ON HORVITZ-THOMPSON AND DES RAJ ESTIMATORS

By K. VIJAYAN

Indian Statistical Institute

SUMMARY. Under the usual superpopulation set up, it was shown by Godambe (1955) that the Horvitz-Thompson estimator with a design where the inclusion probabilities are strictly proportional to the *a priori* expectations will be the best among all the strategies with constant sample sizes, when the super population parameter g is 2. In this paper it is pointed out that when g is different from 2 there exist strategies better than the above strategy.

1. Introduction

Consider a finite population of N units,

$$u_1, u_2, \ldots, u_N.$$

We are interested in estimating the total value, Y of a certain characteristic y for this population. The value of y for u_i is denoted by Y_i.

Consider a sample of size two taken without replacement with probabilities of selection being proportional to

$$p_1, p_2, p_3, \ldots, p_N$$

where

$$\sum_{i=1}^{N} p_i = 1.$$

Let u_i and u_j be the units taken in order of their draw. Des Raj (1956) suggested the following estimate for Y:

$$T_0 = d_1t_1+d_2t_2 \qquad \ldots \quad (1.1)$$

where

$$d_1+d_2 = 1,$$

and

$$t_1 = \frac{Y_i}{p_i}$$

and

$$t_2 = \frac{Y_j}{p_j}(1-p_i)+Y_i.$$

(1.1) being an asymmetric estimator can always be improved by taking the weighted mean of different asymmetric estimators for given unordered sample, the weights being the corresponding conditional probabilities of obtaining the ordered samples given the unordered sample (Halmos, 1946). This estimate happens to be (Murthy, 1957)

$$T_1 = \frac{1}{2-p_i-p_j}\left\{\frac{Y_i}{p_i}(1-p_j)+\frac{Y_j}{p_j}(1-p_i)\right\}. \qquad \ldots \quad (1.2)$$

The variance of this estimator is

$$V(T_1) = \sum_{i=1}^{N} \frac{Y_i^2}{p_i} \sum_{j \neq i} p_j \frac{1-p_i-p_j}{2-p_i-p_j} - \sum\sum_{i \neq j} Y_i Y_j \left(\frac{1-p_i-p_j}{2-p_i-p_j}\right). \quad \dots \quad (1.3)$$

We shall call the design together with the estimator (1.2) as symmetrised Des Raj strategy.

Consider the sampling design, wherein all the samples are of the same size 2. Let us denote the inclusion probability of the i-th unit by π_i and the joint inclusion probabilities of the i-th and j-th units by π_{ij}. Horvitz and Thompson (1955) suggested the following estimate of the population total Y.

$$T_2 = \frac{Y_i}{\pi_i} + \frac{Y_j}{\pi_j}. \quad \dots \quad (1.4)$$

The variance of this estimate is

$$V(T_2) = \sum_{i=1}^{N} \frac{Y_i^2}{\pi_i}(1-\pi_i) + \sum\sum_{i \neq j} \frac{Y_i Y_j}{\pi_i \pi_j}(\pi_{ij} - \pi_i \pi_j). \quad \dots \quad (1.5).$$

In this paper, we consider designs satisfying

$$\pi_i = 2p_i \quad \dots \quad (1.6)$$

together with the estimate (1.5), and call it the Horvitz-Thompson strategy.

It is easy to note that there would be situations in which the Horvitz-Thompson strategy is better than symmetrised Des Raj strategy and other situations where symmetrised Des Raj strategy is better, for a fixed value of $(p_1, p_2, \dots, p_N)$. So, we would compare the two strategies under a superpopulation set up

2. Superpopulation set up

In many practical situations we would be knowing before hand, the values taken by another characteristic $\mathscr{X}$, which is highly correlated with $\mathscr{Y}$. (We denote the value taken by $\mathscr{X}$ on u_i by X_i). Now we consider the $\mathscr{Y}$-values as the values coming from an infinite population such that the expected value taken by Y_i on u_i for a given value of X_i, is proportional to X_i. (Cochran, 1946). We shall denote the conditional variance of Y_i given X_i by σ_i^2.

It was observed in field experiments that $V(Y_i/X_i)$ is of the form (Mahalanobis, 1944; Smith, 1938)

$$V(Y_i/X_i) = \sigma^2 X_i^g. \quad \dots \quad (2.1)$$

Through intuitive arguments it can be seen that in all practical situations, g lies between 1 and 2. We assume the *a priori* distribution satisfies (2.2) and (2.4).

In this paper we will be comparing the two strategies under this model, the criterion for betterness being smaller expected variance, with

$$p_i = \frac{X_i}{X}. \qquad \ldots \quad (2.2)$$

Hanurav (1962) has shown that when $g = 2$, the Horvitz-Thompson strategy is better than the symmetrised Des Raj strategy.

3. The expected variances of the two strategies

We can write down the expected variances of the two strategies as follows :

$$E_1 = \varepsilon(V(T_1)) = \sigma^2 X^g \sum_{i=1}^{N} p_i^{g-1} \sum_{j \neq i} p_j \frac{1-p_i-p_j}{2-p_i-p_j} \qquad \ldots \quad (3.1)$$

$$E_2 = \varepsilon(V(T_2)) = \sigma^2 X^g \, \Sigma p_i^{g-1} \left(\frac{1}{2} - p_i\right). \qquad \ldots \quad (3.2)$$

From (3.1) and (3.2), we get

$$E_1 - E_2 = \sigma^2 X^g \sum_i p_i^{g-1} \left\{ \sum_{j \neq i} p_j \frac{1-p_i-p_j}{2-p_i-p_j} - \frac{1}{2} + p_i \right\} = \sigma^2 X^g \sum_i p_i^{g-1} a_i \qquad \ldots \quad (3.3)$$

where

$$a_i = \frac{1}{2} - \sum_{j \neq i} \frac{p_j}{2-p_i-p_j} \qquad \ldots \quad (3.4)$$

$$= \frac{1}{2} \, \frac{1}{1-p_i} - \sum_j \frac{p_j}{2-p_i-p_j}$$

$$= \frac{1}{2} \sum_j \left\{ \frac{p_j}{1-p_i} - \frac{2p_j}{2-p_i-p_j} \right\}$$

$$= \frac{1}{2} \sum_j \frac{p_j(p_i-p_j)}{(1-p_i)(2-p_i-p_j)}. \qquad \ldots \quad (3.5)$$

4. Main theorems

In the sequel it would be understood that

$$p_1 \leqslant p_2 \leqslant \cdots \leqslant p_N. \qquad \ldots \quad (4.1)$$

Theorem 1 : *In the usual superpopulation model, the symmetrised Des Raj strategy is superior to the Horvitz-Thompson strategy when $g = 1$ and inferior when $g = 2$, except when all the p's are equal in which case the two strategies coincide.*

Proof : From (3.3) and (3.5), we see that the difference in variance between the symmetrised Des Raj strategy and the Horvitz-Thompson strategy, E, is given by

$$E = \frac{\sigma^2X^g}{4}\sum_i\sum_j\left\{\frac{p_i^{g-1}p_j(p_i-p_j)}{(1-p_i)(2-p_i-p_j)}+\frac{p_j^{g-1}p_i(p_j-p_i)}{(1-p_j)(2-p_i-p_j)}\right\}$$

$$= \frac{\sigma^2X^g}{4}\sum_i\sum_j\frac{p_ip_j(p_i-p_j)}{(1-p_i)(1-p_j)(2-p_i-p_j)}\{p_i^{g-2}(1-p_j)-p_j^{g-2}(1-p_i)\}$$

$$= \frac{\sigma^2X^g}{2}\sum_i\sum_{j<i}\frac{p_ip_j(p_i-p_j)}{(1-p_i)(1-p_j)(2-p_i-p_j)}\left\{p_i^{g-2}(1-p_j)-p_j^{g-2}(1-p_i)\right\} \quad \dots \quad (4.2)$$

$$= \begin{cases} -\dfrac{\sigma^2X}{2}\displaystyle\sum_i\sum_{j<i}\frac{(p_i-p_j)^2\,(1-p_i-p_j)}{(1-p_i)(1-p_j)(2-p_i-p_j)} & \text{if } g=1 \quad \dots \quad (4.3) \\[2ex] \dfrac{\sigma^2X^2}{2}\displaystyle\sum_i\sum_{j<i}\frac{p_ip_j(p_i-p_j)^2}{(1-p_i)(1-p_j)(2-p_i-p_j)} & \text{if. } g=2. \quad \dots \quad (4.4) \end{cases}$$

Theorem 1 follows from (4.3) and (4.4).

Lemma : *For given values of $p_1, p_2, \dots, p_N$, if the function*

$$\Sigma\, a_ip_i^{g-1} \quad \dots \quad (4.5)$$

where
$$a_i = \frac{1}{2}-\sum_{j\neq i}\frac{p_j}{2-p_i-p_j}$$

and
$$p_i > 0\ (\Sigma p_i = 1),$$

is negative when $g = g'$, then the function increases with g at that point.

Proof : We take

$$p_1 \leqslant p_2 \leqslant \dots \leqslant p_N.$$

It is clear from (3.5) that

$$a_1 \leqslant a_2 \leqslant \dots \leqslant a_k \leqslant a_{k+1} \leqslant \dots \leqslant a_N$$

where $a_1, a_2, \dots, a_k$ are all negative and $a_{k+1} \dots a_N$ are all positive.

Differentiating (4.5) w.r.t. g we get

$$\Sigma\, a_ip_i^{g-1}\log p_i = \sum_{i=1}^{k} a_ip_i^{g-1}\log p_i+\sum_{i=k+1}^{N} a_ip_i^{g-1}\log p_i$$

$$\geqslant \sum_{i=1}^{k} a_ip_i^{g-1}\log p_k+\sum_{i=k+1}^{N} a_ip_i^{g-1}\log p_k$$

$$\geqslant \left(\sum_{i=1}^{N} a_ip_i^{g-1}\right)\log p_k$$

where the equality sign holds good only if all the p_i's are equal. Since $\log p_i$ is negative we get that when $g = g'$

$$\frac{d}{dg} \Sigma a_i p_i^{g-1} > 0,$$

unless all the p's are equal. Hence the lemma.

Theorem 2 : *Under the usual super population model, given $p_1, p_2, \dots, p_N$, not all equal, there is a value for the super population parameter g, say g_0, where g_0 lies between 1 and 2, such that the Horvitz-Thompson strategy is more precise or less precise (in the expected variance sense) than symmetrised Des Raj's strategy according as $g > g_0$ or $< g_0$. When $g = g_0$ or p's are all equal the two strategies are equally efficient.*

Proof : Theorem follows very easily from Theorem 1 and Lemma 1.

5. Numerical examples

We consider here three populations each consisting of 4 units. The values of the characteristic is

population A	·1	·2	·3	·4
population B	·01	·20	·30	·49
population C	·23	·24	·26	·27

The populations were so chosen that in one x's are moderately spread, in one extremely spread and in the last uniform. The efficiency of the Horvitz-Thompson strategy compared to the symmetrised Des Raj strategy is tabulated for different values of g.

g	population A	population B	population C
1·0	·9398	·8530	·9988
1·1	·9482	·8752	·9990
1·2	·9568	·8993	·9992
1·3	·9658	·9240	·9994
1·4	·9749	·9513	·9995
1·5	·9843	·9805	·9997
1·6	·9939	1·0054	·9999
1·7	1·0037	1·0326	1·0000
1·8	1·0136	1·0601	1·0003
1·9	1·0236	1·0870	1·0005
2·0	1·0336	1·1470	1·0006

6. FINAL REMARKS

It was pointed out by Godambe (1955) that the Horvitz-Thompson stategy is the best strategy whenever the above superpopulation model holds good and g takes the value 2. But when g is nearer to 1, evidently there are other estimators which are better. We have to remember that sometimes in practice g comes near to 1 (Fairfield, 1938).

7. ACKNOWLEDGEMENT

The author wishes to express his thanks to Professor J. Roy and to the referee for their valuable suggestions. Thanks are also due to Mr. T. Parthasarthy for his kind help during the perparation of this paper.

REFERENCES

COCHRAN, W. G. (1946): Relative accuracy of systematic and stratified random samples for a certain class of populations. *Ann. Math. Stat.*, 17, 164-177.

DES RAJ (1956): Some estimators in sampling with varying probabilities without replacement. *J. Amer. Stat. Assoc.*, 51, 269-284.

GODAMBE, V. P. (1955): A unified theory of sampling from finite populations. *J. Roy. Stat. Soc.*, B, 17, 269-277.

HALMOS, P. R. (1946): The theory of unbiased estimation. *Ann. Math. Stat.*, 17, 34-43.

HANURAV, T. V. (1962): On Horvitz-Thompson estimator. Series A, *Sankhyā*, 24, 429-436.

HORVITZ, D. G. and THOMPSON, D. J. (1952): A generalisation of sampling without replacement from a finite universe. *J. Amer. Stat. Assoc.*, 47, 663-685.

MAHALANOBIS, P. C. (1944): On large scale sample surveys. *Roy. Soc. Phil. Trans.*, B, 231, 329-451.

SMITH, H. FAIRFIELD (1938): An empirical law governing soil heterogenity. *Jour. Agr. Sci.*, 28, 1-23.

Paper received : December, 1965.