ON THE CHOICE OF SAMPLE-SIZE FOR A HORVITZ-THOMPSON ESTIMATOR

By

ARIJIT CHAUDHURI

Indian Statistical Institute, Calcutta

(Received : May, 1976)

1. INTRODUCTION

The Horvitz-Thompson [4] estimator (HTE, in brief) and its variance [V(HTE), say]-expressions do not involve the sample-size in an explicit manner and undesirably, the latter may not generally decrease monotonically with the sample-size. A sufficient condition for V(HTE) to decrease with increasing sample-size is noted and checked in respect of a few sampling schemes. Further, we specify a class of sampling schemes for which this requirement is fulfilled and suggest a method of choosing the appropriate sample-size for a design on which to base an HTE.

2. NOTATIONS AND THE RESULTS

Suppose we want to estimate a finite population total on the basis of a sample chosen according to some suitable sampling scheme by using the HTE. Denoting the inclusion-probabilities of the first two orders for a chosen design by π_i, π_{ij}'s respectively and by Y_i the value of a real-variate y assumed on the ith unit of a finite population of N units we have for any sample s actually drawn, the expressions for HTE for population total $Y=\Sigma\, Y_i$ and V(HTE) as respectively

$$e=e(s)=\sum_{i\varepsilon s} \frac{Y_i}{\pi_i} \tag{2.1}$$

and

$$V(e)=\sum_i Y_i^2\left(\frac{1}{\pi_i}-1\right)+\sum_{i\neq j}\sum Y_iY_j\left(\frac{\pi_{ij}}{\pi_i\pi_j}-1\right) \tag{2.2}$$

In the special cases when the sampling design is restricted to have a constant effective sample size ν (say), we shall denote a sample by s (ν) and the inclusion-probabilities by $\pi_i(\nu)$, $\pi_{ij}(\nu)$'s and HTE and V(HTE) by

$$e(\nu) = \sum_{i \varepsilon s(\nu)} \frac{Y_i}{\pi_i(\nu)}$$

$$V(e(\nu)) = V(\nu) = \sum_{i<j}\sum \{\pi_i(\nu)\,\pi_j(\nu) - \pi_{ij}(\nu)\} \left(\frac{Y_i}{\pi_i(\nu)} - \frac{Y_j}{\pi_j(\nu)}\right)^2 \quad (2.3)$$

the formulae (2.2) and (2.3) being due to Horvitz-Thompson [4] and Yates and Grundy [10].

Our main concern here is to study the behaviour of V(HTE) with change in sample-size (and/or effective sample-size and/or everage effective sample-size). If we consider Poisson sampling scheme (vide Hajck [3]),

then we have

$$V(\text{HTE}) = \sum Y^2_i \left(\frac{1}{\pi_i} - 1\right).$$

Here average effective sample-size is

$$\sum_{i=1}^{N} \pi_i = E(\nu(s)), \text{ say.}$$

If we are ready to increase $E(\nu(s))$, then we can do so by increasing π_i's for every i and thereby achieve reduction in the value of V(HTE). However, for the method of sampling with probability proportional to size and with replacement (PPSWR), we have, writing p_i for the normed size-measure of the ith unit, $\pi_i = 1-(1-p_i)^n$ and $\pi_{ij} = 1-(1-p_i)^n-(1-p_j)^n+(1-p_i-p_j)^n$, n being the sample-size (number of draws). Here we cannot say if V(HTE) necessarily diminishes with increasing n or increasing average effective sample-size which in this case is

$$E(\nu(s)) = \sum \pi_i = N - \sum_{1}^{N} (1-p_i)^n, \text{ unless}$$

we have knowledge about specific Y_i—values. It is easily verified that a similar situation obtains if we consider the sampling schemes due to Rao [7] or due to Seth [8].

Consider now the class of πps designs (with constant effective sample size) for which

$$\pi_i(\nu) \propto p_i \; (i = 1, 2, \ldots, N)$$

and

$$V(\nu) = \sum_{i<j}\sum \left(p_i p_j - \frac{\pi_{ij}(\nu)}{\nu^2}\right) \left(\frac{Y_i}{p_i} - \frac{Y_j}{p_j}\right)^2.$$

From this one can immediately conclude that if a 'fixed effective-sample size' design πps, then a sufficient condition for $V(\nu)$ to decrease monotonically with increasing ν is that

$$\frac{\pi_{ij}(\nu)}{\nu^2} \text{ increases monotonically with } \nu$$

$$\text{for all } i, j=1, \ldots, N \ (i\neq j)$$

Let us now check the condition (2·4) for a few well-known schemes. First consider the Midzuno [5] sampling scheme suitably modified to have πps property (vide Chaudhary [1]), so that on the first draw the ith unit is selected with probability θ_i ($O<\theta_i<1$ i, $\Sigma\theta=1$) and on subsequent $(n-1)$ draw selections are made with equal probability without replacement from among the units not already chosen. For this scheme we have

$$\pi_i(n)=np_i=\frac{n-1}{N-1}+\frac{N-n}{N-1}\theta_i$$

and
$$\pi_{ij}(n)=\frac{n(n-1)}{N-2}\left\{(p_i+p_j)-\frac{1}{N-1}\right\}$$

it being noted that in this case we require to assume that

$$\frac{n-1}{n(N-1)}<p_i<\frac{1}{n}\ \forall i \tag{2·5}$$

Obviously, the condition (2·4) is satisfied for this scheme and accordingly, $V(\nu)$ decreases with increasing ν. However, because of the assumptions (2·5) this scheme has a very limited applicability. So, let us consider another πps scheme considered by Chaudhuri [2] which is slightly less restricted than the above requiring, however, that

$\frac{n-2}{n(N-2)}<p_i<\frac{1}{n}\ \forall\ i.$ For this scheme, one has (vide Chaudhuri [2])

$$\pi_i(n)=np_i$$

and

$$\pi_{ij}(n)=\frac{(N-n)(N-n-1)}{(N-2)(N-3)}\pi_{ij}(2)+\frac{2(n-2)(N-n)}{(N-2)(N-3)}(q_i+q_j)+\frac{(n-2)(n-3)}{(N-2)(N-3)}$$

where
$$q_i=\frac{np_i-\frac{n-2}{N-2}}{2\left(\frac{N-n}{N-2}\right)}\ \forall\ i$$

Noting that we may write

$$\frac{\pi_{ij}(n)}{n^2}=\left[\left(\frac{N}{n}-1\right)\left(\frac{N-1}{n}-1\right)\frac{\pi_{ij}(2)}{(N-2)(N-3)}\right.$$

$$\left.+\frac{\left(1-\frac{2}{n}\right)}{N-3}(p_i+p_j)-\frac{\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)}{(N-2)(N-3)}\right]$$

it follows that the condition (2·4) may not obtain in this case. In this case the variance of HTE is vide Chaudhuri [2])

$$V(\text{HTE})=\frac{1}{n}\Sigma p_i\left(\frac{Y_i}{p_i}-Y\right)^2\left\{1+\left(\frac{n-1}{N-2}\right)\left(\frac{n-2}{N-3}\right)\frac{1}{np_i}-2\left(\frac{n-2}{N-2}\right)\right\}$$

$$-\frac{1}{n^2}\left(\frac{n-1}{N-2}\right)\left(\frac{n-2}{N-3}\right)\left\{\Sigma\left(\frac{Y_i}{p_i}-Y\right)\right\}^2$$

$$+\frac{1}{n^2}\left(\frac{N-n}{N-2}\right)\left(\frac{N-n-1}{N-3}\right)\sum_{i\neq j}\sum\pi_{ij}(2)\left(\frac{Y_i}{p_i}-Y\right)\left(\frac{Y_i}{p_j}-Y\right)$$

Now, observing that

$$\frac{d}{dn}\frac{1}{n}\left[1+\left(\frac{n-1}{N-2}\right)\left(\frac{n-2}{N-3}\right)\frac{1}{np_i}-2\left(\frac{n-2}{N-2}\right)\right]<0 \text{ for } n\geqslant 4$$

$$\frac{d}{dn}\left\{-\frac{1}{n^2}\left(\frac{n-1}{N-2}\right)\left(\frac{n-2}{N-3}\right)\right\}<0$$

and $\frac{d}{dn}\left\{\frac{1}{n^2}\left(\frac{N-n}{N-2}\right)\left(\frac{N-n-1}{N-3}\right)\right\}<0$

it follows that the behaviour of

$$\sum_{i\neq j}\sum\pi_{ij}(2)\left(\frac{Y_i}{p_j}-Y\right)\left(\frac{Y_j}{p_j}-Y\right) \quad \text{...(2.6)}$$

will determine if V (HTE) may decrease monotonically with n. However, the nature of the quantity (2·6) cannot be known generally for a sampling scheme unless all the variate-values *viz.* Y_i's are known. However, if for the above πps design $\pi_{ij}(2)$ is of the form

$$\pi_{ij}(2)=\frac{1}{N-1}(\beta_i+\beta_j) \qquad i,j=1, \quad , N\ (i\neq j)$$

then the scheme reduces to πps Midzuno scheme for which V (HTE) decreases with increasing n as we have already noted.

Let us next consider the scheme which is derived by modifying Rao's [7] scheme making the latter πps as was studied in Chaudhuri [1]. Here, let us write $\alpha_i(n)$ as the selection-probability of the i-th unit on the first draw, $\frac{\alpha_j(n)}{1-\alpha_i(n)}$ as the selection-probability of the j-th units on the second draw assuming that the i-th unit was chosen on the first draw $\left[0<\alpha_i(n)<1, \forall\ i, \sum_1^N \alpha_i(n)=1\right]$ and $\frac{1}{N-r+1}$ as the selction-probability of a unit on the r-th draw ($r=3, 4, \ldots, n$) provided it was not selected earlier. Here $d_i(n)$'s are so chosen that.

$$np_i=\pi_i(n)=\frac{n-2}{N-2}+\frac{N-n}{N-2}\ \alpha_i(n)\ (1+T(n)-T_i(n))\ i$$

where $\quad T_i(n)=\frac{\alpha_i(n)}{1-\alpha_i(n)},\ T(n)=\Sigma T_i(n).$

From this the value of $V(HTE)$ becomes (vide Chaudhuri [1])

$V(HTE)=\sigma^2(n)$, say,

$$=\frac{1}{n}\ \Sigma p_i\left(\frac{Y_i}{p_i}-Y\right)^2\left\{\left(\frac{n-2}{N-2}\right)\frac{N-n}{N-3}\ \frac{1}{np_i}\right.$$

$$\left.+\frac{N-2n+1}{N-3}\ \frac{1}{np_i}\left(np_i-\frac{n-2}{N-2}\right)\right\}-\frac{2}{n^2}\left(\frac{n-2}{N-2}\right)^2$$

$$\left\{\left(\Sigma\frac{Y_i}{p_i}-Y\right)\right\}^2+\frac{1}{n^2}\left(\frac{n-2}{N-2}\right)\left(\frac{n-3}{N-3}\right)$$

$$\Sigma\alpha_i(n)\left(\frac{Y_i}{p_i}-Y\right)\ \Sigma T_i(n)\ \alpha_i(n)\left(\frac{Y_i}{p_i}-Y\right)$$

$$-\frac{2}{n^2}\left(\frac{N-n}{N-2}\right)\left(\frac{N-n-1}{N-3}\right)\left[\Sigma T_i(n)\ \alpha_i(n)\left(\frac{Y_i}{p_i}-Y\right)^2\right.$$

$$\left.-\sum T_i(n)\left(\frac{Y_i}{p_i}-Y\right)\Sigma\alpha_i(n)\left(\frac{Y_i}{p_i}-Y\right)\right].$$

Considering a particular case where

$$N=4,\ U=(1, 2, 3, 4)$$
$$Y_1=1, Y_2=-1$$
$$Y_3=Y_4=0,$$
$$p_1=p_2=.17,$$
$$p_3+p_4=.66, \alpha_1(2)=\alpha_2(2)=.25,$$
$$2p_i=(1+T(2)-T_i(2))\ \alpha_i(2)$$
$$i=1, \ldots, 4$$
$$3p_i=\tfrac{1}{2}+\tfrac{1}{2}\ \alpha_i(3)\ (1+T(3)-T_i(3)),$$
$$i=1, \ldots, 4$$

we have $T_1(2) = T_2(2) = \frac{1}{3}$

and $$\sigma^2(2) = \tfrac{1}{2}\Sigma p_i \left(\frac{Y_i}{p_i}\right)^2 - \tfrac{1}{2}\Sigma T_i(2)\,\alpha_i(2)\left(\frac{Y_i}{p_i}\right)^2$$

$$\sigma^2(3) = \tfrac{1}{3}\Sigma p_i \left(\frac{Y_i}{p_i}\right)^2 \left(\frac{1}{3p_i} - 1\right)$$

so that

$$\sigma^2(2) - \sigma^2(3) = \Sigma\left(\frac{Y_i}{p_i}\right)^2 \tfrac{5}{6}p_i - \tfrac{1}{9} - \tfrac{1}{2}T_i(2)\,\alpha_i(2)) < 0.$$

Thus, $\sigma^2(n)$ is not necessarily a decreasing function of n.

For this sampling scheme we have

$$\begin{aligned}\frac{\pi_{ij}(n)}{n^2} &= \frac{1}{n^2}\,[\{\alpha_i(n)\,T_j(n) + \alpha_j(n)\,(T_i(n)\} \\ &\quad + \frac{n-2}{N-2}\,(1-\alpha_i(n) - \alpha_j(n))\,(T_i(n) + T_j(n)) \\ &\quad + \frac{n-2}{N-2}\,(\alpha_i(n) + \alpha_j(n))\,\{T(n) - T_i(n) - T_j(n)\} \\ &\quad + \left(\frac{n-2}{N-2}\right)\left(\frac{n-3}{N-3}\right)\{(1-\alpha_i(n) - \alpha_j(n)) - (\alpha_i(n) \\ &\quad + \alpha_j(n))\,(T(n) - T_i(n) - T_j(n))\}]\end{aligned}$$

and it is difficult to conclude generally whether this increases with increasing n.

So, observing that $V(v)$ may not often show a tendency to decrease uniformly (*i.e.* for all variate-values) with increasing v for a proposed sampling scheme for employing Horvitz-Thompson method of estimation we present a particular class of sampling schemes for which, in particular the condition (2.4) holds good.

Let us consider the class C (say) of 'fixed sample-size' sampling designs for which

$$\pi_i(v) = \frac{v}{2}\,\pi_i(2) \;\forall\; i \tag{I}$$

and $$\pi_{ij}(v) = \frac{v(v-1)}{2}\,\pi_{ij}(2) \;\forall\; i, j = 1, \ldots N\ (i \neq j) \tag{II}$$

Clearly, for this scheme (2.4) is satisfied and hence $V(v)$ decreases with increasing v. Such a scheme may be applied as follows.

Adopting the procedure described by Chaudhuri [2] or otherwise one may start with a set of $\pi_i(2)$, $\pi_{ij}(2)$'s subject to the restrictions

$$0 < \pi_i(2) < 1 \;\forall\; i \;\; \Sigma\pi_i(2) = 2,$$

$$0 < \pi_{ij}(2) < min\,\{\pi_i(2), \pi_j(2)\} \;\forall\; i, j\ (i \neq j)$$

and $$\sum_{j \neq i} \pi_{ij}(2) = \pi_i(2) \;\forall\; i$$

Then, the relations (I) and (II) above specify the $\pi_i(v)$ and $\pi_{ij}(v)$'s for a chosen value of v. Finally one may apply the actual sampling procedures described by Mukhopadhyay [6] or Sinha [9] in realizing the values of $\pi_i(v)$, $\pi_{ij}(v)$'s specified earlier, provided they are self-consistent. For such a scheme we have

$$V(v)=A(2)-\frac{2}{v}(v-1)B(2) \tag{2.7}$$

where $$A(2)=\sum_{i<j}\sum \pi_i(2)\,\pi_j(2)\left(\frac{Y_i}{\pi_i(2)}-\frac{Y}{\pi_j(2)}\right)^2$$

$$B(2)=\sum_{i<j}\sum \pi_{ij}(2)\left(\frac{Y_i}{\pi_i(2)}-\frac{Y_j}{\pi_j(2)}\right)^2$$

so that $$V(2)=A(2)-B(2)$$

and $$V(v)=\{A(2)-2B(2)\}+\frac{2}{v}B(2) \tag{2.8}$$

Now one may use the relations (2.7) and (2.8) in choosing a suitable sample-size for the class C of sampling schemes in the following manner.

Supposing that

$$\hat{A}(2)=\sum_{i<j\,\varepsilon\, s(2)}\sum\left(\frac{\pi_i(2)\,\pi_j(2)}{\pi_{ij}(2)}\right)\left(\frac{Y_i}{\pi_i(2)}-\frac{Y}{\pi_j(2)}\right)^2$$

and $$\hat{B}(2)=\sum_{i<j\,\epsilon\, s(2)}\sum\left(\frac{Y}{\pi_i(2)}-\frac{Y}{\pi_j(\ )}\right)^2$$

are unbiased estimates of $A(2)$ and $B(2)$ available from a preliminary sample $s(2)$ selected according to a design with inclusion-probabilities $\pi_i(2)$, $\pi_{ij}(2)$'s one may proceed to decide on the choice of sample-size v if one wants to realize a level of efficiency corresponding to a stipulated value of $V(v)$ as V_0 (say) by using the formula

$$V_0=\hat{A}(2)-2\hat{B}(2)+\frac{2}{v}\hat{B}(2)$$

$$\Rightarrow v=\frac{2\hat{B}(2)}{V_0-\hat{A}(2)+2\hat{B}(2)} \tag{2 9}$$

and taking v as the integer nearest to the right-hand-size expression in (2.9), provided the denominator in (2.9) is non-zero.

ACKNOWLEDGEMENT

The author is grateful to the referee for his valuable suggestions.

Summary

Nothing that the variance of Horvitz-Thompson Estimator may not decrease monotonically with increasing sample-size for some sampling schemes a sufficient condition for this requirement is pointed out and a simple class of sampling schemes satisfying this condition is specified. A method of choosing optimal sample-size for Horvitz-Thompson estimation is also suggested in the light of the above findings.

References

[1] Chaudhri, Arijit (1974) : On some properties of the sampling scheme due to Midzuno, *Cal. Statist. Bull.*, Vol. 23, pp. 1-19.

[2] ——(1975) : A simple method of sampling without replacement with inclusion-probabilities exactly proportional to size, Metrika. Band 22, 147-52.

[3] Hajek, J. (1964) : Asymptotic theory of rejective sampling with varying probabilities from a finite population, *Ann. Math. Statist.*, Vol. 35, pp. 1491-1523.

[4] Horvitz, D.G. and Thompson D.J. (1952) : A generalization of sampling without replacement from a finite universe, *Jour. Amer. Statist. Assoc.*, Vol. 47, pp. 663-685.

[5] Midzuno, H. (1950) : An outline of the theory of sampling systems, *Ann. Inst. Statist. Math*, Vol. i, pp, 149-156.

[6] Mukhopadhyay, Parimal (1972) : A sampling scheme to realize a preassigred set of inclusion-probabilities to first two orders, *Cal. Statist. Assoc. Bull*, Vol. 21, pp. 87-122.

[7] Rao, J.N.K. (1961) : On the estimate of variance in unequal probability sampling, *Ann. Inst. Statist. Math.*, Vol. 13, pp. 57-60.

[8] Seth, G.R. (1966) : On estimates of variance of estimate of population total in varying probabilities, *Jour. Ind. Soc. Agr. Statist.* Vol. 18, pp. 52-56.

[9] Sinha, B.K. (1973) : On sampling schemes to realize pre-assigned sets of inclusion-probabilities of first two orders, *Cal. Statist. Assoc. Bull*, Vol. 22, pp. 89-100.

[10] Yates, F. and Grundy, P.M. (1953) : Selection without replacement from within strata with probability proportional to size. *Jour. Roy. Statist. Soc. Ser. B.* Vol. 15, pp. 253-261.