Research Journal of Mathematical and Statistical Sciences ______________________________E-ISSN 2320-6047
Vol. **4(6),** 1-5, July **(2016)** *Res. J. Mathematical and Statistical Sci.*

Algorithm-I on Splitting of Sizes for Sampling Procedure with Inclusion Probabilities Proportional to Size

Dwivedi Vijai Kumar
Department of Statistics, University of Botswana, BOTSWANA

Available online at: www.isca.in, www.isca.me
Received 3rd May 2016, revised 9th June 2016, accepted 7th July 2016

Abstract

The sampling scheme proposed by Srivastava and Singh depends upon a specific split of the sizes and most of the π_{ij}'s do not satisfy the condition of non-negativity of variance estimates as suggested by Hanurav. Using the information about the nature of non-negativity condition (ϕ_{ij}>0) approach, Dwivedi provided split of sizes with less number of trials which gives a set of π_{ij} 's satisfying the condition of non-negativity of variance estimates. This paper provides an algorithm for a proper split of sizes which gives a set of π_{ij} 's such that the condition of non-negativity of variance estimates is nearly satisfied. It is also shown that on an average the relative efficiency of proposed algorithm shows the superiority over PPSWR.

Keywords: Selection probabilities, unequal probabilities, inclusion probabilities.

Introduction

The efficiency of πPS sampling schemes differs from one another only due to different sets of joint inclusion probabilities (i.e. π_{ij} ′s). Srivastava and Singh[1]suggested a πPS sampling scheme which depends upon a particular split of the sizes and each split provides a πPS sampling design with different sets of π_{ij} ′s. Further most of the π_{ij}′s do not satisfy the condition of non-negativity of variance estimates as suggested by Hanurav[2]. Using the information about the nature of non-negativity condition (ϕ_{ij}>0) approach, Dwivedi[3] provided split of sizes with less number of trials which gives a set of π_{ij} ′s satisfying the condition of non-negativity of variance estimates.

However a systematic approach is needed for getting suitable splits which may provide desirable sets of π_{ij} ′s. The aim of this paper is to provide an algorithm for a proper split of sizes which gives a set of π_{ij} ′s such that the condition $\left(\phi_{ij} = \dfrac{\pi_{ij}}{\pi_i \pi_j} < 1\right)$ of non-negativity of variance estimates is nearly satisfied, Hanurav[2]. Dwivedi[3] has described Srivastava and Singh[1]sampling procedure; however for completeness it is given in brief as follows: Let the population under consideration consists of N distinct and identifiable units and a sample of size n is desired to be drawn from it. Let X_i be the values of auxiliary characters for the unit U_i (i=1,2,3,...,N) in the population. It is assumed that X_i's are known for all i's and set $P_i = X_i/X$, where $X = \sum_{i \in U} X_i$. It assumed that population unit X_i's are rearranged in ascending order such that;

$$0 < X_i \le X_{i+1} \qquad \text{for } i=1,2,...,N-1 \tag{1}$$

and $nX_i < X < 1$ i.e. $nP_i < 1$ for all i

under such inequality these sizes are split upand named as initial split

$$X'_i = X_i - X_{i-1} \quad (i= 1,2,3,\ldots,N)$$
$$X'_1 = X_1, \text{ since } \quad X_0 = 0$$

The equalities of some of X_i's will result in reducing the number of columns.

The *N* sizes are expressed as

$$\underline{X} = \underline{A}\, \underline{Z} \tag{2}$$

where $\underline{Z}$ is a vector of M non-zero elements and $\underline{A}$ is a N x M matrix whose i,j-th element is 0 or 1 accordingly as cell is filled or empty. If A_{ij} satisfies

$$\sum_{i=1}^{N} A_{i1} = k_1 = N$$
$$\sum_{i=1}^{N} A_{it} = k_t \ \ge\ n; \ \ (t > 1) \tag{3}$$

Then,

$$\pi_i = np_i$$
$$\pi_{ij} = \frac{n(n-1)}{X} \sum_{t=1}^{M} \frac{Z_t A_{it} A_{jt}}{(k_t - 1)} \tag{4}$$

Corresponding to every split satisfying condition (3), one πPS sampling design can be obtained. The controls provided on π_{ij} ′s help in choices of proper splits. Kaur[4] has studied the effect on

π_{ij} 's when only one element is shifted as such from one to other column such that columns means for both the effected columns are same. Thus, each split provides a πPS sampling design with different sets of π_{ij} 's. The sample spaces generated either by Jessen's[5] methods I, II, and III would also be possible to obtain by this method. Also the selection procedure 'A' suggested by Sengupta[6]is a particular case of this method.

Dwivedi[3] has studied the nature of ϕ_{ij} 's for the initial split. Since X_i's are arranged in increasing order it is clear that ϕ_{ij} 's will go on decreasing as we proceed from left to right in upper half of matrix Φ while for the range of ϕ_{ij} 's considered here ϕ_{ij} 's increase as we proceed downwards in Φ i.e. $\phi_{i+1,j} / \phi_{ij} > 1$. However, it remains fairly stable along the diagonal of the matrix Φ. Thus it appears thatin Φ, ϕ_{ij} decreases as $|i-j|$ increases. However these results holds only for i<j and i<N-n+1.

Using the above results an algorithm is proposed in the next section which arranges the initial split in such a way that there are at least *n* non-zero elements in each column along with satisfying condition mentioned in equation (3).

Algorithm to control on $\phi_{i,i+1}$'s

Before writing the steps of proposed algorithm, the following theorem on splitting is given below.

Theorem: The revised value of X_i' (say X_i'') for t-th column of t-th unit are:

$$X_t'' = 1 + R(N-t)\left[\frac{(nX_t X_{t+1})}{(n-1)X} - \sum_{i=1}^{t-1} \frac{X_i''}{(N-i)}\right]; \quad (t = 1, 2, ..., N-n) \qquad (5)$$

$$X_t'' = 1 + R\left[\frac{(nX_t X_{t+1})}{X} - (n-1)\sum_{i=1}^{t-1} \frac{X_i''}{(N-i)}\right]; \quad (t = N-n+1) \qquad (6)$$

Where R is the desired value of $\phi_{t,t+1}$

Proof: Let there are N units in the population, and a sample size of *n* is to be drawn from it. X_t is the size of the t-th unit and these are arranged in ascending order. Such that

$$X_t = \sum_{j=1}^{t} X_j' ; \qquad (t = 1, 2, ..., N)$$

and for j=N-n+2,…,N the X_j' is given zero values; where X_j' is the value of the initial split in j-th column (j=1,2,…,t). Now with condition mentioned in equation (3) we can only obtain the revised split valuesup to column N-n+1 for the sampling scheme proposed by Srivastava and Singh[1], thus

$$\pi_{12} = \frac{n(n-1)}{X}\left[\frac{X_1'}{N-1}\right] = \pi_{13} = \quad ... \quad = \pi_{1N}$$

$$\pi_{23} = \frac{n(n-1)}{X}\left[\frac{X_1'}{N-1} + \frac{X_2'}{N-2}\right] = \pi_{23} = \quad ... \quad = \pi_{2N}$$

$$\pi_{t,t+1} = \frac{n(n-1)}{X}\left[\frac{X_1'}{N-1} + \frac{X_2'}{N-2} + ... + \frac{X_t'}{N-t}\right] = \pi_{t,t+2} = \quad ... \quad = \pi_{tN}$$

i.e. $$\pi_{t,t+1} = \frac{n(n-1)}{X}\left[\sum_{i=1}^{t} \frac{X_i'}{N-i}\right]; \qquad (t = 1, 2, ..., N-n) \qquad (7)$$

and

$$\pi_{t,t+1} = \frac{n(n-1)}{X}\left[\sum_{i=1}^{t-1} \frac{X_i'}{N-i}\right] + \frac{n}{X} X_{t+1}' \qquad (8)$$

also,

$$\pi_{t,t+1} = \frac{n^2}{X^2} X_t X_{t+1} \qquad (9)$$

Now the condition of non-negative variance is satisfied if

$$\phi_{t,t+1} = \frac{\pi_{t,t+1}}{\pi_t \pi_{t+1}} < 1 \qquad (10)$$

From (7), (9) and (10) for t= 1,2,…,N-n; we have

$$\sum_{i=1}^{t} \frac{X_i'}{N-i} < \frac{n X_t X_{t+1}}{(n-1)X}$$

or

$$\frac{X_i'}{N-i} < \frac{n X_t X_{t+1}}{(n-1)X} - \sum_{i=1}^{t-1} \frac{X_i'}{N-i} \; ....$$

Therefore, revised value X_t'' is obtained as

$$X_t'' = R(N-t)\left[\frac{n X_t X_{t+1}}{(n-1)X} - \sum_{i=1}^{t-1} \frac{X_i''}{N-i}\right] \qquad (11)$$

Where R is the desired value of $\phi_{t,t+1}$.

As we are to retain the integral part of the right hand side of (11) and sometimes left hand side of (11) may take value less than one, in that situation $X_t^{''}$ will become zero. Therefore we have added unity to (11) to avoid the zero value of $X_t^{''}$. The final value of $X_t^{''}$ is thus

$$X_t^{''} = 1 + R(N-t)\left[\frac{n X_t X_{t+1}}{(n-1)X} - \sum_{i=1}^{t-1} \frac{X_i^{''}}{N-i}\right] \quad ; t = 1,2,...,N-n \qquad (12)$$

On similar lines as in (8), (9) and (10) for t=N-n+1, we have revised value of $X_t^{''}$ as

$$X_t^{''} = 1 + R\left[\frac{n X_t X_{t+1}}{X} - (n-1)\sum_{i=1}^{t-1} \frac{X_i^{''}}{N-i}\right] \qquad (13)$$

Q.E.D.

Steps involved in Algorithm -I

As the values of ϕ_{ij} 's depend on the values of initial splits ($X_i^{'} = X_i - X_{i-1}$) so we revise the initial split in such a way that $\phi_{i,i+1}$ (i=1,2,…,N-1) values are closed to the desired level, say R, then the other ϕ_{ij} 's (j>i+1) values will automatically will be less than R. However we can revise the initial split up to N-n+1 columns because of initial restrictions imposed on X_i 's. Thus the values of ϕ_{ij} 's (i=1,2,…,N-n+1 ; j=i+1,…,N)could be brought to the level of R. The proposed algorithm consists of the following steps:

Step-1 : Arrange the X_i 's in the ascending order in the column of sizes and place the i-th unit in the i-th row. Calculate $\pi_i = n\frac{X_i}{X}$; ($X = \sum_{i=1}^{N} X_i$).

Step-2 : Select a value of R lying between 0 and 1, being the desired level of

$\phi_{i,i+1}$; (i=1,2,…, N-n+1).

Let t=1.

Step-3 The revised values $X_t^{''}$ of $X_t^{'}$ for the t-th column of t-th unit can be obtained using the equations (12) for t=1,2,…,N-n and (13) for t=N-n+1.

Because of the computational simplicity, only integer position is retained in the right hand side in the above equations (12) and (13).

Step-4 : Calculate $L_t = \sum_{j=1}^{t} X_j^{''}$.

Step-5 : If $L_t > X_t$ proceed to Step-6. Otherwise put $X_t^{''}$ in the t, t+1,…,N-th rows, in the t-th column. Then take the next higher value of t and return to Step-3.

Step-6 : If $L_t > X_t$, put $X_t^{''} = X_t - \sum_{i=1}^{t-1} X_i^{''}$ and return to Step-5, otherwise proceed to the next step.

Step-7: In order to satisfy the following condition for a valid split,

$$X_t = \sum_{j=1}^{t} X_j^{''} \ ; \ (t = 1,2,...,N) \qquad (14)$$

We calculate, $M_t = X_t - \sum_{j=1}^{t} X_j^{''}$; $(t = 1,2,...,N)$

and put it in the t-th row in the (N-n+2)-th column. Here for j=N-n+2,…,N; $X_j^{''}$ is given the value zero.

It is clear that in the table of split, the condition in equation (3) satisfied in the columns 1 to N-n+1. However, this condition may not be fulfilled in column N-n+2. Therefore, to satisfy the condition this last column is suitably adjusted from one to another column. Consequently final split is obtained. The final split may be put in matrix form (2) and π_{ij} 's values are calculated using formula (4).

Therefore with the help of above algorithm we get the split values up to (N-n+1)-th column, which satisfy condition (3) and resulting values of π_{ij} 's satisfy condition $\phi_{ij} < 1$.

However (N-n+2)-th column is to be adjusted in such a way that the above conditions are satisfied. It is noted here that in splitting method of Srivastava and Singh[1], to satisfy condition (3) elements were shifted from one to another columns consequently some resulting values of π_{ij} 's did not satisfy condition $\phi_{ij} < 1$ and to satisfy this condition large number of shifting were required.

However in the proposed algorithm for some set of π_{ij} 's values less shifting are required. One of the major advantage of this algorithm is that the values of $\phi_{i,i+1}$ is in control of sampler, while for the same desired level of $\phi_{i,i+1}$ Srivastava and Singh's splitting method requires more steps. The proposed algorithm works satisfactorily for small values of N and n.

Example of splitting and corresponding set of π_{ij} 's

We present the population considered by Cochran[7] along with X and Y. Though for calculation of π_{ij}'s Y values are not needed, however presented for the sake of future reference.

Table-1
Three artificial populations of size N = 5

Population	**U_i**	**U_1**	**U_2**	**U_3**	**U_4**	**U_5**
Size	X_i	1	1	2	3	3
Population A	Y_i	0.3	0.5	0.8	0.9	1.5
Population B	Y_i	0.3	0.3	0.8	1.5	1.5
Population C	Y_i	0.5	0.5	0.8	0.9	0.9

Without loss of generality and for the sake of computational ease, let us change the scale of X_i's by a constant multiplier (say, 10). Therefore the new values of X_i's are 10,10,20,30 and 30. Here, N=5; n=2; N-n+1= 4 and N-n+2 = 5. The initial split is given in Table-2.

Table-2
Initial Split (Srivastava and Singh[1])

Population Units	**Sizes (X_i's)**	X_t'				
		1	**2**	**3**	**4**	**5**
1	10	10				
2	10	10	0			
3	20	10	0	10		
4	30	10	0	10	10	
5	30	10	0	10	10	0
Total	100	50	0	30	20	0

It is clear that the initial split does not satisfy that property of having at least 2 non-zero elements in every column. Now below we will demonstrate the split using algorithm such that every column satisfy that property of having at least 2 non-zero elements.

The value of R is chosen as 0.5 (Step-2), since the value of N-n+1 is 4, we obtain the values of X_1'', X_2'', X_3'', and X_4'' using Steps 3 to 6.These values are 5, 5, 10 and 2 respectively. Values of X_t'' are placed in their respective columns and rows (Step-5).

Using Step-7 the values of M_1, M_2, M_3, M_4 and M_5 are obtained as 5, 0, 0,8, and 8 respectively. The above is summarised in the following table.

Table-3
A split of sizes with at least 2 nonzero elements in each column

Population Units	**Sizes (X_i's)**	X_t''				***M_i*'s**
		1	**2**	**3**	**4**	**5**
1	10	5				5
2	10	5	5			0
3	20	5	5	10		0
4	30	5	5	10	2	8
5	30	5	5	10	2	8
Total	100	25	20	30	4	21

The elements in 5-th column do not satisfy condition in equation (3), to satisfy this the values of M_i's are adjusted in columns 4 and 5. The final split is thus obtained as follows:

Table-4
Final split of sizes with at least 2 nonzero elements in each column

Population Units	**Sizes (X_i's)**	**Columns (Groups)**				
		1	**2**	**3**	**4**	**5**
1	10	5				5
2	10	5	5			0
3	20	5	5	10		0
4	30	5	5	10	5	5
5	30	5	5	10	5	5
Total	100	25	20	30	10	15

These values are put in the matrix form (2) and corresponding π_{ij} 's values are calculated using formula depicted in equation (4).

The Φ matrix corresponding to split in Table-4

$$\begin{array}{c} \quad\;\; 1 \quad\;\; 2 \quad\;\; 3 \quad\;\; 4 \quad\;\; 5 \\ \Phi = \begin{bmatrix} - & 0.63 & 0.31 & 0.62 & 0.62 \\ & - & 0.73 & 0.49 & 0.49 \\ & & - & 0.66 & 0.66 \\ & & & - & 0.85 \\ & & & & - \end{bmatrix} \end{array} \qquad (15)$$

The Φ matrix corresponding initial split in Table-2

$$\begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \Phi = \begin{bmatrix} - & 1.25 & 0.62 & 0.42 & 0.42 \\ & - & 0.62 & 0.42 & 0.42 \\ & & - & 0.62 & 0.62 \\ & & & - & 0.97 \\ & & & & - \end{bmatrix} \end{array} \quad (16)$$

From (15) and (16) it is clear that proposed algorithm provides a set of π_{ij}'s which satisfy the condition $\phi_{ij} < 1$ whereas the set of π_{ij}'s obtained using the method of Srivastava and Singh[1] does not satisfy this condition. Hence we may conclude that for the same set of π_{ij}'s proposed algorithm requires less number of shifting of elements than the method of Srivastava and Singh.

Estimation Procedure

After obtaining various inclusion probabilities usual Horvitz Thompson (HT) estimator of population total can be used.

$$e_{HT} = \sum_{i \in s} \frac{y_i}{\pi_i} \quad (17)$$

The variance of e_{HT} is given by

$$V(e_{HT}) = \sum_{i<j} \sum_{j \in s} (\pi_i \pi_j - \pi_{ij}) \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2 \quad (18)$$

For the universe one given in Table 1 the exact variance of estimates of the population total Y resulting from proposed algorithm compared with PPS with replacement sampling scheme (PPS WR) is given in Table 5.

Table-5
Relative efficiency of proposed algorithm over PPS WR

Population	**PPS WR**	**Proposed Algorithm**		
		R- Values		
		0.4	**0.5**	**0.6**
A	0.400	0.167	0.357	0.267
B	0.320	0.287	0.258	0.307
C	0.320	0.287	0.258	0.307
Average	0.347	0.247	0.291	0.294
Relative Efficiency	100	140	119	118

Conclusion

It is clear from the above table that on an average the relative efficiency of proposed algorithm shows the superiority over PPSWR. It is observed that proposed algorithm is more sensitive to the population characteristics. The proposed algorithm exercises the control on $\pi_{i,i+1}$ (i=1,2,,…,N-1) i.e. on diagonal values of π_{ij}'s .

Acknowledgment

Author is thankful and grateful to Prof. A. K. Srivastava, Former Principal Scientist, IASRI, New Delhi for his valuable guidance to carry out this work.

References

1. Srivastava A.K. and Singh D. (1981). #A Sampling Procedure with Inclusion Probabilities Proportional to Size.# *Biometrika*, **68**, 3, 732-734.

2. Hanurav T.V. (1967). #Optimum utilization of Auxiliary Information: πPS Sampling of two Units from a Stratum.# *J. R. Statist. Soc.*, B29, 374-391.

3. Dwivedi V.K. (2015). #Inclusion Probabilities Proportional to Size Sampling Scheme based on Splitting of Sizes.# *Res. J. Mathematical and Statistical Sci*., 3(11), 6-11.

4. Kaur Surinder (1980). #Some Contributions to Varying Probability Sampling.# (Unpublished Ph.D. thesis), Indian Agricultural Statistics Research Institute (IASRI), New Delhi, INDIA.

5. Jesson R.J. (1969). #Some Methods of Probability Non-Replacement Sampling.# *J. American Statistical Association*, 64,175-193.

6. Sengupta S. (1981). #A Sampling Scheme to Realize Inclusion Probability Exactly Proportional to Size.# *J. Indian Soc. of Agri. Statistics.*, 33(2), 19-23.

7. Cochran W.G. (1977). #*Sampling Techniques.*# 3rd ed. Wiley Publication, New York.

8. Dwivedi V.K. (1984). #Some investigations in sampling with unequal probabilities.# (Unpublished Ph.D. thesis), Indian Agricultural Statistics Research Institute (IASRI), New Delhi, INDIA.