Sampling Theory

MODULE V

LECTURE - 14

RATIO AND PRODUCT METHODS OF ESTIMATION

DR. SHALABH
DEPARTMENT OF MATHEMATICS AND STATISTICS
INDIAN INSTITUTE OF TECHNOLOGY KANPUR
An important objective in any statistical estimation procedure is to obtain the estimators of parameters of interest with more precision. It is also well understood that incorporation of more information in the estimation procedure yields better estimators, provided the information is valid and proper. Use of such auxiliary information is made through the ratio method of estimation to obtain an improved estimator of population mean. In ratio method of estimation, auxiliary information on a variable is available which is linearly related to the variable under study and is utilized to estimate the population mean.

Let $Y$ be the variable under study and $X$ be any auxiliary variable which is correlated with $Y$. The observation $x_i$ on $X$ and $y_i$ on $Y$ are obtained for each sampling unit. The population mean $\bar{X}$ of $X$ (or equivalently the population total $X_{tot}$) must be known. For example, $x_i$'s may be the values of $y_i$'s from

- some earlier completed census,
- some earlier surveys,
- some characteristic on which it is easy to obtain information etc.

For example, if $y_i$ is the quantity of fruits produced in the $i^{th}$ plot, then $x_i$ can be the area of $i^{th}$ plot or the production of fruit in the same plot in previous year.
Let \((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\) be the random sample of size \(n\) on paired variable \((X, Y)\) drawn, preferably by SRSWOR, from a population of size \(N\). The ratio estimate of population mean \(\bar{Y}\) is

\[
\hat{Y}_R = \frac{\bar{Y}}{\bar{X}} = \hat{R}\bar{X}
\]

assuming the population mean \(\bar{X}\) is known. The ratio estimator of population total \(Y_{tot} = \sum_{i=1}^{N} Y_i\) is

\[
\hat{Y}_{R(tot)} = \frac{Y_{tot}}{X_{tot}} X_{tot}
\]

where \(X_{tot} = \sum_{i=1}^{N} X_i\) is the population total of \(X\) which is assumed to be known, \(Y_{tot} = \sum_{i=1}^{n} Y_i\) and \(X_{tot} = \sum_{i=1}^{n} X_i\) are the sample totals of \(Y\) and \(X\) respectively. The \(\hat{Y}_{R(tot)}\) can be equivalently expressed as

\[
\hat{Y}_{R(tot)} = \frac{\bar{Y}}{\bar{X}} X_{tot}
\]

\[
= \hat{R}X_{tot}.
\]

Looking at the structure of ratio estimators, note that the ratio method estimates the relative change \(\frac{Y_{tot}}{X_{tot}}\) that occurred after \((x_i, y_i)\) were observed. It is clear that if the variation among the values of \(\frac{y_i}{x_i}\) is nearly same for all \(i = 1, 2, \ldots, n\) then values of \(\frac{Y_{tot}}{X_{tot}}\) (or equivalently \(\frac{\bar{Y}}{\bar{X}}\)) vary little from sample to sample and ratio estimate will be of high precision.
Bias and mean squared error of ratio estimator:

Assume that the random sample \((x_i, y_i), i = 1, 2, ..., n\) is drawn by SRSWOR and population mean \(\bar{X}\) is known. Then

\[
E(\hat{Y}_R) = \frac{1}{n \binom{N}{n}} \sum_{i=1}^{N} \frac{\bar{y}_i}{\bar{x}_i} \bar{X}
\]

\[\neq \bar{Y}\] (in general).

Moreover it is difficult to find the exact expression for \(E\left(\frac{\bar{y}}{\bar{x}}\right)\) and \(E\left(\frac{\bar{y}^2}{\bar{x}^2}\right)\). So we approximate them and proceed as follows:

Let

\[
\varepsilon_0 = \frac{\bar{y} - \bar{Y}}{\bar{Y}} \Rightarrow \bar{Y} = (1 + \varepsilon_0)\bar{Y}
\]

\[
\varepsilon_1 = \frac{\bar{x} - \bar{X}}{\bar{X}} \Rightarrow \bar{X} = (1 + \varepsilon_1)\bar{X}.
\]

Since SRSWOR is being followed, so

\[
E(\varepsilon_0) = 0
\]

\[
E(\varepsilon_1) = 0
\]
\[ E(\varepsilon_0^2) = \frac{1}{\bar{Y}^2} E((\bar{Y} - \bar{Y})^2) \]
\[ = \frac{1}{\bar{Y}^2} \frac{N-n}{Nn} S_y^2 \]
\[ = \frac{f}{n} \frac{S_y^2}{\bar{Y}^2} \]
\[ = \frac{f}{n} C_y^2 \]

where \( f = \frac{N-n}{N} \), \( S_y^2 = \frac{1}{N-1} \sum_{i=1}^{N} (Y_i - \bar{Y})^2 \) and \( C_y = \frac{S_y}{\bar{Y}} \) is the coefficient of variation related to \( Y \).

Similarly,
\[ E(\varepsilon_1^2) = \frac{f}{n} C_x^2 \]
\[ E(\varepsilon_0 \varepsilon_1) = \frac{1}{\bar{X}\bar{Y}} E[(\bar{X} - \bar{X})(\bar{Y} - \bar{Y})] \]
\[ = \frac{1}{\bar{X}\bar{Y}} \frac{N-n}{Nn} \frac{1}{N-1} \sum_{i=1}^{N} (X_i - \bar{X})(Y_i - \bar{Y}) \]
\[ = \frac{1}{\bar{X}\bar{Y}} \frac{f}{n} S_{xy} \]
\[ = \frac{1}{\bar{X}\bar{Y}} \frac{f}{n} \rho S_x S_y \]
\[ = \frac{f}{n} \rho \frac{S_x S_y}{\bar{X} \bar{Y}} \]
\[ = \frac{f}{n} \rho C_x C_y \]
where \( C_X = \frac{S_X}{\bar{X}} \) is the coefficient of variation related to \( X \) and \( \rho \) is the correlation coefficient between \( X \) and \( Y \).

Writing \( \hat{Y}_R \) in terms of \( \varepsilon \)'s, we get

\[
\hat{Y}_R = \frac{\bar{Y}}{\bar{X}} = \frac{(1 + \varepsilon_0)\bar{Y}}{(1 + \varepsilon_1)\bar{X}} = (1 + \varepsilon_0)(1 + \varepsilon_1)^{-1}\bar{Y}.
\]

Assuming \( |\varepsilon| < 1 \), the term \((1 + \varepsilon_1)^{-1}\) may be expanded as an infinite series and it would be convergent. Such assumption means that \( \left| \frac{\bar{x} - \bar{X}}{\bar{X}} \right| < 1 \), i.e., possible estimate \( \bar{x} \) of population mean \( \bar{X} \) lies between 0 and 2\( \bar{X} \). This is likely to hold true if the variation in \( \bar{x} \) is not large. In order to ensures that variation in \( \bar{x} \) is small, assume that the sample size \( n \) is fairly large.

With this assumption,

\[
\hat{Y}_R = \bar{Y}(1 + \varepsilon_0)(1 - \varepsilon_1 + \varepsilon_1^2 - ...)
= \bar{Y}(1 + \varepsilon_0 - \varepsilon_1 + \varepsilon_1^2 - \varepsilon_1\varepsilon_0 + ...).
\]

So the estimation error of \( \hat{Y}_R \) is

\[
\hat{Y}_R - \bar{Y} = \bar{Y}(\varepsilon_0 - \varepsilon_1 + \varepsilon_1^2 - \varepsilon_1\varepsilon_0 + ...).
\]

In case, when sample size is large, then \( \varepsilon_0 \) and \( \varepsilon_1 \) are likely to be small quantities and so the terms involving second and higher powers of \( \varepsilon_0 \) and \( \varepsilon_1 \) would be negligibly small.
In such a case
\[
\hat{Y}_R - \bar{Y} = \bar{Y}(\varepsilon_0 - \varepsilon_1)
\]
and
\[
E(\hat{Y}_R - \bar{Y}) = 0.
\]
So the ratio estimator is an unbiased estimator of population mean up to the first order of approximation.

If we assume that only terms of \( \varepsilon_0 \) and \( \varepsilon_1 \) involving powers more than two are negligibly small (which is more realistic than assuming that powers more than one are negligibly small), then the estimation error of \( \hat{Y}_R \) can be approximated as

\[
\hat{Y}_R - \bar{Y} \approx \bar{Y}(\varepsilon_0 - \varepsilon_1 + \varepsilon_1^2 - \varepsilon_1 \varepsilon_0)
\]

and

\[
E(\hat{Y}_R - \bar{Y}) = \bar{Y}\left(0 - 0 + \frac{f}{n} C_X^2 - \frac{f}{n} \rho C_X C_Y\right)
\]

\[
\text{Bias}(\hat{Y}) = E(\hat{Y}_R - \bar{Y}) = \frac{f}{n} \bar{Y} C_X (C_X - \rho C_Y)
\]

upto second order of approximation, the bias generally decreases as the sample size grows large.
The bias of $\hat{Y}_R$ is zero, i.e.,

$$\text{Bias}(\hat{Y}_R) = 0$$

if $E(\epsilon_i^2 - \epsilon_0 \epsilon_i) = 0$

or if

$$\frac{\text{Var}(\overline{x})}{\overline{X}^2} - \frac{\text{Cov}(\overline{x}, \overline{y})}{\overline{X}\overline{Y}} = 0$$

or if

$$\frac{1}{\overline{X}^2} \left[ \text{Var}(\overline{x}) - \frac{\overline{X}}{\overline{Y}} \text{Cov}(\overline{x}, \overline{y}) \right] = 0$$

or if $\text{Var}(\overline{x}) - \frac{\text{Cov}(\overline{x}, \overline{y})}{R} = 0$ (assuming $\overline{X} \neq 0$)

or if $R = \frac{\overline{Y}}{\overline{X}} = \frac{\text{Cov}(\overline{x}, \overline{y})}{\text{Var}(\overline{x})}$

which is satisfied when the regression line of $Y$ on $X$ passes through origin.

Now, to find the mean squared error, consider

$$MSE(\hat{Y}_R) = E(\hat{Y}_R - \overline{Y})^2$$

$$= E \left[ \overline{Y}^2 (\epsilon_0 - \epsilon_1 + \epsilon_i^2 - \epsilon_i \epsilon_0 + \ldots)^2 \right]$$

$$\approx E \left[ \overline{Y}^2 (\epsilon_0^2 + \epsilon_i^2 - 2\epsilon_0 \epsilon_i) \right].$$

Under the assumption $|\epsilon_i| < 1$ and the terms of $\epsilon_0$ and $\epsilon_1$ involving powers more than two are negligible small,

$$MSE(\hat{Y}_R) = \overline{Y}^2 \left[ \frac{f}{n} C_X^2 + \frac{f}{n} C_Y^2 - \frac{2f}{n} \rho C_X C_Y \right]$$

$$= \frac{\overline{Y}^2 f}{n} \left[ C_X^2 + C_Y^2 - 2\rho C_X C_Y \right]$$

up to the second order of approximation.