Ann. Inst. Statist. Math.
28 (1976), Part A, 461-467

BIAS REDUCTION AND EFFICIENCY OF RECONSTRUCTED RATIO ESTIMATORS FOR A FINITE UNIVERSE

J. C. KOOP

(Received Jul. 16, 1973; revised Jan. 5, 1974)

Summary

The bias of ratio estimators based on a simple random sample of n units drawn from a finite universe of N units, and reconstructed according to, or in ways similar to Quenouille's method, is of order $1/n$ only if $N \geqq n^2$, and of order $1/n^Q$, where $1<Q<2$, if $N<n^2$. Further it is shown that the device of splitting samples, either for bias reduction and/or convenience of variance estimation, except for the special case noted in the paper, yields estimators that are inefficient and can be improved.

1. Background

A method of reducing bias of estimators from order $1/n$ to order $1/n^2$, based on splitting a sample into parts, and subsequently termed "jackknife" by Tukey (see Miller [1]), was proposed by Quenouille [2] in the context of infinite populations. Durbin [3] and Rao [4] applied this method for reducing the bias of ratio estimators, frequently used in sample surveys, and considered the problems of efficiency of such estimators with distributional assumptions (relative to infinite populations) stated in their respective papers.

This paper in concerned less with the problem of bias reduction, and more with the problem of efficiency of ratio estimators that are constructed by Quenouille's method, or in ways similar to his method, for the purpose of estimating a ratio pertaining to a finite universe. Needless to elaborate, the problem as to whether reconstructed ratio estimators can or cannot be improved in efficiency has so far received very little attention for the case of sampling a finite universe.

Let s be a simple random sample of $n \geqq 2$ units drawn without replacement from a universe of N units with variate values $\{x_i, y_i: i=1, 2, \cdots, N\}$. To avoid trivialities it will be assumed that there are at least two different pairs of variate values, (x_a, y_a) and (x_b, y_b), in the

universe such that $y_a/x_a \neq y_b/x_b$. This is certainly a realistic assumption for sample surveys. It is desired to estimate the ratio

$$(1) \qquad R=\sum_1^N y_i \Big/ \sum_1^N x_i \equiv N\bar{Y}/N\bar{X}=\bar{Y}/\bar{X} .$$

The sample s with variate values $\{x_i, y_i: i \in s\}$ is split up at random into g subsamples each having m units. Ratio estimators $\hat{R}_j=\bar{y}_j'/\bar{x}_j'$ $(j=1, 2, \cdots, g)$ are computed by omitting the jth subsample where $\bar{y}_j'=(n\bar{y}-m\bar{y}_j)/(n-m)$ and $\bar{x}_j'=(n\bar{x}-m\bar{x}_j)/(n-m)$, in which $\bar{x}$ and $\bar{y}$ are the usual sample means, and $\bar{x}_j$ and $\bar{y}_j$ are the means of the jth subsample. The usual estimator of R is

$$(2) \qquad \hat{R}=\bar{y}/\bar{x} .$$

For an infinite population Rao proposed the estimator

$$(3) \qquad \hat{R}_1=g\hat{R}-(g-1)\Big(\sum_{j=1}^{g} \hat{R}_j/g\Big) .$$

Durbin's estimator is of the same form as $\hat{R}_1$, but specialized to the case $g=2$.

We note that (3) is not of the form proposed by Quenouille; in the context of ratio estimation such an estimator is

$$(4) \qquad \hat{R}_2=\{n\hat{R}-(n-r)\hat{R}_{n-r}\}/r ,$$

where $\hat{R}_{n-r}$ is the arithmetic mean of the $\binom{n}{n-r}$ possible ratio estimates $\bar{y}_{n-r}/\bar{x}_{n-r}$ computed from s, with each random subsample having $n-r$ units.

Motivated by Quenouille's method, Tin [5] proposed

$$(5) \qquad \hat{R}_3=g\hat{R}/(g-1)-\Big\{\sum_{j=1}^{g} (\bar{y}_j/\bar{x}_j)\Big\}\Big/\{g(g-1)\} ,$$

for estimating R.

2. Bias reduction

We shall first determine the bias of $\hat{R}_1$ using exact expressions for the expected value of ratios given by Koop [6], but reproduced with very brief explanations in the appendix for ready reference. We find

$$(6) \qquad \mathrm{E}(\hat{R}_1)=g\,\mathrm{E}(\hat{R})-\{(g-1)/g\}\sum_{j=1}^{g} \mathrm{E}\{\mathrm{E}(\hat{R}_j|s)\} .$$

Because the jth subsample is a sample from s, it is necessary first to determine the conditional expectation of $\hat{R}_j$ before proceeding to the

determination of its unconditional expectation.

With the use of formula (A2) of the appendix we find

$$(7)\qquad \begin{aligned} \mathrm{E}(\hat{R}) &= (\bar{Y}/\bar{X})[1+\{(S_x^2/\bar{X}^2)-(S_{xy}/\bar{X}\bar{Y})\}\{(N-n)/(nN)\} \\ &\quad +(\mu_{21}/\bar{X}^2\bar{Y})\{(N-n)(N-2n)/n^2(N-1)(N-2)\}] \\ &\quad -\mathrm{E}\{\hat{R}(\bar{x}-\bar{X})^3\}/\bar{X}^3 , \end{aligned}$$

where S_x^2 and S_{xy} are the finite universe variance and covariance values with divisor $N-1$, and

$$\mu_{21}=\sum_1^N (x_i-\bar{X})^2(y_i-\bar{Y})/N .$$

Hence on the basis of (7), we have after simplification

$$(8)\qquad \begin{aligned} \mathrm{E}(\hat{R}_j|s) &= \hat{R}+\frac{m}{n(n-m)}\left(\frac{\bar{y}s_x^2}{\bar{x}^3}-\frac{s_{xy}}{\bar{x}^2}\right) \\ &\quad -\frac{m(n-2m)}{(n-1)(n-2)(n-m)^2}\left(\frac{m_{21}}{\bar{x}^3}\right)-\mathrm{E}\{\hat{R}_j(\bar{x}_j'-\bar{x})^3|s\}/\bar{x}^3 \end{aligned}$$

where

$$s_x^2=\sum_1^n (x_i-\bar{x})^2/(n-1) , \qquad s_{xy}=\sum_1^n (x_i-\bar{x})(y_i-\bar{y})/(n-1)$$

$$\text{and} \qquad m_{21}=\sum_1^n (x_i-\bar{x})^2(y_i-\bar{y})/n .$$

In the light of (7) and (8), and also remembering that $n-m=m(g-1)$, we find

$$(9)\qquad \begin{aligned} \mathrm{E}(\hat{R}_1) &= R-\frac{R}{N}\left(\frac{S_x^2}{\bar{X}^2}-\frac{S_{xy}}{\bar{X}\bar{Y}}\right)+\frac{1}{n}\left\{\frac{\bar{Y}S_x^2}{\bar{X}^3}-\mathrm{E}\left(\frac{\bar{y}s_x^2}{\bar{x}^3}\right)-\frac{S_{xy}}{\bar{X}^2}+\mathrm{E}\left(\frac{s_{xy}}{\bar{x}^2}\right)\right\} \\ &\quad +\frac{1}{n^2}\frac{\mu_{21}}{\bar{X}^3}\frac{(N-n)(N-2n)}{(N-1)(N-2)}+\frac{(n-2m)}{(n-1)(n-2)(n-m)}\mathrm{E}\left(\frac{m_{21}}{\bar{x}^3}\right) \\ &\quad -\mathrm{E}\{\hat{R}(\bar{x}-\bar{X})^3\}/\bar{X}^3+(g-1)\,\mathrm{E}[\mathrm{E}\{\hat{R}_j(\bar{x}_j'-\bar{x})^3|s\}/\bar{x}^3] . \end{aligned}$$

With the results of the appendix it can be shown that $\mathrm{E}\{m_{21}/(n-1)\cdot(n-2)\bar{x}^3\}$ and the last expression in (9) are at most of order $1/n^2$; also the leading term of $\mathrm{E}\{\hat{R}(\bar{x}-\bar{X})^3\}$ is at most of order $1/n^2$. Applying the results of (A6) and (A7) it will be found that the expression with the multiplier $1/n$ is of order $1/n^2$. Hence

$$(10)\qquad \mathrm{E}(\hat{R}_1)=R-(R/N)\{(S_x^2/\bar{X}^2)-(S_{xy}/\bar{X}\bar{Y})\}+O(n^{-2}) .$$

Now if $N<n^2$ we should say that

$$(11)\qquad \mathrm{E}(\hat{R}_1)=R+O(n^{-q}) ,$$

where $1<Q<2$. It is only when $N \geqq n^2$ that

$$\mathrm{E}(\hat{R}_1)=R+O(n^{-2}) . \tag{12}$$

If we proceed to determine the expected values of $\hat{R}_2$ and $\hat{R}_3$, then the first three terms on the right-hand side of equation (10), plus slightly different terms, but all of order $1/n^2$, will be obtained. These derivations are omitted to save space. Hence on the question of bias the same conclusions as in the foregoing account apply to these estimators.

Finally, in this connection it must be mentioned that Jones [7] was already aware that for finite populations terms in $1/N$ in the expression for bias were not eliminated by Quenouille's device. However, he arrived at this conclusion not with an approach leading to the results at (10) and (11), but in the context of what he termed "replicated sampling."

3. Improvement of efficiency

The reconstructed ratio estimators $\hat{R}_1$ and $\hat{R}_3$ considered in Section 2 can be improved by determining their respective conditional expectations given the sample s.

By the theorem formalized by Madow [8] we find

$$\mathrm{V}(\hat{R}_1)=\mathrm{V}\{\mathrm{E}(\hat{R}_1|s)\}+\mathrm{E}\{\mathrm{V}(\hat{R}_1|s)\} . \tag{13}$$

Let us denote $\mathrm{E}(\hat{R}_1|s)$ by R_1'. By definition $\mathrm{E}(\hat{R}_1)=\mathrm{E}(R_1')$ so that $\{\mathrm{E}(\hat{R}_1)-R\}^2=\{\mathrm{E}(R_1')-R\}^2$, and hence with (13)

$$\text{M.S.E.}(\hat{R}_1)=\text{M.S.E.}(R_1')+\mathrm{E}\{\mathrm{V}(\hat{R}_1|s)\} . \tag{14}$$

Because of the assumption that there are at least two units with different pairs of variate values, $\mathrm{E}\{\mathrm{V}(\hat{R}_1|s)\}>0$ for all n or $mg \geqq 2$, except when $g=n$, in which case $\mathrm{E}(\hat{R}_1|s)=\hat{R}_1$ and $\mathrm{E}\{\mathrm{V}(\hat{R}_1|s)\}=0$ so that $\text{M.S.E.}(\hat{R}_1)=\text{M.S.E.}\{\mathrm{E}(\hat{R}_1|s)\}$; in this connection it is interesting to note that Rao [4] arrived at an optimum choice of $g=n$ through the twin assumptions of normality for the x-variate and a linear relationship between y and x. Hence except for this case

$$\text{M.S.E.}(\hat{R}_1)>\text{M.S.E.}(R_1') . \tag{15}$$

The determination of the exact expression for the improved estimator R_1' using the result at (8) leads to some interesting results. We find

$$\mathrm{E}(\hat{R}_1|s)=g\hat{R}-(g-1)\,\mathrm{E}(\hat{R}_j|s) ,$$

or

$$(16)\qquad R_1'=\hat{R}\left[1-\{(s_x^2/\bar{x}^2)-(s_{xy}/\bar{x}\bar{y})\}/n+\frac{(n-2m)}{(n-1)(n-2)(n-m)}\left(\frac{m_{21}}{\bar{x}^2\bar{y}}\right)\right]$$
$$+\{(g-1)/\bar{x}^3\}\,\mathrm{E}\,\{\hat{R}_j(\bar{x}_j'-\bar{x})^3\,|\,s\}\ ,$$

in which for computational purposes

$$(17)\qquad \mathrm{E}\,\{\hat{R}_j(\bar{x}_j'-\bar{x})^3\,|\,s\}=\sum{}'\,\hat{R}_j(\bar{x}_j'-\bar{x})^3\Big/\binom{n}{n-m}$$

where the summation $\sum'$ is over all the $\binom{n}{n-m}$ possible subsamples.

We note that the

$$(18)\qquad \{\text{last term in (16)}\}<(g-1)\,|\hat{R}_j|_{\max}\{|(\bar{x}_j'-\bar{x})/\bar{x}|_{\max}\}^3\ .$$

Hence if its upper bound is very small compared to the sum of the terms of order $1/n$ and $1/n^2$ in (16), and this is likely to be so if $|(\bar{x}_j'-\bar{x})/\bar{x}|_{\max}\ll 1$, then this term may be neglected in the computation of the improved estimator R_1', which is just as easy to compute as $\hat{R}_1$. Further under these conditions it is quite likely that (15) may still hold.

The expression for $\mathrm{E}\,(\hat{R}_1\,|\,s)$ given by (16) is not unique. If we use the identity for $\mathrm{E}\,(Y/X)$ given by (A1) in the appendix we find

$$(19)\qquad \mathrm{E}\,(\hat{R}_j\,|\,s)=\hat{R}\left\{1-\frac{m}{n(n-m)}(s_{xy}/\bar{x}\bar{y})\right\}+(1/\bar{x}^2)\,\mathrm{E}\,\{\hat{R}_j(\bar{x}_j'-\bar{x})^2\,|\,s\}\ ,$$

so that, say

$$(20)\qquad R_1''=\mathrm{E}\,(\hat{R}_1\,|\,s)=\hat{R}\left\{1+\frac{1}{n}(s_{xy}/\bar{x}\bar{y})\right\}-\{(g-1)/\bar{x}^2\}\,\mathrm{E}\,\{\hat{R}_j(\bar{x}_j'-\bar{x})^2\,|\,s\}\ .$$

Certainly R_1'' is simpler than R_1', but not better because $\mathrm{M.S.E.}\,(R_1'')=\mathrm{M.S.E.}\,(R_1')$. Actually the basic expression for $\mathrm{E}\,(\hat{R}_1\,|\,s)$ is $g\hat{R}-(g-1)\cdot\sum'\hat{R}_j\Big/\binom{n}{m}$, useful for computational purposes when $\binom{n}{m}$ is not excessively large, but uninteresting as compared to the analytic insight provided by R_1' and R_1''.

The analogues of R_1' and R_1'' corresponding to $\hat{R}_3$ have been derived, but are omitted to save space.

The estimator $\hat{R}_2$ cannot be improved because $\mathrm{E}\,(\hat{R}_2\,|\,s)=\hat{R}_2$ leading to the result $\mathrm{E}\,\{\mathrm{V}\,(\hat{R}_2\,|\,s)\}=0$, and $\mathrm{V}\,(\hat{R}_2)=\mathrm{V}\,\{\mathrm{E}\,(\hat{R}_2\,|\,s)\}$. Incidentally, this result confirms Quenouille's conjecture (contained in the third paragraph of Section 3 of his paper) that with the averaging of estimates from all possible subsamples "it appears likely that little, if any, loss of efficiency will result." However, if instead of $\hat{R}_{n-r}$, an estimate based on a smaller number of subsamples is used (e.g., see Sukhatme and Suk-

hatme [9], p. 162), then the resulting $\hat{R}_2$ can be improved.

In general if any kind of complex probability sample s, drawn from a finite universe, is split up at random into g parts $s_1, s_2, \cdots, s_g$ and an estimator e is constructed as a function of estimates based on the variate values of $s, s_1, \cdots, s_g$, for the purpose of estimating some universe value, then, arguing as in the foregoing account,

$$\text{M.S.E.}\,(e) \geqq \text{M.S.E.}\,\{\text{E}\,(e\,|\,s)\}\,, \tag{21}$$

equality holding if and only if conditions (to be determined in the context of the specific problem) are such that $\text{E}\,\{\text{V}\,(e\,|\,s)\}=0$.

One implication of the foregoing generalization is that e may be an estimator of variance. Thus, the device of splitting a sample either for bias reduction and/or convenience of variance estimation, except for special cases, yields estimators that are inefficient. But, as demonstrated in the paper, the efficiencies of such estimators can be improved by determining their conditional expectations.

Appendix

Let X and Y be pairs of random variables with X taking nonzero values. We assume that $\text{E}\,(X)\neq 0$, $\text{E}\,(Y)\neq 0$ and that all moments exist. Expressions for $\text{E}\,(Y/X)$ are desired in terms of the central moments. Koop [6] has derived the following identities:

$$\begin{aligned}\text{E}\,(Y/X)=&\{\text{E}\,(Y)/\text{E}\,(X)\}[1-\{\text{Cov}\,(X, Y)/\text{E}\,(X)\,\text{E}\,(Y)\}]\\&+\text{E}\,\{(Y/X)(X-\text{E}\,(X))^2\}/\{\text{E}^2\,(X)\}\,,\end{aligned} \tag{A1}$$

$$\begin{aligned}\text{E}\,(Y/X)=&\{\text{E}\,(Y)/\text{E}\,(X)\}[1+\{\text{V}\,(X)/\text{E}^2\,(X)\}\\&-\{\text{Cov}\,(X, Y)/\text{E}\,(X)\,\text{E}\,(Y)\}\\&+\{\text{E}\,(X-\text{E}\,(X))^2(Y-\text{E}\,(Y))/\text{E}^2(X)\,\text{E}(Y)\}]\\&-\text{E}\,\{(Y/X)(X-\text{E}\,(X))^3\}/\{\text{E}^3\,(X)\}\,.\end{aligned} \tag{A2}$$

These identities are useful in a variety of problems and are members of a class of identities.

With the use of (A1) we find

$$\begin{aligned}\text{E}\,(\bar{y}s_x^2/\bar{x}^3)=&\{\text{E}\,(\bar{y}s_x^2)/\text{E}\,(\bar{x}^3)\}-\{\text{Cov}\,(\bar{y}s_x^2,\,\bar{x}^3)/\text{E}^2\,(\bar{x}^3)\}\\&+\text{E}\,\{(\bar{y}s_x^2/\bar{x}^3)(\bar{x}^3-\text{E}\,(\bar{x}^3))^2\}/\{\text{E}^2\,(\bar{x}^3)\}\,.\end{aligned} \tag{A3}$$

It can be shown that the second and third terms of (A3) are at most of order $1/n$. Using results that are already known we find

$$\text{E}\,(\bar{y}s_x^2)=\{\mu_{21}N(N-n)\}/\{n(N-1)(N-2)\}+\bar{Y}S_x^2\,, \tag{A4}$$

and

$$\text{(A5)} \qquad \mathrm{E}(\bar{x}^3) = \{\mu_{30}(N-n)(N-2n)\}/\{n^2(N-1)(N-2)\} + 3\bar{X}\{S_x^2(N-n)/(nN)\} + \bar{X}^3 ,$$

where μ_{30} is the third central moment of X. Hence from (A3), (A4) and (A5)

$$\text{(A6)} \qquad \mathrm{E}(\bar{y}s_x^2/\bar{x}^3) = (\bar{Y}S_x^2)/(\bar{X}^3) + O(n^{-1}) .$$

Similarly it can be shown that

$$\text{(A7)} \qquad \mathrm{E}(s_{xy}/\bar{x}^2) = (S_{xy}/\bar{X}^2) + O(n^{-1}) .$$

RESEARCH TRIANGLE INSTITUTE, NORTH CAROLINA

REFERENCES

[1] Miller, R. G. (1964). A trustworthy jackknife, *Ann. Math. Statist.*, **35**, 1594-1605.

[2] Quenouille, M. H. (1956). Notes on bias in estimation, *Biometrika*, **43**, 353-360.

[3] Durbin, J. (1959). A note on the application of Quenouille's method of bias reduction to the estimation of ratios, *Biometrika*, **46**, 477-480.

[4] Rao, J. N. K. (1965). A note on the estimation of ratios by Quenouille's method, *Biometrika*, **52**, 647-649.

[5] Tin, M. (1965). Comparison of some ratio estimators, *J. Amer. Statist. Ass.*, **60**, 294-307.

[6] Koop, J. C. (1972). On the derivation of expected value and variance of ratios without the use of infinite series expansions, *Metrika*, **19**, 156-170.

[7] Jones, H. L. (1962). Some basic problems in replicated sampling, *Proceedings of 1962 Middle Atlantic Conference*, American Society for Quality Control, Washington.

[8] Madow, W. G. (1949). On the theory of systematic sampling II, *Ann. Math. Statist.*, **20**, 333-354.

[9] Sukhatme, P. V. and Sukhatme, B. V. (1970). *Sampling Theory of Surveys with Applications*, Iowa State University Press, Ames.