A New Calibration Estimator of

Stratified Random Sample

Amany M. Mouhamed
Institute of Statistical Studies and Research
Cairo University, Egypt

Ahmed A. El-sheikh
Institute of Statistical Studies and Research
Cairo University, Egypt

Hamida A. Mohamed
Institute of Statistical Studies and Research
Cairo University, Egypt

Abstract

In this paper, a new calibration estimators for stratified random sample for optimum and Neyman allocation is derived using some classical distances and also, for new suggested distances. A new an estimator of variance of the calibration approach is introduced. The performance of the new calibration estimators compared with Horvitz-Thompson estimator has been established by simulation study.

Keywords: Auxiliary Information, Calibration estimation, Stratified random sample, Horvitz-Thompson estimator

1. Introduction

In survey sampling, the use of auxiliary information can greatly improve the precision of estimates of population total and/or means. Calibration provided a systematic way to incorporate auxiliary information in the procedure. So, it became a
widely used procedure of estimation in sampling survey. Deville and Särndal (1992), estimated a finite population totals in the presence of univariate or multivariate auxiliary information. Théberge (1999), extended the calibration technique to estimate population parameters other than totals and means, and developed the technique when there is no solution to the calibration equation. He developed a new method to compute a calibration estimator that used an arbitrary distance measure. Estevao and Särndal (2002), showed that there are exactly nine different subsets of the complete information, for ten different cases of auxiliary information where more extensive of auxiliary information, the precision of the resulting estimates will be much better.

Tracy, et al (2003), introduced new calibration equations making use of the second order moments of the auxiliary character for estimating the population mean in stratified simple random sampling. Also, different methods for estimating the variance of the proposed estimator are suggested.

Kim, et al (2006), proposed various calibration approach ratio estimators and derived the estimator of the variance of the calibration approach ratio-type estimators of stratified random sample for four estimators to use complete auxiliary information to estimate ratio estimator and derived the estimator of the variance of calibration approach ratio estimators. Also, they showed that the estimator of variance of the combined ratio estimator in stratified sampling using the calibration approach is more efficient than the standard one.

Ranalli (2008), discussed how the most recent developments within the calibration approach may help to use auxiliary information more thoughtfully and, therefore, more efficiently.

Therefore, calibration has established as an important methodological instrument in large-scale production of statistics. Several national statistical agencies have developed software designed to compute weights that usually calibrated to auxiliary information available in administrative registers and other accurate sources.

The main aim of this article is to discuss the case of stratified random sample when only \( X \) (total) is known for optimum and Neyman allocation. The calibration estimator for the parameters has been driven and the result of the simulation study on calibration method has been introduced.

The paper include five sections, in section two, the model of calibration estimation is introduced, section three, the calibration estimators of some classical distances are derived and a new calibration estimators using three new distance measures are also derived with the aid of auxiliary information, in section four, the variance estimator of Horvitz-Thompson will be introduces and a new variance estimator will be suggested. Finally, a simulation study has been conducted to compare different estimators for different sample sizes.

2. Model of Calibration Estimation

Deville and Särndal (1992) introduced a calibration estimator \( Y \), which is constructed as
A new calibration estimator of stratified random sample

\[ \hat{Y}_C = \sum_{i \in s} w_i y_i, \]  

where the calibration weights \( w_i \)'s are chosen to minimize their average distance \( \Phi_s \) from the basic design weights, \( d_i = 1/\pi_i \) ( \( \pi_i \) is the first-order inclusion probability of unit \( i \) in the population) that are used in the Horvitz–Thompson estimator \( \hat{X}_{HT} = \sum_{i \in s} d_i x_i \), subject to the constraint (calibration equation)

\[ \sum_{i \in s} w_i x_i = X \]  

where \( X \) are known population totals of auxiliary variables. The distance measure \( \Phi_s \) is most commonly chosen as

\[ \Phi_s = \sum_{i \in s} \frac{(w_i - d_i)^2}{d_i q_i} \]  

Where, \( q_i \)'s are known positive weights unrelated to \( d_i \). The resulting calibration estimator was:

\[ \hat{Y}_C = \sum_{i \in s} w_i y_i = \hat{Y}_{HT} + (X - \hat{X}_{HT}) \hat{B} \]  

Where,

\[ \hat{X}_{HT} = \sum_{i \in s} d_i x_i \quad \text{and} \quad \hat{B} = [\sum_{i \in s} d_i q_i x_i x'_i]^{-1} \sum_{i \in s} d_i q_i x_i y_i, \]  

The uniform weights \( q_i = 1 \) are used in most applications, but unequal weights can also be motivated. Deville and Särndal (1992).

3. Calibration Estimators for Stratified Sampling (Optimum Allocation)

Suppose the population consists of \( H \) strata with \( N_h \) units in the \( h^{th} \) stratum from which a simple random sample of size \( n_h \) is taken without replacement. Let, total population size be \( N = \sum_{h=1}^{H} N_h \) and sample size be \( n = \sum_{h=1}^{H} n_h \), respectively. Associated with the \( i^{th} \) unit of the \( h^{th} \) stratum there are two values \( y_{hi} \) and \( x_{hi} \) \( (x_{hi} > 0) \) being the covariate which the \( i^{th} \) unit selected from the \( h^{th} \) stratum, where \( i = 1,2,...,n_h \). For the \( h^{th} \) stratum, let \( W_h = N_h / N \) be the stratum weights, \( X = \sum_{h=1}^{H} W_h X_h \) and \( Y = \sum_{h=1}^{H} W_h X_h f = n_h / N_h \) the sample fraction,
Assume, \( x_{hi} \) denote the value of the \( i^{th} \) unit of the auxiliary variable in the \( h^{th} \) stratum for which information known at the unit level or at the stratum level.

### 3.1 Generalized Distance (Chi-Square Distance) (D1)

In some cases it may be necessary to conduct a sample survey with a fixed budget \( c_0 \). But with varying costs of selecting sample units from different strata \( c_h \). The cost function is defined as:

\[
c = \sum_{h=1}^{H} c_h n_h
\]

(5)

Allocating the sample \( n \) among strata where minimize the variance of estimated population mean \( v(\hat{\gamma}_{st}) \) which had the form:

\[
v(\hat{\gamma}_{st}) = \frac{1}{N^2} \sum_{h=1}^{H} \frac{N_h}{N} \frac{N_h - n_h}{n_h} \frac{S^2}{n_h}
\]

(6)

Where \( n_h \) that minimize (6) subject to the linear budget constraint (5) by using Lagrange Multiplier method is

\[
n_h = \frac{N_h S_h / \sqrt{c_h}}{\sum_{h=1}^{H} N_h S_h / \sqrt{c_h}} n
\]

(7)

Now, the weights that will be used in calibration estimator \( \hat{y}_{CST,OA} \) (population total when optimum allocation is used) will be derived as:

\[
\hat{y}_{CST,OA} = \frac{N_h S_h / \sqrt{c_h}}{\sum_{h=1}^{H} N_h S_h / \sqrt{c_h}} \sum_{i=1}^{n_h} w_{hi} y_{hi}
\]

(8)

Where, \( \hat{y}_{CST,OA} \) is the calibration estimator of stratified random sample in case of optimum allocation. So, the chi-square distance will be,

\[
\Phi = \sum_{h=1}^{H} \frac{N_h S_h / \sqrt{c_h}}{\sum_{h=1}^{H} N_h S_h / \sqrt{c_h}} \sum_{i=1}^{n_h} (w_{hi} - d_{hi})^2
\]

(9)
Which subject to the constraint,
\[ \sum_{h=1}^{H} \frac{N_h s_h}{\sqrt{c_h}} \sum_{i=1}^{n_h} w_{hi} x_{hi} = X \] (10)

So, it will be minimized by considering the Lagrange function,
\[
L(\Phi_h; w_{hi}) = \sum_{h=1}^{H} \frac{N_h s_h}{\sqrt{c_h}} \sum_{i=1}^{n_h} (w_{hi} - d_{hi})^2 - \lambda' \sum_{h=1}^{H} \frac{N_h s_h}{\sqrt{c_h}} \sum_{i=1}^{n_h} w_{hi} x_{hi} - X
\]
\[
\frac{\partial L(\Phi_h; w_{hi})}{\partial w_{hi}} = \sum_{h=1}^{H} \frac{2N_h s_h}{\sqrt{c_h}} (w_{hi} - d_{hi}) - \lambda' \sum_{h=1}^{H} \frac{N_h s_h}{\sqrt{c_h}} x_{hi}
\]

Where \( \lambda' \) is a vector of Lagrange factors. By equating the partial derivative to zero to obtain
\[
w_{hi} = d_{hi} (1 + q_{hi} \lambda' x_{hi}) \] (11)

So, by substituting from equation (11) in (10) \Rightarrow
\[
\lambda' = (X - \sum_{h=1}^{H} \frac{N_h s_h}{\sqrt{c_h}} \sum_{i=1}^{n_h} d_{hi} x_{hi}) \sum_{h=1}^{H} \frac{N_h s_h}{\sqrt{c_h}} \sum_{i=1}^{n_h} (d_{hi} q_{hi} x_{hi})^{-1}
\] (12)

And as a result, the weights will have the form,
\[
w_{hi} = [d_{hi} + (X - \sum_{h=1}^{H} \frac{N_h s_h}{\sqrt{c_h}} \sum_{i=1}^{n_h} d_{hi} x_{hi}) \sum_{h=1}^{H} \frac{N_h s_h}{\sqrt{c_h}} \sum_{i=1}^{n_h} (d_{hi} q_{hi} x_{hi})^{-1}] d_{hi} q_{hi} x_{hi} \] (13)

So, by using equation (13), in (8) the calibration estimator of stratified random sample will be obtained.

### 3.2 Multiplicative Distance (D2)

The calibration estimator is defined as above. The multiplicative distance will be,
\[
\Phi_s = \sum_{h=1}^{H} \frac{N_h s_h}{\sqrt{c_h}} \sum_{i=1}^{n_h} (w_{i} \log(\frac{w_{i}}{d_{i}}) - w_{i} + d_{i}) \] (14)

Which subject to the constraint (10), so, it will be minimized by considering the Lagrange function,
\[
L(\Phi, w_{hi}) = \sum_{h=1}^{H} \sum_{k=1}^{n_h} \frac{N_h S_h}{\sqrt{c_h}} \log \left( \frac{w_{hi}}{d_{hi} q_{hi}} \right) - w_{hi} - d_{hi} - \lambda \left( \sum_{h=1}^{H} \sum_{k=1}^{n_h} \frac{N_h S_h}{\sqrt{c_h}} w_{hi} x_{hi} - X \right)
\]

\[
\frac{\partial L}{\partial w_{hi}} = \log \frac{w_{hi}}{d_{hi} q_{hi}} - \lambda' x_{hi},
\]

By solving (15) as above to get,

\[
w_{hi} = d_{hi} q_{hi} e^{\lambda' x_{hi}}
\]

So, from equation (16) in (10) ⇒

\[
\sum_{h=1}^{H} \frac{N_h S_h}{\sqrt{c_h}} \sum_{i=1}^{n_h} d_{hi} q_{hi} e^{\lambda' x_{hi}} x_{hi} = X
\]

For simplicity, the first two terms of McLauren’s series will be used to obtain:

\[
\sum_{h=1}^{H} \frac{N_h S_h}{\sqrt{c_h}} \sum_{i=1}^{n_h} d_{hi} q_{hi} (1 + \lambda x_{hi}) x_{hi} = X
\]

⇒ \(\lambda = \left[ X - \sum_{h=1}^{H} \frac{N_h S_h}{\sqrt{c_h}} \sum_{i=1}^{n_h} d_{hi} q_{hi} x_{hi} \right] \sum_{h=1}^{H} \frac{N_h S_h}{\sqrt{c_h}} \sum_{i=1}^{n_h} d_{hi} q_{hi} x_{hi} \] (17)

Therefore, the weights can be obtained by using equation (17) in (16) and so, the calibration estimator will by obtained for multiplicative distance.

3.3 The Calibration Estimator for New Distances

Three suggested distances (D3, D4 and D5) will be introduced respectively as:
A new calibration estimator of stratified random sample

\[
\Phi_s = \left\{ \begin{array}{l}
\sum_{h=1}^{H} \frac{N_h S_h}{\sqrt{c_h}} \sum_{i=1}^{n_h} \left( \frac{w_{hi} - \sqrt{d_{hi}}}{q_{hi} \rho_h s_h^2} \right)
\sum_{h=1}^{H} \frac{N_h S_h}{\sqrt{c_h}} \sum_{i=1}^{n_h} \left( \frac{-d_{hi} \log(d_{hi}) + w_{hi} - \sqrt{d_{hi}}}{q_{hi} \rho_h s_h^2} \right)
\sum_{h=1}^{H} \frac{N_h S_h}{\sqrt{c_h}} \sum_{i=1}^{n_h} (w_{hi} - \sqrt{d_{hi}})^2 / 2 w_{hi} q_{hi} \rho_h s_h^2
\end{array} \right. \]

Which subject to the same calibration constraint (10), so, by using Lagrange function as above and after simple steps the weights \( w_{hi} \) for the three distances respectively will have the form:

\[
w_{hi} = \left\{ \begin{array}{l}
d_{hi} [1 + q_{hi} \rho_h s_h^2 x_{hi}] (X - \sum_{h=1}^{H} \frac{N_h S_h}{\sqrt{c_h}} \sum_{i=1}^{n_h} d_{hi} x_{hi}) (\sum_{h=1}^{H} \frac{N_h S_h}{\sqrt{c_h}} \sum_{i=1}^{n_h} d_{hi} q_{hi} \rho_h s_h^2 x_{hi} x_{hi})^{-1}
\sum_{h=1}^{H} \frac{N_h S_h}{\sqrt{c_h}} \sum_{i=1}^{n_h} d_{hi} x_{hi}) (\sum_{h=1}^{H} \frac{N_h S_h}{\sqrt{c_h}} \sum_{i=1}^{n_h} d_{hi} q_{hi} \rho_h s_h^2 x_{hi} x_{hi})^{-1}
\sum_{h=1}^{H} \frac{N_h S_h}{\sqrt{c_h}} \sum_{i=1}^{n_h} d_{hi} x_{hi}) (\sum_{h=1}^{H} \frac{N_h S_h}{\sqrt{c_h}} \sum_{i=1}^{n_h} d_{hi} q_{hi} \rho_h s_h^2 x_{hi} x_{hi})^{-1}
\end{array} \right. \]

So, the calibration estimators can be obtained using this equation.

Special Case: Neyman Allocation

Under assumption that \( c_h \) is equal for all strata the form of the weights for all previous distances can be derived and as a result the calibration estimator can be obtained.

4. Variance Estimation

We consider the variance estimator of Horvitz-Thompson (HT) in stratified random sample (STRS) which is the same as the estimator of variance of combined regression estimator that is given by
\[
\text{Var}(\hat{y}_{STRS,HT}) = \sum_{h=1}^{H} \sum_{i=1}^{n_{hi}} \sum_{j=1}^{n_{hj}} \frac{\pi_{hi} - \pi_{hi}\pi_{hj}}{\pi_{hj}} (d_{hi}(y_{hi} - \hat{B}_{hi}x_{hi})) (d_{j}(y_{hj} - \hat{B}_{hj}x_{hj}))
\] (19)

Where, \( \hat{B} \) satisfied the normal equation of \( n_{hi} \) units
\[
(\sum_{h=1}^{H} \sum_{i=1}^{n_{hi}} q_{hi}x_{hi}x_{hi}^T) \hat{B} = \sum_{h=1}^{H} \sum_{i=1}^{n_{hi}} q_{hi}x_{hi}y_{hi}
\]

It is acceptable to use the design weights \( d_{i} \) in the variance estimation but we suggest that the calibration weights \( w_{hi} \) will be used in Equation (19) which makes the variance estimator has a consistent design and nearly unbiased model.

\[
\text{Var}(\hat{y}_{STRS,CAL}) = \sum_{h=1}^{H} \sum_{i=1}^{n_{hi}} \sum_{j=1}^{n_{hj}} \frac{\pi_{hi} - \pi_{hi}\pi_{hj}}{\pi_{hj}} (w_{hi}(y_{hi} - \hat{B}_{hi}x_{hi}))(w_{hj}(y_{hj} - \hat{B}_{hj}x_{hj}))
\]

Moreover, since the calibration estimator is asymptotically equivalent to the generalized regression estimators (GREG), it can be inferred that calibration estimators are more efficient compared to the HT estimator if there is a strong correlation between \( y_{hi} \) and \( x_{hi} \).

**Secondly,** We consider the estimator of variance of HT estimator in STRS as:

\[
\text{Var}(\hat{y}_{STRS,HT}) = W\sum_{h=1}^{H} \sum_{i=1}^{n_{hi}} \sum_{j=1}^{n_{hj}} \left( \frac{y_{hi} - \hat{y}_{hi}}{\pi_{hi} - \pi_{hj}} \right)^2
\] (20)

Where, \( s \) is a fixed sample size.

Instead of \( \pi_{hi} \) (\( \pi_{hj} \)) the calibration weights \( w_{hi} \) will be suggested to be used in (20), so the form of variance estimator will be,

\[
\text{Var}(\hat{y}_{STRS,CJ}) = W\sum_{h=1}^{H} \sum_{i=1}^{n_{hi}} \sum_{j=1}^{n_{hj}} \left( w_{hi}y_{hi} - w_{hi}y_{hj} \right)^2
\]

5- **Simulation Study**

In this section, the performance of the calibration estimator using different distances functions against the HT estimator will be considered.

Monte Carlo simulation will be carried out to investigate the finite sample performance of the estimators of \( \hat{Y}_{STRC} \) that proposed above. A finite population consisting of \( N = 1200 \) units, \( H = 3 \) and \( B = 10000 \) simulation runs in total. For the \( b^{th} \) run \( b = (1,2,\ldots,B) \). The auxiliary variable, \( x_{hi} \) generated as an iid random Gamma sample. The study variable, \( y \), represents a simple regression of the form \( y_{hi} = 2 + x_{hi} + \epsilon \), where \( \epsilon \) distributed as log normal(0,1).

The performance of various estimators were measured by the simulated relative efficiency of standard deviation 1 ( \( RESD1 \)), relative efficiency of
A new calibration estimator of stratified random sample

standard deviation 2 (RESD2), Relative Bias of calibration (RB_c), Relative Bias of HT (RB_HT) and the Weighted Ratio (WR). Each measure is calculated as follows:

\[
RESD1 = \frac{SD(\hat{y}_{STRSCAL})}{SD(\hat{y}_{STRSHT})}, \quad RESD2 = \frac{SD(\hat{y}_{STRSCJ})}{SD(\hat{y}_{STRSHT})},
\]

\[
RB_c = \frac{1}{B} \sum_{b=1}^{B} \frac{\bar{y}_{CST} - \bar{y}_{hi}}{\bar{y}_{hi}}, \quad RB_HT = \frac{1}{B} \sum_{b=1}^{B} \frac{\hat{y}_{HT} - \bar{y}_{hi}}{\bar{y}_{hi}},
\]

\[
WR = \frac{\sum_{b=1}^{B} \text{mean}(w_{CST})}{\sum_{b=1}^{B} \text{mean}(d_{hi})}
\]

Table (1): Performance of Different Distances for Optimum Allocation

<table>
<thead>
<tr>
<th>Sample size</th>
<th>Distances</th>
<th>RESD1</th>
<th>RESD2</th>
<th>RB_c</th>
<th>RB_HT</th>
<th>WR</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>D1</td>
<td>1</td>
<td>1</td>
<td>-0.41701</td>
<td>-0.41701</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>D2</td>
<td>1.0006</td>
<td>1.001082</td>
<td>-0.4791</td>
<td>-0.4791</td>
<td>1.0008</td>
</tr>
<tr>
<td></td>
<td>D3</td>
<td>1</td>
<td>1</td>
<td>-0.5913</td>
<td>-0.5913</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>D4</td>
<td>1</td>
<td>1</td>
<td>-0.579167</td>
<td>-0.579168</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>D5</td>
<td>1.0033</td>
<td>1.0034</td>
<td>-0.5873</td>
<td>-0.5876</td>
<td>1.0004</td>
</tr>
<tr>
<td>60</td>
<td>D1</td>
<td>1</td>
<td>1</td>
<td>-0.4102</td>
<td>-0.4102</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>D2</td>
<td>1.0003</td>
<td>1.0005</td>
<td>-0.5167</td>
<td>-0.5168</td>
<td>1.0005</td>
</tr>
<tr>
<td></td>
<td>D3</td>
<td>1.0005</td>
<td>1.0007</td>
<td>-0.6935</td>
<td>-0.6962</td>
<td>1.0008</td>
</tr>
<tr>
<td></td>
<td>D4</td>
<td>1</td>
<td>1</td>
<td>-0.427466</td>
<td>-0.427466</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>D5</td>
<td>1.0038</td>
<td>1.0007</td>
<td>-0.5534</td>
<td>-0.5538</td>
<td>1.0008</td>
</tr>
<tr>
<td>120</td>
<td>D1</td>
<td>1.000001</td>
<td>1.000001</td>
<td>-0.1934487</td>
<td>-0.1934494</td>
<td>1.000001</td>
</tr>
<tr>
<td></td>
<td>D2</td>
<td>1.00006</td>
<td>1.00053</td>
<td>-0.5968</td>
<td>-0.5969</td>
<td>1.00004</td>
</tr>
<tr>
<td></td>
<td>D3</td>
<td>1</td>
<td>1.0001</td>
<td>-0.6935</td>
<td>-0.6935</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>D4</td>
<td>1</td>
<td>1</td>
<td>-0.0896130</td>
<td>-0.08961342</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>D5</td>
<td>1.0009</td>
<td>1.0015</td>
<td>-0.6099</td>
<td>-0.6104</td>
<td>1.001</td>
</tr>
</tbody>
</table>
According to the results, it can be seen that \( R_{ESD1} \) \(( R_{ESD2} \) and \( WR \) is almost equal to one whatever the distance for all sample sizes while \( RB_c \) and \( RB_{HT} \) are less than zero whatever the distance for all sample sizes.

**Conclusion**

Calibration estimators have been driven for five distances (Three suggested distances have been introduced) according to the problem of auxiliary variable for stratified random sample (optimum allocation (Neyman allocation)). A comparative study has been conducted based on simulation study for different sample sizes and using five criteria to compare between calibration estimators and Horvitz-Thompson estimator. Calibration estimators have a good behavior especially for the suggested distances and that for different sample sizes. Also, it can be seen that using the calibration weights \( w_{hi} \) leaded to the same result as HT estimator.

**References**


*Received: January 14, 2015; Published: March 6, 2015*