ON THE INBETWEEN MODIFICATION OF SIZE MEASURE IN PPSWR SAMPLING

RAVINDRA SINGH AND J.P. GUPTA
Punjab Agricultural University, Ludhiana

(*Received in April, 1971*)

INTRODUCTION

PPSWR (Probability proportional to size—with replacement) sampling is one of the several procedures that are used to get more precise estimates of population total or mean for the estimation variable Y, when information on a highly positively correlated auxiliary variable X is available (Sukhatme and Sukhatme 1970). If the regression of Y on X in the population is of the form $y=a+bx$ where a and $b>0$ are constants, then the efficiency of PPSWR sampling with respect to simple random sampling is high when a is either zero or is nearly so. The efficiency decreases with the increase in the value of $|a|$, although the correlation coefficient remains positive and high (Des Raj : 1958). Then there are cases where the relationship between the variables Y and X is not linear and assumes various different forms. In all such cases usually the PPSWR sampling is not expected to be efficient. This scheme is, therefore, recommended only in those cases where Y values for different population units are approximately proportional to the corresponding X values.

The purpose of the present paper is to suggest a procedure which enables us to use PPSWR sampling efficiently in all cases where the regression of Y on X is of the type $y=\phi(x)$, $\phi(x)$ being some function of x. It is proposed to treat $\phi(x)$ as the new auxiliary variable, although it may not always be possible to ascribe a physieal meaning to it. The regression of Y on $\phi(x)$ will be linear through origin. The PPSWR sampling with $\phi(x)$ as the measure of size can, therefore, be expected to be quite efficient.

If the form of the function $\phi(x)$ is known in advance alongwith the X-values for all the units in the population, the value of $\phi(x)$ can be easily calculated for all population units. A PPSWR sample, with $\phi(x)$ as the size measure, can then be selected and the usual estimation procedure followed.

For the cases where the form of the function $\phi(x)$ is not known, we propose the following sampling scheme.

2. The Proposed Sampling Scheme

Let the finite population Ω consist of N units U_1, U_2, ..., UN. Then in place of drawing a sample of size n straightaway from Ω, we propose to draw an initial sample of size m $(m<n)$ with PPSWR taking $[\rho_i]$ as the selection probabilities. Both the variables Y and X are then observed on these m units. Assuming that we have a rule R which enables us to arrive at an unique function $\phi(x)$ from a given initial sample, let $y=\phi(x, r)$ denote the regression of Y on X obtained from the r th $(r=1, 2, ..., N^m)$ initial sample $S(r, m)$. The initial sample is then augmented by a second sample of $(n-m)$ drawn with PPSWR scheme from $\Omega-S'(r, m)$ where $S'(r, m)$ is the set of distinct units in $S(r, m)$. For selecting the second sample $\phi(x, r)$ is taken as the measure of size.

If the selection probabilities $[\rho_i]$ are taken to be proportional to X values, the use of this method amounts to the inbetween modification of the size measure.

3. Estimate of Population Total

We consider the following estimate of population total Y.

$$\hat{y}=W_1 \hat{y}_1+W_2 \hat{y}_2 \qquad \dots \quad (3.1)$$

where

$$\hat{y}_1=m^{-1} \sum_i \left(y_j-\phi(x_i, r) \right) \rho_i+\sum_{\Omega}\phi(x_i, r),$$

$$\hat{y}_2=(n-m)^{-1} \sum_i yi/p_i'(r)+y'm,$$

$$p'_i(r)=\phi(x_i, r)/ \sum_{\Omega-S'(r, m)} \phi(x_i, r),$$

$$y'm=\sum_{S'(r, m)} y_i,$$

and $W_1+W_2=1$.

Although the estimate $\hat{y}_2$ is unbiased for the total Y, the estimate $\hat{y}$ is biased as it involves the regression estimate $\hat{y}_1$.

Out of N^m possible initial samples of size m, $(N-1)^m$ will not include any particular unit. The probability $p_i'(r)$ is defined only when $S\,(r, m) \not\supset U_i$. Let us define the sets of samples.

$$S_i=\{S(r, m)/S(r, m) \supseteq U_i\}$$

and $$\overline{S_i}=\{S(r, m)/S(r, m) \not\supset U_i\}$$

Following theorem then gives an approximate expression for the variance of the estimate $\hat{y}$.

Theorem 3.1. : The variance $V(\hat{Y})$ of the estimate $\hat{Y}$ is approximately given by

$$V(\hat{Y})=W_1^2\,\sigma_z^2\,(1-\rho^2)/m+W_2^2 \sum_{i\neq j\varepsilon\Omega} \gamma_{ij}\, E_1 p_i(r)p_j\,(r)\,\frac{(y_i/p_i(r)-y_j/p_j(r))^2}{2(n-m)} \qquad \dots \quad (3.2)$$

where $\sigma_z^2=\sum\limits_{\Omega}\; y_i^2/\rho_i-y^2$,

$$\rho=(\sum_{\Omega} y_i x_i/\rho_i-xy)/\surd[(\sum_{\Omega} y_i^2/\rho_i-y^2)\,(\sum_{\Omega} x_i^2/\rho_i--x^2)],$$

$$\gamma_{ij}=(1-\rho_i-\rho_j)^m,$$

$$p_i(r)=\phi(x_i, r)\,\sum_{\Omega}\phi\,(x_{i,r}),$$

and

$$E_1=\text{Expected value over all possible } S(r, m).$$

Proof :— The variance

$$V(\hat{y})=W_1^2V(\hat{y}_1)+2W_1\,W_2\,\text{Cov}\,(\hat{y}_1, \hat{y}_2)+W_2^2V(\hat{y}_2) \qquad \dots \quad (3.3)$$

Now since y_1 is the regression estimate, we have approximately

$$V(\hat{y}_1)=\sigma^2_z\,(1-\rho^2/m) \qquad \dots \quad (3.4)$$

where σ_z^2 and ρ are as defined in (3.2).

Let E_2, Cov.$_2$, V_2 and E_1 Cov.$_1$, V_1 stand for the expectation, covariance and variance for a given $S\,(r, m)$ and over all possible $S(r, m)$ respectively. Then

$$\text{Cov}\,(\hat{y}_1, \hat{y}_2)=E_1\,\text{Cov}_2\,(\hat{y}_1, \hat{y}_2)+\text{Cov}_1\,[E_2(\hat{y}_1),\, E_2(\hat{y}_2)]=0. \qquad \dots \quad (3.5)$$

Also

$$V(\hat{y}_2)=E_1V_2(\hat{y}_2)+V_1E_2(\hat{y}_2)$$

$$=(n-m)^{-1}\,E_1[\sum_{\Omega-S'(r,\,m)} y_i^2/p_i'(r)-(y-y\,m)^2]$$

$$=(n-m)^{-1}E_1[\sum_{\Omega} t_i\,p_i(r)\sum_{\Omega} t_i\,y_i^2/p_i(r)-(\sum_{\Omega} t_i\,y_i)^2]$$

$$=E_1\Big[\sum_{i\neq j\varepsilon\Omega} t_i\,t_j\,p_i(r)p_j(r)\Big(y_i/p_i(r)-y_j/p_j(r)\Big)^2\Big]/2(n-m) \qquad \ldots\ (3.6)$$

where $t_i=1$ if $V_i\varepsilon\Omega-S'\ (r,\,m)$

$=0$ otherwise.

It is easy to see that

$$\left.\begin{aligned}&E(t_i)=E(t_i^2)=(1-\rho_i)^m \text{ and for } i\neq j\\&E(t_it_j)=(1-\rho_i-\rho_j)^m=\gamma_{ij}\end{aligned}\right\} \qquad \ldots\ (3.7)$$

Now from (3.6 and 3.7), we get on taking the epxectation first over a fixed set $\{p_i(r)\}$ and then over all values taken by $\{p_i(r)\}$,

$$V(\hat{y}_2)=\sum_{i\neq j\varepsilon\Omega}\gamma_{ij}E_1\Big\{p_i(r)p_j(r\}\Big(y_i/p_i(r)-y_j/p_j(r)\Big)^2\Big\}/2(n-m)\ldots(3.8)$$

From (3.3), (3.4), (3.5) and (3.8) we get the variance of the estimate $\hat{Y}$ as given in (3.2).

4. Estimation of Variance V($\hat{Y}$)

The estimation of Variance $V(\hat{Y})$ amounts to finding estimates of the variances $V(\hat{Y}_1)$ and $V\ (\hat{Y}_2)$. Let $e_i(r)=y_i-\phi(x_i,\,r)$, then as usual we take

$$\hat{V}(\hat{y}_1)=[\sum_{i=1}^{m} e_i^2(r)/\rho_i^2-m^{-1}(\sum_{i=1}^{m} e_i(r)/\rho_i)^2]/m(m-1) \qquad \ldots\ (4.1)$$

as an estimate of $V\ (\hat{y}_1)$.

Coming to the estimation of variance $V(\hat{y}_2)$, if the coefficient of $W_2^2/(n--m)$ in (3.2) is denoted by $\sigma_z^2(H)$, then it can be easily seen that an unbiased estimate of $\sigma_z^2(H)$ is given by $\hat{\sigma}_z^2(H)$ where

$$\hat{\sigma}_z^2(H)=(n-m-1)^{-1}\left[\sum_i y_i^2/p_i'^2(r)-(n-m)^{-1}\{\sum_i y_i/p'_i(r)\}^2\right] \quad \dots (4.2)$$

where $p'_i(r)$ is as defined in (3.1). Therefore, we have

$$\hat{V}(\hat{y}_2)=\hat{\sigma}_z^2(H)/(n-m). \quad \dots (4.3)$$

It may be remarked here that the estimate of $V(\hat{y}_1)$ has been obtained from the preliminary sample of size m while the second sample selected in any particular case has been used for estimating the variance $V(\hat{y}_2)$. One can easily see that various other estimates can also be obtained for these variances, but it is felt that they may be important only when it is desired to investigate into the accuracy of different estimators. In the present paper, this aspect of the problem is not being considered.

5. Some Further Remarks

1. The optimum values of the weights W_1 and W_2 in 3.1 are, as usual given by

$$W_1=\frac{V(\hat{y}_2)}{V(\hat{y}_1)+V(\hat{y}_2)} \text{ and } W_2=1-W_1 \quad \dots (5.1)$$

If we have some idea about the magnitude of $V(\hat{y}_1)$ and $V(\hat{y}_2)$ these values can be us϶d in (5.1) to get the weights W_1 and W_2. In case no such information is available we have to use in (5.1) the estimated variances. The values of W_1 and W_2 so obtained will, however, become random variables.

2. The second sample of size $(n-m)$ could also be selected from the whole population Ω in place of selecting it from $\Omega-S'(r, m)$. In such a situation we will have

$$\hat{y}_2'=(n-m)^{-1}\sum_i y_i/p_i(r), \quad \dots (5.2)$$

as an unbiased estimate of the total Y. The sampling procedure suggested in the present paper is, however, more efficient since we have

$$2(n-m)[\hat{V}(y_2')-\hat{V}(y_2)]=E_1\left[\sum_{i\neq j\varepsilon\Omega}(1-\gamma_{ij})\, p_i(r)p_j(r)\left(\frac{y_i}{p_i(r)}-\frac{y_j}{p_j(r)}\right)^2\right]$$

$$>0 \quad \ldots \quad (5.3)$$

as $\gamma_{ij}=(1-\rho_i-\rho_j)^m<1.$

3. As m enters in the definition of γ_{ij}, the determination of its optimum value is quite complicated and is not attempted here. Even if the expression for optimum m is available it will involve the function $\phi(x)$ which we know only after the selection of the initial sample. Therefore, the availability of the formula for optimum m will not be of much use.

REFERENCES

1. Raj, Des. (1958) : On the relative accuracy of some sampling techniques. Jour. Amer. State. Assoc. 53, 98-101.

2. Sukhatme, P.V. and Sukhatme, B.V. (1970) : Sampling theory of surveys with applications. Asia Publishing House, New Delhi.