Jour. Ind. Soc. Ag. Statistics
51(2&3), 1998 : 343-348

The Horvitz-Thompson vs. Sen-Yates-Grundy Variance Estimators: Issues in Finite Population Sampling

Subir Ghosh
University of California, Riverside, USA

SUMMARY

There are two well-known estimators of the variance of the Horvitz-Thompson estimator for a population total. They are called Horvitz-Thompson (Horvitz-Thompson [3]) and Sen-Yates-Grundy (Yates and Grundy [7]; Sen [6]) estimators. This paper presents a striking example demonstrating a stunning difference in the numerical values of two estimates. The paper also discusses the comparison between these estimators.

Keywords: Probability sampling design, Horvitz-Thompson estimator, Sen-Yates-Grundy estimator, Inclusion probabilities, Hyper-admissibility criterion.

1. Introduction

Let τ be the unknown total of a characteristic of interest for a population of size N. Consider a probability sampling design d with π_i's and π_{ij}'s as the first-order and second-order inclusion probabilities, respectively. The Horvitz-Thompson estimator of the population total τ is

$$\hat{\tau}_{HT} = \sum_{i \in s} \frac{y_i}{\pi_i}, \qquad s \in S_d \tag{1}$$

where s denotes the sample of size n, S_d is the support of the design d and the y_i's are the values of the characteristic of interest. For the variance $V(\hat{\tau}_{HT})$ of $\hat{\tau}_{HT}$, the Horvitz-Thompson estimator is given by

$$\hat{V}(\hat{\tau}_{HT}) = \sum_{i \in s} \left(\frac{1-\pi_i}{\pi_i} \right) \frac{y_i^2}{\pi_i} + \sum_{\substack{i \\ i,j \in s \\ i \neq j}} \sum_{j} \left(\frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} \right) \frac{y_i y_j}{\pi_{ij}} \tag{2}$$

and the Sen-Yates-Grundy estimator is given by

$$\hat{V}_{SYG}(\hat{\tau}_{HT}) = \sum_{\substack{i \\ i,j \in s \\ i \neq j}} \sum_{j} \left(\frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} \right) \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2 \qquad (3)$$

2. *Issues*

A study is planned to find out the number of left-handed students in eight elementary schools of a town (Example 7.1, page 202, Hedayat and Sinha [2]). The complete list of the left-handed students in all the schools is not readily available but the prior information on the number of registered students in each school is available from the school district office. Three schools (1, 3,7) are selected in the sample using the probability sampling design given in Table 1. Information on the numbers of left-handed students are collected for the schools 1, 3 and 7.

Table 1. The probability sampling design

Sample (s)	Probability, P(s)
1, 2, 4	0.05
2, 3, 5	0.07
3, 4, 6	0.08
4, 5, 7	0.10
1, 5, 6	0.15
2, 6, 7	0.15
1, 3, 7	0.20
1, 2, 8	0.05
3, 4, 8	0.05
5, 6, 8	0.05
2, 7, 8	0.05

Table 2 presents the numbers of registered students in all schools and the first-order inclusion probabilities for schools under the probability sampling design given in Table 1. In fact, it can be seen that

$$\pi_j = 3 \frac{x_i}{\sum_{i=1}^{8} x_i} = \frac{x_i}{1{,}000} \qquad (4)$$

where x_i is the number of registered students for the i-th school. Note that

$$\sum_{i=1}^{8} x_i = 3{,}000$$

Table 2. The number of registered students and the first-order inclusion probabilities π_i's, i = 1, ..., 8

School number, i	Number of registered students, x_i	π_i
1	450	0.45
2	370	0.37
3	400	0.40
4	280	0.28
5	370	0.37
6	430	0.43
7	500	0.50
8	200	0.20

Table 3 presents the numbers of students as well as the numbers of left-handers for the schools selected in the sample.

Table 3. The number of left-handers and the numbers of students in schools 1, 3 and 7

Schools i	Number of students x_i	Number of left-handers y_i
1	450	10
3	400	4
7	500	2

The Horvitz-Thompson estimate of the total number of left-handed students in eight elementary schools is by using (1)

$$\hat{\tau}_{HT} = \frac{10}{0.45} + \frac{4}{0.40} + \frac{2}{0.50} = 36.22 \approx 36$$

Table 4 presents the second order inclusion probabilities π_{ij}'s, i,j= 1, 3, 7, i < j, under the probability sampling design given in Table 1.

Table 4. Second order inclusion probabilities, π_{ij}'s, i, j = 1, 3, 7, i < j

i	j	π_{ij}
1	3	0.20
1	7	0.20
3	7	0.20

The numerical values of variances, standard errors of $\hat{\tau}_{HT}$ by two methods and their ratios are given in Table 5.

Table 5. Variances, standard errors and their ratios for estimated total number of left-handed students

	HT	SYG	Ratio – HT/SYG
Estimated Variance	361.83	26.57	13.62
Standard Error	19.02	5.15	3.69

We observe that $SE_{HT}(\hat{\tau}_{HT})$ is 3.69 times larger than $SE_{SYG}(\hat{\tau}_{HT})$ and $\hat{V}_{HT}(\hat{\tau}_{HT})$ is 13.62 times larger than $\hat{V}_{SYG}(\hat{\tau}_{HT})$. Smaller the numerical value of $\hat{V}(\hat{\tau}_{HT})$ is better in terms of the closeness of $\hat{\tau}_{HT}$ to τ. The numerical values of $\hat{V}_{HT}(\hat{\tau}_{HT})$ and $\hat{V}_{SYG}(\hat{\tau}_{HT})$ are so different that it is impossible to make any sensible interpretation of our findings.

Consider now the information given in Table 6 on the numbers of boys and girls as well as the numbers of left-handed boys and girls are available for the selected schools 1, 3 and 7. The problem is now to estimate the total numbers of left-handed boys and girls in the eight schools.

Table 6. Number of boys and girls and numbers of left-handed boys and girls in selected schools 1, 3 and 7

	Number of Students		Number of Left-handers	
Schools	Boys	Girls	Boys	Girls
1	300	150	4	6
3	200	200	3	1
7	300	200	2	0

The Horvitz-Thompson estimate of the total number of left-handed boys and girls in eight schools are

$$\hat{\tau}^{B}_{HT} = \frac{4}{0.45} + \frac{3}{0.40} + \frac{2}{0.50} = 20.39 \approx 20$$

$$\hat{\tau}^{G}_{HT} = \frac{6}{0.45} + \frac{1}{0.40} + \frac{0}{0.50} = 15.83 \approx 16$$

respectively. In Table 7, we observe that $SE_{HT}\left(\hat{\tau}^{B}_{HT}\right)$ is 5.66 times of $SE_{SYG}\left(\hat{\tau}^{B}_{HT}\right)$ and $\hat{V}_{HT}\left(\hat{\tau}^{B}_{HT}\right)$ is 32.08 times of $\hat{V}_{SYG}\left(\hat{\tau}^{B}_{HT}\right)$. In Table 8, $SE_{HT}\left(\hat{\tau}^{G}_{HT}\right)$ is 3.21 times of $SE_{SYG}\left(\hat{\tau}^{G}_{HT}\right)$ and $\hat{V}_{HT}\left(\hat{\tau}^{G}_{HT}\right)$ is 10.32 times of $\hat{V}_{SYG}\left(\hat{\tau}^{G}_{HT}\right)$. Again, the discrepancies in the numerical values of $\hat{V}_{HT}$ and

$\hat{V}_{SYG}$, for both $\hat{\tau}^{B}_{HT}$ and $\hat{\tau}^{G}_{HT}$, make it hard for a meaningful interpretation of our findings.

Table 7. Variances, standard errors and their ratios for estimated total number of left-handed boys

	HT	SYG	Ratio = HT/SYG
Estimated Variance	89.65	2.79	32.08
Standard Error	9.47	1.67	5.66

Table 8. Variances, standard errors and their ratios for estimated total number of left-handed girls

	HT	SYG	Ratio = HT/SYG
Estimated Variance	108.19	10.49	10.32
Standard Error	10.40	3.24	3.21

3. HT vs SYG Estimators

Which one of $\hat{V}_{HT}(\hat{\tau}_{HT})$ and $\hat{V}_{SYG}(\hat{\tau}_{HT})$ is more reliable? It is known that both are equal for the simple random sampling without replacement design and stratified simple random sampling design (Remark 2.8.4, page 47, Särndal, Swensson and Wretman [5]). For general probability sampling designs, there is no definitive result on reliability. Rao and Singh [4] gave empirical evidence on overall superiority of $\hat{V}_{SYG}(\hat{\tau}_{HT})$ over $\hat{V}_{HT}(\hat{\tau}_{HT})$ for the Brewer's probability sampling design with the sample size two. Considering five artificial populations and 34 natural populations that are known, Rao and Singh [4] observed that the gains in efficiency of $\hat{V}_{SYG}(\hat{\tau}_{HT})$ over $\hat{V}_{HT}(\hat{\tau}_{HT})$ are enormous for several of the populations. For the populations with $\hat{V}_{SYG}(\hat{\tau}_{HT})$ less efficient, the losses in efficiency are small. There is indeed another criterion of non-negative numerical values of $\hat{V}_{HT}(\hat{\tau}_{HT})$ and $\hat{V}_{SYG}(\hat{\tau}_{HT})$. The overall performance of $\hat{V}_{SYG}(\hat{\tau}_{HT})$ is much better than $\hat{V}_{HT}(\hat{\tau}_{HT})$ under the non-negativity criterion. Rao and Singh [4] proved that $\hat{V}_{HT}(\hat{\tau}_{HT})$ is the unique "hyper-admissible" estimator in a wide class of unbiased estimators of $V(\hat{\tau}_{HT})$. But this strength of $\hat{V}_{HT}(\hat{\tau}_{HT})$ has been interpreted as the evidence on the weakness of the "hyer-admissibility" criterion.

4. *Discussions*

In the example given in Section 2, the numerical values of $\hat{V}_{HT}(\hat{\tau}_{HT})$ are strikingly different from the numerical values of $\hat{V}_{SYG}(\hat{\tau}_{HT})$. Considering the fact that $\hat{V}_{HT}(\hat{\tau}_{HT})$ and $\hat{V}_{SYG}(\hat{\tau}_{HT})$ are estimators of the same quantity $V(\hat{\tau}_{HT})$ and yet we just cannot discard one over the other, we face an embarrassing reality of the statistical world. We have not gone to an extreme like Professor D. Basu (Basu [1]) for creating the embarrassment of the circus statistician in his famous elephant example.

REFERENCES

[1] Basu, D., 1971. An essay on the logical foundations of survey sampling, Part one. In: V.P. Godambe and D.A. Sprott (eds.) *Foundations of Statistical Inferences*. Holt, Rinehart and Winston, Toronto, 203-242.

[2] Hedayat, A.S. and Sinha, B.K., 1991. *Design and Inference in Finite Population Sampling*. John Wiley and Sons, New York.

[3] 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.

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

[5] Särndal, C.E., Swensson, B. and Wretman, J. 1992. *Model Assisted Survey Sampling*. Springer-Verlag, New York.

[6] Sen, A.R., 1953. On the estimate of the variance in sampling with varying probabilities. *J. Ind. Soc. Agril. Statist.*, **5**, 119-127.

[7] Yates, F. and Grundy, P.M., 1953. Selection without replacement from within strata with probability proportional to size. *J. Roy. Statist. Soc.*, **B15**, 253-261.