

Estimation of Finite Population Variance Using a Difference-Type Estimator with Optimum and Estimated Optimum Value

Rajiv Saksena, R. Karan Singh and Manu Priya

Department of Statistics, Lucknow University, Lucknow- 226 007

(Received July 12, 2002)

Abstract: An estimator using information on single auxiliary variable is proposed for estimating finite population variance. Its bias and mean square error (MSE) are obtained upto first order of approximation. Optimum value of characteristic scalar 'k' involved in estimator is also found. It has further been shown that the minimum mean square error of the proposed estimator remains unaffected when optimum value is estimated suitably from sample values. A comparative study along with the empirical one is also included.

1. Introduction

Auxiliary information had been very often used for enhancing the efficiency of estimators of population mean in the form of ratio, product and difference-type estimators. This idea has been taken a step further for estimating another very important parameter of the population namely the population variance.

Let y and x denote the main and auxiliary variables having population means $\bar{Y}$ and $\bar{X}$ respectively ($\bar{X}$ assumed to be known). Let a simple random sample of size n with y -values, $y_1, y_2, \dots, y_n$ on y and x -values $x_1, x_2, \dots, x_n$ on x be drawn without replacement from the population of size N and let $\bar{y}$ and $\bar{x}$ denote the sample means of y and x respectively.

$$\text{Let } s_y^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2 \text{ and } \bar{z} = \frac{1}{n} \sum_{i=1}^n z_i = \frac{1}{n} \sum_{i=1}^n y_i^2 \text{ with } z_i = y_i^2.$$

For (Y_i, X_i) ($i = 1, 2, \dots, N$) being the (y, x) - value respectively for the i^{th} unit of the population, let

$$(1.1) \quad \bar{Z} = \frac{1}{N} \sum_{i=1}^N Z_i; \quad Z_i = Y_i^2,$$

$$(1.2) \quad \mu_{rs} = \frac{1}{N} \sum_{i=1}^N (Y_i - \bar{Y})^r (X_i - \bar{X})^s,$$

$$(1.3) \quad S_y^2 = \frac{1}{(N-1)} \sum_{i=1}^N (Y_i - \bar{Y})^2,$$

and

$$(1.4) \quad \sigma_y^2 = \frac{1}{N} \sum_{i=1}^N (Y_i - \bar{Y})^2 = \frac{(N-1)}{N} S_y^2,$$

is the population variance of y .

For estimating finite population variance σ_y^2 , the proposed estimator $\hat{\sigma}_y^2$ is given by

$$(1.5) \quad \hat{\sigma}_y^2 = \frac{f}{(n-1)} \left[\sum_{i=1}^n y_i^2 - n \left\{ \bar{y}^2 + k \left(\frac{\bar{x}}{\bar{X}} - 1 \right) \right\} \right],$$

where $f = (N-1)/N$ and k is the characteristic scalar to be determined in such a way that the mean square error of $\hat{\sigma}_y^2$ is minimum.

It may be mentioned here that for $k = 0$, we get the usual unbiased estimator $s_y^{*2} = fs_y^2 = \frac{f}{(n-1)} \left[\sum_{i=1}^n y_i^2 - n\bar{y}^2 \right]$ of the population variance σ_y^2 as the special case of $\hat{\sigma}_y^2$.

2. Bias and MSE of $\hat{\sigma}_y^2$

Let $\bar{y} = \bar{Y} + e_0$, $\bar{x} = \bar{X} + e_1$, $\bar{z} = \bar{Z} + e_2$ so that $E(e_0) = E(e_1) = E(e_2) = 0$ and

$$E(e_0^2) = \left(\frac{1}{n} - \frac{1}{N} \right) \frac{N}{N-1} \mu_{20},$$

$$E(e_1^2) = \left(\frac{1}{n} - \frac{1}{N} \right) \frac{N}{N-1} \mu_{02},$$

$$E(e_0 e_1) = \left(\frac{1}{n} - \frac{1}{N} \right) \frac{N}{N-1} \mu_{11},$$

$$\begin{aligned} E(e_2^2) &= E(\bar{z} - \bar{Z})^2 \\ &= \left(\frac{1}{n} - \frac{1}{N} \right) S_z^2 \quad \text{where } S_z^2 = \frac{1}{N-1} \sum_{i=1}^N (Z_i - \bar{Z})^2 \end{aligned}$$

$$\begin{aligned}
&= \left(\frac{1}{n} - \frac{1}{N} \right) \frac{1}{N-1} \left(\sum_{i=1}^N Z_i^2 - N\bar{Z}^2 \right) \\
&= \left(\frac{1}{n} - \frac{1}{N} \right) \frac{1}{N-1} \left(\sum_{i=1}^N \{ (Y_i - \bar{Y}) + \bar{Y} \}^2 - N(\mu_{20} + \bar{Y}^2) \right) \\
&= \left(\frac{1}{n} - \frac{1}{N} \right) \frac{N}{N-1} (\mu_{40} + 4\bar{Y}\mu_{30} + 4\bar{Y}^2\mu_{20} - \mu_{20}^2),
\end{aligned}$$

$$\begin{aligned}
E(e_1 e_2) &= E(\bar{x} - \bar{X})(\bar{z} - \bar{Z}) \\
&= \left(\frac{1}{n} - \frac{1}{N} \right) \frac{1}{N-1} \sum_{i=1}^N (Z_i - \bar{Z})(X_i - \bar{X}) \\
&= \left(\frac{1}{n} - \frac{1}{N} \right) \frac{1}{N-1} \sum_{i=1}^N Z_i (X_i - \bar{X}) \\
&= \left(\frac{1}{n} - \frac{1}{N} \right) \frac{1}{N-1} \sum_{i=1}^N Y_i^2 (X_i - \bar{X}) \\
&= \left(\frac{1}{n} - \frac{1}{N} \right) \frac{1}{N-1} \sum_{i=1}^N \{ (Y_i - \bar{Y}) + \bar{Y} \}^2 (X_i - \bar{X}) \\
&= \left(\frac{1}{n} - \frac{1}{N} \right) \frac{N}{N-1} (\mu_{21} + 2\bar{Y}\mu_{11}).
\end{aligned}$$

Similarly,

$$E(e_0 e_2) = \left(\frac{1}{n} - \frac{1}{N} \right) \frac{N}{N-1} (\mu_{30} + 2\bar{Y}\mu_{20}).$$

Now from (1.5), we have

$$\begin{aligned}
\hat{\sigma}_y^2 &= \frac{nf}{n-1} \left[\frac{1}{n} \sum_{i=1}^n y_i^2 - \left\{ \bar{y}^2 + k \left(\frac{\bar{x}}{\bar{X}} - 1 \right) \right\} \right] \\
&= f \left(1 - \frac{1}{n} \right)^{-1} \left[\bar{z} - \left\{ \bar{y}^2 + k \left(\frac{\bar{x}}{\bar{X}} - 1 \right) \right\} \right] \\
&= f \left(1 - \frac{1}{n} \right)^{-1} \left[\bar{Z} + e_2 - \left\{ (\bar{Y} + e_0)^2 + k \frac{e_1}{\bar{X}} \right\} \right] \\
&= f \left(1 - \frac{1}{n} \right)^{-1} \left[\bar{Z} - \bar{Y}^2 + e_2 - 2\bar{Y}e_0 - e_0^2 - \frac{k}{\bar{X}} e_1 \right]
\end{aligned}$$

$$\begin{aligned}
&= \left(1 - \frac{1}{N}\right) \left(1 - \frac{1}{n}\right)^{-1} \left[\sigma_y^2 + e_2 - 2\bar{Y}e_0 - e_0^2 - \frac{k}{X}e_1 \right], \\
(2.1) \quad \hat{\sigma}_y^2 - \sigma_y^2 &= \left(1 - \frac{1}{N}\right) \left(1 + \frac{1}{n} + \frac{1}{n^2} + \dots\right) \\
&\quad \left[\left(\frac{N}{N-1}\right) \left(\frac{1}{n} - \frac{1}{N}\right) \sigma_y^2 + e_2 - 2\bar{Y}e_0 - e_0^2 - \frac{k}{X}e_1 \right].
\end{aligned}$$

Taking expectation on both sides, we have

$$(2.2) \quad E(\hat{\sigma}_y^2) - \sigma_y^2 = \left(1 + \frac{1}{n} + \frac{1}{n^2} + \dots\right) \left[\left(\frac{1}{n} - \frac{1}{N}\right) \sigma_y^2 - \left(\frac{1}{n} - \frac{1}{N}\right) \mu_{20} \right],$$

which is zero since $\mu_{20} = \sigma_y^2$. Therefore $\hat{\sigma}_y^2$ is an unbiased estimator of the finite population variance σ_y^2 .

Further, mean square error (MSE) of $\hat{\sigma}_y^2$ given by $E((\hat{\sigma}_y^2 - \sigma_y^2)^2)$ upto terms of order $O\left(\frac{1}{n}\right)$ is

$$\begin{aligned}
(2.3) \quad MSE(\hat{\sigma}_y^2) &= E\left(e_2 - 2\bar{Y}e_0 - \frac{k}{X}e_1\right)^2 \\
&= E\left(e_2^2 + 4\bar{Y}^2e_0^2 + \frac{k^2}{X^2}e_1^2 - 4\bar{Y}e_0e_2 - \frac{2k}{X}e_1e_2 + 4\bar{Y}\frac{k}{X}e_0e_1\right) \\
&= \left(\frac{1}{n} - \frac{1}{N}\right) \left[\mu_{40} + 4\bar{Y}\mu_{30} + 4\bar{Y}^2\mu_{20} - \mu_{20}^2 + 4\bar{Y}^2\mu_{20} + \frac{k^2}{X^2}\mu_{02} \right. \\
&\quad \left. - 4\bar{Y}(\mu_{30} + 2\bar{Y}\mu_{20}) - \frac{2k}{X}(\mu_{21} + 2\bar{Y}\mu_{11}) + 4\bar{Y}\frac{k}{X}\mu_{11} \right] \\
&= \left(\frac{1}{n} - \frac{1}{N}\right) \left[\mu_{20}^2 \{\beta_2(y) - 1\} + \frac{k}{X} \left\{ \frac{k}{X} \mu_{02} - 2\mu_{21} \right\} \right],
\end{aligned}$$

where $\beta_2(y) = \frac{\mu_{40}}{\mu_{20}^2}$.

3. Results Concerning Optimum and Estimated Optimum Value

(a) From (2.3), the optimum value of k which minimizes $MSE(\hat{\sigma}_y^2)$ is given by

$$(3.1) \quad k_0 = \frac{\bar{X}\mu_{21}}{\mu_{02}},$$

and the minimum mean square error of $\hat{\sigma}_y^2$ for the optimum value k_0 is

$$(3.2) \quad MSE(\hat{\sigma}_y^2)_{\min} = \left(\frac{1}{n} - \frac{1}{N} \right) \left[\mu_{20}^2 \{ \beta_2(y) - 1 \} - \frac{\mu_{21}^2}{\mu_{02}} \right].$$

Also, upto the terms of order $o\left(\frac{1}{n}\right)$, we have

$$(3.3) \quad MSE(s_y^{*2}) = \left(\frac{1}{n} - \frac{1}{N} \right) \mu_{20}^2 \{ \beta_2(y) - 1 \}.$$

Now, from (3.2) and (3.3), we have

$$(3.4) \quad MSE(\hat{\sigma}_y^2)_{\min} = MSE(s_y^{*2}) - \left(\frac{1}{n} - \frac{1}{N} \right) \frac{\mu_{21}^2}{\mu_{02}},$$

showing that the proposed estimator $\hat{\sigma}_y^2$ in this paper is more efficient than the usual conventional estimator s_y^{*2} in the sense of having lesser mean square error than that of s_y^{*2} .

(b) The optimum value k_0 in (3.1) or its guessed value may not always be known in practice, hence the alternative is to replace it by its estimate based on sample values. Thus, replacing μ_{21}, μ_{02} by their estimators

$$\hat{\mu}_{21} = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2 (x_i - \bar{X}) \quad \text{and} \quad \hat{\mu}_{02} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{X})^2 \quad \text{respectively, we obtain}$$

the estimated optimum value k_0 as

$$(3.5) \quad \hat{k}_0 = \frac{\bar{X}\hat{\mu}_{21}}{\hat{\mu}_{02}}.$$

The mean square error in case of estimated optimum k_0 is obtained as follows:

In view of (3.5), we have

$$(3.6) \quad \hat{\sigma}_y^{*2} = \frac{f}{n-1} \left[\sum_{i=1}^n y_i^2 - n \left\{ \bar{y}^2 + \hat{k}_0 \left(\frac{\bar{x}}{\bar{X}} - 1 \right) \right\} \right],$$

as our estimator of population variance σ_y^2 .

Let $\hat{\mu}_{21} = \mu_{21} + e_3$ and $\hat{\mu}_{02} = \mu_{02} + e_4$ so that

$$\begin{aligned}
\hat{\sigma}_y^{*2} &= \frac{f n}{n-1} \left[\frac{1}{n} \sum_{i=1}^n y_i^2 - \left\{ \bar{y}^2 + \frac{\bar{X} \hat{\mu}_{21}}{\hat{\mu}_{02}} \left(\frac{\bar{x}}{\bar{X}} - 1 \right) \right\} \right] \\
&= f \left(1 - \frac{1}{n} \right)^{-1} \left[\bar{z} - \left\{ \bar{y}^2 + \frac{\bar{X} \hat{\mu}_{21}}{\hat{\mu}_{02}} \left(\frac{\bar{x}}{\bar{X}} - 1 \right) \right\} \right] \\
&= f \left(1 - \frac{1}{n} \right)^{-1} \left[\bar{Z} + e_2 - \left\{ \bar{Y}^2 + e_0^2 + 2\bar{Y}e_0 + \frac{\mu_{21}}{\mu_{02}} e_1 \left(1 + \frac{e_3}{\mu_{21}} \right) \left(1 + \frac{e_4}{\mu_{02}} \right)^{-1} \right\} \right] \\
&= f \left(1 - \frac{1}{n} \right)^{-1} \left[\sigma_y^2 + e_2 - \left\{ e_0^2 + 2\bar{Y}e_0 + \frac{\mu_{21}}{\mu_{02}} e_1 \left(1 + \frac{e_3}{\mu_{21}} \right) \left(1 - \frac{e_4}{\mu_{02}} + \dots \right) \right\} \right] \\
&= \left(1 - \frac{1}{N} \right) \left(1 + \frac{1}{n} + \frac{1}{n^2} + \dots \right) \\
&\quad \left[\sigma_y^2 - \frac{\mu_{21}}{\mu_{02}} e_1 + \left\{ e_2 - e_0^2 - 2\bar{Y}e_0 + \frac{\mu_{21}}{\mu_{02}} e_1 e_3 - \frac{\mu_{21}}{\mu_{02}^2} e_1 e_4 + \dots \right\} \right],
\end{aligned}$$

or

$$\begin{aligned}
(3.7) \quad \hat{\sigma}_y^{*2} - \sigma_y^2 &= \left(\frac{1}{n} - \frac{1}{N} + \dots \right) \sigma_y^2 + \left(1 - \frac{1}{N} \right) \left(1 + \frac{1}{n} + \frac{1}{n^2} + \dots \right) \\
&\quad \left(e_2 - 2\bar{Y}e_0 - \frac{\mu_{21}}{\mu_{02}} e_1 - e_0^2 + \dots \right),
\end{aligned}$$

from which, it is clear that the bias of $\hat{\sigma}_y^{*2}$ given by $E(\hat{\sigma}_y^{*2}) - \sigma_y^2$ is of order $o\left(\frac{1}{n^2}\right)$ so that the bias is negligible for large values of n , and $MSE(\hat{\sigma}_y^{*2}) = E(\hat{\sigma}_y^{*2} - \sigma_y^2)^2$ upto terms of order $o\left(\frac{1}{n}\right)$ is

$$\begin{aligned}
(3.8) \quad MSE(\hat{\sigma}_y^{*2}) &= E \left(e_2 - 2\bar{Y}e_0 - \frac{\mu_{21}}{\mu_{02}} e_1 \right)^2 \\
&= \left(\frac{1}{n} - \frac{1}{N} \right) \left[\mu_{40} + 4\bar{Y}\mu_{30} + 4\bar{Y}^2\mu_{20} - \mu_{20}^2 + 4\bar{Y}^2\mu_{20} \right. \\
&\quad \left. + \frac{\mu_{21}^2}{\mu_{02}^2} \mu_{02} - 4\bar{Y}(\mu_{30} + 2\bar{Y}\mu_{20}) - 2\frac{\mu_{21}}{\mu_{02}}(\mu_{21} + 2\bar{Y}\mu_{11}) + 4\bar{Y}\frac{\mu_{21}}{\mu_{02}}\mu_{11} \right]
\end{aligned}$$

$$\begin{aligned}
 &= \left(\frac{1}{n} - \frac{1}{N} \right) \left[\mu_{20}^2 \{ \beta_2(y) - 1 \} - \frac{\mu_{21}^2}{\mu_{02}} \right] \\
 &= MSE(s_y^{*2}) - \left(\frac{1}{n} - \frac{1}{N} \right) \frac{\mu_{21}^2}{\mu_{02}},
 \end{aligned}$$

which is equal to $MSE(\hat{\sigma}_y^{*2})_{\min}$ given by (3.2). Moreover, the expression relating to $MSE(\hat{\sigma}_y^{*2})$ in (3.8), shows that the estimator $\hat{\sigma}_y^{*2}$ with estimated optimum value is more efficient than the usual estimator s_y^{*2} of finite population variance σ_y^2 in the sense of having smaller mean square error.

4. An Illustration

From the data given in Cochran¹ dealing with paralytic polio cases in “placebo” (y) group and paralytic polio cases in “not inoculated” group (X), the required values of μ_{rs} are calculated and relative efficiency of estimator $\hat{\sigma}_y^{*2}$ (or $\hat{\sigma}_y^2$) over the estimator s_y^{*2} for estimating population variance σ_y^2 is worked out as follows:

$$\begin{aligned}
 \mu_{20} &= 9.8894 & \mu_{21} &= 56.556573 \times 10^3 \\
 \mu_{02} &= 24.63885 \times 10^6 & \mu_{40} &= 421.96088.
 \end{aligned}$$

Using above values, we get

$$\begin{aligned}
 MSE(s_y^{*2}) &= \left(\frac{1}{n} - \frac{1}{N} \right) (326.8417), \\
 MSE(\hat{\sigma}_y^{*2}) &= MSE(\hat{\sigma}_y^2) = \left(\frac{1}{n} - \frac{1}{N} \right) (197.0205).
 \end{aligned}$$

The percentage relative efficiency (PRE) of the estimator $\hat{\sigma}_y^{*2}$ (or $\hat{\sigma}_y^2$) over the estimator s_y^{*2} is

$$\begin{aligned}
 PRE &= \frac{MSE(s_y^{*2})}{MSE(\hat{\sigma}_y^{*2})} \times 100\% \\
 &= 165.8922\%
 \end{aligned}$$

showing that the proposed estimator $\hat{\sigma}_y^{*2}$ (or $\hat{\sigma}_y^2$) is better than the estimator s_y^{*2} in the sense of having greater efficiency or lesser mean square error than that of the estimator s_y^{*2} of the finite population variance σ_y^2 .

References

1. W.G. Cochran: *Sampling Techniques*, Wiley Eastern Private Limited, New Delhi, 1977.
2. P.V. Sukhatme, B.V. Sukhatme, S. Sukhatme and C. Asok: *Sampling Theory of Surveys with Applications*, Iowa State University Press, IOWA, USA, 1984.