Sankhyā The Indian Journal of Statistics
1982, Volume 44, Series B, Pt. 1, pp. 92–101

ON UNBIASED VARIANCE-ESTIMATION WITH VARIOUS MULTI-STAGE SAMPLING STRATEGIES

By ARIJIT CHAUDHURI and RAGHUNATH ARNAB
Indian Statistical Institute

SUMMARY. Durbin's (1953, 1967) and Des Raj's (1968) results are generalized to show that a general class of linear homogeneous unbiased estimators for a finite population total based on arbitrary sampling schemes in multistages selecting the first stage units (f.s.u) with replacement admits an unbiased estimator for its variance in terms of homogeneous quadratic functions of the estimators (based on sampling in stages following the first) of the f.s.u totals. In case where the f.s.u's are chosen without replacement or the estimators are based on distinct f.s.u's alone selected with replacement it is also shown that such variance-estimators are available only if the variances of the estimators for the f.s.u totals based on sampling in subsequent stages possess unbiased estimators.

1. Introduction

In multi-stage sampling it is well-known (Durbin 1953, 1967; Des Raj 1968; Cochran 1963; Stuart 1963) that some particular linear unbiased estimators for a finite population total admit unbiased variance-estimators which do or do not involve unbiased estimators (based on sampling at subsequent stages) of the variances of the estimators of the f.s.u. totals according as the f.s.u.'s are selected without or with replacement respectively, provided the f.s.u.'s are selected according to the method of sampling with probability proportional to size (pps) in the latter case. In this paper we extend these results to accommodate more general situations where the estimators belong to classes wider than those so far considered in the literature and in case of sampling with replacement the f.s.u.'s are selected not necessarily following the pps method. In this connection one may note that the results presented by Stuart (1963) are not strictly valid in the generality in which they are apparently claimed because of some obvious algebraic mistakes committed in the paper.

We believe this generalization has a theoretical necessity in view of the recent growth in the literature concerning general classes of estimators in uni-stage sampling and we show that the earlier results in the area are covered as special cases.

Keywords : Finite popultaion ; multi-stage sampling with arbitrary probabilities with and without replacement ; general linear unbiased estimators ; variance-estimation.

2. Sampling strategies

For a population of N first-stage units (f.s.u.) let us have a scheme of sampling where on the r-th draw the i-th f.s.u. is selected with a probability $p_i(r)$, draws being made with replacement ($r = 1, 2, \ldots, n$; $i = 1, \ldots, N$) such that $\sum_{i=1}^{N} p_i(r) = 1$ for every r and every first stage unit so selected is sub-sampled, as it is drawn, in subsequent stages employing arbitrary selection-probabilities. Denoting by Y_i, the i-th f.s.u. total, the customary problem is to unbiasedly estimate $Y = \sum_{1}^{N} Y_i$, the population total, on the basis of a sample so drawn. Here we propose to employ a general linear estimator of the following form namely

$$T = T(s) = \sum_{i=1}^{N} \sum_{r=1}^{n} b_{sir} T_i(r)$$

where

$$b_{sir} = 0 \quad \text{if the } r\text{-th draw does not yield the } i\text{-th f.s.u. of the sample } s;$$

$$b_{sir} \neq 0 \quad \text{otherwise.}$$

such that $T_i(r)$ is a statistic calculated on the basis of the sample values of the variable y under study defined in respect of the sampling units of the ultimate stage sub-sampled from the i-th f.s.u. when it is selected on the r-th draw such that with respect to the sampling design adopted in subsequent stages $T_i(r)$ is an unbiased estimator for Y_i for each r having a variance $V(T_i(r)) = \sigma_{ir}^2$ ($r = 1, \ldots, n$; $i = 1, \ldots, N$) (these expectations and variances relate to the sampling designs adopted in sub-sampling the selected f.s.u.'s).

Also b_{sir}'s are independent of y-values and are so chosen that T is an unbiased estimator for Y. The strategy so described will be called the strategy A. An alternative to be denoted as strategy B that we propose is one where a sample is chosen in the same manner as above but the linear unbiased estimator we consider is based only on the distinct f.s.u.'s included in the selected sample and is of the following form, namely,

$$T_1 = T_1(s) = \sum_{i \epsilon s} c_{si} T_i^*$$

where T_i^* is a function of the sample-values of y for the ultimate stage units sub-sampled from the i-th f.s.u. whenever it is chosen such that with respect to the sampling designs in subsequent stages it is unbiased for Y_i for every i and c_{si}'s are so chosen that T_1 is unbiased for Y, the symbol $\sum_{i \epsilon s}$ denoting summation over the distinct f.s.u.'s only in the sample s. Here it is supposed that from the consideration of sufficiency once the f.s.u.'s are selected with replacement we retain the distinct f.s.u.'s alone and discard the information about which draws produced which f.s.u.'s and each selected f.s.u. is subsampled just once in independent manners.

Denoting by (E_1, E_L), (V_1, V_L) and (C_1, C_L) the operators for mathematical expectations, variances and covariances in 1st and subsequent stages of sampling respectively, and by E and V the corresponding operators for the over-all sampling scheme we have

$$E(T) = E_1\Big[\sum_{i=1}^{N}\sum_{r=1}^{n} b_{sir}E_L(T_i(r))\Big]$$

$$= E_1\Big[\sum_{i=1}^{N}\sum_{r=1}^{n} b_{sir}Y_i\Big]$$

$$= E_1\Big[\sum_{i=1}^{N} Y_i\, d_{is}\Big], \text{ on writing } d_{is} = \sum_{r=1}^{n} b_{sir}$$

$$= \sum_{i=1}^{N} Y_i \sum_{s} d_{is}\, P(s),$$

$P(s)$ denoting the probability of selecting a sample s (say) of f.s.u.'s in n draws as in strategy A. Obviously, because of the unbiasedness condition we require that the d_{is}'s satisfy the condition

$$\sum_{s} d_{is}P(s) = 1 \;\forall\; i = 1, \ldots, N. \qquad \ldots \quad (2.1)$$

Assuming throughout that this is satisfied we have

$$V(T) = E_1\Big[\sum_{i=1}^{N}\Big(\sum_{r=1}^{n} b_{sir}^2\sigma_{ir}^2\Big)\Big] + V_1\Big[\sum_{i=1}^{N} Y_i\, d_{is}\Big]$$

(remembering that $T_i(r)$ and $T_j(r')$ are uncorrelated whenever either $i \neq j$ or $r \neq r'$ or both)

$$= \sum_{s}\Big(\sum_{i=1}^{N}\sum_{r=1}^{n} b_{sir}^2\sigma_{ir}^2\Big)P(s) + \sum_{1}^{N} Y_i^2$$

$$\times\Big(\sum_{s} d_{is}^2P(s) - 1\Big) + \sum_{i \neq j}^{N}\sum^{N}\Big(\sum_{s} d_{is}\, d_{js}P(s) - 1\Big)$$

$$= \sum_{i=1}^{N}\sum_{r=1}^{n} \sigma_{ir}^2\,\alpha_{ir} + \sum_{i}^{N} Y_i^2\beta_i + \sum_{i \neq j}^{N}\sum^{N} Y_iY_j\gamma_{ij}$$

on writing $\alpha_{ir} = \sum_{s} b_{sir}^2 P(s)$

$$\beta_i = \sum_{s} d_{is}^2 P(s) - 1, \qquad i = 1, 2, \ldots, N$$

$$\gamma_{ij} = \sum_{s} d_{is} d_{js} P(s) - 1, \quad i, j = 1, \ldots, N \ (i \neq j).$$

Our problem is to estimate $V(T)$ unbiasedly from the sampled data such that the resulting estimator does not involve any unbiased estimator of $\sigma_{i_r}^2$'s, but a function of $T_i(r)$'s alone. At this stage let us note the special case for which a simple well-known solution exists as described below. Suppose the f.s.u.'s are selected in n draws with replacement with suitable selection probabilities and the selected f.s.u.'s are independently sub-sampled such that for every r-th draw of an f.s.u., t_r is an unbiased estimator for Y for each r such that t_r's are independently distributed. Then T becomes

$$T = \frac{1}{n} \sum_{r} t_r$$

and ... (2.2)

$$E(T) = \frac{1}{n} \sum_{r} E(t_r) = Y$$

and it follows that $\frac{1}{n(n-1)} \sum_{r} (t_r - T)^2$ is an unbiased estimator for $V(T)$ because

$$E \sum_{r} (t_r - T)^2 = E \sum_{r} [(t_r - Y) - (T - Y)]^2$$

$$= \sum_{r} E(t_r - Y)^2 + nE(T - Y)^2 - \frac{2}{n} E(t_r - Y)^2$$

$$= \sum_{r} \sigma_r^2 + \frac{1}{n} \sum_{r} \sigma_r^2 - \frac{2}{n} \sum_{r} \sigma_r^2$$

[where $\sigma_r^2 = E(t_r - Y)^2$]

$$= \frac{n-1}{n} \sum_{r} \sigma_r^2$$

and

$$V(T) = E(T - Y)^2 = E\left[\frac{1}{n} \sum_{r} (t_r - Y)\right]^2$$

$$= \frac{1}{n^2} \sum_{r} \sigma_r^2$$

so that

$$E\,\frac{1}{n(n-1)}\sum_{r}(t_r-T)^2=\frac{1}{n^2}\sum_{r}\sigma_r^2=V(T).$$

However, we shall consider our linear unbiased estimator T so general that b_{sir} is subject to (2.1).

So, in getting an unbiased estimator for $V(T)$ in terms of $T_i(r)$'s we shall seek one within the following class of homogeneous quadratic functions of $T_i(r)$'s as

$$t_1=\sum_{r=1}^{n}C_r\,T_r^2+\sum_{r\neq r'=1}^{n}\sum^{n}C'_{rr'}\,T_rT_{r'} \qquad \ldots \quad (2.3)$$

where $C_r=g_{ir}$ if r-th draw produces the i-th f.s.u. with probability $p_i(r)$ in which case $T_r=T_i(r)$, $C'_{rr}=e_{ii}$ if r-th as well as r'-th draws produce the i-th f.s.u., the probability of which is $p_i(r)p_i(r')$ and $C'_{rr}=f_{ij}$ if r-th draw produces i-th f.s.u. and r'-th draw produces j-th f.s.u., the probability for this being $p_i(r)\,p_j(r')$.

Now

$$E(t_1)=\sum_{r=1}^{n}\sum_{i=1}^{N}g_{ir}(Y_i^2+\sigma_{ir}^2)\,p_i(r)$$

$$+\sum_{r=1}^{n}\sum_{\substack{r=1\\ r\neq r'}}^{n}\sum_{i=1}^{N}e_{ii}Y_i^2p_i(r)\,p_i(r')$$

$$+\sum_{r=1}^{n}\sum_{\substack{r=1\\ r\neq r'}}^{n}\sum_{i=1}^{N}\sum_{\substack{j=1\\ j\neq j}}^{N}f_{ij}Y_iY_jp_i(r)p_i(r')$$

$$=\sum_{i=1}^{N}\sum_{r,=1}^{n}\sigma_{ir}^2g_{ir}p_i(r)+\sum_{i=1}^{N}Y_i^2$$

$$\left\{\sum_{r=1}^{n}g_{ir}p_i(r)+e_{ii}\sum_{r\neq r'=1}^{n}\sum^{n}p_{i(r)}p_{i(r')}\right\}$$

$$+\sum_{i=1}^{N}\sum_{\substack{j=1\\ i\neq j}}^{N}Y_iY_jf_{ij}\sum_{r\neq r'=1}^{n}\sum^{n}p_i(r)p_j(r').$$

Now, if we choose

$$g_{ir} = \frac{\alpha_{ir}}{p_i(r)} \quad \forall\ i, r,\ e_{ii} = \frac{\beta_i - \sum\limits_{r=1}^{n} \alpha_{ir}}{\sum\limits_{r \neq r'=1}^{n} \sum\limits^{n} p_i(r) p_i(r')} \quad \forall\ i,$$

and

$$f_{ij} = \frac{\gamma_{ij}}{\sum\limits_{r \neq r'=1}^{n} \sum\limits^{n} p_i(r) p_j(r')} \ \forall\ i, j\ (i \neq j),$$

then t_1 becomes unbiased for $V(T)$ and such choices are clearly possible. We may cite as examples the following :

Example 1 : Bandyopadhyay, Chottopadhyay and Kundu (1977) considered the estimator $e = \sum\limits_{i=1}^{N} \frac{n_i(s)}{\alpha_i(s)} \frac{Y_i}{P(s)}$ for Y in case of a uni-stage sampling scheme to choose a sample s with a probability $P(s)$ with or without replacement such that

$$\sum_{s} \frac{n_i(s)}{\alpha_i(s)} = 1 \ \forall\ i = 1, \ldots, N \qquad \ldots \quad (2.4)$$

where

$n_i(s)$ = frequency of i-th f.s.u. in the sample s of f.s.u.'s.

In the multi-stage sampling an estimator of this general type should be chosen as

$$T(s) = \sum_{i=1}^{N} \sum_{r=1}^{n} a_i(s) \frac{T_i(r)}{\alpha_i(s) P(s)}$$

(where $a_i(s) = 1$ if i-th f.s.u. is included in s and zero otherwise),

so that

$$E(T(s)) = E_1 \Big[\sum_{1}^{N} Y_i \sum_{r=1}^{n} \frac{a_i(s)}{\alpha_i(s) P(s)} \Big]$$

$$= E_1 \Big[\sum_{i=1}^{N} Y_i \frac{n_i(s)}{\alpha_i(s)} \frac{1}{P(s)} \Big] = Y$$

(provided 2.4 is assumed).

Here clearly, the condition for unbiasedness of $T(s)$ is a condition of the type (2.1) and not of the more restrictive type as (2.2). Here an unbiased estimator for $V(T(s))$ of the form (2.3) is clearly available as one may check following the line discussed above.

Example 2 : If $p_i(r) = p_i$ for every r, then, l_1 will be determined on choosing

$$g_{ir} = \frac{\alpha_{ir}}{p_i} \forall\, i, r, e_{ii} = \frac{\beta_i - \sum\limits_{r=1}^{n} \alpha_{ir}}{n(n-1)p_i^2}$$

and

$$f_{ij} = \frac{\gamma_{ij}}{n(n-1)\, p_i p_j} \forall\, i \neq j = 1, \ldots, N.$$

If, in addition $\sigma_{ir}^2 = \sigma_i^2 \forall\, r = 1, \ldots, n$ then

$$g_{ir} = \frac{\alpha_i}{npi}, e_{ii} = \frac{\beta - \alpha_i}{n(n-1)p_i^2} \text{ and } f_{ij} = \frac{\gamma_{ij}}{n(n-1)p_i p_j}$$

where $\alpha_i = \sum\limits_{r=1}^{n} \alpha_{ir}$.

Example 3 : The situation is well-known (Durbin 1953; Des Raj 1968, where $p_i(r) = p_i \forall\, r, \alpha_i(s) = \frac{np_i}{P(s)} \forall\, i$, $T_i(r) = T_i$ (say) for each $r = 1, \ldots, n; \sigma_{ir}^2 = \sigma_i^2 \forall\, r$.).

Now, considering the strategy B we have

$$E(T_1) = E_1 \sum_{i \in s} c_{si} Y_i = \sum_{i=1}^{N} Y_i \sum_{s \ni i} C_{si} P(s).$$

So, for the sake of unbiasedness of T_1 for Y the sufficiency condition is that

$$\sum_{s \ni i} C_{si}\, P(s) = 1 \forall\, i.$$

Assuming this to be satisfied we have

$$V(T_1) = \sum_{i=1}^{N} \sigma_i^2 (\sum_{s \ni i} C_{si}^2\, P(s) + \sum_{i=1}^{N} Y_i^2 (\sum_{s \ni i} C_{si}^2 P(s) - 1)$$

$$+ \sum_{i \neq j=1}^{N} \sum^{N} Y_1 Y_j \,(\sum_{s \ni i, j} C_{si} C_{sj} P(s) - 1).$$

Now, if possible, let an unbiased estimator for $V(T_1)$ be available as a quadratic function of T_i^*'s as

$$v_1 = \sum_{i \in s} h_{si} T_i^{*2} + \sum_{i \neq j \in s} h_{sij} T_i^* T_j^*. \qquad \ldots \quad (2.5)$$

Then

$$E(v_1) = \sum_i (Y_i^2 + \sigma_i^2) \left(\sum_{s \ni i} h_{si} P(s) \right)$$

$$+ \sum_{i \neq j}\sum Y_i Y_j \sum_{s \ni ij} h_{sij} P(s);$$

and in order for it to be unbiased for $V(T_1)$ we must have

$$\sum_{s \ni i} h_{si} P(s) = \sum_{s \ni ij} C_{si}^2 P(s) \ \forall\ i$$

as well as

$$\sum_{s \ni i} h_{si} P(s) = \sum_{s \ni i} C_{si}^2 P(s) - 1 \ \forall\ i$$

simultaneously.

But this is absurd. So an unbiased estimator of $V(T_1)$ of the form (2.5) does not exist. If, however, an unbiased estimator for σ_i^2 is available as $\hat{\sigma}_i^2$ on the basis of units sub-sampled from the i-th f.s.u. in later stages when it is selected, the unbiased estimator for $V(T_1)$ is available as

$$w_1 = v_1 + \sum_{i \in s} \frac{\hat{\sigma}_i^2}{\pi_i}, \text{ where } \pi_i = \sum_{s \ni i} P(s)$$

provided $\pi_i > 0$ and we have

$$\sum_{s \ni i} h_{si} P(s) = \sum_{s \ni i} C_{si}\, p(s) - 1 \ \forall\ i$$

and

$$\sum_{s \ni ij} h_{sij} P(s) = \sum_{s \ni ij} C_{si} C_{sj} P(s) - 1 \ \forall\ i \neq j.$$

It readily follows that the result if applicable to the strategy B carries over to situations when the f.s.u.'s are selected without replacement because in these cases the general homogeneous linear unbiased estimators for Y of the form T_1 are numerous well-known estimators in the literature.

More generally, for any multistage sampling scheme let

$$e = \sum_1^N b_{si} T_{si}, \text{ (with } E_L(T_{si}) = Y_i \ \forall\ i,$$

$$V_L(T_{si}) = \sigma_{si}^2,\ C_L(T_{si}, T_{sj}) = \sigma_{sij}$$

for a sample s of f.s.u.'s,

$$b_{si} = 0 \text{ if } i \notin s,\ E_L(b_{si}) = 1 \ \forall\ i,$$

$$V_1(b_{si}) = \delta_{ii},\ C_1(b_{si}, b_{sj}) = \delta_{ij})$$

be an unbiased estimator for Y. Then

$$V = \text{var}(e) = \sum_{1}^{N} Y_i^2 \delta_{ii} + \sum\sum_{i \neq j = 1}^{N} Y_i Y_j \delta_{ij}$$

$$+ E_1 \left[\sum_{1}^{N} b_{si}^2 \sigma_{si}^2 + \sum\sum_{i \neq j = 1}^{N} b_{si} b_{sj} \sigma_{sij} \right].$$

Let

$$v = \sum_{1}^{N} C_{si} T_{si}^2 + \sum\sum_{i \neq t = 1}^{N} C_{sij} T_{si} T_{sj}$$

be an estimator for V ($C_{si}, C_{sij} = 0$ for $i \notin s$, $i, j \notin s$ respectively). Necessary conditions for its unbiasedness are

(1) $E_1(\Sigma C_{si} \sigma_{si}^2 + \Sigma\Sigma\, C_{sij} \sigma_{sij}) = E_1[\Sigma\, b_{si}^2 \sigma_{si}^2 + \Sigma\Sigma\, b_{si} b_{sj} \sigma_{sij}]$

and

(2) $V_1(b_{si}) = E_1(C_{si}) \;\forall\; i$, b_{si}, C_{si}, C_{sij} are independent of Y_i's.

In case $\sigma_{si}^2 = \sigma_i^2 \;\forall\; s \ni i$, i and $\sigma_{sij} = 0 \;\forall\; i, j$ $(i \neq j)$ and s, these two conditions cannot hold together implying non-availability of an unbiased estimator for V as a homogeneous quadratic function of T_{si}'s.

Finally let us consider the strategy C where the f.s.u.'s are selected with replacement as in strategies A and B but the homogeneous linear unbiased estimator for Y is of the form

$$T^+ = T^+(s) = \sum_{i \in s} e_{si} T_i(l)$$

which is based on the distinct f.s.u.'s (selected with replacement) only in s but here we retain the information about the particular draws l $(= 1, \ldots, \lambda_i$, say) on which the i-th f.s.u. $(i = 1, 2, \ldots, N)$ occurs in the sample s of f.s.u.'s and decide to sub-sample the i-th f.s.u. with a probability q_{il} (where $0 < q_{il} < 1$ such that $\sum_{l=1}^{\lambda_i} q_{il} = 1$, $i = 1, \ldots, N$) where it is selected on the l-th draw in which case we estimate Y_i by $T_i(l)$ on the basis of the sub-sample so drawn.

Denoting by E_1, V_2 the expectations and variances with respect to the probabilities q_{il}'s and by E_L', V_L' the conditional expectations and variances respectively with respect to the subsampling designs given the estimators $T_i(l)$ $(l = 1, 2, \ldots, \lambda_i;\; i = 1, 2, \ldots, N)$ actually chosen, we have

$$E_2(T_i(l)) = E_2(E_L' T_i(l)) = Y_i, \;\forall\; l$$

assuming $\quad E_L' T_i(l) = Y_i \;\forall\; l$

and

$$V_L(T_i(l)) = E_2[V_L'(T_i(l)]+V_2[E_L'(T_i(l))]$$

$$= E_2\sigma^2 = \sum_{l=1}^{L_i} q_{li}\sigma_l^2 \text{ for every } \begin{array}{l} l = 1, 2, \ldots, \\ i = 1, 2, \ldots, N, \end{array}$$

$$= \xi_i \text{ (say)}.$$

So

$$V(T^+) = E_1 \sum_{i\in s} b_{si}^2 V_L(T_i(l)+V_1 \sum_{i\in s} b_{si}E_1(T_i(l))$$

$$= E_1 \sum_{i\in s} b_{si}^2\xi_i+V_1 \sum_{i\in s} b_{si}Y_i$$

and in estimating $V(T^+)$ unbiasedly we encounter the circumstances similar to those obtaining in case of strategy B and as in case of sampling the f.s.u.'s without replacement.

Acknowledgement. The authors gratefully acknowledge several constructive criticisms by Professor J. Roy that were helpful in preparing this paper. They also wish to thank Dr. Shibdas Bandyopadhyay with whom they had some fruitful discussions.

References

Bandyopadhya, S., Chattopadhyay, A. K. and Kundu, S. C. (1977): On estimation of population total. *Sankhyā*, Ser. C. **39**, 28-42.

Cochran, W. G. (1963): *Sampling Techniques*, 2nd Ed. John Wiley and Sons, New York.

Durbin, J. (1953): Some results in sampling theory when the units are selected with unequal probabilities. *J. Roy. Statist. Soc.*, Ser. B, **15**, 262-269.

——— (1967): Design of multi-stage surveys for the estimation of sampling errors. *Applied Statistics*, **16**, 152-164.

Raj Des (1968): *Sampling Theory*, McGraw-Hill, New York.

Stuart, A. (1963): Some results on sampling with unequal probabilities, Proc. of the 34th sessions, *Bull. Int. Statist. Inst.* **40**, 773-780.

Paper received : October, 1977.

Revised : September, 1981.