ON VARIANCE ESTIMATOR IN UNEQUAL PROBABILITY SAMPLING

S. H. Biyani, East Carolina University

1. INTRODUCTION

Various design-unbiased estimators of the variance of the Horvitz-Thompson estimator are given by Horvitz and Thompson (1952), Yates and Grundy (1952) and Ajgaonkar (1967). Each of these estimators can take values which are either negative, or positive but otherwise known to be impossible (Biyani 1980). Even the most preferred Yates-Grundy estimator can be inadmissible for any sample size, in the class of nonnegative quadratic estimators. Three estimators presented in this paper have the following properties:

1. The values assumed by each are always a posteriori possible values of the true variance. In particular, they are nonnegative, even if the well known inequality $\pi_{ij} \leq \pi_i\pi_j$ does not hold.
2. Each estimator is derived as an optimal estimator under a Random Permutation Model, in a reasonable subclass of nonnegative quadratic estimators. The admissibility of the estimators in the respective subclasses is an immediate consequence of this.

2. PRELIMINARIES

Let $y = (y_1, y_2, \ldots, y_N)$ be the population vector. The Horvitz-Thompson estimator of the population total, based on a sample s, is given by $e_{HT} = \Sigma_{i\varepsilon s} z_i$, where $z_i = y_i/\pi_i$, and its variance is

$$V(e_{HT}) = \Sigma^N_{i<j} c_{ij} f_{ij}, \quad \text{where} \quad c_{ij} = \pi_i\pi_j - \pi_{ij}$$

and $f_{ij} = (z_i - z_j)^2$.

Throughout this paper, we will assume a fixed sample size design with $\pi_i > 0$ for all i. We will consider the class K of estimators of the form $\Sigma_s b_{sij} f_{ij}$, where b_{sij} are constants. All nonnegative quadratic design-unbiased estimators of $V(e_{HT})$ are included in this class (Vijayan 1975). The same result holds when the design-unbiasedness is replaced by the weaker requirement that the estimator vanish when $z_1 = z_2 = \ldots = z_N$ (and hence $V(e_{HT}) = 0$). This seems reasonable because one may expect an estimator to be without error when there is no variability in the population.

We will also consider two subclasses of K, defined for a given function to be estimated and with respect to a given sampling design p and superpopulation model ξ : K_ξ, the class of ξ-unbiased estimators in K, and $K_{p\xi}$, the class of $p\xi$-unbiased estimators in K. (See Cassel, Särndal and Wretman (1977) for definitions.) The subclass of p-unbiased estimators in K is omitted, primarily due to algebraic difficulties which prevent an explicit solution for optimal estimator in this class, and also due to the fact that the "best" p-unbiased estimator would not have the property (1) mentioned in the preceding section at least for sample size two (Biyani 1980).

3. THE MODEL

Optimal estimators, in the sense of minimum expected mean square error, are derived under the following model:

1. $\underset{\sim}{z} = (z_1, z_2, \ldots, z_N)$ is the realization of a random vector $\underset{\sim}{Z}$.
2. The possible values of $\underset{\sim}{Z}$ are the permutations of a fixed unknown vector, each with probability $1/N!$.
3. N is large and $\beta_2 = 3$, where $\beta_2 = [\Sigma^N(z_i - \bar{z})^4/N] \ / \ [\Sigma^N(z_i - \bar{z})^2/N]^2$.

[For any N, the optimality results hold for $\beta_2 = 3(N - 1)/(N + 1)$. This condition is always satisfied for $N = 3$. It is not required for the optimality in K_ξ for the special case $n = 2$.]

In practical terms this model may be interpreted as an assumption that there is no relevant information left in the labels after transforming from y_i's to z_i's. In particular, this means that Z_i's are unrelated to π_i's. Godambe and Thompson (1973) have shown the optimality of e_{HT} under essemtially the same model, except condition (3).

4. THE ESTIMATORS

Notation:

$$c_{ij} = \pi_i\pi_j - \pi_{ij}$$

$$f_{ij} = (y_i/\pi_i - y_j/\pi_j)^2$$

$$f_{is} = \Sigma_{i\varepsilon s} f_{ij} \ / \ (n - 1)$$

$$f_s = \Sigma_s f_{ij} \ / \ (\tfrac{n}{2})$$

Σ_s = sum over i, j in s, $i < j$.

$\Sigma_{s\sim}$ = sum over i in s, j not in s .

$\Sigma_{\sim}$ = sum over i, j not in s, $i < j$.

Theorem 4.1: Under the model of Section 3, for any fixed sample size design, the optimal estimators of $V(e_{HT})$ in the classes K, K_ξ and $K_{p\xi}$ are, respectively,

$$v_0 = \Sigma_s c_{ij} f_{ij} + \Sigma_{s\sim} c_{ij} \{(n-1)/n\} \bar{f}_{is} + \Sigma_{\sim} c_{ij} \{(n-1)/(n+1)\} \bar{f}_s \qquad (4.1)$$

$$v_\xi = \Sigma_s c_{ij} f_{ij} + \Sigma_{s\sim} c_{ij} [\{(n-1)/n\} \bar{f}_{is} + (1/n) \bar{f}_s] + \Sigma_{\sim} c_{ij} \bar{f}_s \qquad (4.2)$$

$$v_{p\xi} = v_0 + A(\tfrac{n}{2}) \bar{f}_s , \qquad (4.3)$$

where

$$A = 2\Sigma^N_{i<j} c_{ij} \left[\frac{2 - \pi_i - \pi_j + 2(1 - \pi_{ij})/(n - 1)}{n^2(n + 1)}\right] \qquad (4.4)$$

Proof: See Biyani (1979).

Interpretation of the estimators: Note that $V(e_{HT}) = \Sigma_s c_{ij} f_{ij} + \Sigma_{s\sim} c_{ij} f_{ij} + \Sigma_{\sim} c_{ij} f_{ij}$.

The first sum involves only the sampled pairs of units, and is completely known. The second and third sums involve pairs with one or both units not in the sample, respectively. Each unknown f_{ij} in these sums is replaced by a suitable estimate in (4.1) and (4.2). When both i and j are not in s, the estimate is based on all sampled pairs. When i is in s and j is not in s, the estimate is based (mostly) on sampled pairs involving i. The last two sums in the expression for v_0 involve shrinkage factors of $(n - 1) / n$ and $(n - 1) / (n + 1)$, respectively, making it negatively biased. The second term in the expression for $v_{p\xi}$ represents a correction for this bias.

For Simple Random Sampling, both v_ξ and $v_{p\xi}$ reduce to the usual unbiased estimator, while v_0 reduces to $(n - 1)(N + 1) / (n + 1)(N - 1)$ times the usual estimator.

Computational considerations: Equations (4.1)-(4.4) represent the heuristic forms of the estimators. Computational forms, involving only sums over the sampled pairs, can be easily derived using the relationship

$$\Sigma^N_{j(\neq i)} c_{ij} = \pi_i(1 - \pi_i) .$$

The actual computation of v_0 and v_ξ involves only slightly more work than for the Yates-Grundy estimator. The computation of $v_{p\xi}$ can become impractical for large N, without the aid of a computer. However, v_ξ and $v_{p\xi}$ have been empirically found to be nearly identical. Thus the former may be used as an approximation for the latter.

5. NUMERICAL RESULTS

The relative efficiencies of different estimators of $V(e_{HT})$ are empirically compared using some "real" populations listed in Table 1. The estimators compared include v_0, v_ξ, $v_{p\xi}$ and the estimators of Horvitz and Thompson (1952) and Yates and Grundy (1953) (v_{HT}, v_{YG}) and the following:

$$v_F = \Sigma_s c_{ij} f_{ij} \pi_{ij}{}^{-1} (\Sigma_s c_{ij} \pi_{ij}{}^{-1})^{-1} \Sigma^N_{i<j} c_{ij}$$

(Fuller 1970), and

$$v_R = \Sigma_s c_{ij} f_{ij} (\Sigma_s c_{ij})^{-1} \Sigma^N_{i<j} c_{ij} .$$

The sampling scheme of Sampford (1967) was used to draw samples with inclusion probabilities proportional to auxiliary variable x. The results, based on 1000 samples, are shown in Table 2.

We note that v_0 is more efficient than v_{YG} in all cases considered, but it is not the most efficient estimator in all cases. In particular, for population 10 which contains an extreme value of z_i (resulting in $\beta_2 = 19.9$), v_0 is much less efficient than v_R, while the design-based estimators are still worse. It is clear that the optimality of v_0 is destroyed by the model breakdown, but the design-based estimators fail to solve the problem.

ACKNOWLEDGMENT

This publication was supported by a grant from East Carolina University Research Committee.

REFERENCES

Ajgaonkar, S. G. P. (1967). "Unbiased Estimator of Variance of Narain, Horvitz and Thompson Estimator," *Sankhya Ser. A*, 29, 55-60.

Biyani, S. H. (1979). "On Estimation of Variance in Unequal Probability Sampling," University of Minnesota Technical Report 349.

Biyani, S. H. (1980). "On Inadmissibility of the Yates-Grundy Estimator in Unequal Probability Sampling," scheduled to appear in the *Journal of the American Statistical Association*, September 1980.

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

Cochran, W. G. (1977). *Sampling Techniques*. New York: John Wiley and Sons.

Fuller, W. A. (1970). "Sampling with Random Stratum Boundaries," *Journal of the Royal Statistical Society, Ser. B*, 32, 209-226.

Godambe, V. P. and Thompson, M. E. (1973). "Estimation in Sampling Theory with Exchangeable Priors," *Annals of Statistics*, 1, 1212-1221.

Hanurav, T. V. (1967). "Optimal Utilization of Auxiliary Information in πps Sampling of Two Units from a Stratum," *Journal of the Royal Statistical Society, Ser. B*, 29, 374-391.

Horvitz, D. G. and Thompson, D. J. (1952). "A Generalization of Sampling without Replacement from a Finite Universe," *Journal of the American Statistical Association*, 47, 663-685.

Rao, J. N. K. (1963). "On Three Procedures of Unequal Probability Sampling without Replacement," *Journal of the American Statistical Association*, 58, 202-215.

Sampford, M. R. (1967). "On Sampling without Replacement with Unequal Probabilities of Selection," *Biometrika*, 54, 499-513.

Sukhatme, P. V. and Sukhatme, B. V. (1970). *Sampling Theory of Surveys with Applications*. Ames, Iowa: Iowa State University Press.

Vijayan, K. (1975). "On Estimating the Variance in Unequal Probability Sampling," *Journal of the American Statistical Association*, 70, 713-716.

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

Yates, F. and Grundy, P. M. (1953). "Selection without Replacement from within Strata with Probability Proportional to Size," *Journal of the Royal Statistical Society, Ser B*, 15, 235-261.

Table 1. Populations Used in the Study

Pop. No.	Source	y	x	N	β_2
1	Hanurav (1967), p. 386)	1960 population	1950 population	20	9.0
2	Yates (1960, p. 159)	number of absentees	total number of persons	43	3.2
3	Sukhatme and Sukhatme (1970, p. 166)	number of banana bunches	number of banana pits	20	3.1
4	Sukhatme and Sukhatme (1970, p. 51)	area under rice	total cultivated area	25	1.9
5	Rao (1963, p. 207)	1960 area under corn	1958 area under corn	14	1.9
6	Cochran (1977, p. 203)	weight of peaches	eye-estimate	10	1.4
7	Cochran (1977, p. 325)	number of persons	number of rooms	10	2.1
8	Sukhatme and Sukhatme (1970, p. 183)	1937 area under wheat	1936 area under wheat	34	3.4
9	Subset of 10 (see text)			23	2.6
10	Yates (1960, p. 163)	volume of timber	eye-estimate	25	19.9

Table 2. Efficiencies Relative to the Yates-Grundy Estimator

Pop. No.	Sample Size	v_{HT}	v_F	v_R	v_0	v_ξ
1	3	.09	1.01	1.07	2.63	1.04
	5	.04	1.06	1.15	1.91	1.12
	10	.004	1.08	1.17	1.44	1.26
2	3	1.02	1.04	1.22	1.98	1.06
	5	1.05	1.06	1.31	1.54	1.14
	10	1.11	1.14	1.86	1.68	1.53
3	3	.59	.98	.92	1.89	.97
	5	.24	.99	.91	1.29	.96
	10	.03	1.05	.93	1.08	.95
4	3	.14	1.07	1.07	1.75	1.11
	5	.05	1.21	1.32	1.56	1.37
5	3	.08	1.19	1.39	1.62	1.33
	5	.02	1.28	1.43	1.28	1.52
6	3	.002	1.09	1.19	1.72	1.15
	5	.0003	1.21	1.34	1.45	1.36
7	3	.15	1.13	1.22	1.75	1.18
	5	.03	1.19	1.37	1.48	1.38
8	3	.23	.88	.58	1.83	.77
	5	.08	.95	.41	1.15	.64
	10	.02	1.39	3.89	2.08	1.04
9	3	.85	1.05	1.11	1.92	1.08
	5	.61	1.07	1.31	1.48	1.20
	10	.22	1.24	1.77	1.80	1.76
10	3	1.00	1.18	6.53	4.43	1.36
	5	1.02	1.22	12.19	3.93	1.97
	10	1.01	1.22	8.38	5.06	4.18