On calibrated weights in stratified sampling

D. K. Rao[1] M. G. M. Khan[2] G. K. Singh[3]

Received 19 November 2017; Revised 04 April 2018

Abstract

In this paper, we propose a calibration estimator of population mean in stratified sampling using the known mean and variance information from multi-auxiliary variables. The problem of determining the optimum calibrated weights is formulated as an optimisation problem and is solved using the Lagrange multiplier technique. A numerical example with real data is presented to illustrate the computational details of the proposed estimator. A comparison study is also carried out using real and simulated data to evaluate the performance and the usefulness of the proposed estimator. The study reveals that the proposed estimator with multi-auxiliary information is the most efficient estimator of the population mean when compared to other estimators as it provides least estimated variance and highest gain in relative efficiency (RE).

Subject class: 62D05; 62-02

DOI:10.21914/anziamj.v59i0.12668, Published September 24, 2018, as part of the Proceedings of the 13th Biennial Engineering Mathematics and Applications Conference. ISSN 1445-8810. (Print two pages per sheet of paper.)

Keywords: Stratified sampling; calibration estimation; auxiliary information

Contents

1 Introduction

Calibration estimation, on which the current research is conducted, dates back to 1992. A large amount of literature is being devoted to it, gaining significant attention in the field of survey methodology and survey practice. It is a technique that uses available auxiliary information to improve the precision of the survey estimates. The technique works by minimising the chi-square distance function subject to some calibration constraints. The notion of calibration estimators was first introduced by [1] in survey sampling.

Since then several survey statisticians have contributed to the study of calibrated estimation in survey sampling [2, 3, 6, 10, 13, 11, 12, 15]. Singh et al. [10] introduced the calibration approach in stratified random sampling

where they proposed the combined generalised regression (GREG) estimator of population mean using the known mean information from a single auxiliary variable. Later, many authors have contributed to the theory of calibration estimation in stratified sampling [5, 8, 7, 11, 9, 14].

The purpose of this paper is to propose a calibration estimator of population mean in stratified sampling using the known mean and variance information from several auxiliary variables. Our main contributions include (1) introducing new calibration constraints; (2) generalising the problem with multi-auxiliary variables; (3) investigating the efficiency of the proposed estimators; and (4) investigating whether the information from several auxiliary variables improves the estimate of population mean.

The problem of determining the optimum calibrated weights is formulated as an optimisation problem that minimises the chi-square type distance, subject to some new calibration constraints. The problem is then solved to determine the calibrated weights using the Lagrange multiplier technique. The computational details of the procedure are illustrated in the presence of two auxiliary variables. A numerical example with real data is presented to demonstrate the computational details of the proposed estimator. To compare the efficiency gain of the proposed multivariate estimator with the other calibration estimators a comparison study is carried out. The study reveals that the proposed multivariate estimator is more efficient than the other calibration estimators.

2 The Problem of Calibrated Weights

Consider that a finite population $U = \{1, 2, ..., i, ..., N\}$ of size N is stratified into L strata $U_h = \{1, 2, ..., i, ..., N_h\}$ containing N_h units in hth stratum $(h = 1, 2, ..., L)$ such that $\sum_{h=1}^{L} N_h = N$ and let $W_h = N_h/N$ be the stratum weights. A sample of size n, comprising of n_h units from strata, h is drawn using simple random sampling without replacement (SRSWR). Let y_{hi} and

x_{hij} denote the value of ith unit from hth stratum for the study variable y and the jth auxiliary variable x_j; $j = 1, 2, ..., i, ..., p$, respectively. For each stratum, h: $\bar{y}_h = \frac{1}{n_h}\sum_{i=1}^{n_h} y_{hi}$ is the sample mean of the study variable. Assume that stratum means $\bar{X}_{hj} = \frac{1}{N_h}\sum_{i=1}^{N_h} x_{hij}$ and the stratum variances

$$S_{hj}^2 = \frac{1}{N_h - 1}\sum_{i=1}^{N_h}\left(x_{hij} - \bar{X}_{hj}\right)^2,$$

of all the p auxiliary variables are accurately known. The purpose of the study is to propose a calibration estimator of the population mean $\bar{Y} = \sum_{h=1}^{L} W_h \bar{Y}_h$ where $\bar{Y}_h = \frac{1}{N_h}\sum_{i=1}^{N_h} y_{hi}$ by using the information from p auxiliary variables x_j.

The stratified estimator of the population mean is given by

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

In the presence of p auxiliary variables x_j; $j = 1, 2, ..., p$ a new calibration estimator of the population mean under stratified sampling is given by

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

where W_h^* are called the calibrated weights. The weights W_h^* are so chosen such that the chi-square type distance function

$$\sum_{j=1}^{p}\sum_{h=1}^{L}\frac{\left(W_h^* - W_h\right)^2}{W_h q_{hj}}, \tag{3}$$

is minimum, subject to the calibration constraints

$$\sum_{h=1}^{L} W_h^* = 1, \tag{4}$$

$$\sum_{h=1}^{L} W_h^* \bar{x}_{hj} = \sum_{h=1}^{L} W_h \bar{X}_{hj}; j = 1, 2, ..., p, \tag{5}$$

$$\sum_{h=1}^{L} W_h^* d_h s_{hj}^2 = \sum_{h=1}^{L} W_h d_h S_{hj}^2; j = 1, 2, ..., p, \tag{6}$$

where $d_h = (1/n_h - 1/N_h)$ are the weights associated with the variance, $\bar{x}_{hj} = \frac{1}{n_h}\sum_{i=1}^{n_h} x_{hij}$, $s_{hj}^2 = \frac{1}{n_h - 1}\sum_{i=1}^{n_h}(x_{hij} - \bar{x}_{hj})^2$ and q_{hj} are suitably chosen constants to obtain different forms of the estimator. Motivated by the calibration constraint given by [14], we have proposed a similar constraint as in (6) by introducing the weights d_h.

Thus, the problem of determining the optimum calibrated weights W_h^* may be formulated as an optimisation problem given below:

$$\begin{aligned} \text{Minimize:} \quad & \sum_{h=1}^{L} \frac{(W_h^* - W_h)^2}{W_h Q_h} \\ \text{subject to} \quad & \sum_{h=1}^{L} W_h^* = 1, \\ & \sum_{h=1}^{L} W_h^* \bar{x}_{hj} = \sum_{h=1}^{L} W_h \bar{X}_{hj}; j = 1, 2, ..., p, \\ & \sum_{h=1}^{L} W_h^* d_h s_{hj}^2 = \sum_{h=1}^{L} W_h d_h S_{hj}^2; j = 1, 2, ..., p, \end{aligned} \tag{7}$$

where $Q_h = \left(\sum_{j=1}^{p} \frac{1}{q_{hj}}\right)^{-1}$ are suitably chosen constants to obtain different forms of the estimator.

3 Determining the Optimum Calibrated Weights

It can be seen that the objective function of problem (7) is convex and the constraints are linear equations, hence the Lagrange multiplier technique will yield an optimum solution. Thus we can use Lagrange multiplier technique to solve the problem (7) and determine the optimum values of W_h^*.

Defining λ_0, and λ_j, φ_j for $j = 1, 2, ..., p$, as Lagrange multipliers, the Lagrange function is given by

$$\begin{aligned}\phi = &\sum_{h=1}^{L} \frac{(W_h^* - W_h)^2}{W_h Q_h} - 2\lambda_0 \left(\sum_{h=1}^{L} W_h^* - 1 \right) \\ &- 2 \sum_{j=1}^{p} \lambda_j \left(\sum_{h=1}^{L} {W_h}^* \bar{x}_{hj} - \sum_{h=1}^{L} W_h \bar{X}_{hj} \right) \\ &- 2 \sum_{j=1}^{p} \varphi_j \left(\sum_{h=1}^{L} W_h^* d_h s_{hj}^2 - \sum_{h=1}^{L} W_h d_h S_{hj}^2 \right). \end{aligned} \tag{8}$$

The necessary and sufficient conditions for solving optimum values of W_h^* are

$$\frac{\partial \phi}{\partial {W_h}^*} = \frac{2\,(W_h^* - W_h)}{W_h Q_h} - 2\lambda_0 - 2 \sum_{j=1}^{p} \lambda_j \bar{x}_{hj} - 2 \sum_{j=1}^{p} \varphi_j d_h s_{hj}^2 = 0, \tag{9}$$

$$\frac{\partial \phi}{\partial \lambda_0} = -2 \left(\sum_{h=1}^{L} W_h^* - 1 \right) = 0, \tag{10}$$

$$\frac{\partial \phi}{\partial \lambda_j} = -2 \left(\sum_{h=1}^{L} W_h^* \bar{x}_{hj} - \sum_{h=1}^{L} W_h \bar{X}_{hj} \right) = 0, \tag{11}$$

and

$$\frac{\partial \phi}{\partial \varphi_j} = -2\left(\sum_{h=1}^{L} W_h^* d_h s_{hj}^2 - \sum_{h=1}^{L} W_h d_h S_{hj}^2\right) = 0. \tag{12}$$

From (9) we have

$$W_h{}^* = W_h + W_h Q_h \left(\lambda_0 + \sum_{j=1}^{p} \lambda_j \bar{x}_{hj} + \sum_{j=1}^{p} \varphi_j d_h s_{hj}^2\right), \tag{13}$$

where λ_0, λ_j and φ_j for $j = 1, 2, ..., p$ will be obtained using [4] by solving a system of nonlinear equations (10)-(12) for the given values of W_h, d_h, $\bar{x}_{hj}$ and s_{hj}^2.

4 Numerical Illustration and Comparison Study

In this section, we illustrate the computational details and demonstrate the performance of the proposed estimator using the tobacco data (Source: Agriculture Statistics 1999 [11]).

4.1 Numerical Illustration

In order to illustrate the computational details of the proposed estimator, we now describe the tobacco population. The population consists of data of $N = 106$ counties with three variables: area (in hectares), yield (in metric tons) and production (in metric tons). The countries were divided into $L = 10$ strata and a sample of $n = 40$ countries using proportional allocation was selected.

Table 1: Population information for tobacco data.

h	N_h	W_h	$\bar{X}_{h1}$	$\bar{X}_{h2}$	S^2_{h1}	S^2_{h2}
1	6	0.05660	3194.5	1.9733	10899652.7	0.0268
2	6	0.05660	14660.0	1.3883	584984730.0	0.2181
3	8	0.07547	18309.4	2.5563	635958094.8	0.3470
4	10	0.09434	14923.5	1.5490	209817189.2	0.2346
5	12	0.11321	5987.8	1.8317	27842810.5	0.5821
6	4	0.03774	3450.0	1.4700	5876666.7	0.1531
7	30	0.28302	11682.7	1.1150	760238523.4	0.3439
8	17	0.16038	145162.3	1.3818	124004506112.8	0.3786
9	10	0.09434	33976.1	1.7210	8340765245.4	2.0183
10	3	0.02830	1333.3	2.0867	2963333.3	0.9746

Suppose that an estimate of average production of tobacco $\left(\bar{Y}\right)$ is of interest using the two auxiliary variables x_1 = area and x_2 = yield. To determine the multivariate calibrated weights and the value of the estimate of $\bar{Y}$ in stratified sampling we use the same sample units as obtained in [11] and we assume that $Q_h = 1$. The information needed for computation is summarised in Table 1 and Table 2. Substituting (13) in equations (10)-(12) and solving the system of nonlinear equations using [4] we obtain $\lambda_0 = 0.63$, $\lambda_1 = 2.64 \times 10^{-6}$, $\lambda_2 = -0.50$, $\varphi_1 = -4.04 \times 10^{-11}$ and $\varphi_2 = 3.12$. The optimum calibrated weights W_h^* are obtained and presented in Column 2 of Table 3.

The calibrated weights of other estimators to be discussed in Subsection 4.2 are also presented in Columns 3, 4, 5 and 6 of Table 3 and will be later used for comparing the efficiency of the estimators.

Using (2), an estimate of the average production of tobacco using the proposed estimator is given by

$$\bar{y}^*_{st} = \sum_{h=1}^{L} W_h^* \bar{y}_h = 53585.53. \tag{14}$$

Table 2: Sample information for tobacco data.

h	n_h	$\bar{x}_{h1}$	$\bar{x}_{h2}$	s^2_{h1}	s^2_{h2}	$\bar{y}_h$
1	3	1304.7	1.9400	722185.3	0.0171	2592.0
2	3	29075.0	1.3767	839008125	0.4757	26763.0
3	3	5191.7	2.7933	74387858.3	0.8010	14559.7
4	3	21700.0	1.4433	6070000.0	0.7362	29900.0
5	4	6808.0	1.7875	63572981.3	1.0698	12462.5
6	2	1800.0	1.7850	1620000.0	0.0612	3375.0
7	11	24481.5	1.3209	1801653230.3	0.4824	38411.8
8	6	294809.2	1.3200	322774101004.2	0.2462	477961.8
9	3	6303.7	1.3267	59939890.3	0.0306	7480.3
10	2	350.0	1.7650	125000.0	1.3285	822.5

Table 3: Calibrated Weights for different methods.

h	W^*_h	$W^{(1)}_h$	$W^{(2)}_h$	$W^{(3)}_h$	$W^{(4)}_h$
1	0.03768	0.05647	0.05758	0.06450	0.05489
2	0.07101	0.05367	0.07769	0.06075	0.07103
3	0.05720	0.07477	0.08057	0.08531	0.07624
4	0.14098	0.09069	0.12140	0.10292	0.11227
5	0.14766	0.11183	0.12330	0.12753	0.11640
6	0.02957	0.03761	0.03864	0.04296	0.03680
7	0.31456	0.27065	0.36750	0.30688	0.34223
8	0.06649	0.07597	0.06273	0.07046	0.06654
9	0.09428	0.09327	0.20212	0.10638	0.09645
10	0.04057	0.02828	0.01640	0.03231	0.02715

4.2 Comparison Study

In this Subsection, using the tobacco data a comparison study is carried out on the efficiency of the following calibration estimators:

1. Singh (1998) estimator, $\bar{y}_{st}^{(1)} = \sum_{h=1}^{L} W_h^{(1)} \bar{y}_h$ in [10].
2. Tracy (2003) estimator, $\bar{y}_{st}^{(2)} = \sum_{h=1}^{L} W_h^{(2)} \bar{y}_h$ in [14].
3. Singh (2003) estimator, $\bar{y}_{st}^{(3)} = \sum_{h=1}^{L} W_h^{(3)} \bar{y}_h$ in [11].
4. A univariate estimator of (2), $\bar{y}_{st}^{(4)} = \sum_{h=1}^{L} W_h^{(4)} \bar{y}_h$ where
   $W_h{}^{(4)} = W_h + W_h Q_h \left(\lambda_0 + \lambda_1 \bar{x}_{h1} + \varphi_1 d_h s_{h1}^2\right)$
   and the auxiliary variable is $x_1 =$ area.
5. Proposed multivariate estimator $\bar{y}_{st}^{*} = \sum_{h=1}^{L} W_h^{*} \bar{y}_h$ in (2).

To compare the efficiency of the above estimators with respect to the stratified estimator $\bar{y}_{st}$, we compute the measure of relative efficiency (RE) as

$$RE = \frac{\hat{v}\left(\bar{y}_{st}\right)}{\hat{v}\left(\hat{\bar{Y}}\right)} \times 100, \tag{15}$$

where

$$\hat{v}\left(\bar{y}_{st}\right) = \sum_{h=1}^{L} W_h^2 d_h s_h^2, \tag{16}$$

is the estimated variance of $\bar{y}_{st}$, $\hat{v}\left(\hat{\bar{Y}}\right)$ is the estimated variance of a calibration estimator and $s_h^2 = \frac{1}{n_h - 1} \sum_{i=1}^{n_h} \left(y_{hi} - \bar{y}_h\right)^2$. The denominator $\hat{v}\left(\hat{\bar{Y}}\right)$ in (15)

Table 4: Comparison results for tobacco data.

Estimator	$\hat{\bar{Y}}$	$\hat{v}\left(\hat{\bar{Y}}\right)$	RE
Stratified: $\bar{y}_{st}$	95373.90	2822731121.0	100.0
Singh (1998): $\bar{y}_{st}^{(1)}$	54329.76	650863727.5	433.7
Tracy (2003): $\bar{y}_{st}^{(2)}$	54321.02	609889505.3	462.8
Singh (2003): $\bar{y}_{st}^{(3)}$	54132.89	569781129.3	495.4
Univariate: $\bar{y}_{st}^{(4)}$	53775.78	517799741.5	545.1
Proposed multivariate: $\bar{y}_{st}^{*}$	53585.53	512333228.2	551.0

is computed using the lower level calibration approach (see [11]) that is by replacing the stratum weights with the calibrated weights in equation (16).

Based on the tobacco population used in Subsection 4.1, we compare the performance of the proposed estimator based on the two auxiliary variables (x_1= Area and x_2= Yield) and other calibration estimators on the single auxiliary variable (x_1= Area). It should be noted that the true average production of the tobacco crop for this population is $\bar{Y} = 52444.56$. In Table 4, the Columns 2, 3 and 4 presents the estimated average production of tobacco $\left(\hat{\bar{Y}}\right)$, the estimated variance $\hat{v}\left(\hat{\bar{Y}}\right)$ and the relative efficiency (RE) for different estimators considered.

Finally, amongst all the estimators, it was found that the proposed estimator $\bar{y}_{st}^{*}$ has the smallest estimated variance and highest RE. Thus, the study reveals that the estimator $\bar{y}_{st}^{*}$ is the most efficient estimator of population mean in stratified sampling using the tobacco data. The gain in efficiency of the proposed estimator over the stratified estimator is 550.96%.

A comparison study was also carried out using a simulated data and similar results were obtained that is the proposed estimator $\bar{y}_{st}^{*}$ has the least estimated variance and highest gain in RE and hence the most efficient estimator (see Table 5 for the results of the simulated data). The gain in efficiency of the proposed estimator over the stratified estimator is 386.85%.

Table 5: Comparison results for simulated data.

Estimator	$\hat{\bar{Y}}$	$\hat{v}\left(\hat{\bar{Y}}\right)$	RE
Stratified: $\bar{y}_{st}$	760.951	45844670263.9	100.00
Singh (1998): $\bar{y}_{st}^{(1)}$	760.867	45835552384.3	100.02
Tracy (2003): $\bar{y}_{st}^{(2)}$	760.869	45745087574.3	100.22
Singh (2003): $\bar{y}_{st}^{(3)}$	760.868	45832943300.4	100.03
Univariate: $\bar{y}_{st}^{(4)}$	760.872	11870824590.0	386.20
Proposed multivariate: $\bar{y}_{st}^{*}$	760.882	11850624683.1	386.85

5 Conclusion

In surveys, the statisticians are often interested to improve the precision of the survey estimates. The calibration approach is one such technique that incorporates the auxiliary information in survey sampling to improve the precision of the survey estimates.

In this paper, we considered the problem of determining the optimum calibrated weights and the optimum calibration estimator of population mean in stratified sampling, when the auxiliary information (mean and/or variance) from several variables are available. The problem is formulated as an optimisation problem that seeks minimisation of the chi-square distance function, subject to the proposed calibration constraints. The problem is then solved using the Lagrange multiplier technique. A numerical example with a real data are presented to illustrate the computational details of the proposed estimator. A comparison study with a real and a simulated data is carried out to determine the performance of the proposed estimator. The results show that the proposed estimator is the most efficient estimator of population mean in stratified sampling. Thus, it can be concluded that the precision of

the survey estimates is further improved when multi-auxiliary information (mean and/or variance) is used as proposed.

References

[1] Jean Claude Deville and Carl Erik Särndal. Calibration estimators in survey sampling. *Journal of the American statistical Association*, 87(418):376–382, 1992. doi:10.1080/01621459.1992.10475217. C191

[2] Victor M Estevao and Carl Erik Särndal. Survey estimates by calibration on complex auxiliary information. *International Statistical Review*, 74(2):127–147, 2006. doi:110.1111/j.1751-5823.2006.tb00165.x C191

[3] Patrick J Farrell and Sarjinder Singh. Model-assisted higher-order calibration of estimators of variance. *Australian & New Zealand Journal of Statistics*, 47(3):375–383, 2005. doi:10.1111/j.1467-842X.2005.00402.x C191

[4] Wolfram Research, Inc. Mathematica, Version 11.3. Champaign, IL, 2018. C196, C197

[5] Jong Min Kim, Engin A Sungur, and Tae Young Heo. Calibration approach estimators in stratified sampling. *Statistics & probability letters*, 77(1):99–103, 2007. doi:10.1016/j.spl.2006.05.015 C192

[6] Phillip S Kott. Using calibration weighting to adjust for nonresponse and coverage errors. *Survey Methodology*, 32(2):133, 2006. C191

[7] Dinesh K Rao. *Mathematical programing in stratified random sampling*. PhD thesis, School of Computing, Information and Mathematical Sciences, The University of the South Pacific, Fiji, February 2017. C192

[8] Dinesh K. Rao, Tokaua. Tekabu, and Mohammad G M Khan. New calibration estimators in stratified sampling. In *Proceedings of*

Asia-Pacific World Congress on Computer Science and Engineering, pages 66–70. IEEE, 2016. C192

[9] Gurmindar K Singh, Dinesh K Rao, and Mohammed GM Khan. Calibration estimator of population mean in stratified random sampling. In *Proceedings of Asia-Pacific World Congress on Computer Science and Engineering (APWC on CSE)*, pages 1–5. IEEE, 2014. C192

[10] Sarjindar Singh, Stephen Horn, and Frank Yu. Estimation of variance of general regression estimator: Higher level calibration approach. *Survey Methodology*, 24:41–50, 1998. C191, C199

[11] Sarjinder Singh. *Advanced Sampling Theory With Applications: How Michael "Selected" Amy*, volume I & II. Kluwer Academic Publishers, Netherlands, 2003. C191, C192, C196, C197, C199, C200

[12] Sarjinder Singh. On the calibration of design weights using a displacement function. *Metrika*, 75(1):85–107, 2012. doi:10.1007/s00184-010-0316-6 C191

[13] Sarjinder Singh, Stephen Horn, Sadeq Chowdhury, and Frank Yu. Theory & methods: Calibration of the estimators of variance. *Australian & New Zealand Journal of Statistics*, 41(2):199–212, 1999. doi:10.1111/1467-842X.00074 C191

[14] D S Tracy, S Singh, and R Arnab. Note on calibration in stratified and double sampling. *Survey Methodology*, 29(1):99–104, 2003. C192, C194, C199

[15] Changbao Wu and Randy R Sitter. A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association*, 96(453):185–193, 2001. doi:10.1198/016214501750333054 C191

Author addresses

1. **D. K. Rao**, School of Computing, Information and Mathematical Sciences, The University of the South Pacific, Suva, Fiji.
mailto:dinesh.rao@usp.ac.fj
2. **M. G. M. Khan**, School of Computing, Information and Mathematical Sciences, The University of the South Pacific, Suva, Fiji.
3. **G. K. Singh**, School of Computing, Information and Mathematical Sciences, The University of the South Pacific, Suva, Fiji.