Revista Colombiana de Estadística - Theoretical Statistics
January 2026, volume 49, issue 1, pp. 77–88
http://doi.org/10.15446/rce.v49n1.115571

A New Modified Post-Stratified Estimator in Finite Population Sampling

Un nuevo estimador post-estratificado modificado en muestreo de poblaciones finitas

Mostafa Hossaini[1,a], Abdolhamid Rezaei Roknabady[1,b], Hussain Ali Abbdllah Aleagobe[2,c]

[1]Department of Statistics, Ferdowsi University of Mashhad, Mashhad, Iran
[2]Department of Economics, University of Thi-Qar, Nasiriyah, Iraq

Abstract

We know that post-stratification sampling is applied in a situation where it is not possible to determine the general framework of each of the categories in population before the selection sample. In this paper, we first use a method for comparing conventional estimators in the subpopulation, presented by Salehi & Seber (2021), and then introduce a new estimator in the post-stratification sampling scheme. We show that this estimator is unbiased and more precise than the estimators in simple random sampling. We have conducted a simulation study to evaluate the performance of the proposed estimator. The simulation results confirmed the theoretical achievements of the article. The data used in this article is a census of agriculture conducted by the US government every five years from all 50 states.

Keywords: Finite population; Post-stratification; Stratified sampling; Subpopulation.

Resumen

Se sabe que el muestreo por post-estratificación se aplica en situaciones en las que no es posible determinar, antes de seleccionar la muestra, el marco general de cada una de las categorías de la población. En este artículo, primero se utiliza un método para comparar estimadores convencionales en la subpoblación, presentado por Salehi & Seber (2021), y luego se introduce un nuevo estimador dentro del esquema de muestreo por post-estratificación. Se demuestra que este estimador es insesgado y más preciso que los estimadores

[a]Ph.D Student. E-mail: mo-hossaini@um.ac.ir
[b]Ph.D. E-mail: rezaei@um.ac.ir
[c]Ph.D. E-mail: hussain.phd@utq.edu.iq

bajo muestreo aleatorio simple. Además, se realiza un estudio de simulación para evaluar el desempeño del estimador propuesto. Los resultados de la simulación confirmaron los logros teóricos del artículo. Los datos utilizados corresponden a un censo agrícola que el gobierno de Estados Unidos realiza cada cinco años en los 50 estados.

***Palabras clave*:** Poblaciones finitas; Post-estratificación; Muestreo estratificado; Subpoblaciones.

1. Introduction

Sometimes we extract a random sample from a finite population with the aim of estimating population parameters such as the mean or the total value. After completing the sampling process, we may be interested in using the same general sample to estimate parameters of a specific subset of the population, referred to as a subpopulation. In this regard, Cochran (1977) presented two different estimators for cases in which the size of the subpopulation is either known or unknown. Salehi & Seber (2021) compared these estimators under different circumstances and presented a strategy for choosing the appropriate one. In addition to Cochran (1977), Thompson (2012) also referred to subpopulation estimation. Hossaini & Rezaei (2021) discussed estimation of subpopulation parameters in one-stage cluster sampling for cases where the subpopulation size is known or unknown. Clark (2009) introduced a regression estimator for subpopulation parameters, and Salehi & Chang (2005) proposed another estimator for the total subpopulation parameter based on reverse sampling.

In some cases, even though stratification of the population is not possible due to the lack of a framework of classes, we may still want to estimate parameters in a desired stratum as well as in the population as a whole. In such situations, post-stratification is recommended. In this paper, we provide an unbiased estimator that performs better than existing estimators by using optimal selection from subpopulations in the post-stratification sampling method. We also provide a step-by-step algorithm for computing both the post-stratified and modified post-stratified estimators, which allows readers to apply the methods in practice. Further discussions of post-stratification and subpopulation estimation can be found in Cochran (1977), Holt & Smith (1979), Jagers et al. (1985), Singh & Chaudhary (1986), Hedayat & Sinha (1991), Levy & Lemeshow (1991), Little (1993), and Thompson (2012).

For clarity, we briefly define the main concepts used throughout the paper. A finite population refers to a set of N distinct sampling units from which observations are drawn. A random sample from a finite population is a subset selected according to a sampling design that specifies selection probabilities for units. A stratum (plural: strata) denotes a homogeneous subgroup into which the population is partitioned in stratified sampling; sampling is then performed independently within each stratum. An estimator is a rule or formula that produces an estimate of a population parameter (such as a mean or total) from sample data. Where relevant, we also refer to inclusion probabilities—the probabilities that given units

are included in the sample under the chosen design—and to standard variance estimators used to quantify estimator precision.

The remainder of this article is organized as follows. Section 2 presents the estimation of subpopulation parameters and reviews related estimators. Section 3 introduces the proposed post-stratified estimator and discusses its theoretical properties. Section 4 describes a simulation study conducted to evaluate the performance of the proposed estimator under several distributions. Section 5 provides an application to real data, including a step-by-step algorithm for practical implementation. Finally, Section 6 concludes the paper with a summary of findings, a critical discussion of the estimator's properties, and possible directions for future research.

2. Estimation of Subpopulation Parameters

Suppose that the units of the finite population are U_1, U_2, ..., U_N, that Y_i is the value of the attribute for unit i, and that the parameters $Y = \sum_{i=1}^{N} Y_i$ and $\overline{Y} = \dfrac{Y}{N}$ represent the total and mean of population, respectively. Now suppose that $u_1, u_2, \ldots, u_n$ are sample members and that random variables y_i are the value of the attribute for the unit i in the sample.

We consider the c as a special property that some members of the population have, and we call the set of such members subpopulation. After simple random sampling, in addition to estimating the parameters of the population, we are interested in estimating subpopulation parameters by the same random sample. For this purpose, we assume that N_c is subpopulation size and that Y_{ci}, for $i = 1, \ldots, N_c$, is the value of the attribute for the unit i in subpopulation, in which case $Y_c = \sum_{i=1}^{N_c} Y_{ci}$ and $\overline{Y}_c = \dfrac{Y_c}{N_c}$ also represent the total and mean of subpopulation, respectively. Also, assume that, for $j = 1, \ldots, n_c$, the random variable y_{cj} represents the value of attribute for the unit j in the sample that is located in the subpopulation c, where n_c is the sample size located in the subpopulation. Obviously n_c is a random variable with $E(n_c) = \dfrac{nN_c}{N}$.

To estimate Y_c, Cochran (1977) introduced the following estimators:

$$\widehat{Y}_c = N_c \bar{y}_c, \tag{1}$$

$$\widehat{\widehat{Y}}_c = \frac{N n_c \bar{y}_c}{n}. \tag{2}$$

In which $\bar{y}_c = \frac{1}{n_c} \sum_{j=1}^{n_c} y_{cj}$ is the sample mean in the subpopulation. Similarly, for the estimate $\overline{Y}_c$, we have

$$\widehat{\overline{Y}}_c = \frac{\widehat{Y}_c}{N_c} = \overline{y}_c,$$

$$\widehat{\widehat{\overline{Y}}}_c = \frac{\widehat{\widehat{Y}}_c}{N_c}.$$

But $\widehat{\overline{Y}}_c$ is an estimate of $\overline{Y}_c$ if N_c is known. Sometimes, N_c is unknown for some populations, so considering that $\widehat{N}_c = \dfrac{Nn_c}{n}$ is an unbiased estimator for N_c, the following estimator can be used to estimate $\overline{Y}_c$:

$$\widehat{\overline{Y}}_c = \frac{Nn_c\bar{y}_c}{\widehat{N}_c n} = \bar{y}_c.$$

It is worth noting that, in accordance with this estimator, to estimate the total subpopulation, $\widehat{Y}_c = N_c\bar{y}_c$ and also it is an only estimator from Y_c, if N_c is known. In cases where N_c is unknown, the estimator

$$\widehat{N}_c\bar{y}_c = \frac{Nn_c}{n}\bar{y}_c = \widehat{\widehat{Y}}_c$$

is obtained, which is the Estimator (2).

It is proved that all estimates $\widehat{Y}_c$, $\widehat{\widehat{Y}}_c$, $\widehat{\overline{Y}}_c$, and $\widehat{\widehat{\overline{Y}}}_c$ are unbiased for their corresponding parameters, and

$$V(\widehat{Y}_c) = N_c^2 S_c^2 \left[E\left(\frac{1}{n_c}\right) - \frac{1}{N_c}\right],$$

where $S_c^2 = \dfrac{1}{N_c - 1}\sum_{j=1}^{N_c}\left(Y_{cj} - \overline{Y}_c\right)^2$ is the subpopulation variance.

For $i = 1, \ldots, N$, we assume that

$$Y_{ci}^* = \begin{cases} Y_i, & U_i \in c, \\ 0, & o.w. \end{cases}$$

in this case

$$S_c^{*2} = \frac{1}{N-1}\sum_{i=1}^{N}\left(Y_{ci}^* - \overline{Y}_c^*\right)^2, \tag{3}$$

in which $\overline{Y}_c^*$ is the average of $Y_{c1}^*, \ldots, Y_{cN}^*$. Accordingly, it can be shown

$$V(\widehat{\widehat{Y}}_c) = N^2\frac{S_c^{*2}}{n}\left(1 - \frac{n}{N}\right).$$

Also, unbiased estimators for $V(\widehat{Y}_c)$ and $V(\widehat{\widehat{Y}}_c)$ are, respectively,

$$v(\widehat{Y}_c) = N_c^2 s_c^2\left(\frac{1}{n_c} - \frac{1}{N_c}\right)$$

and

$$v(\widehat{\widehat{Y}}_c) = N^2\frac{s_c^{*2}}{n}\left(1 - \frac{n}{N}\right),$$

where s_c^2 and s_c^{*2} are also sample variances y_{ci} and

$$y_{ci}^* = \begin{cases} y_i, & u_i \in c, \\ 0, & o.w. \end{cases} \tag{4}$$

for $i = 1, \ldots, n$, respectively.

When N_c is unknown, $\widehat{Y}_c$ and $\widehat{\overline{Y}}$ cannot be used to estimate the subpopulation total and mean. So using $\widehat{\overline{Y}}_c$ and $\widehat{\overline{Y}}_c$ is inevitable for estimation the subpopulation parameters. But if N_c is known, there is no limit to the use of these estimators, and in such circumstances it is important to decide on the choice of the appropriate estimator.

3. Modify the Post-Stratified Estimators

Suppose that a sample with size n is selected by a simple random sampling from population with N members and with L strata. Also, assume that N_c, for $c = 1, \ldots, L$, is the number of members in the stratum c and the random variable n_c is the sample size located on it, so $\sum_{c=1}^{L} N_c = N$ and $\sum_{c=1}^{L} n_c = n$. By assuming that for all stratum N_c are known, the unbiased estimator of post-stratification for population mean is as follows:

$$\overline{y}_{st} = \sum_{c=1}^{L} W_c \widehat{\overline{Y}}_c,$$

where $\widehat{\overline{Y}}_c$ is the sample mean of the stratum c and $W_c = \dfrac{N_c}{N}$. To obtain the variance of this estimator, we need to calculate $E\left(\dfrac{1}{n_c}\right)$. In practice this is difficult, but by taking a positive value for n_c, a good approximation of $E\left(\dfrac{1}{n_c}\right)$ is as follows (Stephan, 1945):

$$E\left(\frac{1}{n_c}\right) \simeq \frac{N}{nN_c} + \frac{N\left(N - N_c\right)}{n^2 N_c^2}. \tag{5}$$

Also under the Approximation (5), the variance of $\overline{y}_{st}$ is

$$V(\overline{y}_{st}) \simeq \frac{N-n}{nN} \sum_{c=1}^{L} \frac{N_c}{N} S_c^2 + \frac{1}{n^2} \sum_{c=1}^{L} \frac{N - N_c}{N} S_c^2, \tag{6}$$

where $S_c^2 = \dfrac{1}{N_c} \sum_{j=1}^{N_c} (Y_{cj} - \overline{Y}_c)^2$ is the variance of stratum c. It is worth noting that without using the stratification scheme, in a simple random sampling, $\overline{y} = \dfrac{1}{n} \sum_{i=1}^{n} y_i$ is a common estimator $\overline{Y}$.

3.1. A New Post-Stratified Estimator

In this section, using the estimators introduced in the previous section, we introduce a new estimator in the post-stratification sampling, which is more precise than $\overline{y}$ and $\overline{y}_{st}$.

Proposition 1. *If in the subpopulation c, the inequality*

$$\overline{Y}_c^2 \leq \frac{NS_c^2}{nN_c} \tag{7}$$

is established, then the $\widehat{\widehat{Y}}_c$ is more precise than $\widehat{Y}_c$ (Salehi & Seber, 2021).

Corollary 1. *Proposition 1 refers to this if the subpopulation mean has small values and members of subpopulation are heterogeneous, then the estimator $\widehat{\widehat{Y}}_c$ is more precise than the estimator $\widehat{Y}_c$. In practice, we need to have the parameters $\overline{Y}_c$ and S_c^2 for checking (7), usually both of them are unknown. Salehi & Seber (2021) suggested the following criterion in practice*

$$(\widehat{\overline{Y}}_c)^2 \leq s_c^2 \left[\left(\frac{2N-n}{nN_c} \right) + \left(\frac{N}{nN_c} \right)^2 \left(\frac{N-N_c}{N} \right) \left(\frac{N-n}{N-1} \right) \right]. \tag{8}$$

We know that in post-stratification every staratum is a subpopulation. Thus, in stratum c, if (8) holds for a selected sample, we can use $\widehat{\widehat{Y}}_c$ to estimate Y_c, otherwise $\widehat{Y}_c$ will be used.

Theorem 1. *a) The estimator*

$$\overline{y}_{mst} = \sum_{c \in G} W_c \widehat{\widehat{\overline{Y}}}_c + \sum_{c \in \bar{G}} W_c \widehat{\overline{Y}}_c$$

is an unbiased estimator for the mean of population, where G is the set index from stratums, in which the inequality (7) is established.

b) The variance of this estimator under the approximation (5) is as follows:

$$V(\overline{y}_{mst}) \simeq \sum_{c \in G} \frac{S_c^{*2}}{n} \left(1 - \frac{n}{N} \right) + \sum_{c \in \bar{G}} \left[\frac{N-n}{Nn} W_c S_c^2 + \frac{1}{n^2} (1 - W_c) S_c^2 \right].$$

c) Unbiased estimation for the approximate variance of b) is

$$v(\overline{y}_{mst}) = \sum_{c \in G} \frac{s_c^{*2}}{n} \left(1 - \frac{n}{N} \right) + \sum_{c \in \bar{G}} \left[\frac{N-n}{Nn} W_c s_c^2 + \frac{1}{n^2} (1 - W_c) s_c^2 \right].$$

Proof. a) Using the conditional expectation, it can be written

$$\begin{aligned} E(\overline{y}_{mst}) &= E\left(E\left(\sum_{c\in G} W_c\widehat{\widehat{\overline{Y}}}_c + \sum_{c\in\bar{G}} W_c\widehat{\overline{Y}}_c|n_1,\ldots,n_L\right)\right) \\ &= E\left(\sum_{c\in G} W_c E\left(\widehat{\widehat{\overline{Y}}}_c|n_c\right) + \sum_{c\in\bar{G}} W_c E\left(\widehat{\overline{Y}}_c|n_c\right)\right) \\ &= E\left(\sum_{c\in G} W_c\overline{Y}_c + \sum_{c\in\bar{G}} W_c\overline{Y}_c\right) \\ &= \overline{Y}. \end{aligned}$$

b) Using the definition of y^*_{ci} in (4), which results in $n_c\overline{y}_c = n\overline{y}^*_c$, the estimator $\overline{y}_{mst}$ can also be written as follows:

$$\begin{aligned} \overline{y}_{mst} &= \sum_{c\in G} W_c\widehat{\widehat{\overline{Y}}}_c + \sum_{c\in\bar{G}} W_c\widehat{\overline{Y}}_c \\ &= \sum_{c\in G}\frac{n_c}{n}\overline{y}_c + \sum_{c\in\bar{G}} W_c\overline{y}_c \\ &= \sum_{c\in G}\overline{y}^*_c + \sum_{c\in\bar{G}} W_c\overline{y}_c. \end{aligned} \tag{9}$$

On the other hand,

$$V(\overline{y}_{mst}) = E\left(V\left(\overline{y}_{mst}|n_1,\ldots,n_L\right)\right) + V\left(E\left(\overline{y}_{mst}|n_1,\ldots,n_L\right)\right). \tag{10}$$

The second expression to the right of (10) is equal to zero, so by using (9) and (10), we have

$$V(\overline{y}_{mst}) = E\left[V\left(\sum_{c\in G}\overline{y}^*_c + \sum_{c\in\bar{G}} W_c\overline{y}_c|n_1,\ldots,n_L\right)\right] \tag{11}$$

$$\begin{aligned} &= E\left[\sum_{c\in G} V\left(\overline{y}^*_c|n_c\right) + \sum_{c\in\bar{G}} W_c^2 V\left(\overline{y}_c|n_c\right)\right] \\ &= E\left[\sum_{c\in G}\frac{S^{*2}_c}{n}\left(1-\frac{n}{N}\right) + \sum_{c\in\bar{G}} W_c^2 S_c^2\left(\frac{1}{n_c}-\frac{1}{N_c}\right)\right] \\ &= \sum_{c\in G}\frac{S^{*2}_c}{n}\left(1-\frac{n}{N}\right) + \sum_{c\in\bar{G}} W_c^2 S_c^2\left[E\left(\frac{1}{n_c}\right)-\frac{1}{N_c}\right]. \end{aligned} \tag{12}$$

Now, by substituting (5) into (12),$V(\overline{y}_{mst})$ is obtained.

c) It is obvious, because s_c^2 and s_c^{*2} are, respectively, unbiased estimators S_c^2 and S_c^{*2}. □

***Note* 1.** It should be noted that if, for every $c = 1, \ldots, L$, the inequality (7) is established, then

$$\overline{y}_{mst} = \frac{1}{n}\sum_{i=1}^{n} y_i = \overline{y}.$$

In this case, $\overline{y}_{mst}$ is the estimator population mean in the simple random sampling, and if, for every $c = 1, \ldots, L$, it is not established, then

$$\overline{y}_{mst} = \sum_{c=1}^{L} W_c \overline{y}_c = \overline{y}_{st}.$$

In this case, the estimator $\overline{y}_{mst}$ is the same as the post-stratified estimator. In the following theorem, we show that $\overline{y}_{mst}$ is more precise than $\overline{y}$ and $\overline{y}_{st}$.

Theorem 2. *The estimator $\overline{y}_{mst}$ is more precise than $\overline{y}$ and $\overline{y}_{st}$.*

***Proof*.** We have

$$\begin{aligned}
V(\overline{y}_{mst}) &= E\left[V\left(\sum_{c\in G} W_c \widehat{\widehat{\overline{Y}}}_c + \sum_{c\in \bar{G}} W_c \widehat{\overline{Y}}_c | n_1, \ldots, n_L\right)\right] \\
&= E\left[\sum_{c\in G} W_c^2 V\left(\widehat{\widehat{\overline{Y}}}_c | n_c\right) + \sum_{c\in \bar{G}} W_c^2 V\left(\widehat{\overline{Y}}_c | n_c\right)\right] \\
&\leq E\left[\sum_{c=1}^{L} W_c^2 V\left(\widehat{\widehat{\overline{Y}}}_c | n_c\right)\right] \\
&= E\left[V\left(\sum_{c=1}^{L} W_c \widehat{\widehat{\overline{Y}}}_c | n_1, \ldots, n_L\right)\right] \\
&= E\left[V\left(\sum_{c=1}^{L} \frac{n_c}{n} \overline{y}_c | n_1, \ldots, n_L\right)\right] \\
&= E\left[V\left(\overline{y} | n_1, \ldots, n_L\right)\right] = V(\overline{y}).
\end{aligned}$$

The inequality of the third line is obtained by using Proposition 1 and the definition of G indexes. Similarly, we can show that $V(\overline{y}_{mst}) \leq V(\overline{y}_{st})$. □

In practice, for the derivation of G indexes, we need the parameters $\overline{Y}_c$ and S_c^2, usually both of them are unknown. Therefore, we can obtain the approximation from G indexes by using unbiased estimator $\overline{Y}_c$ and S_c^2 in a prototype, instead of checking the inequality (7), we suggest that to check inequality (8) for each stratum.

4. Simulation Study

In this section, to compare the estimators discussed in this article, we simulate data for three populations, the first population from the normal distribution, the second from the uniform distribution, and the third from the Laplace distribution. Each population has four strata, each stratum has the same probability density function (pdf) with different parameters such that sets $G = \{1\}$ and $\overline{G} = \{2, 3\}$. Details are given in the Table 1. Although the current study focuses on continuous distributions, the proposed method can be extended to discrete distributions, which will be considered in future work.

TABLE 1: The populations with their pdfs and size of subpopulations.

Population	pdf(1)	pdf(2)	pdf(3)	N_1	N_2	N_3
1	$N(0, 10)$	$N(6, 15)$	$N(10, 20)$	2000	1500	1000
2	$U(-5, 5)$	$U(0, 6)$	$U(2, 10)$	2000	1500	1000
3	Laplace(0,8)	Laplace(4,12)	Laplace(12,20)	2000	1500	1000

For every time with $M = 1000$ repetitions, we select samples with size $n = \{100, 200, 300, 400\}$ from each population using simple random sampling method. For every population, R statistical software was used to compute the following quantities in this simulation study:

$$\begin{aligned} \overline{y}_h &= \frac{1}{M}\sum_{l=1}^{M}\overline{y}_h^{(l)}, \\ MSE &= \frac{1}{M}\sum_{l=1}^{M}\left(\overline{y}_h^{(l)} - \overline{Y}\right)^2. \end{aligned}$$

Where $\overline{y}_h^{(l)}$, for $l = 1, 2, \ldots, M$ and $h = st, mst$, is the introduced estimators of population mean based on lth sample. Details are given in Table 2. Here, h denotes the stratum, st the post-stratified estimator, and mst the modified post-stratified estimator.

TABLE 2: The value of estimators and their MSE for the populations.

population	$\overline{Y}$	n	$\overline{y}_{mst}$	$\overline{y}_{st}$	$\overline{y}$	$MSE(\overline{y}_{mst})$	$MSE(\overline{y}_{st})$	$MSE(\overline{y})$
1	4.22	100	4.186	4.147	4.299	1.322	1.851	1.954
		200	4.216	4.207	4.215	1.066	1.148	1.187
		300	4.218	4.222	4.197	1.008	1.022	1.115
		400	4.223	4.217	4.226	0.428	0.563	0.528
2	2.33	100	2.329	2.330	2.346	0.037	0.048	0.072
		200	2.335	2.337	2.331	0.025	0.028	0.053
		300	2.330	2.334	2.334	0.021	0.027	0.051
		400	2.330	2.329	2.334	0.012	0.014	0.026
3	4.00	100	4.034	4.071	4.057	2.842	3.385	3.597
		200	4.018	4.038	3.939	1.694	1.750	1.799
		300	4.019	4.055	3.980	1.418	1.668	1.639
		400	4.008	4.032	4.021	0.662	0.856	0.819

Plots of the MSEs for all samples are given in Figure 1. According to Table 2 and Figure 1, it can be said that the MSE of $\overline{y}_{mst}$ is less than that of the two estimators in all samples and in all three populations. In terms of MSE after $\overline{y}_{mst}$, $\overline{y}_{st}$ has performed better than $\overline{y}$ in most cases. It is also observed that with increasing n, MSE of all estimators decreases, which is the expected scenario.

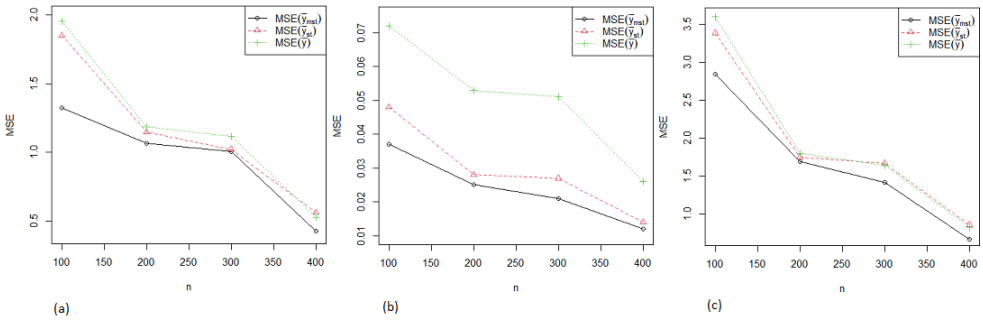

FIGURE 1: The MSE of the estimators in simulation study: (a) population 1 (Normal distribution), (b) population 2 (Uniform distribution) and (c) population 3 (Laplace distribution).

5. Real data application

In this section, we analyse a real dataset taken from the literature to implement the proposed. The data set is used in Lohr (2010, p. 31 and 79) and Salehi & Seber (2021). The data is a Census of Agriculture that the U.S. government conducts it every five years, collecting data on all farms (defined as any place from which 1000 dollars or more of agricultural products were produced and sold) in the 50 states. The acres devoted to farming in 1992 and 1987 are for $N = 3078$ counties in the USA. We are interested in comparing the introduced estimators by estimating the total number of acreage devoted to farming in the USA in 1992. Note that this number is equal to $Y = 943\,881\,045$.

Although agricultural activity occurs in every U.S. state, it is particularly concentrated in the Central Valley of California and in the Great Plains, a vast expanse of flat arable land in the center of the nation, in the region west of the Great Lakes and east of the Rocky Mountains. The eastern wetter half is a major corn and soybean-producing region known as the Corn Belt, and the western drier half is known as the Wheat Belt because of its high rate of wheat production Hatfield (2012). Therefore, it seems that classification of American states can increase the precise of estimators. For this reason, the American states are classified into four strata: northern, southern, eastern, and western states. For each stratum $c = 1, 2, 3, 4$, N_c denotes the number of counties and $\overline{Y}_c$ is the mean acreage devoted to farming. Details are given in Table 3.

TABLE 3: The details of the strata.

Stratum	Subpopulation size (N_c)	$\overline{Y}_c$	Standard deviation (S_c)
(1) northern states	944	295883.0	719263.8
(2) southern states	936	343938.6	377643.3
(3) eastern states	632	22877.17	226539.1
(4) western states	566	579827.9	719263.8

Now, to compare the proposed estimator with previous estimators, we take samples from this population with sizes $n = 25, 50, 100, 250, 500, 1000$ using a simple random sampling method. Based on these samples, the values of the estimators and their standard deviations are listed in the Table 4. It should be noted that according to Table 3, we can say that inequality (7) is valid only for stratum (3) for $n < 478$. So for $n < 478$ we have $G = \{3\}$ and $\overline{G} = \{1, 2, 4\}$.

TABLE 4: The value of estimators and their standard deviations.

n	$\widehat{Y}_{mst} = N\overline{y}_{mst}$	$\widehat{Y}_{st} = N\overline{y}_{st}$	$\widehat{Y} = N\overline{y}$	$SD(\widehat{Y}_{mst})$	$SD(\widehat{Y}_{st})$	$SD(\widehat{Y})$
25	928540854	911184951	867795191	2.248×10^8	2.352×10^8	2.584×10^8
50	850783164	828196001	752905455	1.001×10^8	1.046×10^8	1.137×10^8
100	987723918	1009432136	1085410899	1.090×10^8	1.139×10^8	1.225×10^8
250	954484872	952219418	973141970	8.564×10^7	8.755×10^7	9.516×10^7
500	928647680	928647680	947599673	5.475×10^7	5.475×10^7	5.763×10^7
1000	946926467	946926467	937550957	3.020×10^7	3.020×10^7	3.081×10^7

According to Table 3, we can say that the standard deviation of $\widehat{Y}_{mst}$ is less than that of the two estimators in all samples, for $n = 25, 50, 100, 250$.

6. Conclusions

A new unbiased estimator in the post-stratification sampling scheme is proposed. We have shown that this estimator performs more accurately than conventional and post-stratification estimators under the standard assumptions. A simulation study of three populations confirmed the theoretical findings of the article. During a real data analysis using census data of the acres devoted to farming in American states, the proposed estimator had a lower standard deviation than that of the other two estimators.

We provided a step-by-step algorithm for practical computation of the post-stratified and modified post-stratified estimators, making the methods directly applicable in practice. The proposed estimator performs particularly well when the required conditions are met; however, its behavior under violation of these conditions requires further investigation. This limitation is a potential avenue for future research. Overall, the enhanced discussion and practical guidance improve the applicability and understanding of the proposed methods for practitioners and researchers in finite population sampling.

[Received: February 2024 — Accepted: November 2025]

References

Clark, R. G. (2009), 'Sampling of subpopulations in two-stage surveys', *Statistics in Medicine* **28**(29), 3697–3717.

Cochran, W. G. (1977), *Sampling techniques*, 3 edn, Wiley, New York.

Hatfield, J. (2012), Agriculture in the Midwest, Technical report, U.S. National Climate Assessment. Midwest Technical Input Report.

Hedayat, A. S. & Sinha, B. K. (1991), *Design and inference in finite population sampling*, Wiley, New York.

Holt, D. & Smith, T. M. F. (1979), 'Post stratification', *Journal of the Royal Statistical Society* **142**, 33–46.

Hossaini, M. & Rezaei, A. H. (2021), 'Estimation of subpopulation parameters in one-stage cluster sampling design', *Journal of the Iranian Statistical Society* **20**(2), 65–78.

Jagers, P., Oden, A. & Trulsson, L. (1985), 'Post-stratification and ratio estimation: Usages of auxiliary information in survey sampling and opinion polls', *International Statistical Review* **53**, 221–238.

Levy, P. S. & Lemeshow, S. (1991), *Sampling of populations: Methods and applications*, Wiley, New York.

Little, R. J. A. (1993), 'Post stratification: A modeler's perspective', *Journal of the American Statistical Association* **88**, 1001–1012.

Lohr, S. L. (2010), *Sampling: Design and analysis*, 2 edn, Brooks/Cole, Boston.

Salehi, M. & Chang, K. C. (2005), 'Multiple inverse sampling in post-stratification with subpopulation sizes unknown: A solution for quota sampling', *Statistical Planning and Inference* **131**, 379–392.

Salehi, M. & Seber, G. A. F. (2021), 'A new estimator and approach for estimating the subpopulation parameters', *Journal of Taibah University for Science* **15**(1), 288–294.

Singh, D. & Chaudhary, F. S. (1986), *Theory and analysis of sample survey designs*, Wiley Eastern, New Delhi.

Stephan, F. F. (1945), 'The expected value and variance of the reciprocal and other negative powers of a positive Bernoullian variate', *The Annals of Mathematical Statistics* **16**, 50–61.

Thompson, S. K. (2012), *Sampling techniques*, 3 edn, Wiley, New York.