A new estimator and approach for estimating the subpopulation parameters

Mohammad Salehi M. and George A. F. Seber

ABSTRACT

Based on some theoretical results, we recommend a new algorithm for estimating the total and mean of a subpopulation variable for the case of a known subpopulation size, which is different from the algorithm recommended by most of sampling books. The latter usually recommend the multiplication of the subpopulation sample mean by the subpopulation size rather than the subpopulation total estimator for the unknown subpopulation size. We present a criterion to determine which estimator is more efficient. The criterion shows that the traditional total subpopulation estimator for unknown subpopulation size will be more efficient if the subpopulation mean is close to zero. Using an innovative procedure, we develop a new estimator, and we study its properties using real data. The new estimator is potentially an appropriate direct estimator in a composite estimator for small area estimation.

ARTICLE HISTORY

Received 30 April 2021
Revised 29 July 2021
Accepted 6 September 2021

KEYWORDS

Domain estimation; small area estimation; Taylor series approximation; transformation

1. Introduction

Sampling methods are essential part of many research studies on a population [1–4]. However, sometimes study of a population division will become of researcher’s interest. Subpopulation is a term used in the literature of survey sampling to denote partitions or divisions of population units in order to carry out a separate estimation for an underlying variable associated with each unit (e.g. income per person). Estimators are required for a variety of subpopulations, which should influence the choice of design and estimator [5]. During the last two decades, we have been faced with more metapopulations data sets, so that we frequently need methods of analysis for associated subpopulations. Software developers are adding options for this situation [6,7]. Some examples of specific subpopulations are females, high school graduates, Hispanics, public school students, cannabis users or biomarker positive [8,9], and these motivate researchers to develop more efficient designs and estimators of subpopulations. For instance, Rhone et al. [10] provided estimates of distance to the nearest and the third-nearest food store by age, race, ethnicity, income, vehicle access and participation in United States Department of Agriculture’s Supplementation Nutrition Assistance Program for both individuals and households. Kimani et al. [11] estimated parameters after subpopulation selection in adaptive seamless trials. Salehi and Chang [12] proposed an unbiased total subpopulation estimator under inverse sampling when the subpopulation size is unknown.

Where the implemented sampling design is Simple Random Sampling Without Replacement (SRSWOR), two situations arise depending on whether the number of units in the subpopulation is unknown or known. When the subpopulation size is unknown, the subpopulation sample mean for the variable values will be an unbiased estimator of the subpopulation mean, and the subpopulation total for the variable values is unbiasedly estimated by another estimator. When the subpopulation size is known, many sampling text books recommend using the subpopulation sample mean multiplied by the subpopulation size as an estimator of subpopulation total (say estimator 1) rather than use the other subpopulation total estimator for unknown size case (say estimator 2). In most of the sampling text books, it is noted that there is a reduction in variance in using estimator 1 (e.g. [13–15]). In this paper, we show that this claim is not always correct under certain conditions. We then present a criterion based on sample data that determines whether there will be a reduction in the variance of the mentioned subpopulation estimator or not. The criterion tells us where estimator 1 can be more efficient than estimator 2. Based on these results, we present an algorithm for estimating the mean and total subpopulation that is different from the algorithm recommended in many sampling books.

When the subpopulation size is known and the implemented sampling design is unequal probability sampling design, Särndal et al. [16] also recommended using an estimator in which the known subpopulation
size is used in its formula rather than using the estimator of unknown subpopulation size case. From a simulation study, we also show that the estimator of unknown subpopulation size case can be more efficient than the recommended estimator by Särndal et al. [16], particularly for small sample sizes.

Using a transformation technique, we develop a new estimation procedure providing us new estimators of the total and the mean subpopulation. The estimators can be more efficient than the two previous estimators for possibly any population and a wide range of sample sizes. Sample surveys designed to produce estimates for larger areas may also be used to estimate smaller area parameters. Small Area Estimation (SAE) is recommended estimator by Särndal et al. [16], particularly for small sample sizes.

In Section 3, we derive a criterion that reveals which estimator is an estimator of the total and the mean subpopulation. In Section 5, we introduce new estimators to be used in the composite estimator for SAE, as they are more efficient for smaller sample sizes.

In Section 2, we present the notation and estimators. In Section 3, we derive a criterion that reveals which estimator is more efficient. In Section 4, we present the new algorithm for estimating the mean and total subpopulation, and the result of a simulation study indicating that a similar algorithm can be used for an unequal probability sampling. In Section 5, we introduce new estimators of the total and mean subpopulation.

2. Notation and estimators

Consider a population that consists of $N$ known units, with the index set of the units being $\mathcal{U} = \{1, 2, \ldots, N\}$ with variable of interest $y = \{y_1, y_2, \ldots, y_N\}$. Let $\mathcal{U}_h$ be the index units in subpopulation $h$ with size $N_h$. A sample of size $n$ is selected from the whole population of units using SRS. Let $\mu = (1/N) \sum_{i=1}^{N} y_i$ and $\tau = \sum_{i=1}^{N} y_i$ be the population mean and total, respectively. Suppose that we want to estimate

$$\mu_h = \frac{1}{N_h} \sum_{i \in \mathcal{U}_h} y_i, \quad \tau_h = \sum_{i \in \mathcal{U}_h} y_i.$$  

If $N_h$ is unknown,

$$\hat{\mu}_h = \frac{n_h}{N_h} \sum_{i = 1}^{n_h} y_i = \bar{y}_h$$  

is an estimator of $\mu_h$, where $n_h$ is the random sample size from $\mathcal{U}_h$. Its variance and variance estimator are,

$$V(\hat{\mu}_h) = \left[ E \left( \frac{1}{n_h} - \frac{1}{N_h} \right) \right] \sigma_h^2,$$

$$v(\hat{\mu}_h) = \left[ E \left( \frac{1}{n_h} - \frac{1}{N} \right) \right] \sigma_h^2,$$

where $\sigma_h^2 = 1/(N_h - 1) \sum_{i=1}^{N_h} (y_i - \mu_h)^2$ is the variance of subpopulation $h$, and $s_h^2 = 1/(n_h - 1) \sum_{i=1}^{n_h} (y_i - \hat{\mu}_h)^2$ is the unbiased sample variance of subpopulation $h$. An unbiased estimator of $\tau_h$ is given by

$$\tilde{\tau}_h = \frac{N}{n} n_h \bar{y}_h.$$  

To derive its variance, $\bar{y}_h^*$ is defined as follows:

$$\bar{y}_h^* = \begin{cases} y_i, & i \in \mathcal{U}_h, \\ 0, & i \in \mathcal{U} - \mathcal{U}_h \end{cases}$$

The variance of $\tilde{\tau}_h$ is given by

$$V(\tilde{\tau}_h) = N^2 \left( 1 - \frac{n}{N} \right) \frac{s_h^2}{n},$$

where $s_h^2 = 1/(N - 1) \sum_{i=1}^{N} (y_i^* - \mu_{y^*})^2$ and $\mu_{y^*} = 1/N \sum_{i=1}^{N} y_i^*$. A variance estimator is given by

$$v(\tilde{\tau}_h) = N^2 \left( 1 - \frac{n}{N} \right) \frac{s_h^2}{n},$$

where $s_h^2 = 1/(n - 1) \sum_{i=1}^{n} (y_i^* - \bar{y}_h^*)^2$ and $\bar{y}_h^* = 1/n \sum_{i=1}^{n} y_i^*$.

If $N_h$ is known, Cochran [15, p. 48], Lohr [13, p. 135], Scheaffer et al. [18, p. 339] and Thompson [14, p. 62] recommended using $\tilde{\tau}_h = N_h \bar{y}_h$ rather than $\tilde{\tau}_h$. Cochran [15] showed that $V(\tilde{\tau}_h)$ is approximately smaller than $V(\tilde{\tau}_h)$. He showed that

$$\frac{V(\tilde{\tau}_h)}{V(\tilde{\tau}_h)} \approx \frac{C_h^2}{C_h^2 + \left( \frac{N_h}{N} \right)^2},$$

where $C_h$ is the coefficient of variation of subpopulation $h$. Figure 1 presents the flowchart for estimating the subpopulation mean and total recommended by sampling text books.

For known $N_h$ case, we will show that $\tilde{\tau}_h$ and $\hat{\mu}_h = \tilde{\tau}_h/N_h$ can be more efficient than $\tilde{\tau}_h$ and $\hat{\mu}_h$, respectively.

3. Criterion for comparing the estimators

Since the comparison of the total estimators are quite similar to the comparison of the mean estimators, we will focus only on the total estimators. We present a criterion to compare the variances of $\tilde{\tau}_h$ and $\hat{\tau}_h$. We first note that

$$\mu_{y^*} = \frac{1}{N} \sum_{i \in \mathcal{U}} y_i^* = \frac{1}{N} \sum_{i = 1}^{n_h} y_i = \frac{1}{N} N_h \mu_h.$$  

Then, expanding the quadratic, we obtain after some algebra the following lemma.
Lemma 3.1: The subpopulation variance $\sigma_h^2$ can be written in terms of the subpopulation variance $\sigma_h^2$ as follows:

$$
\sigma_h^2 = \frac{N_h - 1}{N - 1} \sigma_h^2 + \frac{N_h (N - N_h)}{N (N - 1)} \mu_h^2.
$$

Proposition 3.1: If $\mu_h^2 \leq N/(nN_h) \sigma_h^2$ holds, the variance of $\bar{y}_h$ is approximately greater than that of $\bar{\tau}_h$.

Proof: Now $n_h$ has the Hypergeometric distribution with mean $nN_h/N$ and variance

$$
V(n_h) = n \frac{N_h}{N} \left(1 - \frac{N_h}{N}\right) \frac{N - n}{N - 1}.
$$

For $X = n_h$, using a second-order Taylor series approximation for $g(X) = 1/X$, we have

$$
E[g(X) - g(\mu_X)] \leq \frac{1}{2} \sigma_X^2 \left( \frac{d^2 g}{dx^2} \right)_{X = \mu_X},
$$

since the third-order term is negative. Hence, by first conditioning on $n_h$ and then taking an expectation with respect to $n_h$, we get

$$
E \left( \frac{1}{n_h} \right) \leq \frac{N}{nN_h} + \left( \frac{N}{nN_h} \right)^2 \left( \frac{N - N_h}{N} \right) \left( \frac{N - n}{N - 1} \right).
$$

From (2),

$$
V(\bar{\tau}_h) = \frac{N_h^2}{n} \left[ E \left( \frac{1}{n_h} \right) - \frac{1}{N_h} \right] \sigma_h^2.
$$

By the Lemma and Equation (5), we have

$$
V(\bar{\tau}_h) = \frac{N(N-n)}{n} \sigma_h^2.
$$

$$
= \frac{N(N-n)}{n} \left[ \left( \frac{N_h - 1}{N - 1} \right) \sigma_h^2 + \left( \frac{N_h(N - N_h)}{N(N - 1)} \right) \mu_h^2 \right]
$$

$$
\leq \frac{N_h}{nN_h} (N-n) \sigma_h^2 + \frac{N_h(N-N_h)(N-n)}{n(N-1)} \mu_h^2.
$$

(10)

In (10), the variance $V(\bar{\tau}_h)$ is less than or equal to the last equation, as we have $(N_h - 1)/(N - 1) \leq N_h/N$. From (9) and (10), we have

$$
V(\bar{\mu}_h) - V(\bar{\tau}_h) \leq \left[ \frac{(N-N_h)(N-n)}{n^2(N-1)} \right] \left[nN_h \mu_h^2 - N \sigma_h^2\right].
$$

(11)

Since the first term of the right-hand side of (11) is positive, $V(\bar{\tau}_h)$ is less than or equal to $V(\bar{\tau}_h)$, when $\mu_h^2 \leq \sigma_h^2 N/(nN_h)$. In other words, if

$$
\mu_h^2 \leq \frac{N}{nN_h} \sigma_h^2
$$

(12)

holds, the efficiency of $\bar{\tau}_h$ is greater than that of $\bar{\tau}_h$. We may rewrite the criterion (12) as

$$
\mu_h^2 \leq \frac{\sigma_h^2}{n(N_h/N)} \leq \frac{\sigma_h^2}{E(n_h)}
$$

(13)

Criterion (13) tells us that $\bar{\tau}_h$ is the preferable estimator if

(1) the subpopulation mean is close to zero or/and
(2) the subpopulation variance is large or/and
(3) The expected sample size from subpopulation $h$ is small.

However, the criterion (12) depends on the subpopulation parameters that are not available. We may plug in unbiased estimators of $\mu_h^2$ and $\sigma_h^2$ into the criterion (12) to get a new criterion that depends on the sample only. Since $E(\tilde{y}_h^2) = V(\tilde{y}_h) + \mu_h^2$, we can use the unbiased estimators

$$
\tilde{y}_h^2 - \left[ E \left( \frac{1}{n_h} \right) - \frac{1}{N_h} \right] s_h^2
$$

and

$$
\tilde{y}_h^2 - s_h^2
$$

for estimating $\mu_h^2$ and $\sigma_h^2$, respectively. Instead of (12), we may then use the following criterion

$$
\tilde{y}_h^2 - s_h^2 \left[ E \left( \frac{1}{n_h} \right) - \frac{1}{N_h} \right] \leq \frac{N_h^2}{nN_h} s_h^2.
$$

(14)

This leads to

$$
\tilde{y}_h^2 \leq \left[ E \left( \frac{1}{n_h} \right) + \frac{N - n}{nN_h} \right] s_h^2.
$$
Thus, if (15) holds for a selected sample, we can use \( N_h \) if exactly the same criteria (15) and (12).

After we have the observations for a selected sample, we want to decide on using either \( \tilde{\tau}_h \) or \( \tilde{\mu}_h \). Otherwise, \( \tilde{\tau}_h \) will be used. Note that if we want to decide on using either \( \tilde{\tau}_h \) or \( \tilde{\mu}_h \), we should use exactly the same criteria (15) and (12).

4. New algorithm

After we have the observations for a selected sample, if the criterion (15) is satisfied, we will, respectively, estimate \( \tau_h \) and \( \mu_h \) by \( \tilde{\tau}_h \) and \( \tilde{\mu}_h \), respectively. We would estimate them by \( \tilde{\tau}_h \) and \( \tilde{\mu}_h \), where \( N_h \) is known. If \( N_h \) is unknown, we will estimate \( \tau_h \) and \( \mu_h \) by \( \tilde{\tau}_h \) and \( \tilde{\mu}_h \), respectively. Figure 2 presents a new teaching flowchart for estimating the subpopulation mean and total.

Example 4.1: We use a data set that Lohr [13, pp. 31 and pp. 79] used from the Census Agriculture in United States. The acres devoted to farming in 1992 and 1987 are for \( N = 3078 \) counties in the USA. A sample of size \( n = 260 \) counties are randomly selected in which \( n_h = 30 \) counties are in western states. We shall estimate the changes in total and mean acreage devoted to farming from 1987 to 1992 in the western region.

The population and the sample are provided as an online resource. The number of counties in western states is \( N_h = 422 \). The sample mean and variance are, respectively, \( \tilde{h} = 12309.13 \) and \( s^2_h = 3926685773 \). The right-hand side of (15) is 222754517.2, which is greater than \( \tilde{h}^2 = 151514763.4 \). Using the flowchart in Figure 2, the estimates are

\[
\tilde{\tau}_h = \frac{3078}{260} (12309.13) = 4371636.05
\]

\[
\tilde{\mu}_h = \frac{4371636.05}{422} = 10359.32.
\]

The variance estimates are, respectively,

\[
v_\tilde{\tau}_h = N^2 \left( 1 - \frac{n}{N} \right) \frac{s^2_h}{N} = 3078^2 \left( 1 - \frac{260}{3078} \right) (455192494.8) = 1.518 \times 10^{13}
\]

and,

\[
v_\tilde{\mu}_h = \frac{N^2}{N_h} v_\tilde{\tau}_h = \frac{1.52 \times 10^{13}}{178084} = 85271999.6.
\]

If we had used the traditional approach recommended by the sampling text books, the subpopulation mean and total estimates would have been \( \tilde{\mu}_h = \tilde{y}_h = 12309.13 \) and \( \tilde{\tau}_h = N_h \tilde{y}_h = 422 \times 12309.13 = 5194454.27 \), respectively. Using (2) and (8), the variance estimates would have, respectively, been

\[
v_\tilde{\tau}_h = 103293369.6
\]

and,

\[
v_\tilde{\mu}_h = N^2 \left( 1 - \frac{n}{N} \right) \frac{s^2_h}{N} = (422)^2 \times (103293369.6) = 1.84 \times 10^{13}
\]

The estimated relative efficiency gain due to knowing \( N_h \) is

\[
\frac{v_\tilde{\tau}_h}{v_\tilde{\mu}_h} = \frac{1.518 \times 10^{13}}{1.84 \times 10^{13}} = 0.83.
\]

This means that there is not only no gain in efficiency due to knowing \( N_h \), but also the estimated loss in efficiency will be 17%. Furthermore, we suggest using \( \tilde{\mu}_h \) rather than \( \tilde{\tau}_h \) in such a case; see Figure 2.

As we mentioned in Section 3, \( \tilde{\mu}_h \) and \( \tilde{\tau}_h \) are more efficient than their counterparts when the expected sample size from subpopulation \( h \) is small. We now discuss how small the sample size, \( n \), should be such that \( V(\tilde{\tau}_h) \) becomes smaller than \( V(\tilde{\mu}_h) \) in this example. We can either use criterion (12) or we can directly use the variances. In Figure 3 (a), we draw the graph of \( \mu^2_h - n N_h \sigma^2_h \) against the sample size, \( n \). When the graph is negative, \( V(\tilde{\tau}_h) \) is smaller than \( V(\tilde{\mu}_h) \). The graph shows that the variance of \( \tilde{\tau}_h \) is smaller than the variance of \( \tilde{\mu}_h \) for

\[
n \leq 362.
\]
Figure 3. (a) The curve of $\mu^2_h - \frac{N_h}{\tilde{N}_h} \sigma^2_h$ against $n$ which indicates that $V(\tilde{\tau}_h) \leq V(\hat{\tau}_h)$ for $n \leq 362$ according to criterion (12). (b) The curve of $V(\tilde{\tau}_h) - V(\hat{\tau}_h)$ against $n$ which means that $V(\tilde{\tau}_h) \leq V(\hat{\tau}_h)$ for $n \leq 414$.

for $n \leq 414$. The example shows that the inequality in criterion (12) can become even sharper.

From (12), $V(\tilde{\tau}_h)$ is smaller than $V(\hat{\tau}_h)$ for a reasonable sample size, where the population mean is close to zero like the above example. To show that the mentioned results can be extended to unequal probability sampling, we ran a simulation study on the Census Agriculture in United States, where $n$ counties were selected with replacement proportional to their acreage devoted to farming in 1982. We wish to estimate the changes in total acreage devoted to farming from 1987 to 1992 in the western region. For unknown $N_h$, Särndal et al. [16, pp. 390] recommended using

$$\tilde{\tau}_h \pi = \sum_{i=1}^{\nu_h} \frac{y_i^*}{\pi_i},$$

where $\nu$ is the number of distinct counties in the sample, $y_i^*$ is the same as (4) in Section 2, $\pi_i = 1 - (1 - p_i)^n$, and $p_i$ is the acreage devoted to farming in the $i$th county divided by the total acreage devoted to farming in USA in 1982. For known $N_h$, Särndal et al. [16] recommended using

$$\hat{\tau}_h \pi = \frac{N_h}{\tilde{N}_h} \sum_{i=1}^{\nu_h} y_i \pi_i,$$

where $\nu_h$ is the number of distinct counties selected from subpopulation $h$, and $\tilde{N}_h = \sum_{i=1}^{\nu_h} 1/\pi_i$.

In the simulation study, we computed the relative efficiency of $\tilde{\tau}_h \pi$, which is as follows:

$$\text{Reff}(\tilde{\tau}_h \pi) = \frac{\text{MSE}(\hat{\tau}_h \pi)}{\text{MSE}(\tilde{\tau}_h \pi)}.$$  

The relative efficiencies are simulated for $n$ from 50 to 2000, which are given in Figure 4. All relative efficiencies are greater than 1 for all sample sizes and $\tilde{\tau}_h \pi$ was a more efficient estimator, particularly for smaller sample sizes.

5. New estimators of subpopulation total and mean

Unfortunately, criterion (12) is not appropriate for a wide range of populations. We now introduce an estimation procedure that makes criterion (12) appropriate for any population with a reasonable sample size range. This estimation procedure will lead to new estimators for the subpopulation total and subpopulation mean. The procedure is as follows:

(1) We first need to have a reasonable guess at the subpopulation mean. This guess can be a rough estimate related to a similar survey, a previous survey, a pilot survey, the mean of an auxiliary variable or can be purely based on an expert's idea. Let this guess be $c$. 

Figure 4. The relative efficiencies $\tilde{\tau}_h \pi$ are computed for $n = 50–2000$ for an unequal probability sampling design. The relative efficiencies are greater than 1 for all sample sizes and $\tilde{\tau}_h \pi$ was a more efficient estimator, particularly for smaller sample sizes.
The variance of $\tilde{\tau}_h$ is given by,
\[ V(\tilde{\tau}_h) = N^2 \left( \frac{1}{N} \right) \frac{\sigma_{zh}^2}{n} n. \]
(17)
where $\sigma_{zh}^2 = 1/(N-1) \sum_{i=1}^{N}(z_i - \mu_z)^2$ and $\mu_z = 1/N \sum_{i=1}^{N} z_i$. A variance estimator is given by
\[ \hat{V}(\tilde{\tau}_h) = N^2 \left( \frac{1}{N} \right) \frac{\hat{\sigma}_{zh}^2}{n}, \]
(18) where $\hat{\sigma}_{zh}^2 = 1/(n-1) \sum_{i=1}^{n}(z_i - \bar{z})^2$.

(4) Since $\tau_h = \sum_{i=1}^{N}(y_i - c)$, we have $\tau_h = N_h c$. Thus, another estimator of $\tau_h$ is
\[ \tilde{\tau}_h = N_h c + \tilde{\tau}_h = N_h c + \tilde{N}_h. \]
(19)
= $N_h c + \frac{N}{n} \sum_{i=1}^{n}(\tilde{y}_h - c) = \left( N_h - \frac{N}{n} n_h \right) c + \tilde{\tau}_h. \]
(20)
This is just $\tilde{\tau}_h$, plus an adjustment factor of $(N_h - (N_h/n))c$. From (19), the variance of $\tilde{\tau}_h$ and its variance estimator are the same as those of the transformed variable $z_i$, namely
\[ V(\tilde{\tau}_h) = V(\tilde{z}_h), \quad V(\tilde{\tau}_h) = V(\tilde{z}_h). \]
(21)
To evaluate the introduced estimation procedure, we use the Census Agriculture data. We wish to estimate the total acres devoted to farming in 1992 in the western states, where the sample sizes are $n = 20, 30, 40, 50, 75, 100, 125, \ldots, 500, 600, \ldots, 1000, 1500, 2000, 2500, 3000, 3755$. Let guesses on the subpopulation mean be $c = 560000, 600000,640000, 680000, 720000, 735598.62, 760000, 800000$. A very natural guess on the subpopulation mean can be 735598.62 in this example, which is the known subpopulation mean in 1987. Using (5) and (21), $V(\tilde{\tau}_h)$ and $V(\tilde{\tau}_h)$ are computed for each case. We define the relative efficiency of $\tilde{\tau}_h$ as follows:
\[ \text{Reff}(\tilde{\tau}_h) = \frac{V(\tilde{\tau}_h)}{V(\tilde{\tau}_h)}, \]
as both estimator are unbiased. The relative efficiencies are computed and the results are given in Table 1. The mean acres devoted to farming in 1992 in the western states was $\mu_h = 723343.96$. With a very rough guess of 560,000 acres, $\mu_h$, $\tilde{\tau}_h$ can estimate the western states total better than $\tilde{\tau}_h$ for sample sizes of less than or equal to 375. If the guess is 600,000, the relative efficiency is more than 1 for sample sizes of less than or equal to 1000. If the guess is between 640,000 and 740,000, $\tilde{\tau}_h$ is more efficient estimator than $\tilde{\tau}_h$ for any sample size. The subpopulation mean in 1987 was 735598.62 and if we use it as $c$, the variance of $\tilde{\tau}_h$ would be smaller than

### Table 1

| $n$     | 560,000 | 600,000 | 640,000 | 680,000 | 720,000 | 735598.62 | 760,000 | 800,000 |
|--------|---------|---------|---------|---------|---------|-----------|---------|---------|
| 10     | 1.581   | 1.603   | 1.619   | 1.629   | 1.633   | 1.633     | 1.630   | 1.621   |
| 30     | 1.190   | 1.202   | 1.210   | 1.212   | 1.210   | 1.212     | 1.204   | 1.204   |
| 50     | 1.107   | 1.119   | 1.126   | 1.128   | 1.126   | 1.128     | 1.120   | 1.120   |
| 80     | 1.061   | 1.072   | 1.076   | 1.081   | 1.079   | 1.081     | 1.073   | 1.073   |
| 100    | 1.045   | 1.056   | 1.063   | 1.065   | 1.063   | 1.065     | 1.057   | 1.057   |
| 150    | 1.025   | 1.035   | 1.042   | 1.044   | 1.042   | 1.044     | 1.037   | 1.037   |
| 200    | 1.014   | 1.025   | 1.031   | 1.034   | 1.032   | 1.033     | 1.026   | 1.026   |
| 250    | 1.008   | 1.019   | 1.025   | 1.027   | 1.026   | 1.027     | 1.020   | 1.020   |
| 300    | 1.004   | 1.014   | 1.021   | 1.023   | 1.021   | 1.023     | 1.016   | 1.016   |
| 350    | 1.001   | 1.011   | 1.018   | 1.020   | 1.018   | 1.020     | 1.013   | 1.013   |
| 370    | 1.000   | 1.010   | 1.017   | 1.019   | 1.017   | 1.019     | 1.012   | 1.012   |
| 400    | 0.999   | 1.009   | 1.015   | 1.018   | 1.016   | 1.018     | 1.010   | 1.010   |
| 450    | 0.997   | 1.007   | 1.014   | 1.016   | 1.014   | 1.016     | 1.009   | 1.009   |
| 500    | 0.996   | 1.006   | 1.012   | 1.015   | 1.013   | 1.014     | 1.007   | 1.007   |
| 600    | 0.994   | 1.004   | 1.010   | 1.013   | 1.011   | 1.012     | 1.005   | 1.005   |
| 700    | 0.992   | 1.002   | 1.009   | 1.011   | 1.009   | 1.011     | 1.004   | 1.004   |
| 800    | 0.991   | 1.001   | 1.008   | 1.010   | 1.008   | 1.010     | 1.003   | 1.003   |
| 900    | 0.990   | 1.000   | 1.007   | 1.009   | 1.007   | 1.009     | 1.002   | 1.002   |
| 1000   | 0.990   | 1.000   | 1.008   | 1.007   | 1.007   | 1.008     | 1.001   | 1.001   |
| 1500   | 0.988   | 0.998   | 1.004   | 1.006   | 1.005   | 1.006     | 0.999   | 0.999   |
| 2000   | 0.987   | 0.997   | 1.003   | 1.005   | 1.004   | 1.005     | 0.998   | 0.998   |
| 2500   | 0.986   | 0.996   | 1.002   | 1.005   | 1.003   | 1.004     | 0.997   | 0.997   |
| 3000   | 0.986   | 0.996   | 1.002   | 1.004   | 1.002   | 1.004     | 0.997   | 0.997   |
| 3755   | 0.986   | 0.996   | 1.002   | 1.004   | 1.002   | 1.004     | 0.997   | 0.997   |
the variance of $\hat{\tau}_h$ for any sample size. We may even recommend the use of
\[ \hat{\mu}_h = \frac{1}{N_h} \bar{y}_h = c + \frac{N}{N_h} z^2 = \left( 1 - \frac{N_N}{N_h n} \right) c + \frac{1}{N_h} \hat{\tau}_h, \]
to estimate $\mu_h$ as it outperforms $\hat{\mu}_h = \bar{y}_h$, where $n \leq 1000$ for any $c$ between 640,000 and 800,000. Its variance and variance estimators are,
\[ V(\hat{\mu}_h) = \frac{1}{N_h^2} V(\bar{y}_h), \quad V(\hat{\tau}_h) = \frac{1}{N_h^2} V(\hat{\tau}_h). \]

6. Conclusion
Despite the popular impression from the literature, we have shown that $\bar{y}_h$ and $\hat{\mu}_h$ can be more efficient estimators than $\bar{y}_h$ and $\hat{\mu}_h$, respectively, where the subpopulation size is known. A criterion was developed that can be used to choose more appropriate estimators.

The new estimators are introduced to estimate the subpopulation mean and total for a known subpopulation size situation. Criterion (13) for the new estimator is
\[ \mu_h^2 \leq \frac{\sigma^2}{E(n_h)} \quad \text{or} \quad (\mu_h - c)^2 \leq \frac{\sigma^2}{E(n_h)}, \]
for which $\mu_h^2$ is very close to zero for an appropriate $c$. Furthermore, the new estimators are more efficient when $E(n_h)$ is small. This can be interpreted that the new estimators would be potentially appropriate direct estimators to be used in composite estimators for Small Area Estimation.

Acknowledgments
Open Access funding provided by the Qatar National Library. Salehi would like to express his gratitude and appreciation to Prof. J.N.K. Rao, Carleton University, and Prof. R. Chambers, University of Wollongong, for their valuable comments and suggestions on an early draft of this manuscript. The authors are thankful to the editor and two anonymous reviewers for the constructive comments on the initial version.

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

ORCID
Mohammad Salehi M. http://orcid.org/0000-0003-1181-5523

References
[1] Abbasi SA, Qurat-Ul-Ain K, Omar MH, et al. On designing a sequential based EWMA structure for efficient process monitoring. J Taibah Univ Sci. 2020;14(1):177–191. Available From: https://doi.org/10.1080/16583655.2020.1712011.
[2] Alsharif SM, El-Sayed WS, Hanafy AM. Geomorphic distribution and prevalence of potential asymptomatic Staphylococcus spp. in the nasopharyngeal cavity of elementary school boys at al-Madinah, KSA. J Taibah Univ Sci. 2021;15(1):122–132. Available From: https://doi.org/10.1080/16583655.2021.1892991.
[3] Zubair A, Eisa MA, Hamedanib GG, et al. New methods to define heavy-tailed distributions with applications to insurance data. J Taibah Univ Sci. 2020;14(1):359–382. Available from: https://doi.org/10.1080/16583655.2020.1741942.
[4] Javed M, Irfan M. A simulation study: new optimal estimators for population mean by using dual auxiliary information in stratified random sampling. J Taibah Univ Sci. 2020;14(1):557–568. Available from: https://doi.org/10.1080/16583655.2020.1752004.
[5] Kish L. Design and estimation for domains. Stat. 1980;29: 209–222.
[6] West BT, Berglund P, Heeringa SGH. A closer examination of subpopulation analysis of complex-sample survey data. Stat J. 2008;8:520–531.
[7] Zardetto D. ReGenesees: an advanced R system for calibration, estimation and sampling error assessment in complex sample surveys. J Official Stat. 2015;31(2):177–203. Available From: https://doi.org/10.1515/JOS-2015-0013.
[8] Graf AC, Posch M, Koenig F. Adaptive designs for subpopulation analysis optimizing utility functions. Biomet J. 2015;57(1):76–89. DOI: 10.1002/bimj.201300257.
[9] Loflin M, Babson K, Browne K, et al. Assessment of the validity of the CUDIT-R in a subpopulation of cannabis users. The Amer J Drug and Alcohol Abuse. 2018;44(1):19–23. DOI:10.1080/00952990.2017.1376677.
[10] Rhone A, Ver Ploeg M, Williams R, et al. Understanding low-Income and lowAccess census tracts across the nation: subnational and subpopulation estimates of access to healthy food. Econom Inform Bullet. 2019. p. 209. Available From: https://www.ers.usda.gov/webdocs/publications/93141/eib-209.pdf.
[11] Kimani PK, Todd S, Renfro AL, et al. Estimation after subpopulation selection in adaptive seamless trials. Stat Med. 2015;34:2581–2601.
[12] Salehi M, Chang KC. Multiple inverse sampling in post-stratification with subpopulation sizes unknown: a solution for quota sampling. Stat Planning and Inference. 2005;131:379–392.
[13] Lohr SL. Sampling: design and analysis. 2nd ed. Boston: Brooks/Cole; 2010.
[14] Thompson SK. Sampling. 3rd ed. New York: Wiley; 2012.
[15] Cochran WG. Sampling techniques. 3rd ed. Wiley: New York; 1977: 34–39.
[16] Särnal CE, Swensson B, Wretman Y. Model assisted survey sampling. New York: Springer Verlag; 1992.
[17] Rao JNK, Molina I. Small area estimation. 2nd ed. Hobo-Cole; 2010.
[18] Schaeffer RL, Mendenhall W, Ott RL, et al. Elementary survey sampling. 7th ed. Belmont CA: Thomson Brooks/Cole; 2012.