CHANGE OF SCALE IN DIFFERENCE ESTIMATION IN SAMPLING

BY

T. SRIVENKATARAMANA* AND K.P. SRINATH**

(Received ; March, 1975)

SUMMARY

A scale transformation of the variates involved in the difference estimator is suggested. This leads to an estimator more precise than the ratio (product) estimator when the intercept of population regression line with the vertical axis is moderately large in absolute value. This method is not very restrictive from the point of view of survey practice. A numerical example is included for illustration. A multivariate extension is also considered.

INTRODUCTION

It is an usual practice in sample surveys to utilize information on an auxiliary variate x related to the variate of interest y for improving the precision of estimators. Such information can be used either in the sampling scheme or in the estimation procedure or in both. Stratification and ratio methods are two good examples. Mohanty and Das [3] have given a method of reducing the bias of the estimator in the ratio method of estimation when $Y=\alpha+\beta X$ is the line of regression of y on x in the population and α and β are positive. This is done by obtaining approximate estimates a, b of α and β respectively through some pilot enquiry and then changing the origin of measurement of x from 0 to $-a/b$. This clearly necessitates the consideration of the sticky problem of the sampling variation of the ratio-a/b. As an alternative a method of difference estimation after simple changes in scales of measurement of x and y is proposed in this paper. Here it is not necessary that α and β are positive. It is necessary only that their signs be known and that $|\alpha|$ be moderately

*Department of Mathematics, University of Windsor, Windsor, Ontario Canada N9B 3 P 4. On study leave from Bangalore University, India.

**Business Survey Methods Division, Statistics Canada, Ottawa, Canada.

large. These could be judged by the accumulated experience and/or by drawing a scatter diagram for at least a part of the current data.

Change of Scale

Consider a finite population $U=(U_1, U_2, \ldots, U_N)$ of size N. Let the study variate y and the auxiliary variate x take on real values (Y_i, X_i) on the unit U_i, $i=1, 2, \ldots, N$. Assume that $Y_i \geqslant 0$, $X_i \geqslant 0$ and that the population mean $\bar{X}$ of x is known and is positive. Let Y_k be some Y_i, $i=1, 2, \ldots, N$. Assume also that Y_k is known either by design or otherwise and that it is positive. The other restrictions on Y_k are discussed later. Make the transformations

$$W_i=\frac{X_i}{\bar{X}} \text{ and } Z_i=\frac{Y_i}{Y_k},\ i=1, 2, \ldots, N.$$

Estimation of the Population Mean

A simple random sample of $n \leqslant N$ units without replacement is drawn from the population. Let $\bar{y}, \bar{x}$ denote the means of y and x respectively in this sample and let $\bar{w}=\frac{\bar{x}}{\bar{X}}$, $\bar{z}=\frac{\bar{y}}{Y_k}$. Then an estimator of the population mean $\bar{Z}$ of z is

$$\bar{z}_d=\bar{z}+\delta\,(\bar{W}-\bar{w})$$

where $\delta=\pm 1$ according as $\beta \gtrless 0$ and $\bar{W}$ is the population mean of w. In fact $\bar{W}=1$. Next, an estimator of $\bar{Y}$ is

$$\bar{y}_d=Y_k\bar{z}_d. \qquad \ldots(3.1)$$

To evaluate the bias and variance of $\bar{y}_d$ while estimating $\bar{Y}$, rewrite $\bar{y}_d$ in terms of $\bar{y}$ and $\bar{x}$ and then write

$\bar{y}=\bar{Y}\,(1+e_1)$, $\bar{x}=\bar{X}\,(1+e_2)$ with $E\,(e_1)=E\,(e_2)=0$. This leads to

$$\bar{y}_d=\bar{Y}\,(1+e_1)-\delta Y_k e_2$$

so that $E\,(\bar{y}_d)=\bar{Y}$. Thus $\bar{y}_d$ is an unbiased estimator of $\bar{Y}$.

Next,

$$\begin{aligned} V\,(\bar{y}_d) &= E\,(\bar{y}_d-\bar{Y})^2 \\ &= E\,(\bar{Y}e_1-\delta Y_k e_2)^2 \\ &= \bar{Y}^2 E\,(e_1^2)+Y_k^2\ E\,(e_2^2)-2\delta\bar{Y}Y_k\,E\,(e_1e_2) \\ &= \theta\,[S_y^2+C^2S_x^2-2\delta CS_{xy}] \qquad \ldots(3.2) \end{aligned}$$

where $\theta=\left(\frac{1}{n}-\frac{1}{N}\right)$, $C=\frac{Y_k}{\bar{X}}$, $S_y^2=\frac{1}{(N-1)}\sum_1^N Y_i-\bar{Y})^2$ etc.

To obtain the restrictions on Y_k dictated by the consideration that $\bar{y}_d$ be more precise than the traditional estimators we discuss the cases $\beta>0$ and $\beta<0$ separately.

Restrictions on Y_k

Clearly the value of C which minimizes $V(\bar{y}_d)$ with respect to C is $\delta\beta$ and the corresponding $Y_k=\delta\beta\bar{X}$. The crux of the problem is that β is usually unknown and estimating β by the coefficient of regression of y on x in the sample is computationally not convenient especially when many items are to be estimated from the same survey. Also in this case exact expressions for the expected value and variance of the estimator are hard to obtain (Das Raj [2], pp. 100-101).

Case I: $\beta>0$.

In this case the usual ratio method of estimation is to use

$$\bar{y}_r=\bar{y}\left(\frac{\bar{X}}{\bar{x}}\right)$$

as an estimator of $\bar{Y}$. The bias and mean square (mse) of $\bar{y}_r$ are given approximately by

$$B(\bar{y}_r)=\frac{\theta}{\bar{Y}}(RS_x^2-S_{xy})$$

and $$M(\bar{y}_r)=\theta(S_y^2+R^2S_x^2-2RS_{xy}) \qquad ...(4.1)$$

where $R=\frac{\bar{Y}}{\bar{X}}$. Derivation of these expressions depends on rather restrictive assumptions like $|e_2|<1$.

For an increase in precision relative to that of $\bar{y}_r$ we should have $V(\bar{y}_d)<M(\bar{y}_r)$, From (3.2) and (4.1) this needs that

$$(C^2-R^2)S_x^2+2(R-C)S_{xy}<0$$

or equivalently that

$$Q(C)=(C^2-R^2)+2(R-C)\beta<0. \qquad ...(4.2)$$

$Q(C)$ can be considered as a quadratic form in C. The condition (4.2) requires $Q(C)$ to be negative. This will happen iff the roots

of the quadratic equation $Q(C)=0$ are real and distinct and C lies between them. Here these roots are real and, unless $\alpha=0$, distinct. They are $2\beta-R$ and R. Thus (4.2) is the same as

$$C \text{ lies between } 2\beta-R \text{ and } R. \qquad ...(4.3)$$

Here the decrease in sampling variance of the estimator is

$$\left(\frac{\theta S_x^2}{\bar{X}^2}\right)(\bar{Y}-Y_k)(2\alpha+Y_k-\bar{Y}).$$

Next, since $V(\bar{y})=\theta S_y^2$ ...(4.4)

it follows similarly from (3.2) and (4.4) that

$$V(\bar{y}_d)<V(\bar{y})$$

if $$0<C<2\beta. \qquad ...(4.5)$$

To examine when (4.3) and (4.5) hold simultaneously we distinguish between subcases $\alpha>0$ and $\alpha<0$.

Subcase (i). $\alpha>0, \beta>0.$

Here $R>\beta$ and (4.3) can be written as

$$2\beta-R<C<R$$

or, using $\beta=\left(\frac{\bar{Y}-\alpha}{\bar{X}}\right)$, $C=\frac{Y_k}{\bar{X}}$ and $R=\frac{\bar{Y}}{\bar{X}}$ the condition becomes

$$(\bar{Y}-2\alpha)<Y_k<\bar{Y}. \qquad ...(4.3a)$$

Suppose Y_k is chosen to be less than $\bar{Y}$. For instance, since y and x are positively correlated characteristics, and supplementary information is available on x it should be possible to identify the smallest Y-value and use it as Y_k. With such a choice of Y_k, (4.3*a*) needs only that

$$(\bar{Y}-Y_k)<2\alpha. \qquad ...(4.3b)$$

However the other condition (4.5) needs that $(2\bar{Y}-Y_k)>2\alpha$. Thus the estimator $\bar{y}_d$ is more precise than the simple estimator $\bar{y}$ or the estimator $\bar{y}_r$ when

$$\bar{Y}<2\alpha+Y_k<2\bar{Y} \qquad ...(4.6)$$

which is not a stringent condition to be met when α is moderately large. In terms of ρ, the coefficient of correlation between y and x in the population, (4.6) can be written as

$$\frac{1}{2}\left(\frac{Y_k}{\bar{Y}}\right)\left(\frac{C_x}{C_y}\right)<\rho<\frac{1}{2}\left(1+\frac{Y_k}{\bar{Y}}\right)\left(\frac{Cx}{Cy}\right)$$

where C_y, C_x are the coefficients of variation of y and x in the population. This specifies an interval of length $\frac{1}{2}\left(\frac{C_x}{C_y}\right)$ for ρ.

Subcase (ii). $\alpha<0, \beta>0$.

Now $R<\beta$ and (4.3) becomes $R<C<2\beta-R$ which is equivalent to

$$\bar{Y}<Y_k<(\bar{Y}-2\alpha). \qquad \ldots(4.3c)$$

The strategy here is to choose $Y_k>\bar{Y}$ which makes (4.3*c*) to need only

$$(Y_k-\bar{Y})<-2\alpha \qquad \ldots(4.3d)$$

which interestingly enough implies (4.5). Thus any $Y_k>\bar{Y}$ and satisfying (4.3*d*) makes $\bar{y}_d$ superior, with respect to variance, to $\bar{y}$ or $\bar{y}_r$.

Case II : $\beta<0$.

Here it is appropriate to compare $\bar{y}_d$ with the product estimator (Murthy [4])

$$\bar{y}_p=\bar{y}\left(\frac{\bar{x}}{\bar{X}}\right)$$

which has exact bias $\theta S_{xy}/\bar{X}$ and mse (approx.)

$$M(\bar{y}_p)=\theta[S_y^2 \quad +R^2S_x^2 \quad +2RS_{xy}]. \qquad \ldots(4.7)$$

Now $V(\bar{y}_d)<M(\bar{y}_p)$ will need that

C lies between R and $-(2\beta+R)$ $\qquad \ldots(4.8)$

while $V(\bar{y}_d)<V(\bar{y})$ will require

$$0<C<-2\beta. \qquad \ldots(4.9)$$

Since y and x are assumed to be non-negative variates, α and β cannot be both negative. That is $\alpha>0$ in this subcase. However we need distinguish between the situations where $R\gtrless-(2\beta+R)$. These respectively correspond to $A<\bar{Y}$ and $A>\bar{Y}$ where $A=(2\alpha-3\bar{Y})$.

Subcase (iii). $A<\bar{Y}, \beta<0$.

We deliberately choose $Y_k<\bar{Y}$ and such that

$$A<Y_k<A+\bar{Y} \qquad \ldots(4.10)$$

for $\bar{y}_d$ to be more precise than $\bar{y}_p$ or $\bar{y}$,

Subcase (iv) : $A>\bar{Y}$, $\beta<0$.

It is now enough if we choose Y_k such that

$$\bar{Y}<Y_k<A. \qquad ...(4.11)$$

The guidelines for the choice of Y_k are summarized in table 4.1.

TABLE 4.1

Subcase	Y_k *chosen to be*	*Additional restriction on* Y_k
(*i*) $\beta>0, \alpha>0$	less than $\bar{Y}$	$A_0<Y_k<2(A_0+\alpha)$
(*ii*) $\beta>0, \alpha<0$	greater than $\bar{Y}$	$Y<A_0$
(*iii*) $\beta<0, A<\bar{Y}$	less than $\bar{Y}$	$A<Y_k<A+\bar{Y}$
(*iv*) $\beta<0, A>\bar{Y}$	greater than $\bar{Y}$	$Y_k<A$

Note : $A_0=\bar{Y}-2\alpha$, $A=(2\alpha-3\bar{Y})$.

In order to estimate $v(\bar{y}_d)$ given in (3.2) from the sample, we note that

$$s_y^2 = \frac{1}{n-1}\sum_1^n (Y_i-\bar{y})^2$$

is an unbiased estimator of S_y^2 and so on, where $\sum_1^n$ denotes summation over the units in the sample.

Numerical Example

The biases and the mean square errors of $\bar{y}$, $\bar{y}_r$ and $\bar{y}_d$ when simple random sampling two units without replacement from a hypothetical population of 5 units having (9, 1), (13, 2), (15, 3), (15, 4) and (18, 5) as the values of (y, x) are in table 4.2 for comparisons. Here $\bar{Y}=14$, $\bar{X}=3$. $\alpha=8$, $\beta=2$ and $\rho=0.9542$. This corresponds to subcase (*i*) and Y_k is chosen as the Y-value corresponding to the smallest X-value. Hence $Y_k=9$. It is to be noted that $\bar{y}_r$ does even worse than $\bar{y}$ since $\alpha>\bar{Y}/2$.

Use of Multi-Auxiliary Information

Frequently we possess information about several x-variates, and it may be considered important to make use of all the available material to our advantage. Suppose $x_1, \ldots, x_p$ denote these variates

TABLE 4.2

Sl. No.	*Estimator*	*Mse*	*Bias*
1	$\bar{y}$	3.30	0
2	$\bar{y}_r$	33.89	0.82
3	$\bar{y}_d$	1.05	0

each of which is correlated with y. Assume that measurements on all the variates are real, non-negative and that the population means $\bar{X}_t$, $t=1, \ldots, p$ of the x-variates are known and are positive. Let $Y=\alpha_t+\beta_t X_t$ be the population regression line when y is regressed on x_t alone. Consider the set of transformations

$$z_t=\frac{y}{Y_t}, \; w_t=\frac{x_t}{\bar{X}_t}, \; t=1, \ldots, p.$$

The choice of the Y_t is discussed later.

Let $\bar{y}, \bar{x}_t, t=1, \ldots, p$

denote the sample means and let

$$\bar{z}_t=\frac{\bar{y}}{Y_t}, \; \bar{w}_t=\frac{\bar{x}_t}{\bar{X}_t}, \; t=1, \ldots, p.$$

Then an unbiased estimator of the population mean of z_t is provided by

$$\bar{z}^{(t)}=\bar{z}_t+\delta_t \, (\bar{W}_t-\bar{w}_t), \; t=1, \ldots, p,$$

where $\delta_t=\pm 1$ according as $\beta_t \gtrless 0$ and $\bar{W}_t$ is the population mean of w_t. Note that $\bar{W}_t=1$ for each t. Then the estimators

$$\bar{y}^{(t)}=Y_t \bar{z}^{(t)}, \; t=1, \ldots, p \qquad \ldots(5.1)$$

are each unbiased for the population mean $\bar{Y}$. Using an appropriate weighting function $h=(h_1, \ldots, h_p)$,

$\sum_1^p h_t=1$, the estimators $\bar{y}^{(t)}$ can be be combined to give

$$\bar{y}_d=\sum_1^p h_t \bar{y}^{(t)} \qquad \ldots(5.2)$$

as an estimator of $\bar{Y}$ which utilizes the supplementary information on all the x-variates. Clearly $\bar{y}_d$ is unbiased for $\bar{Y}$. Its variance is

$$V(\bar{y}_d) = \sum_{s,t=1}^{p} h_s h_t \operatorname{cov}(\bar{y}^{(s)}, \bar{y}^{(t)}).$$

Defining S_{uv} as the population covariance between u and v and letting 0, 1, ..., p stand for the variates $y, x_1, \ldots, x_p$ respectively, it can be shown that

$$\operatorname{cov}(\bar{y}^{(s)}, \bar{y}^{(t)}) = \theta\,(S_{00} - \delta_s C_s S_{0s} - \delta_t C_t S_{0t} + \delta_s \delta_t C_s C_t S_{st}). \quad \ldots(5.3)$$

where $\theta = \left(\frac{1}{n} - \frac{1}{N}\right)$ and $C_t = \frac{Y_t}{\bar{X}_t}$, $t = 1, \ldots, p$. Thus

$$V(\bar{y}_d) = \sum_{s,t=1}^{p} h_s h_t d_{st} = hDh' \quad \ldots(5.4)$$

where $D = (d_{st})$ is the matrix of covariances defined in (5.3). The matrix D is at least positive semidefinite since it is a covariance matrix. It is positive definite under fairly general assumptions.

Theoretically the C_t can be determined such that $V(\bar{y}_d)$ is minimized with respect to them. But from the point of view of survey practice a simpler alternative is to choose C_t such that d_{tt} is minimized with respect to C_t. This is equivalent to minimizing

$$tr\ D = \sum d_{tt} = \sum_{1}^{p} V(\bar{y} - \delta_t C_t \bar{x}_t)$$

with respect to the C_t. The optimum value of C_t in this sense is

$$\delta_t \beta_t \left(= \frac{\delta_t S_{0t}}{S_{tt}} \right).$$

Suppose it is possible to make a good guess of the α_t, β_t and also identify each of these individual regressions with one of the subcases listed in table 4.1, then the choice of Y_t can be made parallel to that of Y_k there. This may not be hard, for example, when $p = 2$ and the α_t are moderately large in absolute value. In this context it may be mentioned that the case of one or two x-variates is of most frequent application. If any α_t is not moderately large in absolute value then the corresponding Y_t will have to be taken to approximate $\bar{Y}$ (see [1]). The above thumb rules are expected to keep the contribution from d_{tt} to $V(\bar{y}_d)$ under check as described in section 4.

It is of interest to note that if $Y_t=\bar{Y}$ for each t then $V(\bar{y}_d)$ is exactly the same as the large sample approximation for the variance of the generalized multivariate estimator involving the use of p auxiliary variates as given in [6]. The advantages of using a difference estimator in place of a generalized multivariate estimator are that the estimator is unbiased, the variance expression is exact and its derivation does not depend on restrictive assumptions and unbiased variance estimators are easy to obtain. It also appears that the precision of the estimator compares well with that of the generalized multivariate estimator in which case only large sample approximations for the bias and variances are available.

Next, applying the generalized Cauchy inequality (see Olkin [5]), the h_t optimum in the sense of minimizing $V(\bar{y}_d)$ for given D are provided by

$$h_{opt}=\frac{eD^{-1}}{eD^{-1}e'} \qquad ...(5.5)$$

where $e=(1, \ldots, 1)$ is the elementary row vector. Using these optimum weights, the minimized variance is found to be

$$V_{min}(\bar{y}_d)=\frac{1}{eD^{-1}e'}. \qquad ...(5.6)$$

However again in practice h_{opt} can rarely be computed and used since this requires the knowledge of the elements of the matrix D. In this context it may be noted that uniform weighting is obtained if and only if the column sums of D are equal. A hypothetical example of this case occurs when the coefficients of variation of the x-variates are all equal, there is the same correlation between y and the x_t, $t=1, \ldots p$, and C_t are the population ratios

$$R_t=\frac{\bar{Y}}{\bar{X}_t}, \text{ that is } Y_t=\bar{Y}$$

and all pairs of two different x-variates have the same correlation. Usually we have to select the h_t values from experience and theoretical indications regarding the relative influences of the x-variates on y.

In order to estimate $V(\bar{y}_d)$ note that

$$\bar{y}_d=\bar{y}+\sum_{t=1}^{p}\delta_t h_t C_t(\bar{X}_t-\bar{x}_t)$$

and hence

$$V(\bar{y}_d)=V\left(\bar{y}-\sum_{t=1}^{p}\delta_t h_t C_t \bar{x}_t\right)$$

$$=V\left[\frac{1}{n}\sum_{i=1}^{n}\left(Y_i-\sum_{t=1}^{p}\delta_t h_t C_t X_{ti}\right)\right].$$

Therefore an unbiased estimator of $V(\bar{y}_d)$ is

$$\hat{V}(\bar{y}_d)=\frac{\theta}{n-1}\sum_{i=1}^{n}\left[(Y_i-\bar{y})-\sum_{t=1}^{p}\delta_t h_t C_t (X_{ti}-\bar{x}_t)\right]^2.$$

Acknowledgements

The authors wish to thank the referee for the useful comments and Mrs. V. Stein for computational help.

REFERENCES

[1] Des Raj, (1965) : On a method of using multi-auxiliary information in sample surveys, *J. Amer. Stat. Ass.* 60, pp. 270-277.

[2] Des Raj. (1968) : Sampling Theory, McGraw Hill, Inc., N.Y.

[3] Mohanty, S. and M.N. Das, (1971) : Use of transformation in sampling, *J. Ind. Soc. Agri. Stat.* 23 (2)

[4] Murthy, M.N. (1964) : Product method of estimation, *Sankhya* 26A, pp. 69-74.

[5] Olkin, I. (1958) : Multivariate ratio estimation for finite populations, *Biometrika* 45, pp 154-165.

[6] Rao, P.S.R.S. and G.S. Mudholkar, (1967) : Generalized multivariate estimator for the mean of finite populations, *J. Amer. Stat. Ass.* 62, pp. 1009-1012.

[7] Srivenkataramana, T. (1978) : Change of origin and scale in ratio and difference methods of estimation in sampling, *The Canadian Journal of Statistics* 6 (1).