SAMPLING ERRORS IN RATIO ESTIMATION

J. N. K. Rao and Richard A. Kuzik
University of Manitoba and Department of Indian Affairs and Northern Development, Canada

SUMMARY: The results of an extensive semi-empirical study on ratio estimators are reported. Assuming simple random sampling and a linear regression model $y = \alpha+\beta x+e$ with error variance $V(e|x)\propto x^g$, $g\geq 0$, the average mean square errors of seven ratio estimators of the population mean $\bar{Y}$ and the average biases and average mean square errors of two classical variance estimators and of the 'jack-knife' variance estimator, for a given x-population, are derived. Employing a wide variety of x-populations, the performances of these estimators and variance estimators are empirically investigated. The applicability of the results to other sample designs is discussed.

1. INTRODUCTION: In recent years considerable attention has been given to the investigation of the properties of ratio estimators. Suppose a bivariate simple random sample (y_i,x_i), $i=1,2,..,n$, is to be drawn from a population of N units with means $(\bar{Y},\bar{X})$. If $\bar{X}$ is known, the classical ratio estimator of $\bar{Y}$ is $\bar{y}_r = r\bar{X}$ where $r = \bar{y}/\bar{x}$ is the ratio of sample means. Beale's estimator

$$\bar{y}_B = \bar{y}_r[1 + (\frac{1}{n} - \frac{1}{N})\frac{s_{xy}}{\bar{x}\bar{y}}]/[1 + (\frac{1}{n} - \frac{1}{N})\frac{s_x^2}{\bar{x}^2}]$$

and Tin's estimator

$$\bar{y}_T = \bar{y}_r[1 + (\frac{1}{n} - \frac{1}{N})(\frac{s_{xy}}{\bar{x}\bar{y}} - \frac{s_x^2}{\bar{x}^2})]$$

are approximately unbiased in the sense that their asymptotic biases do not contain terms of order n^{-1} and order N^{-1}, where $s_{xy} = (n-1)^{-1}\Sigma(x_i-\bar{x})\cdot(y_i-\bar{y})$ and $s_x^2 = (n-1)^{-1}\Sigma(x_i-\bar{x})^2$. The well known unbiased ratio estimator of Hartley and Ross is given by

$$\bar{y}_H = r\bar{X} + \frac{n(N-1)}{N(n-1)}(\bar{y} - r\bar{x})$$

where $r = n^{-1}\Sigma(y_i/x_i)$. If the sample is divided at random into $g(>2)$ groups each of size p, $(n = pg)$, Jones's estimator of $\bar{Y}$ is

$$\bar{y}_J = w\bar{y}_r - (w-1)\bar{r}'_g\bar{X}$$

where $w = g\{1-(n-p)/N\}$, $\bar{r}'_g = r'_j/g$, $r'_j = (n\bar{y}-p\bar{y}_j)/(n\bar{x}-p\bar{x}_j)$, $\bar{y}_j$ and $\bar{x}_j$ being the sample means from the j^{th} group $(j=1,2,..,g)$. The asymptotic bias of $\bar{y}_J$ does not involve terms of order n^{-1} and order N^{-1}. Quenouille's estimator (originally proposed for infinite populations) is given by $\bar{y}_Q = g\bar{y}_r-(g-1)\bar{r}'_g\bar{X}$, but terms of order N^{-1} appear in its asymptotic bias. Finally, Mickey's unbiased estimator of $\bar{Y}$ is

$$\bar{y}_M = \bar{r}'_g\bar{X} + \frac{(N-n+p)g}{N}(\bar{y} - \bar{r}'_g\bar{x}).$$

$\bar{y}_M$ reduces to $\bar{y}_H$ when n=2. The approximately unbiased or wholly unbiased ratio estimators are useful in surveys with many strata and small samples within strata, especially if 'separate' ratio estimators are appropriate.

We consider three estimators for the mean square error (MSE) of $\bar{y}_r$:

$$v_1 = (n^{-1} - N^{-1})(s_y^2 - 2rs_{xy} + r^2s_x^2)$$

$$v_2 = (n^{-1} - N^{-1})(\bar{X}/\bar{x})^2(s_y^2 - 2rs_{xy} + r^2s_x^2)$$

and the 'jack-knife' estimator

$$v_3 = \bar{X}^2(1 - n/N)\{(g-1)/g\}\Sigma(r'_j - \bar{r}'_g)^2$$

where $s_y^2 = (n-1)^{-1}\Sigma(y_i-\bar{y})^2$. The estimators $\bar{X}^{-2}v_2$ and $\bar{X}^{-2}v_3$ for MSE of r do not require the knowledge of $\bar{X}$.

The behaviour of estimators of $\bar{Y}$ and of estimators of MSE (or variance) of $\bar{y}_r$ (or r) may be investigated in a variety of ways, including the following (McCarthy [6]): (a) Exact analytic, in which the functional form of a distribution or a joint distribution is assumed; (b) approximate analytic, in which Taylor series approximations are used; (c) empirical studies, in which the data from actual surveys are used; and (d) Monte Carlo sampling from synthetic populations. We refer the reader to Hutchison [3], Rao and Rao [10], McCarthy [6] and Frankel [2] for details of the available results under the above categories. Some analytic results, exact for any sample size, have been obtained by assuming simple random sampling and the model

$$y_i = \alpha + \beta x_i + u_i$$

$$\varepsilon(u_i|x_i) = 0,\ \varepsilon(u_i^2|x_i) = \delta x_i^t,\ \delta > 0,\ t \geq 0 \quad (1)$$

$$\varepsilon(u_iu_j|x_i,x_j) = 0,\ i \neq j = 1, 2,\ldots,N$$

with a gamma distribution for the variates x_i, where ε denotes the expectation operator (Rao and Rao [10]). However, it is difficult to obtain analogous exact analytic results for more complex sample designs. On the other hand, the empirical approach, employing actual survey data, permits the use of complex designs, and the properties of estimators of many parameters (including $\bar{Y}$ or $R = \bar{Y}/\bar{X}$) could be investigated with the help of a high speed computer. For instance, Frankel [2] investigated the properties of naive estimators of ratios, regression coefficients, simple, partial and multiple correlation coefficients for three sample designs involving 6, 12 and 30 strata respectively and two clusters per stratum selected by simple random sampling. Employing the data collected by the U.S. Bureau of the Census in the March 1971 Current Population Survey, he generated 300 (or 200) independent samples for each design and then empirically investigated the behaviour of estimators and of variance estimators. Several y-characters and a single x-character (size of cluster) were considered in estimating R. On the other hand, Rao [9] considered a wide variety of (y,x)-populations, but confined himself to simple random sampling. An obvious limitation of the empirical approach is that the results are strictly applicable only to the particular population(s) considered. However, the empirical studies are extremely valuable in providing guidelines on the performances of various methods of estimation.

In this paper we employ a semi-empirical approach by using model (1) and a wide variety of x-populations. This combination of empirical and analytic approaches has obvious advantages as it throws further light on the performances of various methods of estimation. Moreover, it has not been possible to analytically investigate the stabilities of Beale's estimator $\bar{y}_B$ and the 'jack-knife' variance estimator v_3 (except for g=2). Although we confine ourselves to simple random sampling in this paper, it is possible to apply the present approach to more complex designs by employing suitable extensions of model (1). For instance, Konijn [5] has proposed an extension of model (1) suitable for

two-stage sampling involving unequal probability sampling of primaries with or without replacement.

2. ESTIMATION OF $\bar{Y}$: The average MSE of $\bar{y}_r$, under model (1), is given by

$$\varepsilon\, MSE(\bar{y}_r) = \varepsilon\, E_s(\bar{y}_{rs} - \bar{Y})^2$$

$$= \varepsilon\, E_s \{\frac{\bar{X}}{\bar{x}_s}(\alpha+\beta\bar{x}_s+\bar{u}_s) - (\alpha+\beta\bar{X}+\bar{U})\}^2$$

where $\bar{U} = \frac{\Sigma}{1}^{N} u_i$, E_s denotes the average over all the $\binom{N}{n}$ possible samples s each of size n, $\bar{x}_s = \sum_{i\varepsilon s} x_i$ and $\bar{u}_s = \sum_{i\varepsilon s} u_i$. Noting that $n^2\varepsilon(\bar{u}_s^2) = \delta \sum_{i\varepsilon s} x_i^t$, $N^2\varepsilon(\bar{U}^2) = \delta\frac{\Sigma}{1}^{N} x_i^t$ and $nN\varepsilon(\bar{u}_s\bar{U}) = \delta \sum_{i\varepsilon s} x_i^t$, we get

$$\varepsilon\, MSE(\bar{y}_r) = \frac{\delta}{N^2}\frac{\Sigma}{1}^{N} x_i^t + \alpha^2 E_s(A_{r\alpha s}) + \delta E_s(A_{r\delta s})$$

where the coefficients $A_{r\alpha s}$ and $A_{r\delta s}$ are functions of $\bar{X}$ and the x-values in the sample s. The expressions for $A_{r\alpha s}$ and $A_{r\delta s}$ are given in Appendix 1. Following the above lines, we derived the average MSE's of $\bar{y}_B$, $\bar{y}_T$, $\bar{y}_Q$, $\bar{y}_J$, $\bar{y}_H$ and $\bar{y}_M$ and the expressions for the coefficients of α^2 and δ ($A_{B\alpha s}$, $A_{B\delta s}$, etcetra) are presented in Appendix 1. It is possible to entertain more general models than (1) but the derivations become extremely tedious and the interpretation of the results will be difficult. For instance, if the model $y_i = \alpha + \beta x_i^{m+1} + \gamma x_i^{2(m+1)} + u_i$ with the same error structure as in (1) is used, we will have to consider the coefficients of α^2, β^2, γ^2, $\alpha\beta$, $\alpha\gamma$, $\beta\gamma$ for selected values of m and of δ for selected values of t, in order to compare the average MSE's of the estimators.

The computation of the coefficients $E(A_{r\alpha s})$, $E(A_{r\delta s})$, etcetra, for a given x-population, is done on a high speed computer. Table 1 describes the x-populations selected for the present study. The populations numbered 1-12 are natural populations with N ranging from 10 to 270 and the coefficient of variation (C.V.) of x from 0.17 to 1.03. Five artificial populations (nos. 13-17), generated from lognormal and gamma distributions, are also included. A computer program to draw all the $\binom{N}{n}$ possible samples of a given size n is available, but we adopted the following scheme to save computer time: If $\binom{N}{n} \le 2000$, draw all the $\binom{N}{n}$ samples from a given x-population and compute $A_{r\alpha} = E_s(A_{r\alpha s})$, $A_{r\delta} = E_s(A_{r\delta s}) + (\frac{\Sigma}{1}^{N} x_i^t/N^2)$, etc.; otherwise draw 2000 independent simple random samples each of size n when n<20, and approximate $A_{r\alpha}$ by $\frac{\Sigma}{1}^{2000} A_{r\alpha s}/2000$ and so on. We selected 1000 and 500 independent samples for N>100, n=20 and N>100, n=50 respectively. For the estimators involving g we consider the case g=n only, as the previous results indicate that g=n is a good choice. Moreover, three t-values in the range 0 to 2 (t=0, 1, 2) are selected as t is likely to lie between 0 and 2 in practice.

The values of the ratios $E_{rT\alpha} = A_{r\alpha}/A_{T\alpha}$, $E_{HM\delta} = A_{H\delta}/A_{M\delta}$ and so on for n=2, 4, 6, 8, 12 and 50 have been computed (Tables will be published elsewhere). The differences in the average MSE's of the estimators decrease as n increases and/or C.V.(x) decreases. Based on these results, we draw the following conclusions: (1) For n>2, $\bar{y}_M$ is preferable to $\bar{y}_H$ since $E_{HM\alpha}$ is substantially greater than 1, $E_{HM\delta}>1$ for t=0, 1 and $E_{HM\delta}$ for t=2 is >0.95 when C.V.(x) is not too large; the gain in efficiency of $\bar{y}_M$ over $\bar{y}_H$ is considerable when t=0. (2) $E_{MT\alpha}$ and $E_{MT\delta}$ are consistently greater than 1 so that $\bar{y}_T$ is more efficient than $\bar{y}_M$; the gain in efficiency generally increases with t when n>2. (3) $E_{JT\delta}>1$ for t=1, 2 and $E_{JT\alpha}>0.97$; for t=0, $E_{JT\delta}$ is close to 1 excepting for population 10 when n<4. The gain in efficiency of $\bar{y}_T$ over $\bar{y}_J$ is substantial for t=2, especially when n<4 and C.V.(x)>0.75. (4) $E_{rT\delta}$ when t=0 and $E_{rT\alpha}$ are significantly larger than 1; for t>1, $E_{rT\delta}<1$ but close to 1 when t=1. The gain in efficiency of $\bar{y}_r$ over $\bar{y}_T$ could be as high as 20% when t = 2 and $\alpha \doteq 0$. However, when α is significantly different from 0 and mnay strata employed, the approximately unbiased or wholly unbiased 'separate' ratio estimators are preferable to $\bar{y}_r$ (Hutchison [3]). (5) $\bar{y}_J$ could lead to small gains in efficiency over $\bar{y}_Q$. Moreover, the absolute bias of $\bar{y}_J$ is generally smaller than that of $\bar{y}_Q$, especially for small N. (6) $E_{BT\delta}<1$ for all t when n=2 or for t=1, 2 when n>2; $E_{BT\delta}$ for t=0 and $E_{BT\alpha}$ are close to 1, excepting the population 10. Moreover, as Beale has pointed out, $\bar{y}_T$ could occasionally take negative values when all sample pairs (y_i, x_i) are positive, whereas $\bar{y}_B$ is always positive. The conclusions (1)-(4) are in agreement with previous analytic results under a gamma distribution for x.

3. ESTIMATION OF $MSE(\bar{y}_r)$: We turn now to the performances of v_1, v_2 and v_3 as estimators of $MSE(\bar{y}_r)$. The results obtained here are equally applicable to the estimators of MSE(r) but only $\bar{X}^{-2}v_2$ and $\bar{X}^{-2}v_3$ are relevant as $\bar{X}$ is usually unknown when estimating ratios.

The average bias of v_i as an estimator of $MSE(\bar{y}_r)$ is given by

$$B(v_i) = \varepsilon E\{v_i - MSE(\bar{y}_r)\} = B_{i\alpha}\alpha^2 + B_{i\delta}\delta^2 \text{ (say)},$$

i = 1, 2, 3, where $B_{i\alpha}$ and $B_{i\delta}$ are given in Appendix 2. The values of $B_{i\alpha}$ and $B_{i\delta}$ for n=2, 4, 6, 8, 12, 20 and 50 have been computed. Based on these results, a major result is that both v_1 and v_2 (t>0) underestimate $MSE(\bar{y}_r)$ whereas v_3 (with g=n) overestimates $MSE(\bar{y}_r)$ for all t; v_2 leads to overestimation when t=0 and $\alpha \doteq 0$ (this conclusion is in agreement with the analytic result under a gamma distribution for x). Other conclusions are: (1) $|B_{2\delta}|$ is smaller than $|B_{3\delta}|$ for t<1 but the difference is small for t=1, especially when n>4; the comparison between $|B_{2\alpha}|$ and $|B_{3\alpha}|$ is not clear cut. $|B_{3\delta}|$ is substantially smaller than $|B_{2\delta}|$ for t=2. (2) v_3 or v_2 is preferable to v_1 with respect to the absolute bias.

The stabilities of the estimators v_i may be judged by comparing their average MSE's. To simplify the algebra, we assume that the errors u_i are independently and normally distributed, and use the measure $\varepsilon E[v_i - \varepsilon MSE(\bar{y}_r)]^2 = \varepsilon v_i^2 - 2(\varepsilon v_i)(\varepsilon MSE\ \bar{y}_r) + (\varepsilon MSE\ \bar{y}_r)^2$ rather than $\varepsilon E[v_i - MSE(\bar{y}_r)]^2$, but the conclusions are not likely to be different from the latter measure. Let $\varepsilon E[v_i - \varepsilon MSE(\bar{y}_r)]^2 = \alpha^4 M_{i\alpha} + \alpha^2\delta M_{i\alpha\delta} + \delta^2 M_{i\delta}$ and $\varepsilon v_i^2 = \alpha^4 E_s(M'_{i\alpha s}) + \alpha^2\delta E_s(M'_{i\alpha\delta s}) + \delta^2 E_s(M'_{i\delta s})$. The formulae for $M'_{i\alpha s}$, $M'_{i\alpha\delta s}$ and $M'_{i\delta s}$ will be published elsewhere. The values of $M_{i\alpha}$, $M_{i\alpha\delta}$ and $M_{i\delta}$ for n=2, 4, 6, 8, 12, 20 and 50 have been computed. On the basis of these results we draw the following conclusions: (1) $M_{1\alpha}$ is

smaller than $M_{2\alpha}$ which in turn is much smaller than $M_{3\alpha}$ and a similar pattern for $M_{1\alpha\delta}$, $M_{2\alpha\delta}$ and $M_{3\alpha\delta}$ for all t, especially for smaller n and for populations with large C.V.(x). (2) $M_{2\delta}$ is smaller than $M_{3\delta}$ for all t but the difference is small for t=0 and 1, when n>20. (3) $M_{1\delta}$ is smaller than $M_{2\delta}$ for t=0 and 1, but much larger for t=2; $M_{1\delta}$ is slightly smaller than $M_{3\delta}$ for t=2. The results on v_1 and v_2 are in agreement with the analytic results under a gamma distribution for x (no analytic result on the average MSE of v_3 with g=n is available).

4. CONCLUDING REMARKS: Our results are immediately applicable to single-stage cluster sampling within strata, provided simple random sampling and 'separate' ratio estimators are used. We simply replace the variates y_i and x_i by the cluster total Y_i and the cluster size M_i respectively (i=1,...,N). Cochran [1, p.256] proposed a model of the form (1) for cluster sampling: $Y_i=\alpha+\beta M_i+u_i$ with $E(u_i|M_i)=0$ and $E(u_i^2|M_i)\propto M_i^{2-g}$, g>0, where g is likely to lie between 0 and 1 which corresponds to our t between 1 and 2. Our results on estimation of $\bar{Y}$ are also applicable to sub-sampling of clusters. If $\bar{y}_i$ denotes an unbiased estimator of the cluster mean $\bar{Y}_i$, the ratio estimators are as before provided y_i and x_i are replaced by $M_i\bar{y}_i$ and M_i respectively. The between-cluster component of the MSE is the same as the MSE in single-stage cluster sampling. Consequently, the comparisons in Section 2 provide guidelines for the choice of a ratio estimator even when the clusters are sub-sampled. The variance estimators v_1, v_2 and v_3 are applicable to sub-sampling provided the clusters are selected with replacement or n/N is negligible, but their properties remain to be investigated.

If one uses the naive estimators ignoring the design effect and the model (1) with a gamma distribution for x is valid, the average MSE's would be independent of the sample design and, consequently, the previous analytic results for simple random sampling remain valid. However, the model (1) may not be realistic for some of the sample designs. For instance, with stratification it may be more reasonable to assume different intercepts α_h and/or different regression coefficients β_h between strata. Similarly, a model of the form (1) with correlated errors u_i is more realistic when systematic sampling is employed.

This work has been supported by a reasearch grant from the National Research Council of Canada.

REFERENCES

[1] Cochran, W. G. (1963). Sampling Techniques (2nd edition), New York: Wiley.

[2] Frankel, M. R. (1971). An empirical investigation of some properties of multivariate statistical estimates from complex samples. Ph.D. Thesis, University of Michigan.

[3] Hutchison, M. C. (1971). "A Monte Carlo comparison of some ratio estimators", Biometrika, 58, 313-21.

[4] Kish, L. (1965). Survey Sampling, New York: Wiley.

[5] Konijn, H. S. (1962). "Regression analysis in sample surveys", J. Amer. Statist. Assoc., 57, 590-607.

[6] McCarthy, P. (1969). "Pseudoreplication: Further evaluation and application of the balanced half-sample technique", National Center for Health Statistics, Washington, D. C., Series 2, no. 31.

[7] Murthy, M. N. (1967). Sampling Theory and Methods, Calcutta Statistical Publishing Society.

[8] Quenouille, M. H. (1959). "Tables of random observations from standard distributions", Biometrika, 46, 178-202.

[9] Rao, J. N. K. (1969). "Ratio and Regression Estimators", in New Developments in Survey Sampling, eds. N. L. Johnson and Harry Smith, Jr., pp.213-34, New York: Wiley.

[10] Rao, Poduri S. R. S. and Rao, J. N. K. (1971). "Small sample results for ratio estimators", Biometrika, 58, 625-30.

[11] Sampford, M. R. (1962). An Introduction to Sampling Theory, Edinburgh and London: Oliver and Boyd.

[12] Sukhatme, P. V. and Sukhatme, B. V. (1970). Sampling Theory of Surveys with Applications, Rome: F. A. O. of the U. N.

[13] Yates, F. (1960). Sampling Methods for Census and Surveys, London, Griffin.

TABLE 1: Description of x-population

Pop. no.	Source	N	x	C.V.(x)
1	[1], p.204	10	eye est. wt. of peaches	0.17
2	[7], p.131	176	length of timber	0.42
3	[13], p.159	43	no. persons in a kraal	0.45
4	[7], p.127	128	no. persons	0.60
5	[12], p.256	89	no. villages in a circle	0.61
6	[7], p.178	108	geographical area	0.69
7	[11], p.61	35	area under crops and grasses	0.71
8	[7], p.228	80	capital and output	0.75
9	[7], p.228	80	no. workers	0.95
10	[4], p.625	270	no. dwellings	0.99
11	[1], p.156	49	size of cities	1.01
12	[1], p.183	34	no. 'placebo' children	1.03
13	[8]	50	lognormal	0.73
14	[8]	100	gamma	0.79
15	[8]	200	gamma	0.82
16	[8]	100	lognormal	0.82
17	[8]	200	lognormal	0.85

APPENDIX 1: Formulae for the coefficients in the average MSE's of ratio estimators.

The average MSE of $\bar{y}_r$, under the model (1), is given by

$$\varepsilon[MSE(\bar{y}_r)] = \frac{\delta(x^t)_p}{N^2} + \alpha^2 E_s(A_{r\alpha s}) + \delta E_s(A_{r\delta s})$$

where $(x^t)_p = \sum_1^N x_i^t$, E_s denotes the expectation over all possible samples of a given size, and the coefficients $A_{r\alpha s}$, $A_{r\delta s}$ are given below. Similar notation is used for the other estimators $\bar{y}_H$, $\bar{y}_M$, $\bar{y}_Q$, $\bar{y}_J$, $\bar{y}_B$ and $\bar{y}_T$.

Classical ratio:

$$N^2 A_{r\alpha s} = (na_{0s} - N)^2$$

$$N^2 A_{r\delta s} = a_{0s}(x^t)_s(a_{0s}-2)$$

where

$$(x^t)_s = \sum_{i\varepsilon s} x_i^t, \quad a_{0s} = (x)_p/(x)_s.$$