Journal of Reliability and Statistical Studies; ISSN (Print): 0974-8024, (Online):2229-5666
Vol. 7, Issue 1 (2014): 51-60

EMPIRICAL COMPARISON OF VARIOUS APPROXIMATE ESTIMATORS OF THE VARIANCE OF HORVITZ THOMPSON ESTIMATOR UNDER SPLIT METHOD OF SAMPLING

Neeraj Tiwari[1] and Akhil Chilwal[2]
[1,2]Department of Statistics, Kumaun University, S.S.J. Campus,
Almora-263601,Uttarakhand (INDIA)
E Mail: [1]kumarn_amo@yahoo.com, [2]akhil.stat@gmail.com

Received June 21, 2013
Modified May 16, 2014
Accepted May 31, 2014

Abstract

Under inclusion probability proportional to size (IPPS) sampling, the exact second-order inclusion probabilities are often very difficult to obtain, and hence variance of the Horvitz-Thompson estimator and Sen-Yates-Grundy estimate of the variance of Horvitz-Thompson estimator are difficult to compute. Hence the researchers developed some alternative variance estimators based on approximations of the second-order inclusion probabilities in terms of the first order inclusion probabilities. We have numerically compared the performance of the various alternative approximate variance estimators using the split method of sample selection

Key Words: Variance Estimation, Relative Bias, Relative Mean Square Error, Efficiency, Split Method of Sample Selection.

1. Introduction

Unequal probability sampling with inclusion probability, exactly proportional to a measure of size x_i, known for each unit (often called π ps) is extensively used in large-scale surveys. For simplicity, we focus on single stage sampling from a finite population U of size N. The Horvitz-Thompson (1952) (HT) estimator $\hat{Y}_{HT} = \sum_{i \in s} y_i / \pi_i$, with variance $V(\hat{Y}_{HT})$, is used to estimate the population total $Y = \sum_{i \in U} y_i$ of a characteristic of interest y, which is approximately proportional to x, where *s* denotes a sample of fixed size *n* and $\pi_i = nx_i / X$ with $X = \sum_{i \in U} x_i$. Where π_i denotes the first order inclusion probability of unit *i* in the sample. The well known Sen-Yates-Grundy (1953) (SYG) variance estimator

$$v_{SYG}(\hat{Y}_{HT}) = v_{SYG} = \sum_{i \in s} \sum_{j < i \in s} \frac{(\pi_i \pi_j - \pi_{ij})}{\pi_{ij}} \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2 \quad (1)$$

is exactly unbiased. Where π_{ij} denotes the second order inclusion probability for the pair (i, j). The SYG variance estimator is generally preferable to the Horvitz and

Thompson (HT) variance estimator, because the SYG variance estimator is always non-negative, when $\pi_{ij} \leq \pi_i \pi_j, i < j$, whereas the HT variance estimator can take

negative values even when this condition is satisfied. This variance estimator also suffers from two another draw backs. First, it involves the second order inclusion probability π_{ij}, which may not be easy to obtain for some sampling designs. Second, it can be very unstable because of the term $1/\pi_{ij}$ in Eq. (1). This led researchers to develop alternative variance estimators based on some approximations of π_{ij} in terms of π_i. The concept of approximating the joint inclusion probabilities in terms of first order inclusion probabilities only, was introduced by Hartley and Rao (1962) under the randomized systematic IPPS sampling design. Using Conditional Poisson Sampling, Hajek (1964) discussed an approximation of the second order inclusion probabilities in term of first order inclusion probabilities and provided an appropriate variance estimator. Hajek (1964) approximation to π_{ij} works well under a high entropy sampling design. A set of high-entropy variance estimators was presented by Brewer (2002) and by Brewer and Donadio (2003). All these estimators have an important advantage over the standard Sen-Yates-Grundy (SYG) variance estimator that these expressions do not involve the second order inclusion probabilities.

Some empirical studies have been reported in the literature on the performance of approximate variance estimators. First of all Brewer and Donadio (2003) considered a subset of the ten approximate variance estimators and present an empirical study to investigate the performance of the various approximate variance estimators. Matei and Tille (2005) considered a set of eighteen approximate variance estimators. They performed an empirical study under the Conditional Poisson Sampling (CPS) design. Henderson (2006) considered a set of twelve approximate variance estimators and performed a study using the CPS design and randomized IPPS systematic sampling. The goal of our study is to enlarge the scope of previous empirical studies by considering a large set of real populations. In the present article we investigate the performance of the six approximate variance estimators in terms of relative bias and relative mean square error. The performance of various approximate variance estimators is also compare in term of precision. For empirical study we use the Deville and Tille's (1998) split method of sample selection, which is simpler practical choice with respect to the various πps sampling plans. Split method has an advantage over the other sampling methods that it satisfied the Sen-Yates-Grundy condition that if $\pi_{ij} \leq \pi_i \pi_j \ (i \neq j \in U)$, then the variance estimator always takes a positive value. The split method of sample selection also satisfied the Gabler (1984) sufficient condition. In practical life the implementation of Split method is quite easy.

2. Split Method of Sample Selection

To conduct an empirical comparison between the six approximate variance estimators we use split method of sample selection. Deville and Tille (1998) proposed the 'Split Method' of sample selection for unequal probability sampling without replacement. In this method inclusion probabilities are considered as an inclusion probability vector. This inclusion probability vector is split into several new inclusion probability vectors. Out of these vectors one vector is selected randomly; thus, the initial problem is reduced to another sampling problem with unequal probabilities. The splitting of inclusion probability vector is then repeated on these new vectors. At each step, the sampling problem is reduced to a new simpler problem. The basic technique of split method is extremely simple and it is described as follows:

Consider a finite population U of size N, $U = \{1, 2, \dots, l, \dots, N\}$. For each unit of the population consider that the value of y_l of characteristic y can be measured. Suppose that the values of $x_l > 0$ of an auxiliary characteristic x are known for all the units of U and x_l is approximately proportional to y_l. The first order inclusion probabilities are computed using the relation

$$\pi_l = \frac{nx_l}{\sum_{l \in S} x_l}$$

for all $l \in U$, where n is the sample size.

Each π_l is split into two parts $\pi_l^{(1)}$ and $\pi_l^{(2)}$ that satisfy the following conditions

(*i*) $\pi_l = \lambda\pi_l^{(1)} + (1-\lambda)\pi_l^{(2)}$ (*ii*) $0 \le \pi_l^{(1)} \le 1$

(*iii*) $0 \le \pi_l^{(2)} \le 1$ (*iv*) $\sum_{l \in S} \pi_l^{(1)} = \sum_{l \in S} \pi_l^{(2)} = n$

here λ can be chosen freely provided that $0 < \lambda < 1$. The method consists of drawing n units with unequal probabilities

$$\begin{cases} \pi_l^{(1)}, l \in U & \text{with a probability } \lambda \\ \pi_l^{(2)}, l \in U & \text{with a probability } (1-\lambda) \end{cases}$$

Now, the problem is reduced to another sampling problem with unequal probabilities. If the splitting is such that one or several of the $\pi_l^{(1)}$ and $\pi_l^{(2)}$ are equal to 0 or 1, the sampling problem will be simpler at the next step because the splitting is then applied to a smaller population. The splitting is repeated on the $\pi_l^{(1)}$ and $\pi_l^{(2)}$ until all the possible samples are obtained from the population.

3. Approximate Variance Estimators

In this section, we present a set of six approximate variance estimators, which we can write using a common form originally introduced by Haziza, Mecatti and Rao (2008). For fixed sample size $n \ge 2$ and increasing population size $N \to \infty$, Hartley and Rao (1962) proposed first and second order approximations for joint inclusion probabilities under the randomized systematic IPPS sampling design. It may be noted that the exact evaluation of π_{ij}'s for this design (Hidiroglou and Gray, 1980) is cumbersome as sample size n increases, unlike for the Rao-Sampford design. Under the Conditional Poisson Sampling (CPS) design, Hajek (1964) proposed an approximation to π_{ij} by assuming that $\sum_{i\in U}\pi_i(1-\pi_i)\rightarrow\infty$. Hajek's approximation to π_{ij} works well under a high entropy sampling design. Hajek's approximate variance estimator is denoted by v_H in Table 1. Rosen (1991) proposed an alternative approximate variance estimator in the context of Pareto Sampling, which is denoted by v_R in the Table 1. Finally, we considered a family of approximate variance estimators developed by Brewer and Donadio (2003), there estimators are denoted by $v_{B1}, v_{B2}, v_{B3}, v_{B4}$ in Table 1.

All the approximate variance estimators considered in this article can be expressed in the following common form

$$v(\hat{Y}_{HT}) = \sum_{i\in s} t_i \left\{ \frac{y_i}{\pi_i} - \frac{\sum_{i\in s} r_i (y_i/\pi_i)}{\sum_{i\in s} r_i} \right\}^2 \qquad (2)$$

t_i and r_i are constants given in Table 1.

Variance Estimator	Symbol	Coefficient t_i	Coefficient r_i
Hajek	v_H	$\frac{n}{n-1}(1-\pi_i)$	$\frac{n}{n-1}(1-\pi_i)$
Rosen	v_R	$\frac{n}{n-1}(1-\pi_i)$	$\frac{(1-\pi_i)\log(1-\pi_i)}{\pi_i}$
Brewer 1	v_{B1}	$\frac{n}{n-1}(1-\pi_i)$	1
Brewer 2	v_{B2}	$\frac{n}{n-1}\left\{1-\pi_i+\frac{\pi_i}{n}-\frac{1}{n^2}\sum_{l\in U}\pi_l^2\right\}$	1
Brewer 3	v_{B3}	$\frac{n}{n-1}\left\{1-\pi_i-\frac{\pi_i}{n}-\frac{1}{n^2}\sum_{l\in U}\pi_l^2\right\}$	1
Brewer 4	v_{B4}	$\frac{n}{n-1}\left\{1-\pi_i-\frac{\pi_i}{n-1}+\frac{1}{n(n-1)}\sum_{l\in U}\pi_l^2\right\}$	1

Table 1: Coefficient t_i and r_i for the Approximate Variance Estimators

The approximate variance estimator (2) has the following desirable properties:

a) It involves only the first order inclusion probability π_i.
b) It is always positive.
c) It involves a single sum unlike the HT or the SYG variance estimators.

d) It is equal to zero when y is proportional to x; that is, $v(\hat{Y}_{HT}) = 0$ when $y_i = bx_i,\ i = 1,\ldots,N$ where *b* is an arbitrary constant.

e) $v(\hat{Y}_{HT})$ reduces to $v(\hat{Y}_{HT}) = N^2\left(1-\dfrac{n}{N}\right)\dfrac{s_y^2}{n}$ when $\pi_i = \dfrac{n}{N}$, which is the usual expression of the estimated variance of $\hat{Y}_{HT}$ in the special case of simple random sampling without replacement.

4. Empirical Study

We conduct an empirical study, under the Deville and Tille's (1998) design, to investigate the performance of the approximate variance estimators presented in Section 3 and the exact SYG variance estimator Eq. (1).

For empirical study we consider ten real populations given in Table 2. For each population we compute relative bias, relative mean square error and precision. On the basis of these parameters we compare all the approximate variance estimators. The findings of the empirical study presents in Table 3, Table 4 and Table 5. For these populations it will be possible to study the effect of the sampling design on the properties of the variance estimators.

Pop.	Source	N	*y*	*x*
1	Mukhopadhyay [1998, p.157]	06	Output	Fixed capital
2	Mukhopadhyay [1998, p.96]	06	No. of labourers	Quality of raw materials
3	Mukhopadhyay [1998, p.110]	06	Output	Fixed capital
4	Mukhopadhyay [1998, p.131 (7-12)]	06	Yield of paddy	Area
5	Mukhopadhyay [1998, p.131 (1-6)]	06	Yield of paddy	Area
6	Sukhatme & Sukhatme [1970, p.166 (11-20)]	08	No. of banana bunches	No. of banana pits
7	Mukhopadhyay [1998, p.192]	08	Value added	No. of workers
8	Mukhopadhyay [1998, p.110]	08	Output	No. of workers
9	Cochran [1982, p.152]	10	Large United States Cities in 1930	Large United States Cities in 1920
10	Sukhatme & Sukhatme [1970, p.185 (11-20)]	10	Area under wheat in 1937	Area under wheat in 1936

Table 2: Characteristics of the Real Populations

Let S denote the set of all possible samples of size n=2 from the population U. We select the samples using the Deville & Tille's split method. For each sample, we calculate the value of SYG estimator Eq. (1) and six approximate variance estimators, given in Table 1. We compare all the approximate variance estimators, generally denoted by v on the basis of their Relative Bias (RB), Relative Mean Square Error (RMSE) and Precision (R_p). These measures are given as

$$\mathrm{Re}l\, Bias\{v\} = \frac{E(v) - V(\hat{Y}_{HT})}{V(\hat{Y}_{HT})} \tag{4}$$

$$\mathrm{Re}l\, MSE\{v\} = \frac{E[\{v - V(\hat{Y}_{HT})\}^2]}{\{V(\hat{Y}_{HT})\}^2} \tag{5}$$

$$R_P = \frac{1/v}{1/v_{SYG}} \tag{6}$$

The ratio R_P represents a loss in accuracy by using Eq. (1) instead of Eq. (2). When R_P is less than 1, Eq. (2) is better than Eq. (1).

Pop.	v_H	v_R	v_{B1}	v_{B2}	v_{B3}	v_{B4}
Pop. 1	-0.7233	-0.7229	-0.7218	-0.7479	-0.6956	-0.6695
Pop. 2	-0.7567	-0.7570	-0.7566	-0.7588	-0.7521	-0.7521
Pop. 3	-0.5213	-0.5153	-0.4851	-0.5074	-0.4629	-0.4406
Pop. 4	0.5643	0.5643	0.5643	0.5667	0.5618	0.5594
Pop. 5	-0.5099	-0.5099	-0.5099	-0.5049	-0.5148	-0.5198
Pop. 6	-0.5975	-0.5974	-0.5971	-0.6027	-0.5914	-0.5857
Pop. 7	0.3064	0.3069	0.3086	0.2409	0.3763	0.4439
Pop. 8	-0.7759	-0.7755	-0.7739	-0.7900	-0.7578	-0.7417
Pop. 9	-0.4128	-0.4127	-0.4123	-0.5134	-0.3112	-0.2100
Pop. 10	0.0757	0.1010	0.2138	0.1018	0.1459	0.0755

Table 3: Relative Bias

Population	v_H	v_R	v_{B1}	v_{B2}	v_{B3}	v_{B4}
Pop. 1	1.5297	1.5394	1.5736	1.2904	1.9114	2.3038
Pop. 2	0.5263	0.5263	0.5264	0.5263	0.5264	0.5265

Pop. 3	0.5563	0.5469	0.5008	0.3892	0.6506	0.8388
Pop. 4	1.4496	1.4496	1.4497	1.4427	1.4437	2.8757
Pop. 5	0.7521	0.7521	0.7521	0.7503	0.7505	0.7489
Pop. 6	1.4519	1.4520	1.4521	1.3996	1.5064	1.5625
Pop. 7	3.8326	3.8359	3.8466	3.1429	9.2704	5.5089
Pop. 8	0.7014	0.7008	0.6986	0.6983	0.7678	0.7220
Pop. 9	5.1468	5.1468	5.1453	3.5624	7.0839	9.3783
Pop. 10	0.7305	0.7677	0.9775	0.7279	0.8558	0.8079

Table 4: Relative Mean Square Error

Population	v_H	v_R	v_{B1}	v_{B2}	v_{B3}	v_{B4}
Pop. 1	0.7268	0.7247	0.7175	0.8088	0.6447	0.5854
Pop. 2	0.8223	0.8220	0.8209	0.8236	0.8183	0.8156
Pop. 3	0.7224	0.7163	0.6872	0.7683	0.6215	0.5673
Pop. 4	0.8004	0.8004	0.8004	0.8989	0.8018	0.8033
Pop. 5	0.9417	0.9417	0.9414	0.9524	0.9507	0.9601
Pop. 6	0.8242	0.8241	0.8238	0.8398	0.8083	0.7934
Pop. 7	0.8078	0.8075	0.8065	0.8779	0.7458	0.6936
Pop. 8	0.7980	0.7969	0.7933	0.8759	0.7248	0.6673
Pop. 9	0.5758	0.5756	0.5749	0.6974	0.4890	0.4255
Pop. 10	0.8788	0.8695	0.8365	0.8866	0.7962	0.7652

Table 5: Values of Precision (R_P) for the Approximate Variance Estimators

Table 3 shows descriptive statistics regarding the relative bias (RB) of six approximate variance estimators for all the considered real populations. The approximate variance estimators v_{B2} perform very well in terms of relative bias. Table 4 clearly indicates that the approximate variance estimator v_{B2} given by Brewer and Donadio (2003) perform better than all the other approximate variance estimators in terms of relative mean square error. The empirical study also indicates that the Hajek (v_H) estimator performs well in comparison to other approximate variance estimators. Table 5 shows the precision of various approximate variance estimators. In terms of precision, we find that the v_{B2} variance estimator performs quite close to the SYG variance estimator.

5. Conclusion

In this article we consider a set of six approximate variance estimators and the exact SYG variance estimator under the Deville and Tille's split method of sampling. The approximate variance estimators have an advantage over the SYG variance estimator that these estimators do not involve the second order inclusion probabilities. The implementation of such estimators in the practical life is very easy. On the basis of empirical study we conclude that the approximate variance estimators perform as well as the SYG variance estimator. With the help of empirical study, we show that the approximate variance estimators perform relatively well in terms of relative bias and relative mean square error. The comparison in terms of precision indicates that the approximate variance estimators perform quite close to SYG variance estimator.

References

1. Asok, C. and Sukhatme, B.V. (1976). On Sampford's procedure of unequal probability sampling without replacement, J. Amer. Stat. Assoc., 71, p. 912-918.
2. Brewer, K.R.W. (2002). Combined Survey Sampling Inference, Weighing Basu's Elephants, Arnold Publisher.
3. Brewer, K.R.W. and Donadio, M.E. (2003). The high entropy variance of the Horvitz-Thompson estimator, Survey Methodology, 29, p. 189-196.
4. Cochran, W.G. (1982). Sampling Techniques, 3rd Ed., John Wiley & Sons.
5. Deville, J.C. and Tille, Y. (1998). Unequal probability sampling without replacement through a splitting method, Biometrika, 85(1), p. 89-101.
6. Gabler, S. (1984). On unequal probability sampling: Sufficient conditions for the superiority of sampling without replacement, Biometrika, 71, p. 171-175.
7. Hajek, J. (1964). Asymptotic theory of rejective sampling with varying probabilities from a finite population, Annals of Mathematical Statistics, 35, p. 1491-1523.
8. Hartley, H.O. and Rao, J.N.K. (1962). Sampling with unequal probabilities and without replacement, The Annals of Mathematical Statistics, 33, p. 350-374.
9. Haziza, D., Mecatti, F., and Rao, J.N.K. (2008). Evaluation of some approximate variance estimators under the Rao-Sampford unequal probability sampling design, Int. J. Statist. Vol. LXVI, No.1, 91-108.
10. Henderson, T. (2006). Estimating the variance of Horvitz-Thompson estimator, Bachelor's Thesis, Australian National University.
11. Hidiroglou, M.A. and Gray, G. B. (1980). Construction of joint probability of selection of for systematic pps sampling, Applied Statistics, 29, p. 107-112.
12. Horvitz, D.G. and Thompson, D.J. (1952). A generalization of sampling without replacement from finite universe, J. Amer. Statist. Assoc., 47, p. 663-685.

13. Matei, A and Tille, Y. (2005). Evaluation of variance approximations estimators in maximum entropy sampling with unequal probability and fixed sample size, Journal of Official Statistics, 21, No.4, p. 543-570.
14. Mukhopadhyay, P. (1998). Theory and Methods of Survey Sampling. Prentice-Hall of India, New Delhi.
15. Rosen, B. (1991). Variance estimation for systematic pps-sampling, Report 1991:15, Statistics Sweden.
16. Sen, A.R. (1953). On the estimation of variance in sampling with varying probabilities, J. Ind. Soc. Agri. Statist., 5, p. 119-127.
17. Sukhatme, P.V. and Sukhatme, B.V. (1970). Sampling Theory of Surveys with Applications, Asia Publishing House, Calcutta.
18. Yates, F. and Grundy, P.M. (1953). Selection without replacement from within strata with probability proportional to size, J. Roy. Statist. Soc., B15, p.253-261.

Appendix

Example 1: Theory and Methods of Survey Sampling, P. Mukhopadhyay (1998), p. 157(1-6)

Output (*y*)	1451	2800	3890	3520	4700	5712
Fixed Capital (*x*)	112	208	367	450	620	780

Example 2: Theory and Methods of Survey Sampling, P. Mukhopadhyay (1998), p. 96 (1-6)

No. of labourers (*y*)	38	40	41	38	29	31
Quality of raw materials (*x*)	376	387	429	472	503	512

Example 3: Theory and Methods of Survey Sampling, P. Mukhopadhyay (1998), p. 110(1-6)

Output (*y*)	2552	3975	3607	3975	5712	6903
Fixed capital (*x*)	219	352	475	619	775	1412

Example 4: Theory and Methods of Survey Sampling, P. Mukhopadhyay (1998), p. 131(7-12)

Yield of paddy (*y*)	9565	9598	10316	8963	9562	10512
Area (*x*)	995	1031	1043	1054	1078	1089

Example 5: Theory and Methods of Survey Sampling, P. Mukhopadhyay (1998), p. 131(1-6)

Yield of paddy (*y*)	8521	8554	8783	8863	7025	8887
Area (*x*)	870	883	894	901	914	973

Example 6: Sampling Theory of Surveys with Application, P.V. Sukhatme and B.V. Sukhatme, 1970, p.166 (11-20)

No. of banana bunches (*y*)	567	580	867	923	954	952	1051	1138
No. of banana pits (*x*)	460	540	578	608	630	635	688	815

Example 7: Theory and Methods of Survey Sampling, P. Mukhopadhyay (1998), p. 192

Value added (*y*)	3607	3975	5712	6903	6973	7075	7545	8975
No. of workers (*x*)	475	619	775	1412	1675	1935	2515	3512

Example 8: Theory and Methods of Survey Sampling, P. Mukhopadhyay (1998), p. 110

Output (*y*)	31.3	11.2	38.4	21.9	32.2	36.5	15.7	61.7
No. of workers (*x*)	22	43	52	65	67	75	103	116

Example 9: Sampling Techniques, W.G. Cochran, (1982), p.152

Large United States Cities in 1930 (*y*)	48	50	63	69	67	80	115	143	464	459
Large United States Cities in 1920 (*x*)	23	29	37	61	67	76	120	138	381	387

Example 10: Sampling Theory of Surveys with Application, P.V. Sukhatme and B.V. Sukhatme, 1970, p.185 (11-20)

Area under wheat in 1937 (*y*)	79	60	62	103	100	179	141	219	265	330
Area under wheat in 1936 (*x*)	62	71	73	129	137	192	196	236	255	663