New Methodology of Calibration in Stratified Random Sampling

Shameem Alam[*] Sarjinder Singh[†] Javid Shabbir[‡]

Abstract

Calibration is a technique which improves the precision of the estimates of population parameters by using the auxiliary information. In the current investigation, the problem of estimation of population mean in stratified random sampling has been developed by a new calibration technique. It has been shown through simulation studies, the proposed resultant estimators are more efficient than the combined ratio estimator as well as the combined regression estimator.

keywords
Calibration, stratified random sampling, auxiliary information, percent relative efficiency, simulation.

1 Introduction

In survey sampling, the main purpose is to estimate the unknown population parameters such as mean, variance, total, correlation coefficient etc. It is observed that when the auxiliary information is available then the calibration methodology works better in estimating such parameters. Deville and Särndal (1992) were the pioneers of this technique and many other authors such as Arnab and Singh (2005), Farrell and Singh (2005) and Singh and Arnab (2011) also contributed their efforts in this technique. Many authors such as Kim et al. (2007), Koyuncu and Kadilar (2013) and Tracy et al. (2003) also developed calibration estimators in stratified random sampling. In the current investigation, we developed calibration estimators of finite population mean by using a new distance function.

Let $P = 1, 2, ..., N$ be a finite population of size N units that is divided into L homogeneous strata, such that the h^{th} stratum consists of N_h units, where $h = 1, 2, ..., L$ and $\sum_{h=1}^{L} N_h = N$. A sample of size n_h is drawn from the h^{th} stratum by using simple random sampling with replacement(SRSWR) such that $\sum_{h=1}^{L} n_h = n$. Let y_{hi} and x_{hi}, $i = 1, 2, ..., N_h$ be the values of the study variable Y and the auxiliary variable X for the i^{th} unit in the h^{th} stratum, respectively. In order to estimate the finite population mean say $\bar{Y}$, we assume that the complete information on the auxiliary variable X is known. The classical stratified unbiased estimator of the population mean $\bar{Y}$ is given by,

$$\hat{\bar{Y}}_{st} = \sum_{h=1}^{L} w_h \bar{y_h}, \tag{1}$$

[*]Department of Statistics, Quaid-i-Azam University, Islamabad, Pakistan shamimalamqau@gmail.com

[†]Department of Mathematics, Texas A&M University-Kingsville, Kingsville, TX, 78363, USA sarjinder.singh@tamuk.edu

[‡]Department of Statistics, Quaid-i-Azam University, Islamabad, Pakistan javidshabbir@gmail.com

where $w_h = \frac{N_h}{N}$ is the known stratum weight.

2 Proposed calibration estimator of mean in stratified sampling design

The proposed calibrated estimator of finite population mean is defined as,

$$\hat{\bar{Y}}_{c_1} = \sum_{h=1}^{L} w_h^{(1)} \bar{y}_h, \tag{2}$$

where $w_h^{(1)}$ are the new calibrated weights obtained by minimizing a new distance function,

$$\frac{1}{2}\sum_{h=1}^{L} \frac{\left(w_h^{(1)} - w_h\right)^2}{w_h q_h} + \sum_{h=1}^{L} \sum_{h \neq h'=1}^{L} \left(w_h^{(1)} - w_h\right)\left(w_{h'}^{(1)} - w_{h'}\right), \tag{3}$$

where $w_{h'}$ is the h'^{th} stratum weight. The new calibrated weights are obtained by minimizing (3) with respect to the following restriction:

$$\sum_{h=1}^{L} w_h^{(1)} \bar{x}_h = \bar{X} \tag{4}$$

The Lagrange's function labeled as,

$$L_{11} = \frac{1}{2}\sum_{h=1}^{L} \frac{\left(w_h^{(1)} - w_h\right)^2}{w_h q_h} + \sum_{h=1}^{L} \sum_{h \neq h'=1}^{L} \left(w_h^{(1)} - w_h\right)\left(w_{h'}^{(1)} - w_{h'}\right) - \lambda_{11}\left(\sum_{h=1}^{L} w_h^{(1)} \bar{x}_h - \bar{X}\right), \tag{5}$$

where q_h are suitably chosen weights. Taking partial derivative of L_{11} with respect to ${w_h}^{(1)}$ and equating to zero, results,

$$w_h^{(1)} = w_h + \lambda_{11}\left(\frac{w_h q_h \bar{x}_h}{1 - w_h q_h}\right) \tag{6}$$

The value of multiplier λ_{11} can be obtained by substituting the value of (6) in (4) as:

$$\lambda_{11} = \frac{\bar{X} - \sum_{h=1}^{L} w_h^{(1)} \bar{x}_h}{\sum_{h=1}^{L}\left(\frac{w_h q_h \bar{x}_h^2}{1 - w_h q_h}\right)} \tag{7}$$

Replacing back (7) in (6), we get the optimum calibrated weights such as,

$$w_h^{(1)} = w_h + \left[\frac{\bar{X} - \sum_{h=1}^{L} w_h^{(1)} \bar{x}_h}{\sum_{h=1}^{L}\left(\frac{w_h q_h \bar{x}_h^2}{1 - w_h q_h}\right)}\right]\left(\frac{w_h q_h \bar{x}_h}{1 - w_h q_h}\right) \tag{8}$$

Substituting the optimum calibrated weights $w_h^{(1)}$ in (2) results the first proposed calibrated estimator of the form:

$$\hat{\bar{Y}}_{c_1} = \sum_{h=1}^{L} w_h \bar{y}_h + \hat{\beta}_{11}\left(\bar{X} - \sum_{h=1}^{L} w_h \bar{x}_h\right), \tag{9}$$

where

$$\hat{\beta}_{11} = \frac{\sum_{h=1}^{L}\left(\frac{w_h q_h \bar{x}_h \bar{y}_h}{1-w_h q_h}\right)}{\sum_{h=1}^{L}\left(\frac{w_h q_h \bar{x}_h^2}{1-w_h q_h}\right)} \tag{10}$$

Similarly, the second proposed calibrated estimator of population mean is defined as,

$$\hat{\bar{Y}}_{c_2} = \sum_{h=1}^{L} w_h^{(2)} \bar{y}_h, \tag{11}$$

where $w_h^{(2)}$ are the new weights obtained by using the same distance function ,

$$\frac{1}{2}\sum_{h=1}^{L}\frac{\left(w_h^{(2)}-w_h\right)^2}{w_h q_h} + \sum_{h=1}^{L}\sum_{h\neq h'=1}^{L}\left(w_h^{(2)}-w_h\right)\left(w_{h'}^{(2)}-w_{h'}\right) \tag{12}$$

Minimizing (12) by using two constraints:

$$\sum_{h=1}^{L} w_h^{(2)} = \sum_{h=1}^{L} w_h \tag{13}$$

$$\sum_{h=1}^{L} w_h^{(2)} \bar{x}_h = \bar{X} \tag{14}$$

Such constraints can be found in Singh (2003). The associated Lagrange function is written as,

$$\begin{aligned} L_{21} =& \frac{1}{2}\sum_{h=1}^{L}\frac{\left(w_h^{(2)}-w_h\right)^2}{w_h q_h} + \sum_{h=1}^{L}\sum_{h\neq h'=1}^{L}\left(w_h^{(2)}-w_h\right)\left(w_{h'}^{(2)}-w_{h'}\right) - \lambda_{21}\left(\sum_{h=1}^{L} w_h^{(2)} - \sum_{h=1}^{L} w_h\right) \\ & - \lambda_{22}\left(\sum_{h=1}^{L} w_h^{(2)}\bar{x}_h - \bar{X}\right), \end{aligned} \tag{15}$$

where λ_{21} and λ_{22} are the Lagrange multipliers.
Differentiating (15) with respect to $w_h^{(2)}$ and equating to zero, we get

$$w_h^{(2)} = w_h + \lambda_{21}\left(\frac{w_h q_h}{1-w_h q_h}\right) + \lambda_{22}\left(\frac{w_h q_h \bar{x}_h}{1-w_h q_h}\right) \tag{16}$$

Substituting the value of equation (16) in equations (13) and (14) respectively, we get

$$\lambda_{21}\sum_{h=1}^{L}\left(\frac{w_h q_h}{1-w_h q_h}\right) + \lambda_{22}\sum_{h=1}^{L}\left(\frac{w_h q_h \bar{x}_h}{1-w_h q_h}\right) = 0 \tag{17}$$

$$\lambda_{21}\sum_{h=1}^{L}\left(\frac{w_h q_h \bar{x}_h}{1-w_h q_h}\right) + \lambda_{22}\sum_{h=1}^{L}\left(\frac{w_h q_h \bar{x}_h^2}{1-w_h q_h}\right) = \bar{X} - \sum_{h=1}^{L} w_h \bar{x}_h \tag{18}$$

The system of linear equations (17) and (18) can be written as,

$$\begin{bmatrix} \sum_{h=1}^{L}\left(\frac{w_h q_h}{1-w_h q_h}\right) & \sum_{h=1}^{L}\left(\frac{w_h q_h \bar{x}_h}{1-w_h q_h}\right) \\ \\ \sum_{h=1}^{L}\left(\frac{w_h q_h \bar{x}_h}{1-w_h q_h}\right) & \sum_{h=1}^{L}\left(\frac{w_h q_h \bar{x}_h^2}{1-w_h q_h}\right) \end{bmatrix} \begin{bmatrix} \lambda_{21} \\ \\ \lambda_{22} \end{bmatrix} = \begin{bmatrix} 0 \\ \\ \bar{X} - \sum_{h=1}^{L} w_h \bar{x}_h \end{bmatrix} \tag{19}$$

Solving the system (19) for the unknown values of λ_{21} and λ_{22}, we get

$$\lambda_{21} = -\frac{\left(\bar{X} - \sum_{h=1}^{L} w_h \bar{x}_h\right)\sum_{h=1}^{L}\left(\frac{w_h q_h \bar{x}_h}{1-w_h q_h}\right)}{\sum_{h=1}^{L}\left(\frac{w_h q_h}{1-w_h q_h}\right)\sum_{h=1}^{L}\left(\frac{w_h q_h \bar{x}_h^2}{1-w_h q_h}\right) - \left(\sum_{h=1}^{L}\left(\frac{w_h q_h \bar{x}_h}{1-w_h q_h}\right)\right)^2} \tag{20}$$

$$\lambda_{22} = \frac{\left(\bar{X} - \sum_{h=1}^{L} w_h \bar{x}_h\right)\sum_{h=1}^{L}\left(\frac{w_h q_h}{1-w_h q_h}\right)}{\sum_{h=1}^{L}\left(\frac{w_h q_h}{1-w_h q_h}\right)\sum_{h=1}^{L}\left(\frac{w_h q_h \bar{x}_h^2}{1-w_h q_h}\right) - \left(\sum_{h=1}^{L}\left(\frac{w_h q_h \bar{x}_h}{1-w_h q_h}\right)\right)^2} \tag{21}$$

After substituting the above optimum values of λ_{21} and λ_{22} in (16), the optimum calibrated weights are written as:

$$w_h^{(2)} = w_h + \frac{\left(\bar{X} - \sum_{h=1}^{L} w_h \bar{x}_h\right)}{\sum_{h=1}^{L}\left(\frac{w_h q_h}{1-w_h q_h}\right)\sum_{h=1}^{L}\left(\frac{w_h q_h \bar{x}_h^2}{1-w_h q_h}\right) - \left(\sum_{h=1}^{L}\left(\frac{w_h q_h \bar{x}_h}{1-w_h q_h}\right)\right)^2}$$
$$\left[\left(\frac{w_h q_h \bar{x}_h}{1-w_h q_h}\right)\sum_{h=1}^{L}\left(\frac{w_h q_h}{1-w_h q_h}\right) - \left(\frac{w_h q_h}{1-w_h q_h}\right)\sum_{h=1}^{L}\left(\frac{w_h q_h \bar{x}_h}{1-w_h q_h}\right)\right] \tag{22}$$

Replacing the optimum value of $w_h^{(2)}$ in (11), the second calibrated estimator of finite population mean $\bar{Y}$ takes the form as,

$$\hat{\bar{Y}}_{c_2} = \sum_{h=1}^{L} w_h \bar{y}_h + \hat{\beta}_{21}\left(\bar{X} - \sum_{h=1}^{L} w_h \bar{x}_h\right), \tag{23}$$

where

$$\hat{\beta}_{21} = \frac{\sum_{h=1}^{L}\left(\frac{w_h q_h \bar{x}_h \bar{y}_h}{1-w_h q_h}\right)\sum_{h=1}^{L}\left(\frac{w_h q_h}{1-w_h q_h}\right) - \sum_{h=1}^{L}\left(\frac{w_h q_h \bar{x}_h}{1-w_h q_h}\right)\sum_{h=1}^{L}\left(\frac{w_h q_h \bar{y}_h}{1-w_h q_h}\right)}{\sum_{h=1}^{L}\left(\frac{w_h q_h}{1-w_h q_h}\right)\sum_{h=1}^{L}\left(\frac{w_h q_h \bar{x}_h^2}{1-w_h q_h}\right) - \left(\sum_{h=1}^{L}\left(\frac{w_h q_h \bar{x}_h}{1-w_h q_h}\right)\right)^2} \tag{24}$$

3 Simulation study

The properties of the proposed estimators are investigated through simulation study. Both proposed estimators are compared with the general stratified mean estimator, combined ratio estimator and combined regression estimator for $q_h = 1$. The percent absolute relative bias and percent relative efficiency are the criterion used for comparing the estimators are defined as:

$$PARB(\hat{\bar{Y}}_v) = \left|\frac{\frac{1}{K}\sum_{r=1}^{K}\left(\hat{\bar{Y}}_v\right)_r - \bar{Y}}{\bar{Y}}\right| \times 100\%, \quad v = st, crat, creg, c_1, c_2 \tag{25}$$

$$MSE(\hat{\bar{Y}}_v) = \frac{1}{K}\sum_{r=1}^{K}\left(\left(\hat{\bar{Y}}_v\right)_r - \bar{Y}\right)^2, \quad r = 1, 2, ..., K \tag{26}$$

and

$$PRE(\hat{\bar{Y}}_l, \hat{\bar{Y}}_m) = \frac{MSE(\hat{\bar{Y}}_l)}{MSE(\hat{\bar{Y}}_m)} \times 100\%, \quad l = st, crat, creg, \quad m = c_1, c_2 \tag{27}$$

where the symbols st, $crat$ and $creg$ stands for stratified, combined ratio and combined regression estimators.

Based on artificial population, we generated $N = 450$ observations in three different strata of sizes 130, 170 and 150, respectively in each stratum. The population correlation coefficient ρ_{hxy} is taken as 0.45, 0.65 and 0.85 respectively in the h^{th} stratum where $h = 1, 2, 3$. Let $K = 1000$, simple random samples with replacement are drawn by using normal, gamma and beta distribution respectively, from sample of sizes $n = 10$ to $n = 50$ with an increment of 10, where $n = 20$ is rounded to 21 by using proportional allocation. We consider the following transformation,

$$y_{hi} = \mu^*_{y_h} + \rho_{hxy}\frac{\sigma^*_{y_h}}{\sigma^*_{x_h}}(x^*_h - \mu^*_{x_h}) + ({y_h}^* - \mu^*_{y_h})\sqrt{\left(1 - \rho^2_{hxy}\right)} \tag{28}$$

and

$$x_{hi} = \mu^*_{x_h} + ({x_h}^* - \mu^*_{x_h}), \quad h = 1, 2, 3, \quad i = 1, 2, ..., N_h \tag{29}$$

Let,
$x^*_1 \sim N(55, 10)$, $y^*_1 \sim N(45, 13)$,
$x^*_2 \sim G(2.5, 3)$, $y^*_2 \sim G(14.5, 3.2)$,
$x^*_3 \sim B(1.5, 1.7)$, $y^*_3 \sim B(3, 1.5)$,
where $\mu^*_{x_h}, \mu^*_{y_h}, \sigma^*_{x_h}$ and $\sigma^*_{y_h}$ are the respective means and standard deviations of x^*_h and y^*_h.
The results are given in the Table 1 and 2.

Table 1: Percent absolute relative biases based on simulated data in case of stratified random sampling

n	$PARB(\hat{\bar{Y}}_{st})$	$PARB(\hat{\bar{Y}}_{crat})$	$PARB(\hat{\bar{Y}}_{creg})$	$PARB(\hat{\bar{Y}}_{c_1})$	$PARB(\hat{\bar{Y}}_{c_2})$
10	0.62391	0.29642	0.18785	0.25678	0.23530
21	0.22936	0.15070	0.41744	0.08637	0.05503
30	0.15780	0.07021	0.06025	0.02881	0.00868
40	0.03475	0.07383	0.20783	0.04795	0.03545
50	0.33462	0.23395	0.32667	0.22566	0.22107

4 Conclusion

Table 1, gives the percent absolute relative biases of the usual mean estimators and of the proposed estimators. The proposed estimators $\hat{\bar{Y}}_{c_1}$ and $\hat{\bar{Y}}_{c_2}$ have minimum relative biases such as 0.08637% and 0.00868% as compared to the other usual mean estimators for $n = 21$ and $n = 30$, respectively. All the relative biases are negligible.

Table 2: Percent relative efficiencies based on simulated data in case of stratified random sampling

n	$PRE(\hat{\bar{Y}}_{st}, \hat{\bar{Y}}_{c_1})$	$PRE(\hat{\bar{Y}}_{crat}, \hat{\bar{Y}}_{c_1})$	$PRE(\hat{\bar{Y}}_{creg}, \hat{\bar{Y}}_{c_1})$	$PRE(\hat{\bar{Y}}_{st}, \hat{\bar{Y}}_{c_2})$	$PRE(\hat{\bar{Y}}_{crat}, \hat{\bar{Y}}_{c_2})$	$PRE(\hat{\bar{Y}}_{creg}, \hat{\bar{Y}}_{c_2})$
10	141.57	102.05	227.77	142.48	102.71	229.22
21	148.19	102.20	117.31	149.20	102.89	118.11
30	144.53	102.05	107.44	145.46	102.71	108.13
40	145.29	102.32	102.46	146.39	103.09	103.24
50	154.32	102.23	106.09	155.41	102.95	106.83

Table 2, shows the percent relative efficiencies of usual stratified mean estimator, combined ratio estimator, combined regression estimator and both proposed estimators. Both proposed estimators are more efficient than the usual stratified mean estimator with maximum gain of 54.32% and 55.41% respectively for $n = 50$. Similarly, both proposed estimators gain, approximately more than 2% as compared to combined ratio estimator. Comparing with combined regression estimator, the maximum at PRE of $\hat{\bar{Y}}_{c_1}$ is 227.77% and $\hat{\bar{Y}}_{c_2}$ is 229.22%. Finally, both proposed estimators are efficient than the usual estimators of mean. Hence new calibration methodology works better than the usual stratified mean estimator and combined regression estimator.

References

Arnab, R. and Singh, S. (2005). A note on variance estimation for the generalized regression predictor. *Australian & New Zealand Journal of Statistics*, 47(2):231–234.

Deville, J.-C. and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American statistical Association*, 87(418):376–382.

Farrell, P. J. and Singh, S. (2005). Model-assisted higher-order calibration of estimators of variance. *Australian & New Zealand Journal of Statistics*, 47(3):375–383.

Kim, J.-M., Sungur, E. A., and Heo, T.-Y. (2007). Calibration approach estimators in stratified sampling. *Statistics & probability letters*, 77(1):99–103.

Koyuncu, N. and Kadilar, C. (2013). Calibration estimator using different distance measures in stratified random sampling. *International Journal of Modern Engineering Research*, 3(1):415–419.

Singh, S. (2003). *Advanced Sampling Theory With Applications: How Michael"" Selected"" Amy*, volume 2. Springer Science & Business Media.

Singh, S. and Arnab, R. (2011). On calibration of design weights. *Metron*, 69(2):185–205.

Tracy, D., Singh, S., and Arnab, R. (2003). Note on calibration in stratified and double sampling. *Survey Methodology*, 29(1):99–104.