ON THE INCLUSION OF LARGE UNITS IN THE SAMPLE

By

D. Singh and R. Singh

Institute of Agricultural Research Statistics, New Delhi-12.

(Received : August, 1974)

Introduction

1.1. In situations where sampling units vary considerably in size, it is advantageous to select units with probability proportional to a given measure of size, the size measure being generally the value of some auxiliary variable closely related to the variable under study. But in some situations when the population under study is skew, it becomes a necessity to include some of the largest units in the sample with certainty to obtain precise estimate for the population total. In such situations it may be advantageous and simpler to have a sampling plan which calls for :

(a) The complete coverage of a group of units having extreme values ; and

(b) A supplementary sub-sample from the remaining population with probability proportional to size of the remaining units.

1.2. The problem was probably first considered by Dalenius (1952) as one of stratification. He considered the formation of two strata, one for complete enumeration and the other for sampling. Nanjumma et al (1955-56) studied the problem of large plots in crop surveys. Here the sampling of plots with probability proportional to area was considered for estimating the proportion of area in a village under some crop. A simple rule which was free from the knowledge of y's was derived as follows. The largest plot in a village was to be included in the sample if its area $A_1 \geqslant \frac{A}{n}$ where A is the area of the village and n is the number of plots to be included in the sample. The next largest plot is examined and included if its area $A_2 \geqslant \frac{A-A_1}{n-1}$ and so

on, till at some stage

$A\gamma \leqslant \frac{A - \sum_{i=1}^{\gamma-1} A_i}{n-r+1}$. Thus $(r-1)$ largest plots are definitely selected and $n-r+1$ plots are selected from the remaining with probability proportional to area.

Glasser (1962) has also considered such a problem of obtaining the optimum point beyond which all the units are included in the sample definity and a sub-sample from the remaining part is selected by simple random sampling. But this expression being in terms of population mean and variance for the character under study is of little use in practical sample surveys.

To overcome this handicap, the use of an auxiliary variable has been suggested. The usefulness of the new procedure has been illustrated with the help of a numerical example. The paper also gives the range within which the optimum point is expected to lie for most of the practical sample surveys.

2. The Sampling Procedure and Method of Estimation

2.1. Consider a finite population of size N. Let y be the character under study and x be the auxiliary character known for every unit of the population having values $x_1, x_2, \ldots\ldots x_N$ when arranged in the ascending order.

Let n be the sample size to be selected. Now when we know that some of the units in the population are having extremely large values we propose to include all of them in the sample. Let the number of such units be k $(<n)$. Now select a sub-sample of size $n-k(=n')$ from the remaining $N-k(=N')$ population units with ppswr such that the sample size becomes n.

2.2. An estimate of population total from such a sampling design is given by

$$\hat{y} = \frac{1}{n'} \sum_{i=1}^{n'} \frac{y_i}{P_i} + Y_k \qquad \ldots (2.1)$$

where

$$= P_i \frac{x_i}{X'},$$

$$X' = \sum_{i=1}^{N'} x_i \ ;$$

is population total for x for N' units from which the sub-sample is selected,

$$Y_k = \sum_{N'+1}^{N} y_i ;$$

is the population total for y for k large units.

Similarly Y' and X_k may be defined.

The variance of $\hat{y}$ is given by

$$V(\hat{y}) = \frac{1}{n'}\left[X' \Sigma' \frac{y_i^2}{x_i} - Y'^2 \right] \qquad ...(2.2)$$

where Σ' denotes sum over N' units.

Also an unbiased estimate of the variance is given by

$$\hat{V}(\hat{Y}) = \frac{1}{n'(n'-1)}\left[\sum_{i=1}^{n'} \frac{y_i^2}{P_i^2} - n' \hat{y}'^2 \right] \qquad ...(2.3)$$

where $\hat{y}' = \frac{1}{n'} \sum{}' \frac{y_i}{P_i}$ is the estimate of Y'.

3. The Optimum Point

3.1. Our object in this section is to obtain the optimum value of k which minimizes $V(\hat{y})$. This can be obtained if $V(\hat{y}_{n'})$ the variance of the estimate based on n' units, is less than $V(\hat{y})$ based on both $n'+1$ and $n'-1$ units ; respectively, which implies

$$V(\hat{y}_{n'+1}) - V(\hat{y}_{n'}) \geqslant 0 \qquad ...(3.1)$$

and

$$V(\hat{y}_{n'-1}) - V(\hat{y}_{n'}) \geqslant 0 \qquad ...(3.2)$$

Now we have $V(\hat{y}_{n'}) = \frac{1}{n'}\left[X' \Sigma' \frac{y_i^2}{x_i} - Y'^2 \right]$, so the variance of $\hat{y}$ based on $n'+1$ and $n'-1$ units may be written as ;

$$V(\hat{y}_{n'+1}) = \frac{X'+x_{N'+1}}{n'+1}\left(\Sigma' \frac{y_i^2}{x_i} + \frac{y^2_{N'+1}}{x_{N'+1}} \right) - \frac{(Y'+y_{N'+1})^2}{n'+1}$$

$$= \frac{1}{n'+1}\left[X' \Sigma' \frac{y_i^2}{x_i} + x_{N'+1} \Sigma' \frac{y_i^2}{x_i} + \frac{X' y^2_{N'+1}}{x_{N'+1}} - Y'^2 - 2Y' y_{N'+1} \right] \qquad ...(3.3)$$

and

$$V\left(\hat{\bar{y}}'_{n-1}\right)=\frac{1}{n'-1}\left[X'\,\Sigma'\frac{y_i^2}{x_i}-x_{N'}\,\Sigma'\frac{y_i^2}{x}-\frac{X'\,y_{N'}^2}{x_{N'}}-Y'^2+2Y'y_{N'}\right] \qquad ...(3.4)$$

From (3·3) and (3.4) we may see after some simplifications that the optimum value of n' lies in the interval

$$\frac{X'\,\Sigma'\frac{y_i^2}{x_i}-Y'^2}{\frac{X'\,y^2_{N+1}}{X'_{N+1}}+X_{N+1}\Sigma'\frac{y_i^2}{x_i}-2Y'y_{N'+1}}\leqslant n'\leqslant\frac{X'\,\Sigma'\frac{y_i^2}{x_i}-Y'^2}{\frac{X'\,y^2_{N'}}{x_{N'}}+x_{N'}\,\Sigma\frac{y_i^2}{x_i}-2Y'y_{N'}} \qquad ...(3.5)$$

3.2. Now we see that (3.5) involves the study variable y. So to avoid this difficulty and to express (3.5) in terms of the auxiliary variable x only; we assume that the finite population, under study is a sub-sample from an infinite super population, for which the following model, considered by Des Raj (1958) and others, is assumed to hold true.

$$\left.\begin{aligned} &y_i=\beta\,x_i+e_i \\ \text{where}\quad &E(e_i/x_i)=0 \\ &V(e_i/x_i)=a\,x_i^g \qquad g\geqq 0 \\ &\qquad\qquad\qquad a>0 \end{aligned}\right\} \qquad ...(3.6)$$

Now under the model (3.6) the inequalities (3.1) and (3.2) for obtaining the optimum value of n' may be written as

$$E\,V\,(\hat{\bar{y}}_{n'-1})-EV\,(\hat{\bar{y}}_{n'})\geqslant 0 \qquad ...(3.1')$$

and

$$E\,V\,(\hat{\bar{y}}_{n'-1})-E\,V\,(\hat{\bar{y}}_{n'})\geqslant 0 \qquad ...(3.2')$$

Now taking conditional expectations we obtain

$$E\,V\,(\hat{\bar{y}}_{n'})=\frac{a}{n'}\,(X'\,\Sigma'\,x_i^{g-1}-\Sigma'\,x_i^g)$$

$$E\,V\,(\hat{\bar{y}}_{n+1})=\frac{a}{n'+1}\left[X'\,\Sigma\,x_i^{g-1}+x_{N'+1}\,\Sigma\,x_i^{g-1}+X'\,x_{N'+1}^{g-1}-\Sigma\,x_i^g\right] \qquad ...(3.3')$$

$$\text{and } E\,V\,(\hat{\bar{y}}_{n-1})=\frac{a}{n'-1}\left[X'\,\Sigma\,x_i^{g-1}-x_{N'}\Sigma\,x_i^{g-1}-X'\,x_{N'}^{g-1}-\Sigma\,x_i^{g}+2\,x_{N'}^{g}\right] \quad \ldots(3.4)$$

Now using these results in inequalities (3.1′) and (3.2′) we obtain as earlier after some simplifications that the optimum value of n' lies in the interval

$$\frac{X'\,\Sigma'\,x_i^{g-1}-\Sigma'\,x_i^{g}}{X'\,x_{N'+1}^{g-1}+x_{N'+1}\,\Sigma'\,x^{g-1}} \leqslant n' \leqslant \frac{X'\,\Sigma'\,x_i^{g-1}-\Sigma'\,x_i^{g}}{X'\,x_N^{g-1}\,\Sigma'\,x_i^{g-1}-2\,x_{N'}^{g}} \quad \ldots(3.7)$$

3.3. From (3.7) we see that an advance value of k may be obtained only if we have knowledge of g value for the population under study. Thus any estimate of g value obtained from some earlier surveys or other sources may be utilised to great advantage. In the absence of any knowledge of g value it is possible to have approximately an upper limit of k since for most of the survey populations g value is expected to lie between 0 and 2. Thus an upper limit for k (*i.e.* a lower limit for n') is given by (3.7) for $g=2$ as

$$\frac{X'^2-\Sigma'\,x_i^2}{2\,X'\,x_{N'+1}} \leqslant n' \quad \ldots(3.8)$$

and a lower limit for k is obtained from (3.7) for $g=0$ as

$$n' \leqslant \frac{x_{N'}N'\,(X'\,H'-1)}{X'+\,x_{N'}^2\,N'\,H'-2x_{N'}} \quad \ldots(3.9)$$

where H' is the harmonic mean given by

$$\frac{1}{H'}=\frac{1}{N'}\left[\frac{1}{x_1}+\frac{1}{x_2}+\ldots\ldots\ldots+\frac{1}{x_{N'}}\right]$$

because for the values of g lying between 0 to 2. *LHS* of (3.7) is least for $g=2$ and the *RHS* of (3.7) is maximum for $g=0$. Thus from (3.8) and (3.9) we see that n' will generally lie in the interval

$$\frac{X'^2-\Sigma'\,x_i^2}{2\,X'\,x_{N+1}} \leqslant n' \leqslant \frac{x_{N'}\,N'\,(X'\,H'-1)}{X'+x_{N'}^2N'\,H-2\,x_{N'}} \quad \ldots(3.10)$$

4. Illustrative Example

To show the usefulness of the suggested sampling procedure, we have considered the following example which shows that the

inclusion of some of the largest units in the sample provides far more efficient estimates than those obtained from unrestricted sampling.

Example : Horvitz-Thompson (1952) gave the following data of eye estimated number of households (x) and the actual number of households (y) in an area containing 20 Blocks. Units after arranging in ascending order according to x are given below :

i	1	2	3	4	5	6	7	8	9	10
x	9	12	12	13	14	15	17	18	19	19
y	9	12	12	12	12	14	14	18	18	19
i	11	12	13	14	15	16	17	18	19	20
x	20	21	22	25	25	27	27	35	37	47
y	17	24	25	21	26	23	27	24	40	30

Now consider the selection of samples of size varying from 2 to 10 for all possible values of k. The variances of the estimates of population total obtained for different sample sizes are given in Table 4.1. Also the upper and lower limits of k $(=n\text{-}n')$ as obtained from (3.10) are given in the table 4.2. From the Table 4.1 it is seen that the efficiency of the estimate increases, as more and more units up to a certain point, called the optimum point, are included in the sample. The units included beyond that point do not add to the efficiency a all. We can see from here that the efficiency of the estimate based on a sample of size 10 selected such that 3 large units are included in the sample and 7 others are selected with ppswr from the remaining 17 units is about 250 per cent more as compared to the estimate based on a sample of size 10 selected with ppswr from all the 20 units. From Table (4.2) we see that the optimum value of k as obtained in the Table (4.1) is contained in the interval as obtained in the the Table (4.2) for all sample sizes greater than 6 but for sample sizes less than 6 the two values do not agree. The reason for this is not far to seek since (3.10) has beed obtained by assuming the super population model (3.6) and for the range of g from 0 to 2 and it is known that the super population models results need not hold true for finite populations.

SUMMARY

In the present paper a new sampling procedure which consists in including some of the large units with certainty and selecting the remaining sub-sample with ppswr from the remaining population has been discussed. This scheme is always superior to the ppswr sampling scheme and is simpler to operate. Practical example has also been given to illustrate the efficiency of this scheme.

TABLE 4.1

Variance of the estimate of population total for example I for different values of n and k

n \ *k*	*0*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
10	495.07	323.52	295.04	200.63*	211.13	214.22	218.02	251.37	271.03	—
9	550.07	363.96	337.19	234.07*	253.36	267.78	290.70	375.05	—	
8	618.83	415.93	393.39	280.88*	316.70	357.04	436.52	—		
7	707.24	485.29	472.07	351.10*	422.27	553.55	—			
6	825.11	582.35	590.08	468.14*	633.40	—				
5	990.13	727.93	786.78	702.20*	—					
4	1237.67	970.57*	1180.17	—						
3	1650.23	1455.86	—							
2	2475.35	—								

* Denotes the value of k for which the variance is minimum

TABLE 4.2

Values of n′ as obtained from (3.10) for different values of n

n	*5*	*14*	*13*	*12*	*11*	*10*	*9*	*8*	*7*	*6*	*5*	*4*
Upper limit of k	12	10	9	8	7	5	4	3	3	1	1	1
lower limit of k	5	3	3	3	1	1	1	1	1			
n' lies between	3—10	4—11	4—10	4—9	4—10	5—9	5—8	5—7	4—6	5	4	3

ACKNOWLEDGEMENT

The authors are greatly thankful to the referee for his valuable comments for improving the quality of the paper.

REFERENCES

[1] Dalenius (1952) : "The problem of optimum stratification in a special type of design" Sk. and Akt 35.

[2] Des Raj (1958) : "On the accuracy of some sampling techniques" JASA V. 53, 98-101.

[3] Glasser, G. J. (1962) : "On the complete coverage of large units in a statistical study. Rev. International Stat. Instt. V. 30, 28-32.

[4] Nanjamma, N S., Murthy, M.N. and Sethi, V.K. (1955-56) : "On the problem of large plots in crop surveys" paper presented to the review committee of NSS.