Jour. Ind. Soc. Ag. Statistics
Vol. XLI, No. 1 (1989), pp. 104-112

MODIFIED HORVITZ-THOMPSON ESTIMATOR TRANSFORMING STUDY VARIATE UNDER IPPS SAMPLING SCHEME

B. V. S. SISODIA* and R. C. BHARATI
Deptt. of Statistics and Mathematics, R.A.U., Pusa (Samastipur)-848125

SUMMARY

Use of transformation of study variate under IPPS sampling schemes is investigated. Horvitz-Thompson (HT) estimator is modified and its properties are discussed. It is empirically shown that efficiency of modified Horvitz-Thompson estimator is quite substantial as compared to the usual HT estimator.

Keywords: Linear transformation, Sampling with varying Probabilities, IPPS sampling schemes.

Introduction

Horvitz and Thompson [5] developed a general theory of estimation in sampling with varying probabilities without replacement from finite population. Since then the development in sampling with varying probabilities without replacement centred around various aspects of the Horvitz and Thompson (*HT*) estimator.

Consider a population $U = (U_1, U_2, \ldots, U_N)$ consisting of N units. Associated with U_i $(i = 1, 2, \ldots, N)$ are two variables y_i (study variate) and x_i (auxiliary variate). It is assumed that x_i's are known. An

*Present address : Deptt. of Statistics, N.D.U.A.T., Kumarganj Faizabad-224229

unbiased estimator of $Y = \sum_{i=1}^{N} y_i$, given by Horvitz and Thompson is

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

where π_i is inclusion probability of ith population unit in the sample and n is fixed sample size. The Yates and Grundy (1953) form of variance of HT estimator is

$$V(\hat{Y}_{HT}) = \sum_{i>j=1}^{N} (\pi_i\pi_j - \pi_{ij})\,(y_i/\pi_i - y_j/\pi_j)^2 \tag{2}$$

where π_{ij} is inclusion probability of ith and jth population units together in the sample. An unbiased variance estimator of HT estimator is

$$\hat{V}(\hat{Y}_{HT}) = \sum_{i>j=1}^{n} \frac{(\pi_i\,\pi_j - \pi_{ij})}{\pi_{ij}}\,(y_i/\pi_i - y_j/\pi_j)^2 \tag{3}$$

It was, however, pointed out by Durbin [3] that the HT estimator may be less efficient than that based on PPS sampling with replacement for some set of inclusion probabilities. Another major drawback of the HT estimator is that its variance estimator my assume negative values for some samples.

If y_i/π_i remains constant for all i then $V(\hat{Y}_{HT})$ reduces to zero. Keeping this in mind and assuming $y_i = \beta x_i$, recently, Prasad and Srivenkataramana [7] considered a linear transformation of study variate as

$$Z_i = y_i + (n-1)\,b/(N-n)\,; \tag{4}$$

where b is some appropriate scalar quantity. Under Midzuno sampling scheme [6] this makes Z_i/π_i almost constant for all i. They accordingly modified the HT estimator and shown that it is more efficient than usual HT estimaror as long as $0 < b < 2\beta X$, $X = \sum_{i=1}^{N} x_i$. The use of transformation suggested by them is, however, limited to only Midzuno sampling scheme. Recently, Stuart [10] has dealt with the general theory of location shifts is sampling with unequal probabilities. In fact, one of the practical considerations while using HT estimator is that $\pi_i \pi_j > \pi_{ij}$; $\pi_{ij} > 0$, which ensures non-negative variance estimation. There is presently no dearth of inclusion probability proportional to size (*IPPS*) sampling schemes in the literature, which satisfy this condition. In this paper, a linear transformation of study variate under *IPPS* sampling schemes is proposed and consequently a modified *HT* (*MHT*) estimator is developed and its properties are studied.

2. Proposed Transformation and Modified HT Estimator

For any *IPPS* sampling scheme we know that $\pi_i = nX_i \Big/ \sum_{i=1}^{N} X_i$. Now, we propose to transform y to z by

$$Z_i = y_i - a \tag{5}$$

The modified Horvitz-Thompson (*MHT*) estimator of Y under *IPPS* sampling scheme with the transformation (5) is proposed as follows :

$$\hat{Y}_{MHT} = \sum_{i=1}^{N} Z_i/\pi_i + Na \tag{6}$$

We now prove the following theorem

THEOREM 2.1. The estimator $\hat{Y}_{MHT}$ is unbiased and its minimum variance is given by

$$V(\hat{Y}_{MHT}) \text{ min.} = V(\hat{Y}_{HT}) - W_2^2/W_1^2$$

where $W_1 = \sum_{i>j=1}^{N} (\pi_i \pi_j - \pi_{ij}) (1/\pi_i - 1/\pi_j)^2$, and

$$W_2 = \sum_{i>j=1}^{N} (\pi_i \pi_j - \pi_{ij}) (y_i/\pi_i - y_j/\pi_j) (1/\pi_i - 1/\pi_j)$$

Proof: Let us express $\hat{Y}_{MHT}$ as under

$$\hat{Y}(_{MHT}) = \sum_{i=1}^{N} \frac{t_i z_i}{\pi_i} + Na$$

where t_i is a random variable taking value one if the population ith unit occurs in the sample, otherwise it is zero. Obviously $E(t_i) = \pi_i$. Now taking expectation of $\hat{Y}_{MHT}$, and using (5) it follows that

$$E(\hat{Y}_{MHT}) = \sum_{i=1}^{N} Z_i + Na = \sum_{i=1}^{N} Y_i = Y \tag{7}$$

Thus, $\hat{Y}_{MHT}$ is unbiased estimator of Y.

Following the Yates-Grundy from of variance HT estimator, the variance of $\hat{Y}_{MHT}$ can easily be written as under

$$V(\hat{Y}_{MHT}) = \sum_{i>j=1}^{N} (\pi_i \pi_j - \pi_{ij}) (Z_i/\pi_i - Z_j/\pi_j)^2$$

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

from (5).

After little algebraic simplification, the final expression could be obtained as follows :

$$V(\hat{Y}_{MHT}) = V(\hat{Y}_{HT}) + a^2 W_1 - 2a W_2 \tag{8}$$

The optimum value of a is obtained by differentiating (8) w.r.t. a and equating is to zero, which is as follows :

$$a_{opt.} = W_2/W_1 \tag{9}$$

Therefore, the minimum variance of $\hat{Y}_{MHT}$ with optimum value of a from (9) is obtained as

$$V(\hat{Y}_{MHT_{min.}}) = V(\hat{Y}_{HT}) - W_2^2/W_1 \tag{10}$$

3. Efficiency of $\hat{Y}_{MHT}$

It is obvious from the expression (10) that the modified Horvitz-Thompson estimator $\hat{Y}_{MHT}$ will be more efficient than the HT estimator as W_1 is always positive quantity. When π_i is proportional to Y_i, W_2

reduces to zero and hence both the estimators are equally efficient. Therefore, the proposed modified (HT) estimator will fetch more precision as against HT estimator in case of Y_i departs from proportionality to the π_i.

It can further easily be shown from (8) that the $\hat{Y}_{MHT}$ will remain efficient than the HT estimator as long as a lies between 0 and $2a_{opt}$. Thus, we have wide range for a to be appropriately chosen.

4. Choice of a

Since in practice Y_i is not known for all units in the population, the optimum value of a can not be known. But, a reasonable choice of a can be made by following methods.

(i) Consider the simple linear regression equation of Y on x

$$Y_i = \alpha + \beta x_i \times \varepsilon_i \tag{11}$$

where α and β are intercept and regression coefficient respectively, and ϵ_i is an error term. Now, from (9) we see that

$$a_{opt} = \frac{\sum\limits_{i>j=1}^{N} (\pi_i \pi_j - \pi_{ij}) (y_i/\pi_i - y_j/\pi_j) (1/\pi_i - 1/\pi_j)}{\sum\limits_{i>j=1}^{N} (\pi_i \pi_j - \pi_{ij}) (1/\pi_i - 1/\pi_j)^2} \tag{12}$$

Substituting the value of y_i from (11) in the above expression and noting that $x_i/\pi_i = \sum\limits_{i=1}^{N} x_i/n$, i.e. constant for all $i = 1, 2, \ldots N$ in case of IPPS sampling, we get

$$a_{opt.} = \alpha + \frac{\sum\limits_{i>j=1}^{N} (\pi_i \pi_j - \pi_{ij}) (\epsilon_i/\pi_i - \epsilon_j/\pi_j) (1/\pi_i - 1/\pi_j)}{\sum\limits_{i>j=1}^{N} (\pi_i \pi_j - \pi_{ij}) (1/\pi_i - 1/\pi_j)^2} \tag{13}$$

In case of perfect correlation i.e. $y_i = \alpha + \beta x_i$, the equation (13) reduces to $a_{opt.} = \alpha$. Also, if y and x are highly correlated, the last term of the expression (12) is expected to be very small quantity near to almost zero and, therefore, under such situation the optimum value of a would be very close to α.

Using the following relations :

$$\sum_{\substack{j=1 \\ j \neq i}}^{N} \pi_{ij} = (n-1)\pi_i \; ; \; \sum_{i=1}^{N} \pi_i = n \text{ and } \sum_{i} \sum_{\substack{j \\ i \neq j}} \pi_{ij} = n(n-1)$$

the expression (12) can easily be written as

$$a_{opt} = \frac{\sum_{i=1}^{N} y_i/\pi_i = N^2 \bar{Y} + \sum_{i \neq j}^{N} \sum \pi_{ij}\, y_i/\pi_i\, \pi_j}{\sum_{i=1}^{N} 1/\pi_i - N^2 + \sum_{i \neq j}^{N} \sum \pi_{ij}/\pi_i\, \pi_j} \tag{14}$$

where $\bar{Y} = \sum_{i=1}^{N} y_i/N$. It has been shown by Stuart (10) that a_{opt} is regression coefficient of $\sum_{i=1}^{n} y_i/\pi_i$ on $\sum_{i=1}^{n} 1/\pi_i - N$. However, this can not lead to a practical solution of a_{opt} in practice.

Thus, an appropriate value of a could be obtained on the basis of past experience gathered in repeated surveys or by plotting y_i against x_i for the sample units and gauging the intercept of the best fitting line.

(ii) The numerator of (14) can be estimated unbiasedly by

$$e = \sum_{i=1}^{n} y_i/\pi_i^2 - N^2 \hat{\bar{Y}} + \sum_{i \neq j = 1}^{n} \sum y_i/\pi_i \pi_j$$

which can further be simplified as

$$e = \sum_{i=1}^{n} y_i/\pi_i \; \left(\sum_{i=1}^{n} 1/\pi_i - N \right)$$

Therefore, an unbiased estimate of a_{opt}. can be obtained by

$$a_{opt} \doteq \frac{\sum_{i=1}^{n} y_i/\pi_i \; \left(\sum_{i=1}^{n} 1/\pi_i - N \right)}{\sum_{i=1}^{N} 1/\pi_i - N^2 + \sum_{i \neq j = 1}^{N} \sum \pi_{ij}/\pi_i\, \pi_j} \tag{15}$$

Alternatively, another estimate of a_{opt}, though biased, can be obtained by taking summation only over sample units in the denominator of (15).

Thus, reasonably a good choice of a can be made from (15).

5. Empirrical Study

In all eight populations are considered to illustrate the efficiency of *MHT* estimator Y_{MHT} in comparison to *HT* estimator. First three artificial populations are due to Yates and Grundy [11], 4, 5 and 6th populations considered by Cochran ([2], page : 268), population 7 considered by Sampford [9] and a real poulation 8 (Y : no. of cattle, x : no. of farms) considered by Prasad and Srivenkataramana [7].

Consider three IPPS sampling schemes suggested by Brewer [1], Rao [8] and Durbin [4] for sample size $n = 2$, we assume every $P_i > 1/2$. Using different approaches, methods produced by them gave the same expression for π_i and π_{ij} which are as follow :

$$\pi_i = 2P_i$$

$$\pi_{ij} = \frac{2P_i P_j(1 - P_i - P_j)}{D(1 - 2P_i)(1 - 2P_j)}$$

where $D = 1/2 \left(1 + \sum_{i=1}^{P} \frac{P_i}{1 - 2P_i}\right)$

The values of W_1, W_2 and optimum value of a are computed for all the populations and are depicted in Table 1. The variance of *HT* and *MHT* estimators and relative efficiency of latter one over former, $E = 100\ \hat{V}(Y_{HT})/\hat{V}(Y_{MHT})$, are also computed and are presented in this table. It may be sèen from this table that relative efficiency of *MHT* estimator is substantially high for almost all the populations. The *MHT* and *HT* estimators are observed to be equally efficient in case of population 4, because the intercept a is zero, i.e. the linear regression line of y on x passes through the origin indicating thereby no need of the proposed transformation for the population.

Three IPPS sampling schemes for sample size two are considered owing to simplicity in illustrating results empirically.

To investigate the sensitivity of the relative efficiency of *MHT* estimator to departure from the optimum choice of a, the values of relative efficiency of *MHT* estimator are computed for different deviation from the optimum value of a for all the populations and are given in Table 2. It can be seen from this table that even if there is up to eighty percent deviation from the optimum value of a, the *MHT* estimator remains efficient than *HT* estimator.

ACKNOWLEDGEMENT

Authors are very much thankful to the referee for his valuable suggestions which brought improvement in the paper.

TABLE 1—RELATIVE EFFICIENCY OF MHT ESTIMATOR OVER USUAL HT ESTIMATOR

Population	W_1	W_2	a_{opt}	$V(\hat{Y}_{MHT})$	$V(\hat{Y}_{MHT})$	*Efficiency*(%)
1.	1.84227	—0.6633	0.36	0.282125	0.04331	651
2.	1.84227	0.6633	0.36	0.282125	0.04331	651
3.	1.84227	—0.1327	0.0721	0.059375	0.051069	116
4.	3.26236	0	0	0.270250	0.27025	100
5.	3.26236	—0.9584	0.294	0.296450	0.01489	1991
6.	3.26236	2.0183	0.6184	1.450175	0.20154	720
7.	9.694	98.8650	10.199	1.374014×10^{2}	3.65698×10^{2}	376
8.	12.359	—1257.1520	101.234	8.112571×10^{5}	6.84598×10^{5}	118

TABLE 2—SENSITIVITY OF EFFICIENCY OF $\hat{Y}_{MHT}$ TO DEPARTURE FROM THE OPTIMUM CHOICE OF a

$100 \mid 1-a/a_{opt} \mid$	*Values of efficiency for population*							
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*
0	651	651	116	100	1991	720	726	118
20	310	310	112	100	416	321	242	114
40	203	203	109	100	232	207	179	110
60	151	151	105	100	161	153	142	106
80	120	120	102	100	123	121	117	103
100	100	100	100	100	100	100	100	100

REFERENCES

[1] Brewer, K. R. W. (1963) : A model of systematic sampling with unequal probabilities, *Aust. J. Stat.*, **5** : 5-13.

[2] Cochran, W. G. (1977) : *Sampling Techniques*, Third Edition, Wiley Est. Ltd.

[3] Durbin, J. (1953) : Some results in sampling theory when the units are selected with unequal probabilities, *J. Roy. Stat. Soc.*, Series B, **15** : 252-269.

[4] Durbin, J. (1967) : Design of multi-stage surveys for the estimation of sampling errors. *Applied Statistics*, **16** : 152-164.

[5] Horvitz, D. G. and Thompson, D. J. (1952) : A generalisation of sampling without replacement from a finite universe, *Journal of the American Statistical Association*, **47** : 663-685.

]6] Midzuno, H. (1950) : An outline of the theory of some sampling system, *Annals of the Institute of Statistical Mathematics*, **1** : 49-56.

[7] Prasad, N. G. N. and Srivenkataramana, T. (1980) : A modification of Horvitz-Thompson estimator under Midzuno sampling scheme, *Biometrika*, **67**(3) : 709-11.

[8] Rao, J. N. K. (1965) : On two sample schemes of unequal probability sampling without replacement, *Journal of the Indian Statistical Association*, **3** : 173-180.

[9] Sampford, M. R. (1967) : On sampling without replacement with unequal probabilities of selection, *Biometrika*, **54** : 499-513.

[10] Stuart, A. (1986) : Location shifts in sampling with unequal probabilities, *J. R. Statist. Soc.* series A, 149(4) : 349-365.

[11] Yatas, F. and Grundy, P. M. (1953) : Selection without replacement from within strata with probability proportional to size, *Journal of the Royal Statistical Society*, Series B, **15** : 153-261.