New calibration estimators in stratified sampling

D.K.Rao and T.Tekabu
School of Computing, Information and Mathematical Sciences,
The University of the South Pacific,
Suva, Fiji,
rao_di@usp.ac.fj and tekabu_t@usp.ac.fj

M.G.M Khan
School of Computing, Information and Mathematical Sciences,
The University of the South Pacific,
Suva, Fiji,
khan_mg@usp.ac.fj

***Abstract*—Calibration approach is widely used survey sampling that incorporates auxiliary information to increase the precision of survey estimates. In this manuscript, we propose two new calibration estimators of population mean in stratified sampling, using the known auxiliary information on mean and coefficient of variation in each stratum. A numerical example is presented to illustrate the application and computational details of the proposed calibration estimators. Moreover, a simulation study is carried out to compare the performance of the proposed calibration estimators.**

Keywords*—*Calibration estimator, Stratified random sampling, Auxiliary information, Lagrange multiplier technique, Coefficient of variation, Generalized linear regression.

I. INTRODUCTION

Calibration approach is well known in the sampling literature to increase the precision of the population parameters, using the known auxiliary information. The method works by minimizing the distance measure between the design and the calibrated weights subject to some calibration constraints on the auxiliary information.

The notion of calibration estimators was first introduced in survey sampling by Deville and Sarndal (1992) and since then several survey statisticians such as Singh et al. (1998, 1999, 2006, 2011), Singh (2001, 2003 , 2006, 2011, 2012), Farrell and Singh (2002, 2005), Wu and Sitter (2001), Sarndal (2007), Estevao and Sarndal (2000, 2003), Kott (2003), Montanari and Ranalli (2005), Rueda et al. (2010), Kim (2009, 2010) have contributed to improve the calibration approach.

Singh et al. (1998) introduced the calibration approach in stratified random sampling and later many others contributed such as Tracy et al. (2003), Singh (2003), Kim et al. (2007), Singh et al. (2014), Rao et al. (2012) and Rao et al. (2015)

Motivated by the estimators proposed by Singh et al. (1998), Tracy et al. (2003) and Singh (2003), we propose two new calibration estimators in stratified sampling, which incorporates the use of mean and coefficient of variation information in each stratum. The problem of determining the optimum calibrated weights is to minimize the chi-square type distance measure subject to a new calibration constraint.

Let the population of N units be divided into L non-overlapping, homogeneous sub-population called strata, such that the hth stratum consists of N_h units, where $h=1,2,\ldots,L$ and $\sum_{h=1}^{L} N_h = N$, the population size. For the hth population stratum, Y_{hi} and X_{hi} are the ith population unit of the study variable (Y) and the auxiliary variable (X), respectively, for $i=1,2,\ldots,N_h$. The population means of the study and auxiliary variable in the hth stratum are given by $\bar{Y}_h = N_h^{-1}\sum_{i=1}^{N_h} Y_{hi}$ and $\bar{X}_h = N_h^{-1}\sum_{i=1}^{N_h} X_{hi}$, respectively, for $h=1,2,\ldots,L$.

From the hth population stratum consisting of N_h units, a sample size of n_h units is drawn by simple random sampling without replacement (SRSWOR) such that $\sum_{h=1}^{L} n_h = n$, the total sample size. Also denote y_{hi} and x_{hi} to be the ith sample unit of the study and auxiliary variable, respectively, in the hth stratum. The sample means of the study and auxiliary variable in the hth stratum are given by $\bar{y}_h = n_h^{-1}\sum_{i=1}^{n_h} y_{hi}$ and $\bar{x}_h = n_h^{-1}\sum_{i=1}^{n_h} x_{hi}$ respectively, for $h=1,2,\ldots,L$.

Let the estimation of unknown population mean $\bar{Y} = \sum_{h=1}^{L} W_h \bar{Y}_h$, where $W_h = N_h / N$ be of interest, assuming some auxiliary information is known.

The stratified estimator of population mean $\bar{Y}$ is given by

$$\bar{y}_{st} = \sum_{h=1}^{L} W_h \bar{y}_h, \qquad (1)$$

where, $W_h = N_h / N$ and $\bar{y}_h = n_h^{-1}\sum_{i=1}^{n_h} y_{hi}$.

In the presence of single auxiliary variable X, we propose two calibration estimators of the population mean $\bar{Y}$.

978-1-5090-5753-5/16 $31.00 © 2016 IEEE
DOI 10.1109/APWC.on.CSE.2016.20

II. CALIBRATION ESTIMATOR I

A new calibration estimator of $\bar{Y}$ under stratified sampling is proposed as

$$\bar{y}_{st}^{*} = \sum_{h=1}^{L} W_h^{*} \bar{y}_h, \tag{2}$$

where, W_h^{*}, the calibrated weights are chosen in such a way that the chi-square distance function

$$D^{*} = \sum_{h=1}^{L} \frac{\left(W_h^{*} - W_h\right)^2}{W_h Q_h} \tag{3}$$

is minimum, subject to a new calibration constraint:

$$\sum_{h=1}^{L} W_h^{*}\left(\bar{x}_h + c_{hx}\right) = \sum_{h=1}^{L} W_h\left(\bar{X}_h + C_{hx}\right), \tag{4}$$

where,

$c_{hx} = \frac{s_{hx}}{\bar{x}_h}$, $s_{hx}^2 = \frac{1}{n_h - 1}\sum_{i=1}^{n_h}\left(x_{hi} - \bar{x}_h\right)^2$, $C_{hx} = \frac{S_{hx}}{\bar{X}_h}$, and $S_{hx}^2 = \frac{1}{N_h - 1}\sum_{i=1}^{N_h}\left(X_{hi} - \bar{X}_h\right)^2$. Also $Q_h > 0$ in (3) is the suitability chosen weights which determine the different form of the estimator.

The Lagrange multipliers technique can be used to compute the calibrated weights W_h^{*}, where the Lagrange function L is formed as:

$$L = \sum_{h=1}^{L} \frac{\left(W_h^{*} - W_h\right)^2}{W_h Q_h} - 2\lambda^{*}\left[\sum_{h=1}^{L} W_h^{*}\left(\bar{x}_h + c_{hx}\right) - \sum_{h=1}^{L} W_h\left(\bar{X}_h + C_{hx}\right)\right], \tag{5}$$

where λ^{*}, is a Lagrange multiplier. The necessary conditions for the solution of the problem are:

$$\frac{\partial L}{\partial W_h^{*}} = \frac{\partial L}{\lambda_1} = 0. \tag{6}$$

Using (6), W_h^{*} can be written as

$$W_h^{*} = W_h + \lambda^{*} W_h Q_h\left(\bar{x}_h + c_{hx}\right). \tag{7}$$

Using (7) and (4), we obtain

$$\lambda^{*} = \frac{\sum_{h=1}^{L} W_h\left(\bar{X}_h + C_{hx}\right) - \sum_{h=1}^{L} W_h\left(\bar{x}_h + c_{hx}\right)}{\sum_{h=1}^{L} W_h Q_h\left(\bar{x}_h + c_{hx}\right)^2}. \tag{8}$$

On substituting (8) in (7) the calibrated weights can be written as

$$W_h^{*} = W_h + \frac{W_h Q_h\left(\bar{x}_h + c_{hx}\right)}{\sum_{h=1}^{L} W_h Q_h\left(\bar{x}_h + c_{hx}\right)^2} \times\left[\sum_{h=1}^{L} W_h\left(\bar{X}_h + C_{hx}\right) - \sum_{h=1}^{L} W_h\left(\bar{x}_h + c_{hx}\right)\right]. \tag{9}$$

Substituting (9) in (2), we obtain the GREG type estimator as

$$\bar{y}_{st}^{*} = \bar{y}_{st} + \hat{\beta}^{*}\left[\sum_{h=1}^{L} W_h\left(\bar{X}_h + C_{hx}\right) - \sum_{h=1}^{L} W_h\left(\bar{x}_h + c_{hx}\right)\right] \tag{10}$$

where,

$$\hat{\beta}^{*} = \frac{\sum_{h=1}^{L} W_h Q_h \bar{y}_h\left(\bar{x}_h + c_{hx}\right)}{\sum_{h=1}^{L} W_h Q_h\left(\bar{x}_h + c_{hx}\right)^2}. \tag{11}$$

Remark:

1. The auxiliary information is combined as a single calibration constraint to form the estimator.

2. If $Q_h = 1$, then the estimator in (10) reduces to a Linear Regression (LREG) estimator.

3. If $Q_h = \frac{1}{\bar{x}_h + c_{hx}}$, then the estimator in (10) reduces to a new combined ratio estimator in stratified sampling defined as

$$\bar{y}_{st}^{*} = \sum_{h=1}^{L} W_h \bar{y}_h \frac{\sum_{h=1}^{L} W_h\left(\bar{X}_h + C_{hx}\right)}{\sum_{h=1}^{L} W_h\left(\bar{x}_h + c_{hx}\right)}. \tag{12}$$

III. CALIBRATION ESTIMATOR II

Similarly, another new calibration estimator of $\bar{Y}$ under stratified sampling is proposed as

$$\bar{y}_{st}^{\otimes} = \sum_{h=1}^{L} W_h^{\otimes} \bar{y}_h, \tag{13}$$

where, $W_h^{\otimes}$, the calibrated weights are chosen in such a way that the chi-square distance function is minimum, subject to a new calibration constraint

$$\sum_{h=1}^{L} W_h^{\otimes}\left(1 + \bar{x}_h + c_{hx}\right) = \sum_{h=1}^{L} W_h\left(1 + \bar{X}_h + C_{hx}\right). \tag{14}$$

Here, the Lagrange function L is defined as

$$L = \sum_{h=1}^{L} \frac{\left(W_h^{\otimes} - W_h\right)^2}{W_h Q_h} - 2\lambda^{\otimes}\left[\sum_{h=1}^{L} W_h^{\otimes}\left(1 + \bar{x}_h + c_{hx}\right) - \sum_{h=1}^{L} W_h\left(1 + \bar{X}_h + C_{hx}\right)\right]. \tag{15}$$

Minimizing the Lagrange function L, we obtain

$$W_h^{\otimes} = W_h + \lambda^{\otimes} W_h Q_h\left(1 + \bar{x}_h + c_{hx}\right). \tag{16}$$

Using (16) and (14) the langrage multiplier is obtained as

$$\lambda^{\otimes} = \frac{\sum_{h=1}^{L} W_h\left(\bar{X}_h + C_{hx}\right) - \sum_{h=1}^{L} W_h\left(\bar{x}_h + c_{hx}\right)}{\sum_{h=1}^{L} W_h Q_h \left(1+\bar{x}_h + c_{hx}\right)^2}. \tag{17}$$

Further, the calibrated weights in (16) can be written as

$$W_h^{\otimes} = W_h + \frac{W_h Q_h\left(1+\bar{x}_h + c_{hx}\right)}{\sum_{h=1}^{L} W_h Q_h\left(1+\bar{x}_h + c_{hx}\right)^2} \times\left[\sum_{h=1}^{L} W_h\left(\bar{X}_h + C_{hx}\right) - \sum_{h=1}^{L} W_h\left(\bar{x}_h + c_{hx}\right)\right]. \tag{18}$$

and hence the estimator in (13) can be written as

$$\bar{y}_{st}^{\otimes} = \bar{y}_{st} + \hat{\beta}^{\otimes}\left[\sum_{h=1}^{L} W_h\left(\bar{X}_h + C_{hx}\right) - \sum_{h=1}^{L} W_h\left(\bar{x}_h + c_{hx}\right)\right] \tag{19}$$

where,

$$\hat{\beta}^{\otimes} = \frac{\sum_{h=1}^{L} W_h Q_h \bar{y}_h\left(1+\bar{x}_h + c_{hx}\right)}{\sum_{h=1}^{L} W_h Q_h\left(1+\bar{x}_h + c_{hx}\right)^2}. \tag{20}$$

Remark:

1. The auxiliary information is combined as a single calibration constraint, which also incorporates that the sum of design weights be equal to the sum of calibrated weights, to form the estimator.

2. If $Q_h = 1,$ then the estimator in (19) reduces to a linear regression (LREG) estimator.

3. If $Q_h = \frac{1}{\bar{x}_h + c_{hx}},$ we obtain another new combined ratio estimator defined as

$$\bar{y}_{st}^{\otimes} = \sum_{h=1}^{L} W_h \bar{y}_h \frac{\sum_{h=1}^{L} W_h\left(1+\bar{X}_h + C_{hx}\right)}{\sum_{h=1}^{L} W_h\left(1+\bar{x}_h + c_{hx}\right)}. \tag{21}$$

IV. Numerical Illustration

In order to illustrate the application and computational details of the proposed estimators, we use a tobacco population data of $N = 106$ countries with three variables: area (in hectares), yield (in metric tons) and production (in metric tons). The data are obtained from the Agriculture Statistics 1999 reported in Singh (2003). The tobacco data was divided into $L = 10$ strata and a sample of $n = 40$ countries using proportional allocation was selected. Suppose that an estimate of average production ($\bar{Y}$) of tobacco crop is of interest using auxiliary variable $X =$ area. The same sample units as obtained in Singh (2003) are used for the computation.

Assuming $Q_h = 1$ and using the information given in Table I the following sample information are obtained:

$\sum_{h=1}^{L} W_h(\bar{x}_h + c_{hx}) = 59812.62$ and

$\sum_{h=1}^{L} W_h(\bar{x}_h + c_{hx})^2 = 14212155280.47.$

Assume that the known population information for the tobacco data is

$\sum_{h=1}^{L} W_h(\bar{X}_h + C_{hx}) = 34440.43.$

Using (8) and (17) the Lagrange multipliers for the calibration estimators were computed to be $\lambda^{*} = -1.78525\mathrm{E}-06,$ and $\lambda^{\otimes} = -0.000001785.$ The calibrated weights, W_h^{*} and $W_h^{\otimes}$ are calculated and displayed in Table II.

The estimates of the average production of tobacco using the proposed calibration estimators are

$$\bar{y}_{st}^{*} = 54330.87 \tag{22}$$

and

$$\bar{y}_{st}^{\otimes} = 54331.04. \tag{23}$$

V. Simulation Study

In this section, a simulation study is carried out to investigate the efficiency and the performance of the proposed estimators.

To carry out the simulation study, we used the same tobacco population, where, the population size $N = 106,$ the number of strata $L = 10,$ the stratum size $N_h = \{6,6,8,10,12,4,30,17,10,3\}$. We selected 5000 different samples of size $n = 40,$ that is, $n_h = \{3,3,3,3,4,2,11,6,3,2\}$ units from each stratum, respectively, using proportional allocation.

The correlation coefficient between the study ($Y =$ production) and the auxiliary variable ($X =$ area) is 0.991.

We calculated empirical mean square error (*MSE*) and percent relative efficiency (*PRE*) as follows:

$$MSE = \frac{1}{5000}\sum_{j=1}^{5000}\left[\left(\hat{\bar{Y}}\right)_j - \bar{Y}\right]^2, \quad \hat{\bar{Y}} = \left\{\bar{y}_{st}^{\Diamond}, \bar{y}_{st}^{\dagger}, \bar{y}_{st}^{*}, \bar{y}_{st}^{\otimes}\right\} \tag{24}$$

and

$$PRE = \frac{MSE\left(\bar{y}_{st}^{\Diamond}\right)}{MSE\left(\hat{\bar{Y}}\right)} \times 100\%, \quad \hat{\bar{Y}} = \left\{\bar{y}_{st}^{\dagger}, \bar{y}_{st}^{*}, \bar{y}_{st}^{\otimes}\right\} \tag{25}$$

where, $\bar{y}_{st}^{\Diamond}, \bar{y}_{st}^{\dagger}$ are Singh (2003) and Tracy et al. (2003) estimators and $\bar{y}_{st}^{*}, \bar{y}_{st}^{\otimes}$ are the proposed estimators, respectively.

The true average production of the tobacco crop for this population is $\bar{Y} = 52444.6$. The values of MSE, and PRE were obtained using a computer program developed in `MATLAB` and are presented in Table III for the different estimators.

Thus, from the Table III, it is evident from the *PRE* that the proposed calibration estimators are always more efficient than the Singh (2003) and Tracy (2003) for the tobacco population.

VI. CONCLUSION

In this paper, we propose two calibration estimators to estimate the population mean using known auxiliary information on mean and coefficient of variation in the stratum. A numerical example is presented to illustrate the computational details of the proposed estimators. The simulation study reveals that the proposed calibration estimators are more efficient than Singh (2003) and Tracy (2003). Moreover, the Tracy (2003) estimator performs most poorly and the proposed estimator I performs the best in this simulation study.

REFERENCES

[1] Deville, J.C, and Sarndal, C.E.(1992). Calibration Estimators in Survey Sampling. J.Amer.Statist.Assoc., 87, 376-382.

[2] Estevao, V.M, and Sarndal, C.E.(2000). A functional form approach to calibration. Journal of Official Statistics, 16, 379-399.

[3] Estevao, V.M, and Sarndal, C.E.(2003). A new perspective on calibration estimators. Joint Statistical Meeting-Section on Survey Research Methods, 1346-1356.

[4] Farrell, P.J, and Singh, S. (2002). Penalized chi-square distance function in survey sampling. Proc. of the Joint Statist. Meet-New York.

[5] Farrell, P.J, and Singh, S. (2005). Model-assisted higher order calibration of estimators of variance. Aust. Nz J. Stat., 47(3), 375-383.

[6] Kim, J.K. (2009) Calibration estimation using empirical likelihood in unequal probability sampling. Statist. Sinnica., 19, 145-157.

[7] Kim, J.K., (2010). Calibration estimation using exponential tilting in sample surveys. Statistics Canada, Catalogue No. 12-001, Vol. 36, 2, 145-155.

[8] Kim, J.M., Sungur, E.A. and Heo, T.Y. (2007). Calibration Approach Estimators in Stratified Sampling. Statistics & Probability Letters; Vol. 77, 1, 99-103.

[9] Kott, P.S. (2003). An overview of calibration weighting. Joint Statistical Meeting-Section on Survey Research Methods, 22241-2252.

[10] Montanari, G.E, and Ranalli, G.(2005). Nonparametric model calibration Estimation in Survey Sampling. J.Amer.Statist.Assoc., 100(472), 1429-1442.

[11] Rao, D.K, Khan, M.G.M. and Khan, S. (2012). Mathematical Programming on Multivariate Calibration Estimation in Stratified Sampling. World Academy of Science, Engineering and Technology 72, 58-62.

[12] Rao, D.K., Khan, M.G.M., and Reddy, K.G. (2015). *Stratified Calibration Estimator of Population Mean using Multi-auxiliary Information.* IEEE Proceedings of 2015 Asia-Pacific World Congress on Computer Science and Engineering, December 2-4, 2015, Shangri-La Fijian resort, Fiji.

[13] Rueda, M., Martinez, S. and Sanchez-Borrego, I. (2010). Model calibration estimation of the distribution function using nonparametric regression, Metrika, 71, 33-44.

[14] Sarndal, C.-E., (2007). The calibration approach in survey in theory and practice, Statistics Canada, Catalogue No. 12-001, Vol. 33, 2, 99-119.

[15] Singh, G.K., Rao, D.K. and Khan, M.G.M., (2014). Calibration Estimator of population mean in stratified sampling. IEEE Proceedings of 2014 Asia-Pacific World Congress on Computer Science and Engineering (APWC on CSE), 1-5, DOI: 10.1109/APWCCSE.2014.7053875.

[16] Singh, S. (2001). Generalized calibration approach for estimating the variance in survey sampling. Ann. Ins. Stat. Math. 53(2), 404-417.

[17] Singh, S. (2006). Survey Statisticians celebrate Golden and Silver Jubilee-2003 of the linear regression estimator. Metrika, 1-18.

[18] Singh, S. (2011). A dual problem of calibration of design weights. Statistics, 47, 566-574.

[19] Singh, S. (2012). Calibration of design weights using a displacement function. Metika, 85-107.

[20] Singh, S. and Arnab, R. (2006). A bridge between GREG and the linear regression estimators, Proc. of the ASA. Section on Survey Research Methods, Seattle, 3689-3693.

[21] Singh, S. and Arnab, R. (2011). On calibration of design weights, Metron, LXIX, 185–205.

[22] Singh, S., Horn, S., Chaudhuri, S. and Yu, F. (1999). Calibration of the estimator of variance, Aust. Nz J. Stat., 41(2), 199–212.

[23] Singh, S., Horn, S. and Yu, F. (1998). Estimation of variance of the general regression estimator: higher level calibration approach, Survey Methodology 24, 41–50.

[24] Tracy, D.S., Singh, S. and Arnab, R., (2003). Note on Calibration in Stratified and double Sampling. Survey Methodology 29, 99-104.

[25] Wu, C. & Sitter, R.R. (2001). A model-calibration approach to using complete auxiliary information from survey data. J. Amer. Statist. Assoc., 96, 185-193.

APPENDIX

TABLE I. INFORMATION FOR TOBACCO POPULATION

h	$\bar{x}_h$	c_{hx}	$\bar{y}_h$	W_h	$\bar{X}_h$	C_{hx}
1	1304.7	0.65137	2592.0	0.05660	3194.5	1.03348
2	29075.0	0.99624	26763.0	0.05660	14660.0	1.64983
3	5191.7	1.66129	14559.7	0.07547	18309.4	1.37734
4	21700.0	0.11354	29900.0	0.09434	14923.5	0.97062
5	6808.0	1.17116	12462.5	0.11321	5987.8	0.88123
6	1800.0	0.70711	3375.0	0.03774	3450.0	0.70266
7	24481.5	1.73379	38411.8	0.28302	11682.7	2.36010
8	294809.2	1.92712	477961.8	0.16038	145162.3	2.42586
9	6303.7	1.22819	7480.3	0.09434	33976.1	2.68800
10	350.0	1.01015	822.5	0.02830	1333.3	1.29108

TABLE II. CALIBRATED WEIGHTS

h	W_h^*	$W_h^{\otimes}$
1	0.056471869	0.056471769
2	0.053665597	0.053665521
3	0.074771972	0.074771844
4	0.090684903	0.090684766
5	0.111831392	0.111831201
6	0.037614539	0.037614473
7	0.270648493	0.270648091
8	0.075969131	0.075969555
9	0.093277756	0.093277597
10	0.028284152	0.028284101

TABLE III: SIMULATION RESULTS

Estimator	*MSE*	*PRE*
$\bar{y}_{st}^{\diamond}$	46721494	100.000
$\bar{y}_{st}^{\dagger}$	1.59E+10	0.295
$\bar{y}_{st}^{*}$	38797728	120.423
$\bar{y}_{st}^{\otimes}$	38799034	120.419