Jour. Ind. Soc. Ag. Statistics
58(2), 2004 : 153-162

Conditional Inference under Two-phase Sampling

Girish Kumar Jha, A.K. Srivastava and Anil Rai
Indian Agricultural Statistics Research Institute, New Delhi-110012
(Received : February, 2002)

SUMMARY

The most common criticism of inferences based on randomization is that it does not take enough account of the available partial knowledge regarding the sample while drawing the inferences about the parameter of interest. In this paper, some aspect of this criticism has been addressed and an attempt has been made to show how to use approximations to counter it in the particular case of double sampling for ratio estimation in sample surveys. Here, a conditional bias adjusted ratio type estimator has been proposed under two-phase sampling. A conditionally adjusted variance estimator has also been conjectured for the proposed conditional estimator, which has been supported through simulation results. Further, a simulation study of two populations, one where the model on which ratio estimator is based holds and one where it does not, has been undertaken to show the improvements in the performance of proposed estimator as compared to existing estimator under conditional framework.

Key words : Conditional inference, Ratio estimator, Two-phase sampling, Simulation.

1. Introduction

In case of ratio method of estimation, the efficiency for the estimation of population parameter, i.e., mean or total etc. has been improved by exploiting the correlation between the auxiliary variable X with the character under study Y. In a number of situations, it happens that the population mean $\overline{X}$ is not known and the ratio estimator cannot be used to estimate the population mean $\overline{Y}$. Obviously one of the procedure in such situations is to use the method of two-phase or double sampling. The first-phase sample gives a good estimate $\overline{x}'$ of the population mean $\overline{X}$ while the second-phase subsample in which Y is measured is employed to estimate the population mean $\overline{Y}$ through ratio estimation using $\overline{x}'$. Suppose a simple random sample s′ of size n′ is taken without replacement from the population of N elements and X_i alone is observed for all elements $i \in s'$. A simple random subsample s of size n is then drawn without replacement from s′ and Y_i is observed for all $i \in s$.

The usual ratio type estimator of $\overline{Y}$ in case of double sampling is

$$\bar{y}_{rd} = \frac{\bar{y}}{\bar{x}}\bar{x}' = \hat{R}\,\bar{x}' \tag{1}$$

where $\bar{y}$ and $\bar{x}$ are the means based on s and $\bar{x}'$ is the mean based on s′. This estimator is design consistent or p-consistent for $\overline{Y}$.

Clearly, this is a biased estimator of $\overline{Y}$ and its relative bias is given by

$$\text{Relative Bias} = B(\bar{y}_{rd}) = \left(\frac{1}{n} - \frac{1}{n'}\right)\left(C_x^2 - \rho C_x C_y\right) \tag{2}$$

where, C_x and C_y denote the coefficient of variation of X and Y respectively and ρ denotes the correlation coefficient between X_i and Y_i. It has been shown that the bias will be negligible if n is sufficiently large but it will be serious for small sample size. Royall and Cumberland (1981) pointed out that in a number of examples of real data where a ratio estimator might be used, the estimator and its estimated variance can be badly biased unless the sample is balanced with respect to the X variable, in particular, when $\bar{x}$ and $\overline{X}$ are not close.

To the first order of approximation, the variance of this estimator is given by

$$V(\bar{y}_{rd}) = \left(\frac{1}{n} - \frac{1}{n'}\right)S_D^2 + \left(\frac{1}{n'} - \frac{1}{N}\right)S_y^2 \tag{3}$$

with

$$S_D^2 = \frac{1}{N-1}\sum_{i=1}^{N} D_i^2, \quad S_y^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(Y_i - \overline{Y}\right)^2$$

$$S_X^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(X_i - \overline{X}\right)^2 \text{ and } R = \frac{\overline{Y}}{\overline{X}}$$

where $D_i = Y_i - RX_i$

A design – consistent estimator of $V(\bar{y}_{rd})$ is expressed as

$$v_{rd} = \hat{V}(\bar{y}_{rd}) = \left(\frac{1}{n} - \frac{1}{n'}\right)s_d^2 + \left(\frac{1}{n'} - \frac{1}{N}\right)s_y^2 \tag{4}$$

with

$$s_d^2 = \frac{1}{n-1}\sum_{i\in s} d_i^2, \quad s_y^2 = \frac{1}{n-1}\sum_{i\in s}\left(y_i - \bar{y}\right)^2$$

where, $d_i = y_i - \hat{R}\,x_i$. The last term in (4) is obtained by using the sample variance s_y^2 to estimate the population variance S_y^2. In this paper, the conditionally unbiased ratio estimator and estimation of its conditional variance have been investigated under two phase sampling design.

2. A Conditional Adjustment

The most common criticism of making inferences under randomization framework is due to non-utilizing the available partial knowledge regarding the sample while drawing the inference about the parameter of interest. In conditional approach estimators are adjusted for conditional biases utilizing the information available in the sample. In this study, an attempt will be made to develop conditionally unbiased estimators in the particular case of double sampling for ratio estimation in sample surveys. In case of double sampling information regarding auxiliary variable X for the first and second phase samples can be used to make inferences conditional on $\bar{x}'$ and $\bar{x}$. The estimates of the bias will be derived and so an adjusted estimator and an estimate of the variance will be obtained under the assumption of the asymptotic distribution of $(\bar{y}, \bar{x})$ as bivariate normal. Of course, $\bar{x}'$ and $\bar{x}$ are considered in some small interval rather than a fixed value in the exact finite case, but the approximation will still be valid for large enough n and n′.

Following the notations used by Robinson (1987), let $\sigma_{yy} = \frac{(1-f)}{(N-1)} \sum_{i=1}^{N} (Y_i - \bar{Y})^2$ and define σ_{xx} and σ_{yx} similarly where $f = \frac{n}{N}$.

Let $\beta = \frac{\sigma_{yx}}{\sigma_{xx}}$ and let $\sigma_{yy.x} = \sigma_{yy} - \beta\sigma_{yx}$. Accordingly, in case of double sampling β', σ'_{yy}, σ'_{xx} and σ'_{yx} are defined for large sample n'.

Then it is known that

$$n^{1/2}\left(\frac{\bar{y} - \bar{Y} - \beta(\bar{x} - \bar{X})}{\sigma_{yy.x}^{1/2}}, \frac{\bar{x} - \bar{X}}{\sigma_{xx}^{1/2}}\right)$$

tends in distribution to a pair of independent standard normal variates.

Thus, approximately

$$E\{E(\bar{y}_{rd}|\bar{x})|s'\} = E\left\{E\left(\frac{\bar{y}}{\bar{x}}\bar{x}'\middle|\bar{x}\right)\middle|s'\right\}$$

$$= E\left[\{\bar{y}' + \beta'(\bar{x} - \bar{x}')\}\frac{\bar{x}'}{\bar{x}}\middle|\bar{x}'\right]$$

$$= E\left[\left\{\bar{y}' - \frac{\bar{y}'}{\bar{x}'}(\bar{x} - \bar{x}')\frac{\bar{x}'}{\bar{x}} + \beta'(\bar{x} - \bar{x}')\frac{\bar{x}'}{\bar{x}}\right\}\middle|\bar{x}'\right]$$

$$= \mathrm{E}\left[\left\{\bar{y}' - \left(\frac{\bar{y}'}{\bar{x}'} - \beta'\right)(\bar{x} - \bar{x}')\frac{\bar{x}'}{\bar{x}}\right\} \middle| \bar{x}'\right]$$

$$= \mathrm{E}(\bar{y}' \mid \bar{x}') - \mathrm{E}\left\{\left(\frac{\bar{y}'}{\bar{x}'} - \beta'\right)(\bar{x} - \bar{x}')\frac{\bar{x}'}{\bar{x}} \middle| \bar{x}'\right\}$$

$$= \left[\bar{Y} + \beta(\bar{x}' - \bar{X})\right]\frac{\bar{x}'}{\bar{x}} + \beta(\bar{x} - \bar{x}')\frac{\bar{x}'}{\bar{x}}$$

$$= \frac{\bar{Y}\,\bar{x}'}{\bar{x}} + \beta\,\bar{x}'\left(1 - \frac{\bar{X}}{\bar{x}}\right)$$

$$= \bar{Y} - \mathrm{R}(\bar{x} - \bar{x}')\frac{\bar{X}}{\bar{x}} + \beta\,\bar{x}'\left(1 - \frac{\bar{X}}{\bar{x}}\right)$$

$$= \bar{Y} - \mathrm{R}\left(1 - \frac{\bar{x}'}{\bar{x}}\right)\bar{X} + \beta\,\bar{x}'\left(1 - \frac{\bar{X}}{\bar{x}}\right)$$

So, the asymptotic bias is

$$\mathrm{B}(\bar{y}_{rd}) \cong -\left\{\mathrm{R}\bar{X}\left(1 - \frac{\bar{x}'}{\bar{x}}\right) - \beta\,\bar{x}'\left(1 - \frac{\bar{X}}{\bar{x}}\right)\right\}$$

The bias-adjusted estimator of $\bar{Y}$ can be written as

$$\bar{y}_{rdc} = \bar{y}_{rd} + (\hat{R} - b)(\bar{x} - \bar{x}')\frac{\bar{x}'}{\bar{x}} \qquad (5)$$

where, $\hat{R} = \frac{\bar{y}}{\bar{x}}$ and b is the regression coefficient for Y_i on X_i (i = 1, 2, ..., n).

Then, using the asymptotic results on functions of moments [see, e.g., Cramer (1946, (27.7.3))], we have

$$\bar{x} - \bar{X} = O_p(n^{-1/2})$$

so $$\mathrm{E}\left(\hat{R} \middle| \bar{x}\right) = \mathrm{E}\left(\frac{\bar{y}}{\bar{x}} \middle| \bar{x}\right) = \frac{\bar{Y} + \beta(\bar{x} - \bar{X})}{\bar{x}} = \frac{\bar{Y} + \beta\{O_p(n^{-1/2})\}}{\bar{X} + O_p(n^{-1/2})}$$

$$= \frac{\bar{Y} + \beta\{O_p(n^{-1/2})\}}{\bar{X}}\left\{1 - \frac{O_p(n^{-1/2})}{\bar{X}} + \frac{O_p(n^{-1})}{\bar{X}^2} + \ldots\right\}$$

$$= \mathrm{R} + \frac{\beta\{O_p(n^{-1/2})\}}{\bar{X}}\left\{1 - \frac{O_p(n^{-1/2})}{\bar{X}}\right\}$$

Hence, $E\left(\hat{R}\,\middle|\,\bar{x}\right) = R + O_p(n^{-1/2})$

Similarly, it can be shown that

$$E\left(b\,\middle|\,\bar{x}\right) = \beta + O_p(n^{-1/2})$$

Thus, using the above results, it can be shown that approximately

$$E\{E(\bar{y}_{rdc}\,|\,\bar{x})\,|\,\bar{x}'\} = \bar{Y} + O_p(n^{-1}) \tag{6}$$

where the term $O_p(n^{-\alpha})$ denotes a quantity that when multiplied by n^{α} remains bounded in probability.

A conditionally adjusted variance estimate for $\bar{y}_{rd}$ is conjectured as

$$v_{rdc} = v_{rd}\frac{\bar{x}'^2}{\bar{x}^2} \tag{7}$$

This expression for variance estimate will be used in the empirical study for making comparison with the simulated estimate of variance for proposed double sampling ratio estimator.

3. A Simulation Study

To study the performance of the proposed estimator and its conditionally adjusted variance estimate relative to the standard one a limited simulation study is conducted. The properties of these estimators have been studied for the two populations, one is based on the linear form of model, i.e., model (A) and other is based on the quadratic form of model, i.e., model (B). Model (A) and (B) are presented below

$$Y = X + Z\frac{X^{1/2}}{5} \tag{A}$$

and $$Y = X + 0.05\,X^2 + \frac{ZX^{1/2}}{5} \tag{B}$$

where Z is a standard normal variate.

Under the first population the usual ratio estimator is the best and under the second population it does not possesses its desirable properties. The populations of size 1,000 each are generated by obtaining 1,000 X values from the distribution of $\frac{1}{2}\chi_6^2$ or Gamma (3) and for each value of X the value of Y has been obtained using model (A) and (B). In this investigation gamma distribution is used to generate the data because it produces variability in the population. Hence, it is worth mentioning that while deriving the theoretical results the distribution of ($\bar{y}, \bar{x}$) is assumed as asymptotically bivariate normal, but for the simulation study the population is generated with the help of gamma distribution, which is a deviation from the assumption.

From each of the above populations 1,000 independent two-phase random samples for n′ = 100, n = 20 and n′ = 400, n = 80 have been selected with the help of simple random sampling without replacement. For each sample the actual bias corresponding to the traditional estimator y_{rd} and suggested estimator $\bar{y}_{rdc}$ as well as their variance estimates have been calculated. To see how the performance of these estimators and of their variance estimates depends on $\bar{x}'/\bar{x}$, the 1,000 samples from each population have been arranged in order of increasing values of $\bar{x}'/\bar{x}$. It is to note that $\bar{x}'/\bar{x}$ may be treated as approximately ancillary since $\bar{x}'$ is based on a large preliminary sample. Then the samples were grouped in 10 sets of 100 each so that the first group contains the 100 samples where $\bar{x}'/\bar{x}$ values are smallest, the next group contains the samples with the next 100 smallest $\bar{x}'/\bar{x}$ values and so on. For each of these 10 groups, the averages of biases and variance estimates corresponding to y_{rd} and $\bar{y}_{rdc}$ have been computed.

To provide the empirical evidence to the conjectured formula for conditionally adjusted variance estimate, i.e., v_{rdc} the simulated variance, denoted by v_s, for the proposed estimator has been obtained and it is found very close to the v_{rdc}. Lastly, the average biases and the values of $(\bar{v}_{rd})^{½}$ and $(\bar{v}_{rdc})^{½}$ have been plotted against the group average of $\bar{x}'/\bar{x}$ in Figure 1 and 2 .

Table 1 and Figure 1 correspond with n′ = 400 and n = 80 under model (A). The case n′ = 100 and n = 20 gave a qualitatively similar results and thus it is not presented. Similarly, under model (B), the cases n′ = 100 and n = 20 as well as n′ = 400 and n = 80 gave a qualitatively similar results. Hence, Table 2 and Figure 2 refers to n′ = 400 and n = 80 and the other cases are omitted. Each plotted point represents 100 samples and each figure shows the results for 1,000 independent two-phase random samples for n′ = 400 and n = 80 from one population. For each population there are four trajectories. The trajectories showing average biases are labelled as $\bar{y}_{rd} - \bar{Y}$ and $\bar{y}_{rdc} - \bar{Y}$. The other two are showing standard error of group variance estimate, i.e., $\sqrt{\bar{v}_{rd}}$ and $\sqrt{\bar{v}_{rdc}}$. The group average of $\bar{x}'/\bar{x}$ is shown on the abscissa. In each case $\sqrt{\bar{v}_{rdc}}$ and simulated variance estimate, $\sqrt{\bar{v}_s}$ were virtually indistinguishable, so only $\sqrt{\bar{v}_{rdc}}$ was plotted.

Table 1. Results of simulation under model (A): $Y = X + ZX^{1/2}/5$ for 1,000 two-phase random samples with n = 80 and n' = 400

Avg. in 100 samples			(Avg. in 100 samples)$^{1/2}$		
$\bar{x}'/\bar{x}$	$\bar{y}_{rd} - \bar{Y}$	$\bar{y}_{rdc} - \bar{Y}$	v_{rd}	v_{rdc}	v_s
0.909	0.008	0.007	0.083	0.075	0.080
0.946	0	0	0.080	0.075	0.075
0.965	-0.008	-0.008	0.079	0.076	0.078
0.982	-0.002	-0.002	0.078	0.077	0.074
0.997	0.007	0.007	0.077	0.077	0.081
1.011	-0.002	-0.003	0.075	0.076	0.081
1.026	0.001	0.001	0.075	0.077	0.073
1.043	-0.009	-0.009	0.074	0.077	0.082
1.069	-0.006	-0.007	0.072	0.077	0.078
1.114	0.001	0	0.070	0.078	0.074

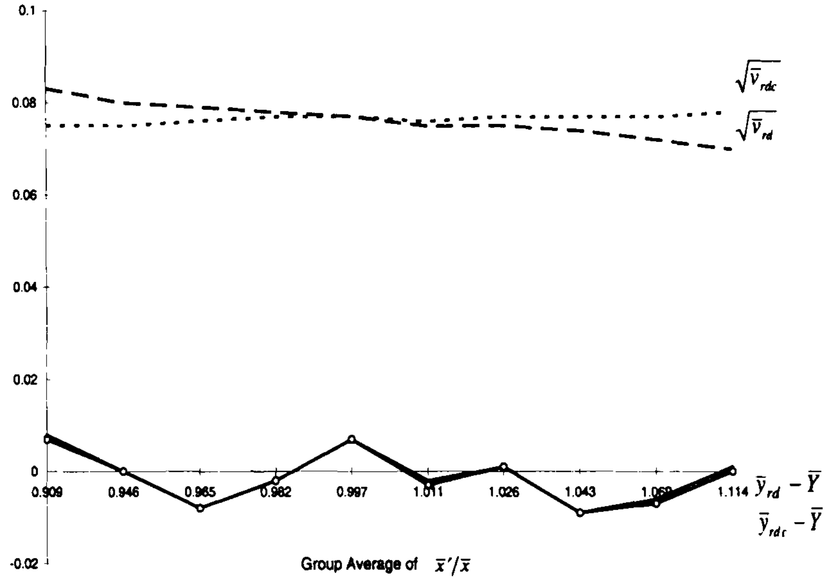

Fig. 1. Comparison of Conditional Bias and Conditional Standard Error of group variance estimate for ratio estimator under two-phase sampling (Model A, n' = 400, n = 80)

Table 2. Results of simulation under model (B): $Y = X + 0.05X^2 + ZX^{1/2}/5$ for 1,000 two-phase random samples with n' = 400 and n = 80

	Avg. in 100 samples		(Avg. in 100 samples)$^{1/2}$		
$\bar{x}'/\bar{x}$	$\bar{y}_{rd} - \bar{Y}$	$\bar{y}_{rdc} - \bar{Y}$	v_{rd}	v_{rdc}	v_s
0.909	0.065	0.008	0.121	0.110	0.107
0.946	0.032	0.002	0.115	0.108	0.099
0.965	0.009	-0.012	0.113	0.109	0.107
0.982	0.011	0.001	0.112	0.110	0.101
0.997	0.009	0.007	0.109	0.109	0.108
1.011	-0.013	-0.008	0.105	0.106	0.110
1.026	-0.013	0.002	0.106	0.109	0.103
1.043	-0.038	-0.015	0.103	0.108	0.113
1.069	-0.049	-0.012	0.100	0.107	0.104
1.114	-0.062	-0.002	0.096	0.107	0.101

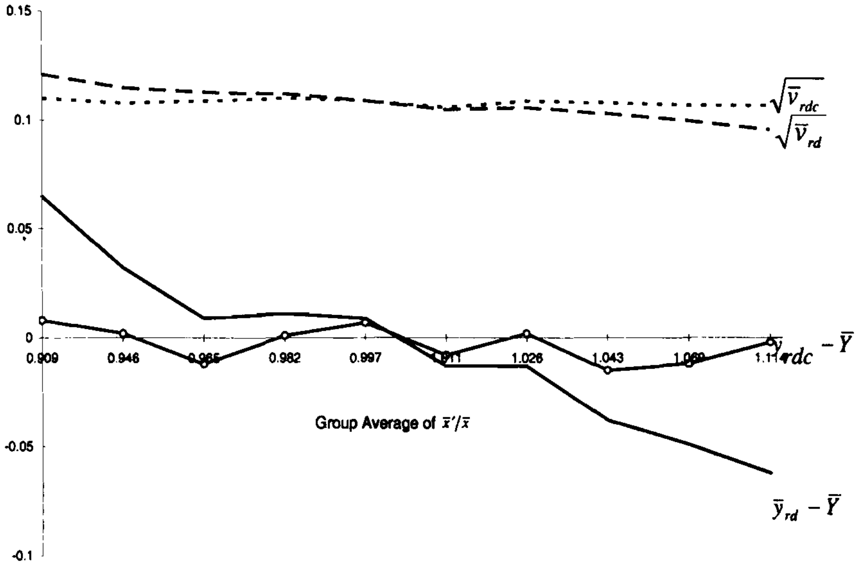

Fig. 2. Comparison of Conditional Bias and Conditional Standard Error of group variance estimate for ratio estimator under two-phase sampling. (Model B, n' = 400, n = 80)

It is clear from Figure 1 that under the model (A), there is little bias in the usual estimator $\bar{y}_{rd}$ because it is based on this particular model. The proposed estimator $\bar{y}_{rdc}$ is close to the population mean $\bar{Y}$ due to its conditionally unbiased nature. The $\sqrt{v_{rdc}}$ is almost constant whereas $\sqrt{v_{rd}}$ decreases as $\bar{x}'/\bar{x}$ increases. .Only in balanced samples for which $\bar{x}'/\bar{x} \cong 1$ both $\sqrt{v_{rd}}$ and $\sqrt{v_{rdc}}$ are equal. Surprisingly, there appears to be some bias in $\bar{y}_{rd}$ which may be due to the relatively small "finite population" considered for this study.

It is well known that the ratio estimator in case of double sampling is almost unbiased in simple random sampling. However, the average bias curve labelled as $\bar{y}_{rd} - \bar{Y}$ in Figure 2 reveals that in this population it is true only because a positive bias when $\bar{x}'/\bar{x} < 1$ is matched by a negative bias when $\bar{x}'/\bar{x} > 1$. In effect, probability sampling theory refers to the projection of the biases on to the vertical axis. From Figure 2 it is apparent that under model (B) usual estimator $\bar{y}_{rd}$, is badly biased except in balanced samples for which $\bar{x}'/\bar{x} \cong 1$. But this bias is largely corrected by suggested estimator $\bar{y}_{rdc}$ because it is conditionally unbiased. Here, again $\sqrt{v_{rd}}$ decreases with $\bar{x}'/\bar{x}$ whereas $\sqrt{v_{rdc}}$ is almost constant. In case of balanced samples, both $\sqrt{v_{rd}}$ and $\sqrt{v_{rdc}}$ are equal. This implies that in case of unbalanced samples, for both populations the existing variance estimate is either over-estimating or under-estimating the true population variance.

4. Conclusions

The standard randomization analysis does not take enough account of the observations on the auxiliary variable, so estimators and its estimated variance can be conditionally biased unless the sample is balanced with respect to the auxiliary variable. Hence, in this article appropriate adjustment for the bias of the estimators based on asymptotic approximations has been made in case of double sampling design under the randomization theory, which also leads to conditionally valid inferences.

To conclude, following points may be noted on the basis of simulation studies made in this paper.

(i) Proposed conditional estimator as well as its unconditional counterpart are almost unbiased under the first population, which is based on the linear form of model. However, in some cases there appear some biases that may be due to relatively small "finite population" considered for this study.

(ii) Existing estimators are conditionally biased under the second population unless the sample is balanced with respect to the auxiliary variable. However, this conditional bias is largely corrected by the proposed conditional estimators.

(iii) The standard error (S.E.) of group variance estimate of proposed conditional estimator, i.e., $\sqrt{\mathrm{v}_{\mathrm{rdc}}}$ is almost constant whereas S.E. of group variance estimate of existing estimator, i.e., $\sqrt{\mathrm{v}_{\mathrm{rd}}}$ decreases as $\bar{\mathrm{x}}'/\bar{\mathrm{x}}$ increases. Only in balanced samples for which $\bar{\mathrm{x}}'/\bar{\mathrm{x}} \cong 1$ both $\sqrt{\mathrm{v}_{\mathrm{rd}}}$ and $\sqrt{\mathrm{v}_{\mathrm{rdc}}}$ are equal.

ACKNOWLEDGEMENT

Authors are thankful to the referee for critically reviewing the original version of the manuscript.

REFERENCES

Casady, R.J. and Valliant, R. (1993). Conditional properties of post-stratified estimators under normal theory. *Survey Methodology*, **19**, 183-192.

Cramer, H. (1946). *Mathematical Methods of Statistics*. Princeton University Press, Princeton.

Holt, D. and Smith, T.M.F. (1979). Post-stratification. *J. Roy. Statist. Soc.*, **A142**, 33-46.

Rao, J.N.K. (1985). Conditional inference in survey sampling. *Survey Methodology*, **11**, 15-31.

Rao, J.N.K. (1997). Developments in sample survey theory: An appraisal. *Canad. J. Statist.*, **25(1)**, 1- 21.

Robinson, J. (1987). Conditioning ratio estimates under simple random sampling. *J. Amer. Statist. Assoc.*, **82**, 826-831.

Royall, R.M. and Cumberland, W.G. (1981). An empirical study of the ratio estimator and estimators of variance. *J. Amer. Statist. Assoc.*, **76**, 66-88.