ESTIMATION OF PARAMETERS IN THE DOMAIN OF STUDY WITH CLUSTER SAMPLING

by Ing-Tzer Wey

1. Introduction

It is usually planned in sample surveys to estimate parameters not only for the entire population but also for its subdivisions, which may be called subpopulations, or domains of study. If the domains represent classifications by variables such as income, expenditure, age, sex, and so on, the domain to which a particular sampling unit belongs is not known until the sample has been taken. Thus, the number of sampled units falling into each domain is itself a random variable. This will naturally affect the precision of estimates in the domain of study and is perhaps the most characteristic difference between the estimations of parameters in the domain of study and in the entire population.

For generality, we shall only consider the cluster units containing different numbers of elementary units. Let M_i be the number of elementary units in the i-th cluster unit, and let y_{ik} be the observed value on the k-th elementary unit within the i-th cluster unit. Let y_{ikJ} denote the observed value which belongs to the J-th domain.

N: number of clusters in the population,

n: number of clusters drawn from the population,

N_J : number of clusters in the population that fall into the J-th domain,

n_J : number of clusters in the sample of n that happen to fall into the J-th domain,

$M = \sum_{i=1}^{N} M_i$: total number of elementary units in the population,

M_{iJ} : number of elementary units in the i-th cluster that fall into the J-th domain,

$M_J = \sum_{i=1}^{N_J} M_{iJ}$: total number of elementary units falling into the J-th domain in the population,

$y_{iJ} = \sum_{k=1}^{M_{iJ}} y_{ikJ} = M_{iJ}\bar{y}_{iJ}$: total of y-variates in the J-th domain within the i-th cluster,

$Y_J = \sum_{i=1}^{N_J} y_{iJ}$: population total of y-variates in the J-th domain,

$\bar{Y}_J = Y_J / N_J$: population mean per cluster of the J-th domain,

$\bar{\bar{Y}}_J = Y_J / M_J$: population mean per elementary unit of the J-th domain,

$\bar{y}_J = \sum_{i=1}^{n_J} y_{iJ}/n_J$: sample mean per cluster of the J-th domain.

Various estimators of the domain mean and total will be derived in different sampling schemes, and their precisions will be compared. An estimator of the gain in precision in the estimation of the domain total due to the use of one sampling scheme as compared to another sampling scheme will be constructed on the basis of the former sample itself.

II. Sampling without Replacement with Equal Probability

A sample of n clusters is taken from a population of N clusters by simple random sampling without replacement. Consider first the estimation of the population mean per elementary unit of the J-th domain from the sample. We shall consider a ratio estimator in which M_{iJ} is taken as the auxiliary variate.

$$\bar{\bar{y}}_J = \sum_{i=1}^{n_J} y_{iJ} \Big/ \sum_{i=1}^{n_J} M_{iJ} \tag{1}$$

In the notation of the ratio estimator the population ratio $R_J = Y_J / X_J = Y_J / M_J = \bar{\bar{Y}}_J$. By theoram 2.3 given by Wey (1970), assuming that the number of clusters in the sample is large, we have

$$\mathrm{Var}(\bar{\bar{y}}_J) = \frac{1-f}{nP_J}\left(1+\frac{Q_J}{nP_J}\right)\frac{1}{\bar{M}_J^2}\,\frac{1}{N_J-1}\sum_{i=1}^{N_J} M_{iJ}^2(\bar{y}_{iJ}-\bar{\bar{Y}}_J)^2 \tag{2}$$

where $f=n/N$, $P_J=1-Q_J=N_J/N$, and $\bar{M}_J=M_J/N_J$.

Since it is often not known the values of N_J and M_J, the population total Y_J of the J-th domain is estimated by

$$\hat{Y}_J = \frac{N}{n}\sum_{i=1}^{n_J} y_{iJ}$$

with variance being the same as expression (1.6) given by Wey (1970),

$$\text{Var}(\hat{Y}_J) = \frac{N^2(1-f)}{n(N-1)}\left[\sum_{i=1}^{N_J} y_{iJ}^2 - Y_J^2/N\right] \qquad (3)$$

III. Sampling with Replacement and with Unequal Probabilities

If all the cluster sizes M_i are known, we may take a sample of n clusters with probabilities proportional to their sizes M_i. In this case the probability that the i-th cluster will be selected with replacement is $z_i = M_i / M$. This sampling scheme is called sampling with probability proportional to size (pps sampling). In some applications the cluster sizes are known only approximately. In others the "size" is not the number of elementary units in the cluster but simply a measure of its bigness that is thought to be highly correlated with the cluster total y_i. Consequently, we shall consider sampling with probability proportional to an estimate or measure of size M'_i (ppes sampling). The probability of selecting the i-th cluster in ppes sampling is $z_i = M'_i / M'$, where $M' = \sum_{i=1}^{N} M'_i$.

(A) Estimation of the Domain Total

Before the derivation of an unbiased estimator for the domain total, we define that to each cluster in the population is attached a variate y^*_i for the character y such that

$$y^*_i = \begin{cases} y_{iJ} & \text{if the i-th cluster is in the domain J,} \\ 0 & \text{otherwise.} \end{cases} \qquad (4)$$

Now, if a sample of n clusters is drawn with probabilities z_i's and with replacement, then a general linear estimator of the domain total Y_J can be written as

$$\hat{Y}_{Jppes} = \sum_{i=1}^{N} t_i c_i y^*_i \qquad (5)$$

where c_i 's are unknown constants to be determined such that the estimator be unbiased, and t_i 's are random variables defined as the number of times that the i-th cluster appears in a specific sample of size n, where t_i may have any of the values 0, 1,……, n. Obviously, the joint distribution of t_i 's is the multinomial expression

$$f(t_1,\cdots\cdots,t_N)=\frac{n!}{t_1!\,t_2!\cdots\cdots t_N!}\prod_{i=1}^{N} z_i^{t_i}$$

where $0\leq t_1,\cdots\cdots,t_N\leq n$ and $\sum_{i=1}^{N} t_i = n$. For the multinomial, the following properties of the distribution of the t_i are well known:

$$\left.\begin{aligned} &E(t_i)=nz_i \\ &Var(t_i)=nz_i(1-z_i) \\ &Cov(t_i,t_j)=-nz_i z_j \end{aligned}\right\} \qquad (6)$$

Taking the expectation of the expression (5) with respect to the t_i, we have

$$E(\hat{Y}_{Jppes})=\sum_{i=1}^{N} c_i y_i^* E(t_i)=n\sum_{i=1}^{N} c_i z_i y_i^*$$

In order that $\hat{Y}_{Jppes}$ be an unbiased estimator of Y_J, we must have

$$E(\hat{Y}_{Jppes})=\sum_{i=1}^{N} y_i^* = \sum_{i=1}^{N_J} y_{iJ}=Y_J$$

which gives $c_i=(nz_i)^{-1}$.

Thus, an unbiased estimator of Y_J is given by

$$\hat{Y}_{Jppes}=\frac{1}{n}\sum_{i=1}^{N} t_i y_i^*/z_i = \frac{1}{n}\sum_{i=1}^{n_J} y_{iJ}/z_i \qquad (7)$$

The variance of $\hat{Y}_{Jppes}$ is obtained as follows:

$$Var(\hat{Y}_{Jppes})=\frac{1}{n^2}\left[\sum_{i=1}^{N}\left(\frac{y_i^*}{z_i}\right)^2 Var(t_i)+2\sum_{i<k}^{N}\frac{y_i^*}{z_i}\frac{y_k^*}{z_k}Cov(t_i,t_k)\right]$$

$$=\frac{1}{n}\left[\sum_{i=1}^{N}\left(\frac{y_i^*}{z_i}\right)^2 z_i(1-z_i)-2\sum_{i<k}^{N}\frac{y_i^*}{z_i}\frac{y_k^*}{z_k}z_i z_k\right]$$

$$= \frac{1}{n} \left[\sum_{i=1}^{N} \frac{y_i^{*2}}{z_i} - Y^{*2} \right]$$

$$= \frac{1}{n} \sum_{i=1}^{N} z_i \left(\frac{y_i^*}{z_i} - Y^* \right)^2 \tag{8}$$

$$= \frac{1}{n} \left[\sum_{i=1}^{N_J} z_i \left(\frac{y_{iJ}}{z_i} - Y_J \right)^2 + \left(1 - \sum_{i=1}^{N_J} z_i \right) Y_J^2 \right]$$

since $1 = \sum_{i=1}^{N} z_i = \sum_{i=1}^{N_J} z_i + \left(1 - \sum_{i=1}^{N_J} z_i\right)$.

The foregoing can be summed up by the following theorem.

Theorem 1 If a sample of n clusters is drawn with probabilities z_i and with replacement, then

$$\hat{Y}_{Jppes} = \frac{1}{n} \sum_{i=1}^{n_J} y_{iJ}/z_i$$

is an unbiased estimator of the domain total Y_J with variance

$$Var(\hat{Y}_{Jppes}) = \frac{1}{n} \left[\sum_{i=1}^{N_J} z_i \left(\frac{y_{iJ}}{z_i} - Y_J \right)^2 + \left(1 - \sum_{i=1}^{N_J} z_i \right) Y_J^2 \right] \tag{9}$$

Taking $z_i = M_i / M$ in theorem 1 gives the corresponding results for sampling with probability proportional to size.

The next theorem shows how to estimate the variance of $\hat{Y}_{Jppes}$ from the sample.

Theorem 2 Under the conditions of Theorem 1, an unbiased estimator of $Var(\hat{Y}_{Jppes})$ is provided by

$$var(\hat{Y}_{Jppes}) = \frac{1}{n(n-1)} \left[\sum_{i=1}^{n_J} \left(\frac{y_{iJ}}{z_i} - \hat{Y}_{Jppes} \right)^2 + (n - n_J) \hat{Y}_{Jppes}^2 \right] \tag{10}$$

Proof: We note that in sampling with replacement and with unequal probabilities the

random variable y_i^* / z_i $(i=1,\cdots\cdots, n)$ are n independent unbiased estimators of $Y^* = Y_J$ having the same variance, i.e. for each i $(=1,\cdots\cdots, n)$

$$E\left(\frac{y_i^*}{z_i}\right)=\sum_{i=1}^{N} z_i \left(\frac{y_i^*}{z_i}\right)=\sum_{i=1}^{N} y_i^* = Y_J$$

$$\mathrm{Var}\left(\frac{y_i^*}{z_i}\right)=E\left(\frac{y_i^*}{z_i}-Y^*\right)^2=\sum_{i=1}^{N} z_i \left(\frac{y_i^*}{z_i}-Y^*\right)^2 = n\ \mathrm{Var}\left(\hat{Y}_{Jppes}\right)$$

and $\mathrm{Cov}\left(\frac{y_i^*}{z_i}, \frac{y_k^*}{z_k}\right)=0$ for $i \neq k$.

Hence, it is obvious that by setting $w_i^* = y_i^* / z_i$, an unbiased estimator of $\mathrm{Var}(w_i^*)$ is given by

$$\frac{1}{n-1}\sum_{i=1}^{n}(w_i^*-\bar{w}^*)^2$$

where $\bar{w}^* = \frac{1}{n}\sum_{i=1}^{n} w_i^* = \hat{Y}_{Jppes}$.

Thus, an unbiased estimator of the variance of $\hat{Y}_{Jppes}$ is provided by

$$\mathrm{Var}\left(\hat{Y}_{Jppes}\right)=\frac{1}{n(n-1)}\sum_{i=1}^{n}(w_i^*-\bar{w}^*)^2$$

$$=\frac{1}{n(n-1)}\left[\sum_{i=1}^{n_J}\left(\frac{y_{iJ}}{z_i}-\hat{Y}_{Jppes}\right)^2+(n-n_J)\hat{Y}_{Jppes}^2\right]$$

This completes the proof of the theorem.

(B) Comparison with Sampling with Equal Probability

From the expression (8), it is clear that the variance of $\hat{Y}_{Jppes}$ will be small if z_i is roughly proportional to y_i^*. In fact, the variance is zero when z_i and y_i^* are exactly proportional. Thus, it may be expected that if the regression of y_i^* on M_i', the size measure is found to be a straight line passing through the origin,

the ppes sampling will be very efficient. We define M_i^* as M_i' if the i-th cluster is in domain J and zero otherwise. Suppose the relationship between y_i^* and M_i^* is of the form

$$y_i^* = A + BM_i^* + e_i \tag{11}$$

where A and B are constants, and e_i is a random variable with its conditional expected value and variance for a given M_i^* as

$$E(e_i \mid M_i^*) = 0, \quad Var(e_i \mid M_i^*) = KM_i^{*g} \tag{12}$$

where K and g are constants, This model is based on the work of Des Raj (1958), Zarkovich (1960), and Cochran (1953) who used the similar models.

Note that the variance of the estimator of Y_J based on simple random sampling with replacement is obtained by omitting the finite population correction term in the expression (3) as

$$V_{srs} = Var(\hat{Y}_J) = \frac{1}{n}\left[N\sum_{i=1}^{N_J} y_{iJ}^2 - Y_J^2\right] = \frac{1}{n}\left[N\sum_{i=1}^{N} y_i^{*2} - Y^{*2}\right]$$

where $\hat{Y}_J = \frac{N}{n}\sum_{i=1}^{n_J} y_{iJ}$,

and the variance of the estimator of Y_J based on ppes sampling is given by

$$V_{ppes} = Var(\hat{Y}_{Jppes}) = \frac{1}{n}\left[\sum_{i=1}^{N}\frac{y_i^{*2}}{z_i} - Y^{*2}\right]$$

$$= \frac{1}{n}\left[M'\sum_{i=1}^{N}\frac{y_i^{*2}}{M_i'} - Y^{*2}\right]$$

The difference between the two variances is

$$D = V_{srs} - V_{ppes} = \frac{N}{n}\sum_{i=1}^{N} y_i^{*2}\left(1 - \frac{\overline{M'}}{M_i'}\right)$$

Substituting the value given in (11) for y_i^* in the expression for D and taking its expected value using (12), we get after simplification

$$E(D) = \frac{N^2}{n}\left[A^2 P_J\left(1-\frac{\overline{M}'}{N_J}\sum_{i\epsilon J} M_i'^{-1}\right)+B^2\,\mathrm{Cov}\,(M_i^*, M_i')\right.$$

$$\left.+2AB\,(\overline{M}^* - \overline{M}'P_J)+K\,\mathrm{Cov}(M_i^{*g-1}, M_i')\right]$$

where $P_J = N_J/N$, and

$$\mathrm{Cov}\,(M_i^*, M_i') = \frac{1}{N}\sum_{i=1}^{N} M_i^*\,(M_i' - \overline{M}') = \frac{1}{N}\sum_{i\epsilon J}(M_i' - \overline{M}')^2 > 0$$

Hence, ppes sampling will be more efficient than srs, if

$$K\mathrm{Cov}(M_i^{*g-1}, M_i') > A^2 P_J\left(\frac{\overline{M}'}{N_J}\sum_{i\epsilon J} M_i'^{-1} - 1\right)$$

$$+2AB\,(\overline{M}'P_J - \overline{M}^*) - B^2\,\mathrm{Cov}\,(M_i^*, M_i') \qquad (13)$$

From the inequality (13), it is clear that even if y_i^* and M_i^* are perfectly linearly related, that is, $\mathrm{Var}\,(e_i \mid M_i^*) = 0$, ppes sampling is not necessarily more efficient than srs, for the inequality becomes

$$B^2\,\mathrm{Cov}\,(M_i^*, M_i') > A^2 P_J\left(\frac{\overline{M}'}{N_J}\sum_{i\epsilon J} M_i'^{-1} - 1\right) + 2AB\,(\overline{M}'P_J - \overline{M}^*)$$

which may not be satisfied always. Thus, linearity of regression is not a sufficient condition for ppes sampling to be better than srs. If the regression line passes through the origin, that is, if $A = 0$, the inequality (13) becomes

$$K\mathrm{Cov}(M_i^{*g-1}, M_i') > -B^2\mathrm{Cov}(M_i^*, M_i') \qquad (14)$$

which is obviously satisfied if $g > 1$, since $K \geq 0$ and $\mathrm{Cov}(M_i^{*g-1}, M_i') > 0$. Empirical studies conducted by different authors have shown that the value of g is likely to lie between 1 and 2 in practice.

(C) Gain due to PPES Sampling

It is of interest to note that it is possible to estimate unbiasedly the gain due to ppes sampling as compared to simple random sampling with replacement (srs sampling) from the ppes sample itself. An unbiased estimator of the variance of the ppes estimator is already given in (10) as

$$\mathrm{var}(\hat{Y}_{Jppes}) = \frac{1}{n(n-1)} \left[\sum_{i=1}^{n_J} \left(\frac{y_{iJ}}{z_i} - \hat{Y}_{Jppes} \right)^2 + (n - n_J) \hat{Y}^2_{Jppes} \right] \tag{15}$$

The variance of the estimator of Y_J based on srs sampling with replacement is

$$\mathrm{Var}(\hat{Y}_J) = \frac{1}{n} \left[N \sum_{i=1}^{N_J} y^2_{iJ} - Y^2_J \right], \quad \hat{Y}_J = \frac{N}{n} \sum_{i=1}^{n_J} y_{iJ}$$

and an unbiased estimator of this on the basis of a ppes sample can be obtained by noting that unbiased estimators of the terms $\sum y^2_{iJ}$ and Y^2_J are respectively given by

$$\frac{1}{n} \sum_{i=1}^{n_J} y^2_{iJ}/z_i \quad \text{and} \quad \hat{Y}^2_{Jppes} - \mathrm{var}(\hat{Y}_{Jppes}).$$

Thus, an unbiased estimator of $\mathrm{Var}(\hat{Y}_J)$ is given by

$$\mathrm{var}(\hat{Y}_J)_{ppes} = \frac{1}{n^2} \left[N \sum_{i=1}^{n_J} \frac{y^2_{iJ}}{z_i} - n\hat{Y}^2_{Jppes} \right] + \frac{1}{n} \mathrm{var}(\hat{Y}_{Jppes}) \tag{16}$$

By comparing this variance estimator with (15), we can make an estimator of the gain G_{ppes} due to the use of ppes sampling as follows:

$$G_{ppes} = \mathrm{var}(\hat{Y}_J)_{ppes} - \mathrm{var}(\hat{Y}_{Jppes}) = \frac{1}{n^2} \sum_{i=1}^{n_J} \frac{y^2_{iJ}}{z_i} \left(N - \frac{1}{z_i} \right) \tag{17}$$

Similarly, it is also possible to estimate the gain in precision by estimating unbiasedly $\mathrm{Var}(\hat{Y}_{Jppes})$ on the basis of a random sample selected with srs. Since

$$\mathrm{Var}(\hat{Y}_{Jppes}) = \frac{1}{n} \left[\sum_{i=1}^{N} y^{*2}_i / z_i - Y^{*2} \right]$$

and unbiased estimators of $\Sigma\, y_i^{*2}/z_i$ and Y^{*2} based on n units selected with srs are respectively given by

$$\frac{N}{n}\sum_{i=1}^{n} y_i^{*2}/z_i = \frac{N}{n}\sum_{i=1}^{n_J} y_{iJ}^2/z_i$$

and $\hat{Y}_J^2 - \mathrm{var}\,(\hat{Y}_J) = \hat{Y}_J^2 - \frac{N^2}{n(n-1)}\left[\sum_{i=1}^{n_J} y_{iJ}^2 - \frac{1}{n}\left(\sum_{i=1}^{n_J} y_{iJ}\right)^2\right]$

Thus, an unbiased estimator of $\mathrm{Var}(\hat{Y}_{Jppes})$ on the basis of a srs sample is given by

$$\mathrm{var}\,(\hat{Y}_{Jppes})_{srs} = \frac{1}{n}\left[\frac{N}{n}\sum_{i=1}^{n_J} y_{iJ}^2/z_i - \hat{Y}_J^2 + \mathrm{var}\,(\hat{Y}_J)\right] \tag{18}$$

Hence, an estimator of the gain G_{srs} due to the use of srs sampling is as follows:

$$G_{srs} = \mathrm{var}\,(\hat{Y}_{Jppes})_{srs} - \mathrm{var}\,(\hat{Y}_J) = \frac{N}{n^2}\sum_{i=1}^{n_J} y_{iJ}^2\left(\frac{1}{z_i} - N\right) \tag{19}$$

It should be noted that in the above estimation of the gain in precision due to the use of ppes or srs sampling, the costs involved in conducting both methods of sampling were assumed to be the same. We summarize the foregoing in the following theorem.

Theorem 3 Let a sample of n clusters be drawn with probabilities proportional to estimates of size with replacement (ppes sampling). Then, an unbiased estimator of the gain in precision in the estimation of the domain total due to the use of ppes sampling as compared to simple random sampling with replacement (srs) is provided by

$$G_{ppes} = \frac{1}{n^2}\sum_{i=1}^{n_J} \frac{y_{iJ}^2}{z_i}\left(N - \frac{1}{z_i}\right)$$

Conversely, if the sample is selected with srs, then, an unbiased estimator of the gain in precision due to srs as compared to ppes sampling is given by

$$G_{srs} = \frac{N}{n^2}\sum_{i=1}^{n_J} y_{iJ}^2\left(\frac{1}{z_i} - N\right).$$

(D) Estimation of the Domain Mean

If the total number M_J of elementary units belonging to the J-th domain in the

population is known, then an unbiased estimator of the domain mean $\bar{Y}_J$ is given by

$$\bar{y}^*_{Jppes} = \frac{1}{nM_J} \sum_{i=1}^{n_J} y_{iJ}/z_i = \hat{Y}_{Jppes}/M_J$$

However, since it is often not known the value of M_J, we shall use an alterrative estimator that is a ratio estimator in which the M_{iJ} are treated as auxiliary variates.

$$\bar{y}_{Jppes} = \left[\frac{1}{n}\sum_{i=1}^{n_J}\frac{y_{iJ}}{z_i}\right] / \left[\frac{1}{n}\sum_{i=1}^{n_J}\frac{M_{iJ}}{z_i}\right] = \hat{Y}_{Jppes}/\hat{M}_{Jppes} \tag{20}$$

In general, the estimator is biased. It can be shown that the ratio of the absolute value of the bias to the standard deviation of the estimator is less than or equal to the coefficient of variation of $\hat{M}_{Jppes}$, i.e.

$$\frac{|\text{Bias}(\bar{y}_{Jppes})|}{\sqrt{\text{Var}(\bar{y}_{Jppes})}} \leq \frac{1}{M_J}\sqrt{\text{Var}(\hat{M}_{Jppes})} = CV(\hat{M}_{Jppes}) \tag{21}$$

where $\text{Var}(\hat{M}_{Jppes}) = \frac{1}{n}\left[\sum_{i=1}^{N_J} z_i\left(\frac{M_{iJ}}{z_i} - M_J\right)^2 + \left(1 - \sum_{i=1}^{N_J} z_i\right)M_J^2\right]$ (22)

This result is extremely useful in practice when it is considered important that the bias of the estimator be negligibly small in order that proper confidence statements be made. In that case the sample size n is to be so chosen that the coefficient of variation of $\hat{M}_{Jppes}$ is less than 0.1 [Wey (1968), Sec. 6.3].

In the derivation of an approximate variance of $\bar{y}_{Jppes}$, we define analogously to the variate y_i^* given in (4) a new variate M_i^* on each cluster in the population as follows:

$$M_i^* = \begin{cases} M_{iJ} & \text{if the i-th cluster belongs to the J-th domain,} \\ 0 & \text{otherwise.} \end{cases} \tag{23}$$

Then, the deviation of the estimator $\bar{y}_{Jppes}$ from the true mean $\bar{Y}_J$ is given by

$$\bar{y}_{J\text{ppes}} - \bar{Y}_J = \frac{1}{n\hat{M}_{J\text{ppes}}} \sum_{i=1}^{n} (y_i^* - \bar{Y}_J M_i^*)/z_i$$

Assume the sample size n is large enough so that the terms of order n^{-1} in the expansion of $\hat{M}^{-1}_{J\text{ppes}}$ by the Taylor's series around M_J are negligible. This gives

$$\bar{y}_{J\text{ppes}} - \bar{Y}_J \doteq \frac{1}{nM_J} \sum_{i=1}^{n} (y_i^* - \bar{Y}_J M_i^*)/z_i$$

Therefore, by Theorem 1 an approximate variance of $\bar{y}_{J\text{ppes}}$ can be obtained as follows:

$$\text{Var}(\bar{y}_{J\text{ppes}}) = \frac{1}{nM_J^2} \sum_{i=1}^{N} z_i \left[\frac{y_i^* - \bar{Y}_J M_i^*}{z_i} - (Y^* - \bar{Y}_J M^*) \right]^2$$

$$= \frac{1}{nM_J^2} \sum_{i=1}^{N_J} (y_{iJ} - \bar{Y}_J M_{iJ})^2/z_i \qquad (24)$$

An alternative expression for the approximate variance is

$$\text{Var}(\bar{y}_{J\text{ppes}}) = \frac{1}{nM_J^2} \sum_{i=1}^{N} z_i \left[\left(\frac{y_i^*}{z_i} - Y^* \right) - \bar{Y}_J \left(\frac{M_i^*}{z_i} - M^* \right) \right]^2$$

$$= \frac{1}{M_J^2} [\text{Var}(\hat{Y}_{J\text{ppes}}) + \bar{Y}_J^2 \text{Var}(\hat{M}_{J\text{ppes}})$$

$$- 2\bar{Y}_J \text{Cov}(\hat{Y}_{J\text{ppes}}, \hat{M}_{J\text{ppes}})] \qquad (25)$$

We summarize the above result in the following theorem.

Theorem 4 If a sample of n clusters is drawn with probabilities z_i and with replacement, then the J-th domain mean $\bar{Y}_J$ per elementary unit is estimated by

$$\bar{y}_{J\text{ppes}} = \sum_{i=1}^{n_J} y_{iJ}/z_i \Big/ \sum_{i=1}^{n_J} M_{iJ}/z_i$$

with an approximate variance (to terms of order n^{-1})

$$\mathrm{Var}(\bar{\bar{y}}_{Jppes}) = \frac{1}{nM_J^2} \sum_{i=1}^{N_J} (y_{iJ} - \bar{\bar{Y}}_J M_{iJ})^2 / z_i \cdot$$

Using Theorem 2, the approximate variance of $\bar{\bar{y}}_{Jppes}$ may be estimated by

$$\mathrm{var}(\bar{\bar{y}}_{Jppes}) = \frac{1}{n(n-1)\hat{M}_{Jppes}^2} \sum_{i=1}^{n_J} (y_{iJ} - \bar{\bar{y}}_{Jppes} M_{iJ})^2 / z^2{}_i \qquad (26)$$

We shall examine whether or not the ratio estimator $\bar{\bar{y}}_{Jppes}$ should be used even when M_J, the total number of elementary units belonging to the J-th domain in the population is known. If M_J is known, then an unbiased estimator of the domain mean $\bar{\bar{Y}}_J$ is given by

$$\bar{\bar{y}}^*_{Jppes} = \frac{1}{nM_J} \sum_{i=1}^{n_J} y_{iJ}/z_i = \hat{\bar{Y}}_{Jppes} / M_J$$

with variance

$$\mathrm{Var}(\bar{\bar{y}}^*_{Jppes}) = \mathrm{Var}(\hat{\bar{Y}}_{Jppes}) / M_J^2$$

We assume that the sample size n is sufficiently large so as to make the bias of $\bar{\bar{y}}_{Jppes}$ negligible. Then, the approximate variance of $\bar{\bar{y}}_{Jppes}$ given in (25) is valid. Hence, the estimator $\bar{\bar{y}}_{Jppes}$ will give a more precise result whenever

$$\rho > \frac{CV(\hat{M}_{Jppes})}{2CV(\hat{Y}_{Jppes})} \qquad (27)$$

where ρ is the correlation coefficient between $\hat{Y}_{Jppes}$ and $\hat{M}_{Jppes}$, $CV(\hat{M}_{Jppes})$ and $CV(\hat{Y}_{Jppes})$ are coefficients of variation of $\hat{M}_{Jppes}$ and $\hat{Y}_{Jppes}$ respectively. Thus we see that it may be profitable to use the ratio estimator $\bar{\bar{y}}_{Jppes} = \hat{\bar{Y}}_{Jppes} / \hat{M}_{Jppes}$, though biased, even in the case where the actual value of M_J is known, provided the correlation coefficient between the estimator of the numerator and the

denominator satisfies the condition in (27). It may be mentioned from the expression (24) that the variance of $\bar{\bar{y}}_{Jppes}$ would be small if the set of probabilities z_1, ……, z_n used for selection is roughly proportional to both the numerator and the denominator variables and if y_i^* / z_i and M_i^* / z_i are positively correlated.

V. Summary

1. If a sample of n clusters is drawn with probabilities z_i and with replacement (ppes sampling), then

$$\hat{Y}_{Jppes} = \frac{1}{n} \sum_{i=1}^{n_J} y_{iJ}/z_i$$

is an unbiased estimator of the domain total Y_J with variance

$$Var(\hat{Y}_{Jppes}) = \frac{1}{n} \left[\sum_{i=1}^{N_J} z_i \left(\frac{y_{iJ}}{z_i} - Y_J \right)^2 + \left(1 - \sum_{i=1}^{N_J} z_i \right) Y_J^2 \right]$$

and an unbiased estimator of Var $(\hat{Y}_{Jppes})$ is given by

$$var(\hat{Y}_{Jppes}) = \frac{1}{n(n-1)} \left[\sum_{i=1}^{n_J} \left(\frac{y_{iJ}}{z_i} - \hat{Y}_{Jppes} \right)^2 + (n - n_J) \hat{Y}_{Jppes}^2 \right]$$

2. Under the model $y_i^* = A + BM_i^* + e_i$, where $E(e_i \mid M_i^*) = 0$ and $E(e_i^2 \mid M_i^*) = KM_i^{*g}$, $K(>0)$ and g are constants, a sufficient condition for ppes sampling with replacement to be better than simple random sampling with replacement $(\hat{Y}_J = \frac{N}{n} \sum_{i=1}^{n_J} y_{iJ})$ is that $A=0$ and $g>1$.

3. An unbiased estimator of the gain in precision in the estimation of the domain total due to the use of ppes sampling with replacement as compared to simple random sampling with replacement is provided by

$$G_{ppes} = \frac{1}{n^2} \sum_{i=1}^{n_J} \frac{y_{iJ}^2}{z_i} \left(N - \frac{1}{z_i} \right).$$

4. In ppes sampling with replacement, the domain mean $\bar{Y}_J$ may be estimated by either

$$\bar{y}_{Jppes} = \sum_{i=1}^{n_J} \frac{y_{iJ}}{z_i} \Big/ \sum_{i=1}^{n_J} \frac{M_{iJ}}{z_i}$$

or $$\bar{y}^*_{Jppes} = \frac{1}{nM_J} \sum_{i=1}^{n_J} y_{iJ}/z_i$$

But the estimator $\bar{y}_{Jppes}$ will be more efficient than the estimator $\bar{y}^*_{Jppes}$ if the following condition is satisfied:

$$\rho > \frac{CV(\hat{M}_{Jppes})}{2CV(\hat{Y}_{Jppes})}$$

where ρ is the correlation coefficient between $\hat{Y}_{Jppes}$ and $\hat{M}_{Jppes}$, $CV(\hat{Y}_{Jppes})$ and $CV(\hat{M}_{Jppes})$ are coefficients of variation of $\hat{Y}_{Jppes}$ and $\hat{M}_{Jppes}$ respectively.

VI. References

1. Cochran, W. G. (1963): Sampling Techniques. Second edition. John Wiley and Sons, New York.

2. Raj, Des (1958): On the relative accuracy of some sampling techniques. Journal of American Statistical Association 53:98-101.

3. Wey, I. T. (1970): Estimation of Parameters in Domains of Study I. Report to the National Science Council, Republic of China.

4. Zarkovich, S. S. (1960): On the efficiency of sampling with varying probabilities and the selection of units with replacement. Metrika 3:53-59.