THE HORVITZ-THOMPSON ESTIMATOR AND ESTIMATES OF ITS VARIANCE

Chand Midha
The University of Akron, Akron, Ohio 44325

Introduction: Several estimators for V_{HT}, the variance of Horvitz-Thompson estimator have been proposed in the recent years. For comparing variance estimators, the criterion of mean square error is appealing, but not necessarily best. An estimator with a relatively large mean square error may come fairly close to the true value in a large proportion of samples. Therefore, within the framework of conventional sampling theory, one is most interested in the accuracy of confidence intervals and the average width of confidence intervals based on a given variance estimator. Therefore, in this paper, we have emperically compared the distribution of standarized (studentized) estimates and average widths of confidence intervals obtained using different estimators of the variance of the Horvitz-Thompson estimator of the population total.

Formulae: The Horvitz-Thompson estimator of the population total is given by

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

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

$$V_{HT} = \sum_{i,j}^{N} (\pi_{ij} - \pi_i\pi_j)\, z_i z_j$$

and $$V_{HT} = \sum_{i<j}^{N} (\pi_i\pi_j - \pi_{ij})(z_i - z_j)^2$$

Some of the proposed estimators for the variance of the Horvitz and Thompson estimator, V_{HT}, are defined below:

$$V_{HT}(1) = \sum_{i\varepsilon s} (1-\pi_i) z_i^2 + 2 \sum_{i<j\varepsilon s} \frac{(\pi_{ij}-\pi_i\pi_j)}{\pi_{ij}} \cdot z_i z_j$$

(due to Horvitz-Thompson (1952))

$$V_{HT}(2) = \sum_{i<j\varepsilon s} \frac{(\pi_i\pi_j-\pi_{ij})}{\pi_{ij}} (z_i-z_j)^2$$

(due to Yates and Grundy (1953))

$$V_{HT}(3) = \Big(\sum_{s} \frac{c_{ij}f_{ij}}{\pi_{ij}}\Big)\Big(\sum_{s}\frac{c_{ij}}{\pi_{ij}}\Big)^{-1} \sum_{i<j}^{N} c_{ij}$$

(due to Fuller (1970))

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

(due to Biyani (1978))

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

(due to Biyani (1978))

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

(Ratio type estimator)

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

(Simplified version of V_{HT} (5))

where, the sum $\sum_{i\varepsilon s}$ is over all distinct units of samples s, and $f_{ij} = (z_i - z_j)^2$,

$$\bar{f}_{is} = \sum_{j\varepsilon s} \frac{f_{ij}}{(n-1)},$$

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

$$f_s = \frac{1}{n} \sum_{i\varepsilon s} \bar{f}_{is} \text{ ,}$$

$$c_{si} = \sum_{k\notin s} c_{ik} = \pi_i(1-\pi_i) - \sum_{k\in s} c_{ik} \text{ ,}$$

$$c_{\tilde{s}s} = \sum_{i\notin s} c_{\tilde{s}i} \text{ and}$$

$$c_s = \sum_{k<l\varepsilon s} c_{kl} = \sum_{i=1}^{N} \frac{\pi_i(1-\pi_i)}{2} - c_{s\tilde{s}} - \sum_{i<j\varepsilon s} c_{ij}$$

Method: Define $t_i = \dfrac{e_{HT} - Y\cdot}{\sqrt{V_{HT}(i)}}$, i = 1,2 --- 7

where, $Y\cdot$ = true population total

The populations used in the study are listed in Table 1.

Table 1

Popn. No.	Source	X	Y	Pop. Size
1	Cochran (1963)	1920 pop. area of farm	1930 pop. area under corn	20
2	Jessen (1978)	area of farm	area under corn	20
3	Scheaffer, et al. (1979)	real estate value 2 Yrs. ago	current value	20

The sampling scheme of Sampford (1967) was used to draw samples with inclusion probabilities proportional to x. For this design, all estimators, except $V_{HT}(1)$, are non-negative, as the relation $\pi_i\pi_j - \geq 0$ holds. Since $V_{HT}(1)$ can take negative values, $V_{HT}(1)$ was not included in the study. One thousand independent samples were drawn from each population, for each of the sample sizes, 3,5, and 7, the last one being restricted to the last two populations. For first population, sample size 7 was not used because the condition $\pi_i \alpha x_i$ would have forced the largest π_i to exceed unity for n = 7. For each sample and each variance estimator $V_{HT}(i)$, the quantities above were calculated.

From the resulting empirical distribution of each t_i, and for selected percentile, p, the lower (1000p)th percentile of the distribution of t_i is

$t_{i(p)}$ = (1000p)th observation in increasing order, and the

upper (1000p)th percentile is

$t_{i(1-p)}$ = (1000p)th observation in decreasing order

a 100(1-2p)% confidence interval for Y. based on the i-th variance estimator is

$$(e_{HT} - t_{i(p)} \sqrt{V_{HT}(i)} , e_{HT} + t_{i(-p)} \sqrt{V_{HT}(i)})$$

The width of the above confidence interval is

$$[t_{i(1-p)} - t_{i(p)}] \sqrt{V_{HT}(i)}$$

and hence the average width over 1000 samples can be computed by multiplying the average of $\sqrt{V_{HT}(i)}$ with $[t_{i(1-p)} - t_{i(p)}]$.

The percentiles of the distribution of the t_i's are shown in Table 2 and the average widths of the confidence intervals are shown in Table 3.

Table 2 Percentiles of the distribution of Studentized estimates with p = .05

Popn.	Sample Size	Empirical distributions					
		t2	t3	t4	t5	t6	t7
1	3	-6.23	-5.81	-7.82	-6.08	-6.21	-6.64
	5	-6.72	-5.37	-5.92	-5.27	-5.09	-5.31
2	3	-2.56	-2.39	-3.04	-2.32	-2.24	-2.35
	5	-2.19	-2.18	-2.17	-1.89	-1.86	-1.9
	7	-2.07	-1.96	-1.79	-1.63	-1.72	-1.58
3	3	-7.29	-7.2	-9.66	-7.2	-7.2	-7.2
	5	-2.15	-2.12	-2.5	-2.14	-2.17	-2.13
	7	-1.94	-1.94	-2.16	-1.96	-1.97	-1.96

Table 3 Average widths of confidence intervals based on different variance estimators (.95 Confidence Coefficients).

Popn	Sample Size	average width based on					
		$V_{HT}(2)$	$V_{HT}(3)$	$V_{HT}(4)$	$V_{HT}(5)$	$V_{HT}(6)$	$V_{HT}(7)$
1	3	2124	1820	1600	1654	1469	1688
	5	1396	1321	1014	1024	950.9	1048
2	3	1638	1636	1667	1659	1813	1643
	5	863.2	885.5	893.1	898	892.7	901.8
	7	533.6	543.3	549.1	552.7	540.8	567.6
3	3	97.63	97.61	97.62	97.66	98.17	97.61
	5	21.15	21.06	21.19	21.19	21.53	21.12
	7	14.15	14.10	14.02	14.0	14.16	13.94

Table 4 Coverage probabilities for 95% confidence interval based on Student's t

Popn. No.	Sample Size	Coverage probability based on					
		$V_{HT}(2)$	$V_{HT}(3)$	$V_{HT}(4)$	$V_{HT}(5)$	$V_{HT}(6)$	$V_{HT}(7)$
1	3	.86	.90	.81	.95	.95	.90
	5	.76	.79	.70	.76	.75	.76
2	3	.96	.96	.94	.96	.97	.96
	5	.95	.95	.94	.95	.97	.94
	7	.96	.96	.96	.97	.97	.97
3	3	.88	.88	.85	.87	.87	.87
	5	.95	.95	.93	.95	.95	.95
	7	.95	.95	.94	.96	.96	.96

Discussion: Based on this study, the following conclusions may be drawn. The distributions of t_2, t_3, t_5, t_6, and t_7 are very similar, and are somewhat closer to the t-distribution than those of t_4. This may be expected as the estimators $V_{HT}(2)$, $V_{HT}(3)$ and $V_{HT}(5)$ through $V_{HT}(7)$ are either design-unbiased or model-unbiased, whereas $V_{HT}(4)$ is a shrinkage estimator and inflate the value of t_4 by appearing in the denominator.

Although the relative efficiences of different variance estimators differ considerably, rather surprisingly, we find few major differences among the average widths of confidence intervals based on them. However, there are a few cases where the ratio estimator, $V_{HT}(6)$, gives considerably narrower confidence intervals than other estimators. However, these comparisons depend on the knowledge of the distribution of the t_i's which may not be available in practice. The construction of confidence intervals based on unequal probability sampling, when we do not know the distribution of the t_i's is an open problem.

In case Student's t is used for constructing confidence intervals, the coverage probabilities would not generally differ much for different variance estimators provided shrinkage is compensated for in case of shrinkage estimator. However, the coverage probabilities may differ considerable from the nominal confidence coefficient as shown in Table 4.

References:

1. Biyani, S. H. (1980). "On Inadmissability of the Yates Grundy Estimator in Unequal Probability Sampling", Journal of the American Statistical Association, 75, 709-712.

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

3. 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.

4. Midha, C. (1985). "An Empirical Study of Stability of Variance Estimators", Proceedings of the International Statistical Institute, 1, 251-252.

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