An "approximately unbiased" estimator may be uniformly larger than an" overestimate"

Robert G. Edson[1] and Gary M. Shapiro[2]

[1]VA Palo Alto Health Care System, Cooperative Studies Program Coordinating Center (151K), 701 N. Shoreline Blvd, Mountain View, CA 94043, U.S.A.; email bob.edson@va.gov
[2]Statistics Without Borders; email g.shapiro4@verizon.net

Abstract
This paper considers the best form of the collapsed stratum variance estimator. If you select 1 primary sampling unit per stratum and collapse 2 strata (say, A and B) together for each group, the usual estimator (Hansen et al., 1953) assigns a weight to stratum A's estimate equal to twice the measure associated with stratum B divided by the sum of the measures for the 2 strata. This estimator is known to overestimate the variance. We developed an alternative estimator that is approximately unbiased and assigns a weight to stratum A (B) of the square root of the ratio of stratum B's (A's) measure to its measure. We expected it would be a better estimator as an "approximately unbiased" estimate might be expected to produce better results than one known to be an overestimate. However, this paper shows the approximately unbiased estimator never results in a lower variance estimate than the overestimating variance estimator. A general cautionary conclusion from these results is that an "approximately unbiased estimator" can be worse and actually larger than an "overestimate estimator".

Keywords: collapsed stratum variance estimator, primary sampling unit, multi-stage sampling

1. Standard Collapsed Stratum Variance Estimator

The standard collapsed stratum variance estimator (Hansen et al, 1953, equation 5.2, p. 218) used when there is only one primary sampling unit (PSU) selected from each stratum is

$$s_x^2 = \sum_{g=1}^{G} \frac{L_g}{L_g - 1} \sum_{h=1}^{L_g} \left(x_{gh} - \frac{A_{gh}}{A_g} x_g \right)^2 , \text{where} \tag{1}$$

G is the number of groups,
L_g is the number of strata in group g,
$x_{gh} = \frac{x_{ghi}}{P_{ghi}}$ where x_{ghi} is an unbiased estimate of a total for and P_{ghi} is the probability of selecting PSU i in the sample from stratum h in group g,
$E(x_{gh}) = X_{gh}$, that is, x_{gh} is an unbiased estimator of the population total in stratum h within group g,
A_{gh} is a measure associated with stratum h within group g that tends to be highly correlated with X_{gh} (e.g., the population size),

$$A_g = \sum_h A_{gh}$$

$$x_g = \sum_h x_{gh}$$

and

$$E(s_x^2) = \sum_g \frac{L_g}{L_g - 1} \left[\sum_h \left(X_{gh} - \frac{A_{gh}}{A_g} X_g \right)^2 + \sigma_{x_g}^2 \left(1 + \sum_h \frac{A_{gh}^2}{A_g^2} \right) - 2 \sum_h \frac{A_{gh}}{A_g} \sigma_{x_{gh}}^2 \right] , \text{where} \tag{2}$$

$$X_g = \sum_h X_{gh} , \qquad \sigma_{x_g}^2 = \sum_h \sigma_{x_{gh}}^2$$

and $\sigma_{x_{gh}}^2$ is the variance of x_{gh} (Hansen et al., 1953, equation 5.13, page 221.)
When two strata are chosen from each group and collapsed together, that is, when $L_g = 2$ ∀ g, expanding equation (1) by each value of h produces the following simplified expression for s_x^2.

$$s_x^2 = \sum_g 2 \left[\left(x_{g1} - \frac{A_{g1}}{A_g} x_g \right)^2 + \left(x_{g2} - \frac{A_{g2}}{A_g} x_g \right)^2 \right] = \sum_g \left(w_{g2} x_{g1} - w_{g1} x_{g2} \right)^2 , \tag{3}$$

where $w_{g1} = \frac{2A_{g1}}{A_{g1} + A_{g2}}$ and $w_{g2} = \frac{2A_{g2}}{A_{g1} + A_{g2}} = 2 - w_{g1}$
Similarly, when $L_g = 2$ ∀ g, expanding equation (2) by each value of h produces the following simplified expression for $E(s_x^2)$.

$$E(s_x^2) = 2\sum_g \left\{ \frac{1}{A_g^2}\left[\left(A_{g2}X_{g1} - A_{g1}X_{g2}\right)^2 + \left(A_{g1}X_{g2} - A_{g2}X_{g1}\right)^2\right] \right. \quad (4)$$
$$\left. + \frac{\sigma_{x_{g1}}^2}{A_g^2}\left(A_g^2 + A_{g1}^2 + A_{g2}^2 - 2A_gA_{g1}\right) \right.$$
$$\left. + \frac{\sigma_{x_{g2}}^2}{A_g^2}\left(A_g^2 + A_{g1}^2 + A_{g2}^2 - 2A_gA_{g2}\right)\right\}$$
$$= 2\sum_g \left\{ \frac{2}{A_g^2}\left(A_{g2}X_{g1} - A_{g1}X_{g2}\right)^2 + \frac{\sigma_{x_{g1}}^2}{A_g^2}\left[\left(A_g - A_{g1}\right)^2 + A_{g2}^2\right] \right.$$
$$\left. + \frac{\sigma_{x_{g2}}^2}{A_g^2}\left[\left(A_g - A_{g2}\right)^2 + A_{g1}^2\right]\right\}$$
$$= \sum_g \left[\left(w_{g2}X_{g1} - w_{g1}X_{g2}\right)^2 + w_{g2}^2\sigma_{x_{g1}}^2 + w_{g1}^2\sigma_{x_{g2}}^2\right]$$

Hansen et al. (1953, Equation 5.5, page 220) shows that s_x^2 produces an overestimate of $\sigma_{x_g}^2$ so there is interest in determining if there are weights which would produce more accurate estimates of $\sigma_{x_g}^2$.

2. Alternative Collapsed Stratum Variance Estimator

$$E\left(x_{g1} - x_{g2}\right)^2 = Var\left(x_{g1} + x_{g2}\right) \quad (5)$$

if x_{g1} and x_{g2} are assumed to be independent, each is an estimate of one-half of the population total of interest, and $Ex_{g1} = Ex_{g2}$.

For non-self-representing (NSR) PSU's where the two strata within each group are collapsed and the estimates from the two strata are treated as x_{g1} and x_{g2} their expected values are not equal since the purpose of stratification is to maximize the difference between strata. One may instead propose to satisfy the modified condition

$E\left(p_{g2}x_{g1}\right) = E\left(p_{g1}x_{g2}\right)$ which is consistent with (6)

$$E\left(p_{g2}x_{g1} - p_{g1}x_{g2}\right)^2 = Var\left(p_{g2}x_{g1} + p_{g1}x_{g2}\right) \quad (7)$$

The goal is to determine values of p_{gh} which minimize

$$Var\left[\sum_g \left(p_{g2}x_{g1} + p_{g1}x_{g2}\right)\right] - Var\left[\sum_g \left(x_{g1} + x_{g2}\right)\right] \text{ where} \quad (8)$$

$\sum_g \left(x_{g1} + x_{g2}\right)$ is the unbiased estimate. (9)

The goal in pairing collapsed strata is to satisfy the following conditions as nearly as possible within each group.

$$E\left(\frac{x_{g1}}{A_{g1}}\right) = E\left(\frac{x_{g2}}{A_{g2}}\right) \text{ and} \tag{10}$$

$$S^2_{x_{g1}} = S^2_{x_{g2}} \text{ where} \tag{11}$$

$S^2_{x_{gh}}$ is the population variance among final stage sampling units in stratum h within group g.

Equations (6) and (10) show that we want the p_{gh} to satisfy the following equation.

$$p_{g2} = \frac{E(x_{g2})}{E(x_{g1})} p_{g1} = \frac{A_{g2}}{A_{g1}} p_{g1} \tag{12}$$

Let a_{gh} be the sample size of PSU's in stratum h within group g, then

$$Var(x_{g1}) \doteq \frac{A^2_{g1}}{a_{g1}} S^2_{x_{g1}} \text{ and } Var(x_{g2}) \doteq \frac{A^2_{g2}}{a_{g2}} S^2_{x_{g2}} \tag{13}$$

If we assume that the sample sizes are proportionate to the population, then

$$\frac{a_{g1}}{A_{g1}} = \frac{a_{g2}}{A_{g2}} = k \tag{14}$$

Equations (11), (13) and (14) imply that

$$Var(x_{g1}) \doteq \frac{A_{g1}}{A_{g2}} Var(x_{g2}) \tag{15}$$

Using the assumptions described above and equations (12) and (15),

$$\begin{aligned} B &= Var(p_{g2}x_{g1} + p_{g1}x_{g2}) - Var(x_{g1} + x_{g2}) \\ &\doteq Var(x_{g2}) \left[\left(\frac{A_{g2}}{A_{g1}} p_{g1} \right)^2 \frac{A_{g1}}{A_{g2}} + p^2_{g1} - \left(\frac{A_{g1}}{A_{g2}} + 1 \right) \right] \\ &\doteq Var(x_{g2}) \left[p^2_{g1} \left(\frac{A_{g1} + A_{g2}}{A_{g1}} \right) - \frac{A_{g1} + A_{g2}}{A_{g2}} \right] \end{aligned} \tag{16}$$

Equations (12) and (16) shows that $B \doteq 0$ when

$$p_{g1} = \frac{1}{p_{g2}} = \sqrt{\frac{A_{g1}}{A_{g2}}} = \sqrt{\frac{w_{g1}}{w_{g2}}} \tag{17}$$

3. Comparison of Expected Value of Variance Estimators

Since the p_{gh} produce an estimate that is almost unbiased, one might assume that the resulting estimate is better than the one produced by the w_{gh}. The estimator and expected value of the estimator resulting from substituting the p_{gh} for the w_{gh} in equations (3) and (4) are

$$v^2_x = \sum_g (p_{g2}x_{g1} - p_{g1}x_{g2})^2 \text{ and} \tag{18}$$

$$E v_x^2 = \sum_g \left[\left(p_{g2} X_{g1} - p_{g1} X_{g2} \right)^2 + p_{g2}^2 \sigma_{x_{g1}}^2 + p_{g1}^2 \sigma_{x_{g2}}^2 \right] \tag{19}$$

Noting that $\left(1 - w_{g1} w_{g2}\right) = \frac{\left(A_{g1} + A_{g2}\right)^2 - 4 A_{g1} A_{g2}}{\left(A_{g1} + A_{g2}\right)^2} = \left(\frac{A_{g1} - A_{g2}}{A_{g1} + A_{g2}} \right)^2$,
it can be shown from expanding equations (4) and (19) by term and the relationship in equation (17) that

$$\begin{aligned} E v_x^2 - E s_x^2 &= \sum_g \left(1 - w_{g1} w_{g2}\right) \left(\frac{w_{g2}}{w_{g1}} X_{g1}^2 + \frac{w_{g1}}{w_{g2}} X_{g2}^2 - 2 X_{g1} X_{g2} + \frac{w_{g2}}{w_{g1}} \sigma_{x_{g1}}^2 \right. \\ &\quad \left. + \frac{w_{g1}}{w_{g2}} \sigma_{x_{g2}}^2 \right) \\ &= \sum_g \left(\frac{A_{g1} - A_{g2}}{A_{g1} + A_{g2}} \right)^2 \left[\left(p_{g2} X_{g1} - p_{g1} X_{g2} \right)^2 + p_{g2}^2 \sigma_{x_{g1}}^2 + p_{g1}^2 \sigma_{x_{g2}}^2 \right] \\ &\geq 0 \end{aligned} \tag{20}$$

with the difference equal to 0 only when $A_{g1} = A_{g2}$ ∀ g .
Thus, the approximately unbiased estimator using the p_{gh} actually produces overestimates at least as extreme as the estimator that uses the w_{gh}.

4. Comparison of Variance Estimators

There is a similar relationship between $v_x^2$and s_x^2. For a given group g, there are positive numbers c_g and d_g such that

$$c_g A_{g1} = A_{g2} \text{ and } d_g x_{g1} = x_{g2} \tag{21}$$

Let $v_{x_g}^2$ and $s_{x_g}^2$ be the components from group g of v_x^2 and s_x^2, respectively, then from equation (21)

$$v_{x_g}^2 = \frac{\left(c_g - d_g\right)^2}{c_g} x_{g1}^2 \text{ and } s_{x_g}^2 = \left[\frac{2\left(c_g - d_g\right)}{\left(c_g + 1\right)} \right]^2 x_{g1}^2 \tag{22}$$

$$\text{So, } v_{x_g}^2 = s_{x_g}^2 \text{ if } c_g = d_g \text{ and} \tag{23}$$

$$v_{x_g}^2 > s_{x_g}^2 \text{ if } \frac{1}{c_g} > \frac{4}{\left(c_g + 1\right)^2}, \text{ i.e., if } \left(c_g - 1\right)^2 > 0. \tag{24}$$

From equations (23) and (24) it follows that

$$v_x^2 \geq s_x^2 \text{ with equality only if } \forall \text{ g}, \quad c_g = d_g \text{ or } c_g = 1 \tag{25}$$

$\left(\text{i.e., if } A_{g1} x_{g2} = A_{g2} x_{g2} \text{ or } A_{g1} = A_{g2}\right)$.

5. Conclusion

The general lesson here is that an estimator that is approximately unbiased is not necessarily preferable to one that is a known overestimate. It is possible that the approximately unbiased estimator is always an overestimate and indeed is always larger than the overestimate estimator.

Acknowledgments

The authors completed this research while employed by the US Bureau of the Census. Charles H. Alexander and Beverly D. Causey (both of whom passed away well before their time) provided important comments on the internal Bureau of the Census memoranda on which this paper is based.

References

[1] Hansen, Hurwitz, Madow, Sample Survey Methods and Theory, Volume II – Theory, 1953, John Wiley and Sons, New York