Jour. Ind. Soc. Ag. Statistics
44 (2), 1992 : 95–105

Estimation of Weighted Mean of Finite Populations

Shambhu Dayal and Rajesh Kumar*
Central Water Commission, New Delhi
(Received : August, 1983)

Summary

Consider estimating the weighted mean of a characteristic under study, the weights being the straum level totals of another characteristic. The estimaters, their biases and efficiencies are given under two situations; (I) when only the estimates of weights are known and (II) when the estimates of weights are known at the planning stage but their actual values are known at the estimation stage.

Key word: Weighted mean, estimation of yield rate of a crop, allocation of the sample in stratified sampling.

Introduction

Sampling schemes for estimating weighted means were considered by Dayal [1]. In many surveys under stratified sampling the object is to estimate the weighted mean of a characteristic under study where the weights are stratum level totals of another characteristic. Sometimes (situation I), stratum level totals of other characteristic are not available but their estimates are known on the basis of an independent survey and these estimates are used. There is also a situation (situation II) where only estimates of totals of other characteristic are available at the time of planning of the survey, but at estimation stage, actual values of totals are known.

For example, in surveys carried out in India for estimating yield rate and production of a crop, crop cutting experiments are conducted to estimate the yield rate and for estimating area under crop in a district an independent survey is undertaken and complete enumeration of a random sample of villages is done. At the stratum (Tehsil/Taluk) level, the simple mean yield rate of crop and area under crop alongwith estimate of total area are obtained. The

* Indian Institute of Sugarcane Research, Lucknow

estimate of yield rate of crop at district level is obtained by taking the weighted mean of yield rates of crop, the weights being stratum level estimates of total area under crop. The estimate of production of crop in a district is obtained by multiplying the weighted mean yield rate with estimate of area at the district level (NSSO, 1978). The example of situation II is also available, in certain States of India, where advance estimates of areas, on the basis of sampling of villages under "Timely Reporting scheme of Area Statistics", are available at the beginning of season but towards end, estimates of areas on basis of complete enumeration of all the villages are available.

There are thus two separate characteristics (i) the characteristic under study and (ii) the auxiliary one and each characteristic has its own population. Both the population are divided into the number of strata. The sampling scheme adopted is stratified simple random sampling with or without replacement where independent samples are drawn for each characteristic within each stratum in order to estimate the stratum level means/totals of the two characteristics. Finally, the object is to estimate the weighted mean of the characteristic under study where weights are stratum level totals of the auxiliary characteristic.

Let y_{h_i} denote the value of the characteristic under study for the i-th unit in the h-th stratum, i = 1, 2, . . ., N_h and h = 1, 2, . . ., L where N_h and n_h are the population size and the sample size respectively in the h-th stratum such that $\sum_{h=1}^{L} N_h = N$ and $\sum_{h=1}^{L} n_h = n$ and L denotes the number of strata. Also let a_{h_j} denote the value of the auxiliary characteristic, which is used as weight, for the j-th unit in the h-th stratum, j=1, 2, ..., M and h=1, 2, ..., L where M_h and m_h are the population size and the sample size respectively in the h-th stratum such that $\sum_{h=1}^{L} M_h = M$ and $\sum_{h=1}^{L} m_h = m$. Strata for estimating the two characteristics are same.

As usual, we denote

$$\bar{Y}_h = \frac{1}{N_h} \sum_{i=1}^{N_h} Y_{h_i}, \quad \bar{y}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} Y_{h_i}$$

$$S^2_{y_h} = \frac{1}{N_h - 1} \sum_{i=1}^{N_h} (Y_{h_i} - \bar{Y}_h)^2$$

$$\bar{A}_h = \frac{1}{M_h} \sum_{j=1}^{M_h} a_{h_j}, \quad \bar{a}_h = \frac{1}{m_h} \sum_{i=1}^{m_h} a_{h_i}$$

$$\bar{A} = \frac{1}{M} \sum_{h=1}^{L} M_h \bar{A}_h, \quad \bar{a} = \frac{1}{M} \sum_{h=1}^{L} M_h \bar{a}_h$$

$$S^2_{a_h} = \frac{1}{M_h - 1} \sum_{i=1}^{M_h} (a_{h_i} - \bar{A}_h)^2$$

$$\bar{Q} = \frac{1}{M\bar{A}} \sum_{h=1}^{L} M_h \bar{A}_h \bar{Y}_h \text{ and } \bar{q}_{st} = \frac{1}{M\bar{a}} \sum_{h=1}^{L} M_h \bar{a}_h \bar{y}_h$$

If we put

$$W_h = \frac{M_h \bar{A}_h}{M\bar{A}} \quad \text{and} \quad w_h = \frac{M_h \bar{a}_h}{M\bar{a}}$$

we get

$$\bar{Q} = \sum_{h=1}^{L} W_h \bar{Y}_h \text{ and } \bar{q}_{st} = \sum_{h=1}^{L} w_h \bar{y}_h$$

The object is to estimate $\bar{Q}$. The estimates $\bar{q}_{st}$ of $\bar{Y}$ can be interpreted in two ways, viz. (A) $\bar{q}_{st}$ can be taken as the sum of products of two variables, w_h and $\bar{y}_h$ and (B) $\bar{q}_{st}$ can be taken as a ratio of two variables $\Sigma M_h \bar{a}_h \bar{y}_h$ and $M\bar{a}$. As stated earlier, there can also be two situation viz (i) when at the planning as well as at the estimation stage only $\bar{a}_h$ is known and (ii) at the planning stage $\bar{a}_h$ is known but at the estimation stage $\bar{A}_h$ is known. Thus the following three cases arise:

Situation I : only $\bar{a}_h$ is known at the planning as well as at the estimation stages:

(A) $\bar{q}_{st}$ is taken as the product of two variables

(B) $\bar{q}_{st}$ is taken as the ratio of two variables.

Situation II : $\bar{a}_h$ is known at the planning stage but $\bar{A}_h$ is known at the estimation stage.

(A) $\bar{q}_{st}$ is taken as the product of two variables.

2. *Situation I – Method (A)*

Under this case

$$\bar{q}_{st} = \sum_{h=1}^{L} w_h \bar{y}_h = \sum_{h=1}^{L} \left(\frac{M_h \bar{a}_h}{\sum M_h \bar{a}_h} \right) \bar{y}_h$$

and

$$E(\bar{q}_{st}) = E\left(\sum w_h E(\bar{y}_h) \right) = \sum \bar{Y}_h E(w_h) \qquad (2.1)$$

Since w_h is not an unbiased estimator of W_h, $\bar{q}_{st}$ is not an unbiased estimator of $\bar{Q}$. In order to find the bias of w_h, let $\bar{a}_h = \bar{A}_h + \bar{e}_h$ such that $E(\bar{e}_h) = 0$.

Hence
$$w_h = \frac{M_h \bar{a}}{\sum M_h \bar{a}_h}$$

$$= \frac{M_h(\bar{A}_h + \bar{e}_h)}{\sum M_h(\bar{A}_h + \bar{e}_h)}$$

$$= M_h \bar{A}_h \left(1 + \frac{\bar{e}_h}{\bar{A}_h}\right) \left[\sum M_h \bar{A}_h \left(\frac{1 + \sum M_h \bar{e}_h}{\sum M_h \bar{A}_h} \right) \right]^{-1}$$

$$= W_h \left[1 + \left(\frac{\bar{e}_h}{\bar{A}_h}\right) - \left(\frac{\sum M_h \bar{e}_h}{\sum M_h \bar{A}_h} \right) - \left(\frac{\bar{e}_h}{\bar{A}_h} \right) \left(\frac{\sum M_h \bar{e}_h}{\sum M_h \bar{A}_h} \right) + \left(\frac{\sum M_h \bar{e}_h}{\sum M_h \bar{A}_h} \right)^2 + \ldots \right] \qquad (2.2)$$

Taking expectation of w_h from (2.2) and substituting in (2.1), we get, ignoring terms of higher powers than the second of e_h,

$$E(\bar{q}_{st}) = \sum W_h \bar{Y}_h \left[1 - \frac{M_h S_{y_h}^2}{n_h \bar{A}_h \sum M_h \bar{A}_h} + \frac{\sum M_h^2 S_{y_h}^2}{n_h \left(\sum M_h \bar{A}_h \right)^2} \right] \tag{2.3}$$

(2.3) shows that the bias in $\bar{q}_{st}$ will be small when sample sizes in each stratum are large.

In order to find the variance of $\bar{q}_{st}$, we first take average of $\bar{q}_{st}$ over samples in which w_h were fixed. Over these samples, the mean of $\bar{q}_{st}$ is $\sum w_h \bar{Y}_h$ so that there is a bias of amount $\sum (w_h - W_h) \bar{Y}_h$. The mean square error, ignoring finite population correction (f.p.c.), is given by

$$V\left(\bar{q}_{st} \mid w_h \right) = E\left[\frac{\left(\bar{q}_{st} - \bar{Q} \right)^2}{w_h} \right]$$

$$= \sum \left(\frac{w_h^2 S_{y_h}^2}{n_h} \right) + \left[\sum (w_h - W_h) \bar{Y}_h \right]^2 \tag{2.4}$$

Averaging over selections of w_h

$$V(\bar{q}_{st}) = E \left[V (\bar{q}_{st} \mid w_h) \right]$$

$$= \sum \left[V(w_h) + W_h^2 \right] \frac{S_{y_h}^2}{n_h} + \sum \bar{Y}_h^2 V(w_h) + \sum_{h \neq j} \bar{Y}_h \bar{Y}_j \, Cov(w_h, w_j)$$

(2.5)

From (2.2), ignoring terms of higher powers, we get

$$V(w_h) = W_h^2 E \left[\left(\frac{\bar{e}_h}{\bar{A}_h} \right) - \frac{\sum M_h \bar{e}_h}{\sum M_h \bar{A}_h} \right]^2$$

$$= W_h^2 \left[\left(\frac{S_{a_h}^2}{m_h \bar{A}_h^2} \right) + \frac{\sum M_h^2 S_{a_h}^2}{m_h \left(\sum M_h \bar{A}_h \right)^2} - \frac{2M_h S_{a_h}^2}{m_h \bar{A}_h \sum M_h \bar{A}_h} \right] \quad (2.6)$$

Similarly

$$\text{Cov}(w_h, w_j) = W_h W_j \left[- \frac{M_h S_{a_h}^2}{m_h \bar{A}_h \sum M_h \bar{A}_h} - \left\{ \frac{M_j S_{a_j}}{m_j \bar{A}_j \sum M_h \bar{A}_h} \right\} + \frac{\sum M_h S_{a_h}^2}{m_h \left(\sum M_h \bar{A}_h \right)^2} \right] \quad (2.7)$$

From (2.7), we get

$$\sum_{h \neq j} \bar{Y}_h \bar{Y}_j \text{Cov}(w_h, w_j) = - \left[\frac{\sum W_h \bar{Y}_h M_h S_{a_h}^2}{m_h \bar{A}_h \sum M_h \bar{A}_h} \sum W_h \bar{Y}_h - \frac{\sum W_h^2 \bar{Y}_h^2 M_h S_{a_h}}{m_h \bar{A}_h \sum M_h \bar{A}_h} - \left\{ \frac{\sum W_h \bar{Y}_h M_h S_{a_h}^2}{m_h \bar{A}_h \sum M_h \bar{A}_h} \sum W_h \bar{Y}_h - \frac{\sum W_h^2 \bar{Y}_h^2 M_h S_{a_h}^2}{m_h \bar{A}_h \sum M_h \bar{A}_h} \right\} + \left\{ \left(\sum W_h \bar{Y}_h \right)^2 - \sum W_h^2 \bar{Y}_h^2 \right\} \frac{\sum M_h^2 S_{a_h}^2}{m_h \left(\sum M_h \bar{A}_h \right)^2} \right] \quad (2.8)$$

Substituting from (2.6) and (2.8) in (2.5), we get

$$V_1(\bar{q}_{st}) = \sum \left(\frac{W_h^2 S_{y_h}^2}{n_h} \right) + \sum \left(\frac{W_h^2 S_{y_h}^2 S_{a_h}^2}{n_h m_h A_h^2} \right)$$

$$+ \sum \frac{W_h^2 S_{y_h}^2}{n_h} \sum \frac{M_h^2 S_{a_h}^2}{m_h} \left(\sum M_h \bar{A}_h \right)^2$$

$$- 2 \sum \left(\frac{W_h^2 S_{y_h}^2 M_h S_{a_h}^2}{n_h m_h \bar{A}_h \sum M_h \bar{A}_h} \right) + \sum \frac{W_h^2 \bar{Y}_h^2 S_{a_h}^2}{m_h \bar{A}_h^2}$$

$$+ \left(\sum W_h \bar{Y}_h \right)^2 \sum \frac{\frac{M_h^2 S_{a_h}^2}{m_h}}{\left(\sum M_h \bar{A}_h \right)^2}$$

$$- 2 \sum M_h \bar{Y}_h \frac{W_h \bar{Y}_h M_h S_{a_h}^2}{m_h \bar{A}_h \sum M_h \bar{A}_h} \qquad (2.9)$$

Allocation of the sample

From (2.9), it can be seen that $V_1(\bar{q}_{st})$ will be minimum when

$$n_h \alpha W_h S_{y_h} \left[1 + \frac{S_{a_h}^2}{m_h \bar{A}_h^2} + \frac{\sum M_h^2 S_{a_h}^2}{m_h \left(\sum M_h \bar{A}_h \right)^2} - \frac{2 M_h S_{a_h}^2}{m_h \bar{A}_h \sum M_h \bar{A}_h} \right]^{1/2} \qquad (2.10)$$

or

$$n_h \alpha W_h S_{y_h} K_h, \text{ say.}$$

It may be noted that values of $W_h S_{y_h} K_h$ are not known. However, estimates of $W_h K_h$ are available. It may be desirable to allocate n_h proportional to estimates of $W_h K_h$ since such an allocation may not be far off from optimality if S_{y_h} do not vary considerably over strata.

If we allocate n_h proportional to w_h, since values of W_h are not known, substituting $n_h \alpha W_h$ in (2.4) and then taking expectation with respect to w_h we get,

$$V_{1\,prop}(\bar{q}_{st}) = \frac{1}{n} \sum W_h S_{y_h}^2 + \sum \bar{Y}_h^2 V(w_h) + \sum_{h \neq j} \bar{Y}_h \bar{Y}_j \, Cov(w_h w_j) \qquad (2.11)$$

$$= \frac{1}{n}\sum W_h^2 S_{y_h}^2 + \sum \frac{W_h^2 \bar{Y}_h^2 S_{a_h}^2}{m_h \bar{A}_h^2}$$

$$+ \left(\sum W_h \bar{Y}_h\right)^2 \sum \frac{M_h^2 S_{a_h}^2}{m_h} \Big/ \left(\sum M_h \bar{A}_h\right)^2$$

$$- \sum W_h \bar{Y}_h \sum \frac{W_h \bar{Y}_h S_{a_h}^2}{m_h \bar{A}_h \sum M_h \bar{A}_h} \qquad (2.12)$$

A comparison of (2.5) and (2.11) shows that the positive term, $\sum V(w_h) \frac{S_{y_h}^2}{n_h}$ occuring in (2.5), does not occur in (2.11). The other difference is that in (2.5) we get $\sum (w_h^2) / \frac{S_{y_h}^2}{n_h}$ whereas in (2.11), the corresponding term is $\frac{1}{n}\sum w_h S_{y_h}^2$. The minimum value of $\sum w_h^2 \frac{S_{y_h}^2}{n_h}$ is $\frac{1}{n}\left(\sum w_h S_{y_h}\right)^2$ and $\frac{1}{n}\sum w_h S_{y_h}^2$ may not be far from $\frac{1}{n}\left(\sum w_h S_{y_h}\right)^2$ and S_{y_h} do not very much vary over strata. Thus, it may even be desirable to allocate $n_h \alpha w_h$.

3. *Situation I – Method (B)*

Under this case

$$\bar{q}_{st} = \frac{\sum M_h \bar{a}_h \bar{Y}_h}{\sum M_h \bar{a}_h} \qquad (3.1)$$

is taken as a ratio of two variables. Let

$$\bar{y}_h = \bar{Y}_h + \bar{d}_h \qquad \text{and} \qquad \bar{a}_h = \bar{A}_h + \bar{e}_h$$

Such that $E(d_h) = E(e_h) = O$. We get

$$\bar{q}_{st} = \frac{\sum M_h \bar{A}_h \bar{Y}_h \left[1 + \frac{\bar{e}_h}{\bar{A}_h}\right]\left[1 + \frac{\bar{d}_h}{\bar{Y}_h}\right]\left[1 + \frac{\sum M_h \bar{e}_h}{\sum M_h \bar{A}_h}\right]^{-1}}{\sum M_h \bar{A}_h}$$

$$= \sum M_h \bar{A}_h \bar{Y}_h \left[1 + \frac{\bar{e}_h}{\bar{A}_h} + \frac{\bar{d}_h}{\bar{Y}_h} - \frac{\sum M_h \bar{e}_h}{\sum M_h \bar{A}_h} + \frac{\bar{e}_h \bar{d}_h}{\bar{A}_h \bar{Y}_h} \right.$$

$$\left. - \frac{\bar{e}_h \sum M_h \bar{e}_h}{\bar{A}_h \sum M_h \bar{A}_h} - \frac{\bar{d}_h \sum M_h \bar{e}_h}{\bar{Y}_h \sum M_h \bar{A}_h} + \left(\frac{\sum M_h \bar{e}_h}{\sum M_h \bar{A}_h} \right)^2 \right] \frac{1}{\sum M_h \bar{A}_h} \tag{3.2}$$

after neglecting terms of higher powers than the second of $\bar{e}_h$ and $\bar{d}_h$. Taking expectation from (3.2), we get

$$E(\bar{q}_{st}) = \frac{\left[\sum M_h \bar{A}_h \bar{Y}_h \left\{ 1 - \frac{M_h S^2_{y_h}}{n_h \bar{A}_h \sum M_h \bar{A}_h} + \frac{\sum M_h S^2_{y_h}}{n_h (\sum M_h \bar{A}_h)^2} \right\} \right]}{\sum M_h \bar{A}_h} \tag{3.3}$$

(3.3) shows that the bias of $\bar{q}_{st}$ under this method will be small when sample sizes in each stratum are large.

The variance of $\bar{q}_{st}$ will be given by

$$V_2(\bar{q}_{st}) = \frac{\left(\sum M_h \bar{A}_h \bar{Y}_h \right)^2}{\left(\sum M_h \bar{A}_h \right)^2} \left[\frac{V\left(\sum M_h \bar{a}_h \bar{y}_h \right)}{\left(\sum M_h \bar{A}_h \bar{Y}_h \right)^2} \right.$$

$$\left. + \frac{V\left(\sum M_h \bar{a}_h \right)}{\left(\sum M_h \bar{A}_h \right)^2} - \frac{2 \sum \bar{Y}_h V(M_h \bar{a}_h)}{\sum M_h \bar{A}_h \bar{Y}_h \sum M_h \bar{A}_h} \right]$$

$$= \frac{1}{\left(\sum M_h \bar{A}_h \right)^2} \left[\frac{\sum M_h^2 \bar{A}_h^2 S^2_{y_h}}{n_h} + \frac{\sum M_h^2 \bar{Y}_h^2 S^2_{a_h}}{m_h} + \frac{\sum M_h^2 S^2_{a_h} S^2_{y_h}}{m_h n_h} \right.$$

$$\left. + \frac{\left(\sum M_h \bar{A}_h \bar{Y}_h \right)^2 \sum M_h^2 S^2_{a_h}}{m_h \left(\sum M_h \bar{A}_h \right)^2} - 2 \left(\sum M_h \bar{A}_h \bar{Y}_h \right)^2 \frac{\sum M_h^2 \bar{Y}_h S^2_{a_h}}{m_h \left(\sum M_h \bar{A}_h \right)} \right] \tag{3.4}$$

Allocation of the sample

From (3.4), it can be seen that $V_2(\bar{q}_{st})$ will be minimum when

$$n_h \alpha W_h S_{y_h} \left(1 + \frac{S_{a_h}^2}{m_h \bar{A}_h^2}\right)^{1/2}$$

It may be desirable to allocate n_h proportional to estimates of $W_h \left(1 + \frac{S_{a_h}^2}{m_h \bar{A}_h^2}\right)^{1/2}$ which are known. It may be noted that if we allocate n_h proportional to w_h, we get formula (2.12) under this method also.

Comparison of the two methods

A comparison of (2.3) and (3.3) will show that the bias of q_{st} under the two methods will be the same if we ignore terms with powers higher than the second of e_h and d_h. It can be easily varified that the bias will be the same even if we ignore terms with powers higher than the third of e_h and d_h.

From (2.9) and (3.4), it can be seen that

$$V_1(\bar{q}_{st}) - V_2(\bar{q}_{st}) = \left[\frac{\sum \frac{W_h S_h^2}{n_h} \sum \frac{M_h^2 S_{a_h}^2}{m_h}}{\left(\sum M_h \bar{A}_h\right)^2}\right] - \frac{2\sum W_h^2 S_h^2 M_h S_{a_h}^2}{n_h m_h \bar{A}_h \sum M_h \bar{A}_h} \quad (3.6)$$

The difference between the two formulae of variances is due to different types of approximations involved in the two methods. However, if we allocate $n_h \alpha w_h$, the two methods give the same formula.

4. Situation II – Method (A)

under this case,

$$\bar{q}_{st} = \sum W_h \bar{y}_h \quad (4.1)$$

and $$V\,\bar{q}_{st} = \sum \frac{W_h^2 S_{y_h}^2}{n_h} \quad (4.2)$$

if we ignore f.p.c. The optimum allocation of the sample will be when

$$n_h \;\alpha\; W_h \,.\, S_{y_h} \tag{4.3}$$

However, since S_{y_h} are not known but w_h are known, we can allocate $n_h \alpha w_h$. Putting $n_h \alpha w_h$ in (4.2) and taking expectation with respect to w_h, we get

$$V(\bar{q}_{st}) = \frac{1}{n} \sum W_h \; S_{y_h}^2 \; E\left(\frac{1}{w_h}\right)$$

Also

$$\frac{1}{w_h} = \left(\frac{\sum M_h \bar{A}_h}{M_h \bar{A}_h}\right)\left(1 + \frac{\sum M_h \bar{e}_h}{\sum M_h \bar{A}_h}\right)\left(1 + \frac{\bar{e}_h}{\bar{A}_h}\right)^{-1}$$

$$= \frac{1}{W_h}\left[1 - \frac{\bar{e}_h}{\bar{A}_h} + \frac{\sum M_h \bar{e}_h}{\sum M_h \bar{A}_h} + \left(\frac{\bar{e}_h}{\bar{A}_h}\right)^2 - \frac{\bar{e}_h \sum M_h \bar{e}_h}{\bar{A}_h \sum M_h \bar{A}_h}\right] \tag{4.5}$$

after ignoring terms of higher powers than the second of $\bar{e}_h$. Taking expectation of (4.5) and substituting in (4.4), we get

$$V(\bar{q}_{st}) = \frac{1}{n}\left[\frac{\sum W_h \, S_{y_h}^2 \, S_{a_h}^2}{m_h \bar{A}_h^2} - \frac{\sum W_h \, S_{y_h}^2 \, M_h^2 \, S_{a_h}^2}{m_h \bar{A}_h \sum M_h \bar{A}_h}\right] \tag{4.6}$$

ACKNOWLEDGEMENT

The authors are grateful to the referee for his valuable coments.

REFERENCES

[1] Dayal, Shambhu, 1981. Sampling for estimating weighted totals and averages, *Ann. Inst. Stat. Math.*, **33**, 1, A, 165-76.

[2] National Sample Survey Organisation (NSSO), India, 1978. Consolidated Resultes of Crop Estimation Surveys on Principal Crops, 1975-76.