

**A SIMPLE GENERAL PROCEDURE FOR UNEQUAL PROBABILITY
SAMPLING WITHOUT REPLACEMENT**

Muhammad Hanif

National College of Business Administration & Economics
40/E-I Gulberg III, Lahore, Pakistan
drhanif@ncbae.edu.pk

M. Samiuddin

94/1, 7th street, Khyaban-e-Rahat DHA Phase VI
Karachi, Pakistan
msamiuddin@sat.net.pk

and

Muhammad Qaiser Shahbaz

Department of Statistics
Government College University, Lahore, Pakistan
hafiz_shahbaz@yahoo.com

ABSTRACT

A general class of selection procedure is developed for use with Horvitz and Thomson estimator for sample size $n = 2$. Known special cases of this class are Yates and Grundy (1953) draw-by-draw selection procedure, Yates and Grundy (1953) rejective procedure and Brewer (1963) draw-by-draw procedure. Empirical study based on 50 populations is carried out to compare performance of different member of the class.

KEY WORDS

Unequal Probability Sampling, Horvitz and Thompson estimator

1. INTRODUCTION

Horvitz and Thomson (1952) gave theoretical framework of unequal probability sampling without replacement Consider N units in a population. Associated with each unit I are values Y_I and X_I ($I = 1, 2, \dots, N$). X_I and Y_I are thought to be highly correlated. Further each X_I is exactly known for all I where as only those Y_I 's are made known which happen to have been drawn in the sample under any sampling plan. The problem is

to estimate the total $Y = \sum_{I=1}^N Y_I$ given the sample. The sampling scheme yields the probability of inclusion of unit I in the sample π_I and probability of joint inclusion of units J and I ($J \neq I$) in the sample π_{IJ} . Horvitz and Thompson (1952) gave an unbiased estimator of the population total.

$$y'_{HT} = \sum_{i \in s} y_i / \pi_i \quad (1.1)$$

(We will use smaller letter $x_i, y_i, \pi_i, \pi_{ij}$ for units in the sample)

The Variance of y'_{HT} is given by

$$\text{Var}(y'_{HT}) = \sum \frac{1 - \pi_i}{\pi_i} Y_i^2 + \sum \frac{\pi_{IJ} - \pi_i \pi_j}{\pi_i \pi_j} Y_i Y_j \quad (1.2)$$

An alternative form of (1.2) was given by Sen (1953) and Yates and Grundy (1953) independently:

$$\text{Var}(y'_{HT}) = \frac{1}{2} \sum_{\substack{I, J=1 \\ J \neq I}}^n (\pi_i \pi_j - \pi_{ij}) \left(\frac{Y_i}{\pi_i} - \frac{Y_j}{\pi_j} \right)^2 \quad (1.3)$$

These two forms lead to two unbiased estimates of $\text{Var}(y'_{HT})$ which may not be equal. These are:

$$\text{var}(y'_{HT}) = \sum_{i=1}^n \frac{1 - \pi_i}{\pi_i^2} y_i^2 + \sum_{\substack{i, j=1 \\ j \neq i}}^n \sum \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j \pi_{ij}} y_i y_j \quad (1.4)$$

$$\text{var}_{sYG}(y'_{HT}) = \frac{1}{2} \sum_{\substack{i, j=1 \\ j \neq i}}^n \sum \frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2 \quad (1.5)$$

A number of selection procedures have since been developed which can be used with the Horvitz and Thompson estimator. A comprehensive bibliography can be found in Hanif and Brewer (1980), Brewer and Hanif (1983) and Hanif et-al (1990).

2- THE GENERAL SELECTION PROCEDURE FOR $n=2$

This general procedure can be described as follows

- Select first units with probability proportional to $p_i(1-p_i)/(1-2ap_i)$ where $p_i < 1/2a$ for all i
- Select second unit with probability proportional to size of the remaining unit where $P_1 = X_1/X$ and $p_i = x_i/x$ and $X = \sum_1^N X_i$

The Probability of inclusion of i^{th} unit simplifies to

$$\pi_i = \frac{p_i}{b} \left[\frac{1-2p_i}{1-2ap_i} + \sum_{j=1}^N \frac{P_j}{1-2aP_j} \right], \quad (2.1)$$

where

$$b = \sum_{j=1}^N \frac{P_j(1-P_j)}{1-2aP_j}$$

π_{ij} works out to be

$$\pi_{ij} = \frac{2p_i p_j (1-ap_i - ap_j)}{b(1-2ap_i)(1-2ap_j)} \quad (2.2)$$

It can be checked that these values do satisfy the usual results such as

$$\sum_1^N \pi_i = n, \sum_{\substack{J=1 \\ J \neq I}}^N \pi_{IJ} = (n-1)\pi_i \text{ and } \sum_{I=1}^N \sum_{\substack{J=1 \\ J \neq I}}^N \pi_{IJ} = n(n-1) \text{ for } n = 2 \text{ (in the present case)}$$

It is known that $\pi_i \pi_j > \pi_{ij}$ for all i and j . Also for 50 real populations we have checked that for all \mathbf{a} values considered here $\pi_i \pi_j > \pi_{ij}$ for all i and j . However we do not yet have a general proof for this.

It is important to recognize that some very familiar sampling schemes are special cases of this general procedure. For example for $\mathbf{a} = 0$ the sampling scheme is equivalent to that of Yates and Grundy (1953) rejective procedure [in the sense of leading to the same values of π_i and π_{ij} , in fact it is interesting to note that Yates and Grundy (1953) rejective procedure has been converted to a draw by draw procedure]. For $\mathbf{a} = 0.5$ and 1 it reduces Yates-Grundy (1953) draw by draw procedure and Brewer (1963) draw by draw procedure respectively.

3. EMPIRICAL STUDY

In this section we have provided the comparative study of procedures based on different values of \mathbf{a} . Some explanation is needed however to explain the table and in particular why have we resorted to using ranks. Ranks correlations are effectively used to detect the monotone relation between two quantities (see Jeffreys (1961)). This is obvious because a perfect monotone relation between two quantities is reduced to a perfect linear relation in terms of ranks. Departure and extents of departure from this is reflected in the magnitude of ranks correlation and full regression analysis based on ranks. We have based our study on 50 real population frequently studied in the literature. These provide values of X_i and Y_i coefficient of variation of X , $[CV(X)]$, correlation coefficient between X and Y $[\rho_{XY}]$, skewness $S(X)$ and kurtosis $K(X)$ of X . We believe that X and Y are nearly proportional that ρ_{XY} is generally high. $CV(X)$, $K(X)$, $S(X)$ are exactly known or can be calculated and we may have a general idea about ρ_{XY} as prior information. With this background for each value of \mathbf{a} , we calculate variance of the estimate of Y using Horvitz and Thompson estimator. In this study we have selected

$\mathbf{a} = 0, 0.5, 1.0, 2.0, 2.5, 3.0, 3.5, 4$. Using the variance of estimator for different \mathbf{a} for the same population we have first converted these into ranks such that the smallest variance is given rank 1 and the largest rank 9 because we have 9 estimators one each for different \mathbf{a} . Also each population has the values of $CV(X)$, ρ_{XY} , $K(X)$, $S(X)$. These are then ranked from 1 to 50 in increasing order of magnitude. We will then end up with only ranks for 50 populations. From these the consolidated frequency tables are prepared separately for each \mathbf{a} . This is given in Table 1, the last row gives the mean rank for each \mathbf{a} .

Table 1: Frequency Table of Ranks of General Selection Procedures for various \mathbf{a} along with the Average Rank

Rank	Values of "a"								
	0.0	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0
1	23	5	3	4	2	3	2	0	8
2	7	27	3	2	3	0	1	7	0
3	4	3	30	3	2	1	6	0	1
4	3	6	4	30	1	6	0	0	0
5	1	2	4	3	34	3	1	1	1
6	4	0	1	8	5	30	1	1	0
7	1	2	5	0	3	5	33	1	0
8	2	5	0	0	0	2	3	38	0
9	5	0	0	0	0	0	3	2	40
Average Ranks	3.16	3.12	3.52	4.00	4.77	5.49	6.20	6.76	6.82

For an initial investigation of any relation between variance of estimator and ρ_{XY} or $CV(X)$ we have calculated average rank of estimator for each \mathbf{a} within ranges of ranks of ρ_{XY} and $CV(X)$. These are given in Tables 2 and 3 respectively. The last row in each case gives the exact rank correlation between the quantities involved, calculated from the ungrouped data of ranks.

Table 2: Average Ranks of Various Values of \mathbf{a} with ranks of Coefficient of Variation.

CV	Values of "a"								
	0.0	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0
1 – 10	3.60	3.40	3.60	3.70	4.60	5.50	6.40	7.20	7.00
11 – 20	4.60	4.50	4.40	4.30	4.40	4.80	5.40	6.00	6.60
21 – 30	2.90	2.80	3.20	3.90	4.60	5.50	6.50	7.40	8.20
31 – 40	2.20	2.20	3.00	3.80	4.90	6.10	7.30	7.30	8.20
41 – 50	2.50	2.70	3.40	4.30	5.40	5.70	5.90	7.50	7.60
Correlation	-0.08	-0.09	-0.08	0.05	0.30	0.15	0.12	0.13	0.12

Table 3: Average Ranks of Values of a with ranks of Correlation Coefficient.

ρ_{XY}	Values of "a"								
	0.0	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0
1 – 10	2.70	3.10	3.90	4.40	4.50	5.20	6.40	7.00	7.80
11 – 20	2.40	2.90	3.60	4.10	4.70	5.60	6.60	7.50	7.60
21 – 30	4.70	4.40	4.40	4.30	4.40	4.80	5.40	6.00	6.60
31 – 40	2.40	2.60	3.10	4.10	5.30	6.50	6.80	6.80	7.40
41 – 50	3.60	2.60	2.60	3.10	5.00	5.50	6.30	8.10	8.20
Correlation	0.16	-0.03	-0.25	-0.21	0.15	0.11	-0.01	0.13	0.07

We have also tried to see if Skewness and Kurtosis of X have any bearing on the average rank of the estimators. For this we have formed 4 groups such as positive S(X) and positive K(X), positive S(X) and negative K(X), negative S(X) and positive K(X) and negative S(X) and negative K(X). The results are given in Tables 4 and 5.

Table 4: Average Ranks and Total Frequency for Negative Kurtosis and Positive or Negative Skewness for X.

	Freq- uency	Values of "a"								
		0.0	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0
Positive S(X)	28	2.71	2.75	3.21	3.82	4.74	5.67	6.75	7.54	7.79
Negative S(X)	3	3.33	3.67	4.00	4.00	4.33	4.33	5.33	7.00	9.00
Combined S(X)	31	2.77	2.84	3.29	3.84	4.70	5.53	6.61	7.48	7.90

Table 5: Average Ranks and Total Frequency for Positive Kurtosis and Positive or Negative Skewness for X.

	Freq- uency	Values of "a"								
		0.0	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0
Positive S(X)	16	3.06	3.00	3.63	4.25	5.00	5.63	6.06	6.88	7.50
Negative S(X)	3	7.67	6.67	5.33	4.33	4.33	4.67	4.33	4.00	3.67
Combine d S(X)	19	3.79	3.58	3.89	4.26	4.89	5.47	5.79	6.42	6.89

Finally we have also made a regression analysis of rank of estimator for different 'a' on those of rank of CV(X) and rank of ρ_{XY} . The model considered is:

$$\text{Rank (a)} = \beta_0 + \beta_1(\text{Rank CV(X)}) + \beta_2(\text{Rank } \rho_{XY}) + \varepsilon \tag{3.1}$$

The results are presented in Table 6

Table 6: Regression Summary for Ranks of Various Values of a for model

	Values of "a"								
	0.0	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0
β_0	3.89	4.20	4.55	4.37	4.06	4.93	5.99	6.32	6.69
p-Value	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
β_1	-0.07	-0.04	-0.01	0.02	0.02	0.02	0.02	0.02	0.03
p-Value	0.03	0.06	0.46	0.18	0.23	0.39	0.41	0.35	0.38
β_2	0.04	-0.00	-0.03	-0.03	0.01	0.01	-0.01	0.01	0.00
p-Value	0.21	0.98	0.06	0.02	0.43	0.65	0.80	0.72	0.91
F	2.65	2.17	3.06	3.12	1.59	0.71	0.34	0.74	0.52
p-Value	0.08	0.13	0.06	0.05	0.22	0.50	0.71	0.48	0.60

4. CONCLUSIONS

Some tentative conclusions can be drawn from Table 1. Mean and average Rank is lowest for $a = 0.5$ (Yates-Grundy draw by draw procedure) and for $a = 1.0$ Brewer (1963) procedure. There is a strong case for Brewer's procedure because the Horvitz and Thompson estimator in this case is also model unbiased. This is closely followed by $a = 0$ (Yates-Grundy rejective procedure). However procedures other than Brewer do not enjoy the property of model unbiasedness. This property of model unbiasedness looks even more compelling if we try to see its implications as follows. Suppose we were to use the estimator to estimate the total X . We feel that since X is known our estimate should equal X . A strange but true picture also emerges. Maximum frequency for different 'a' seem to be moving diagonally starting at rank 1 for $a = 0$ (23) to rank 9 for $a = 4$ (40). This puts some premium for the case $a = 0$. We see for example 23, 5 and 3 cases out of 50 turn out to be best performers when $a = 0$, $a = 0.5$ and $a = 1$ respectively. We can also try some negative values of a , with permissible limits in further investigations. The average rank given in Table 1 also indicates that the real competitors are $a = 0$, $a = 0.5$ and $a = 1.0$. Looking to Table 2 one can perhaps suggest that for very high values of $CV(X)$ one can choose either $a = 0$, or $a = 0.5$. The last row in Table 2 also indicated that as $CV(X)$ increases. The cases covered by $a = 0$, $a = 0.5$ and $a = 1.0$ perform better since in all three cases ranks correlation is negative. Table 3 indicates that a very strong correlation coefficient between X and Y may lead to the choice of $a = 0.5$ or perhaps $a = 1.0$. The last case is particularly intuitive because Brewer's estimator is model unbiased also. Tables 4 and 5 do not provide any guide lines for choosing between $a = 0$, $a = 0.5$ and $a = 1.0$ Table 6 seem to suggest that for $a = 0$, rank $CV(X)$ seem to significantly affects the average rank where as for $a = 1.0$. Rank (ρ_{xy}) seem to significantly affect the average rank.

5. REFERENCES

1. Brewer, K.R.W. (1963). A model of systematic sampling with unequal probabilities. *Aust. J. Stat.* 5, 5 – 13.
2. Brewer, K.R.W. and Hanif, M. (1983). *Sampling with Unequal Probabilities*, Lecture notes to Statistics, No. 15, Springer – Verlag.
3. Hanif, M. and Brewer, K.R.W. (1980). Sampling with unequal probabilities without replacement; a review. *Inter. Stat. Rev.* 48(3), 317 – 335.
4. Hanif, M., Beg, M. A. and Khawaja I. (1990), Sampling with unequal probabilities A historical review. Proceedings of the first Iranian Statistics Conference held at Isfahan University, May 1992 Vol. 1. Invited paper, 19 – 38.
5. 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.
6. Jeffreys, H. (1961). *Probability Theory*, 3rd Edition, Oxford University Press, Oxford.
7. Sen, A.R. (1953). On some estimate of the variance in sampling varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, 5, 119 – 27.
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 – 161.305