On the Feasibility of Basing Horvitz and Thompson's Estimator on a Sample by Rao, Hartley, and Cochran's Scheme

ARIJIT CHAUDHURI, KAJAL DIHIDAR, AND MAUSUMI BOSE

Applied Statistics Unit, Indian Statistical Institute, Kolkata, India

Formulae for the first and second order inclusion probabilities for the Rao et al. (1962) (RHC) scheme of sampling are derived. They enable one to evaluate, for a sample drawn according to the RHC scheme, the Horvitz and Thompson's (1952) estimator (HTE) along with its unbiased variance estimator given by Yates and Grundy (1953). So, for a sample at hand thus drawn one may choose between the RHCE and the HTE for use on finding which one has the smaller coefficient of variation.

Keywords First and second order inclusion probabilities; Variance estimation.

Mathematics Subject Classification 62D05.

1. Preliminaries

Rao et al. (1962) have given a practically useful (RHC) scheme of selecting a sample with a pre-assigned number of distinct units from a survey population when certain positive-valued size measures for the units of the population are available. They have also given an unbiased estimator, called the Rao, Hartley, and Cochran's estimator (RHCE), for the population total along with a uniformly nonnegative unbiased variance estimator for it. They have specified the situations when the variance of the RHCE attains its minimum possible value. We derive, in such situations, the formulae for the inclusion probabilities of units and distinct pairs of units in a sample drawn according to this scheme. Consequently, for an RHC sample at hand, we may evaluate both the RHCE and the Horvitz and Thompson's estimator (HTE) along with their variance estimates. For the variance of the HTE, the Yates and Grundy's estimator (YGE), though unbiased, is not yet known to have its value to be uniformly nonnegative. However, we tried several numerical

examples and each time the YGE turned out to be positive. So, the estimated coefficients of variation (CV) may be computed and one may choose the estimate with the smaller CV for actual use in practice.

Let $U = (1, \ldots, i, \ldots, N)$ denote a survey population with known size measures $x_i (>0)$ and unknown variate values of interest y_i with respective totals X and Y. The main objective is to unbiasedly estimate Y on taking an RHC sample of size n, $(2 \le n < N)$ using the normed size measures $p_i = x_i/X$, $i \in U$. In the RHC scheme, U is first randomly divided into n non overlapping groups taking N_i units into the ith group $(i = 1, \ldots, n)$, N_i chosen as positive integers subject to $\sum_n N_i = N$, $\sum_n$ denoting sum over the n groups. Labeling the units in the ith group as ij $(j = 1, \ldots, N_i)$, one unit is chosen with probability p_{ij}/Q_i, where $Q_i = p_{i1} + \cdots + p_{iN_i}$. This is independently repeated over all the n groups. Let us write the ith sample value of y along with the corresponding normed size measure as (y_i, p_i) instead of (y_{ij}, p_{ij}), by suppressing the subscript j for simplicity. Then, the RHC unbiased estimator for Y, say t_{RHC}, its variance, say $V(t_{RHC})$, and an unbiased estimator for it, say $v(t_{RHC})$, are respectively, given by

$$t_{RHC} = \sum_{i \in s} y_i \frac{Q_i}{p_i}, \quad V(t_{RHC}) = A \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} p_i p_j \left(\frac{y_i}{p_i} - \frac{y_j}{p_j} \right)^2,$$

$$v(t_{RHC}) = B \sum_{i \in s} \sum_{j>i} Q_i Q_j \left(\frac{y_i}{p_i} - \frac{y_j}{p_j} \right)^2, \tag{1.1}$$

where

$$A = \frac{\sum_{i=1}^{n} N_i^2 - N}{N(N-1)} \quad \text{and} \quad B = \frac{\sum_{i=1}^{n} N_i^2 - N}{N^2 - \sum_{i=1}^{n} N_i^2}.$$

The HT unbiased estimator for Y, say t_{HT}, Yates and Grundy's form of its variance, say $V(t_{HT})$, and an unbiased estimator for it, say $v(t_{HT})$, are, respectively, given by

$$t_{HT} = \sum_{i \in s} \frac{y_i}{\pi_i}, \quad V(t_{HT}) = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} (\pi_i \pi_j - \pi_{ij}) \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2,$$

$$v(t_{HT}) = \sum_{i \in s} \sum_{j>i} \frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2, \tag{1.2}$$

writing s for a sample chosen by RHC method. In the next section, we shall derive formulae for π_i and π_{ij} and note that $\pi_i > 0$ and $\pi_{ij} > 0 \; \forall i, j \; (i \ne j)$ so that formulae (1.2) can be applied.

Rao et al. (1962) have shown that the choice of N_i's which minimizes $V(t_{RHC})$ is given by

$$N_i = [N/n] = m \quad \text{for } i = 1, \ldots, k; \quad N_i = [N/n] + 1 = m + 1 \quad \text{for } i = k+1, \ldots, n$$

such that $N = mk + (m+1)(n-k)$, with $1 < k \le n$.

2. Derivation of Formulae

Case 1. N/n = an integer. $N/n = m$, (say).

Here $k = n$. For integers T and M, $T \geq M$, we use the following notation:

$$\alpha(T, M) = \binom{T}{M}.$$

Let G denote the total number of possible random n groups by RHC scheme. A group of m units containing unit i can be formed from the N units of U in $\alpha(N-1, m-1)$ ways. Let S_{m-1}^{N-1} denote summation over these $\alpha(N-1, m-1)$ m-tuples. Once a group containing unit i is formed, let G_1 denote the total number of possible random $(n-1)$ groups by RHC scheme. Then, clearly,

$$G = \frac{N!}{(m!)^n n!} \quad \text{and} \quad G_1 = \frac{(N-m)!}{(m!)^{n-1}(n-1)!}. \tag{2.1}$$

Let p_{r_l} be the normed size measure of the lth unit r_l in the ith group consisting of m distinct units, $(r_l \neq i)$. Then, π_i can be written as

$$\pi_i = \frac{G_1}{G}\left[S_{m-1}^{N-1}\left(\frac{p_i}{p_i + \sum_{l=1}^{m-1} p_{r_l}}\right)\right]. \tag{2.2}$$

To derive formulae for π_{ij} we first note that the number of m-tuples of distinct units of U with unit i included but unit j not included is equal to $\alpha(N-2, m-1)$. Again, the number of m-tuples $(j, s_1, \ldots, s_{m-1})$ of distinct units of U such that $j \neq i$, $s_t \neq i \neq j \neq r_l$, $l = 1, \ldots, m-1$, $t = 1, \ldots, m-1$ is $\alpha(N-m-1, m-1)$. Let S_{m-1}^{N-2} and S_{m-1}^{N-m-1}, respectively, denote summations over these $\alpha(N-2, m-1)$ and $\alpha(N-m-1, m-1)$ m-tuples.

Let G_2 denote the number of possible ways of forming $(n-2)$ random groups of m units other than (i) any random group containing i but excluding j and (ii) any random group containing j but not any unit in a random group formed as in (i). Then,

$$G_2 = \frac{(N-2m)!}{(m!)^{n-2}(n-2)!}. \tag{2.3}$$

Hence,

$$\pi_{ij} = \frac{G_2}{G}\left[S_{m-1}^{N-2}\frac{p_i}{p_i + \sum_{l=1}^{m-1} p_{r_l}}\left(S_{m-1}^{N-m-1}\frac{p_j}{p_j + \sum_{t=1}^{m-1} p_{s_t}}\right)\right]. \tag{2.4}$$

Note that

$$\frac{G_1}{G} = \frac{nm!(N-m)!}{N!}, \quad \frac{G_2}{G} = \frac{(N-2m)!(m!)^2 n(n-1)}{N!}.$$

One possible check for the correctness of the above rather complicated formulae is to verify if the well-known consistency conditions, namely, (I) $\sum_{i=1}^{N} \pi_i = n$ and (II) $\sum_{j=1, j\neq i}^{N} \pi_{ij} = (n-1)\pi_i$ implying that $\sum\sum_{j\neq i} \pi_{ij} = n(n-1)$ are satisfied by (2.2) and (2.4).

From (2.1) and (2.2), it follows that

$$\sum_{i=1}^{N}\pi_i = \frac{(N-m)!}{(m!)^{n-1}(n-1)!}\frac{(m!)^n n!}{N!}\binom{N-1}{m-1}\frac{N}{m}1$$

$$= \frac{(N-m)!m!n}{N!}\frac{(N-1)!}{(m-1)!(N-m)!}\frac{N}{m} = n. \tag{2.5}$$

Thus condition (I) is satisfied.

Again, from (2.1), (2.3), and (2.4),

$$\sum_{j=1,j\neq i}^{N}\pi_{ij} = \frac{(N-2m)!(m!)^n n!}{(m!)^{n-2}(n-2)!N!}\left[\sum_{j=1j\neq i}^{N}\left(S_{m-1}^{N-2}\frac{p_i}{p_i+\sum_{l=1}^{m-1}p_{r_l}}\left(S_{m-1}^{N-m-1}\frac{p_j}{p_j+\sum_{t=1}^{m-1}p_{s_t}}\right)\right)\right].$$

On simplification, we get

$$\sum_{j=1,j\neq i}^{N}\pi_{ij} = \frac{(N-2m)!(m!)^2 n(n-1)}{N!}$$

$$\times\left[\binom{N-m-1}{m-1}\left(\frac{N-1}{\binom{N-1}{m-1}/\binom{N-2}{m-1}}\right)\frac{1}{m}\left(S_{m-1}^{N-1}\frac{p_i}{p_i+\sum_{l=1}^{m-1}p_{\eta}}\right)\right],$$

and finally, it follows that

$$\sum_{j=1,j\neq i}^{N}\pi_{ij} = \frac{(n-1)n(m!)(N-m)!}{N!}\left(S_{m-1}^{N-1}\frac{p_i}{p_i+\sum_{l=1}^{m-1}p_{r_l}}\right) = (n-1)\pi_i. \tag{2.6}$$

(2.5) and (2.6) imply that $\sum\sum_{j\neq i}\pi_{ij} = n(n-1)$ and so condition (II) is satisfied.

Case 2. N/n is not an integer.

Here we recall that in each of the k groups, the number of units is m and the remaining $(n-k)$ groups have $(m+1)$ units in each. These two types of sizes of groups in turn lead us to re-form here the values of G, G_1, and G_2 as defined in Case 1.

In this case, let G reduce to G'; G_1 correspond to two terms namely G_1' and G_1'', and G_2 correspond to three terms namely $G_2^{(1)}$, $G_2^{(2)}$, and $G_2^{(3)}$. Their values may be written as follows:

$$G' = \frac{N!}{(m!)^k k!(m+1)!^{n-k}(n-k)!} \tag{2.7}$$

$$G_1' = \frac{(N-m)!}{(m!)^{k-1}(k-1)!(m+1)!^{n-k}(n-k)!}, \quad G_1'' = \frac{(N-m-1)!}{(m!)^k k!(m+1)!^{n-k-1}(n-k-1)!} \tag{2.8}$$

$$G_2^{(1)} = \frac{(N-2m)!}{(m!)^{(k-2)}(k-2)!(m+1)!^{n-k}(n-k)!}$$

$$G_2^{(2)} = \frac{(N-2m-1)!}{(m!)^{(k-1)}(k-1)!(m+1)!^{n-k-1}(n-k-1)!}$$

$$G_2^{(3)} = \frac{(N-2m-2)!}{(m!)^k k!(m+1)!^{n-k-2}(n-k-2)!}. \tag{2.9}$$

Then, applying similar notations and arguments as in Case 1, it follows on simplification that

$$\pi_i = \frac{G_1'}{G'}\left[S_{m-1}^{N-1}\left(\frac{p_i}{p_i + \sum_{l=1}^{m-1} p_{r_l}}\right)\right] + \frac{G_1''}{G'}\left[S_m^{N-1}\left(\frac{p_i}{p_i + \sum_{l=1}^{m} p_{r_l}}\right)\right] \tag{2.10}$$

$$\pi_{ij} = A_1 + A_2 + A_3 + A_4 \tag{2.11}$$

where

$$A_1 = \frac{G_2^{(1)}}{G'}\left[S_{m-1}^{N-2}\frac{p_i}{p_i + \sum_{l=1}^{m-1} p_{r_l}}\left(S_{m-1}^{N-m-1}\frac{p_j}{p_j + \sum_{t=1}^{m-1} p_{s_t}}\right)\right],$$

$$A_2 = \frac{G_2^{(2)}}{G'}\left[S_{m-1}^{N-2}\frac{p_i}{p_i + \sum_{l=1}^{m-1} p_{r_l}}\left(S_{m}^{N-m-1}\frac{p_j}{p_j + \sum_{t=1}^{m} p_{s_t}}\right)\right],$$

$$A_3 = \frac{G_2^{(2)}}{G'}\left[S_{m}^{N-2}\frac{p_i}{p_i + \sum_{l=1}^{m} p_{r_l}}\left(S_{m-1}^{N-m}\frac{p_j}{p_j + \sum_{t=1}^{m-1} p_{s_t}}\right)\right],$$

$$A_4 = \frac{G_2^{(3)}}{G'}\left[S_{m}^{N-2}\frac{p_i}{p_i + \sum_{l=1}^{m} p_{r_l}}\left(S_{m}^{N-m-2}\frac{p_j}{p_j + \sum_{t=1}^{m} p_{s_t}}\right)\right].$$

As in Case 1, we verify the correctness of the above expressions by checking conditions (I) and (II). From (2.7) and (2.8) it follows that

$$\sum_{i=1}^{N} \pi_i = \frac{m!kN!}{N!m!} + \frac{(m+1)!(n-k)N!}{N!(m+1)!} = k + (n-k) = n$$

and so condition (I) is satisfied. Again, using arguments as in Case 1 to sum the first two terms of π_{ij} over $j = 1, \ldots, N$ $(j \neq i)$ and then doing the same for next two terms, we get the following results:

$$\sum_{j=1,\, j\neq i}^{N} (A_1 + A_2) = (n-1)\frac{G_1'}{G'}\left[S_{m-1}^{N-1}\left(\frac{p_i}{p_i + \sum_{l=1}^{m-1} p_{r_l}}\right)\right].$$

$$\sum_{j=1,\, j\neq i}^{N} (A_3 + A_4) = (n-1)\frac{G_1''}{G'}\left[S_{m}^{N-1}\left(\frac{p_i}{p_i + \sum_{l=1}^{m} p_{r_l}}\right)\right].$$

Hence, $\sum_{j=1, j\neq i}^{N} \pi_{ij} = (n-1)\pi_i$ and so $\sum\sum_{j\neq i} \pi_{ij} = n(n-1)$, showing that condition (II) is satisfied.

Remark. Since π_i, π_{ij}'s as worked out are all positive we may straightaway employ the HTE based on the RHC scheme to unbiasedly estimate Y and the YGE to

unbiasedly estimate the variance of the HTE. The estimated coefficients of variation (CV) are $\frac{+\sqrt{v(t_{RHC})}}{|t_{RHC}|}$ and $\frac{+\sqrt{v(t_{HT})}}{|t_{HT}|}$.

3. Concluding Remarks

When a sample is drawn by the RHC scheme, the only estimator which could be employed so far in practice was RHCE. But on deriving the requisite formulae, we find the HTE as a feasible alternative. Our recommendation is to employ, for a realized sample, the RHC estimate or the HT estimate preferring the one with the smaller CV, provided the YG estimate turns out nonnegative. With several numerical examples varying N, n, p_i's we observed '$\pi_i\pi_j > \pi_{ij}$' in all our cases ensuring positivity of the YG estimate of the variance of the HTE.

Acknowledgments

The authors would like to thank Prof. J. N. K. Rao for some helpful discussions over email. The research of the first author was supported by the Grant No. 21(0539)/02/EMR-II of CSIR, India.

References

Horvitz, D. G., Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *J. Amer. Statist. Assoc.* 47:663–685.

Rao, J. N. K., Hartley, H. O., Cochran, W. G. (1962). On a simple procedure of unequal probability sampling without replacement. *J. Roy. Statist. Soc. B* 24:482–491.

Yates, F., Grundy, P. M. (1953). Selection without replacement from within strata with probability proportional to size. *J. Amer. Statist. Assoc.* 75:206–211.