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 \(h=1, \ldots, L\). Within stratum \(h\), let \(Y_h\) and \(X_h\) denote the stratum means for variables \(y\) and \(x\) respectively and let \(y_h\) and \(x_h\) denote the population means. The overall population means are then \(\bar{Y}=\sum_{h=1}^{L} w_h Y_h\) and \(\bar{X}=\sum_{h=1}^{L} w_h X_h\) where \(W_N=w/n\) and \(N=N_h\).

The classical combined ratio estimator of \(R = \frac{\bar{Y}}{\bar{X}}\) is \(r = \frac{\sum_{h=1}^{L} w_h y_h}{\sum_{h=1}^{L} w_h x_h}\) where \(y_h = \frac{\sum_{i=1}^{n_h} y_{hi}}{n_h}\) and \(x_h = \frac{\sum_{i=1}^{n_h} x_{hi}}{n_h}\). 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

\[
\begin{align*}
r_2 &= (L+1)r_1 - \frac{\sum_{i=1}^{L} (r_1 - r_i^h)}{L} \quad (1.1) \\
\end{align*}
\]

respectively, where \(r_1 = \frac{r_1 + r_2}{2}\) and \(r_i = \frac{r_1 + r_i}{2}\) denotes the combined ratio estimator omitting \((y_{hi}, x_{hi})\).

The linearization variance estimator of the mean square error of \(r_1\) is

\[
\sigma^2 = \left\{ \frac{\sum_{i=1}^{L} \left[ (\bar{y}_{hi} - \bar{y}_h)^2 + (\bar{x}_{hi} - \bar{x}_h)^2 \right]}{L} \right\}
\]

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

\[
\begin{align*}
v_1 &= \frac{\sum_{i=1}^{k} (r_1(i) - r_1)^2}{k}, \\
v_2 &= \frac{\sum_{i=1}^{k} (r_1(i) - r_2(i))^2}{k}, \\
v_3 &= \frac{\sum_{i=1}^{k} (r_1(i) - r_i)^2}{k}, \\
v_4 &= \frac{\sum_{i=1}^{k} (r_1(i) - r_i(i))^2}{4k}
\end{align*}
\]

where \(r_1(i) = \frac{r_1 + r_i}{2}\). When all \(n_h=2\), two jackknife variance estimators proposed by Jones (1974) and Kish & Frankel (1974) are given by

\[
\begin{align*}
v_5 &= \sum_{i=1}^{k} \left[ (r_1(i) - r_1)^2 / 2 \right], \\
v_6 &= \sum_{i=1}^{k} \left[ (r_1(i) - r_1)^2 / 2 \right].
\end{align*}
\]

Exact small sample properties of these estimators may be obtained in the case of proportional allocation \((n_h = \bar{X}_h)\) under the linear regression model (Rad & Ramachandran, 1974)

\[
\begin{align*}
y_{hi} &= \hat{y}_h + \hat{\beta}_h x_{hi}, \\
E(e_{hi} | x_{hi}) &= 0, E(e_{hi}^2 | x_{hi}) &= \sigma^2_h, \\
E(e_{hi} e_{hj} | x_{hi} x_{hj}) &= 0 \quad (i \neq j) \\
E(e_{hi} e_{hi} | x_{hi} x_{hi}) &= 0 \quad (h \neq h')
\end{align*}
\]

(1.9)

The coefficient of variation of the auxiliary variable \(x_{hi}\) has often ranged between 0.4 and 1.0. Since the coefficient of variation of \(x_{hi}\) is \(a_h^{-2}\), small values of \(a_h\) will be of interest in what follows. In practice, \(a_h\) 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

\[
B(r_1) = E(r_1) - R = \frac{-n_a}{m(m-1)}
\]

(2.1)

provided \(m \geq 1\), where \(m=\bar{X}_h, m_1=n_h a, \sigma = \bar{X} \bar{y}_h/n, \sigma_a = \bar{X} a_h/x_{hi}/n_h\).

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

\[
\begin{align*}
B(r_2) &= -\frac{1}{m(m-1)} + \frac{1}{m} - L(l(b,a;2)) \sigma_a \\
B(r_3) &= -\frac{4L}{m(m-1)} - \frac{n}{m} - 2L(l(b,a;2)) \sigma_a
\end{align*}
\]

(2.2)

(2.3)
provided m>1. Here b=2(a(L-l)) and I(a, a; λ) =E(X_1 +X_2) -1 where λ>0 (X>a) and X_1 and X_2 are independent gamma variates with means a 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 λ1 and integral values of a and b

\[ I(a, b; λ) = E(X_1 + X_2) - I(λ1) \]

where λ1 >0 and X_1 and X_2 are independent gamma variates with means a 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 X*l and integral values of a and b

\[ I(a, b; λ) = E(X_1 + X_2) - I(1) - I(2) - I(3) \]

where

\[ I(1) = \sum_{k=1}^{a-1} (-1)^{k+1} \frac{Γ(b+k+1) - Γ(a-k)}{Γ(b) Γ(a)} \]

\[ I(2) = \frac{b-a+2}{a-1} \sum_{k=1}^{a-1} (-1)^{k+1} \frac{Γ(b+a-k)}{Γ(b) Γ(a)} \]

\[ I(3) = \frac{b-a+1}{a-1} \sum_{k=1}^{a-1} (-1)^{k+1} \frac{Γ(b+a-1)}{Γ(b) Γ(a)} \]

Since the expressions for the coefficients D_1 and D_2 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_3 appears particularly effective as a means of bias reduction in this case while the bias of r_1 approaches that of r_3 as L increases.

Where λ_1 = λ (h=1,...,L), the biases of both jackknife estimators do not depend on λ_1 regardless of the value of the a, h. When both the a_h and λ_h are not constant, however, the biases both r_1 and r_3 will in general involve the λ_h. In this case B(r_2) = λB(r_3) for λ_1 = λ. This implies that r_3 may be preferable to r_2 with respect to bias when the a, h are constant (Table 1), r_2 may be preferable to r_3 when λ_1 = λ and the a_h and λ_h are constant. Further, B(r_2) is always biased in the case of unequal a, 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

\[ MSE(r_1) = \frac{n^2(m-2)2^2}{m^2(m-1)(m-2)} + \frac{m^2(b-h)2^2}{m^2(m-1)(m-2)} + \sum_{h=1}^{L} \frac{f_h(t_h)5^2}{(m+2th-2)^2} \]

provided m>2, where b = 2(a(L-l)) and f_h(t) = \frac{Γ(a_h+t)}{Γ(a_h)}. After considerable algebra the bias of the linearization variance estimator under (1.9) with proportional allocation may be expressed as

\[ B(v_1) = E(V_1) - MSE(r_1) \]

\[ = - \frac{n^2(m-2)2^2}{m^2(m^2-1)(m-2)} + \frac{m^2(b-h)2^2}{m^2(m^2-1)(m-2)} \]

\[ - \sum_{h=1}^{L} \frac{f_h(t_h)5^2}{(m+2th-2)^2} \]

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

\[ B(v_2) = \frac{2n^2(m+2)(m-4)2^2}{m^2(m-1)(m-2)^2} \]

\[ + 4 \sum_{h=1}^{L} \frac{f_h(t_h)5^2}{(m+2th-2)^2} \]

\[ - \frac{1}{m+2th-2} \]

\[ \frac{1}{m+2th-2} \]

\[ = \frac{1}{m+2th-2} \]

provided m>4. From (3.2), it is easily seen that B(v_1) > 0 when t_h > 2/2 for all h. From (3.3), B(v_2) > 0 when t_h > 2/2 for all h. From (3.4), B(v_4) > 0 when t_h > 2/2 for all h, provided m>5.

Comparing (3.2) - (3.4) when h = 2 and λ_h = λ_h for all h shows B(v_1) > B(v_2) > B(v_4) when all t_h = 1 (provided m>5). For λ_1 = 0 and n_1 = 2 for all h, B(v_1) > B(v_2) > B(v_4) when all t_h = 1.

In the special case n_h = 2, a_1 = a_2, λ_1 = λ_2 and t_h = t for all h, the biases of the jackknife variance estimators v_1 and v_2 may be expressed as

\[ B(v_1) = F_1 / F_H \frac{B(v_1)}{B(v_2)} \]

\[ B(v_2) = F_2 / F_H \frac{B(v_2)}{B(v_4)} \]

where F_1 and F_2 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_m may also be expressed as B(v_m) = F_m / F_H, although the expressions for F_3 and F_4 are again not in closed form.

From (3.2) - (3.4), the biases of the three alternative variance estimators were compared for selected values of a and L and for t = 0, 1, and 2.
This analysis indicated that $B(v_i) = 0$ when $t=1$ or $2$ as indicated earlier. Both jackknife variance estimators $v_1$ and $v_2$ also underestimate $MSE(v_1)$ when $t=1$ or $2$ and $L >\frac{3}{2}$ with $B(v_1) > B(v_2) > B(v_3)$ in this case. For $t=3/2$, it was shown previously that $v_2$ and $v_3$ are both overestimates. The present analysis showed that $v_4$ is also an overestimate when $t=0$ or $1$ with $B(v_2) > B(v_3) > B(v_4)$. When $t=1$, $B(v_2) > B(v_3) > B(v_4) > B(v_5) > B(v_6)$ in this case. For $t=2$, all six variance estimators are underestimates when $a=0$ 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 $n=2$, $a_r = a$, $\beta_r = B$, $t = t_1$ and $\delta_r = \delta$ for all $h$. After considerable algebra, the mean square errors of these two variance estimators may be expressed as

$$\text{MSE}(v_i) = \frac{1}{t} [J_i^2 a^2 + K_i^2 \delta^2 + L_i^2 \beta^2] \quad (4.1)$$

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

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 $r_2$ is comparable to that of the combined ratio estimator $r$ for moderate values of $L$. When the distribution of the x population is not the same in all strata, however, $r_2$ can have smaller absolute bias than $r$. Moreover, both jackknife estimators may be biased in situations where the classical estimator is unbiased.

When $t_i > 1/2$ in each stratum $h$, the linearization variance estimator $v_i$ underestimates the mean square error of the combined ratio estimator. When $t_i = 1$ in each stratum $h$, the balanced half-sample variance estimator $v_4$ is an overestimate.

Further results on the biases of the alternative variance estimators considered here may be obtained under some simplifying assumptions for the important case $n=2$ ($h=1, \ldots, L$). When the distribution of the x population and the slopes $\beta_i$ are the same in each stratum $h$, both jackknife variance estimators $v_2$ and $v_3$ tend to underestimate $MSE(v_i)$ for $t_i = 1$ or $2$. When $t_i = 0$ 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 $q_i$), the mean square error of $v_1$ was found to be less than that of the balanced half-sample variance estimator $v_3$, although the two variance estimators are of somewhat comparable stability when $a=0$ 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


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

(Original values multiplied by 1000)

<table>
<thead>
<tr>
<th>L</th>
<th>$a=1$</th>
<th>$a=2$</th>
<th>$a=3$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$D_1$</td>
<td>$D_2$</td>
<td>$D_3$</td>
</tr>
<tr>
<td>2</td>
<td>333</td>
<td>-90</td>
<td>121</td>
</tr>
<tr>
<td>3</td>
<td>200</td>
<td>-13</td>
<td>129</td>
</tr>
<tr>
<td>4</td>
<td>143</td>
<td>-1.8</td>
<td>107</td>
</tr>
<tr>
<td>5</td>
<td>111</td>
<td>0.9</td>
<td>89</td>
</tr>
<tr>
<td>6</td>
<td>91</td>
<td>1.6</td>
<td>76</td>
</tr>
<tr>
<td>7</td>
<td>77</td>
<td>1.6</td>
<td>66</td>
</tr>
<tr>
<td>8</td>
<td>67</td>
<td>1.5</td>
<td>59</td>
</tr>
<tr>
<td>9</td>
<td>59</td>
<td>1.4</td>
<td>52</td>
</tr>
<tr>
<td>10</td>
<td>53</td>
<td>1.3</td>
<td>47</td>
</tr>
<tr>
<td>11</td>
<td>48</td>
<td>1.1</td>
<td>43</td>
</tr>
<tr>
<td>12</td>
<td>43</td>
<td>1.0</td>
<td>40</td>
</tr>
</tbody>
</table>

* Values obtained are not accurate
Table 2. Coefficients $L_i$ in MSE($v_i$), $i=1,2$

(original values multiplied by $10^6$)

<table>
<thead>
<tr>
<th>L</th>
<th>t</th>
<th>$a=1$</th>
<th></th>
<th>$a=2$</th>
<th></th>
<th>$a=3$</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$L_1$</td>
<td>$L_2$</td>
<td>$L_1$</td>
<td>$L_2$</td>
<td>$L_1$</td>
<td>$L_2$</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>555714</td>
<td>**</td>
<td>5438</td>
<td>28070</td>
<td>632</td>
<td>1437</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>40476</td>
<td>**</td>
<td>6753</td>
<td>14196</td>
<td>2645</td>
<td>4283</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>54374</td>
<td>**</td>
<td>56635</td>
<td>56518</td>
<td>30715</td>
<td>42734</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>99495</td>
<td>**</td>
<td>1683</td>
<td>5365</td>
<td>220</td>
<td>429</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>15584</td>
<td>**</td>
<td>2781</td>
<td>5587</td>
<td>1103</td>
<td>1767</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>30212</td>
<td>**</td>
<td>18634</td>
<td>30142</td>
<td>14890</td>
<td>21350</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>7717</td>
<td>25421</td>
<td>217</td>
<td>340</td>
<td>32</td>
<td>43</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>2677</td>
<td>5016</td>
<td>508</td>
<td>697</td>
<td>204</td>
<td>252</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>8845</td>
<td>12944</td>
<td>4602</td>
<td>5814</td>
<td>3393</td>
<td>4024</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>4534</td>
<td>13162</td>
<td>137</td>
<td>213</td>
<td>21</td>
<td>28</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>1776</td>
<td>3389</td>
<td>339</td>
<td>476</td>
<td>136</td>
<td>172</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>6482</td>
<td>9883</td>
<td>3247</td>
<td>4205</td>
<td>2356</td>
<td>2846</td>
</tr>
<tr>
<td>11</td>
<td>0</td>
<td>1370</td>
<td>2504</td>
<td>47</td>
<td>61</td>
<td>7</td>
<td>*</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>672</td>
<td>996</td>
<td>130</td>
<td>159</td>
<td>52</td>
<td>*</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>3006</td>
<td>3939</td>
<td>1387</td>
<td>1626</td>
<td>973</td>
<td>*</td>
</tr>
<tr>
<td>12</td>
<td>0</td>
<td>1001</td>
<td>1824</td>
<td>35</td>
<td>46</td>
<td>6</td>
<td>*</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>516</td>
<td>787</td>
<td>100</td>
<td>125</td>
<td>40</td>
<td>*</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>2421</td>
<td>3254</td>
<td>1094</td>
<td>1305</td>
<td>762</td>
<td>*</td>
</tr>
</tbody>
</table>

* Values obtained are not accurate.

** Not defined for La < 4.