Probability Sampling Scheme for Variance Minimization

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

Dongguk University
&
Survey Research Center
University of Michigan

Overview

- Some Probability Sampling Schemes
- Horvitz-Thompson Estimator
- Jessen’s Method 4
- Non-linear Programming (NLP) Approaches
- Implementation of the Alternatives
- Numerical Examples
- Discussion

Some Probability Sampling Schemes

- **Sampling without replacement**

  Simple Random Sampling

  Yates and Grundy Method (1953)

  Raj's Method (1956)

  Murthy's Method (1957)

  Hartley and Rao Method (1962)

  Brewer's Method (1963)

- **Sampling with replacement**

  Probability proportional to size sampling

- **Rao, Hartley and Cochran Method (1962)**

Horvitz-Thompson (1952) Estimator

Notation

Selecting a sample of n out of the N units, without replacement, by some method

$P(S)$: selection probability of a sample S

π_i : probability that the ith unit is in the sample S

or

$\sum P(s)$ over all samples containing the ith unit

π_{ij} : probability that the ith and jth units are in the sample S

or

$\sum P(s)$ over all samples containing the ith and jth units

- **H-T estimator of the population total Y:**

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

where y_i is the characteristic of interest on the i unit

- **Several equivalent forms for the variance of H-T estimator, $Var(\hat{Y}_{HT})$:**

  - $\displaystyle \sum_{i=1}^{N} \frac{y_i^2}{\pi_i} + 2\sum_{i=1}^{N}\sum_{j>i}^{N} \frac{\pi_{ij}}{\pi_i \pi_j} y_i y_j - Y^2$

  - $\displaystyle \sum_{i=1}^{N} \frac{(1-\pi_i)}{\pi_i} y_i^2 + 2\sum_{i=1}^{N}\sum_{j>i}^{N} \frac{(\pi_{ij} - \pi_i \pi_j)}{\pi_i \pi_j} y_i y_j$

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

Jessen (1969)'s Method 4

- **Examined the influence of π_{ij} on the following:**

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

where $\pi_i = nP_i$**,** the P_i is the relative size of the ith unit,
$W_{ij} = \pi_{ij} - \pi_i\pi_j$

- **Considering the situation where the weight W_{ij} is a constant** $\overline{W} = \sum_{i=1}^{N}\sum_{j>i}^{N}W_{ij} /(N(N-1)/2)$ **with** $n = 2$

- **The desired π_{ij} for selecting samples of $n = 2$**

$$\pi_{ij} \doteq \pi_i \pi_j - \overline{W}$$

- **High statistical efficiency shown through several examples from the literature**

- **Some disadvantages of Method 4**

  ✓ **Difficult to employ in practical problems due to the arbitrariness and complexities of trials to meet the requirement that $\pi_i = nP_i$**

  ✓ **Repeated to find the exact variance of estimates**

  ✓ **Limited to samples of size $n = 2$**

Non-linear Programming (NLP) Approaches

Alternative I

Designates a set of π_{ij} such that the following is minimized :

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

which is equivalent to minimizing

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

subject to

$$\sum_{i\in s,s\in S^*}P(s)=\pi_i,\ i=1,\cdots,N$$

where S^* is a set of all possible samples

Alternative II

Finds a set of π_{ij} such that the following is directly minimized :

$$\sum_{i=1}^{N}\sum_{j>i}^{N}(\pi_i\pi_j - \pi_{ij}) = \sum_{i=1}^{N}\sum_{j>i}^{N}(n^2 P_i P_j - \pi_{ij})$$

which amounts to maximizing

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

under the same constraints as Alternative I

Implementation of the Alternatives

- Using SAS/OR NLP Procedure to optimize non-linear or linear objective functions under linear constraints

- Available several different non-linear programming algorithms by some options in finding a solution

- Not restricted to the sample of size $n = 2$

- Needed calculation of the selection probability of each sample

Numerical Examples

Table 1. Yates and Grundy (1953)

Three artificial populations with $N = 4$, $n = 2$

Unit	i:	1	2	3	4
Relative Sizes	P_i:	0.1	0.2	0.3	0.4
Population A	y_i:	0.5	1.2	2.1	3.2
	y_i/P_i	5	6	7	8
Population B	y_i:	0.8	1.4	1.8	2.0
	y_i/P_i	8	7	6	5
Population C	y_i:	0.2	0.6	0.9	0.8
	y_i/P_i	2	3	3	2

Table 2. Yates and Grundy (1953)

Second inclusion probabilities over all units

Units (i, j)	π_{ij}		
	Jessen Method 4	Alternative I	Alternative II
1, 2	0.010	0.013	0.000
1, 3	0.050	0.053	0.050
1, 4	0.140	0.133	0.150
2, 3	0.140	0.133	0.150
2, 4	0.250	0.253	0.250
3, 4	0.410	0.413	0.400

Table 3. Yates and Grundy (1953)

Comparison of variances of Alternative I and Alternative II with other schemes

Population	$Var(\hat{Y})$						
	PPS	**R**	**J 4**	**A I**	**A II**	**Y-G**	**H-R**
A	**0.500**	**0.200**	**0.245**	**0.253**	**0.225**	**0.323**	**0.367**
B	**0.500**	**0.200**	**0.245**	**0.253**	**0.225**	**0.269**	**0.367**
C	**0.125**	**0.100**	**0.070**	**0.067**	**0.075**	**0.057**	**0.033**
Average	**0.375**	**0.167**	**0.187**	**0.191**	**0.175**	**0.216**	**0.256**
Rel. Eff.	**100**	**225**	**201**	**196**	**214**	**173**	**147**

PPS: Sampling with probability proportional to size
R : Raj's method
J4 : Jessen's method 4
A 1 : Alternative I
A II: Alternative II
Y-G: Yates and Grundy method
H-R: Hartley and Rao method

Table 4. Cochran (1977)

Three artificial populations with $N=5,\ n=2$

Unit	$i:$	1	2	3	4	5
Relative Sizes	$P_i:$	0.1	0.1	0.2	0.3	0.3
Population A	$y_i:$	0.3	0.5	0.8	0.9	1.5
	y_i/P_i	3	5	4	3	5
Population B	$y_i:$	0.3	0.3	0.8	1.5	1.5
	y_i/P_i	3	3	4	5	5
Population C	$y_i:$	0.7	0.6	0.4	0.9	0.6
	y_i/P_i	7	6	2	3	2

Table 5. Cochran (1977)

Comparison of variances of Alternative I and Alternative II with other schemes

Population	$Var(\hat{Y})$						
	SRS	PPS	B	A I	A II	M	RHC
A	1.575	0.400	0.246	0.247	0.278	0.267	0.320
B	2.715	0.320	0.248	0.247	0.184	0.237	0.256
C	0.248	1.480	1.251	1.290	1.160	1.130	1.184
Average	1.513	0.733	0.582	0.594	0.541	0.545	0.587
Rel. Eff.	100	206	260	254	280	278	258

SRS : Simple random sampling
PPS : Sampling with probability proportional to size
B : Brewer's method
A 1 : Alternative I
A II : Alternative II
M : Murthy's method
RHC: Rao, Hartley and Cochran method

Discussion

- Alternative II preferable to Alternative I and other methods with respect to statistical efficiency and conveniences to carry out
- Both alternatives favored in the stratified multistage cluster sampling design, where two clusters are drawn from each stratum
- followed empirical comparisons for $n \geq 2$ in different populations
- Needed the examinations of stabilities and non-negativeness of the variance estimators such as Yates-Grundy form