AN EMPIRICAL STUDY OF STABILITIES OF VARIANCE ESTIMATORS II

Chand Midha, University of Akron

Several estimators for the variance of the Horvitz-Thompson estimator of the population total are available in the literature. The stability of the several estimators is compared numerically.

KEY WORDS: Horvitz-Thompson Estimator.

1. Introduction.

Several estimators for V_{HT}, the variance of Horvitz-Thompson estimator have been proposed in the recent years. Some of these proposed estimators are $V_{HT}(1)$, due to Horvitz-Thompson (1952); $V_{HT}(2)$, due to Yates and Grundy (1953); $V_{HT}(3)$, due to Fuller (1970); $V_{HT}(4)$, $V_{HT}(5)$, and $V_{HT}(6)$, due to Biyani (1978). However, not much is known with regard to stability of these variance estimators. Therefore, in this paper, the stability of the several estimators is compared numerically, by means of analytic expressions for the variances of these estimators. The estimators chosen for this study are $V_{HT}(1)$, $V_{HT}(2)$, $V_{HT}(3)$, $V_{HT}(6)$, together with $V_{HT}(7)$ (simplified version of $V_{HT}(5)$ and $V_{HT}(8)$ (simplified version of $V_{HT}(4)$). The estimators, $V_{HT}(4)$ and $V_{HT}(5)$ are excluded because analytic expressions for the mean square error of these estimators are very cumbersome. However, we have compared the estimators, $V_{HT}(4)$ and $V_{HT}(5)$ with other estimators, using the criterion of mean square error based on all possible samples. The rejective sampling scheme of Sampford (1967) was used to draw the samples, with inclusion probabilities proportional to some auxillary variable x, related to the variable of interest y.

2. Formulae.

The Horvitz-Thompson estimator of the population total based on a given sample, s, is given by

$$e_{HT} = \sum_{i\varepsilon s} z_i, \text{ where } z_i = y_i/\Pi_i$$

The variance of e_{HT} can be expressed in two equivalent forms as

$$V_{HT} = \sum_{ij}^{N} (\Pi_{ij}-\Pi_i\Pi_j)z_iz_j$$

and

$$V_{HT} = \sum_{i\varepsilon j}^{N} (\Pi_i\Pi_j-\Pi_{ij})(z_i-z_j)^2.$$

Some of the proposed estimators for V_{HT} are given below:

$$V_{HT}(1) = \sum_{i<j\varepsilon s} (\Pi_{ij}-\Pi_i\Pi_j)/\Pi_{ij}\cdot z_iz_j$$

(due to Horvitz-Thomspon (1952))

$$V_{HT}(2) = \sum_{i<j\varepsilon s} c_{ij}/\Pi_{ij}\cdot f_{ij}$$

(due to Yates and Grundy (1953))

$$V_{HT}(3) = (\sum_s c_{ij}f_{ij}/\Pi_{ij})(\sum_s c_{ij}/\Pi_{ij})^{-1} \sum_{i<j}^{N} c_{ij}$$

(due to Fuller (1970))

$$V_{HT}(4) = \sum_{i<j\varepsilon s} c_{ij}f_{ij} + \sum_{i\varepsilon s} c_{\tilde{s}i}\cdot \bar{f}_{is} + (n-1)/(n+1)\cdot c_{\tilde{s}}\bar{f}_s$$

(due to Biyani (1978))

$$V_{HT}(5) = \sum_{i<j\varepsilon s} c_{ij}f_{ij} + \sum_{i\varepsilon s} c_{\tilde{s}i}[(n-1)\bar{f}_{is} + \bar{f}_s]/n + c_{\tilde{s}}f_s$$

(due to Biyani(1978))

$$V_{HT}(6) = (\sum_s c_{ij}f_{ij})(\sum_s c_{ij})^{-1} \sum_{i<j}^{N} c_{ij}.$$

(due to Biyani (1978))

and

$$V_{HT}(7) = 2\sum_{i<j}^{N} c_{ij} \sum_{i<j\varepsilon s} f_{ij}/n(n-1)$$

(simplied version of $V_{HT}(5)$)

$$V_{HT}(8) = 2(n-1)/(n+1)\cdot \sum_{i<j}^{N} c_{ij}\cdot \sum_{i<j\varepsilon s} f_{ij}/n(n-1)$$

(simplied version of $V_{HT}(4)$), where

$$f_{ij} = (z_i-z_j)^2,\ \phi_{ij} = z_iz_j,$$

$$\bar{f}_{is} = \sum_{j\varepsilon s} f_{ij}/(n-1),\ c_{ij} = \Pi_i\Pi_j-\Pi_{ij},$$

$$\bar{f}_s = \sum_{i\varepsilon s} \bar{f}_{is}/n,\ c_{\tilde{s}i} = \sum_{k\notin s} c_{ik}$$

and

$$c_{\tilde{s}} = \sum_{k<l\varepsilon\tilde{s}} c_{ke}$$

The estimators $V_{HT}(1)$ and $V_{HT}(2)$ are design-unbiased, while $V_{HT}(6)$ is approximately so. The estimators $V_{HT}(3)$, $V_{HT}(5)$, and $V_{HT}(6)$ are model-unbiased, based on the model used in Biyani (1978).

The stability of the estimators, $V_{HT}(1)$, $V_{HT}(2)$, $V_{HT}(3)$, $V_{HT}(6)$, together with $V_{HT}(7)$ and $V_{HT}(8)$, is compared numerically, by using the analytic expressions for the variances of these estimators. These expressions are given in Midha (1980).

3. Empirical Results.

The populations used in the study are listed in Table 1. In each case, we have an auxillary variable x approximately proportional to the variable y of interest. For most of the populations, the Horvitz-Thompson estimator can be expected to perform reasonably well as an estimator of the population total.

Table 2 gives the efficiencies of the estimators over Yates-Grundy estimator (i.e. [MSE (Yates-Grundy est.)/MSE (est.)]) for the populations of Table 1 for sample sizes n = 3, 5, 7. The following tentative conclusions can be drawn from Table 2:

(1) The Horvitz-Thompson estimator $V_{HT}(1)$ is the worst performer among all the estimators considered.

(2) Fuller's estimator $V_{HT}(3)$ and the ratio-type estimator $V_{HT}(6)$ are generally better than the Yates-Grundy estimator for sample size n=7. $V_{HT}(3)$ seems to be more efficient even for sample sizes 3 and 5.

(3) $V_{HT}(7)$ and $V_{HT}(8)$ have performed generally better than the rest; however, $V_{HT}(7)$ appears less efficient than the latter. In fact, $V_{HT}(8)$ is consistently more efficient than the rest and the gains are considerable for some populations. Therefore, it seems that the estimator $V_{HT}(8)$ compares favorably with those of $V_{HT}(1)$, $V_{HT}(3)$, $V_{HT}(6)$, and $V_{HT}(7)$.

Biyani (1978) showed that the estimators $V_{HT}(4)$, $V_{HT}(5)$, and $V_{HT}(6)$ are more efficient than the rest, using the criterion of mean square error. While comparing the stabilities of variance estimators, we have considered the simplified versions of $V_{HT}(4)$ and $V_{HT}(5)$, namely $V_{HT}(7)$ and $V_{HT}(8)$. In order to see how good the simplied versions are compared to the original estimators, we have compared all the estimators using the criterion of mean square error based on all possible samples. We have not included the Horvitz-Thompson estimator $V_{HT}(1)$ because it was found to be the worst among the estimators compared, using the analytic expressions of variance. The populations and samples used in the study are the same as in the empirical comparison of analytic expressions. Table 3 gives the efficiencies of the estimators over Biyani's estimator, $V_{HT}(4)$. The following tentative conclusions can be drawn from Table 3:

(1) The Yates-Grundy estimator is generally poor compared to other estimators.

(2) The efficiencies of $V_{HT}(3)$, $V_{HT}(5)$, and $V_{HT}(7)$ are essentially identical; however, $V_{HT}(3)$ appears slightly less efficient than the latter.

(3) $V_{HT}(8)$ is almost as efficient as $V_{HT}(4)$ for most of the populations considered, but in some cases it is considerably less efficient.

(4) The ratio type estimator, $V_{HT}(6)$ is considerably more efficient thatn $V_{HT}(4)$ in many cases; however, it is considerably less efficient in several other cases. It appears on the whole that $V_{HT}(4)$ (or its simplified version $V_{HT}(8)$) and the ratio-type estimator $V_{HT}(6)$ are more efficient than the other estimators, at least for the populations considered.

TABLE 1 POPULATIONS USED IN THE STUDY

Popn. No.	Source	y	x	N
1	Scheaffer, Mendenhall, and Ott (1979, p. 134)	Present sales	Precampaign sales	20
2	Jessen (1978, p. 264)	Production of corn	Area in corn	20
3	Johnson and Smith (1969, p. 182)	Acreage under oats in 1957	Acreage of crops and grass in 1947	20
4	Des Raj (1972, p. 48)	Number of dwellings occupied by renters	Total number of dwellings	20
5	Sukhatme and Sukhatme (1970, p. 256)	Area under wheat	Number of villages	20

TABLE 2 EFFICIENCY OF THE ESTIMATORS RELATIVE TO $V_{HT}(2)$ ESTIMATOR

Popn. No.	Sample Size	$V_{HT}(1)$	$V_{HT}(3)$	$V_{HT}(6)$	$V_{HT}(7)$	$V_{HT}(8)$
1	3	0.06	1.01	5.59	1.44	5.30
	5	0.03	1.02	5.52	1.93	4.04
	7	0.02	1.03	6.03	2.72	4.38
2	3	0.45	1.04	0.34	1.13	1.93
	5	0.34	1.17	0.91	1.56	1.99
	7	0.33	1.33	2.61	2.57	2.83
3	3	0.27	1.10	0.34	1.25	1.55
	5	0.11	1.23	0.50	1.55	1.22
	7	0.05	1.24	1.10	1.98	1.17
4	3	0.01	1.06	0.45	1.22	1.79
	5	0.04	1.14	0.82	1.59	1.64
	7	0.02	1.20	1.72	2.27	1.99
5	3	0.66	1.06	1.13	1.28	2.83
	5	0.53	1.11	1.42	1.58	2.27
	7	0.45	1.17	1.92	1.99	2.18

TABLE 3 EFFICIENCY OF THE ESTIMATORS RELATIVE TO $V_{HT}(4)$ ESTIMATOR

Popn. No.	Sample Size	$V_{HT}(2)$	$V_{HT}(3)$	$V_{HT}(5)$	$V_{HT}(6)$	$V_{HT}(7)$	$V_{HT}(8)$
1	3	0.196	0.249	0.304	1.079	0.282	1.038
	5	0.219	0.282	0.537	2.019	0.422	0.882
	7	0.188	0.246	0.709	1.636	0.509	0.819
2	3	0.499	0.553	0.569	0.600	0.563	0.962
	5	0.455	0.619	0.759	0.862	0.713	0.906
	7	0.315	0.509	0.929	1.037	0.812	0.897
3	3	0.596	0.702	0.756	0.731	0.743	0.923
	5	0.704	0.925	1.106	1.064	1.088	0.825
	7	0.627	0.864	1.262	1.220	1.247	0.732
4	3	0.522	0.597	0.652	0.717	0.637	0.936
	5	0.531	0.663	0.897	1.013	0.843	0.871
	7	0.414	0.561	0.980	0.961	0.934	0.821
5	3	0.354	0.419	0.474	0.678	0.455	1.002
	5	0.420	0.519	0.744	1.075	0.667	0.955
	7	0.422	0.547	0.933	1.139	0.835	0.917

REFERENCES

Biyani, S.H., 1978. "On some estimation problems in finite population sampling," Unpublished Ph.D. Dissertation, Iowa State University, Ames, Iowa.

Des, Raj, 1972. *The Design of Sample Surveys*, McGraw-Hill, New York.

Fuller, W.A., 1970. "Sampling with random stratum boundaries," *J. Roy. Stat. Soc., Ser. B*, 32:209-226.

Horovitz, D.G. and D.J. Thompson, 1952. "A generalization of sampling without replacement from a finite universe, *J. Amer. Stat. Assoc.*, 47:663-685.

Jessen, R.A., 1978. *Statistical Survey Techniques*, Wiley, New York.

Johnson, N.L. and H. Smith, 1969. *New Developments in Survey Sampling*, Wiley, New York.

Midha, C.K., 1980. "Contributions to survey sampling and design of experiments," Unpublished Ph.D. Dissertation, Iowa State University, Ames, Iowa.

Sampford, M.R., 1967. "On sampling without replacement with unequal probabilities of selection," *Biometrika*, 54:499-513.

Scheaffer, R.L., W. Mendenhall, and L. Ott, 1979. *Elementary Survey Sampling*, Duxbury, California.

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

Yates, F. and P.M. Grundy, 1953. "Selection without replacement from within strata with probability proportional to size," *J. Roy. Stat. Soc., Ser. B*, 15:235-261.