ISSN 1684 – 8403
Journal of Statistics
Vol: 12, No.1 (2005)

An Empirical Study of some Unequal Probability Sampling Estimators

Usman Ali Khan*

Muhammad Qaiser Shahbaz**

Abstract

An empirical study has been carried out to decide about the performance of various estimators used in unequal probability sampling without replacement and a sample of size 2. The Hansen–Hurwitz estimator and simple random sampling method has also been compared in this study. Some suggestions have been given at the end.

Key Words: Unequal probability sampling without replacement.

1. Introduction

The unequal probability sampling has its emergence in early forties, when Hansen and Hurwitz (1943) first introduced the concept. The sampling design proposed by them was used with replacement sampling only. The estimator proposed by Hansen and Hurwitz to estimate the population total is given as:

$$y'_{HH} = \frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{p_i} \qquad (1.1)$$

where p_i is probability of selection of i-th unit. The sampling variance of Hansen – Hurwitz estimator has different forms given as:

* AC-Nilson Aftab Associates, Lahore.
** Department of Statistics, Government College University, Lahore.

$$Var\ (y'_{HH}) = \frac{1}{2n}\sum_{\substack{i=1\\ j\neq i}}^{N}\sum_{j=1} P_i P_j \left(\frac{Y_i}{P_i} - \frac{Y_j}{P_j}\right)^2 \quad (1.2)$$

$$= \frac{1}{n}\sum_{i=1}^{N}\frac{1}{P_i}\left(Y_i - P_i Y_i\right)^2 \quad (1.3)$$

The concept of unequal probability sampling without replacement was first introduced by Madow (1949) but no theoretical framework was given. Horvitz and Thompson (1952) gave the first theoretical framework of unequal probability sampling. They also proposed their selection procedure and an estimator of population total. The estimator proposed by Horvitz and Thompson is given as:

$$y'_{HT} = \sum_{i\in s}\frac{Y_i}{\pi_i}\ , \quad (1.4)$$

where π_i is probability of inclusion of i-th unit in the sample.

Horvitz and Thompson gave following variance formula for estimator (1.4).

$$V(y'_{HT}) = \sum_{i=1}^{N}\frac{(1-\pi_i)}{\pi_i}Y_i^2 + \sum_{\substack{i,j=1\\ j\neq i}}^{N}\sum \frac{(\pi_{ij} - \pi_i\pi_j)}{\pi_i\pi_j}Y_i Y_j \quad (1.5)$$

an alternative expression, for fixed n, given by Sen (1953) and independently by Yates and Grundy (1953), is:

$$V(y'_{HT}) = \sum_{\substack{i=1\\ j>i}}^{N}\sum_{j=1}\left(\pi_i\pi_j - \pi_{ij}\right)\left(\frac{Y_i}{\pi_i} - \frac{Y_j}{\pi_j}\right)^2 \quad (1.6)$$

Since the emergence of Horvitz and Thompson estimator a number of selection procedures have been developed that can be used with this estimator. Raj (1956a) introduced his estimator based on the order of selection of units. The estimator proposed by Raj (1956a) is given as:

$$t_{mean} = \frac{1}{n}\sum_{r=1}^{n} t_r \qquad (1.7)$$

where

$$t_r = \sum_{i=1}^{r-1} y_i + \frac{y_r}{p_r}\left(1 - \sum_{i=1}^{r-1} p_i\right) \qquad (1.8)$$

The Raj (1956a) estimator for a sample of size 2 is given as:

$$t_{mean} = \frac{1}{2}\left[\frac{y_i}{p_i}\left(1 + p_i\right) + \frac{y_i}{p_i}\left(1 - p_j\right)\right] \qquad (1.9)$$

the sampling variance of estimator given in (1.9) is given as:

$$Var\left(t_{mean}\right) = \frac{1}{8}\sum_{i=1}^{N}\sum_{\substack{j=1 \\ j \neq i}} P_i P_j \left(2 - P_i - P_j\right)\left(\frac{Y_i}{P_i} - \frac{Y_j}{P_j}\right)^2 \qquad (1.10)$$

Raj's estimator has defect that it is based on order of the units in which they are selected. Murthy (1957) uses the idea of sufficiency to overcome the defect of Raj estimator. He symmetrized the Raj estimator to produce an un-ordered estimator. The estimator proposed by Murthy has general form:

$$t_{symm} = \frac{1}{P(s)}\sum_{i=1}^{n} P\left(s \mid i\right) y_i \qquad (1.11)$$

where $P\left(s \mid i\right)$ is the probability of obtaining a sample "s" given that ith unit has been already selected and $P(s)$ is the probability of obtaining a sample "s"

The Murthy (1957) estimator for a sample of size 2 is given as:

$$t_{symm} = \frac{1}{2 - p_i - p_j}\left[\frac{y_i}{p_i}\left(1 - p_j\right) + \frac{y_j}{p_j}\left(1 - p_i\right)\right] \qquad (1.12)$$

The sampling variance for estimator given in (1.12) is given as:

$$Var\left(t_{symm}\right)=\frac{1}{2}\sum_{i=1}^{N}\sum_{\substack{j=1\\ j\neq i}}\frac{P_i\,P_j\left(1-P_i-P_j\right)}{2-P_i-P_j}\cdot\left(\frac{Y_i}{P_i}-\frac{Y_j}{P_j}\right)^2 \qquad (1.13)$$

The estimators given so far, for unequal probability sampling without replacement, are very hard to apply for a sample of size more than 2. To overcome this defect Rao – Hartley and Cochran (1962) proposed an estimator that can be used with a sample of any size. The estimator proposed by them is given as:

$$y'_{RHC}=\sum_{i=1}^{n}\frac{\pi_i\,y_{iT}}{p_{iT}} \qquad (1.14)$$

where p_{iT} is the probability of T-th unit selected from the i-th group. Also $\pi_i=\sum_{T=1}^{N_i}p_{iT}$ *and* $\sum_{i=1}^{n}\pi_i=1.$ The sampling variance of estimator given in (1.14) is:

$$Var\left(y'_{RHC}\right)=\frac{n\left(\sum_{i=1}^{n}N_i^2-N\right)}{N\left(N-1\right)}\cdot\left[\sum_{i=1}^{n}\sum_{T=1}^{N_i}\frac{Y_{iT}^2}{n\,P_{iT}}-\frac{Y^2}{n}\right] \qquad (1.15)$$

2. The Empirical Study

In this section the empirical study has been given in order to decide about the performance of various estimators in unequal probability sampling without replacement. To carry out the study fifty natural populations have been used, which are given in standard texts on sampling techniques. The sampling variance of estimators given in section 1 has been obtained for all the populations. After evaluating the sampling variance, ranking has been done for each estimator according to the sampling variance. The average rank of each estimator has been calculated for various ranges of ranks of coefficient of variation of measure of size and correlation coefficient between actual variable of study and the measure of size. The average ranks have also been calculated for various actual ranges of the coefficient of variation and correlation coefficient. It should be noticed that an estimator with smaller average rank will have better performance as compared to some other estimator with a larger average rank. The results of the empirical study have been given in following tables.

Table 1: Average Ranks of Various Estimators with ranks of Coefficient of Variation.

CV (Z)	SRS	HH	HT (YG)	HT Brewer	RHC	Raj	Murthy
1 – 10	5.8	6.1	2.6	3.3	2.8	4.7	2.4
11 – 20	5.2	6.3	3.3	3.1	3.6	4.1	2.4
21 – 30	7.0	6.0	2.2	3.8	3.4	3.5	2.1
31 – 40	5.2	6.2	2.3	4.1	4.4	3.5	2.3
41 – 50	5.8	6.0	2.6	3.5	4.5	3.5	2.1

Table 2: Average Ranks of Various Estimators with ranks of Correlation Coefficient.

ρ_{YZ}	SRS	HH	HT (YG)	HT Brewer	RHC	Raj	Murthy
1 – 10	4.0	6.5	2.0	3.7	3.9	4.9	2.9
11 – 20	5.8	6.2	2.0	4.2	3.6	3.9	2.3
21 – 30	6.4	6.1	3.2	2.7	4.1	3.4	2.1
31 – 40	5.8	6.2	2.0	3.7	3.8	3.9	2.5
41 – 50	7.0	5.6	3.8	3.5	3.3	3.2	1.5

Table 3: Average Ranks of Various Estimators with various ranges of Correlation Coefficient.

CV (Z)	SRS	HH	HT (YG)	HT (Brewer)	RHC	Raj	Murthy
Less than 0.5	5.89	6.15	2.74	3.41	3.19	4.19	2.33
0.5<CV<1.0	5.67	6.11	2.28	3.78	4.44	3.44	2.28
Greater than 1	5.80	6.00	3.00	3.60	4.20	3.60	1.80

Table 4: Average Ranks of Various Estimators with various ranges of Coefficient of Variation.

ρ_{YZ}	**SRS**	**HH**	**HT (YG)**	**HT (Brewer)**	**RHC**	**Raj**	**Murthy**
$\rho_{YZ} < 0.5$	2.71	6.71	2.00	4.14	4.14	5.00	3.14
$0.5 < \rho_{YZ} < 0.9$	6.14	6.14	2.67	3.38	3.67	3.81	2.19
$0.9 < \rho_{YZ} < 1.0$	6.45	5.91	2.73	3.55	3.68	3.55	2.05

3. Conclusions

The empirical study of various estimators has been given in section 2. The results of this study have been given in Table–1 through Table–4.

Table–1 contains the average ranks of various estimators along with the group ranks of coefficient of variation of measure of size. From this table we can see that the Murthy estimator clearly outperform other estimators and is closely followed by the Horvitz – Thompson estimator under the Yates–Grundy draw-by-draw procedure for all ranges of coefficient of variation. Table–2 contains the average ranks along with the group ranks of correlation coefficient. From this table we can see that for smaller range of correlation coefficient the Horvitz–Thompson estimator under Yates–Grundy draw-by-draw procedure performs reasonably well as compared to other estimators. For other ranges the Murthy estimator again outperforms other estimators. Table–3 contains the average ranks of various estimators along with the actual values of coefficient of variation. The value of coefficient of variation has been divided in three ranges, that is, less than 0.5, between 0.5 and 1.0 and greater than 1.0. From this table we can again see that Murthy estimator outperform all other estimators and is again closely followed by the Horvitz–Thompson estimator under the Yates–Grundy draw-by-draw procedure. Table – 4 contains the average ranks of various estimators along with the actual values of correlation coefficient. Again the value of correlation coefficient has been divided in three ranges, that is, less than 0.5, between 0.5 and 0.9 and between 0.9 and 1.0. From this table it can bee seen that for smaller values of correlation coefficient, that is less than 0.5, the Horvitz – Thompson estimator under Yates – Grundy draw-by-draw procedure clearly outperforms all other estimators and is closely followed by the simple random sampling procedure. For other values of correlation coefficient the Murthy estimator outperforms all other estimators.

In general, we can see that the Murthy estimator performs reasonably well then all other estimators for almost all criterions and hence this estimator should be used for estimation of population total. For populations that have smaller correlation coefficient between the measure of size and correlation coefficient, the Horvitz – Thompson estimator under Yates – Grundy draw-by-draw procedure can produce reasonably good results. The simple random sampling procedure can also produce efficient results for populations having smaller correlation coefficient between variable of study and measure of size.

References

1. Brewer, K. R. W. (1963a) "A model of systematic sampling with unequal probabilities", *Aust. J. Stat.* 5, 5 – 13.
2. Das, A. C. (1951) "On two phase sampling and sampling with varying probabilities", *Bull. Intr. Stat. Inst.*, 33, Book 2, 105 – 112.
3. Hansen, M. H. and Hurwitz, W. N. (1943) "On the theory of sampling from a finite population", *Ann. Math. Stat*. 14, 333 – 362.
4. Horvitz, D. G. and Thompson, D. J. (1952) "A generalization of sampling without replacement from a finite universe", *J. Amer. Stat. Assoc*. 47, 663 – 685.
5. Murthy, M. N. (1957) "Ordered and unordered estimators in sampling without replacement", *Sankhya*, 18, 379 – 390.
6. Raj, D. (1956a) "Some estimators in sampling with varying probabilities without replacement", *J. Amer. Stat. Assoc.* 51, 269 – 284.
7. Rao, J. N. K., Hartley, H. O. and Cochran, W. G. (1962) "On a simple procedure of unequal probability sampling without replacement", *J. Roy. Stat. Soc*., B, 24, 482 – 491.
8. Yates, F. and Grundy, P. M. (1953) "Selection without replacement from within strata with probability proportional to size", *J. Roy. Stat. Soc*., B, 15, 153 – 161.