Model based Analysis of the Variance Estimators for the Combined Ratio Estimator

Charles Wafula

Mathematics Department, Kenyatta University, P.O. Box 43844, Nairobi, KENYA.

ABSTRACT

In this paper we study the variance estimators for the combined ratio estimator under an appropriate asymptotic framework. An alternative bias-robust variance estimator, different from that suggested by Valliant (1987), is derived. Several variance estimators are compared in an empirical study using a real population.

1.0 INTRODUCTION

The problem of variance estimation is an important one in sample surveys. It is made much more important by the fact that the two main competing approaches to sample survey theory suggest different variance estimators for a given population mean estimator. This has aroused a lot of interest among researchers to find out which method of variance estimation is appropriate for a given population mean estimator.

One such study is that by Valliant (1987). Valliant studied, among other things, the prediction properties of the variance estimators for the combined ratio estimator. However, we note that the asymptotic framework used by Valliant is not appropriate for the combined ratio estimator. He considered the case when the sample size within each stratum tends to infinity and the number of strata is fixed. This is a situation where the separate ratio estimator, rather than the combined ratio estimator, is used in practice. We also note that some of the variance estimators were not considered in Valliant's study. For example, the BRR (Balance Repeated Replication) variance estimation technique was not considered.

In this paper we examine the prediction properties of the variance estimators for the combined ratio estimator. We consider the case when the number of strata tends to infinity while the sample size in each stratum remains bounded. We also derive an alternative bias-robust variance estimator for the combined ratio estimator.

2.0 THE COMBINED RATIO ESTIMATOR AND ITS ERROR - VARIANCE

The population consists of H strata with N_h elements in the h - th stratum from which a simple random sample of size n_h is taken without replacement. The total sample size $n = \sum n_h$ and population size $N = \sum N_h$. Throughout this paper the summation sign is from I to H. Associated with the i - th unit of the h - th stratum are two values y_{hi} and x_{hi}. Y is the variable under investigation while x is a variable assumed known for each unit of the population. For the h - th stratum let $W_h = N_h/N$ be the stratum weight, $f_h = n_h/N$ the sampling fraction, $\bar{y}_h, \bar{x}_h, \bar{Y}_h, \bar{X}_h$ the y and x sample and population means respectively. One common estimator of $\bar{Y} = \sum \bar{W}_h \bar{Y}_h$ is the combined ratio estimator, $\bar{y}_{CR}$, given by $\bar{y}_{CR} = \dfrac{\bar{x}\,\bar{y}_{st}}{\bar{x}_{st}}$

where $\bar{y}_{st} = \sum W_h \bar{y}_h$, $\bar{x}_{st} = \sum W_h \bar{x}_h$ and $\bar{x} = \sum W_h \bar{x}_h$.

A framework (Krewski and Rao, 1981) for asymptotic calculations in stratified random sampling is to let H tend to infinity and n_h remain bounded. To prevent any stratum from playing a dominant role, it is often further assumed that $W_h n/n_h$ is bounded uniformly in h (Wu, 1985). The followings assumptions will be made in all the asymptotic calculations in the rest of this paper.

(i.) $\underset{h}{\text{Max}} \left(n_h\right) = 0(1)$ .....(1)

(ii.) $\underset{h}{\text{Max}} \left(W_h\right) = 0\left(n^{-1}\right)$

(iii.) As N_h, N and n grow, $f_n \to 0$ and both the sample and population remain stable in the sense that the sample and population averages $\bar{x}$, $\bar{x}_{st}$, $\bar{x}_h$, $\bar{x}_h^{(t)}$, $\bar{X}_h^{(t)}$ all converge to non zero constants.

The combined ratio estimator is often studied under the following model (Wu, 1985):

$$EY_{hi} = \alpha_h + \beta x_{hi}$$

$$\text{Cov}\left(Y_{hi}, Y_{hj}\right) = \begin{cases} \sigma^2 x_{hi} & i = j \\ 0 & i \neq j \end{cases} \qquad \text{.....(2)}$$

$$D = \sum W_h \alpha_h = 0$$

The side constraint D = 0 is unnatural but indispensable. Under model (2.2), the error-variance of $\bar{y}_{CR}$ is

$$\mathrm{Var}\left(\bar{y}_{st}-\bar{Y}\right)=\left(\bar{X}/\bar{x}_{st}\right)^2\left\{\sum W_h^2\frac{N_h-2n_h}{N_h n_h}\bar{x}_h^{(t)}+\sum W_h^2\bar{X}_h^{(t)}/N_h\right.$$

$$-\frac{2}{x}\sum\frac{W_h^3}{N_h}\left(\bar{x}_h-\bar{x}_h\right)\left(\bar{x}_h^{(t)}-\bar{X}_h^{(t)}\right)$$

$$\left.+\frac{1}{\bar{x}^2}\sum W_h^2\left(\bar{x}_h-\bar{x}_h\right)^2\sum\frac{W_h^2\bar{X}_h^{(t)}}{N_h}\right\}\sigma^2 \qquad \ldots(3)$$

where $\bar{x}_h^{(t)}=n_h^{-1}\sum_1^{n_h}x_{hi}^t$ and $\bar{X}_h^{(t)}=N_h^{-1}\sum_1^{N_h}x_{hi}^t$ Under conditions (2.1), this error-variance reduces to

$$\mathrm{Var}\left(\bar{y}_{CR}-\bar{Y}\right)=\left(\bar{X}/x_{st}\right)^2\sum\frac{W_h^2}{n_h}\bar{x}_h^{(t)}\sigma^2+0\left(1/nN\right) \qquad \ldots\ldots\ldots\ldots(4)$$

3.0 Biases of the Variance Estimators

We now investigate the biases of the variance estimators in estimating (3)

3.1 The Conventional Estimator

A variance estimator that is conventionally associated with the combined ratio estimator is given by

$$V_{ost}=\sum W_h^2\frac{1-f_h}{n_h}S_{eh}^2$$

here $S_{eh}^2=\left(n_h-1\right)^{-1}\sum_1^{n_h}\left(y_{hi}-x_{hi}\bar{y}_y\frac{\bar{y}_h}{\bar{x}_h}\right)$

Under model (2) and using assumptions in (1) the model bias of V_{ost} is obtained as

$$\mathrm{Bias}\left(V_{ost}\right)=\left[1-\left(\bar{x}/\bar{x}_{st}\right)^2\right]\sigma^2\sum\frac{W_h^2}{n_h}\bar{x}_h^{(t)}+0\left(1/nN\right)$$

Hence the relative bias of V_{ost} is $\left(\bar{x}_{st}/\bar{x}\right)^2-1$ which is an increasing function of $\bar{x}_{st}$ and vanishes only in samples balanced on x, i.e. when $\bar{x}_{st}=\bar{X}$. The condition $\bar{x}_{st}=\bar{X}$ is satisfied if the sample from each stratum is balanced on x.

3.2 THE LINEARIZATION ESTIMATOR

This estimator was originally suggested by Wu(1985) and is given by $V_{2st} = \left(\bar{x}/\bar{x}_{st}\right)^2 V_{ost}$. Under model (2) and to order $0\left(1/nN\right)$, this estimator can easily be shown to be unbiased.

3.3 THE JACKKNIFE ESTIMATOR

A commonly used version of the jackknife estimate of variance for some estimator $\hat{\theta}$ in stratified sampling is given by $V_j\left(\hat{\theta}\right) = \sum \frac{n_h - 1}{n_h}\left(1 - f_h\right)\sum\left(\hat{\theta}^{hi} - \hat{\theta}\right)^2$ where $\hat{\theta}^{hi}$ has the same form as $\hat{\theta}$ but omits the h - th sample observation. Under (2.2), using assumptions (1) and with the extra assumption that $\left|W_h\left(x_{hi} - \bar{x}\right)/\left(n_h - 1\right)\bar{x}_{st}\right| < 1$ for all h and i, it can easily be shown that to order 0(1/nN), $V_j\left(\hat{\theta}\right)$ for the combined ratio estimator is unbiased.

3.4 BRR VARIANCE ESTIMATOR

Assume we have a sample obtained by selecting two units from each stratum. Let $\hat{\theta}$ be the estimate of the population parameter $\hat{\theta}$ based on the entire sample S. Then the BRR method of estimating the variance of $\hat{\theta}$ is as follows. From each stratum one out of the two sample units is selected to form an r - th sample denoted by H_r.

This process is repeated R times to form R replicates. Define $d_h^r = 1$ or $d_h^r = -1$ if the first or the second sample unit, respectively, in the h - th stratum is in the r - th half sample. Also let

$$\bar{y}_h^{(r)} = \begin{cases} y_{h_1} & \text{if } d_h^r = 1 \\ y_{h_2} & \text{if } d_h^r = -1. \end{cases}$$

Then $\bar{y}_h^{(r)} = \bar{y}_h + \Delta y_h$ where $\Delta y_h = \frac{1}{2}\left(y_{h_1} - y_{h_2}\right)$ Now let $\bar{y}^{(r)}$ be the estimate of $\bar{Y}$ based on the r - th half sample where

$$\bar{y}^{(r)} = \sum W_h \bar{y}_h^{(r)}$$

$$= \bar{y}_{st} + W_h d_h^r \Delta y_h$$

Hence the estimate of $\theta = g(\bar{Y})$ based on the r - th half sample is $\hat{\theta}^{(r)} = g(\bar{y}^{(r)})$ and one version of the BRR estimator of the variance of $\hat{\theta}$ is given by

$$v_B(\hat{\theta}) = \frac{1}{R}\sum^{R} (\hat{\theta}^{(r)} - \hat{\theta})^2$$

In the case of the combined ratio estimator,

$$\hat{\theta}^{(r)} - \hat{\theta} = \frac{\bar{x}}{\bar{x}_{st}}\left\{\frac{\bar{y}_{st} + \sum W_h d_h^r \Delta y_h}{1 + \Delta_{hx}^r} - \bar{y}_{st}\right\}$$

where $\Delta_{hx}^r = \dfrac{\sum W_h d_h^r \Delta x_h}{\bar{x}_{st}}$, $\Delta x_h = \frac{1}{2}(x_{h_1} - x_{h_2})$.

Assuming all the Δ_{hx}^r are less than one in absolute value and using the orthogonality of the d_h^r's i.e. $\sum_{\substack{r=1 \\ h \neq h_1}}^{R} d_h^r d_{h_1}^r = 0$, it can be shown that under model (2)

$$Ev_B(\hat{\theta}) = (\bar{x}/x_{st})^2 \sigma^2 \sum W_h / 2\bar{x}_h^{(r)} + \beta^2 (\bar{x}/\bar{x}_{st})^2 \sum W_h^2 (\Delta x_h)^2 + 0(1/nN) \quad \ldots\ldots(5)$$

Comparing (5) and (4) (with $n_h = 2$) shows that to order 0(1/nN), $v_B(\hat{\theta})$ has bias

$$\beta^2 (\bar{x}/\bar{x}_{st})^2 \sum W_h^2 (x_h)^2 \sum W_h^2 (\Delta x_h)^2$$

Thus $V_B(\ )$ is positively biased and the bias is an increasing function of the model parameter β The bias does not vanish in balanced samples.

3.5 Bias - Robust Variance Estimator

In this section we obtain a bias-robust variance estimator using the procedure of Royall and Cumberland (1978). To do this we assume a special case of model (2) i.e.

$$Ey_{hi} = \beta x_{hi}$$

$$Cov(Y_{hi}, Y_{hj}) = \begin{cases} \sigma^2 x_{hi} & i = j \\ 0 & i \neq j \end{cases} \quad \ldots\ldots\ldots(6)$$

Valliant (1987) has also obtained a bias-robust variance estimator using (6) as a working model. However the estimator he obtained is different from the one we obtain here. The difference arises because of the difference in the way residuals under

6 are defined. Valliant uses the BLU estimator of β under 6 to define the residual. On our part we have rewritten the combined ratio estimator in the predictive for from which we have obtained the implied estimator of β and then used this implied estimator to define the residual.

Under 6 the estimate of the population mean is of the form. $\hat{T} = N^{-1}\left\{\sum n_h \bar{y}_h + \hat{\beta}\sum\sum_r x_{hi}\right\}$ where summation over r indicates summation over non-sample units. Rewriting $\bar{y}_{cr}$ in the predictive form

$$\bar{y}_{cr} = N^{-1}\left\{\sum n_h \bar{y}_h + \frac{1}{T_{sx}}\left(T_x \bar{y}_{st}/\bar{x}_{st} - \sum n_h \bar{y}\right)T_x\right\}$$

$\bar{y}_{CR}$ where $T_x = \sum\sum_r x_{hi}$ and $T_{sx} = \sum N_h x_h$, we conclude that under 6, $\bar{y}_{CR}$ uses

$\hat{\beta} = \frac{1}{T_{sx}}\left(T_x \bar{y}_{st}/\bar{x}_{st} - \sum n_h \bar{y}_h\right)$ as an estimate of β. Hence the residual is given by

$r_{hi} = y_{hi} - \hat{\beta}x_{hi}$

Note that $Er_{hi} = 0$. Under a more general model in which the variance of the hi - th unit is v_{hi}, the expected value of the squared residual is $Er_{hi}^2 = V_{hi}\left(1 - d_{hi}\right)$

where $d_{hi} = \frac{2x_{hi}}{T_x}\left(\bar{X}/f_h \bar{x}_{st} - 1\right) - \frac{x_{hi}}{T_x^2 v_{hi}}\left\{\left[\frac{N_h \bar{x}}{\bar{x}_{st}} - n_h\right]^2 \bar{v}_h/n_h\right\}$ and $\bar{v}_h = n_h^{-1}\sum_1^{n_h} v_{hi}$.

When $v_{hi} = \sigma^2 x_{hi}$, d_{hi} is simply

$$K'_{hi} = \frac{2x_{hi}}{T_x}\left(\bar{X}/f_h \bar{x}_{st} - 1\right) - \frac{x_{hi}}{T_x^2 v_{hi}}\left\{\left[\frac{N_h \bar{x}}{\bar{x}_{st}} - n_h\right]^2 n_h \bar{x}_h\right\}$$

Therefore under 6 an estimator of σ^2 is given by $\hat{\sigma}^2 = \sum\sum_1^{n_h} \frac{r_{hi}^2}{\left(1 - K'_{hi}\right)/\sum n_h \bar{x}_h}$

Substituting this value in 3 (with t =1) we obtain a bias-robust variance estimator, v_{R_o}, of the combined ratio estimator. Another estimator can be obtained by using 4.

4.0 EMPIRICAL STUDY

We tested the theory of the preceding section using a population of 64 cities in the United States. The variable y is the population size of each city in 1930 and the auxiliary x is the corresponding population size of each city in 1920. this data set is

given in table 5.1 of Cochran (1977). Based on the scatter plot of y verses x (not given here), a combined ratio estimator seems a reasonable estimator to use.

The population was divided into two strata by sorting units in ascending order on x and then splitting it into two equal parts; the first stratum consisting of the first 32 units with the smallest values of x and the second consisting of the 32 units with the largest values of z. Simple random samples of equal size were selected without replacement from each stratum. This sample selection procedure was repeated 2000 times for two sample sizes - 2 units per stratum for a sample size of 4 and 16 units per stratum for a total sample size of 32.

The following variance estimators were included in the study v_{ost}, v_{2st}, v_{R_o}, v_j and v_B. The estimator v_B was not included in the sets of samples with total samples with total sample size 32 because it would have required grouping of their units before applying it. For each sample, the combined ratio estimator and the five variance estimators were computed.

Table 1. Summary statistics for standardized errors and estimators of variance from 2000 stratified simple random samples

Estimator	Sample Size	$\left(\frac{\text{Aug. VAr Est.}}{\text{MSE}}\right)^2$	SZE≤-2.201	SZE≥2.201	|SZE|≤2.201
v_{ost}	4	0.95	26.8	0.0	73.2
	32	0.98	23.4	0 0	76.6
v_{2st}	4	0.83	26.9	0.0	73.1
	32	0.97	23.1	0.0	76.9
v_j	4	0.89	25.5	0.0	74.5
	32	1.00	22.6	0.0	77.4
v_{R_o}	4	0.89	23.6	0.0	76.4
	32	0.98	22.7	0.0	77.3
v_B	4	0.93	24.3	0.0	75 7
	32	–	–	–	–

Relative error of $\bar{y}_{CR}$ was 0.0015 for sample size 4 and 0.0006 for sample size 32.

As guaranteed by probability sampling theory $\bar{y}_{CR}$ is approximately unbiased over all samples for all the two sample sizes. At sample size 32 all the variance estimators are nearly unbiased with v_j being somewhat conservative. All the variance

estimators are under estimates as sample size 4. Table 1 also contains the 95% confidence coverage results over the 2000 samples. The standardized error (SZE) defined as $(\bar{y}_{CR} - \bar{Y})/v^{1/2}$ was computed for each sample and each variance estimator v. The percentages of samples with SZE $\leq$-2.201, SZE>2.201 and |SZE| <2.201 were computed. All the variance estimators gave relatively poor coverage rates at all sample sizes. Interestingly, all the excess SZEs are negative. A possible reason for this is given later.

We also performed conditional analysis of the variance estimators. Samples were sorted in ascending order of $\bar{x}_{st}$ and divided into 10 groups of 200 samples each. In each group the averages of the error $\bar{y}_{CR} - \bar{Y}$, the biases of the variance estimators and the percentage of samples with |SZE|<2.201 were computed. Results for the conditional coverage rates are given in table 3 and those for the conditional biases in Table 2. Note that the results in tables 2 and 3 are given in ascending order of $\bar{x}_{st}$. The theory in the preceding sections showed that under 2.2 the bias of v_{ost} is an increasing function of $\bar{x}_{st}$. This result is well illustrated by the conditional biases given in Table 2. For sample size 32, other estimators also tend to underestimate when $\bar{x}_{st}$ is small and overestimate when $\bar{x}_{st}$ is large.

In general all the variance estimators gave better coverage probability rates in the lower tails of the $\bar{x}_{st}$ distribution than in the upper tails. This seems to suggest that the biases of the variance estimators are not the major determinants of the poor coverage rates of the associated confidence intervals. If they were, it would have been expected that the coverage rates are better in the upper tail of the $\bar{x}_{st}$ distribution than in the lower tail.

The major determinant of the poor performance seems to be the large positive correlations between the numerators and denominators of the standardized errors. The correlations associated with the five variance estimators were computed for each of the groups. The correlations range from 0.39 to 0.97 in samples of size 4 and from 0.92 to 0.95 in samples of size 32. The correlations also tend to be small in the lower tails than in the upper tails of the $\bar{x}_{st}$ distribution.

As noted earlier, all the excess SZEs are negative. The reason for this lies in the large positive correlations between the numerators and denominators of SZEs.

Thus when $\bar{y}_{CR}$ is small, giving negative error, the variance estimators also tend to be small producing large negative SZEs. On the other hand, when $\bar{y}_{CR}$ is large, giving positive error, the variance estimators tend to be large preventing a large positive SZE.

Table 2. Conditional Biases of the Combined Ratio Estimator and its Variance Estimators

Sample Size 4

$\bar{y}_{CR}$	v_{ost}	v_{2st}	v_j	v_{R_o}	v_P
1.65	-297.0	33.2	37.1	51.8	74.3
-5.67	-187.6	-47.6	-42.6	-25.6	-17.9
-5.00	-67.2	44.7	55.8	49.8	87.5
-3.05	-91.9	-15.1	8.8	59.5	55.4
-5.77	-113.6	-78.1	-59.5	21.0	-25.2
13.10	-1443.0	-1441.5	-982.4	-1058.9	-855.5
23.98	-1574.0	-1916.0	-1338.3	-1606.5	-1126.3
1.60	131.4	-496.0	-280.3	-472.8	-115.2
-10.96	145.0	-296.4	-218.7	-286.0	-125.6
-5.89	-773.8	258.5	308.9	-267.5	422.3

Sample Size 32

$\bar{y}_{CR}$	v_{ost}	v_{2st}	v_j	v_{R_o}
1.87	-39.3	-21.5	-16.2	-10.3
-0.01	-53.9	-42.7	-37.8	-35.6
1.59	-15.2	-7.3	-2.1	-1.5
-0.48	-16.5	-11.8	-7.1	-8.2
0.3	-16.6	-10.0	-5.1	-7.9
0.94	4.3	2.8	7.9	3.2
-0.53	5.1	-9.8	-5.3	-11.3
-0.79	5.9	-2.8	1.4	-6.3
-0.37	33.3	19.2	23.3	13.1
00.69	56.1	37.5	36.2	21.8

Table 3. Conditional coverage rates

Sample Size 4

v_{ost}	v_{2st}	v_j	v_{R_o}	v_B
75.5	82.0	82.0	87.5	83.0
71.5	78.5	78.5	82.0	78.5
75.5	79.5	79.5	85.0	80.5
79.0	83.0	83.0	86.5	83.5
67.5	71.0	71.5	72.5	72.5
72.5	72.5	73.5	78.0	75.0
76.5	74.5	76.0	76.5	76.5
80.0	74.5	80.0	78.0	82.5
66.0	56.5	62.0	58.0	65.0
68.0	59.0	59.5	60.0	60.0

Sample Size 32

v_{ost}	v_{2st}	v_j	v_{R_o}
86.5	89.0	89.0	90.5
72.0	75.5	76.5	78.0
84.0	86.5	86.5	86.5
75.5	76.5	77.5	77.5
74.5	74.5	74.5	75.0
76.0	76.0	76.5	75.5
71.0	70.5	70.5	70.0
73.0	71.0	72.0	71.5
76.5	75.5	76.0	74.5
77.0	74.0	75.0	74.0

5.0 CONCLUSIONS

In this paper we have shown that the commonly used variance estimators v_{ost} and v_B for the combined ratio estimator are biased. The bias of v_{ost} is an increasing function of $\bar{x}_{st}$ and vanishes in samples balanced on x. On the other hand, the bias of v_B is an increasing function of the model parameter β. Neglecting terms of

order $0(1/nN)$ v_{2st} and v_j are unbiased. The unbiasedness of v_{2st} and v_j was also reported in Valliant (1987).

REFERENCES

Cochran W. G. (1977) *Sampling Techniques*, **156**. Wiley, New York.

Krewski D. and Rao J.N.K. (1981) *Inference for stratified samples: Properties of linearization, Jackknife and Balanced Repeated Replication Methods*. Annals of Statistics, **9**, 1010 - 1019.

Royall, R. M. and Cumberland W.G. (1978) *Variance Estimation in Finite Population Sampling*. Journal of American Statistical Association, **73**, 351 - 358.

Valliant, R. (1987) *Conditional properties of some estimators in stratified sampling*. Journal of American Statistical Association, **73**, 351 - 358.

Wu, C.F.J. (1985) *Variance Estimation for the combined ratio and the combined regression estimators*. Journal Royal Statistics Society . **B**, 147 - 154.