ON SAMPLING WITH UNEQUAL PROBABILITIES

By P. K. PATHAK

Indian Statistical Institute

SUMMARY. This paper deals with the problem of deriving improved estimators in sampling schemes with unequal probabilities of selection. The improved estimator of the population total, Y, (Basu, 1958), is derived. In addition, two sets of estimators of Y and Y^2 are given. The first set of estimators is unwieldy to compute, while the second set is simple. The second set of estimators, though less efficient than the first, is more efficient than the usually employed estimators.

It is proved in subfield terminology that if $\mathcal{S}_1$ and $\mathcal{S}_2$ are two sufficient subfields and K is a set common to $\mathcal{S}_1$ and $\mathcal{S}_2$, then $\mathcal{S}_1K+\mathcal{S}_2K'$ is also a sufficient subfield. Hence the subfield $\mathcal{S}_1K+\mathcal{S}_2K'$ can be used to derive improved estimators by Rao-Blackwell theorem. Generalisation of this is also given in case of countable number of subfields. Application of this result to sampling with unequal probabilities is given.

1. Introduction

Consider a population containing N units. Let y_j be some real-valued characteristic of the j-th population unit in which we are interested.[1] Suppose that a sample of size n is drawn from the above population with unequal probabilities of selection (with replacement), P_j being the probability of selection associated with the j-th population unit ($\Sigma P_j = 1$). If for the i-th sample unit, we record its Y-characteristic y_i, probability of selection p_i and unit-index u_i, then the sample of observation is

$$S = (x_1, x_2, \ldots, x_n),$$

where $x_i = (y_i, p_i, u_i)$.[2]

It has been shown by Basu (1958) that the 'order-statistic'

$$T = (x_{(1)}, x_{(2)}, \ldots, x_{(r)})$$

(where $x_{(1)}, x_{(2)}, \ldots x_{(r)}$, are the distinct units in the sample arranged in ascending order of their unit-indices) is sufficient.

Therefore, if $g(S)$ is some estimator depending on the sample S, for any convex (downwards) loss function, an estimator uniformly better than $g(S)$ is given by $E[g(S)|T]$. In the subsequent sections this result is used to derive improved estimators of the population total and its square.

2. Estimation of the population total

The usual estimator of the population total

$$Y = \Sigma\, Y_j$$

is given by

$$\bar{z} = \frac{1}{n}\,\Sigma\, z_i, \qquad \ldots \quad (2.1)$$

where

$$z_i = \frac{y_i}{p_i}.$$

[1] j varies from 1 to N, i from 1 to n and (i) from (1) to (r) unless otherwise stated.

[2] Capital letters refer to the population and small letters to the sample.

Theorem 1 : *For any convex (downwards) loss function, an estimator uniformly better than $\bar{x}$ is given by*

$$\bar{x}_\nu = E[\bar{x}|T] = \Sigma\, C_{(i)} \frac{y_{(i)}}{p_{(i)}} \qquad \ldots \quad (2.2)$$

where $C_{(i)} = \dfrac{p_{(i)}[(p_{(1)}+\ldots+p_{(\nu)})^{n-1} - \Sigma_1^{i}(p_{(1)}+\ldots+p_{(\nu-1)})^{n-1}+\ldots(-)^{\nu-1}p_{(i)}^{n-1}]}{[(p_{(1)}+\ldots+p_{(\nu)})^{n} - \Sigma_1(p_{(1)}+\ldots+p_{(\nu-1)})^{n}+\ldots(-)^{\nu-1}\Sigma_1 p_{(i)}^{n}]} \quad \ldots \quad (2.3)$

the summations Σ_1 *and* Σ_1^{i} *stand for all combinations of p's and all combinations of p's containing* $p_{(i)}$ *(chosen out of* $p_{(1)}, p_{(2)}, \ldots, p_{(\nu)}$*) respectively.*

Proof : Obviously by Rao-Blackwell theorem, an estimator uniformly better than $\bar{x}$ is given by

$$E(\bar{x}|T) = E\left(\frac{y_1}{p_1}\middle|T\right) = \Sigma\, \frac{y_{(i)}}{p_{(i)}}\, P[x_1 = x_{(i)}|T]. \qquad \ldots \quad (2.4)$$

But
$$P[x_1 = x_{(i)}|T] = \frac{p_{(i)}\Sigma'' \dfrac{(n-1)!}{\alpha_{(1)}!\ldots\alpha_{(\nu)}!}\, p_{(1)}^{\alpha_{(1)}} \ldots p_{(\nu)}^{\alpha_{(\nu)}}}{\Sigma' \dfrac{n!}{\alpha_{(1)}!\ldots\alpha_{(\nu)}!}\, p_{(1)}^{\alpha_{(1)}} \ldots p_{(\nu)}^{\alpha_{(\nu)}}} \qquad \ldots \quad (2.5)$$

where Σ' means summation over all integral α's such that

$$\alpha_{(i)} > 0 \text{ for } i = 1, \ldots, \nu \quad \text{and} \quad \alpha_{(1)}+\alpha_{(2)}+\ldots, +\alpha_{(\nu)} = n,$$

and Σ'' means summation over all integral α's such that

$$\alpha_{(i)} \geqslant 0,\ \alpha_{(i')} > 0 \quad \text{for} \quad i' \neq i = 1, 2, \ldots, \nu$$

and
$$\alpha_{(1)}+\alpha_{(2)}+\ldots,+\alpha_{(\nu)} = (n-1).$$

It can be seen by induction over ν that

$$\Sigma' \frac{n!}{\alpha_{(1)}!\ldots\alpha_{(\nu)}!}\, p_{(1)}^{\alpha_{(1)}} \ldots p_{(\nu)}^{\alpha_{(\nu)}} = [(p_{(1)}+\ldots+p_{(\nu)})^{n} - \Sigma_1 (p_{(1)}+\ldots+p_{(\nu-1)})^{n} + \ldots(-)^{\nu-1}\, \Sigma_1\, p_{(1)}^{n}];$$

and

$$\Sigma'' \frac{(n-1)!}{\alpha_{(1)}!\ldots\alpha_{(\nu)}!}\, p_{(1)}^{\alpha_{(1)}} \ldots p_{(\nu)}^{\alpha_{(\nu)}} = [(p_{(1)}+\ldots+p_{(\nu)})^{n-1} - \Sigma_1^{i}(p_{(1)}+\ldots+p_{(\nu-1)})^{n-1} + \ldots(-)^{\nu-1}\, p_{(i)}^{n-1}]. \qquad \ldots \quad (2.6)$$

Using (2.4), (2.5) and (2.6), we get

$$\bar{x}_\nu = E(\bar{x}|T) = \Sigma\, C_{(i)} \frac{y_{(i)}}{p_{(i)}}$$

Hence the theorem is proved.

The above estimator, though better, is not very useful in large samples because of cumbersome computation of $C_{(i)}$'s. In Section 4, we shall derive a simpler estimator of Y better than s. Table 1 gives exact expressions of s_r for $n = 3$, 4 and 5. For $n = 1$ and 2, s_r and s are identical.

TABLE 1. $\tilde{s}_r$ FOR $n=3$, 4 AND 5

$n \rightarrow$ r	3	4	5
1	$\frac{y_{(1)}}{p_{(1)}}$	$\frac{y_{(1)}}{p_{(1)}}$	$\frac{y_{(1)}}{p_{(1)}}$
2	$\frac{\Sigma\,(2p_{(1)}+p_{(2)})\frac{y_{(1)}}{p_{(1)}}}{3(p_{(1)}+p_{(2)})}$	$\frac{\Sigma\left[(p_{(1)}+p_{(2)})^2-p_{(1)}^2\right]y_{(1)}}{\left[(p_{(1)}+p_{(2)})^4-p_{(1)}^4-p_{(2)}^4\right]}$	$\frac{\Sigma\left[(p_{(1)}+p_{(2)})^4-p_{(1)}^4\right]y_{(1)}}{\left[(p_{(1)}+p_{(2)})^5-p_{(1)}^5-p_{(2)}^5\right]}$
3	$\Sigma\frac{y_{(1)}}{p_{(1)}}$	$\frac{\Sigma\left[2p_{(1)}+p_{(2)}+p_{(3)}\right]\frac{y_{(1)}}{p_{(1)}}}{4(p_{(1)}+p_{(2)}+p_{(3)})}$	$\frac{\Sigma\left[12p_{(1)}(p_{(1)}+p_{(2)}+p_{(3)}) +4(p_{(2)}^3+p_{(3)}^3)+6p_{(2)}p_{(3)}\right]\frac{y_{(1)}}{p_{(1)}}}{5\left[4(p_{(1)}^3+p_{(2)}^3+p_{(3)}^3) +6(p_{(1)}p_{(2)}+p_{(2)}p_{(3)}+p_{(1)}p_{(3)})\right]}$
4	—	$\Sigma\frac{y_{(1)}}{p_{(1)}}$	$\frac{\Sigma\left[2p_{(1)}+p_{(2)}+p_{(3)}+p_{(4)}\right]\frac{y_{(1)}}{p_{(1)}}}{5\left[p_{(1)}+p_{(2)}+p_{(3)}+p_{(4)}\right]}$
5	—	—	$\Sigma\frac{y_{(1)}}{p_{(1)}}$

3. Estimation of Y^2

The problem of finding an unbiased estimator of Y^2 arises in most problems of variance estimation of estimators of Y. The usual estimator of Y^2 is

$$z_p = \frac{1}{n(n-1)} \sum_{i \neq i'=1}^{n} z_i z_{i'} . \qquad \ldots \quad (3.1)$$

Theorem 2: *For any convex (downwards) loss function, an estimator uniformly better than z_p is given by*

$$E(z_p \mid T) = \sum_{i=1}^{r} C_{(i,i)}\, z_{(i)}^2 + \sum_{i \neq i'=1}^{r} C_{(i,i')}\, z_{(i)}\, z_{(i')} \qquad \ldots \quad (3.2)$$

Remark: $\tilde{s}_r$ when $r=(n-1)$ may be expressed in a simple form as $s_{(n-1)} = \frac{1}{n}\left\{ \Sigma\frac{y_{(i)}}{p_{(i)}} + \frac{\Sigma}{\Sigma}\frac{y_{(i)}}{p_{(i)}} \right\}$

where $$C_{(i,i')} = \frac{p_{(i)}^2[(p_{(1)}+\ldots+p_{(\nu)})^{n-2} - \Sigma_1^i (p_{(1)}+\ldots+p_{(\nu-1)})^{n-2}+\ldots(-)^{\nu-1}p_{(i)}^{n-2}]}{[(p_{(1)}+\ldots+p_{(\nu)})^n - \Sigma_1(p_{(1)}+\ldots+p_{(\nu-1)})^n+\ldots(-)^{\nu-1}\Sigma_1 p_{(1)}^n]}$$

and $$C_{(i,i')} =$$

$$\frac{p_{(i)}p_{(i')}[(p_{(1)}+\ldots+p_{(\nu)})^{n-2} - \Sigma_1^{ii'}(p_{(1)}+\ldots+p_{(\nu-1)})^{n-2}+\ldots(-)^{\nu-2}(p_{(i)}+(p_{(i')})^{n-2}]}{[(p_{(1)}+\ldots+p_{(\nu)})^n - \Sigma_1(p_{(1)}+\ldots+p_{(\nu-1)})^n+\ldots(-)^{\nu-1}\Sigma_1 p_{(1)}^n]} \quad \ldots (3.3)$$

the summations Σ_1 *and* Σ_1^i *have been defined in* (2.3) *and the summation* $\Sigma_1^{ii'}$ *stands for all combinations of* p*'s containing* $p_{(i)}$ *and* $p_{(i')}$.

Proof : Obviously

$$E(z_p|T) = E\left[\frac{1}{n(n-1)}\sum_{i\neq i'=1}^{n} z_i z_{i'}|T\right] = E(z_1 z_2|T)$$

$$= \Sigma\, z_{(i)}\, z_{(i')}\, P[x_1 = x_{(i)}, x_2 = x_{(i')}|T]. \quad \ldots (3.4)$$

It is easy to see that

$$P[x_1 = x_{(i)}, x_2 = x_{(i)}|T] = \frac{p_{(i)}^2 \Sigma'' \frac{(n-2)!}{\alpha_{(1)}!\ldots\alpha_{(\nu)}!} p_{(1)}^{\alpha_{(1)}}\ldots p_{(\nu)}^{\alpha_{(\nu)}}}{\Sigma' \frac{n!}{\alpha_{(1)}!\ldots\alpha_{(\nu)}!} p_{(1)}^{\alpha_{(1)}}\ldots p_{(\nu)}^{\alpha_{(\nu)}}} \quad \ldots (3.5)$$

and $$P[x_1 = x_{(i)}, x_2 = x_{(i')}|T] = \frac{p_{(i)}p_{(i')}\Sigma''' \frac{(n-2)!}{\alpha_{(1)}!\ldots\alpha_{(\nu)}!} p_{(1)}^{\alpha_{(1)}}\ldots p_{(\nu)}^{\alpha_{(\nu)}}}{\Sigma' \frac{n!}{\alpha_{(1)}!\ldots\alpha_{(\nu)}!} p_{(1)}^{\alpha_{(1)}}\ldots p_{(\nu)}^{\alpha_{(\nu)}}}, \quad \ldots (3.5.1)$$

where Σ', Σ'' have been defined in (2.5) and Σ''' means summation over all integral α's such that

$$\alpha_{(1)}+\ldots+\alpha_{(\nu)} = (n-2),\ \alpha_{(i)} \geqslant 0,\ \alpha_{(i')} \geqslant 0 \text{ and } \alpha_{(k)} > 0 \text{ for } k \neq i \neq i' = 1, \ldots, \nu.$$

It can be proved on lines similar to (2.6) (by induction over ν) that

$$\left.\begin{aligned} &P[x_{(1)} = x_{(i)}, x_{(2)} = x_{(i)} \mid T] = c_{(i,i)} \\ \text{and} \quad &P[x_{(1)} = x_{(i)}, x_{(2)} = x_{(i')} \mid T] = c_{(i,i')} \end{aligned}\right\} \quad \ldots (3.6)$$

Using (3.4) and (3.6), we get

$$E[z_p \mid T] = \sum_{i=1}^{p} C_{(i,i)}\, z_{(i)}^2 + \sum_{i\neq i'=1}^{p} C_{(i,i')}\, z_{(i)}\, z_{(i')} \quad \ldots (3.7)$$

which proves the theorem.

Improved estimator of σ_x^2 : The usual estimator of $\sigma_x^2 = \Sigma\, P_j \left(\frac{Y_j}{P_j} - Y\right)^2$ is

given by $$s_x^2 = \frac{1}{(n-1)}\Sigma(z_i - \bar{z})^2 = \frac{1}{2n(n-1)}\sum_{i\neq i'=1}^{n}(z_i - z_{i'})^2.$$

Corollary 1 : *Thus an estimator uniformly better than s_z^2 is given by*

$$E[s_z^2 \mid T] = E\left[\frac{(z_1-z_2)^2}{2} \mid T\right] = \sum_{i\neq i'=1}^{r} O_{(i,i')}(z_{(i)}-z_{(i')})^2 \qquad \ldots (3.8)$$

Corollary 2 : *An unbiased estimator of $V(\bar{z}_r)$ is given by*

$$v(\bar{z}_r) = s_r^2 - \sum_{i=1}^{r} O_{(i+i)}\, z_{(i)}^2 - \sum_{i\neq i'=1}^{r} O_{(i,i')}\, z_{(i)}\, z_{(i')}. \qquad \ldots (3.9)$$

Since this estimator is quite complicated for use in large samples, Basu (1958) has suggested the use of

$$\frac{1}{n(n-1)} \Sigma\, (z_i - \bar{z})^2 \qquad \ldots (3.10)$$

as an estimator of $V(\bar{z}_r)$. As it over-estimates $V(\bar{z}_r)$, we are always on the safe side to use (3.10) as our estimate.

The estimators derived in this and preceding sections, though superior to the usually employed estimators, are not of much use for large scale sample surveys owing to their cumbersome nature. In the next section, we give simpler estimators of Y and Y^2. These estimators, though less efficient than the above derived estimators, are superior to the usually employed estimators.

4. Simple improved estimators of Y and Y^2

Let us suppose that the observed samples are segregated into groups of equal p_i's. For instance, consider the problem of estimating the total yield of a crop from a sample of farms. Every sample-farm is to be selected with probability proportional to its area. Here, if some crude approximation (say correct to an acre) is used to measure their areas, we expect to get number of farms with same p_i in the sample. In the sequel, by the p-value of a unit, we mean the probability of selection associated with that unit. Let $p_{(1)}, p_{(2)}, \ldots, p_{(k)}$ be the distinct p-values of the sample units arranged in an increasing order of their magnitude. Let $n_{(i)}$ be the number of sample units having $p_{(i)}$ as their p-value. However, not all these $n_{(i)}$ units will be distinct, let $\nu_{(i)}$ be the number of distinct units among them. Now, if we arrange these $\nu_{(i)}$ distinct units in an increasing order of their unit-indices and call them $x_{(i1)}, x_{(i2)}, \ldots, x_{(i\nu_{(i)})}$, then it is not difficult to see that the statistic

$$T^* = [\{x_{(11)}, \ldots, x_{(1\nu_{(1)})}; n_{(1)}\}, \ldots, \{x_{(k1)}, \ldots, x_{(k\nu_{(k)})}; n_{(k)}\}] \qquad \ldots (4.1)$$

is sufficient.

It should be noted that if we take away the ancillary statistics $n_{(1)}, \ldots, n_{(k)}$ from the sufficient statistic T^*, then it reduces to the 'order-statistic' T defined in the earlier section. The 'unnecessarily wide' sufficient statistic T^* is used here for the purpose of deriving estimators of Y and Y^2 that are much simpler (though somewhat less efficient) than those considered in the previous section. Theorems 3 and 4 below give simple improved estimators of Y and Y^2 respectively.

Theorem 3 : *For any convex (downwards) loss function, an estimator uniformly better than $\bar{x}$ is given by*

$$\bar{x}_\nu^* = \frac{1}{n}\sum_{i=1}^{k}\frac{n_{(i)}}{p_{(i)}}\, y_{\nu_{(i)}}, \qquad \ldots (4.2)$$

where

$$\bar{y}_{\nu_{(i)}} = \frac{1}{\nu_{(i)}}\sum_{r=1}^{\nu_{(i)}} y_{(r)}.$$

Proof : Evidently an estimator uniformly better than $\bar{x}$ is given by

$$E(\bar{x}\,|\,T^*) = E\left(\frac{y_1}{p_1}\,\middle|\, T^*\right) \qquad \ldots (4.3)$$

Further, the probability of getting a sample with a given T^* is

$$P(T^*) = \frac{n!}{n_{(1)}!\ldots n_{(k)}!}\, p_{(1)}^{n_{(1)}} \ldots p_{(k)}^{n_{(k)}} C_{\nu_{(1)}}(n_{(1)}) \ldots C_{\nu_{(k)}}(n_{(k)}), \qquad \ldots (4.4)$$

where $C_{\nu_{(i)}}(n_{(i)}) = \nu_{(i)}^{n_{(i)}} - \binom{\nu_{(i)}}{1}(\nu_{(i)}-1)^{n_{(i)}} + \ldots (-)^{\nu_{(i)}-1}\binom{\nu_{(i)}}{\nu_{(i)}-1}1^{n_{(i)}};$

$(i = 1, \ldots, k)$

and

$$P[x_1 = x_{(r)} \mid T^*] = \frac{p_{(i)}\dfrac{(n-1)!}{n_{(1)}!\ldots(n_{(i)}-1)!\ldots n_{(k)}!}\, p_{(1)}^{n_{(1)}} \ldots p_{(i)}^{n_{(i)}-1} \ldots p_{(k)}^{n_{(k)}}}{\dfrac{n!}{n_{(1)}!\ldots n_{(i)}!\ldots n_{(r)}!}\, p_{(1)}^{n_{(1)}} \ldots p_{(i)}^{n_{(i)}} \ldots p_{(k)}^{n_{(k)}}}$$

$$\times \frac{C_{\nu(1)}(n_{(1)})\ldots\dfrac{C_{\nu(i)}(n_{(i)})}{\nu_{(i)}}\ldots C_{\nu(k)}(n_{(k)})}{C_{\nu(1)}(n_{(1)})\ldots C_{\nu(i)}(n_{(i)})\ldots C_{\nu(k)}(n_{(k)})}$$

$$= \frac{n_{(i)}}{n}\cdot\frac{1}{\nu_{(i)}}. \qquad \ldots (4.5)$$

From (4.3) and (4.5), it follows that

$$E(\bar{x}\,|\,T^*) = \frac{1}{n}\sum_{i=1}^{k}\frac{n_{(i)}}{p_{(i)}}\,\bar{y}_{\nu_{(i)}}$$

which completes the proof of the theorem.

A simple comparison of $\bar{x}_\nu^*$ and $\bar{x}$ will show that $\bar{x}_\nu^*$ will be superior to $\bar{x}$ if and only if the sample size is greater than two and at least three population units have the same p-value, otherwise $\bar{x}_\nu^*$ and $\bar{x}$ will be identical. It is not difficult to give a direct proof of the fact that $V(\bar{x}_\nu^*) \leqslant V(\bar{x})$. The strict sign of inequality holds only when the above condition is satisfied.

Theorem 4 : *For any convex (downwards) loss function, an estimator uniformly better than*

$$z_p = \frac{1}{n(n-1)} \sum_{i \neq i'=1}^{n} \frac{y_i}{p_i} \cdot \frac{y_{i'}}{p_{i'}}$$

is given by

$$z_p^* = \frac{1}{n(n-1)} \Big[\Big\{ \Big(\sum_{i=1}^{k} n_{(i)} \frac{y_{\nu(i)}}{p_{(i)}} \Big)^2 - \sum_{i=1}^{k} n_{(i)} \frac{y_{\nu(i)}^2}{p_{(i)}^2} \Big\} - \sum n_{(i)}(n_{(i)}-1) \frac{C_{\nu_{(i)}-1}(n_{(i)}-1)}{c_{\nu_{(i)}}(n_{(i)})} \cdot \frac{s_{\nu(i)}^2}{p_{(i)}^2} \Big], \quad \ldots \quad (4.6)$$

where $s_{\nu_{(i)}}^2 = \frac{1}{(\nu_{(i)}-1)} \sum_{r=1}^{\nu_{(i)}} (y_{\nu_{(i)}} - \bar{y}_{\nu_{(i)}})^2$, *and* $C_{\nu_{(i)}}(n_{(i)})$

and $C_{\nu_{(i)}-1}(n_{(i)}-1)$ *have meaning similar to those defined in* (4.4).

Proof : Obviously, an estimator uniformly better than z_p is given by

$$E[z_p | T^*] = E\Big[\frac{y_1}{p_1} \cdot \frac{y_2}{p_2} \,|\, T^* \Big]. \quad \ldots \quad (4.7)$$

Further, it can be shown that

$$P[x_1 = x_{(ir)},\ x_2 = x_{(ir)} | T^*] = \frac{n_{(i)}(n_{(i)}-1)}{n(n-1)} \cdot \frac{C_{\nu_{(i)}}(n_{(i)}-1)}{\nu_{(i)} C_{\nu(i)}(n_{(i)})},$$

$$P[x_1 = x_{(ir)},\ x_2 = x_{(i'r')} | T^*] = \frac{n_{(i)}(n_{(i)}-1)}{n(n-1)} \times \Big[\frac{C_{\nu_{(i)}}(n_{(i)}) - C_{\nu_{(i)}}(n_{(i)}-1)}{\nu_{(i)}(\nu_{(i)}-1) C_{\nu(i)}(n_{(i)})}, \Big], \ (\nu \neq \nu')$$

and $$P[x_1 = x_{(ir)},\ x_2 = x_{(i'r')} | T^*] = \frac{n_{(i)} n_{(i')}}{n(n-1)} \cdot \frac{1}{\nu_{(i)}} \cdot \frac{1}{\nu_{(i')}} \ (i \neq i'). \quad \ldots \quad (4.8)$$

Therefore,

$$E[z_p | T^*] = \sum_{i=1}^{k} \frac{n_{(i)}(n_{(i)}-1)}{n(n-1)} \sum_{r=1}^{\nu(i)} \frac{y_{(ir)}^2}{p_{(i)}^2} \frac{C_{\nu(i)}(n_{(i)}-1)}{\nu_{(i)} C_{\nu(i)}(n_{(i)})}$$

$$+ \sum_{i=1}^{k} \frac{n_{(i)}(n_{(i)}-1)}{n(n-1)} \sum_{r \neq r'=1}^{\nu(i)} \frac{y_{(ir)} y_{(ir')}}{p_{(i)}^2} \frac{[C_{\nu(i)}(n_{(i)}) - C_{\nu(i)}(n_{(i)}-1)]}{\nu_{(i)}(\nu_{(i)}-1) C_{\nu(i)}(n_{(i)})}$$

$$+ \sum_{i \neq i'=1}^{k} \frac{n_{(i)} n_{(i')}}{n(n-1)} \sum_{r=1}^{\nu(i)} \sum_{r'=1}^{\nu(i')} \frac{y_{(ir)}}{p_{(i)}} \frac{y_{(i'r')}}{p_{(i')}} \frac{1}{\nu_{(i)}} \frac{1}{\nu_{(i')}}. \quad \ldots \quad (4.9)$$

Using the equality

$$\frac{C_{\nu_{(i)}-1}(n_{(i)}-1)}{C_{\nu_{(i)}}(n_{(i)})} s_{\nu_{(i)}}^2 = \bar{y}_{\nu(i)}^2 - \frac{C_{\nu(i)}(n_{(i)}-1)}{\nu_{(i)} C_{\nu_{(i)}}(n_{(i)})} \sum_{r=1}^{\nu(i)} y_{(ir)}^2$$

$$- \frac{[C_{\nu(i)}(n_{(i)}) - C_{\nu(i)}(n_{(i)}-1)]}{\nu_{(i)}(\nu_{(i)}-1) C_{\nu(i)}(n_{(i)})} \sum_{r \neq r'=1}^{\nu(i)} y_{(ir)} y_{(ir')},$$

and simplifying (4.9) we get

$$z_p^* = E[z_p|T^*] = \frac{1}{n(n-1)}\left[\left\{\left(\sum_{i=1}^{k} n_{(i)}\frac{\bar{y}_{v_{(i)}}}{p_{(i)}}\right)^2 - \sum_{i=1}^{k} n_{(i)}\frac{\bar{y}^2_{v_{(i)}}}{p^2_{(i)}}\right\}\right.$$
$$\left. - \sum_{i=1}^{k} n_{(i)}(n_{(i)}-1)\cdot\frac{C_{v_{(i)}-1}(n_{(i)}-1)}{C_{v_{(i)}}(n_{(i)})}\cdot\frac{s^2_{v_{(i)}}}{p^2_{(i)}}\right].$$

This completes the proof.*

Corollary 3 : *It is easy to see that*

$$E[s_i^2|T^*] = E\left[\frac{1}{(n-1)}\Sigma(z_i-\bar{z})^2 \mid T^*\right] = E\left[\frac{y_1^2}{p_1^2}|T^*\right] - E\left[\frac{y_1}{p_1}\cdot\frac{y_2}{p_2}|T^*\right]$$
$$= \sum_{i=1}^{k}\frac{n_{(i)}}{n}\cdot\frac{1}{v_{(i)}}\sum_{r=1}^{v_{(i)}}\frac{y^2_{(ir)}}{p^2_{(i)}} + \sum_{i=1}^{k}\frac{n_{(i)}(n_{(i)}-1)}{n(n-1)}\cdot\frac{C_{v_{(i)}-1}(n_{(i)}-1)}{C_{v_{(i)}}(n_{(i)})}\cdot\frac{s^2_{v_{(i)}}}{p^2_{(i)}}$$
$$- \frac{1}{n(n-1)}\left[\left(\sum_{i=1}^{k} n_{(i)}\frac{\bar{y}_{v_{(i)}}}{p_{(i)}}\right)^2 - \sum_{i=1}^{k} n_{(i)}\frac{\bar{y}^2_{v_{(i)}}}{p^2_{(i)}}\right] \quad \ldots (4.10)$$

is a simple improved estimator of s_z^2.

Corollary 4 : *An unbiased estimator of* $V(\bar{z}_p^*)$ *is given by*

$$v(\bar{z}_p^*) = \bar{z}_p^{*2} + \sum_{i=1}^{k}\frac{n_{(i)}(n_{(i)}-1)}{n(n-1)}\cdot\frac{C_{v_{(i)}-1}(n_{(i)}-1)}{C_{v_{(i)}}(n_{(i)})}\cdot\frac{s^2_{v_{(i)}}}{p^2_{(i)}}$$
$$- \frac{1}{n(n-1)}\left[\left(\sum_{i=1}^{k} n_{(i)}\frac{\bar{y}_{v_{(i)}}}{p_{(i)}}\right)^2 - \sum_{i=1}^{k} n_{(i)}\frac{\bar{y}^2_{v_{(i)}}}{p^2_{(i)}}\right]. \quad \ldots (4.11)$$

However, in practice it seems reasonable to use

$$\frac{1}{n(n-1)}\sum_i\left(\frac{z_i}{p_i} - \bar{z}\right)^2 \quad \ldots (4.12)$$

as an estimator of $V(\bar{z}_p^*)$. First, because it is simple to compute, secondly, because it is always non-negative. Besides this, we are always on the safe side as it over-estimates the variance of $\bar{z}_p^*$.

---

* The estimator Z_p requires the computation of the ratio $\frac{C_{p-1}(n-1)}{C_p(n)}$. Values of $\frac{C_{p-1}(n-1)}{C_p(n)}$ can be obtained from the relation $\frac{C_{p-1}(n-1)}{C_p(n)} = \frac{1}{p} - \frac{C_p(n-1)}{C_p(n)}$.

Values of $\frac{C_p(n-1)}{C_p(n)}$ have been tabulated for all p and $n = 1$ to 50 in a paper published elsewhere in the same issue (Pathak, 1962), pp. 287-302.

5. A remark on Rao-Blackwell theorem

Let X be the sample space of all possible outcomes x. Let $\mathcal{S}$ be the field of subsets of X on which a set $P = \{p\}$ of probability measures is defined. Any statistic $T(x)$ [an $\mathcal{S}$-measurable function defined from X onto another space Y] generates a subfield $\mathcal{S}_T \subseteq \mathcal{S}$. The statistic T or the subfield $\mathcal{S}_T$ is called sufficient for P (Bahadur, 1954) if corresponding to each $\mathcal{S}$-measurable set A, there exists an $\mathcal{S}_T$-P-integrable function ϕ_A such that

$$P[x \epsilon A_0 \bigcap A] = \int_{A_0 \cap A} dp = \int_{A_0} \phi_A(x) dp \text{ for } A_0 \epsilon \mathcal{S}_T,\ p \epsilon P. \quad \ldots \quad (5.1)$$

It is known that any estimator $g(x)$ based on the sample x can be uniformly improved by taking the conditional expectation of $g(x)$ given a sufficient subfield. If several sufficient subfields are available, the minimum condensation of $g(x)$ is mostly obtained by employing the minimal sufficient subfield ($\mathcal{S}_1$ say). There are situations, e.g., in sample surveys where this minimal condensation is unwieldy and it is not possible for practical reasons to use this condensation. But it is sometimes possible to divide X into subsets $K(\epsilon \mathcal{S}_1)$ and K' such that the condensation is simple on K and unwieldy on K'. In such cases simpler condensation can be achieved with the help of some other subfield, $\mathcal{S}_2$, which contains the minimal one. It follows as a consequence of Theorem 5 (stated below) that condensation smaller than that of $\mathcal{S}_2$ (but larger than that of $\mathcal{S}_1$) can be obtained by employing $\mathcal{S}_1 K + \mathcal{S}_2 K'$ as the sufficient subfield, and this condensation will still have the merit of simplicity.

Theorem 5: *Let $\mathcal{S}_1$ and $\mathcal{S}_2$ be two sufficient subfields of $(X, \mathcal{S}, P)$, and K a set common to $\mathcal{S}_1$ and $\mathcal{S}_2$. Then the subfield**

$$\mathcal{S}_3 = \mathcal{S}_1 K + \mathcal{S}_2 K' \quad \ldots \quad (5.2)$$

is also sufficient.

Proof: Since $\mathcal{S}_1$ and $\mathcal{S}_2$ are two sufficient subfields, there exist for each $\mathcal{S}$-measurable set A, an $\mathcal{S}_1$-P-integrable function ϕ_{1A} and an $\mathcal{S}_2$-P-integrable function ϕ_{2A} such that

$$\int_{A_i \cap A} dp = \int_{A_i} \phi_{iA}(x) dp \quad \text{for } A_i \epsilon \mathcal{S}_i \quad (i = 1, 2,), p \epsilon P. \quad \ldots \quad (5.3)$$

Now for any $A_3 = A_1 K + A_2 K' \epsilon \mathcal{S}_3$, and for each $A \epsilon \mathcal{S}$

$$\int_{A_3 \cap A} dp = \int_{A_3} \phi_{3A}\, dp. \quad \ldots \quad (5.4)$$

where $\phi_{3A} = \phi_{1A}\chi_k + \phi_{2A}\chi_{k'}$, and $\chi_k = 1 - \chi_{k'}$ is the characteristic function of the set K.

It is easily seen that $\phi_{1A}(x)\chi_k$ and $\phi_{2A}\chi_{k'}$ are both $\mathcal{S}_3$-P-integrable and thus ϕ_{3A} is $\mathcal{S}_3$-P-integrable. This with (5.4) implies that $\mathcal{S}_3$ is a sufficient subfield.

Corollary 5: *If $\mathcal{S}_1 \subseteq \mathcal{S}_2$, then $\mathcal{S}_1 \subseteq \mathcal{S}_3 \subseteq \mathcal{S}_2$.*

---

* $\mathcal{S}_3$ is the field of sets of the form $\mathcal{S}_3 = \{A_1 K + A_2 K'\}$, $A_i \epsilon \mathcal{S}_i$ $(i = 1, 2)$.

Corollary 6 : *Let* $\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_k$ *be* k *sufficient subfields. If* $A_1, A_2, \ldots, A_k$ $[\bigcup_{i=1}^{k} A_i = X, A_i \bigcap A_j = \phi\ (i \neq j = 1, \ldots, k)]$ *are* k *sets such that* $A_i \in \mathcal{S}_i\ (i = 1, \ldots, k)$, *then*

$$\mathcal{S}^* = A_1\mathcal{S}_1 + A_2\mathcal{S}_2 + \ldots + A_k\mathcal{S}_k \qquad \ldots (5.5)$$

is also a sufficient subfield.

Corollary 7 : *Let* $g(x)$ *be an estimator based on* x. *Let* $\mathcal{S}_1$ *and* $\mathcal{S}_2$ *be two sufficient subfields and* K *a set common to* $\mathcal{S}_1$ *and* $\mathcal{S}_2$. *Then for any convex (downwards) loss function, an estimator uniformly better than* $g(x)$ *is given by*

$$E[g(x)|\mathcal{S}_3] = \begin{cases} E[g(x)|\mathcal{S}_1] & \text{if } x \in K \\ E[g(x)|\mathcal{S}_2] & \text{otherwise,} \end{cases} \qquad \ldots (5.6)$$

where $\mathcal{S}_3 = \mathcal{S}_1 K + {}_2K'$.

This result is useful in estimation problems when the improved estimator $E[g(x)|\mathcal{S}_1]$ is difficult to compute for all $x \in X$. In such cases, this result may be utilized by employing subfield $\mathcal{S}_1$ and another subfield $\mathcal{S}_2$ such that $E[g(x)|\mathcal{S}_1]$ is simple to compute when $x \in K$ and $E[g(x)|\mathcal{S}_2]$ is simple to compute when $x \in K'$ ($K \in \mathcal{S}_i$ $i = 1, 2$) and using (5.6) as the improved estimator. The resulting estimator will still be better than $g(x)$ and in addition will be simple to compute. Further if $\mathcal{S}_1 \subseteq \mathcal{S}_2$, this estimator will also be better than $E[g(x)|\mathcal{S}_2]$. For completeness a straightforward generalisation of Theorem 5 to the case of countable number of sufficient subfields is given below.

Theorem 6 : *Let* $\{\mathcal{S}_i\}$ *be a countable number of sufficient subfields. Let* $\{K_i\}$ *be a sequence of mutually exclusive and exhaustive sets such that* $K_i \in \mathcal{S}_i\ (i = 1, 2, \ldots)$. *Then*

$$\mathcal{S}^* = \sum_{i=1}^{\infty} K_i\mathcal{S}_i$$

is a sufficient subfield.

Proof : Since $\mathcal{S}_i$ is a sufficient subfield, there exists for each $A \in \mathcal{S}$ an $\mathcal{S}_i$-P-integrable function ϕ_{iA} such that

$$P[x \in A_i \bigcap A] = \int_{A_i \cap A} dp = \int_{A_i} \phi_{iA}(x)dp \text{ for } A_i \in \mathcal{S}_i\ (i = 1, \ldots, \infty), p \in P. \quad \ldots (5.7)$$

Now for any $A^* = \Sigma A_iK_i \in \mathcal{S}^*$ and for each $A \in \mathcal{S}$, we have from (5.7)

$$P[x \in A^* \bigcap A] = \lim_{n \to \infty} P[x \in (\sum_{i=1}^{n} A_iK_i) \bigcap A]$$

$$= \lim_{n \to \infty} \int_{\left(\sum_{i=1}^{n} A_iK_i\right)} \left[\sum_{i=1}^{n} \phi_{iA}(x)\chi_{K_i}(x)\right] dp$$

$$= \lim_{n \to \infty} \int_{A^*} \left[\sum_{i=1}^{n} \phi_{iA}(x)\chi_{K_i}(x)\right] dp. \qquad \ldots (5.8)$$

Since $0 \leqslant f_n = \left(\sum_{i=1}^{n} \phi_{iA}(x)\chi_{K_i}(x) \right) \leqslant 1$ a.e. (P), $\lim_{n\to\infty} f_n = f_A^*$ exists a.e. (P). We, thus, have by Lebesgue's monotone convergence theorem

$$P[x \in A^* \bigcap A] = \int_{A^*} f_A^*(x)dp. \qquad \ldots \quad (5.9)$$

Since for each n, f_n is S^*—P-integrable, $f_A^*(x)$ is S^*—P-integrable. Hence the theorem.

In the following section Theorem 5 is used in sampling with unequal probabilities to derive some simple improved estimators of Y.

6. Application to sampling with unequal probabilities

We have seen that $\bar{z}_\nu$ is uniformly better than $\bar{z}$. For $n = 3$, $\bar{z}_\nu$ can be expressed as

$$\bar{z}_\nu = \begin{cases} z_{(1)} & \text{if } \nu = 1, \\ \frac{1}{3}\left[z_{(1)} + z_{(2)} + \frac{y_{(1)}+y_{(2)}}{p_{(1)}+p_{(2)}}\right] & \text{if } \nu = 2 \\ \frac{1}{3}[z_{(1)}+z_{(2)}+z_{(3)}] & \text{if } \nu = 3. \end{cases} \qquad \ldots \quad (6.1)$$

It is not simple to compute $\bar{z}_\nu$ when $n > 3$ owing to cumbersome computation of $C_{(i)}$'s. However, if in a sample of size n, $\nu=(n-1)$, $\bar{z}_\nu$ is expressible in the simple form

$$\bar{z}_{(n-1)} = \frac{1}{n}\left[\sum_{i=1}^{(n-1)} \frac{y_{(i)}}{p_{(i)}} + \frac{\sum_{i=1}^{(n-1)} y_{(i)}}{\sum_{i=1}^{(n-1)} p_{(i)}} \right]. \qquad \ldots \quad (6.2)$$

As a direct consequence of Corollary 7, it follows that a simple estimator uniformly better than $\bar{z}_\nu^*$ (and hence better than $\bar{z}$) is given by

$$\bar{z}'_\nu = \begin{cases} \bar{z}_\nu & \text{if } \nu = (n-1) \\ \bar{z}_\nu^* & \text{otherwise.} \end{cases} \qquad \ldots \quad (6.3)$$

Two points in favour of utilising z'_ν are: (i) it is as simple as $\bar{z}$ or $\bar{z}_\nu^*$ and (ii) it is more efficient than $\bar{z}_\nu^*$.

Another simple improved estimator of Y. Another simple improved estimator of Y can be derived by using the following sufficient statistic

$$T_2 = [(x_{(1)}, \alpha_{(1)}), \ldots, (x_{(\nu)}, \alpha_{(\nu)})] \qquad \ldots \quad (6.4)$$

where

$$\alpha_{(i)} = \begin{cases} \lambda_{(i)} & \text{if } \lambda_{(i)} > 2 \\ 1 & \text{otherwise} \quad (i = 1, \ldots, \nu) \end{cases}$$

and $\lambda_{(i)}$ is the number of times $x_{(i)}$ is included in the sample.

Assuming without any loss of generality that $\alpha_{(i)} = 1$ if $i = 1, \ldots, k$, and $\alpha_{(i)} > 1$ if $i = k+1, \ldots, \nu$, and $\sum_{i=1}^{k} \lambda_{(i)} = m$, it can be shown that an estimator better than $\bar{z}$ is given by

$$\bar{z}_{\nu(2)} = E[\bar{z} \mid T_2] = \frac{1}{n}\left[\sum_{i=1}^{\nu} \alpha_{(i)} z_{(i)} + \sum_{i=1}^{k} d_{(i)} z_{(i)} \right], \qquad \ldots \quad (6.5)$$

where $d_{(i)} = \dfrac{\Sigma_i' p_{(1)} \cdots p_{(m-k)}}{\Sigma_1 p_{(1)} \cdots p_{(m-k)}}$ the summations Σ_1 and Σ_i' have been defined in (2.3) and are taken over $p_{(1)} \ldots, p_{(k)}$.

In practical situations, it is much simpler to compute this estimator than to compute $\hat{s}_r$. For $m-k=1$ and $m-k=2$, $\hat{s}_{r_{(2)}}$ is given by

$$\hat{s}_{r_{(2)}} = \begin{cases} \dfrac{1}{n}\left[\sum_{i=1}^{n} \alpha_{(i)} z_{(i)} + \dfrac{\sum_{i=1}^{k} y_{(i)}}{\sum_{i=1}^{k} p_{(i)}}\right] \text{ if } m-k=1 \\ \dfrac{1}{n}\left[\sum_{i=1}^{n} \alpha_{(i)} z_{(i)} + \dfrac{2k(y,p)}{k(p,p)}\right] \text{ if } m-k=2 \end{cases} \quad \ldots (6.6)$$

where $k(y, p) = \sum_{i=1}^{k} y_{(i)} p_{(i)} - (\Sigma y_{(i)})(\Sigma p_{(i)})$ and $k(p, p)$ is defined similarly.

In general when $(m-k)$ is large, this estimator may also involve some extra computation. If the statistician is not even in favour of this extra computation, the author, as a consequence of Corollary 7, recommends the following improved procedure of estimation

$$\begin{aligned} &\text{(i)} \quad \text{use } \bar{s} \quad \text{if} \quad (m-k) \qquad > 2 \\ &\text{(ii)} \quad \text{use} \quad E[\bar{s} \mid T_2] \quad \text{if} \quad (m-k) \leqslant 2. \end{aligned} \quad \ldots (6.7)$$

For estimating the variance of these estimators, author suggests (4.12) as an estimator.

7. Concluding remark

In case of large samples if one is interested in altogether dispensing with the extra computation, the observed sample of size n may be divided into sub-samples of sizes $n_1, n_2, \ldots, n_i$ etc., ($\Sigma n_i = n$ and $n_i = 3, 4$ or 5 etc.). This division should, however, be independent of sample observations. Each sub-sample may then be treated as a sample in itself and simple improved estimators may be obtained for each sub-sample by using the estimators given in the preceding section. The over-all improved estimator can now be obtained by averaging the estimators obtained from each sub-sample with weights proportional to the sub-sample sizes.

Acknowledgment

The author wishes to express his gratefulness to Dr. D. Basu for his guidance and encouragement in writing the paper.

References

Bahadur, R. R. (1954) : Sufficiency and statistical decision function. *Ann. Math. Stat.*, **25**, 423-462.

Basu, D. (1958) : On sampling with and without replacement. *Sankhyā*. **20**, 287-294.

Fraser, D. A. S. (1957) : *Nonparametric Methods in Statistics*, John Wiley and Sons, New York.

Pathak, P. K. (1962) : On simple random sampling with replacement. *Sankhyā*, **24**, 287-302.

Paper received : August, 1961.