Research Journal of Mathematical and Statistical Sciences ISSN 2320–6047
Vol. **3(11),** 6-11, November **(2015)** *Res. J. Mathematical and Statistical Sci.*

Inclusion Probabilities Proportional to Size Sampling Scheme Based on Splitting of Sizes

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

Available online at: www.isca.in, www.isca.me
Received 19th August 2015, revised 17th September 2015, accepted 19th October 2015

Abstract

Srivastava and Singh suggested inclusion probabilities proportional to size (IPPS or πPS) sampling scheme which depends upon a specific split of the sizes named as initial split. Each split provides a πPS sampling design with different sets of joint inclusion probabilities (π_{ij}'s). The sampling scheme proposed by them is not exactly πPS unless the largest n units are having same sizes. Furthermore most of the π_{ij}'s do not satisfy the condition of non-negativity of variance estimates as suggested by Hanurav. The scheme has got potential of providing π_{ij}'s satisfying desirable properties with suitable splits; however a systematic approach is needed for getting such splits. Using the information about the nature of non-negativity condition (ϕ_{ij} 's >0) approach, this paper provides 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

Keywords: Selection probabilities, unequal probabilities, inclusion probabilities.

Introduction

Sampling scheme with inclusion probabilities proportional to sizes (IPPS or π PS) has got an efficient appeal, when Horvitze Thompson (HT) estimator is used. There is no dearth of IPPS sampling scheme available in literature and most of them deal with the case for sample size two. However, these sampling schemes have got some inherent limitations. In general these techniques are complicated for sample size greater than two or else provide joint inclusion probabilities (π_{ij} 's) only in approximate form and consequently asymptotic variances are generally provided. It is known that the efficiency of πPS sampling schemes differ from one another only due to different sets of joint inclusion probabilities (i.e. π_{ij} 's). Thus, a proper choice of π_{ij} 's will help in getting a good πPS sampling scheme. Srivastava and Singh[1] suggested a πPS sampling scheme which depends upon a specific split of the sizes. Thus each split provides a πPS sampling design with different sets of π_{ij} 's. The sampling scheme suggested by them is not exactly πPS unless the largest n units are having same sizes however the scheme has got potential of providing π_{ij} 's satisfying desirable properties especially satisfying the condition of non-negativity of variance estimates. The splits which provide desirable sets of π_{ij} 's; trail method are followed which might take a number of trials to achieve the proper splits. Thus a systematic approach is required for getting splits which may provide desirable sets of π_{ij} 's. Using the information about the nature of ϕ_{ij}'s the aim of this paper is to provide a proper split of sizes with less number of trials which gives a set of π_{ij} 's such that the condition

$$\phi_{ij} = \pi_{ij}/\pi_i\pi_j < 1 \qquad (1)$$

of non-negativity of variance estimates is nearly satisfied, Hanurav[2]. The sampling scheme considered by Srivastava and Singh[1] is given in brief as follows.

Sampling scheme by Srivastava and Singh[2]: 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 is assumed that population unit X_i's are rearranged in ascending order such that.

$0 < X_i \le X_{i+1}$ for i=1,2,...,N-1 (2)

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

under such inequality these sizes are split up and is presented in Table 1.

The sampling scheme with initial split given in Table 1 consists of the following steps: i. A column is selected with probability proportional to column total. ii. If the selected column has more than *n* non-zero entries, *n* units are selected with SRS without replacement. Otherwise select all the units with non-zero entries in the column.

The inclusion probability π_i of i-th unit is given by

$$\pi_i = \begin{cases} nP_l & (i \le N-n+1) \\ nP_l - (n+i-N-1)P_l + \sum_{j=N-n+1}^{i-1} P_j & (i > N-n+1) \end{cases} \qquad (3)$$

and the joint inclusion probability of i-th and j-th units (j>i) in the sample is

$$\pi_{ij} = \begin{cases} n(n-1)P_1/(N-1) & (i=1) \\ \pi_{i-1,j} + n(n-1)(P_i - P_{i-1})/(N-1) & (1<i\le N-n+1) \\ \pi_{i-1,j} + (n-i+1)(P_i - P_{i-1}) & (i>N-n+1) \end{cases} \quad (4)$$

In general, the *N* sizes are expressed as

$$\underline{X} = \underline{A}\,\underline{Z} \quad (5)$$

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

$$\left.\begin{aligned} \sum_{i=1}^{N} A_{i1} = k_1 = N \\ \sum_{i=1}^{N} A_{it} = k_t \ge n; (t>1) \end{aligned}\right] \quad (6)$$

Then,

$$\left.\begin{aligned} \pi_i = \frac{n}{X}\sum_{j=1}^{M} A_{ij} Z_j = np_i \\ \pi_{ij} = \frac{n(n-1)}{X}\sum_{t=1}^{M} \frac{A_{it} A_{jt} Z_t}{(k_t - 1)} \end{aligned}\right] \quad (7)$$

Corresponding to every split satisfying condition in equation (6), one πPS sampling design can be obtained. So the splitting procedure has got lot of flexibility with respect to choice of π_{ij} ′s. Thus through a proper choice of splitting, controls may be exercised on π_{ij} ′s. Also from efficiency point of view πPS sampling schemes differ from one another only due to different sets of π_{ij} ′s.

In order to get a set of desired π_{ij}'s the approach followed here is to start with an initial split and then shifting the individual elements keeping in view the effect of each change, such that the desired properties of π_{ij}'s are satisfied. It is therefore necessary to study the nature of ϕ_{ij}'s for the initial split. This is presented in the next section.

An approach to get IPPS sampling schemes through initial split: It is obvious from the initial split proposed by Srivastava and Singh[2]that control on π_{ij} ′s are to be carried through the shifting of elements from one to another columns in initial split on trial and error basis. Here a basic question arises that how these shifting should be operated such that the condition in equation (6) is satisfied? For this, an approach of split 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 conditions mentioned in equations (1) and (6) in less number of trails than Srivastava and Singh[2].

Nature of ϕ_{ij}'sfor the initial split: From the initial split the first N-n+1 columns have at least n elements in each column and the last (n-1) columns have less than n elements. Then the joint inclusion probabilities of i-th and j-th (i<j and i<N-n+1) is given

$$\pi_{ij} = \frac{n(n-1)}{X}\sum_{t=1}^{N-n+1} \frac{\overline{X}_{.t}\, A_{it}\, A_{jt}}{(k_t - 1)} \quad (8)$$

It is clear from the π_{ij} matrix that π_{ij} = is constant for all j=i+1, N and for a given i≤N-n+1; which shows that all the columns to a given row are same.

The value in the π_{ij} matrix increases as i increase,

$$\pi_{2j} - \pi_{1j} = \frac{n(n-1)\left(\overline{X}_{.2} - \overline{X}_{.1}\right)}{X(N-2)} > 0 \quad (9)$$

This shows that for the initial split in upper half of Φmatrix (i.e. i<j and i<N-n+1) the values of ϕ_{ij}'s increases as we proceed downwards while it remains constant in the rows. These properties of π_{ij} for the initial split may be used to study the nature of Φ matrix. Since X_i's are arranged in increasing order, the X_{ij}'s will go on decreasing as we proceed from left to right in upper half of Φ matrix while for the range of ϕ_{ij}'s considered here ϕ_{ij}'s increase as we proceed downwards in Φ i.e.$\phi_{ij} = \pi_{ij}/\pi_i\pi_j$<1.However, it remains fairly stable along the diagonal of the matrix Φ. Thus it appears that in Φ, ϕ_{ij} decreases as $|i - j|$ increases. However, these results holds only for i< j and i< N-n+1. Now information about the nature of ϕ_{ij}' son the initial split have been used to do the further splits (i.e. control on π_{ij}'s) such that condition of non-negativity of variance estimate is nearly satisfied. The above results are summarized in table-2 for ready reference.

Numerical example of splitting and corresponding set of ϕ_{ij}'s: The example considered by Srivastava and Singh[1]. Sukhatme and Sukhatme[3] is considered for illustration that is a village has 10 orchards containing 150, 50, 80, 100, 200, 160, 40, 220 and 140 trees. A sample of size n=4 is to be drawn using number of trees as the sizes, and with inclusion probability proportional to size.

The method which involves splitting the sizes of the population units arranged in ascending order. The splitting forms colums. We call this split as initial split. From the initial split in table-3 if one of the columns 8-10 is selected, the achieved sample size would be less than four. To achieve the required sample size they have shown that original size values can be split in different ways so as to ensure at least *n* non zero values in each column (table-4).

Now the next step is to examine whether the resulting values of π_{ij}'s from this split (table-4) follow condition given in equation (1) or not? For this the Φ matrix corresponding to split is examined in table-5.The components of ϕ_{ij}' sare calculated using equation (7). The Φ matrix in table-5 indicate that this split does not provide a set of π_{ij}'s which satisfy condition (1), required for non-negative variance estimator. Here if we control ϕ_{12}, the other elements namely ϕ_{13}, ϕ_{14} and ϕ_{15} will be

adjusted accordingly. As ϕ_{12} is greater than unity, then evidently 1st column itself requires to be split up into more than one columns and retaining first part as it is, some of elements from the other part may be shifted to subsequent columns. Similarly further split, as and when required can be done for the columns corresponding to elements ϕ_{23}, ϕ_{34} and ϕ_{45}. This will result in reducing the values ϕ_{ij}'s for the upper left hand corner of Φ matrix. Subsequent shifts may be incorporated in order to get a split for which ϕ_{ij}'s are maintained within the reasonable limits.

Hence in this approach after necessary shifting of elements in initial-split (table-3) the corresponding ϕ_{ij}'s are examined. This process continues till a set of π_{ij}'s is obtained which satisfies condition (1) or corresponding ϕ_{ij}'s values remain in the reasonable limits. A final split for the initial split (table-3) is given in table-6, for which resulting ϕ_{ij}'s values are within the reasonable limits (table-7). The final split has been obtained after three successive splitting processes along with control on π_{ij}'s. Here the final split is named in the sense that a split of sizes which satisfy condition (6) exactly and condition (1) nearly. The matrix for the final split (table-6) is givenin Table-7.

Table-1
Initial split of sizes

Pop. Unit	Sizes	C_1	C_2	…	C_{N-n+1}	…	C_{N-1}	C_N
1	$X_{1=}$	X'_1		…		….		
2	$X_{2=}$	X'_1	$+X'_2$	…		….		
.		X'_1	$+X'_2$	…		….		
.		X'_1	$+X'_2$	…		….		
N-n+1	$X_{N-n+1=}$	X'_1	$+X'_2$	…	$+X'_{N-n+1}$	….		
.		X'_1	$+X'_2$	…	$+X'_{N-n+1}$	….		
.		X'_1	$+X'_2$	…	$+X'_{N-n+1}$	….		
N-1	$X_{N-1=}$	X'_1	$+X'_2$	…	$+X'_{N-n+1}$	….	$+X'_{N-1}$	
N	$X_{N=}$	X'_1	$+X'_2$	…	$+X'_{N-n+1}$	….	$+X'_{N-1}$	$+X'_N$
Total	X=	NX'_1	$+(N-1)X'_2$	…	$+nX'_{N-n+1}$	….	$+2X'_{N-1}$	$+X'_N$

$X'_i = X_i - X_{i-1}$ (i= 1,2,3,…,N). $X'_1 = X_1$, since $X_0 = 0$. The equalities of some of X_i's result in reducing the number of columns.

Table-2
Nature of Φ matrix for initial split in Table 1; (i<j and i ≤ N-n+1)

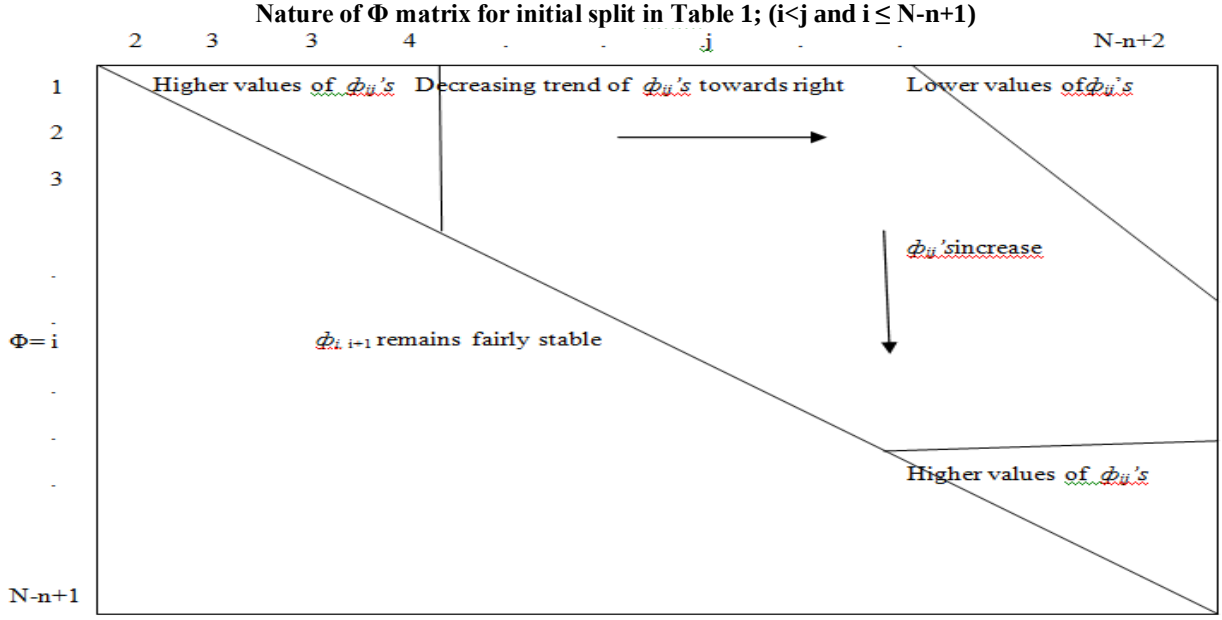

Table-3
Initial split of sizes for numerical example

Pop. units	Sizes	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}
1	40=	40									
2	50=	40	+10								
3	60=	40	+10	+10							
4	80=	40	+10	+10	+20						
5	100=	40	+10	+10	+20	+20					
6	140=	40	+10	+10	+20	+20	+40				
7	150=	40	+10	+10	+20	+20	+40	+10			
8	160=	40	+10	+10	+20	+20	+40	+10	+10		
9	200=	40	+10	+10	+20	+20	+40	+10	+10	+40	
10	220=	40	+10	+10	+20	+20	+40	+10	+10	+40	+20
Total	1200=	400	+90	+80	+140	+120	+200	+40	+30	+80	+20

Table-4
A split sizes with at least 4 non-zero values in each column Srivastava and Singh[2]

Pop. units	Sizes	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}
1	40=	40									
2	50=	40	+10								
3	60=	40	+10	+10							
4	80=	40	+10	+10	+20						
5	100=	40	+10	+10	+20	+20					
6	140=	40	+10	+10		+20				+40	+20
7	150=	40	+10			+20	+40	+10	+10		+20
8	160=	40	+10	+10			+40	+10	+10	+40	
9	200=	40	+10	+10	+20		+40	+10	+10	+40	+20
10	220=	40	+10	+10	+20	+20	+40	+10	+10	+40	+20
Total	1200=	400	+90	+70	+80	+80	+160	+40	+40	+160	+80

Table-5
ϕ_{ij} 's (i<j) values corresponding to split in Table 4

$$\Phi = \begin{bmatrix} 2.00 & 1.67 & 1.25 & 1.00 & 0.71 & 0.67 & 0.62 & 0.50 & 0.45 \\ & 1.71 & 1.28 & 1.02 & 0.73 & 0.68 & 0.64 & 0.51 & 0.46 \\ & & 1.38 & 1.10 & 0.79 & 0.57 & 0.70 & 0.55 & 0.50 \\ & & & 1.58 & 0.59 & 0.43 & 0.52 & 0.79 & 0.72 \\ & & & & 0.90 & 0.74 & 0.41 & 0.63 & 0.85 \\ & & & & & 0.82 & 0.83 & 0.88 & 0.99 \\ & & & & & & 0.96 & 0.97 & 1.06 \\ & & & & & & & 1.14 & 1.04 \\ & & & & & & & & 1.11 \end{bmatrix}$$

Table-6
Final split corresponding to initial split (Table 3)

Pop. Units	Sizes	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16	C17	C18	C19	C20	C21
1	40=	15										+10	+5	+5	+5							
2	50=	15	+10								+10					+5	+5					+5
3	60=	15	+10	+10					+15	+10												
4	80=	15	+10	+10	+20									+5				+10	+10			
5	100=	15	+10	+10	+20	+25							+5		+5	+5				+5	+5	
6	140=	15	+10	+10	+20	+25	+30					+10						+10		+5		+5
7	150=	15			+20	+25	+30	+10		+10	+10						+5		+10		+5	
8	160=	15	+10	+10	+20	+25	+30	+10	+15				+5		+5	+5		+10				+5
9	200=	15	+10	+10	+20	+25	+30	+10	+15	+10	+10	+10		+5	+5		+5		+10	+5	+5	
10	220=	15	+10	+10	+20	+25	+30	+10	+15	+10	+10	+10	+5	+5		+5	+5	+10	+10	+5	+5	+5
Total	1200=	150+	80+	70+	140+	150+	150+	40+	60+	40+	40+	40+	20+	20+	20+	20+	20+	40+	40+	20+	20+	20

Table-7
ϕ_{ij} 's (i<j) values corresponding to split in Table 6

$$\Phi = \begin{bmatrix} 0.75 & 0.62 & 0.94 & 0.75 & 0.80 & 0.75 & 0.47 & 0.94 & 0.85 \\ & 0.93 & 0.70 & 0.86 & 0.61 & 0.80 & 0.72 & 0.73 & 0.94 \\ & & 0.89 & 0.71 & 0.51 & 0.50 & 0.92 & 0.98 & 0.89 \\ & & & 0.91 & 0.91 & 0.62 & 0.80 & 0.73 & 0.84 \\ & & & & 0.95 & 0.80 & 0.83 & 0.74 & 0.80 \\ & & & & & 0.75 & 1.03 & 0.82 & 0.89 \\ & & & & & & 0.84 & 1.07 & 0.98 \\ & & & & & & & 0.87 & 0.91 \\ & & & & & & & & 1.00 \end{bmatrix}$$

Comparing the values of ϕ_{ij}'s given in table-5 and table-7 it is apparent that ϕ_{ij}'s values in latter has been narrowed down considerably. Hence we may conclude that examining the nature of ϕ_{ij}'s in initial stage may produce the desirable split in less number of trails. However the above study gives an insight to develop the systematic splitting process such that resulting ϕ_{ij}'s satisfy condition(1).

Conclusion

In the present scenario where estimation of variance is one of the most desirable properties of the estimation procedure, attempts have been made to provide the proper split in getting non-zeros π_{ij}'s and also satisfying the condition of non-negative estimation of variance.

Acknowledgement

Author is grateful to former Principal Scientist, IASRI, New Delhi for his valuable guidance to carry out this research work.

References

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

2. Hanurav T.V., Optimum Utilization of Auxiliary Information: IIPs Sampling of Two Units from a Stratum, *J.R. Statist. Soc.*, **B(29)**, 374-391, **(1967)**

3. Sukhatme P.V. and Sukhatme B.V., Sampling Theory of Surveys with Applications, Ames. Iowa: Iowa State College Press, **(1970)**