SMALL SAMPLE PROPERTIES OF THE LINEARIZATION, JACKKNIFE AND BALANCED HALF-SAMPLE METHODS FOR RATIO ESTIMATION IN STRATIFIED SAMPLES

D. Krewski and J.N.K. Rao

Health & Welfare Canada and Carleton University

Exact small sample properties of the linearization, jackknife and balanced half-sample methods applied to ratio estimation in stratified samples are investigated under a suitable linear regression model. In particular, the bias of the classical combined ratio estimator is compared with that of two different jackknife estimators that have been proposed. In addition, the biases of the linearization variance estimator and several alternative variance estimators based on the jackknife and balanced half-sample methods are evaluated. The stability of the linearization variance estimator is also compared with that of a particular balanced half-sample variance estimator. The analytical results reported here are compared with empirical results reported previously by other investigators.

1. INTRODUCTION. Suppose (y_{hi}, x_{hi}) $(i=1,\ldots,n_h)$ denotes a simple random sample of size n_h from the N_h units in stratum $h=1,\cdots,L$. Within stratum h, let $\bar{y}_h$ and $\bar{x}_h$ denote the stratum means for variables y and x respectively and let $\bar{Y}_h$ and $\bar{X}_h$ denote the population means. The overall population means are then $\bar{Y}=\Sigma W_h\bar{Y}_h$ and $\bar{X}=\Sigma W_h\bar{X}_h$ where $W_h=N_h/N$ and $N=\Sigma N_h$.

The classical combined ratio estimator of $R=\bar{Y}/\bar{X}$ is $r_1=\bar{y}_{st}/\bar{x}_{st}$ where $\bar{y}_{st}=\Sigma W_h\bar{y}_h$ and $\bar{x}_{st}=\Sigma W_h\bar{x}_h$. For the important special case $n_h=2$ $(h\ 1,\ldots,L)$, Jones' jackknife estimator r_2 (Jones, 1974) and McCarthy's jackknife estimator r_3 (McCarthy, 1966) are given by

$$r_2=(L+1)r_1 - \Sigma r_1^h \text{ and} \tag{1.1}$$

$$r_3=2r_1-(\Sigma r_1^h/L) \tag{1.2}$$

respectively, where $r_1^h=(r_1^{h1}+r_1^{h2})/2$ and r_1^{hi} denotes the combined ratio estimator omitting (y_{hi},x_{hi}).

The linearization variance estimator of the mean square error of r_1 is

$$v_1=(\bar{x}_{st})^{-2}\ \{v(\bar{y}_{st})+r_1^2\ v(\bar{x}_{st})-2r\text{cov}(\bar{y}_{st},\bar{x}_{st})\} \tag{1.3}$$

where, for example, $v(\bar{y}_{st})=\sum_{h=1}^{L} W_h^2 \sum_{i=1}^{n_h} (y_{hi}-\bar{y}_h)^2/ n_h(n_h-1)$. When (h=1,...,L), a set of k balanced half samples may be formed (McCarthy, 1966) with the estimators $r_{1(i)}$ and $r^c_{1(i)}$ associated with the ith half-sample and its complement $(i=1,\ldots,k)$. Three alternative balanced half-sample variance estimators are then given by

$$v_2=\Sigma(r_{1(i)}-r_1)^2/k, \tag{1.4}$$

$$v_3=\Sigma(r_{1(i)}-r_{1(.)})^2/k \tag{1.5}$$

$$v_4=\Sigma(r_{1(i)}-r^c_{1(i)})^2/(4k) \tag{1.6}$$

where $r_{1(.)}=\Sigma r_{(i)}/k$. When all $n_h=2$, two jackknife variance estimators proposed by Jones (1974) and Kish & Frankel (1974) are given by

$$v_5=\Sigma\Sigma(r_1^{hi}-r_1^h)^2/2 \text{ and} \tag{1.7}$$

$$v_6=\Sigma\Sigma(r_1^{hi}-r_1)^2/2. \tag{1.8}$$

Exact small sample properties of these estimators may be obtained in the case of proportional allocation $(n_h=W_hn)$ under the linear regression model (Rao & Ramachandran, 1974)

$$y_{hi}=\alpha_h+\beta_hx_{hi}+e_{hi}, \tag{1.9}$$
$$E(e_{hi}|x_{hi})=0,\ E(e_{hi}^2|x_{hi})=\delta_hx_h^{t_h},$$
$$E(e_{hi}e_{hj}|x_{hi}x_{hj})=0\quad (i\neq j)$$
$$E(e_{hi}e_{h'j}|x_{hi}x_{h'j})=0\quad (h\neq h')$$

$(i,j=1,\ldots,n_h;h,h'=1,\ldots,L)$ where x_{hi} has a gamma distribution with mean a_h. For a variety of natural and synthetic populations Rao & Kuzik (1974) found the coefficient of variation of the auxiliary variable x_{hi} to lie between 0.4 and 1.0. Since the coefficient of variation of x_{hi} is $a_h^{-\frac{1}{2}}$, small value of a_h will be of interest in what follows. In practice, t_h has often been found to lie between 0 and 2 and is assumed here to lie in this range. For simplicity, it will be assumed that the strata sizes $\{N_h\}$ are effectively infinite.

Derivations of the analytical results presented here are similar to those of Rao & Webster (1966) and Rao (1974) and are omitted. Further details of the results summarized here may be found in Krewski (1977).

2. Bias of ratio estimators. Under the model (1.9) with proportional allocation, the bias of the combined ratio estimator r_1 may be expressed as

$$\begin{aligned} B(r_1)&=E(r_1)-R\\ &=\frac{n\bar{\alpha}}{m(m-1)}\\ &=D_1\bar{\alpha}, \end{aligned} \tag{2.1}$$

provided m>1, where $m=\Sigma m_h$, $m_h=n_ha_h$ and $\bar{\alpha}=\Sigma n_h\alpha_h/n$.

For the special case $n_h=2$ (h=1,...,L), the biases of the jackknife ratio estimators r_2 and r_3 may also be evaluated under (1.9) with proportional allocation. Assuming $a_h=a$ (h=1,...,L) so that $B(r_2)$ and $B(r_3)$ do not depend on the β_h, the biases of these two estimators may be expressed as

$$\begin{aligned} B(r_2)&=n\{\frac{1}{m(m-1)}+\frac{L}{(m-1)}-LI(b,a;2)\}\bar{\alpha}\\ &=D_2\bar{\alpha}\ \text{ and} \end{aligned} \tag{2.2}$$

$$\begin{aligned} B(r_3)&=\{\frac{4L}{(m-1)}-\frac{n}{m}-2LI(b,a;2)\}\bar{\alpha}\\ &=D_3\bar{\alpha}, \end{aligned} \tag{2.3}$$

provided $m>1$. Here $b=2a(L-1)$ and $I(a_1,a_2;\lambda) = E(X_1+\lambda X_2)^{-1}$ where $\lambda>0$ ($\lambda\neq1$) and X_1 and X_2 are independent gamma variates with means a_1 and a_2 respectively. This expectation is evaluated explicitly in the following Theorem.

Theorem. Let X_1 and X_2 be independent gamma variates with means a and b respectively. Then for any positive constant $\lambda\neq1$ and integral values of a and b

$$I(a,b;\lambda)=E(X_1+\lambda X_2)^{-1} = I^{(1)}+I^{(2)}+I^{(3)}$$

where

$$I^{(1)}=\sum_{k=1}^{a-1}(-1)^{k+1}\lambda^{k-1}\frac{\Gamma(b+k-1)}{\Gamma(b)}\frac{\Gamma(a-k)}{\Gamma(a)},$$

$$I^{(2)}=(-1)^{a+1}\lambda^{a-1}\frac{\Gamma(b+a-1)}{\Gamma(b)\Gamma(a)}\sum_{k=1}^{b+a-2}\frac{(-1)^{k+1}}{(\lambda-1)^k(b+a-k-1)} \text{ and}$$

$$I^{(3)}=(-1)^{b+1}\lambda^{a-1}\frac{\Gamma(b+a-1)}{\Gamma(b)\Gamma(a)}\frac{\ell n\lambda}{(\lambda-1)^{b+a-1}}.$$

Since the expressions for the coefficients D_2 and D_3 are not in closed form, the biases of the three alternative ratio estimators were compared by evaluating these coefficients for selected values of a and L (Table 1). While both jackknife estimators have smaller absolute bias than the classical estimator, r_2 appears particularly effective as a means of bias reduction in this case while the bias of r_3 approaches that of r_1 as L increases.

Where $\beta_h=\beta$ $(h=1,\ldots,L)$, the biases of both jackknife estimators do not depend on β, regardless of the value of the a_h. When both the a_h and β_h are not constant, however, the biases both r_2 and r_3 will in general involve the β_h. In this case $B(r_2) = LB(r_3)$ for $\bar{\alpha}=0$.

Thus, while r_2 may be preferable to r_3 with respect to bias when the a_h are constant (Table 1), r_3 may be preferable to r_2 when $\bar{\alpha}=0$ and the a_h and β_h are not constant. Further, $B(r_1)=0$ when $\bar{\alpha}=0$ while both r_2 and r_3 will in general be biased in the case of unequal a_h and β_h.

3. Bias of variance estimators. The mean square error of the combined ratio estimator under the model (1.9) with proportional allocation is given by

$$MSE(r_1)=\frac{n^2(m+2)\;\bar{\alpha}^{-2}}{m^2(m-1)(m-2)}+\Sigma\frac{m_h(\beta_h-\bar{\beta})^2}{m(m+1)} + \Sigma\frac{n_h\,f_h(t_h)\delta_h}{(m+t_h-1)(m+t_h-2)} \qquad (3.1)$$

provided $m>2$, where $\bar{\beta}=\Sigma m_h\beta_h/m$ and $f_h(t)=\Gamma(a_h+t)/\Gamma(a_h)$.

After considerable algebra the bias of the linearization variance estimator under (1.9) with proportional allocation may be expressed as

$$B(v_1)= E(v_1)-MSE(r_1) = \frac{n^2(3m+2)\bar{\alpha}^{-2}}{m^2(m^2-1)(m-2)}-\frac{3\Sigma m_h(\beta_h-\bar{\beta})^2}{m(m+1)(m+3)} -\Sigma\frac{n_h\;f_h(t_h)(t_h^2+t_h(2m+1)-m)\;\delta_h}{(m+t_h+1)(m+t_h)\;\;(m+t_h-1)(m+t_h-2)} \qquad (3.2)$$

provided $m>2$. Similarly, after some tedious algebra, the biases of the balanced half-sample variance estimators v_2 and v_4 may be expressed as

$$B(v_2)=\frac{2n^2(m+2)(3m-4)\bar{\alpha}^{-2}}{m^2(m-1)(m-2)^2(m-4)} + 4\;\Sigma\frac{f_h(t_h)\delta_h}{(m+2t_h-2)}\left\{\frac{1}{(m+2t_h-4)}-\frac{1}{(m+t_h-2)}\right\}$$

and

$$B(v_4)=\frac{n^2(3m^2+4m-16)\bar{\alpha}^{-2}}{m^2(m-1)(m-2)^2(m-4)}-\frac{\Sigma\;m_h(\beta_h-\bar{\beta})^2}{m(m+1)(m+2)} + 2\;\Sigma\;f_h(t_h)\delta_h\left\{\frac{1}{(m+2t_h-2)(m+2t_h-4)} - \frac{1}{(m+t_h-1)(m+t_h-2)}\right\} \qquad (3.3),\ (3.4)$$

provided $m>4$.

From (3.2), it is easily seen that $B(v_1)>0$ when $t_h\geq1/2$ for all h. From (3.3), $B(v_2)\geq0$ when $t_h\leq2$ for all h. From (3.4), $B(v_4)\geq0$ when $t_h\leq3/2$ and $\beta_h=\beta$ for all h, provided $m\geq5$.

Comparing (3.2) - (3.4) when $n_h=2$ and $\beta_h=\beta$ for all h shows $B(v_2)>B(v_4)>0$ when all $t_h\leq3/2$ and $B(v_2)>B(v_4)>|B(v_1)|$ when all $t_h=1$ (provided $m\geq5$). For $\bar{\alpha}=0$ and n_h 2 for all h, $|B(v_1)|>|B(v_4)|>B(v_2)=0$ when all $t_h=2$.

In the special case $n_h=2$, $a_h=a$, $\beta_h=\beta$ and $t_h=t$ for all h, the biases of the jackknife variance estimators v_5 and v_6 may be expressed as $B(v_i)=F_i\bar{\alpha}^{-2}+H_i\bar{\delta}$, $(i=5,6)$, where $\bar{\delta}=\Sigma\delta_h/L$. As in (2.2) and (2.3), however, the coefficients F_i and H_i $(i=5,6)$ are not in closed form. If a set of k balanced replicates is constructed with the properties that (i) the number of observations common to each pair of half-samples is constant and (ii) each observation is included in precisely half of the half-samples, then the bias of v_3 may also be expressed as $B(v_3)=F_3\bar{\alpha}^{-2}+H_3\bar{\delta}$, although the expressions for F_3 and H_3 are again not in closed form.

From (3.2) - (3.4), the biases of v_1, v_2 and v_4 may also be expressed as $B(v_i)=F_i\bar{\alpha}^{-2}+H_i\bar{\delta}$ $(i=1,2,4)$ in the case $n_h=2$, $a_h=a$, $\beta_h=\beta$ and $t_h=t$ for all h. Since the expressions for $B(v_i)$ $(i=3,5,6)$ are not in closed form, the biases of the six alternative variance estimators considered here were compared for selected values of a and L and for t=0, 1 and 2.

This analysis indicated that $B(v_1)=0$ when $t=1$ or 2 as indicated earlier. Both jackknife variance estimators v_5 and v_6 also underestimate $MSE(r_1)$ when $t=1$ or 2 and $L>4$ with $|B(v_1)| > |B(v_5)| > |B(v_6)|$ in this case. For $t\leq 3/2$, it was shown previously that v_2 and v_4 are both overestimates. The present analysis showed that v_3 is also an overestimate when $t=0$ or 1 with $B(v_2)>B(v_3)>B(v_4)$ in this case. When $t=1$, $B(v_2)>B(v_3)>B(v_4)>|B(v_1)| >|B(v_5)| >|B(v_6)|$ provided $L>4$.

When in addition $\bar{\alpha}=0$, it was found that all six variance estimators overestimate $MSE(r_1)$ for $t=0$ with $B(v_2)>B(v_3)>B(v_4)>B(v_6)>B(v_5)>B(v_1)$. For $t=2$, all six variance estimators are underestimators when $\bar{\alpha}=0$ with the absolute biases following the reverse order to that for $t=0$.

4. Stability of variance estimators. The mean square error of the linearization variance estimator v_1 and the balanced half-sample variance estimator v_2 may be derived under (1.9) with normally distributed errors and proportional allocation in the special case $n_h=2$, $a_h=a$, $\beta_h=\beta$, $t_h=t$ and $\delta_h=\delta$ for all h. After considerable algebra, the mean square errors of these two variance estimators may be expressed as

$$MSE(v_i) = J_i\bar{\alpha}^4+K_i\bar{\alpha}^2\delta+L_i\delta^2 \qquad (4.1)$$

($i=1,2$), provided that the set of k balanced half-samples is selected so that the number of observations common to each pair is constant.

Since the expressions for J_i, K_i and L_i ($i=1,2$) in (4.1) are not in closed form, these coefficients were evaluated for selected values of a and L and for $t=0$, 1 and 2. (The results in the case of L_i, for example, are shown in Table 2.)

Each of the coefficients J_i, K_i and L_i ($i=1,2$) was found to decrease as L or a increase so that the mean square errors of both variance estimators decrease as the number of strata increase or as the coefficient of variation of the x population decreases. Since $J_1<J_2$, $K_1<K_2$ and $L_1<L_2$ for all values of a, L and t, v_1 is more stable than v_2. The ratios J_2/J_1, K_2/K_1 and L_2/L_1 all decrease, however, as L or a increase. The ratio L_2/L_1 is particularly close to one for moderate values of La, indicating that the stability of v_2 is comparable to that of v_1 when $\bar{\alpha}=0$. Since L_2/L_1 decreases as t increases, $MSE(v_2)/(MSE(v_1)$ decreases as t increases when $\bar{\alpha}=0$.

5. Discussion. In contrast to the case of simple random sampling (Rao & Webster, 1966), results concerning the jackknife method as a means of bias reduction in stratified samples are not clearcut. When the distribution of the x population is the same in all strata, r_2 is particularly effective as a means of bias reduction while the bias of r_3 is comparable to that of the combined ratio estimator r_1 for moderate values of L. When the distribution of the x population is not the same in all strata, however, r_3 can have smaller absolute bias than r_2. Moreover, both jackknife estimators may be biased in situations where the classical estimator is unbiased.

When $t_h\geq 1/2$ in each stratum h, the linearization variance estimator v_1 *underestimates* the mean square error of the combined ratio estimator. When $t_h\leq 2$ in each stratum h, the balanced half-sample variance estimator v_2 is an *overestimate*.

Further results on the biases of the alternative variance estimators considered here may be obtained under some simpliying assumptions for the important case $n_h=2$ ($h=1,\ldots,L$). When the distribution of the x population and the slopes β_h are the same in each stratum h, both jackknife variance estimators v_5 and v_6 tend to *underestimate* $MSE(r_1)$ for $t_h=t=1$ or 2. When $t_h=t=0$ or 1, the three balanced half-sample variance estimators are all *overestimates*.

Under the assumption that all parameters in (1.9) are the same in each stratum h (with the exception of the intercepts α_h), the mean square error of v_1 was found to be less than that of the balanced half-sample variance estimator v_2, although the two variance estimators are of somewhat comparable stability when $\bar{\alpha}=0$ provided L is moderately large or the coefficient of variation of the x population is relatively small. (Results on the stabilities of the remaining variance estimators could not be obtained.)

In an empirical study using data from the Current Population Survey, Kish & Frankel (1974) found that for a variety of nonlinear statistics the variance estimators based on the balanced half-sample technique were less stable than those based on the jackknife method, which in turn were less stable than the linearization variance estimator, although the differences encountered were small. Related studies by Bean (1975) and Lemeshow & Levy (1978) have subsequently confirmed this finding in the special case of ratio estimation. On the basis of the present analysis, Kish & Frankel's finding may be attributable to regression approximately through the origin and the small coefficients of variation (0.076 - 0.19) of the x populations involved. The models employed by Lemeshow & Levy (1978) were in fact limited to the case of regression through the origin with the coefficients of variation of the x populations in the range 0.01 - 0.3.

REFERENCES

Bean, J.A. (1975). Distribution and properties of variance estimators for complex multistage probability samples. *Vital and Health Statistics*, Series 2, No. 65. Washington, D.C.: U.S. Government Printing Office.

Jones, H.L. (1974). Jackknife estimation of functions of stratum means. *Biometrika 58*, 313-21.

Kish, L. & Frankel, M.R. (1974). Inference from complex samples (with discussion). *Journal of the Royal Statistical Society* B 36, 1-37.

Krewski, D. (1977). Linearization and replication methods in finite population sampling. Ph.D. thesis, Carleton University, Ottawa.

Lemeshow, S.A. & Levy, P.S. (1978). Estimating the variance of ratio estimates in complex sample surveys with two primary units per stratum - a comparison of balanced replication and jackknife techniques. *Journal of Statistical Computation and Simulation* 8, 191 - 205.

McCarthy, P.J. (1966). Replication: an approach to the analysis of data from complex surveys. *Vital and Health Statistics*, Series 2, No. 31. Washington, D.C.: U.S. Government Printing Office.

Rao, J.N.K. & Kuzik, R.A. (1974). Sampling errors in ratio estimation. *Sankhyā* C 36, 43-58.

Rao, J.N.K. & Ramachandran, V. (1974). Comparison of the separate and combined ratio estimate. *Sanklyā* C 36, 151-6.

Rao, J.N.K. & Webster, J.T. (1966). On two methods of bias reduction in the estimation of ratios. *Biometrika* 53, 571-7.

Rao, PS.R.S. (1974). Jackknifing the ratio estimator. *Sankhyā* C 36, 84-97.

Table 1. Coefficients D_i in $B(r_i)$, i=1,2,3

(original values multiplied by 1000)

L	a=1			a=2			a=3		
	D_1	D_2	D_3	D_1	D_2	D_3	D_1	D_2	D_3
2	333	-90	121	71	-6.7	32	30	-1.7	14
3	200	-13	129	45	-.78	30	20	-.15	13
4	143	-1.8	107	33	.13	25	14	.07	11
5	111	0.9	89	26	.30	21	11	.11	9.2
6	91	1.6	76	22	.31	18	9.5	.10	8.0
7	77	1.6	66	19	.28	16	8.1	.09	7.0
8	67	1.5	59	16	.24	14	7.1	.08	6.2
9	59	1.4	52	14	.21	13	6.3	.07	5.6
10	53	1.3	47	13	.19	12	5.6	.06	5.1
11	48	1.1	43	12	.16	11	5.1	.04	4.7
12	43	1.0	40	11	.14	10	4.7	*	*

* Values obtained are not accurate

Table 2. Coefficients L_i in MSE(v_i), i=1,2

(original values multiplied by 10^6)

L	t	a=1		a=2		a=3	
		L_1	L_2	L_1	L_2	L_1	L_2
3	0	555714	**	5438	28070	632	1437
	1	40476	**	6753	14196	2643	4283
	2	54374	**	36635	56518	30715	42734
4	0	99495	**	1683	5365	220	429
	1	15584	**	2781	5587	1103	1767
	2	30212	**	18634	30142	14890	21350
7	0	7717	25421	217	340	32	43
	1	2677	5016	508	697	204	252
	2	8845	12944	4602	5814	3393	4024
8	0	4534	13162	137	213	21	28
	1	1776	3389	339	476	136	172
	2	6482	9883	3247	4205	2356	2846
11	0	1370	2504	47	61	7	*
	1	672	996	130	159	52	*
	2	3006	3939	1387	1626	973	*
12	0	1001	1824	35	46	6	*
	1	516	787	100	125	40	*
	2	2421	3254	1094	1305	762	*

* Values obtained are not accurate.

** Not defined for La $\leq$ 4.