A NOTE ON THE SMALL SAMPLE THEORY OF THE RATIO ESTIMATOR IN CERTAIN SPECIFIED POPULATIONS*

R.P. CHAKRABARTY
University of Georgia

(*Received in January, 1971*)

It is well known that under certain conditions the ratio estimator is more efficient than the sample mean in large samples but little is known about the efficiency of the ratio estimator in small samples. In this note the exact bias and variance of the ratio estimator are given assuming a linear regression of y on x where x has a gamma distribution. It is shown that the ratio estimator is generally more efficient than the sample mean in small samples. The variance estimator of the ratio estimator is shown to be generally more stable than the variance estimator of the sample mean. Results are exact for any sample size.

1. INTRODUCTION

In sample surveys ratio estimators are often used for estimating the population mean $\bar{Y}$ of a characteristic of interest 'y' or the population ratio $R=\bar{Y}/\bar{X}$ utilizing an auxiliary variate 'x' that is positively correlated with 'y'. It is well known that the ratio method increases the precision of estimators in large samples if $\rho>C_x/(2C_y)$ where ρ is the coefficient of correlation between y, x, C_y and C_x are coefficients of variation of y and x respectively. However, not much is known about the exact efficiency of ratio estimator in small samples (Cochran, 1963 p. 157). Therefore, in this paper, we investigate the exact efficiency of the ratio estimator assuming a model. The stability of the variance estimator of the ratio estimator is also compared with the stability of the variance estimator of the sample mean.

We confine ourselves to simple random sampling and assume the population size is infinite, to simplify the discussion. From a simple random sample of n pairs (y_i, x_i) we have the unbiased estimator of $\bar{Y}$, the population mean of y, as

$$\bar{y}=\sum_{i=1}^{n} y_i/n. \qquad \ldots(1.1)$$

*Revised version of a paper presented at the annual meeting of the Indian Soeiety of Agricultural Statistics held at Madras in December, 1970.

The unbiased estimator of $V(\bar{y})$, the variance of $\bar{y}$, is given by

$$v_0 = s_y^2/n \qquad \ldots (1.2)$$

where s_y^2 is the sample mean square of y. The ratio estimator of $\bar{Y}$ is

$$\bar{y}_r = (\bar{y}/\bar{x})\,\bar{X} = r\,\bar{X} \qquad \ldots (1.3)$$

where $\bar{x}$ is the sample mean and $\bar{X}$ is the known population mean of x and

$$r = \bar{y}/\bar{x} \qquad \ldots (1.4)$$

is the ratio estimator of the ratio $R = \bar{Y}/\bar{X}$. As an estimator of $V(\bar{y}_r)$, the variance of $\bar{y}_r$, it is customary to take

$$\begin{aligned} v_1 &= (s_y^2 - 2rs_{yx} + r^2 s_x^2)/n \\ &= \bar{X}^2 v(r) \end{aligned} \qquad \ldots (1.5)$$

where s_x^2 is the sample mean square of x and s_{yx} is the sample covariance. It is known that v_1 is consistent but biased; bias is of order $1/n$. It may be noted that the unbiased estimator of the population ratio $R = \bar{Y}/\bar{X}$ is $\bar{y}/\bar{X}$ and its variance estimator is $v_0/\bar{X}^2$ (assuming the population mean $\bar{X}$ is known). Therefore, without loss of generality, we shall discuss in the sequel the efficiencies of estimators, $\bar{y}$ and $\bar{y}_r$, of the population mean $\bar{Y}$, and stabilities of their variance estimators v_0 and v_1 respectively.

The stability of a variance estimator may be judged by its coefficient of variation. Rao and Beegle (1967) have made a Monte Carlo study of the small-sample properties of v_0 and v_1. Assuming a linear regression of y on x, with x normal, they have demonstrated that (1) the coefficients of variation of v_0 and v_1 are of the same order when the regression is through the origin and C_x is small and (2) the coefficient of variation of v_1 is considerably larger than that of v_0 when the regression does not pass through the origin and C_x is large. Recently Rao (1968) has investigated the performances of v_0 and v_1 using several sets of live data which represent a wide variety of populations. His emperical results indicate that for small samples stability of v_1 compare favorably with that of v_0; in fact considerably better for most of the populations.

2. THE EXACT THEORY

We assume the following model for the comparison of estimators:

$$\begin{aligned} &y_i = \alpha + \beta x_i + u_i \;;\; \beta > 0 \\ &E(u_i/x_i) = 0,\; E(u_i,\, u_j/x_i,\, x_j) = 0 \\ &V(u_i/x_i) = n\delta \;(\delta \text{ is a constant of order } n^{-1}) \end{aligned} \qquad \ldots (1)$$

where the variates x_i/n have the gamma distribution with parameter h so that $\bar{x}=\Sigma x_i/n$ has the gamma distribution with parameter $m=nh$. To compare the stabilities of variance estimators we further assume that u_i's are normally and independently distributed with mean zero and variance $n\delta$. This model was used by Durbin (1959) and Rao and Webster (1966) to investigate the bias in estimation of ratios. This model is quite suitable to describe many situations met in practice. An example would be the estimation of production rate of a manufacturing process where varing amounts (random variable y) are produced at varying time intervals (random yariable x); the latter usually follows a gamma distribution. It may be noted that all our results under this model are exact for any sample size, n.

2.1. *The exact efficiency of the ratio estimator.*

Under the model (I) we have

$$Y=\alpha+\beta m \qquad \ldots(2.1)$$

and

$$\bar{y}_r=\beta m+\frac{(\alpha+\bar{u})m}{\bar{x}} \qquad \ldots(2.2)$$

Consequently the bias of $\bar{y}_r$ is

$$\text{Bias } (\bar{y}_r)=\alpha/(m-1) \qquad \ldots(2.3)$$

The variances of $\bar{y}_r$ and $\bar{y}$ are obtained as

$$V(\bar{y}_r)=\frac{\alpha^2m^2}{(m-1)^2(m-2)}+\frac{\delta m^2}{(m-1)(m-2)} \qquad \ldots(2.4)$$

which exists for $m > 2$, and

$$V(\bar{y})=\delta+\beta^2m \qquad \ldots(2.5)$$

respectively. The exact efficiency of $\bar{y}_r$ relative to that of $\bar{y}$ is given by

$$E=\frac{V(\bar{y})}{MSE(\bar{y})_r} \qquad \ldots(2.6)$$

Now, we note that in terms of the model (1)

$$\begin{aligned}\alpha&=\overline{Y}\,[(K-\rho)/K],\\ \beta&=\overline{Y}\,[\rho/(Km)],\\ \delta&=\overline{Y^2}\,[(1-\rho^2)/(K^2m)]\end{aligned} \qquad \ldots(2{\cdot}7)$$

where $K=C_x/C_y$.

Therefore using (2·3) through (2·5) and substituting the values α, β and δ given by (2·7) E can be expressed as a function of K, ρ and m. It may be noted that $K=C_x/C_y$ and the coefficient of variation of $\bar{x}$, $C_{\bar{x}}=m^{-1/2}$ and consequently E is independent of the units of measurement of the variables x and y as it should be. The numerical values of E as percentages are presented in Table 1 for selected

values of K, ρ and $m=nh > 2$. The results of Table 1 may be summarized as follows :

The efficiency or $\bar{y}_r$ increases as ρ increases for given K and m, and for given ρ and K it increases as $m=nh$ increases. The ratio estimator $\bar{y}_r$ is more efficient than the unbiased estimator $\bar{y}$ for the following values of $\rho(>K/2)$ and m: (*a*) $\rho > .8$, $m > 8$ (*b*) $\rho \geqslant .5$, $m \geqslant 16$, (*c*) $\rho \geqslant .4$, $m > 20$.

Noting that in our model $C_x=h^{-1/2}$, $C_{\bar{x}}=m^{-1/2}$ and $n \leqslant m$ for $h \geqslant 1$ we may conclude that for $\rho \geqslant .4$ and $K < 2\rho$, the ratio estimator $\bar{y}_r$ is efficient in small samples if $h \geqslant 1$.

Finally, it is of interest to note that the large-sample theory (*viz.* $\bar{y}_r$ is superior to $\bar{y}$ if $\rho > K/2$) is generally applicable in this case if $m=nh > 32$.

Now, we consider the case of the linear regression through the origin (*i.e.*, $\alpha=0$ in model 1). Putting $\alpha=0$ in (2·3) we get as a check the well-known result that $\bar{y}_r$ is unbiased for $\bar{Y}$. We note that in this case $K=\rho$.

The variances of $\bar{y}_r$ and $\bar{y}$ are given by

$$V(\bar{y}_r)=\frac{\delta m^2}{(m-1)(m-2)} \qquad \ldots(2\cdot8)$$

and

$$V(\bar{y}) = \frac{\delta}{(1-\rho^2)} \qquad \ldots(2\cdot9)$$

respectively. Therefore the exact efficiency of $\bar{y}_r$ relative to that of $\bar{y}$ is given by

$$E_0=\frac{(m-1)(m-2)}{m^2(1-\rho^2)} \; ; \; \rho > 0 \qquad \ldots(2\cdot10)$$

Clearly E_0 increases as ρ increases for mixed $m(>2)$. For a given ρ the value of m for which $E_0=1$ is

$$m=\frac{3+(9-8\rho^2)^{1/2}}{2\rho^2} \; ; \; \rho > 0 \qquad \ldots(2\cdot11)$$

The values of m have been calculated using (2·11) for different values of ρ and are presented as integers in Table 2 so that $E_0 \geqslant 1$.

We find from Table 2 that in the case of the linear regression through the origin the ratio estimator $\bar{y}_r$ is superior to $\bar{y}$ for $\rho \geqslant .4$ in small samples ($m \leqslant 18$; $n \leqslant 18$ if $h \geqslant 1$). For low correlation, $\rho < .4$, $\bar{x}_r$ is efficient if $m > 32$. Further, the comparison of these results with those given in Table 1 shows that the sample size needed for the ratio estimator to be efficient in the case of the regression through the origin is smaller than in the case of the general regression model *i*.

TABLE 1

The exact efficiency of $\overline{y}_r$ for selected values of K, $\rho(\rho>K/2)$ and m

K	ρ	m			
		8	*16*	*20*	*32*
·25	·4	76	95	99	105
	5	79	100	104	111
	·7	86	111	117	125
	·9	91	123	131	142
·50	·4	77	96	100	107
	·5	87	109	114	121
	·7	117	148	154	164
	·9	168	222	234	252
1·00	·6	71	100	105	112
	·7	105	134	140	150
	·9	324	408	425	453
1·50	·8	67	90	95	103
	·9	103	138	146	159

TABLE 2

The values of ρ and m for which $E_0 \geqslant 1$

ρ	·1	·2	·3	·4	·5	·6	·7	·8	·9
m	300	75	33	18	12	8	6	4	3

2.2. *The exact stabilities of the variance estimators v_0 and v_1*

The formulae for the variance estimators v_0 and v_1 were given by (1·2) and 1·5) respectively. It can be shown that in terms of model I they are

$$v_0=(s_u^2+\beta^2 s_x^2+2\beta s_{ux})/n \qquad (2\cdot12)$$

and

$$v_1=\left[s_u^2-2(\alpha+\bar{u})\frac{s_{ux}}{\bar{x}}+(\alpha+\bar{u})^2\ \frac{s_x^2}{\bar{x}^2}\right]/n \qquad \ldots(2\cdot13)$$

Now, we have the following expectations

$$E\left(\frac{s_{ux}}{\bar{x}}\right)=E(s_{ux})=0,\ E(s_x^2)=mn$$

and

$$E\left(\frac{s_x^2}{\bar{x}}\right)=\frac{1}{n-1}\left\{n^3E\left[\frac{z_i^2}{(\Sigma z_i)}\right]-mn\right\}$$

where $z_i=x_i/n$. From Rao and Webster (1966) we have

$$E\left(\frac{z_i^2}{\Sigma z_i}\right)=\frac{h(h+1)}{(m+1)}\ ;$$

hence

$$E\left(\frac{s_x^2}{\bar{x}}\right)=\frac{mn}{m+1}\ .$$

Similarly we obtain

$$E\left(\frac{s_x^2}{\bar{x}^2}\right)=\frac{n}{m+1}\ .$$

Using these expected values we obtain the well-known result that v_0 is unbiased for $V(\bar{y})$. The expected value of v_1 is obtained as

$$E(v_1)=\frac{\alpha^2}{m+1}+\frac{(m+2)\delta}{m+1} \qquad \ldots(2\cdot14)$$

Consequently the bias of v_1 as an estimator of $V(\bar{y}_r)$, given by (2·4) is

$$\text{Bias}\ (v_1)=-\frac{(5m^2-5m+2)\alpha^2}{(m^2-1)(m-1)(m-2)}-\frac{2(m^2+2m-2)\delta}{(m^2-1)(m-2)} \qquad \ldots(2.15)$$

We note that for finding the variances of v_0 and v_1 expected values of some functions of sample moments are needed. Following the method of Rao and Webster (1966) Chakrabarty (1968) has evaluated these expectations. The details of evaluating these expectations, which involve some tedious algebra, are omitted and only the final results are given here. The variances of v_0 aud v_1 are obtained as

$$V(v_0)=\frac{2\delta^2}{(n-1)}+\frac{4\beta^2\delta m}{(n-1)}+\beta^4[\theta m(m+1)(m+2)(m+3)-m^2] \ldots(2.16)$$

and

$$V(v_1)=\delta^2\left[3\theta\ +\frac{(n+1)(m+3)}{(n-1)(m+1)}-\frac{(m+2)^2}{(m+1)^2}\right]+\alpha^4\left[\theta\ -\frac{1}{(m+1)^2}\right]$$
$$+2\alpha^2\delta\left[3\theta\ +\frac{(2m-n+3)}{(n-1)(m+1)^2}\right] \ldots(2\cdot17)$$

respectively, where

$$\theta=\frac{[(n+1)(m+6)-12]}{(n-1)(m+3)(m+2)(m+1)} \qquad \ldots(2.18)$$

The relative variance of v_0 is

$$CV^2(v_0)=\frac{V(v_0)}{[V(\bar{y})]^2}=T_1 \qquad \text{(say)} \qquad \ldots(2.19)$$

where $V(v_0)$ and $V(\bar{y})$ are given by (2·16) and (2·5) respectively. The relative variance of v_1 is

$$CV^2(v_1)=\frac{V(v_1)+[\text{Bias }(v_1)]^2}{[V(\bar{y}_r)]^2}=T_2 \quad \text{(say)} \qquad \ldots(2\cdot20)$$

where $V(v_1)$, Bias (v_1) and $V(\bar{y}_r)$ are given by (2·17), (2·15) and (2·4) respectively. Finally, the stability of v_1 relative to that of v_0 is given by

$$S=T_1/T_2. \qquad \ldots(2\cdot21)$$

We note that substituting the values of α, β and δ given by (2·7) in (2·21) S can be expressed explicitly as a function of K, ρ, m and n. However, the resulting expression is rather complicated for analytical investigation of the behavior of S. The comparison of Stability of v_1 with that of v_0 is of interest when $\bar{y}_r$ is more efficient than $\bar{y}$. Therefore, we have computed the values S for selected values of m, n, K and ρ for which the efficiency of $\bar{y}_r$ is greater than or equal to that of $\bar{y}$. The results are given as percentages in Table 3. We find from Table 3 that—

TABLE 3

The Value of S for Selected Values of m. n, ρ and K

m	n	$K=\cdot25$				$K=\cdot50$				$K=1\cdot00$			$K=1\cdot50$	
		$\rho=\cdot4$	$\rho=\cdot5$	$\rho=\cdot7$	$\rho=\cdot9$	$\rho=\cdot4$	$\rho=\cdot5$	$\rho=\cdot7$	$\rho=\cdot9$	$\rho=\cdot6$	$\rho=\cdot7$	$\rho=\cdot9$	$\rho=\cdot8$	$\rho=9$
8	4							188	254		185	258		262
8	8							182	250		172	310		240
16	4		130	142	189		131	142	176	134	141	172		186
16	8		124	140	196		129	151	193	127	146	216		195
16	16		115	137	204		127	165	214	118	152	287		204
20	4	112	123	133	170	122	124	133	159	126	132	156		168
20	10	117	119	133	184	118	125	145	183	121	139	210		183
20	20	109	112	132	195	112	125	161	206	115	148	283		196
32	4	113	114	120	142	114	115	120	135	116	120	133	130	141
32	8	112	113	121	152	113	116	126	148	115	124	157	133	151
32	16	110	112	123	165	112	118	138	168	113	131	201	136	165
32	32	107	109	126	181	109	122	157	194	110	143	278	140	182

(1) The variance estimator v_1 is more stable than v_0 ; the gain in stability is considerable for $\rho \geqslant .5$

(2) For fixed K, m, and n the stability of v_1 increases as ρ increases.

(3) For the special case where x has the exponential distribution with mean $m=n$ (*i.e.*, $h=1$) in model I, S decreases as $m=n$ increases.

It may be noted that Rao's (1968) emperical results for small-sample stabilities of v_0 and v_1 obtained from several sets of live data generally agree with the exact results obtained here namely v_1 is more stable than v_0 and the gain in stability in considerable for $\rho \geqslant .5$.

Turning to the case of the regression through the origin (*i.e.* $\alpha=0$ in model I) we get the relative variance of v_0 using (2·7), (2·9), (2·16) and (2·19) as

$$CV^2(v_0) = \frac{2(1-\rho^4)}{(n-1)} + \frac{\rho^4}{m}\left[\theta\,(m+1)(m+2)(m+3) - m\right] \quad \ldots(2\cdot22)$$

$$= T_3 \qquad \text{(say)}$$

From (2·7), (2·15), (2·17) and (2·20) we get the relative variance of v_1 as

$$CV^2(v_1) = \frac{(m-1)^2(m-2)^2}{m^4}\left[3\theta + \frac{(n+1)(m+3)}{(n-1)(m+1)} - \frac{(m+2)^2}{(m+1)^2} + \frac{4(m^2+2m-2)^2}{(m-1)^2(m-2)^2}\right]$$

$$= T_4 \text{ (say).} \qquad \ldots(2\cdot23)$$

Consequently the stability of v_1 relative to that of v_0 is given by

$$S_0 = \frac{T_3}{T_4} \qquad \ldots(2\cdot24)$$

which is a function of ρ, $m = [CV^2(X)]^{-1}$ and n. The numerical evalution S_0 shows that the results for the relative stability of v_1 are similar to those obtained in the case of the general regression model I (Table 3), and hence the numerical values of S_0 are not given here. In this case also, the variance estimator v_1 is more stable than v_0, its stability increases as ρ increases for fixed m and n. For the special case where x has the exponential distribution S_0 decreases as $m=n$ increases. Further, the comparison of S values of and S_0 reveals that the variance estimator v_1 is slightly more stable in the case of regression through the origin than in the general regression model I. For example, when $m=n=8$ and $\rho=.9$, $S_0=325$ where as S ranges from 240 to 310 depending on the values of K and when $m=20$, $n=10$ and $\rho=.7$, $S_0 = 152$ where as S ranges from 133 to 145 depending on the values of K (Table 3). The implication of these findings is that the ratio method of estimation would frequently lead to an improvement in the accuracy of estimators and variance estimators even in small samples if the auxiliary variable x follows a gamma distribution.

ACKNOWLEDGEMENT

I am grateful to Professor J.N.K. Rao for suggesting the problem and to the referee for his valuable comments.

REFERENCES

Chakrabarty, R.P., (1968) : *Contributions to the Theory of Ratio-Type Estimators.* Ph. D. Thesis, Texas A and M University.

Cochran, W.G., (1963) : *Sampling Techniques.* John Wiley and Sons, New York.

Durbin, J., (1959) : A note on the application of Quenouille's meihod of bias reduction to the estimation of ratios. *Biometrika*, 46, 477-80.

Rao J.N.K., (1968) : Ratio and regression estimators. *New Developments in Survey Sampling*, edited by N.L. Johnson and H. Smith. Wiley interscience, New York.

Rao, J.N.K. and Beegle, L.D. (1967) : A Monte-Carlo study of some ratio estimators. *Sankhya* Series B, 29, 47-56.

Rao, J.N.K. and Webster, J.T., (1966) : On two methods of Bias reduction in estimation of ratios. *Biometrika*, 53, 571-77.