Jour. Ind. Soc. Ag. Statistics
44 (2), 1992 : 138–148

On Comparing Sampling Strategies

P. Mukhopadhyay and Tapati Bagchi*
Indian Statistical Institute, Calcutta
(Received : March, 1990)

Summary

A comparison among several sampling strategies involving Deshpande's sampling design and Hansen-Hurwitz estimator has been made under a superpopulation model.

Key words : Horvitz Thompson estimator, Midzuno's sampling scheme, Deshpande's sampling scheme, PPS sampling.

Introduction

Consider a finite population U of N identifiable units labelled 1, 2, . . ., i, . . ., N. Associated with i are two real quantities (y_i, x_i), values of a study variable 'y' and a related auxiliary variable 'x' called size measure on unit i.

Deshpande [1] considered the following modification of Midzuno's [2] sampling design for estimating the population total $y = \sum_{i=1}^{n} y_i$. His sampling design P_D is as follows :

A subset s of n distinct units out of possible $\binom{N}{n}$ subsets of U and a number R in (O, Q) where $Q = \max_s \sum_{\bar{s}} p_k$, $\bar{s} = U - s$, $p_k = \frac{x_k}{X}$, $X = \sum_1^N x_k$, are chosen at random. If $R \le \sum_{\bar{s}} p_k$'s is selected as a sample; otherwise, the process is repeated involving fresh choices of a subset and a random number in (O, Q). Here and subsequently

* T.D.B. College, Raniganj (presently at Ashutosh College, Calcutta)

Table 2 (Contd.)

Cattle Crossbreds 3/4				
Order of lactation	2	3	4	5
2	0.09 (0.09)	0.89 (0.10)	0.85 (0.12)	0.96 (0.05)
3	0.58 (0.03)	0.21 (0.10)	0.91 (0.06)	0.67 (0.26)
4	0.36 (0.04)	0.64 (0.03)	0.21 (0.10)	**
5	0.23 (0.04)	0.42 (0.04)	0.72 (0.03)	0.10 (0.09)

Cattle Crossbreds 7/8				
2	0.19 (0.21)	**	0.13 (0.69)	–0.37 (0.36)
3	0.58 (0.06)	**	0.96 (0.17)	–0.52 (1.26)
4	0.21 (0.07)	0.44 (0.07)	0.24 (0.22)	–0.60 (0.24)
5	0.05 (0.07)	0.28 (0.07)	0.52 (0.06)	**

Note : ** Indicate inadmissible estimates. Figures in parentheses denote standard errors. Diagonal terms are heritabilities of stayability or survival. Values below diagonal are phenotypic correlations. Values above diagonal are genotypic correlations.

$\sum_s$ will denote sum over all $k \varepsilon s$. Similarly $\sum_s'$ will denote summation $k \neq k \varepsilon s$. For this scheme,

$$p(s) = \left[\binom{N-1}{n}\right]^{-1} \sum_s p_k$$

$$\pi_i = \left[\frac{n}{N-1}\right] (1 - p_i), \tag{1.1}$$

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

where $\pi_i = \sum_{s \ni i} p(s)$, $\pi_{ij} = \sum_{s \ni i,j} p(s)$.

An unbiased estimator of Y is

$$e_D = \frac{\bar{y}_s}{\bar{x}_{\bar{s}}} X \tag{1.2}$$

with $\bar{y}_s = \left(\frac{1}{n}\right) \sum_s y_k$, $\bar{x}_{\bar{s}} = \frac{1}{N-n} \sum_{\bar{s}} x_k$. Subsequently $\sum$ and $\sum'$ will denote $\sum_1^N$ and $\sum^N_{k \neq} \sum_{k'=1}$ respectively.

We have $$\hat{V}(e_D) = \frac{NX}{\binom{N}{n}} \sum_{s \ni \zeta} \frac{\bar{y}_s^2}{\bar{x}_{\bar{s}}} - Y^2 \tag{1.3}$$

with an unbiased variance estimator

$$v(e_D) = e_D^2 - \frac{X}{n\bar{x}_{\bar{s}}} \left[\sum_s y_i^2 + \frac{N-1}{n-1} \sum_s' y_i y_j\right], \tag{1.4}$$

ζ denoting the sample space.

By using $p'_i = 1 - (N-1) p_i$ in place of p_i the above design can be made a πps ($\pi_i \alpha p_i$, $i = 1, 2, \ldots, N$) design, p_D, (say) for which it is required,

$$\bar{x}_s > \frac{N(n-1)}{n(N-1)}\bar{X} \quad \forall \quad s, \tag{1.5}$$

$\bar{x}_s = \frac{1}{n}\sum_s x_k$, $\bar{X} = \frac{1}{N}\sum x_k$. We note that both Midzuno's design and D-design, when made πps have identical values of π_i and π_{ij} and hence give the same values of $V(e_{HT})$, e_{HT} denoting the Horvitz-Thompson estimator.

In what follows we shall compare the strategies (denoting by ppswr, e_{pps}, probability proportional to size with replacement scheme and the corresponding Hansen-Hurwitz estimator respectively),

(i) $(p_D, e_D) = H_o$

(ii) $(ppswr, e_{pps}) = H_1$

(iii) $(p_D, e_{HT}) = H_2$

(iv) $(p_{D'}, e_{HT}) = H_3$

under the following superpopulation model. It is assumed $\mathbf{y} = (y_1, y_2, \ldots, y_N)$ is a realisation of a vector of random variables $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_N)$, ($Y_i$ being the random variable corresponding to y_i), following a joint probability distribution ξ with

$$\begin{aligned} &\mathcal{E}\,(Y_i \mid x_i) = \beta(1-p_i) \\ &\mathcal{V}\,(Y_i \mid x_i) = \sigma_i^2 \qquad\qquad i = 1, 2, \ldots, N \\ &\mathcal{C}\,(Y_i\,,\,Y_{i'} \mid x_i, x_{i'}) = 0 \quad i \neq i', \end{aligned} \tag{1.6}$$

$\mathcal{E}, \mathcal{V}, \mathcal{C}$ denoting respectively expectation, variance and covariance with respect to ξ.

A strategy H_i will be said to be better than H_j ($H_i \succ H_j$) if $\mathcal{E}\,[V(H_i)] \leq \mathcal{E}\,[V(H_j)]$, $V[(H_k)]$ denoting design-variance of H_k, $k = 0, 1,2,3$.

2. Main Results

We have

$$V(H_o)=\left(\frac{N-n}{n}\right)^2\left[\binom{N-1}{n}\right]^{-1}\left[\sum \lambda_i Y_i^2+\sum{}' \lambda_{ij} Y_i Y_j\right]-Y^2 \tag{2.1}$$

where,

$$\lambda_i=\sum_{s\ni i}\left(1-\sum_{s} p_k\right)^{-1},\ i=1,2,\ldots,N \tag{2.2}$$

$$\lambda_{ij}=\sum_{s\ni i,j}\left(1-\sum_{s} p_k\right)^{-1},\ i\neq j=1,2,\ldots,N \tag{2.3}$$

$$V(H_1)=\frac{1}{n}\left[\sum \frac{Y_i^2}{p_i}-Y^2\right] \tag{2.4}$$

$$V(H_2)=\left(\frac{N-1}{n}\right)\sum\left[\frac{Y_i^2}{(1-p_i)}\right]+\left(\frac{N-1}{n}\right)^2\sum{}'\left[\frac{Y_iY_j}{(1-p_i)(1-p_j)}\right]$$
$$\left[\frac{n(n-1)}{(N-1)(N-2)}\right](1-p_i-p_j)-Y^2 \tag{2.5}$$

$$V(H_3)=\frac{1}{n}\sum\left(\frac{Y_i^2}{p_i}\right)+\frac{n-1}{n(N-1)(N-2)}$$
$$\sum{}'\left[\frac{Y_iY_j}{p_ip_j}\left\{(N-1)(p_i+p_j)-1\right\}\right]-Y^2 \tag{2.6}$$

Writing $\delta_k=\mathcal{E}\,[V(H_k)]$, $k=0,1,2,3$,

$$\delta_o=\beta^2\left[\left(\frac{N-n}{n}\right)^2\left\{\binom{N-1}{n}\right\}^{-1}\left\{\sum \lambda_i(1-p_i)^2+\sum{}'\lambda_{ij}(1-p_i)(1-p_j)\right\}-(N-1)^2\right]$$
$$+\sum \sigma_i^2\left[\left(\frac{N-n}{n}\right)^2\lambda_i\left\{\binom{N-1}{n}\right\}^{-1}-1\right] \tag{2.7}$$

$$\delta_1 = \frac{\beta^2}{n}\left[\sum\left(\frac{1}{p_i}\right) - N^2\right] + \frac{1}{n}\sum \sigma_i^2\left[\sum\left(\frac{1}{p_i}\right) - 1\right] \tag{2.8}$$

$$\delta_2 = \sum \sigma_i^2\left[\frac{(N-1)}{n(1-p_i)} - 1\right] \tag{2.9}$$

$$\delta_3 = \beta^2\left[\frac{\frac{1}{n}\sum(1-p_i)^2}{p_i} + \frac{n-1}{n(N-1)(N-2)}\right.$$

$$\left.\sum{}' \frac{(N-1)(p_i+p_j)-1}{p_i\, p_j}(1-p_i)(1-p_j) - (N-1)^2\right] + \sum \sigma_i^2\left[\frac{1}{np_i} - 1\right] \tag{2.10}$$

Lemma 1. For the sampling design p_D,

$$\lambda_i \geq \binom{N-1}{n-1}\frac{N-1}{(N-n)(1-p_i)} = \alpha_i \quad \text{(say)}$$

$$i = 1, 2, \ldots, N; \tag{2.11}$$

$$\lambda_{ij} \geq \binom{N-2}{n-2}\frac{N-2}{(N-n)(1-p_i-p_j)} = \alpha_{ij} \quad \text{(say)}$$

$$i \neq j = 1, 2, \ldots, N; \tag{2.12}$$

Proof. Since arithmetic mean ≥ harmonic mean,

$$\left[\binom{N-1}{n-1}\right]^{-1} \sum_{s \ni i}\left(1 - \sum_{s} p_k\right) \geq \binom{N-1}{n-1}\left[\sum_{s \ni i}\left(1 - \sum_{s} p_k\right)^{-1}\right]^{-1}$$

Hence the lemma.

Inequality (2.12) follows similarly.

Theorem 1. $H_o \succ H_1$ if

$$\alpha_i \leq \lambda_i \leq \binom{N-1}{n-1}\left[\frac{n}{N-n}\right]\left[\frac{1+(n-1)\,p_i}{np_i}\right] = \mu_i \quad \text{(say)} \tag{2.13}$$

and

$$\frac{\binom{N-1}{n-1}}{N-n}\left[(N-1)^2+\frac{(N-2)(n-1)}{N-1}\sum{}'\frac{(1-p_i)(1-p_j)}{1-p_i-p_j}\right]$$

$$\leq \sum_{i=1}^{N}\sum_{j=1}^{N}\lambda_{ij}(1-p_i)(1-p_j)$$

$$\leq \frac{\binom{N-1}{n-1}}{N-n}\left[\sum\frac{(1-p_i)^2}{p_i}+(N-1)^2(n-1)\right] \tag{2.14}$$

where $\lambda_{ii} = \lambda_i, \quad i = 1, 2, \ldots, N.$

Proof.

$$\delta_1 - \delta_0 = b\beta^2 + \sum c_i\ \sigma_i^2 \tag{2.15}$$

where

$$b = \sum\left[\frac{(1-p_i)^2}{np_i}-\frac{(N-1)^2}{n}\right]-$$

$$\left[\frac{N-n}{n}\right]^2\left[\binom{N-1}{n}\right]^{-1}\sum_{i=1}^{N}\sum_{j=1}^{N}\lambda_{ij}(1-p_i)(1-p_j)+(N-1)^2 \tag{2.16}$$

$$c_i = 1+(1/n)\left[\left(\frac{1}{p_i}\right)-1-\frac{(N-n)^2}{n}\lambda_i\left\{\binom{N-1}{n}\right\}^{-1}\right] \tag{2.17}$$

Now $\quad \delta_1-\delta_0 \geq 0$ if

$$b \geq 0 \text{ and } c_i \geq 0\ \forall\, i = 1, 2, \ldots, N. \tag{2.18}$$

(2.18) combined with lemma 1 gives (2.13) and (2.14).

Example 1. Consider the following values.

$N=5$, $n=2$; $p_1=0.14$, $p_2=0.20$, $p_3=0.21$, $p_4=0.22$ and $p_5=0.23$

Here

λ_1	6.20341	α_1	6.20	μ_1	10.86
λ_2	6.68859	α_2	6.67	μ_2	8.00
λ_3	6.77348	α_3	6.75	μ_3	7.68
λ_4	6.85921	α_4	6.83	μ_4	7.39
λ_5	6.94558	α_5	6.92	μ_5	7.13
λ_{12}	1.51515	λ_{23}	1.69492	λ_{34}	1.75439
λ_{13}	1.53846	λ_{24}	1.72414	λ_{35}	1.78571
λ_{14}	1.56250	λ_{25}	1.75439	λ_{45}	1.81818
λ_{15}	1.58730				

(Note that $\lambda_{ij} = \lambda_{ji}$, $i, j = 1, 2, \ldots, N$). Here

$$\sum_{1}^{5} \sum_{1}^{5} \lambda_{ij} \; (1 - p_i)(1 - p_j) = 43.31206.$$

Which is non-negative if (2.22) holds.

Example 4. For the population in Example 1,

$$\sum{}' \left\{(N-1)(p_i+p_j)-1\right\} \frac{(1-p_i)(1-p_j)}{p_i p_j} = 190.29306.$$

Again $(N-1)^3 (N-2) = 192$

Therefore, (2.22) is satisfied.

Actually, $\delta_1 - \delta_3 = 0.07112\beta^2 + 0.5 \sum \sigma_i^2$

i.e., $H_3 \succ H_1$.

Considering H_2 and H_3

$$\delta_3 - \delta_2 = b_4\beta^2 + \sum m_i \sigma_i^2 \tag{2.23}$$

Where $b_4 = \frac{1}{n}\left[\sum \left\{\frac{(1-p_i)^2}{p_i} + \frac{n-1}{(N-1)(N-2)}\right\} \sum{}' [(N-1)(p_i+p_j)-1] \left\{\frac{(1-p_i)(1-p_j)}{(p_i p_j)}\right\} - n(N-1)^2\right]$

and $m_i = \frac{1-Np_i}{np_i(1-p_i)}$

Now $m_i \geq 0 => 1-Np_i \geq 0 => p_i \leq \left(\frac{1}{N}\right) \forall\ i = 1, 2, \ldots, N$; but since $\sum_{1}^{N} p_i = 1$, the only possible value of $p_i = \frac{1}{N} \forall\ i = 1, 2, \ldots, N$, when $\delta_2 = \delta_3$.

Thus when σ_i^2 is arbitrary we can not come to any definite conclusion about superiority of one to the other.

Theorem 5. If $\sigma_i^2 \ \alpha\ p_i(1-p_i)$ and $p_i + p_j \geq \frac{1}{N-1}$, $i \neq j = 1, 2, \ldots, N$. (2.24)

then $H_2 \succ H_3$.

Proof. Putting $\sigma_i^2 = K p_i (1 - p_i)$, K being a constant, in (2.23) we get $\sum m_i \sigma_i^2 = 0$. Also when $p_i + p_j \geq \frac{1}{N-1}$, $1 \leq i \neq j \leq N$, $b_4 \geq 0$. Hence the theorem.

Example. 5. For the population in Example 1., the second set of conditions in (2.24) holds. Here

$$\delta_3 - \delta_2 = 0.32790\beta^2 > 0. \text{ Hence } H_2 \succ H_3.$$

3. *Discussion*

Under some situations the value of the main variable may be inversely related to its (only available) size-measure x, when the model (1.6) may be applicable. Under this model the performance of H_1 is worst among the strategies considered, which is not surprising, as H_1 should be used when y values vary directly with x-values. The main point of interest is comparison among H_o, H_2 and H_3 all of which involve D-sampling design, in which p(s) being proportional to the total size-measure of the units in the complementary subset, the model considered here seems to be of interest. Under certain conditions H_2 seems to be the best choice among H_o, H_2 and H_3 in the sense of minimum average variance.

REFERENCES

[1] Deshpande, M.N., 1977. A new sampling procedure with varying probabilities. *Jour. Ind. Soc. Agr. Stat.* **30**, 110-114.

[2] Midzuno, M., 1952. On the sampling system with probability proportional to sum of sizes. *Ann. Inst. Stat. Math.* **3**, 99-107.