Introduction to Sampling Theory

Lecture 18
Ratio Method of Estimation in Stratified Sampling

Shalabh

Department of Mathematics and Statistics

Indian Institute of Technology Kanpur

Slides can be downloaded from

http://home.iitk.ac.in/~shalab/sp

Ratio Estimator in Stratified Sampling:

Suppose a population of size *N* is divided into *k* strata.

The objective is to estimate the population mean $\bar{Y}$ using ratio method of estimation.

In such situation, a random sample of size n_i is being drawn from the i^{th} strata of size N_i on variable under study *Y* and auxiliary variable *X* using SRSWOR.

Ratio Estimator in Stratified Sampling:

Ratio Estimator in Stratified Sampling:

Let

y_{ij} : j^{th} observation on Y from i^{th} strata

x_{ij} : j^{th} observation on X from i^{th} strata $i = 1, 2, \ldots, k$; $j = 1, 2, \ldots, n_i$.

An estimator of $\bar{Y}$ based on the philosophy of stratified sampling can be derived in following two possible ways:

1. **Separate ratio estimator**
2. **Combined ratio estimator**

1. Separate Ratio Estimator

- **Employ first the ratio method of estimation separately in each strata and obtain ratio estimator $\hat{\bar{Y}}_{R_i}$ $i = 1, 2, .., k$ assuming the stratum mean $\bar{X}_i$ to be known.**

- **Then combine all the estimates using weighted arithmetic mean.**

This gives the separate ratio estimator.

1. Separate Ratio Estimator

This gives the separate ratio estimator as

$$\hat{\bar{Y}}_{Rs} = \sum_{i=1}^{k} \frac{N_i \hat{\bar{Y}}_{R_i}}{N} = \sum_{i=1}^{k} w_i \hat{\bar{Y}}_{R_i} = \sum_{i=1}^{k} w_i \frac{\bar{y}_i}{\bar{x}_i} \bar{X}_i$$

where

$\bar{y}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}$: **sample mean of *Y* from *i*th strata**

$\bar{x}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij}$: **sample mean of *X* from *i*th strata**

$\bar{X}_i = \frac{1}{N_i}\sum_{j=1}^{N_i} x_{ij}$: **mean of all the *X* units in *i*th strata**

No assumption is made that the true ratio remains constant from stratum to stratum. It depends on information on each $\bar{X}_i$.

2. Combined Ratio Estimator

- **Find first the stratum mean of $Y's$ and $X's$ as**

$$\overline{y}_{st} = \sum_{i=1}^{k} w_i \overline{y}_i \quad , \qquad \overline{x}_{st} = \sum_{i=1}^{k} w_i \overline{x}_i .$$

- **Then define the combined ratio estimator as**

$$\hat{\overline{Y}}_{Rc} = \frac{\overline{y}_{st}}{\overline{x}_{st}} \overline{X}$$

where $\overline{X}$ is the population mean of *X* based on all the $N = \sum_{i=1}^{k} N_i$ units.

It does not depend on individual stratum units.

It does not depend on information on each $\overline{X}_i$ but only on $\overline{X}$.

Properties of Separate Ratio Estimator: Bias

Note that there is an analogy between $\bar{Y} = \sum_{i=1}^{k} w_i \bar{Y}_i$ **and** $\bar{Y}_{Rs} = \sum_{i=1}^{k} w_i \bar{Y}_{Ri}.$

We already have derived the approximate bias of $\hat{\bar{Y}}_R = \frac{\bar{y}}{\bar{x}} \bar{X}$ **as**

$$E(\hat{\bar{Y}}_R) = \bar{Y} + \frac{\bar{Y}f}{n}(C_x^2 - \rho C_X C_Y).$$

So for $\hat{\bar{Y}}_{Ri}$ **, we can write**

$$E(\hat{\bar{Y}}_{Ri}) = \bar{Y}_i + \bar{Y}_i \frac{f_i}{n_i}(C_{ix}^2 - \rho_i C_{iX} C_{iY})$$

Properties of Separate Ratio Estimator: Bias

where $\bar{Y}_i = \frac{1}{N_i}\sum_{j=1}^{N_i} y_{ij}, \quad \bar{X}_i = \frac{1}{N_i}\sum_{j=1}^{N_i} x_{ij}$

$$f_i = \frac{N_i - n_i}{N_i}, C_{iy}^2 = \frac{S_{iy}^2}{\bar{Y}_i^2}, \quad C_{ix}^2 = \frac{S_{ix}^2}{\bar{X}_i^2},$$

$$S_{iy}^2 = \frac{1}{N_i - 1}\sum_{j=1}^{N_i}(Y_{ij} - \bar{Y}_i)^2, \quad S_{ix}^2 = \frac{1}{N_i - 1}\sum_{j=1}^{N_i}(X_{ij} - \bar{X}_i)^2,$$

ρ_i : **correlation coefficient between the observation on *X* and *Y* in *i*th stratum**

C_{ix} : **coefficient of variation of *X* values in *i*th sample.**

Properties of Separate Ratio Estimator: Bias

Thus $$E(\hat{\bar{Y}}_{Rs}) = \sum_{i=1}^{k} w_i E(\hat{\bar{Y}}_{Ri})$$

$$= \sum_{i=1}^{k} w_i \left[\bar{Y}_i + \bar{Y}_i \frac{f_i}{n_i} (C_{ix}^2 - \rho_i C_{ix} C_{iy} \right]$$

$$= \bar{Y} + \sum_{i=1}^{k} \frac{w_i \bar{Y}_i f_i}{n_i} (C_{ix}^2 - \rho_i C_{ix} C_{iy})$$

$$Bias(\hat{\bar{Y}}_{Rs}) = E(\bar{Y}_{Rs}) - \bar{Y}$$

$$= \sum_{i=1}^{k} \frac{w_i \bar{Y}_i f_i}{n_i} C_{ix} (C_{ix} - \rho_i C_{iy}).$$

up to the second order of approximation.

Properties of Separate Ratio Estimator: MSE

Now we derive the *MSE* of $\hat{\bar{Y}}_{Rs}$.

We already have derived the approximate *MSE* of $\hat{\bar{Y}}_R$ earlier as

$$MSE(\hat{\bar{Y}}_R) = \frac{\bar{Y}^2 f}{n}(C_X^2 + C_Y^2 - 2\rho C_x C_y)$$

$$= \frac{f}{n(N-1)}\sum_{i=1}^{N}(Y_i - RX_i)^2 \quad \text{where} \quad R = \frac{\bar{Y}}{\bar{X}}.$$

Thus the *MSE* of ratio estimator up to the second order of approximation based on the i^{th} stratum is

$$MSE(\hat{\bar{Y}}_{Ri}) = \frac{f_i \bar{Y}_i^2}{n_i(N_i - 1)}(C_{iX}^2 + C_{iY}^2 - 2\rho_i C_{iX} C_{iY})$$

$$= \frac{f_i}{n_i(N_i - 1)}\sum_{i=1}^{N_i}(Y_{ij} - R_i X_{ij})^2.$$

Properties of Separate Ratio Estimator: MSE

and so

$$MSE(\hat{\bar{Y}}_{Rs}) = \sum_{i=1}^{k} w_i^2 \, MSE(\hat{\bar{Y}}_{Ri})$$

$$= \sum_{i=1}^{k} \left[\frac{w_i^2 f_i}{n_i} \bar{Y}_i^2 (C_{iX}^2 + C_{iY}^2 - 2\rho_i C_{iX} C_{iY}) \right]$$

$$= \sum_{i=1}^{k} \left[w_i^2 \frac{f_i}{n_i(N_i - 1)} \sum_{j=1}^{N_i} (Y_{ij} - R_i X_{ij})^2 \right].$$

Properties of Separate Ratio Estimator: Estimator of MSE

An estimate of $MSE(\hat{\bar{Y}}_{Rs})$ can be found by substituting the unbiased estimators of S_{iX}^2, S_{iY}^2 and S_{iXY}^2 as s_{ix}^2, s_{iy}^2 and s_{ixy}, respectively for *i*th stratum and $R_i = \bar{Y}_i / \bar{X}_i$ can be estimated by $r_i = \bar{y}_i / \bar{x}_i$.

$$\widehat{MSE}(\hat{\bar{Y}}_{Rs}) = \sum_{i=1}^{k} \left[\frac{w_i^2 f_i}{n_i} (s_{iy}^2 + r_i^2 s_{ix}^2 - 2 r_i s_{ixy}) \right].$$

Also

$$\widehat{MSE}(\hat{\bar{Y}}_{Rs}) = \sum_{i=1}^{k} \left[\frac{w_i^2 f_i}{n_i(n_i - 1)} \sum_{j=1}^{n_i} (y_{ij} - r_i x_{ij})^2 \right].$$

Properties of Combined Ratio Estimator:

Here

$$\hat{\bar{Y}}_{RC} = \frac{\sum_{i=1}^{k} w_i \bar{y}_i}{\sum_{i=1}^{k} w_i \bar{x}_i} \bar{X} = \frac{\bar{y}_{st}}{\bar{x}_{st}} \bar{X} = \hat{R}_c \bar{X}.$$

It is difficult to find the exact expression of bias and mean squared error of $\hat{\bar{Y}}_{Rc}$, so we find their approximate expressions.

Properties of Combined Ratio Estimator:

Define

$$\varepsilon_1 = \frac{\bar{y}_{st} - \bar{Y}}{\bar{Y}}, \qquad \varepsilon_2 = \frac{\bar{x}_{st} - \bar{X}}{\bar{X}}$$

$$E(\varepsilon_1) = 0, \qquad E(\varepsilon_2) = 0$$

$$E(\varepsilon_1^2) = \sum_{i=1}^{k} \frac{N_i - n_i}{N_i n_i} \frac{w_i^2 S_{iY}^2}{\bar{Y}^2} = \sum_{i=1}^{k} \frac{f_i}{n_i} \frac{w_i^2 S_{iY}^2}{\bar{Y}^2}$$

$$E(\varepsilon_2^2) = \sum_{i=1}^{k} \frac{f_i}{n_i} \frac{w_i^2 S_{iX}^2}{\bar{X}^2}$$

$$E(\varepsilon_1 \varepsilon_2) = \sum_{i=1}^{k} \frac{w_i^2 f_i}{n_i} \frac{S_{iXY}}{\bar{X}\bar{Y}}.$$

Properties of Combined Ratio Estimator:

Thus assuming $|\varepsilon_2| < 1,$

$$\hat{\bar{Y}}_{RC} = \frac{(1+\varepsilon_1)\bar{Y}}{(1+\varepsilon_2)\bar{X}}\bar{X}$$

$$= \bar{Y}(1+\varepsilon_1)(1-\varepsilon_2+\varepsilon_2^2-...)$$

$$= \bar{Y}(1+\varepsilon_1-\varepsilon_2-\varepsilon_1\varepsilon_2+\varepsilon_2^2-...).$$

Expanding and retaining the terms up to order two .

$$\hat{\bar{Y}}_{RC} \simeq \bar{Y}(1+\varepsilon_1-\varepsilon_2-\varepsilon_1\varepsilon_2+\varepsilon_2^2)$$

$$\hat{\bar{Y}}_{RC} - \bar{Y} \simeq \bar{Y}(\varepsilon_1-\varepsilon_2-\varepsilon_1\varepsilon_2+\varepsilon_2^2).$$

Properties of Combined Ratio Estimator: Bias

The approximate bias of $\hat{\bar{Y}}_{Rc}$ up to second order of approximation is

$$Bias(\hat{\bar{Y}}_{Rc}) = E(\hat{\bar{Y}}_{Rc} - \bar{Y})$$

$$\simeq \bar{Y}E(\varepsilon_1 - \varepsilon_2 - \varepsilon_1\varepsilon_2 + \varepsilon_2^2)$$

$$= \bar{Y}\left[0 - 0 - E(\varepsilon_1\varepsilon_2) + E(\varepsilon_2^2)\right]$$

$$= \bar{Y}\sum_{i=1}^{k}\left[\frac{f_i}{n_i}w_i^2\left(\frac{S_{iX}^2}{\bar{X}^2} - \frac{S_{iXY}}{\bar{X}\bar{Y}}\right)\right]$$

$$= \bar{Y}\sum_{i=1}^{k}\left[\frac{f_i}{n_i}w_i^2\left(\frac{S_{iX}^2}{\bar{X}^2} - \frac{\rho_i S_{iX}S_{iY}}{\bar{X}\bar{Y}}\right)\right]$$

$$= \frac{\bar{Y}}{\bar{X}}\sum_{i=1}^{k}\left[\frac{f_i}{n_i}w_i^2 S_{iX}\left(\frac{S_{iX}}{\bar{X}} - \frac{\rho_i S_{iY}}{\bar{Y}}\right)\right]$$

$$= R\sum_{i=1}^{k}\left[\frac{f_i}{n_i}w_i^2 S_{iX}\left(C_{iX} - \rho_i C_{iY}\right)\right]$$

Properties of Combined Ratio Estimator: Bias

Here $R = \frac{\bar{Y}}{\bar{X}}, \ \rho_i$ **is the correlation coefficient between the observations on *Y* and *X* in the *i*th stratum,**

C_{ix} **and** C_{iy} **are the coefficients of variation of *X* and *Y* respectively in the *i*th stratum.**

Properties of Combined Ratio Estimator: MSE

The mean squared error up to second order of approximation is

$$
\begin{aligned}
MSE(\hat{\bar{Y}}_{Rc}) &= E(\hat{\bar{Y}}_{Rc} - \bar{Y})^2 \\
&\simeq \bar{Y}^2 E(\varepsilon_1 - \varepsilon_2 - \varepsilon_1\varepsilon_2 + \varepsilon_2)^2 \\
&\simeq \bar{Y}^2 E(\varepsilon_1^2 + \varepsilon_2^2 - 2\varepsilon_1\varepsilon_2) \\
&= \bar{Y}^2 \sum_{i=1}^{k} \left[\frac{f_i}{n_i} w_i^2 \left(\frac{S_{iX}^2}{\bar{X}^2} + \frac{S_{iY}^2}{\bar{Y}^2} - \frac{2S_{iXY}}{\bar{X}\bar{Y}} \right) \right] \\
&= \bar{Y}^2 \sum_{i=1}^{k} \left[\frac{f_i}{n_i} w_i^2 \left(\frac{S_{iX}^2}{\bar{X}^2} + \frac{S_{iY}^2}{\bar{Y}^2} - 2\rho_i \frac{S_{iX}}{\bar{X}} \frac{S_{iY}}{\bar{Y}} \right) \right] \\
&= \frac{\bar{Y}^2}{\bar{Y}^2} \sum_{i=1}^{k} \left[\frac{f_i}{n_i} w_i^2 \left(\frac{\bar{Y}^2}{\bar{X}^2} S_{iX}^2 + S_{iY}^2 - 2\rho_i \frac{\bar{Y}}{\bar{X}} S_{iX} S_{iY} \right) \right] \\
&= \sum_{i=1}^{k} \left[\frac{f_i}{n_i} w_i^2 (R^2 S_{iX}^2 + S_{iY}^2 - 2\rho_i R S_{iX} S_{iY}) \right].
\end{aligned}
$$

Properties of Combined Ratio Estimator: Estimator of MSE

An estimate of $MSE(\hat{\bar{Y}}_{Rc})$ **can be obtained by replacing** S_{iX}^2, S_{iY}^2 **and** S_{iXY}

by their unbiased estimators s_{ix}^2, s_{iy}^2 **and** s_{ixy} **respectively whereas**

$R = \frac{\bar{Y}}{\bar{X}}$ **is replaced by** $r = \frac{\bar{y}}{\bar{x}}$.

Thus the following estimate is obtained:

$$\widehat{MSE}(\hat{\bar{Y}}_{Rc}) = \sum_{i=1}^{k} \left[\frac{w_i^2 f_i}{n_i} \left(r^2 s_{ix}^2 + s_{iy}^2 - 2 r s_{ixy} \right) \right]$$

Comparison of Combined and Separate Ratio Estimators:

An obvious question arises that which of the estimates $\hat{\bar{Y}}_{Rs}$ or $\hat{\bar{Y}}_{Rc}$ is better.

So we compare their *MSE*s.

Note that the only difference in the term of these *MSE*s is due to the form of ratio estimate. It is

* $R_i = \dfrac{\bar{y}_i}{\bar{x}_i}$ **in** $MSE(\hat{\bar{Y}}_{Rs})$

* $R = \dfrac{\bar{Y}}{\bar{X}}$ **in** $MSE(\hat{\bar{Y}}_{Rc})$.

Comparison of Combined and Separate Ratio Estimators:

Thus $\Delta = MSE(\hat{\bar{Y}}_{Rc}) - MSE(\hat{\bar{Y}}_{Rs})$

$$= \sum_{i=1}^{k}\left[\frac{w_i^2 f_i}{n_i}\left[(R^2 - R_i^2)S_{iX}^2 + 2(R_i - R)\rho_i S_{iX} S_{iY}\right]\right]$$

$$= \sum_{i=1}^{k}\left[\frac{w_i^2 f_i}{n_i}\left[(R - R_i)^2 S_{iX}^2 + 2(R - R_i)(R_i S_{iX}^2 - \rho_i S_{iX} S_{iY})\right]\right].$$

The difference Δ depends on

i. **The magnitude of the difference between the strata ratios** (R_i) **and whole population ratio (*R*).**

ii. **The value of** $(R_i S_{ix}^2 - \rho_i S_{ix} S_{iy})$ **is usually small and vanishes when the regression line of *y* on *x* is linear and passes through origin within each stratum.**

Comparison of Combined and Separate Ratio Estimators:

The value of $(R_i S_{ix}^2 - \rho_i S_{ix} S_{iy})$ is usually small and vanishes when the regression line of *y* on *x* is linear and passes through origin within each stratum.

See as follows:

$$R_i S_{ix}^2 - \rho_i S_{ix} S_{iy} = 0$$

$$R_i = \frac{\rho_i S_{ix} S_{iy}}{S_{ix}^2}$$

which is the estimator of the slope parameter in the regression of *y* on *x* in the *i*^th^ stratum.

In such a case $MSE(\hat{\bar{Y}}_{Rc}) > MSE(\hat{\bar{Y}}_{Rs})$

but $Bias(\hat{\bar{Y}}_{Rc}) < Bias(\hat{\bar{Y}}_{Rs}).$

Comparison of Combined and Separate Ratio Estimators

So unless R_i varies considerably, the use of $\hat{\bar{Y}}_{Rc}$ would provide an estimate of $\bar{Y}$ with negligible bias and the precision as good as $\hat{\bar{Y}}_{Rs}$.

- **If $R_i \neq R,\quad \hat{\bar{Y}}_{Rs}$ can be more precise but bias may be large.**
- **If $R_i \simeq R,\quad \hat{\bar{Y}}_{Rc}$ can be as precise as $\hat{\bar{Y}}_{Rs}$ but its bias will be small. It also does not require knowledge of $\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_k$.**