(p_{(N)}-p_{(N-1)})}{1-p_{(N)}-p_{(N-1)}},$$

$p_{(N)}$ is the largest p_i and $p_{(N-1)}$ is the second largest p_i .

Note that since $\alpha < 1$, we would like $\frac{1}{2}\pi_i\pi_j < \pi_{ij,NLP}$.

(4) NLP Approaches vs. Murthy's method

1) Variance

If $\frac{\mathbf{1}}{\mathbf{2}-p_i-p_j}\pi_i\pi_j < \pi_{ij,NLP}$ all i, j, the smaller variance than in Murthy's method is achieved.

Note that $\frac{\mathbf{1}}{\mathbf{2}} < \frac{\mathbf{1}}{\mathbf{2}-p_i-p_j} < \mathbf{1}$ for $p_i > 0$.

2) Variance Estimator

For the smaller variance estimator the following is Needed:

$$\frac{\mathbf{1}}{c_{ij,M}+\mathbf{1}}\pi_i\pi_j < \pi_{ij,NLP}$$

$$c_{ij,M} = \frac{\mathbf{4}(\mathbf{1}-p_i)(\mathbf{1}-p_j)(\mathbf{1}-p_i-p_j)}{(\mathbf{2}-p_i-p_j)^{\mathbf{2}}}$$

Note that $\frac{\mathbf{1}}{\mathbf{2}} < \frac{\mathbf{1}}{c_{ij,M}+\mathbf{1}} < \mathbf{1}$.

In summary, we would prefer using the following linear constraints

$$c_{NLP}\pi_i\pi_j < \pi_{ij} \le \pi_i\pi_j, \text{ where } \frac{1}{2} < c_{NLP}.$$

Note that $c_{NLP} \approx \min \pi_{ij,NLP} \big/ \pi_i\pi_j$

Feasibility of NLP Approaches

16 natural populations (Rao and Bayless (1969)):

The population sizes from 9 to 20
The coefficients of variation from 0.14 to 0.98

Table 1. Comparison of $\min \pi_{ij} / \pi_i \pi_j$

No.	$CV(x)$	NLP	Brewer	Hanurav
1	0.14	0.54*	0.53	0.54*
2	0.17	0.54*	0.53	0.53
3	0.30	0.51*	0.50	0.51*
4	0.40	0.51*	0.49	0.50
5	0.43	0.52*	0.49	0.50
6	0.44	0.50*	0.47	0.50*
7	0.46	0.51*	0.48	0.50
8	0.50	0.51*	0.48	0.50
9	0.52	0.50*	0.48	0.50*
10	0.59	0.51*	0.45	0.49
11	0.65	0.51*	0.47	0.49
12	0.65	0.52*	0.44	0.47
13	0.71	0.47	0.49	0.50*
14	0.91	0.51*	0.45	0.49
15	0.93	0.50*	0.39	0.48
16	0.98	0.51*	0.43	0.50

Note. *: The largest value

Comparison of Efficiency

The percent gain in variance over Brewer's method:

$$\left[\left(Var(Brewer's\ \ est.)/Var(est.)\right)-1\right]\times 100$$

Table 2

No.	$CV(x)$	**N1**	**N2**	**N3**	**H**	**M**	**RHC**	**PPS**
1	0.14	+0	+0	+0	+0	1	1	–10
2	0.17	–0	–0	–0	–0	+0	+0	–11
3	0.30	+0	+0	+0	–0	+0	+0	–5
4	0.40	+0	+0	+0	+0	–1	–2	–7
5	0.43	–1	–1	–0	–0	1	1	–7
6	0.44	+0	+0	+0	+0	–0	–2	–7
7	0.46	+0	+0	+0	+0	+0	–0	–6
8	0.50	–0	+0	–0	–0	1	–0	–5
9	0.52	+0	+0	+0	+0	–0	–2	–8
10	0.59	–0	–0	3	1	–2	–6	–17
11	0.65	1	+0	2	+0	–0	–3	–10
12	0.65	–1	–1	–0	–1	6	5	–7
13	0.71	–1	–0	–1	–0	+0	–1	–4
14	0.91	–1	–1	–1	–0	4	3	–3
15	0.93	1	1	2	1	7	3	–9
16	0.98	+0	+0	–1	+0	6	4	–3

Note. +0 : positive value –0: negative value

. N1,N2, N3:NLP approaches, H: Hanurav's method
M: Murthy's method, RHC: Rao-Hartley –Cochran method
PPS: PPSWR

Among the three NLP approaches, the third is slightly better.

The third NLP approach is slightly better than non-optimization methods such as those of Brewer, Hanurav and Rao-Hartley-Cochran and compares favorably with Murthy's method when $CV(x)$ is smaller.

The percent gain in variance estimator over Brewer's method:

$$\left[\left(CV^2(Brewer's \ \text{var } est.)/CV^2(\text{var. } est.)\right)-1\right]\times 100$$

Table 3

No.	$CV(x)$	**N1**	**N2**	**N3**	**H**	**M**	**RHC**	**PPS**
1	0.14	+0	+0	+0	+0	3	5	–12
2	0.17	–1	–1	–1	–0	3	7	–12
3	0.30	+0	+0	1	+0	1	2	–5
4	0.40	2	1	1	1	–3	–5	–13
5	0.43	–2	–2	–0	–0	6	10	–6
6	0.44	–0	–0	–0	–0	1	1	–9
7	0.46	–0	–1	+0	–0	5	9	+0
8	0.50	–2	–1	–1	–0	4	8	–3
9	0.52	–2	–1	–2	–1	4	9	–3
10	0.59	–7	–5	–12	2	13	20	–6
11	0.65	–4	–3	–2	–2	4	5	–9
12	0.65	5	5	4	–1	16	27	3
13	0.71	+0	–1	2	1	2	4	–3
14	0.91	1	1	1	2	8	15	2
15	0.93	27	23	27	17	38	75	36
16	0.98	16	15	17	15	22	39	19

Murthy's method and Rao-Hartley-Cochran method are best.

Among the three NLP approaches, the third is slightly better.

The third NLP approach is slightly better than Brewer's method and compares favorably with Hanurav's method.

Discussion

- We have shown that the bounded linear constraints in NLP approaches are directly related to the variance as well as the variance estimator.

- We have suggested the criteria in NLP approaches to establish the smaller variance to the alternatives and they can be used for more stability of variance estimator.

- We have shown that the criteria can be achieved in practice when using NLP approaches.

- **The optimization sampling method using NLP appears to be better than other IPPS sampling methods such as the methods of Brewer and Hanurav.**

- **The optimization method would be more efficient when the strata have smaller** $CV(x)$**.**

- **In our next study we deal with the cases where the sample size is more than two.**