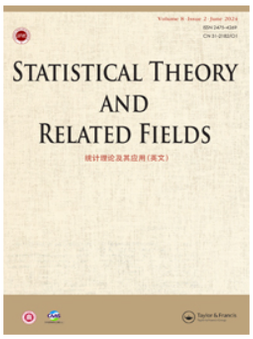

Statistical Theory and Related Fields

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/tstf20

Adjusted variance estimators based on minimizing mean squared error for stratified random samples

Guoyi Zhang & Bruce Swan

To cite this article: Guoyi Zhang & Bruce Swan (2024) Adjusted variance estimators based on minimizing mean squared error for stratified random samples, Statistical Theory and Related Fields, 8:2, 117-123, DOI: 10.1080/24754269.2024.2303915

To link to this article: https://doi.org/10.1080/24754269.2024.2303915

Published online: 16 Jan 2024.

STATISTICAL THEORY AND RELATED FIELDS
2024, VOL. 8, NO. 2, 117–123
https://doi.org/10.1080/24754269.2024.2303915

Adjusted variance estimators based on minimizing mean squared error for stratified random samples

Guoyi Zhang[a] and Bruce Swan[b]

[a]Department of Mathematics and Statistics, University of New Mexico, Albuquerque, NM, USA; [b]Department of Mathematics, SUNY Buffalo State College, Buffalo, NY, USA

ABSTRACT

In the realm of survey data analysis, encountering substantial variance relative to bias is a common occurrence. In this study, we present an innovative strategy to tackle this issue by introducing slightly biased variance estimators. These estimators incorporate a constant c within the range of 0 to 1, which is determined through the minimization of Mean Squared Error (MSE) for $c \times$ (variance estimator). This research builds upon the foundation laid by Kourouklis (2012, A new estimator of the variance based on minimizing mean squared error. *The American Statistician*, *66*(4), 234–236) and extends their work into the domain of survey sampling. Extensive simulation studies are conducted to illustrate the superior performance of the adjusted variance estimators when compared to standard variance estimators, particularly in terms of MSE. These findings underscore the efficacy of our proposed approach in enhancing the precision of variance estimation within the context of survey data analysis.

ARTICLE HISTORY
Received 31 May 2023
Revised 27 November 2023
Accepted 3 January 2024

KEYWORDS
Biased variance estimator; mean squared error; simulations; stratified random sampling; survey data

1. Introduction

Consider a random sample $X_1, X_2, \ldots, X_n$ from a population with distribution function $F \in \mathcal{F}$. Assume that X_i has finite fourth moment. In general, the population variance σ^2 is estimated by the sample variance $s^2 = \sum_{i=1}^{n}(X_i - \overline{X})^2/(n-1)$. Many researchers believe that there are estimators of the form cs^2 (where c is a constant between 0 and 1) that have smaller MSE than the sample variance s^2. These work include Stein (1964), Brown (1968), Brewster and Zidek (1974), Strawderman (1974), Maruyama (1998), Yatracos (2005) and Maruyama and Strawderman (2006). Kourouklis (2012) proposed a variance estimator $c_1 s^2$ and showed that this estimator has the smallest MSE among the estimators of the form cs^2.

In this research, we extend the methodology of Kourouklis (2012) to survey data. Survey data typically spans large geographical areas, resulting in substantial inherent variability. To address this challenge, we have developed adjusted variance estimation techniques that effectively handle the variability commonly associated with survey data. These adjustments result in reduced variance, leading to narrower confidence intervals and enhancing the precision of our estimates.

This research is organized as follows: Section 2 introduces notation in a general survey frame with simple random sample without replacement (SRS) and stratified random sample design; Section 3 proposes the adjusted variance estimator for stratified random samples; Section 4 performs simulation comparisons among the estimators; and Section 5 gives conclusions of the research.

2. Notation

Let $U = \{1, 2, \ldots, N\}$ be the index set of the finite population with size N, and $y_1, y_2, \ldots, y_N$ be the values of the character of the sampling units in the population. Let $\bar{y}_U$ be the population mean: $\bar{y}_U = \sum_{i=1}^{N} y_i/N$, and S^2 be the population variance: $S^2 = \sum_{i=1}^{N}(y_i - \bar{y}_U)^2/(N-1)$. Also let $\mu_2 = \sum_{i=1}^{N}(y_i - \bar{y}_U)^2/N$ and $\mu_4 = \sum_{i=1}^{N}(y_i - \bar{y}_U)^4/N$ be the centralized second and fourth moments respectively. At the sample $\mathcal{S}$ level, let n be the sample size. Sample mean $\bar{y}$ and sample variance s^2 are defined as $\bar{y} = \sum_{i\in\mathcal{S}} y_i/n$, and $s^2 = \sum_{i=1}^{n}(y_i - \bar{y})^2/(n-1)$.

Under an SRS, $E(\bar{y}) = \bar{y}_U$, and

$$\mathrm{Var}(\bar{y}) = \left(1 - \frac{n}{N}\right)\frac{S^2}{n}, \tag{1}$$

where $(1 - n/N)$ is called the finite population correction coefficient.

CONTACT Guoyi Zhang gzhang123@gmail.com Department of Mathematics and Statistics, University of New Mexico, Albuquerque, NM 87131-0001, USA

This article has been corrected with minor changes. These changes do not impact the academic content of the article.

In a stratified random sample, population with size N is divided into H non-overlapping strata with size $N_h, h = 1, 2, \ldots, H$, such that $N = \sum_{h=1}^{H} N_h$. Let y_{hj} be the value of the character for the jth sampling unit within stratum h. Let $\bar{y}_{hU}$ be the population mean of stratum h with $\bar{y}_{hU} = \sum_{j=1}^{N_h} y_{hj}/N_h$, and S_h^2 be the population variance with $S_h^2 = \sum_{j=1}^{N_h}(y_{hj} - \bar{y}_{hU})^2/(N_h - 1)$. Population mean $\bar{y}_U$ can also be written as a weighted average of the stratum means such as $\bar{y}_U = \sum_{h=1}^{H} N_h \bar{y}_{hU}/N$.

Within each stratum h, an SRS with size n_h is taken independently. Assume that $n_h \geq 2$ throughout the paper, and $\sum_{h=1}^{H} n_h = n$. Let $\mathcal{S}_h$ be the set of n_h units in the SRS within stratum h. Stratum sample mean $\bar{y}_h$ and sample variance s_h^2 are defined as $\bar{y}_h = \sum_{j\in\mathcal{S}_h} y_{hj}/n_h$ and $s_h^2 = \sum_{j\in\mathcal{S}_h}(y_{hj} - \bar{y}_h)^2/(n_h - 1)$. An unbiased estimator of the population mean $\bar{y}_U$ is

$$\bar{y}_{\text{str}} = \sum_{h=1}^{H} \frac{N_h \bar{y}_h}{N}. \tag{2}$$

By Equation (1) and independent sampling within each stratum, variance of $\bar{y}_{\text{str}}$ is

$$\text{Var}(\bar{y}_{\text{str}}) = \sum_{h=1}^{H} \left(1 - \frac{n_h}{N_h}\right)\left(\frac{N_h}{N}\right)^2 \frac{S_h^2}{n_h}, \tag{3}$$

and is estimated by

$$\widehat{\text{Var}}(\bar{y}_{\text{str}}) = \sum_{h=1}^{H} \left(1 - \frac{n_h}{N_h}\right)\left(\frac{N_h}{N}\right)^2 \frac{s_h^2}{n_h}. \tag{4}$$

3. Proposed adjusted variance estimator in a stratified random sample

Estimating population mean and total are two main topics in survey sampling. In this section, we first propose an adjusted variance estimator of the mean that minimizes MSE under an SRS setting. Next, we extend the estimator to a stratified random sample. Last, we discuss how to estimate the optimal value c in practice.

3.1. Lemma and theorem

In an SRS, we adjust the sample variance s^2 by $cs^2, 0 < c < 1$, where c is determined by minimizing the MSE of cs^2. This is equivalent to minimize MSE of $\widehat{\text{Var}}(\bar{y})$. We state the result as the following lemma and give a brief proof.

Lemma 3.1: *For a size n SRS selected from a population with size N, the optimal value c that minimizes* $\text{MSE}(cs^2)$ *is*

$$c_{\text{srs}} = S^4/E(s^4), \tag{5}$$

where

$$E(s^4) = \frac{n^2}{(n-1)^2}(aN\mu_4 + bN^2\mu_2^2), \tag{6}$$

with

$$a = \frac{e_1 - e_2}{n^2} - \frac{2(e_1 - 3e_2 + 2e_3)}{n^3} + \frac{e_1 - 7e_2 + 12e_3 - 6e_4}{n^4},$$

$$b = \frac{e_2}{n^2} - \frac{2(e_2 - e_3)}{n^3} + \frac{3(e_2 - 2e_3 + e_4)}{n^4},$$

$$e_1 = n/N, \quad e_2 = \frac{n(n-1)}{N(N-1)}, \quad e_3 = \frac{n(n-1)(n-2)}{N(N-1)(N-2)}, \quad e_4 = \frac{n(n-1)(n-2)(n-3)}{N(N-1)(N-2)(N-3)},$$

and μ_4 *and* μ_2 *are the centralized moments defined in Section 2.*

Proof:

$$\begin{aligned}\text{MSE}(cs^2) &= E(cs^2 - S^2)^2 \\ &= E(c^2s^4) - 2E(cs^2S^2) + S^4 \\ &= c^2E(s^4) - 2cS^4 + S^4.\end{aligned}$$

Let $g(c) = c^2E(s^4) - 2cS^4 + S^4$. By setting $g'(c) = 0$, and using the fact that $g''(c) = 2E(s^4) > 0$, the optimal value of c that minimizes $\text{MSE}(cs^2)$ is $c_{\text{srs}} = S^4/E(s^4)$.

The remaining challenge is to calculate $E(s^4)$ under an SRS. Utilizing the formulas for $E(s^2)$ from Sukhatme (1984, p. 28) and $V(s^2)$ from Sukhatme (1984, p. 36), we can determine $E(s^4)$ as shown in Equation (6). The computation of the sampling variance $V(s^2)$ is intricate, primarily due to its fourth-order term. Sukhatme's approach benefits significantly from the use of partitional notation and monomial symmetric functions. Readers interested in a comprehensive understanding of this derivation can refer to Sukhatme's book for detailed insights. ■

We extend the adjusted variance estimator $c_{\text{srs}}s^2$ from SRS to stratified random sampling. One direct extension is to consider

$$\widehat{\text{Var}}_2(\bar{y}_{\text{str}}) = \sum_{h=1}^{H} c_h \left(1 - \frac{n_h}{N_h}\right)\left(\frac{N_h}{N}\right)^2 \frac{s_h^2}{n_h}, \tag{7}$$

where c_h represents a constant optimized within stratum h under SRS calculated as

$$c_h = S_h^4 / E(s_h^4), \tag{8}$$

where S_h^2 and s_h^2 are defined in Section 2.

Alternatively, we can consider a unified constant adjustment c_{str}, which aims to balance bias and variance across all strata collectively. The derivation of c_{str} is presented in the following theorem.

Theorem 3.1: *In a stratified random sample, population is divided into H non-overlapping strata, and an SRS is taken independently from each stratum. Let N_h and n_h be the population and sample size, and S_h^2 and s_h^2 be the population and sample variance within stratum h as defined in Section 2. The optimal value of c that minimizes* $\text{MSE}(c\widehat{\text{Var}}(\bar{y}_{\text{str}}))$ *is*

$$c_{\text{str}} = \frac{\left(\sum_{h=1}^{H} k_h S_h^2\right)^2}{\sum_{h=1}^{H} k_h^2 E(s_h^4) + \sum_{i=1}^{H}\sum_{j=1,j\neq i}^{H} k_i k_j S_i^2 S_j^2}, \tag{9}$$

where $E(s_h^4)$ can be derived by Equation (6) *by taking an SRS of size n_h from stratum h, and $k_h = (1 - n_h/N_h)(N_h/N)^2/n_h$.*

Proof: By Equation (2), $\bar{y}_{\text{str}} = \sum_{h=1}^{H} N_h \bar{y}_h / N$. Recall Equation (4) gives estimator of $\text{Var}(\bar{y}_{\text{str}})$ as

$$\widehat{\text{Var}}(\bar{y}_{\text{str}}) = \sum_{h=1}^{H} \left(1 - \frac{n_h}{N_h}\right)\left(\frac{N_h}{N}\right)^2 \frac{s_h^2}{n_h},$$

which can be written as $\widehat{\text{Var}}(\bar{y}_{\text{str}}) = \sum_{h=1}^{H} k_h s_h^2$ with expected value of $\sum_{h=1}^{H} k_h S_h^2$. Now we want to find a constant c such that MSE of $c\widehat{\text{Var}}(\bar{y}_{\text{str}})$ reaches the minimum. After some algebra,

$$\begin{aligned}
&E\left\{(c\widehat{\text{Var}}(\bar{y}_{\text{str}}) - \text{Var}(\bar{y}_{\text{str}}))^2\right\} \\
&= E\left\{\left(\sum_{h=1}^{H} k_h c s_h^2 - \sum_{h=1}^{H} k_h S_h^2\right)^2\right\} \\
&= \sum_{h=1}^{H} k_h^2 (c^2 E(s_h^4) - 2cS_h^4 + S_h^4) + \sum_{i=1}^{H}\sum_{j=1,j\neq i}^{H} k_i k_j (c-1)^2 S_i^2 S_j^2 \\
&= h(c).
\end{aligned}$$

Setting $h'(c) = 0$, the local extreme value is obtained at

$$c_{\text{str}} = \frac{\left(\sum_{h=1}^{H} k_h S_h^2\right)^2}{\sum_{h=1}^{H} k_h^2 E(s_h^4) + \sum_{i=1}^{H}\sum_{j=1,j\neq i}^{H} k_i k_j S_i^2 S_j^2}, \tag{10}$$

where $E(s_h^4)$ can be derived by Equation (6). Notice that $h''(c) = \sum_{h=1}^{H} 2k_h^2 E(s_h^4) + \sum_{i=1}^{H}\sum_{j=1,j\neq i}^{H} k_i k_j S_i^2 S_j^2 > 0$. c_{str} is the optimal value of c that minimizes $\text{MSE}(c\widehat{\text{Var}}(\bar{y}_{\text{str}}))$. ■

3.2. Estimating c_{srs}, c_h and c_{str}

In practice, the constant c needs to be estimated using a larger survey or using sample information. We can use $(n-1)s^2/n$ to estimate μ_2. But estimating the fourth moment μ_4 is challenging. Some recent estimators of the fourth moment are not unbiased, or are based on h-statistics and U-statistics (Heffernan, 1997), which can be computationally expensive. Espejo et al. (2013) proposed estimating the fourth population central moment under distribution-free setting, which involves variance and covariance among the lower sample moments.

Most practitioners may not have the mathematical and statistical background to understand or use the general estimators given in literature. Assume that an SRS or a stratified random sample is with large size, and that the selected sample is representative of the finite population and estimation bias is small. We then use the fourth sample moment and plugin method to estimate the optimal values of c_{srs} and c_{str} as follows.

$$\hat{c}_{srs} = s^4/\widehat{E}(s^4), \tag{11}$$

where

$$\widehat{E}(s^4) = \frac{n^2}{(n-1)^2}(aN\hat{\mu}_4 + bN^2\hat{\mu}_2^2), \tag{12}$$

where $\hat{\mu}_4 = \sum_{i=1}^n (y_i - \bar{y})^4/n$, $\hat{\mu}_2 = \sum_{i=1}^n (y_i - \bar{y})^2/n$, and a and b are defined as in Section 3.

Similarly, extending SRS to a stratified random sample, we have

$$\hat{c}_h = s_h^4/\widehat{E}(s_h^4), \tag{13}$$

and

$$\hat{c}_{str} = \frac{\left(\sum_{h=1}^H k_h s_h^2\right)^2}{\sum_{h=1}^H k_h^2 \widehat{E}(s_h^4) + \sum_{i=1}^H \sum_{j=1, j\neq i}^H k_i k_j s_i^2 s_j^2}, \tag{14}$$

where $\widehat{E}(s_h^4)$ can be derived using Equation (12) when an SRS of size n_h is taken from stratum h, and $k_h = (1 - n_h/N_h)(N_h/N)^2/n_h$.

4. Simulation studies

In this section, we conduct a simulation study to assess the performance of the proposed adjusted variance estimator. The constant c can be determined using population data from `agpop.csv` through Equations (5), (8) and (10). Alternatively, it can be estimated from samples using Equations (11), (13) and (14). We evaluate the bias, variance and mean squared error (MSE) of the adjusted variance estimators under two scenarios: simple random samples (SRS) and stratified random samples. The adjusted variance estimators are as follows.

- Estimator 2: $c \cdot$ variance estimator (using the population constant c),
- Estimator 3: $\hat{c} \cdot$ variance estimator (using the estimated constant $\hat{c}$),
- Estimator 4: stratum-specific adjustment estimator (7), where c_h is estimated as $\hat{c}_h$ using Equation (13), i.e.,

$$\widehat{\text{Var}}_3(\bar{y}_{str}) = \sum_{h=1}^H \hat{c}_h \left(1 - \frac{n_h}{N_h}\right)\left(\frac{N_h}{N}\right)^2 \frac{s_h^2}{n_h}, \tag{15}$$

is compared with the unadjusted variance estimator (Estimator 1) across both sampling scenarios.

4.1. Simulation set up

The population data used in our simulation study is sourced from `agpop.csv`, which is available in the supplementary material of the textbook by Lohr (2010). The U.S. government conducts a census of agriculture every five years, collecting data on all farms in the 50 states. The census of agriculture provides data `agpop.csv` on number of farms, total acreage devoted to farms (*acres92* is the total acreage devoted to farms in 1992 and is the variable of interest in the study), farm size, yield of different crops, and a wide variety of other agriculture measures for $N = 3078$ counties and county-equivalents in the United States. These 3078 counties are divided into four regions (strata) with stratum size N_h: North Central (NC, stratum 1, $N_1 = 1054$), North East (NE, stratum 2, $N_2 = 220$), South (S, stratum 3, $N_3 = 1382$) and West (W, stratum 4, $N_4 = 422$).

Table 1. Simulation Results under SRS and stratified random sample settings (variable of interest is *acres92*).

Sampling method	SRS			Stratified random sampling			
Estimator	1	2	3	1	2	3	4
c	N/A	0.8606	N/A	N/A	0.8538	N/A	N/A
$\hat{c}$	N/A	N/A	0.9196	N/A	N/A	0.9440	N/A
Bias	6.876e+03	−7.564e+07	−5.315e+07	1.1490e+06	−6.4230e+07	−3.2259e+07	−4.9763e+07
Variance	4.731e+16	3.504e+16	2.874e+16	3.4425e+16	2.5098e+16	2.2128e+16	1.9392e+16
MSE	4.731e+16	4.076e+16	3.156e+16	3.4426e+16	2.9224e+16	2.3169e+16	2.1868e+16

Note: Estimators 1, 2 and 3 are variance estimators of mean that are unadjusted, adjusted by a constant c_{srs} for an SRS or by c_{str} for a stratified sample, and adjusted by $\hat{c}_{srs}$ and $\hat{c}_{str}$ respectively. Estimator 4 is the domain specific estimate (15) using $\hat{c}_h$ adjustment for a stratified sample.

Simulation does $L = 100000$ times for each setting. Each time, we draw a sample from the population data `agpop.csv` using either SRS with sample size $n = 300$ or stratified proportional allocated random sample with $(n_1, n_2, n_3, n_4) = (103, 21, 135, 41)$. In a general notation, let $\hat{\theta}$ be an estimator of θ. Assume $\hat{\theta}^{(i)}$ represents the estimator of θ from the ith sample, $i = 1, \ldots, L$. The Monte Carlo mean E_{MC}, Monte Carlo bias B_{MC}, Monte Carlo variance V_{MC} and Monte Carlo MSE are given by the following formulas

$$E_{\text{MC}}\{\hat{\theta}\} = L^{-1}\sum_{i=1}^{L}\hat{\theta}^{(i)}, \tag{16}$$

$$B_{\text{MC}}\{\hat{\theta}\} = E_{\text{MC}}\{\hat{\theta}\} - \theta, \tag{17}$$

$$V_{\text{MC}}\{\hat{\theta}\} = L^{-1}\sum_{m=1}^{L}[\hat{\theta}^{(i)} - E_{\text{MC}}\{\hat{\theta}\}]^2, \tag{18}$$

and the main criterion for determining efficiency: Monte Carlo MSE is defined by

$$\text{MSE}_{\text{MC}}\{\hat{\theta}\} = L^{-1}\sum_{i=1}^{L}\{\hat{\theta}^{(i)} - \theta\}^2. \tag{19}$$

True mean $\bar{y}_U$ is the average of y_i's from the population. For SRS, true variance of $\bar{y}$ is calculated by $\text{Var}(\bar{y}) = (1 - n/N)S^2/n$ (Equation (1)). For a stratified random sample, variance of $\bar{y}_{\text{str}}$ is $\text{Var}(\bar{y}_{\text{str}}) = \sum_{h=1}^{H} k_h S_h^2$ (Equation (3)). The unadjusted variance estimators of $\text{Var}(\bar{y})$ and $\text{Var}(\bar{y}_{\text{str}})$ from the ith sample are $\widehat{\text{Var}}^{(i)}(\bar{y}) = (1 - n/N)s^2/n$ and $\widehat{\text{Var}}^{(i)}(\bar{y}_{\text{str}}) = \sum_{h=1}^{H} k_h s_h^2$ respectively.

Optimal values of c_{srs}, c_h and c_{str} are calculated by Equations (5), (8) and (10) using population data `agpop.csv`. $\hat{c}_{\text{srs}}$ and $\hat{c}_{\text{str}}$ are estimated by averages of the L estimates $\hat{c}_{\text{srs}}^{(i)}$ and $\hat{c}_{\text{str}}^{(i)}$ from the ith sample using Equations (11) and (14). The adjusted variance estimates of $\text{Var}(\bar{y})$ for Estimators 2 and 3 from the ith sample are $c_{\text{srs}}\widehat{\text{Var}}^{(i)}(\bar{y})$ and $\hat{c}_{\text{srs}}^{(i)}\widehat{\text{Var}}^{(i)}(\bar{y})$. The adjusted variance estimates of $\text{Var}(\bar{y}_{\text{str}})$ for Estimators 2, 3 and 4 from the ith sample are $c_{\text{str}}\widehat{\text{Var}}^{(i)}(\bar{y}_{\text{str}})$, $\hat{c}_{\text{str}}^{(i)}\widehat{\text{Var}}^{(i)}(\bar{y}_{\text{str}})$ and $\widehat{\text{Var}}_3^{(i)}(\bar{y}_{\text{str}})$ respectively.

4.2. Simulation results

Table 1 gives simulation results under SRS and stratified random sampling settings. Bias, variance and MSE are calculated by Equations (16) –(19). Note that using population data, we have $V(\bar{y}) = 542599828$ and $\text{Var}(\bar{y}_{\text{str}}) = 446220740$. Based on this large scale, bias, variance and MSE of $\widehat{\text{Var}}(\bar{y})$ are all huge.

Table 1 shows that under SRS and stratified random samples, (1) biases of Estimators 2, 3 and 4 are all larger than that of Estimator 1, since Estimator 1 is unbiased; (2) the trade-off of biased estimators are smaller variance of Estimators 2, 3 and 4 compared to that of Estimator 1; (3) the overall measurement MSE of Estimators 2, 3 and 4 are both smaller than that of Estimator 1. For example, under SRS, the percentage of MSE reduction by Estimator 2 (defined as [MSE of Estimator 1 − MSE of Estimator 2]/MSE of Estimator 1) is $(4.731\text{e}+16 - 4.076\text{e}+16)/(4.731\text{e}+16) = 13.8\%$, and percentage of MSE reduction by Estimator 3 (defined as [MSE of Estimator 1 − MSE of Estimator 3]/MSE of Estimator 1) is $(4.731\text{e}+16 - 3.156\text{e}+16)/(4.731\text{e}+16) = 33.3\%$. For stratified random sample, the percentage of MSE reduction by Estimator 2 is 15.1%, by Estimator 3 is 32.7% and by Estimator 4 is 36.48%.

Upon closer examination of Estimators 3 and 4 for stratified samples, we note that Estimator 4 exhibits a somewhat larger bias compared to Estimator 3. However, it also displays a smaller variance and mean squared error (MSE). While there are discernible differences between these two estimators, they are not statistically significant.

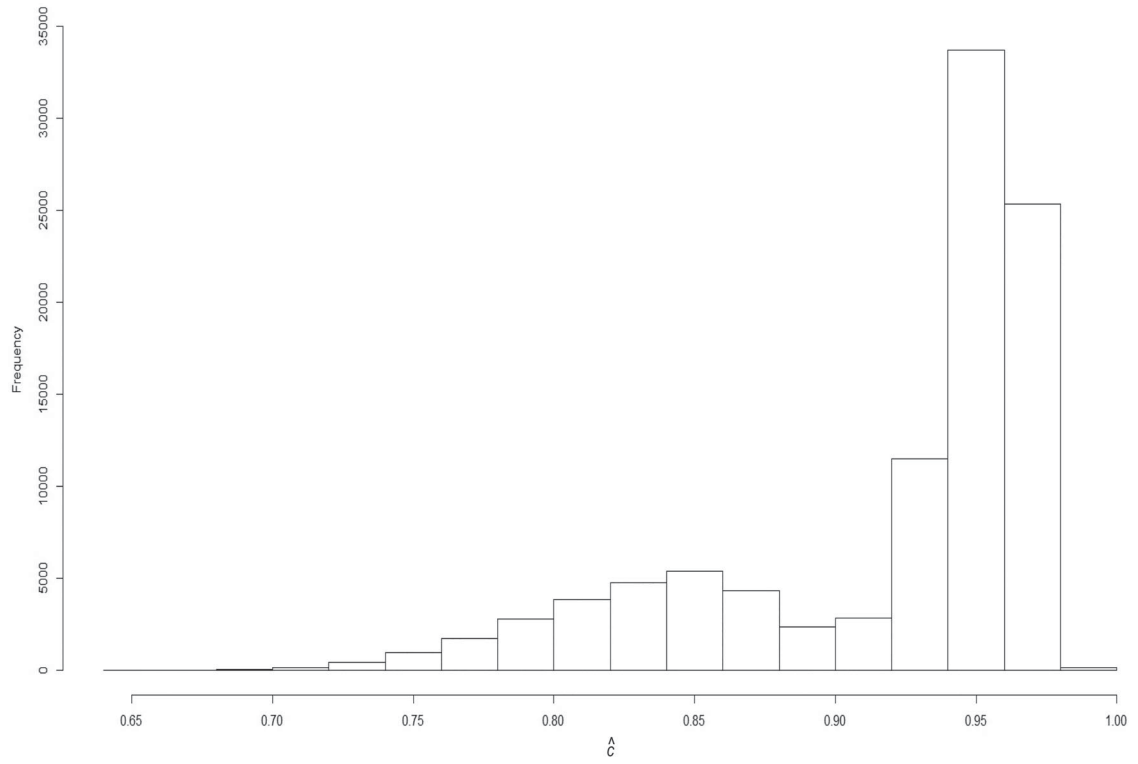

Figure 1. Histogram plot of $\hat{c}_{srs}$ from the 100000 simulations.

Estimator 4, which addresses each stratum individually, may offer certain advantages in specific cases. Nevertheless, based on the results of our simulation study, it is evident that Estimator 3, utilizing a common adjustment factor, strikes a reasonable balance between bias and variance. Therefore, for its simplicity of implementation, we recommend adopting Estimator 3.

Now, let's examine Estimator 3 more closely. In the case of SRS, we have $c = 0.8606$, while $\hat{c}_{srs} = 0.9196$ with a standard error of 0.0592. For stratified SRS, $c_{str} = 0.8538$, and $\hat{c}_{str} = 0.9440$ with a standard error of 0.0588. It's worth noting that all bias, variance and mean squared error (MSE) values of Estimator 3 are consistently smaller than those of Estimator 2 under the same settings.

In Figure 1, we present the histogram of $\hat{c}$ derived from the results of $L = 100000$ simulations conducted under the SRS setting. This bimodal histogram exhibits one peak at 0.85 and another at 0.95, ultimately yielding an estimate of $\hat{c}_{srs} = 0.9196$.

Figure 2 shows the sample variance $s^{2(i)}$ versus $\hat{c}_{srs}^{(i)}$ from the ith simulation. Unlike Estimator 2 with a constant adjustment c, $\hat{c}$ seems like a dynamic adjustment with large $\hat{c}$ associated with small s^2 and small $\hat{c}$ associated with large s^2. This makes $\hat{c}s^{2(i)}$ tend to get closer to the true value S^2 and to get closer to each other. Therefore, bias, variance and MSE of Estimator 3 are smaller than those of Estimator 2.

5. Conclusions and future study

In this research, we extended Kourouklis (2012)'s work to encompass simple random samples (SRS) and stratified random samples. Theoretically, the proposed variance Estimator 2, adjusted by c_{srs} for SRS and c_{str} for stratified samples, demonstrates the smallest Mean Squared Error (MSE) among estimators of the form $c \times$ (variance estimator). In practice, we employ sample statistics to estimate the constant c and introduce Estimator 3, which is adjusted by $\hat{c}_{srs}$ or $\hat{c}_{str}$. Simulation studies consistently show that both Estimators 2 and 3 yield lower overall MSE compared to Estimator 1 (the unadjusted estimator). Notably, $\hat{c}$ behaves as a dynamic adjustment factor, with larger or smaller $\hat{c}$ values corresponding to smaller or larger variance estimates. Consequently, Estimator 3 exhibits smaller bias, variance and MSE than Estimator 2.

The unified constant adjustment c_{str} is designed to strike a balance between bias and variance across all strata collectively. In contrast, Estimator 4 employs stratum-specific optimal constant c_h values for each associated stratum h under SRS, presenting another valid option. Our simulation study indicates that the stratum-specific c_h method offers certain advantages by individually handling each stratum. However, these advantages are not statistically significant.

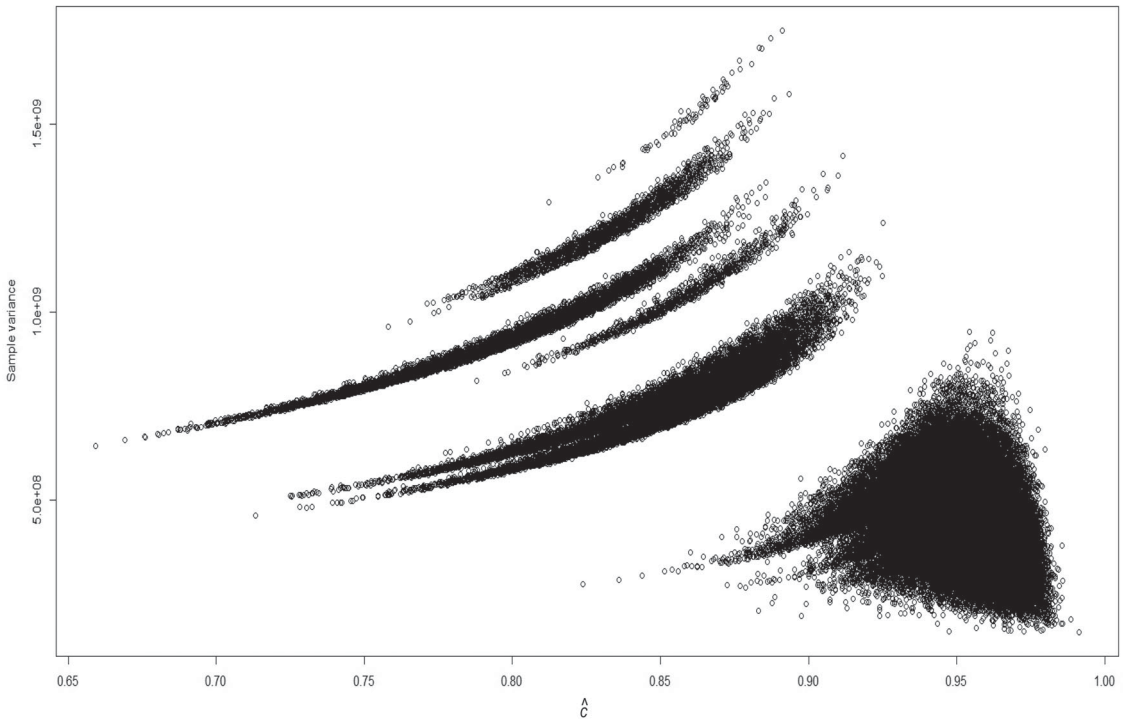

Figure 2. Sample variance versus $\hat{c}_{srs}$ from the 100000 simulations.

In practical applications, we recommend using Estimator 3 to adjust variance estimators of the mean and total in both SRS and stratified random samples, as it consistently yields narrower confidence intervals. Future research avenues may explore the extension of adjusted variance estimators to complex survey designs, such as two-stage stratified cluster surveys.

Acknowledgements

The authors gratefully thank to the Referee for the constructive comments and recommendations which definitely help to improve the readability and quality of the paper.

Disclosure statement

No potential conflict of interest was reported by the author(s).

References

Brewster, J. F., & Zidek, J. V. (1974). Improving on equivariant estimators. *The Annals of Statistics*, *2(1)*, 21–38.
Brown, L. D. (1968). Inadmissibility of the usual estimators of scale parameters in problems with unknown location and scale parameters. *Annals of Mathematical Statistics*, *39*(1), 29–48. https://doi.org/10.1214/aoms/1177698503
Espejo, M., Pineda, M., & Nadarajah, S. (2013). Optimal unbiased estimation of some population central moments. *Metron*, *71*(1), 39–62. https://doi.org/10.1007/s40300-013-0006-z
Heffernan, P. (1997). Unbiased estimation of central moments by using U-statistics. *Journal of the Royal Statistical Society Series B*, *59*(4), 861–863. https://doi.org/10.1111/1467-9868.00102
Kourouklis, S. (2012). A new estimator of the variance based on minimizing mean squared error. *The American Statistician*, *66*(4), 234–236. https://doi.org/10.1080/00031305.2012.735209
Lohr, S. (2010). *Sampling: Design and Analysis* (2nd ed.). New York: CRC Press.
Maruyama, Y. (1998). Minimax estimators of a normal variance. *Metrika*, *48*(3), 209–214. https://doi.org/10.1007/PL00003974
Maruyama, Y., & Strawderman, W. E. (2006). A new class of minimax generalized Bayes estimators of a normal variance. *Journal of Statistical Planning and Inference*, *136*(11), 3822–3836. https://doi.org/10.1016/j.jspi.2005.05.005
Stein, C. (1964). Inadmissibility of the usual estimator for the variance of a normal distribution with unknown mean. *Annals of the Institute of Statistical Mathematics*, *16*(1), 155–160. https://doi.org/10.1007/BF02868569
Strawderman, W. E. (1974). Minimax estimation of powers of the variance of a normal population under squared error loss. *The Annals of Statistics*, *2*(1), 190–198. https://doi.org/10.1214/aos/1176342625
Sukhatme, P. V. (1984). *Sampling Theory of Surveys with Applications*. Iowa: The Iowa State College Press.
Yatracos, Y. (2005). Artificially augmented samples, shrinkage, and mean squared error reduction. *Journal of the American Statistical Association*, *100*(472), 1168–1175. https://doi.org/10.1198/016214505000000321