J. Ind. Soc. Agril. Statist.
59(3), 2005 : 199–208

Horvitz-Thompson Variance Estimation when Auxiliary Information is Available

Praful A. Patel and Raju D. Chaudhari[1]
Sardar Patel University, Vallabh Vidyanager-388120, Gujarat
(Received : May, 2002)

SUMMARY

This article considers variance estimation of the Horvitz-Thompson estimator of the finite population total when information on some relevant auxiliary variable is available. A model-based variance estimator is suggested and it is shown that this estimator has smaller expected mean square error among the class of all model-unbiased estimators. A small simulation study is presented to compare the performance of the suggested estimator with the Yates-Grundy variance estimator for a fixed size sampling and with the Horvitz-Thompson variance estimator for a random size sampling. In simulation study, it was observed that in most of the cases the proposed estimator is more efficient than one given by Horvitz and Thompson under the specified assumptions.

Key words : Variance estimation, Model-based estimation, Auxiliary information, Super population model.

1. INTRODUCTION

Accurate estimation of forest resources over large geographical area is of significant interest to forest managers and forestry scientists. In forest surveys, design-based estimates of the parameters like total tree volume, growth and mortality, or area by forest type are required. Design-based estimation of such parameters, based on information gathered during ground visits of sample plots, can be made more precise by incorporating auxiliary information available from remote sensing. The ratio of means estimator, mean of ratios estimator and Horvitz-Thompson estimator (mean of ratios under πps – sampling scheme) are often used in forest inventory to estimate population totals with their standard errors (Zarnoch and Behtold (2000)).

Here we consider variance estimation of the Horvitz-Thompson (HT) estimator of the population total.

Based on empirical and limited theoretic evidence, the Yates-Grundy variance estimator (v_{YG}) (1953) of the HT estimator of a finite population total is generally considered superior to the Horvitz-Thompson variance estimator (v_{HT}) (1952) because of fewer negative estimates and smaller sampling variance (Cumberland and Royall (1981), Rao and Singh (1973)). However, v_{YG} requires fixed sample size, whereas v_{HT} does not. This restriction of v_{YG} to fixed sample size design eliminated this variance estimator from consideration of many applications in surveys (Stehman and Overton (1994)). Several design-based variance estimators of HT estimator which incorporate knowledge of an auxiliary variable known for every unit in the population have been proposed and their performances examined and compared. See, for example, Isaki (1983), Singh *et al.* (1999).

In this article a model-based estimator of the Horvitz-Thompson variance is suggested that incorporates the auxiliary information and that does not require fixed sample size. A fully developed approach to model-based variance estimation did originate with Royall and Eberhardt (1975), but an earlier version may be found in Royall (1971).

Consider the finite population of units $U = \{1, ..., i, ..., N\}$. Let Y_i and x_i be the values of the

[1] *V.P. & R.P.T.P. Science College, Vallabh Vidyanagar-388120, Gujarat*

main variable y and the auxiliary variable x, respectively, for the i^{th} unit. Let s be a sample of size n drawn from U with the probability p(s) having positive inclusion probabilities $\pi_i = \sum_{s \ni i} p(s)$ and $\pi_{ij} = \sum_{s \ni i,j} p(s)$. For short, $\sum_A$ and $\sum\sum_A$ will be used for $\sum_{i \in A}$ and $\sum\sum_{i \neq j \in A}$, where A is an arbitrary set.

The Horvitz-Thompson estimator $\hat{Y}_{HT} = \sum_S \frac{Y_i}{\pi_i}$ is unbiased for the population total, $T_y = \sum_U Y_i$, and has the Horvitz-Thompson and Yates-Grundy variance expressions

$$V_{HT}(\hat{Y}_{HT}) = \sum_U \left(\frac{1}{\pi_i} - 1\right) Y_i^2 + \sum\sum_U \left(\frac{\pi_{ij}}{\pi_i \pi_j} - 1\right) Y_i Y_j \qquad (1.1)$$

and

$$v_{YG}(\hat{Y}_{HT}) = -\frac{1}{2}\sum\sum_U (\pi_{ij} - \pi_i \pi_j)\left(\frac{Y_i}{\pi_i} - \frac{Y_j}{\pi_j}\right)^2 \qquad (1.2)$$

Equation (1.2) holds only for fixed sample size designs. Their unbiased estimators are

$$v_{HT}\left(\hat{Y}_{HT}\right) = \sum_S \left(\frac{1}{\pi_i} - 1\right)\frac{Y_i^2}{\pi_i} + \sum\sum_S \left(\frac{\pi_{ij}}{\pi_i \pi_j} - 1\right)\frac{Y_i Y_j}{\pi_{ij}}$$

and

$$v_{YG}\left(\hat{Y}_{HT}\right) = -\frac{1}{2}\sum\sum_S \left(\frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}}\right)\left(\frac{Y_i}{\pi_i} - \frac{Y_j}{\pi_j}\right)^2$$

It is well known that (Cassel *et al.* (1977)) under the super population model ξ with

$$\mathcal{E}(Y_i) = \mu x_i,\ \mathcal{V}(Y_i) = \sigma^2 x_i^2,\ Cov(Y_i, Y_j) = \rho\sigma^2 x_i x_j \quad (i \neq j) \qquad (1.3)$$

where μ, $\sigma > 0$ and $\rho \in (-1/(N-1), 1)$ are the parameters and $\mathcal{E}(.), \mathcal{V}(.)$ and $Cov(.,.)$ denote respectively, ξ-expectation, ξ-variance and ξ-covariance, the best strategy for estimating the population total is the HT estimator with first-order inclusion probabilities π_i proportional to $x_i (x_i \leq \frac{T_x}{n}, \forall i \in U)$.

Here, our objective is to obtain a model-based estimator of V_{HT} given in (1.1), that will work better than the usual estimator. The basic methodology for constructing such an estimator will be to:

1. obtain the conditional ξ-expectation of V_{HT} given the data $d = \{(y_i, x_i) : i \in s\}$ from the observed sample, i.e. $\mathcal{E}\left\{V_{HT}(\hat{Y}_{HT}) \middle| d\right\}$ and
2. substitute the sample estimates of functions of unknown parameters in the expression $\mathcal{E}\left\{V_{HT}(\hat{Y}_{HT}) \middle| d\right\}$.

The efficiency of this estimator, as compared to the standard estimator, will depend on the goodness of fit. More details and illustrative example concerning the above idea was given in Wolter (1985). The design-based estimators v_{HT} and v_{YG} use the x variable only at the sampling stage, whereas the model-based estimator also uses the x variable at the estimation stage.

In the next section the theoretic development is done using the Royall's (1970) prediction approach for finite population sampling under the model (1.3). Motivated by this, a model-based estimator of v_{HT} is suggested in Section 3. Section 4 presents a limited simulation study.

2. OPTIMAL ξ-UNBIASED PREDICTION

A general theory of prediction that includes quadratic form, the population variance in particular, has been formulated, under a more general model, by Rodrigues *et al.* (1985). The population non-negative definite (n.n.d.) quadratic form is defined as $V_y = \underline{Y}'\Delta\underline{Y}$ where $\underline{Y} = (Y_1, ..., Y_N)'$ and $\Delta = (\Delta_{ij})$ is an $N \times N$ n.n.d. and symmetric matrix of known constants.

A predictor Q is p-unbiased for V_y if, for a given design p(s)

$$E(Q) = \sum_{s \in S} p(s) Q = V_y \qquad \forall \underline{Y} \in R_N$$

and is ξ-unbiased if, for a given ξ

$$\mathcal{E}(Q - V_y) = 0 \quad \forall s \in \mathcal{S}$$

where $\mathcal{S} = \{s : s \subseteq U\}$

The ξ-expected p-mean square error, denoted by $\mathcal{E}MSE$, of an arbitrary strategy (a pair of sampling design p and a predictor Q) (p, Q) is given by

$$\mathcal{E}MSE(p,Q) = \mathcal{E}E(Q - V_y)^2 \qquad (2.1)$$

The goal of this section is to obtain an optimal predictor of V_y. Here, the optimality is interpreted in the sense of minimizing (2.1) subject to the ξ-unbiasedness. The main result is contained in Theorem 1 below.

The parameter ρ in model (1.3) was shown to be a quite generally redundant by Brewer and Tam (1990). Therefore in the remaining article, we assume that $\rho = 0$.

Theorem 1. Let p be any given design, and let $Y_1, \ldots, Y_N$ be normally distributed. Then, under model (1.3), among all predictors Q of $V_y = \underline{Y}'\Delta\underline{Y}$ satisfying $\mathcal{E}(q - V_y) = 0$, the $\mathcal{E}$MSE is minimized by

$$Q_{PR} = \Sigma_S \Delta_{ii} Y_i^2 + \Sigma\Sigma_S \Delta_{ij} Y_i Y_j + \{\Sigma_U \Delta_{kk} x_k^2 - \Sigma_S \Delta_{kk} x_k^2\} \frac{1}{n} \Sigma_S \frac{Y_i^2}{x_i^2} + \{\Sigma\Sigma_U \Delta_{kl} x_k x_l - \Sigma\Sigma_S \Delta_{kl} x_k x_l\} \frac{1}{n(n-1)} \Sigma\Sigma_S \frac{Y_i Y_j}{x_i x_j} \quad (2.2)$$

Proof. See Appendix A.

Remark 1. The above theorem is valid even if the sampling design is not fixed size. In this case, replace n by n_s, the random sample size.

There are two special cases of V_y that are of interest.

1. The population variance itself given by
$$S_y^2 = \frac{1}{N}\Sigma_U Y_i^2 - \frac{1}{N(N-1)}\Sigma\Sigma'_U Y_i Y_j$$
2. The Horvitz-Thompson variance, $V_{HT}(\hat{Y}_{HT})$ that was given in (1.1).

Corollary 1. Let p be any given design, and let $Y_1, \ldots, Y_N$ be normally distributed. Then under model (1.3), the optimal ξ-unbiased predictor of the population variance S_y^2 is given by

$$Q_{PR} = \frac{1}{N}\Sigma_S \left(\frac{1 + \Sigma_U x_k^2 - \Sigma_S x_k^2}{n x_i^2} \right) Y_i^2 - \frac{1}{N(N-1)} \Sigma\Sigma_S \left(\frac{1 + \Sigma\Sigma_U x_k x_l - \Sigma\Sigma_S x_k x_l}{n(n-1) x_i x_j} \right) Y_i Y_j$$

Corollary 2. For any given design p, the predictor $v_{PR}(\hat{Y}_{HT})$ given by

$$v_{PR}(\hat{Y}_{HT}) = \Sigma_S \left(\frac{1}{\pi_i} - 1 \right) Y_i^2 + \Sigma\Sigma_S \left(\frac{\pi_{ij}}{\pi_i \pi_j} - 1 \right) Y_i Y_j + \left\{ \Sigma_U \left(\frac{1}{\pi_k} - 1 \right) x_k^2 - \Sigma_S \left(\frac{1}{\pi_k} - 1 \right) x_k^2 \right\} \Sigma_S \frac{Y_i^2}{n x_i^2} + \left\{ \Sigma\Sigma_U \left(\frac{\pi_{kl}}{\pi_k \pi_l} - 1 \right) x_k x_l - \Sigma\Sigma_S \left(\frac{\pi_{kl}}{\pi_k \pi_l} - 1 \right) x_k x_l \right\} \Sigma\Sigma_S \frac{Y_i Y_j}{n(n-1) x_i x_j}$$

is optimal, under model (1.3), in the class of all ξ-unbiased predictors of $V_{HT}(\hat{Y}_{HT})$ when $Y_i (i = 1,2,\ldots,N)$ are normally distributed.

3. VARIANCE ESTIMATION

Motivated by the prediction theory we suggest the following estimator for estimating the variance of Horvitz-Thompson estimator, $V_{HT}(\hat{Y}_{HT}) = \underline{Y}' \Delta \underline{Y}$, where $\Delta = (\Delta_{ij})$ is an N×N n.n.d. and symmetric matrix with $\Delta_{ii} = \pi_i^{-1} - 1$ and $\Delta_{ij} = \pi_{ij}\pi_i^{-1}\pi_j^{-1} - 1$, that was given in (1.1). This estimator is general in that it applies for both fixed size and non-fixed size sampling design.

$$v_{PR}(\hat{Y}_{HT}) = A_{HT}(s, \underline{Y}) + U_{PR}(s, \underline{Y}) \quad (3.1)$$

where

$$A_{HT}(s, \underline{Y}) = \Sigma_S \left(\frac{1}{\pi_i} - 1 \right) Y_i^2 + \Sigma\Sigma_S \left(\frac{\pi_{ij}}{\pi_i \pi_j} - 1 \right) Y_i Y_j$$

and

$$U_{PR}(s, \underline{Y}) = \left\{ \Sigma_U \left(\frac{1}{\pi_k} - 1 \right) x_k^2 - \Sigma_S \left(\frac{1}{\pi_k} - 1 \right) x_k^2 \right\} \Sigma_S \frac{Y_i^2}{n x_i^2} + \left\{ \Sigma\Sigma_U \left(\frac{\pi_{kl}}{\pi_k \pi_l} - 1 \right) x_k x_l - \Sigma\Sigma_S \left(\frac{\pi_{kl}}{\pi_k \pi_l} - 1 \right) x_k x_l \right\} \times \Sigma\Sigma_S \frac{Y_i Y_j}{n(n-1) x_i x_j}$$

This estimator is so intuitively plausible, following directly from Royall's approach of replacing a population sum by the sample sum plus a model-based estimator of the non-sample sum.

The necessary and sufficient conditions for v_{PR} to be non-negative are derived in the following theorem.

Theorem 2. The PR –estimator v_{PR} is n.n.d. if and only if $\pi_{ij} \leq \pi_i \pi_j$ for all $i \neq j \in U$.

Proof. See Appendix B.

4. SIMULATION STUDY

In this section, we present the results of limited simulation studies on the small sample performance and the large sample performance of the variance estimator given in (3.1) of V_{HT}. The populations used in this simulation were listed in Table 1 and Table 4.

4.1 Small Sample Performance

Two kinds of comparison will be considered. They are

1. Comparison of v_{PR} and v_{YG} for a fixed size sampling
2. Comparison of v_{PR} and v_{HT} for a random size sampling

Table 1. Study population

No.	Source	X	Y	N	CV(x)	CV (y)	ρ_{YX}
1.	Yates (1960), p. 163	Eye estimates	Measured volume	25	0.443	0.470	0.535
2.	Murthy (1967), p. 126	Area in 1951	Workers at household industry in 1961	40	0.632	0.951	0.578
3.	Sukhatme *et al.* (1970) p. 256 (51-89)	Number of villages	Area under wheat	40	0.675	0.631	0.599
4.	Murthy (1967), p. 178	Geographical area (in acres)	Area under winter paddy	58	0.347	0.584	0.631
5.	Murthy (1967), p. 126	Number of persons in 1961	Workers at household industry in 1961	40	0.595	0.952	0.635
6.	Murthy (1967), p. 398	Number of workers	Number of absentees	43	0.459	0.681	0.661
7.	Sukhatme (1970) p. 256 (1-40)	Number of villages	Area under wheat	40	0.482	0.613	0.662
8.	Yates (1960), p. 159	Total number of persons	Number of absentees	43	0.455	0.706	0.666
9.	Murthy (1967), p. 126	Number of persons in 1961	Number of cultivators in 1961	40	0.595	0.645	0.871
10.	Murthy (1967), p. 126	Area in 1951	Number of cultivators in 1961	40	0.632	0.645	0.882
11.	Sukhatme (1970), p. 185	Area under wheat in 1936	Area under wheat in 1937	34	0.768	0.756	0.930
12.	Murthy (1967), p. 228	Fixed capital for factories	Output for factories	40	0.353	0.252	0.951
13.	Murthy (1967), p. 131	Length (in chain units)	Timber volume	48	0.338	0.547	0.954
14.	Kish (1965), p. 625	Number of dwellings	Dwellings occupied by renters	40	0.612	0.746	0.975

Comparison under Sunter's method

From each population a sample of size n = 10 was drawn using Sunter's (1986) method (see also, Särndal *et al.* (1992)). The variance estimators v_{YG} and v_{PR} were computed from each sample. This process was repeated M = 10,000 times. We then reconducted the above simulation with n = 15.

The performance of different variance estimators was measured and compared in terms of relative percentage bias (RB%) and relative percentage standard error (RSE%). The simulated values of RB % and RSE % for a particular variance estimator v were compared as

$$RB\%(v) = 100 \times \frac{\bar{v} - V_{HT}}{V_{HT}}$$

and

$$RSE\% = 100 \times \sqrt{V_{Sim}} / V_{HT}$$

where

$$\bar{v} = \frac{1}{M}\sum_{j=1}^{M} v_{(j)} \text{ and } V_{Sim}(v) = \frac{1}{M-1}\sum_{j=1}^{M}(v_{(j)} - \bar{v})^2$$

Table 2 presents the values of RB% and RSE% of variance estimators v_{YG} and v_{PR}.

A scatter plot of each of the populations 1-4 exhibits lack of fit of straight line due to some outliers whereas populations 5-8 (with slight intercept), 9-12 and 14 reveals that a linear model $y_i = \beta x_i + \varepsilon_i$ with $v(y_i) \propto x_i^{\gamma} (1 \le \gamma \le 2)$ might be appropriate and the relationship between y and x is strong, and population 13 seems to obey the intercept model $y_i = \alpha + \beta x_i + \varepsilon_i$ with $v(y_i) \propto x_i^2$.

The following results were obtained through simulation:

1. The PR-estimator has smaller RSE% than the YG-estimator for the populations 5-14 (except 13) as the assumptions of the underlying model viz. y_i and x_i is straight line passing through origin and the variance of y_i is proportional to x_i^2 along the straight line are satisfied.

Table 2. Relative percentage bias and RSE (under Sunter's scheme)

Popl. No.	n	RB %		RSE %	
		v_{YG}	v_{PR}	v_{YG}	v_{PR}
1.	10	– 7.0	48.4	28.68	44.44
	15	–	–	–	–
2.	10	13.3	35.5	56.35	366.74
	15	–	–	–	–
3.	10	13.3	35.5	56.35	366.74
	15	–	–	–	–
4.	10	1.9	0.6	35.03	37.46
	15	0.7	11.9	26.35	29.59
5.	10	31.3	– 13.0	115.72	31.99
	15	168.5	9.4	57.35	20.29
6.	10	6.7	– 6.3	53.96	38.28
	15	6.0	– 10.8	43.10	25.17
7.	10	18.7	– 7.4	76.60	34.86
	15	8.9	1.4	36.60	23.31
8.	10	6.1	– 6.8	50.48	37.43
	15	4.9	– 11.9	41.23	24.40
9.	10	2.5	21.9	54.60	37.50
	15	200.5	– 5.4	70.82	14.02
10.	10	19.4	6.5	107.29	59.85
	15	14.5	0.5	45.01	21.88
11.	10	– 1.1	32.8	79.80	12.62
	15	–	–	–	–
12.	10	16.0	-15.0	57.14	20.23
	15	– 11.3	17.2	31.71	9.16
13.	10	9.6	30.7	44.94	46.77
	15	10.0	-22.2	53.59	20.24
14.	10	22.1	6.6	96.75	29.98
	15	– 13.3	– 7.6	37.00	12.30

' – ' indicates that Sunter method fails to provide inclusion probabilities.

2. v_{PR} may have smaller RB% than v_{YG} (Populations: 3, 5, 7, 10, 14), but the pattern is not very clear or general because of small sample size.

Remark 2. The variance estimator (due to Stehman and Overton (1994))

$$v^{o}_{HT} = n(s_z^2 - s_{zy}^2)$$

where $z_i = y_i/\pi_i$ $(i = 1,...,N)$; s_z^2 and s_{zy}^2 are the sample variance of z and covariance of z and y respectively, of V_{HT} takes frequently negative values under Sunter's method for most of the populations considered in this article.

Comparison under Poisson Sampling

In this section, we compare v_{HT}, v^{o}_{HT} (defined in Remark 2) and v_{PR} under Poisson sampling (see, e.g. Särndal *et al.* (1992).

Table 3. Relative percentage bias and RSE (under Poisson sampling)

Popl. No	n	RB %			RSE %		
		v_{HT}	v_{PR}	v^{o}_{HT}	v_{HT}	v_{PR}	v^{o}_{HT}
1.	10	0.529	−0.250	0.699	103.426	21.495	114.918
	15	0.117	−0.328	0.196	53.068	16.904	56.859
2.	10	0.842	−0.094	1.046	427.304	176.970	474.782
	15	0.658	−0.266	0.777	304.922	100.465	326.702
3.	10	0.346	−0.195	0.495	100.193	46.122	111.326
	15	0.350	−0.313	0.446	82.039	31.102	87.899
4.	10	0.873	−0.007	1.081	103.924	32.934	115.471
	15	0.926	−0.014	1.063	76.575	25.787	82.045
5.	10	0.729	−0.191	0.921	476.713	36.743	529.681
	15	0.509	−0.236	0.616	170.820	27.980	183.021
6.	10	1.102	0.248	1.335	146.653	39.451	162.947
	15	1.364	0.284	1.533	117.087	30.494	125.451
7.	10	0.603	−0.109	0.782	184.024	31.238	204.471
	15	0.593	−0.122	0.707	122.000	22.697	130.714
8.	10	0.949	0.254	1.165	134.449	39.630	149.388
	15	1.199	0.270	1.356	114.317	30.727	122.483
9.	10	1.201	−0.130	1.445	298.389	20.101	331.544
	15	1.021	−0.167	1.166	157.453	14.362	168.700
10.	10	1.140	−0.024	1.377	236.131	25.460	262.368
	15	1.150	−0.124	1.303	186.132	16.240	199.427
11.	10	0.896	−0.153	1.106	204.829	19.477	227.588
	15	1.061	−0.089	1.208	169.982	13.854	182.124
12.	10	0.529	−0.207	0.699	54.013	14.662	60.014
	15	0.574	−0.214	0.687	38.696	7.774	41.460
13.	10	−0.334	−0.167	−0.260	27.193	27.888	30.215
	15	−0.355	−0.137	−0.309	22.032	21.929	23.606
14.	10	0.299	−0.154	0.444	105.593	21.959	117.326
	15	0.595	−0.014	0.709	84.940	17.628	91.008

Table 4

Popl.No.	Source	X	Y	N	CV(x)	CV(y)	ρ
1.	Särndal *et al.* (1986)	CS82: Number of conservative seats in Municipal Council	RMT × 10^{-4}: Revenue from the 1985 municipal taxation	281	0.52	1.06	0.657
2.	Chambers *et al.* (1992)	Area assigned for sugarcane farms	Gross value of sugarcane	338	0.59	0.61	0.902
3.	Valliant *et al.* (2000)	Number of beds	Number of patients discharged	393	0.78	0.72	0.910
4.	Valliant *et al.* (2000)	Adult female population, 1960	Breast cancer mortality, 1950-69 (white female)	301	1.22	1.28	0.967
5.	Valliant *et al.* (2000)	Number of households, 1960	Population, excluding residents of group quarters, 1960	304	1.30	1.38	0.982
6.	Valliant *et al.* (2000)	Number of households, 1960	Population, excluding residents of group quarters, 1960	304	1.30	1.24	0.998

Some noteworthy results in Table 3 are:

(i) The excellent performances of RB% (except population 1 for n = 15) and RSE% associated with v_{PR} are evident for all populations. That is, v_{PR} outperforms the convention HT variance estimator v_{HT} under Poisson sampling.

(ii) The worst performer is v^{o}_{HT}.

4.2 Large Sample Performance

In this subsection, for the case n = 30 and for the large population (listed in Table 4), the above simulation procedure was repeated. The simulated values of the summary statistics under Sunter's sampling and Poisson sampling were presented in Table 5 and Table 6 respectively.

Results of Table 5 and Table 6 can be summarized as follows:

Under Sunter's sampling v_{YG} has responsible RB%, varying from 1% to 10%, whereas v_{PR} has large absolute RB%. Both the assumptions underlying the model are satisfied by populations 1 to 5 and therefore v_{PR} has considerably smaller RSE% than v_{YG}. For population 6, v_{PR} is worst. The reason is that the relation between y_i and x_i is a straight line through the origin but

Table 5. Relative percentage bias and RSE (under Sunter's scheme)

Popl. No.	RB%		RSE%	
	v_{YG}	v_{PR}	v_{YG}	v_{PR}
1.	3.087	4.310	40.655	35.804
2.	0.056	– 7.297	28.410	25.591
3.	3.201	– 16.452	44.979	29.643
4.	1.297	– 17.110	39.656	27.571
5.	3.974	13.576	62.286	52.257
6.	-8.601	– 71.259	48.884	73.432

Table 6. Relative percentage bias and RSE (under Poisson sampling)

Popl. No.	RB%			RSE%		
	v_{HT}	v_{PR}	v^{o}_{HT}	v_{HT}	v_{PR}	v^{o}_{HT}
1.	–0.383	2.668	–65.958	30.362	29.805	67.492
2.	–0.395	–1.668	–90.887	19.715	17.428	91.066
3.	–0.396	–3.246	–89.903	20.015	17.638	90.089
4.	–0.456	–1.493	–91.180	20.133	14.878	91.339
5.	–0.167	1.594	–95.966	18.773	14.056	96.088
6.	–0.279	–3.497	–95.776	18.363	13.419	95.880

the variance of y_i about this line is proportional to x_i which violate the assumption of the model namely $v(y_i) \propto x_i^2$.

Under Poisson sampling the absolute values of RB% of v_{HT} are all less than ½% for all the six populations, while v_{PR} out performs the conventional HT-variance estiamator v_{HT}.

5. CONCLUSION

Based on the empirical study and the theoretical discussion we arrive at the following conclusions :

1. It is clear from (3.1) that the PR-estimator will reflect the true variance closely when the best linear fit between y_i and x_i goes through the origin and the residuals from it are small. A high correlation between y_i and x_i is a necessary (though not a sufficient) condition for that kind of close fit which is evident from the following.

   Since $\rho_{YX} = \mathrm{Cov}(x_i, y_i)/\{\mathrm{SE}(x_i)\mathrm{SE}(y_i)\}$ and the regression coefficient of y_i on x_i is $\beta = \mathrm{Cov}(x_i, y_i)/\{\mathrm{SE}(x_i)\}^2$, the proportion of the variability in y_i explained by the regression of y_i on x_i is simply

   $$\frac{\beta^2\{\mathrm{SE}(x_i)\}^2}{\{\mathrm{SE}(y_i)\}^2} = \frac{\{\mathrm{Cov}(x_i, y_i)\}^2}{\{\mathrm{SE}(x_i)\}^2\{\mathrm{SE}(y_i)\}^2} = \rho_{YX}^2$$

   which depends only on ρ_{YX}.

2. The estimator v_{PR} was constructed assuming that the underlying super population model is a straight line passing through the origin and the variance of y_i is proportional to x_i^2. Therefore, the efficiency of the estimator depends on the goodness of fit of the model.

3. Implementation of the suggested estimator v_{PR} requires the complete auxiliary information, that is, values of x variable for the entire finite population. With strong auxiliary information the gain for using v_{PR} can be substantial compared to the HT and YG variance estimators under the specific assumptions.

4. The excellent performances of RB% and RSE% associated with v_{PR} are evident for most of the populations under Sunter's sampling and Poisson sampling when underlying assumptions are met. The worst performer is v_{HT}°.

5. Further empirical works through Monte Carlo for different populations under different sampling schemes are needed to assess the possible impact on the suggested estimator.

Appendix A. Proof of Theorem 1.

Let us begin by assuming that the random vector $\underline{Z} = (Z_1, Z_2, \ldots, Z_N)'$ have normal distribution with first two moments

$$\mathcal{E}(\underline{Z}) = \underline{\mu}$$
$$\mathcal{V}(\underline{Z}) = \Sigma \qquad \text{(A.1)}$$

where $\underline{\mu} = \mu \mathbf{1}_N, \Sigma = \sigma^2 \mathbf{I}_N$ and $\mathbf{1}_N$ is an N × 1 vector of unities, $\mathbf{I}_N$ is an N × N identity matrix and $\mathbf{J}_N = \mathbf{1}_N \mathbf{1}_N'$. Further assume that

$$\underline{Z} = \begin{bmatrix} \underline{Z}_s \\ \underline{Z}_r \end{bmatrix}, \Sigma = \begin{bmatrix} \Sigma_s & 0 \\ 0 & \Sigma_r \end{bmatrix}, \mathbf{J}_N = \begin{bmatrix} \mathbf{J}_s & \mathbf{J}_{sr} \\ \mathbf{J}_{rs} & \mathbf{J}_r \end{bmatrix}$$

where s and r denote the sets of sample and non-sample units.

The problem is to estimate $V_z = \underline{Z}' \Lambda \underline{Z}$ where Λ is an N × N n.n.d. and symmetric matrix of known constants.

Assuming that the non-informative sampling design is given, V_z and its quadratic predictor Q_z can be represented as

$$V_z = A(s, \underline{Z}) + B(r, \underline{Z})$$

and

$$Q_z = A(s, \underline{Z}) + U(s, \underline{Z}) \qquad \text{(A.2)}$$

where $A(s, \underline{Z}) = \underline{Z}_s' \Lambda_s \underline{Z}_s$, $B(r, \underline{Z}) = \underline{Z}_r' \Lambda_r \underline{Z}_r + \underline{Z}_s' \Lambda_{sr} \underline{Z}_r$ Λ is partitioned accordingly as $\underline{Z}$ and $U(s, \underline{Z})$ is considered a predictor of $B(r, \underline{Z})$. Further assuming that U is ξ-unbiased for $B(r, \underline{Z})$, (2.1) simplifies to

$$\mathcal{E}E(Q - V_z)^2 = E[\mathcal{V}(U) + \mathcal{V}(B(r, \underline{Z})) - 2\mathcal{C}\mathrm{ov}(U, B(r, \underline{Z}))] \qquad \text{(A.3)}$$

If a quadratic and x-unbiased predictor Q_z is of the form (A.2), then U must have the form

$$U(s, \underline{Z}) = \underline{Z}_s' \mathbf{H}_s \underline{Z}_s \qquad \text{(A.4)}$$

where $\mathbf{H}_s = (h(s,ij))$ is an $n \times n$ symmetric matrix of constants. Since Q_z is assumed to be a ξ-unbiased for V_z, we have

$$\mathcal{E}[U(s,\underline{Z}) - B(r,\underline{Z})] = 0 \quad \forall s \in S$$

from which, under model (A.1), we obtain

$$\mathcal{E}(U) = \sigma^2 tr(\Lambda_r) + \mu^2 \{tr(\Lambda_r \mathbf{J}_r) + 2tr(\Lambda_{sr}\mathbf{J}_{rs})\} \tag{A.5}$$

Moreover, from (A.4) and (A.5) we obtain

$$\left.\begin{aligned} tr(\mathbf{H}_s) &= tr(\Lambda_r) \\ tr(\mathbf{H}_s\mathbf{J}_s) &= tr(\Lambda_r\mathbf{J}_r) + tr(\Lambda_{sr}\mathbf{J}_{rs}) \end{aligned}\right\} \tag{A.6}$$

Since $\underline{Z}_s$ and $\underline{Z}_r$ are independent, it follows that $Cov(U, B(r,\underline{Z}) = 0$. Therefore, the $\mathcal{E}$MSE given by (A.3) is minimized if, for every fixed $s \in S$, we choose U to minimize $\mathcal{V}(U)$ subject to (A.5), where σ^2 and μ^2 are the unknown quantities.

By generalized least square theory, the uniformly minimum ξ-variance and ξ-unbiased estimators of μ^2 and σ^2 (Arnold (1979)) are

$$\hat{\mu}^2 = \frac{1}{n(n-1)} \Sigma\Sigma_s Z_i Z_j ; \hat{\sigma}^2 = s_z^2$$

where $s_z^2 = \frac{1}{n}\Sigma_s Z_i^2 - \frac{1}{n(n-1)}\Sigma\Sigma_s Z_i Z_j$

Substituting these estimators in (A.5) we get with the help of (A.6)

$$U^* = tr(\Lambda_r)s_z^2 + [tr(\Lambda_r\mathbf{J}_r) + tr(\Lambda_{sr}\mathbf{J}_{rs})] \frac{1}{n(n-1)}\Sigma\Sigma_s Z_i Z_j$$

Clearly

$$Q_z^* = A(s,\underline{Z}) + U^*(s,\underline{Z}) \tag{A.7}$$

minimizes $\mathcal{E}(Q - V_z)^2$ for every $s \in S$ and hence also $\mathcal{E}E(Q - V_z)^2$.

The proof of Theorem 1 follows from (A.7) by putting $\underline{Z} = D^{-1}\underline{Y}$ and $\Lambda = \mathbf{D}'\Delta\mathbf{D}$ where $\underline{Y} = (Y_1, Y_2, ..., Y_N)'$, $\mathbf{D} = diag\,(x_1, x_2, ..., x_N)$ and Δ is an $N \times N$ n.n.d. and symmetric matrix.

Appendix B. Proof of Theorem 2.

The proof is based on the following results.

Result 1. If a quadratic form $\underline{X}'\mathbf{A}\underline{X} \geq 0$ (≥ 0 means n.n.d.), every principal minor determinant of A is ≥ 0. Moreover, if x_i actually appears in $\underline{X}\,\mathbf{A}\underline{X}$ then $a_{ii} > 0$.

Result 2. A pattern matrix $\mathbf{A} = (a-b)\mathbf{I}_n + b\mathbf{J}_n$ is n.n.d. either a = b or a = –(n – 1)b. Further

$\mathbf{A} \geq 0$ if and only if $b \leq 0$, for $a = -(n-1)b$

Assume without loss of generality that the first n population units have been selected in the sample. With the help of Result 1, since $V_{HT} = \underline{Y}'\Delta\underline{Y} \geq 0$ it follows that $A_{HT}(s,\underline{Y}) = \underline{Y}_s'\Delta_s\underline{Y}_s \geq 0$ where Δ_s is an $n \times n$ diagonal block matrix of Δ.

Next, using Result 2 with

$$a = \frac{1}{n}\left\{\Sigma_U\left(\frac{1}{\pi_k} - 1\right)x_k^2 - \Sigma_S\left(\frac{1}{\pi_k} - 1\right)x_k^2\right\}$$

$$b = \frac{1}{n(n-1)}\left\{\Sigma\Sigma_U\left(\frac{\pi_{kl}}{\pi_k\pi_l} - 1\right)x_k x_l - \Sigma\Sigma_S\left(\frac{\pi_{kl}}{\pi_k\pi_l} - 1\right)x_k x_l\right\}$$

and $z_i = Y_i/x_i$, $i \in s$, one can see that $U_{PR}(s,\underline{Y}) \geq 0$ if and only if $b \leq 0$ i.e. if and only if

$$\Sigma\Sigma_U\left(\frac{\pi_{kl}}{\pi_k\pi_l} - 1\right)x_k x_l \leq \Sigma\Sigma_S\left(\frac{\pi_{kl}}{\pi_k\pi_l} - 1\right)x_k x_l$$

i.e. if and only if $\pi_{kl} \leq \pi_k\pi_l \;\forall\, k \neq l \in U$ which complete the proof.

ACKNOWLEDGEMENT

The authors would like to express their thanks to anonymous referee for his helpful and constructive observations on earlier version of this article.

REFERENCES

Arnold, S.F. (1979). Linear models with exchangeably distributed errors. *J. Amer. Statist. Assoc.*, **74**, 194-199.

Brewer, K.R.W. and Tam, S.M. (1990). Is the assumption of uniform intra-class correlation ever justified? *Austr. J. Statist.*, **32(3)**, 411-423.

Cassel, C.M., Särndal, C.E., and Wretman, J.H. (1977). *Foundation of Inference in Survey Sampling*. John Wiley, New York.

Chambers, R.L. and Dunstan, R. (1986). Estimating distribution function from survey data. *Biometrika,* **73,** 597-604.

Cumberland, W.G. and Royall, R.M. (1981). The finite-population linear regression estimator and estimators of its variance - An empirical study. *J. Amer. Statist. Assoc.,* **76,** 924-930.

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

Isaki, C.T. (1983). Variance estimation using auxiliary information. *J. Amer. Statist. Assoc.,* **78(381),** 117-123.

Kish, L. (1965). *Survey Sampling*. John Wiley, New York.

Murthy, M.N. (1967). *Sampling Theory and Methods*. Statistical Publishing Society, Calcutta.

Rao, J.N.K. and Singh, M.P. (1973). On the choice of estimator in survey sampling. *Austr. J. Statist.,* **15(2),** 95-104.

Rodrigues, J., Bolfarine, H. and Rogatko, A. (1985). A general theory of prediction in finite population. *Inter. Statist. Rev.,* **53,** 239-254.

Royall, R.M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika,* **57,** 377-387.

Royall, R.M. (1971). Linear regression models in finite population sampling (with discussion). In : *Foundation of Statistical Inference,* V.P. Godambe and D.A. Sprott (Editors), 259-279. Holt, Rinehart and Winston, Montreal, Toronto.

Royall, R.M. and Eberhardt, K.R. (1975). Variance estimates for the ratio estimator. *Sankhya,* **C37,** 43-52.

Särndal, C.E., Swensson, B. and Wretman, J.H. (1992). *Model Assisted Survey Sampling*. Springer Verlag, New York.

Singh, S., Horn, S., Chowdhury, S. and Yu, F. (1999). Calibration of the estimators of variance. *Austr. & New Zealand J. Statist.,* **41(2),** 199-212.

Stehman, S.V. and Overton, W.S. (1994). Comparison of variance estimators of the Horvitz-Thompson estimator for randomized variable probability systematic sampling. *J. Amer. Statist. Assoc.,* **89,** 30-43.

Sukhatme, P.V. and Sukhatme, B.V. (1970). *Sampling Theory of Survey with Application,* 2nd Edition. Asia Publishing House, London.

Sunter, A.B. (1986). Solution to the problem unequal probability sampling without replacement. *Inter. Statist. Rev.,* **54,** 33-50.

Valliant, R., Dorfman, A.H. and Royall, R.M. (2000). *Finite Population Sampling and Inference*. John Wiley and Sons Inc., New York.

Wolter, K.M. (1985). *Introduction to Variance Estimation*. Springer Verlag, New York.

Yates, F. (1960). *Sampling Methods for Censuses and Surveys,* 3rd Edition. Hafner Publishing Company, New York.

Yates, F. and Grundy, P.M. (1953). Selection without replacement from within strata with probability proportional to size. *J. Roy. Statist. Soc.,* **B15,** 235-261.

Zarnoch, S.J. and Bechtold, W.A. (2000). Estimating mapped plot forest attributes with ratios of means. *Can. J. For. Res.,* **30,** 688-697.