The Calibration for Two-Phase Randomized Response Estimator

CHANG-KYOOON SON¹, KI-HAK HONG², GI-SUNG LEE³, AND JONG-MIN KIM⁴

¹Korea Institute for Health and Social Affairs, Seoul, Republic of Korea
²Department of Computer Science, College of Engineering, Dongshin University, Jeollanam-Do, Republic of Korea
³Department of Children’s Welfare, College of Health and Welfare, Woosuk University, Jeollabuk-Do, Republic of Korea
⁴Statistics Discipline, Division of Science and Mathematics, University of Minnesota-Morris, Morris, Minnesota, USA

This article presents the calibration procedure of the two-phase randomized response (RR) technique for surveying the sensitive characteristic. When the sampling scheme is two-phase or double sampling, auxiliary information known from the entire population can be used, but the auxiliary information should be information available from both the first and second phases of the sample. If there is auxiliary information available from both the first and second phases, then we can improve the ordinary two-phase RR estimator by incorporating this information in the estimation procedure. In this article, we used the new two-step Newton’s method for computing unknown constants in the calibration procedure and compared the efficiency of the proposed estimator through some numerical study.

Keywords  Auxiliary information; Calibration; Generalized regression estimator; Randomized response technique; Two-phase sampling.

Mathematics Subject Classification  62D05.

1. Introduction

When the sampling is carried out in several phases, there is often usable auxiliary information that is distributed by a complex survey design. Using the auxiliary information, the estimators of the population mean, population total, and population proportion can be improved. For the related literature review, Chaudhuri and Roy (1997) studied the optimal properties of the well-known regression estimators used in two-phase sampling. Sarndal and Swensson (1987)
proposed an estimator that uses all auxiliary information available for two-phase sampling, with different auxiliary information for the total population and the population obtained from the first-phase sampling. Dupont (1995) suggested the calibration approach with the several levels of auxiliary information, and he sought the joint combination method of the regression estimator and the calibration estimator according to the levels of auxiliary information. Hidiroglou and Sarndal (1995) suggested the domain estimation and considered the calibration and regression estimation in two-phase sampling scheme. Recently, Wu and Luan (2003) suggested the optimal calibration for two-phase sampling design. They didn’t consider the overall two-phase calibration procedure, even if they considered the calibration at the second-phase only.

Meanwhile, in cases of surveys for sensitive characters such as drug, tax evasion, and so on, it is difficult to obtain the truthful answer from the respondents. Warner (1965) was the first to suggest a randomized response (RR) technique, which minimizes underreporting of data on socially undesirable or incriminating behavior questions. In Warner’s original proposal the statements are (refers to Cochran, 1977, Ch. 13):

“I am a member of Group A.” (with probability P)
“I am not a member of Group A.” (with probability 1 − P).

The Warner model required the respondent to give a “Yes” or “No” answer either to the sensitive question or to its negative depending on the outcome of a randomizing device not reported to the interviewer. Many researchers have developed the variants of the Warner’s RR technique. If two-phase RR technique is used for sensitive trait surveys, then a calibration procedure using the auxiliary information for reducing the sampling error or nonresponse bias of the sampling units can be applied to two-phase RR technique. In this article, we will use the several levels of auxiliary information obtained from the first and second phase for getting the calibration two-phase RR estimator.

The article is organized as follows. Section 2 explains how calibration is carried out in each of two phases using the generalized least squares distance. Section 3 describes two-phase RR estimator. Section 4 describes the calibration of two-phase RR estimators according to auxiliary information. Section 5 provides the variance of the calibration two-phase RR estimator. Section 6 deals with some numerical study of the proposed estimator, and Sec. 7 provides concluding remarks.

2. Calibration on Two-Phase Sampling

2.1. Notations

In the two-phase sampling, it is possible to use the auxiliary information of the survey estimators such as population total, mean, and proportion. Then the auxiliary value may be one of two types: (i) values obtained by observing the units in first-phase sample \( s_1 \), that is, values that appear in the frame used for the second phase; (ii) values available at the outset for all \( N \) units of the population \( U \), that is, values given in the initial frame. In this framework, we assume the auxiliary information is available at two different levels, the target population and population obtained from the first-phase sample.
Notation will be as follows: (a) Let $x_k$ be a vector of $J$ auxiliary values available for all $k \in s_1$; (b) Let $x_{1k}$ be the vector of $J_1$ auxiliary values available for all $k \in U$. We assume that $x_k$ contains variable values known beforehand for all of $U$ as well as variable values known for $k \in s_1$ only. In other words, we can write the auxiliary vector $x_k = [x_{1k}, x_{2k}]$, where $x_{1k}$ be the vector of $J_1$ values known for all $k \in U$, and $x_{2k}$ is the vector of $J_2 = J - J_1$ values recorded by observation of units $k$ in the first-phase sample $s_1$ only. More precisely, we denote the auxiliary vectors as follows:

(i) $x_{1k}$ is known for all units $k \in U$, or that $\sum_{k \in l} x_{1k}$ is known and $x_{1k}$ observed for $k \in s_1$;
(ii) $x_{2k}$ is observed for all $k \in s_1$ and $k \in s_2$;
(iii) $y_k$ is observed for all $k \in s_2$.

From Sarndal et al. (1992), we define the inclusion probabilities of the first-phase and second-phase sampling.

Let us define $v_{1k}$ and $v_{1kl}$ as the inclusion probabilities of the first-phase with sampling design $p_1(\bullet)$ such that $p_1(s_1)$ as follows:

$$v_{1k} = \sum_{s_k} p_1(s_1), \quad v_{1kl} = \sum_{s_k \& l} p_1(s_1),$$

(2.1)

where $v_{1k}$ is the first-order inclusion probability for $k \in s_1$ and $v_{1kl}$ is the second-order inclusion probability for $k \& l \in s_1$, respectively.

For second-phase, the sampling design $p_2(\bullet \mid s_1)$ such that $p_2(s_2 \mid s_1)$, given $s_1$ as follows:

$$v_{2k \mid s_1} = \sum_{s_2} p_2(s_2 \mid s_1), \quad v_{2kl \mid s_1} = \sum_{s_2 \& l} p_2(s_2 \mid s_1),$$

(2.2)

where $v_{2k \mid s_1}$ is the first-order inclusion probability for $k \in s_2$ and $v_{2kl \mid s_1}$ is the second-order inclusion probability for $k \& l \in s_1$ given the first phase sample $s_1$, respectively. Hereafter, we use $v_{2k}$ and $v_{2kl}$ instead of $v_{2k \mid s_1}$ and $v_{2kl \mid s_1}$, for the second-phase inclusion probabilities, respectively.

So that the inclusion probability of $k$ element in the final sample is $v_k = v_{1k}v_{2k}$, letting the variable of interest be $y_k$ then the estimate of the population total is defined as

$$\hat{\tau}_y = \sum_{k \in k} \frac{y_k}{v_{1k}v_{2k}} = \sum_{k \in s_2} \frac{y_k}{v_k}.$$

This estimator is defined as $\pi^*$ estimator in Sarndal et al. (1992). Now, the first-phase sampling weight of unit $k$ is denoted as $d_{1k} = 1/v_{1k}$ and the second-phase sampling weight as $d_{2k} = 1/v_{2k}$. The overall sampling weight for a selected element $k \in s_2$ will be $d_k = d_{1k}d_{2k}$. Therefore, the estimate of the population total is rewritten as

$$\hat{\tau}_y = \sum_{k \in s_2} d_{1k}d_{2k}y_k = \sum_{k \in s_2} d_ky_k.$$

(2.3)

If the first-phase sampling design considers that a large sample of fixed size $n'$ is simple random sampling without replacement (SRSWOR), and a second-phase
The Calibration for Two-Phase Sample

sample $s_2$ of size $n$ is selected by a general sampling design $d_{2k} = 1/v_{2k}$, then the estimator (2.3) can be reduced as follows:

$$\hat{t}_y = \sum_{k \in s_2} \left( \frac{N}{n'} \right) d_{2k} y_k.$$  

(2.4)

2.2. Two-Phase Calibration Procedure

Basically, the calibration improves the original sampling weight, which has non coverage or nonresponse bias due to population change or refusal to response using auxiliary information on population or sample level. In this sub-section we can consider the two-step calibration procedure to improve the original sampling weight according to the original sampling scheme.

If there is available auxiliary information defined in Sec. 2.1, then we can calibrate the original sampling weights $d_{1k}$ and $d_{k} = d_{1k} d_{2k}$ to improve the original estimator. The new weights are denoted as $w_{1k}$ and $w_k = w_{1k} w_{2k}$ for each phase, then these values of the new weights are very close to the values of original sampling weights from the calibration procedure. As usual, calibration requires the specification of a distance function measuring the distance between the original weights and the new weights. Several distance functions have been proposed; see Deville and Sarndal (1992). Hidiroglou and Sarndal (1995) used the generalized least square (GLS) distance function. They derived the new calibrated weights, which minimize the GLS function successively in each phase and are subject to benchmark constraints.

At first-phase calibration, let the first-phase sampling weight $d_{1k}$ be original weights. Then determine the first-phase calibrated weights $w_{1k}$ by minimizing

$$G(w_{1k}/d_{1k}) = \sum_{s_1} \frac{(w_{1k} - d_{1k})^2}{2d_{1k}}.$$  

(2.5)

subject to the calibration equation

$$\sum_{s_1} w_{1k} x_{1k} = \sum_{U} x_{1k} = \tau_{s_1},$$  

(2.6)

where the total $\sum_{U} x_{1k}$ be assumed to known.

From (2.5) and (2.6), we can define the Lagrange function $L(w_{1k}/d_{1k})$ as follows:

$$L(w_{1k}/d_{1k}) = \sum_{s_1} G(w_{1k}/d_{1k}) - \lambda \left( \sum_{s_1} w_{1k} x_{1k} - \tau_{1k} \right).$$

Differentiating with respect to $w_{1k}$ and we set it equal to 0, then we can obtain the calibrated weights as follows:

$$w_{1k} = d_{1k} (1 + x_{1k} \lambda),$$  

(2.7)

where $F(\cdot) = (1 + x'_{1k} \lambda)$ and $\lambda$ is the unknown vector value of Lagrange multiplier in first-phase sampling.
From (2.7), the calibrated total of \( x_{1k} \) equals to the population total of \( x_{1k} \) as follows:

\[
\sum_{s_i} w_{ik} x_{ik} = \sum_{s_i} d_{1k} F(x'_{ik} \lambda_1) x_{ik} = \tau_{s_i},
\]  

(2.8)

We also use the initial weights \( w_{ik} d_{2k} \) at second-phase calibration, so that it is reasonable to use for second-phase weight because the final weight depends on the first-phase weight.

Similar to the first-phase calibration, the final (second-phase) weight \( w_k \) is determined by minimizing

\[
G(w_k/d_k) = \sum_{s_i} \frac{(w_k - w_{ik} d_{2k})^2}{2w_{ik} d_{2k}}
\]

subject to

\[
\sum_{s_i} w_k x_k = \sum_{s_i} w_{ik} x_k,
\]

(2.10)

where the auxiliary vector value \( x_k = (x'_{ik}, x'_{ik})' \).

The final weights can be obtained as follows:

\[
w_k = w_{ik} d_{2k}(1 + x'_{ik} \lambda_2)
\]

\[
= d_{1k} d_{2k}(1 + x'_{ik} \lambda_1)(1 + x'_{ik} \lambda_2)
\]

\[
= d_k F(x'_{ik} \lambda_1) F(x'_{ik} \lambda_2),
\]

(2.11)

where \( \lambda_2 \) is the unknown vector value of Lagrange multiplier in second-phase sampling.

If we consider the linear function in calibration as \( F(x' \lambda) = 1 + x' \lambda \), then the calibrated weight can be written as \( g \)-factor. In this respect, Hidiroglou and Sarndal (1995) considered the multiplicative and additive cases. From their works, the distance function is given by

\[
\sum_{x} \frac{(w_k - w_{ik} d_{2k})^2}{2d_{1k} d_{2k}},
\]

(2.12)

for the additive case for second-phase calibration.

The final calibrated weight resulting from minimizing (2.12) subject to (2.10) is

\[
w_k = d_k [F(x'_{ik} \lambda_1) + F(x'_{ik} \lambda_2) - 1].
\]

(2.13)

### 2.3. New Two-Step Newton’s Method

Here, we propose the new two-step Newton’s method for determining the values of the Lagrange multipliers, \( \lambda_1 \) and \( \lambda_2 \), in the calibrated weights (2.7) and (2.11). Our proposed iterative method is an extension of Deville et al. (1993) to two-step...
according to each phase of sample. Let’s define a function of $\lambda_1$ for the first-phase as follows:

$$
\phi_{s_1}(\lambda_1) = \sum_{s_1} d_{ik} F(x_{ik}', \lambda_1) x_{ik} - \sum_{s_1} d_{ik} x_{ik}
= \sum_{s_1} d_{ik} [F(x_{ik}', \lambda_1) - 1] x_{ik} = \tau_{1k} - \hat{\tau}_{1k}.
$$

(2.14)

Using the Newton’s method, we can find the unknown $\lambda_1$. Let $\phi'_1(\lambda_1) = \frac{\partial \phi_{s_1}(\lambda_1)}{\partial \lambda_1}$. The initial value of $\lambda_1$ is $\lambda_1^{(0)} = 0$. Subsequent iterative values $\lambda_1^{(v)}$, $v = 2, 3, \ldots$, are obtained from

$$
\lambda_1^{(v+1)} = \lambda_1^{(v)} + \left[ \phi'_1(\lambda_1^{(v)}) \right]^{-1} \left[ \tau_{1k} - \hat{\tau}_{1k} - \phi_{s_1}(\lambda_1^{(v)}) \right],
$$

(2.15)

where $\phi_{s_1}(0) = 0$ and $\phi'_1(0) = \sum d_{ik} x_{ik}' x_{ik}$, subsequent iterative values $\lambda_1^{(v)}$, $v = 2, 3, \ldots$, follow (2.15) until convergence.

For GLS distance function, (2.15) converges after a single iteration as follows:

$$
\lambda_1^{(1)} = \left[ \phi'_1(0) \right]^{-1} \left[ \tau_{1k} - \hat{\tau}_{1k} \right] = \left[ \sum_{s_1} d_{ik} x_{ik}' x_{ik} \right]^{-1} \left[ \tau_{1k} - \hat{\tau}_{1k} \right].
$$

(2.16)

From (2.16), we can obtain the function of $\lambda_1$, $F(x_{ik}', \lambda_1)$ as follows:

$$
F(x_{ik}', \lambda_1) = 1 + x_{ik}' \lambda_1
= 1 + \left[ \sum_{s_1} d_{ik} x_{ik}' x_{ik} \right]^{-1} \left[ \tau_{1k} - \hat{\tau}_{1k} \right] x_{ik}.
$$

(2.17)

If the auxiliary variable for the first phase is $x_{ik} = 1$ with simple random sampling without replacement (SRSWOR), then $F(x_{ik}', \lambda_1)$ becomes

$$
F(\lambda_1) = 1 + \left[ \sum_{s_1} d_{ik} x_{ik}' x_{ik} \right]^{-1} \left[ \tau_{1k} - \hat{\tau}_{1k} \right] x_{ik}
= 1 + \left[ \sum_{s_1} d_{ik} \right]^{-1} \left[ N - \sum_{s_1} d_{ik} \right] = N/\bar{N}_1,
$$

(2.18)

where $\bar{N}_1 = \sum_{s_1} d_{ik}$ is the sum of first-phase weights.

Similar to (2.14), we define a function of $\lambda_2$ as follows:

$$
\phi_{s_2}(\lambda_2) = \sum_{s_2} d_k F(x_{ik}', \lambda_2) [F(x_{ik}', \lambda_2) - 1] x_{ik} = \tau_{2k} - \hat{\tau}_{2k}.
$$

(2.19)

We set $\phi'_s(\lambda_2) = \frac{\partial \phi_{s_2}(\lambda_2)}{\partial \lambda_2}$ and the initial value of $\lambda_2$ is $\lambda_2^{(0)} = 0$. Subsequent iterative values $\lambda_2^{(v)}$, $v = 2, 3, \ldots$ are obtained from

$$
\lambda_2^{(v+1)} = \lambda_2^{(v)} + \left[ \phi'_2(\lambda_2^{(v)}) \right]^{-1} \left[ \tau_{2k} - \hat{\tau}_{2k} - \phi_{s_2}(\lambda_2^{(v)}) \right],
$$

(2.20)

where $\phi_{s_2}(0) = 0$ and $\phi'_s(0) = \sum_{s_2} d_k F(x_{ik}', \lambda_2) x_{ik}' x_{ik}$.
Then the first iteration gives
\begin{equation}
\lambda_2^{(1)} = \left[ \phi_1'(0) \right]^{-1} [\hat{\tau}_1 - \hat{\tau}_2] = \left[ \sum_{s_2} d_i F(x'_{1k}\lambda_1) x_k \right]^{-1} [\hat{\tau}_1 - \hat{\tau}_2],
\end{equation}
and subsequent iterative values \( \lambda_2^{(v)} \), \( v = 2, 3, \ldots \), follow (2.20) until convergence.

From (2.21), we can obtain the function of \( \lambda_2 \), \( F(x'_1\lambda_2) \) as follows:
\begin{equation}
F(x'_1\lambda_2) = 1 + x'_1\lambda_2
= 1 + \left[ \sum_{s_2} w_{1k} d_{2k} x'_1 x_k \right]^{-1} [\hat{\tau}_1 - \hat{\tau}_2] x_k.
\end{equation}

If the auxiliary variable for the second phase is \( x_k = x_{1k} = x_{2k} = 1 \) with SRSWOR, then the \( F(x'_k\lambda_2) \) becomes
\begin{equation}
F(\lambda_2) = 1 + \left[ \sum_{s_2} w_{1k} d_{2k} x'_1 x_k \right]^{-1} [\hat{\tau}_1 - \hat{\tau}_2] x_k
= 1 + \left[ \sum_{s_2} d_k \frac{N}{N_1} \right]^{-1} \left[ N - \sum_{s_2} d_k \right]
= 1 + \hat{N}_1 \left( \frac{1}{N_2} - \frac{1}{N} \right),
\end{equation}
where \( \hat{N}_2 = \sum_{s_2} d_k \) is the sum of second phase weights.

**Theorem 2.1.** Under the assumption of the auxiliary variable with \( x_k = x_{1k} = x_{2k} = 1 \) and the sampling design is SRSWOR in both phases and the first-phase sample of size \( n' \) is as large as \( N \), the multiplicative case of the final weight \( w_{mk} \) (say) suggested by Hidiroglou and Sarndal (1995) is equal to the additive case \( w_{ak} \) (say):
\begin{equation}
w_{mk} = d_k F(x'_{1k}\lambda_1) F(x'_2\lambda_2) \approx d_k [F(x'_{1k}\lambda_1) + F(x'_2\lambda_2) - 1] = w_{ak}.
\end{equation}

**Proof.** If the first-phase sample of size \( n' \to N \), that is, the size of first sample is as large as the size of population, using (2.18) and (2.23), we can obtain the final calibrated weight \( w_{ak} \) by the additive case as follows:
\begin{align*}
w_{ak} &= d_k [F(x'_{1k}\lambda_1) + F(x'_2\lambda_2) - 1]
= d_k \left[ \frac{N}{N_1} + \frac{\hat{N}_1}{N_2} - \frac{\hat{N}_1}{N} \right] = d_k (N/\hat{N}_2).
\end{align*}
Also, the multiplicative form is given by
\begin{align*}
w_{mk} &= d_k F(x'_{1k}\lambda_1) F(x'_2\lambda_2)
= d_k \frac{N}{N_1} \left[ 1 + \hat{N}_1 \left( \frac{1}{N_2} - \frac{1}{N} \right) \right] = d_k (N/\hat{N}_2).
\end{align*}
From (2.24), we know the additive case of Hidiroglou and Sarndal (1995) asymptotically equals the multiplicative case, so that it is reasonable to make use of the multiplicative case weight as the calibrated final weight with respect to the auxiliary variable vector \( x_k = x_{1k} = x_{2k} = 1 \).

### 2.4. Variance and Variance Estimator

Deville and Sarndal (1992) showed that the calibration estimator is asymptotically equivalent with the generalized regression (GREG) estimator. In this point of view, Hidiroglou and Sarndal (1995) used the variance and variance estimator of GREG estimator for the calibration variance and variance estimator in two-phase sampling.

The variance estimator for calibration estimator is given by

\[
\hat{V} = \sum_{s_1} \sum_{s_2} d_{1k}(d_{1k}d_{1l} - d_{1kl})(F(x'_{1k}e_{1k})(F(x'_{1l}e_{1l})) + \sum_{s_2} d_{1k}d_{1l}(d_{2k}d_{2l} - d_{2kl})(F(x'_{2k}e_{2k})(F(x'_{2l}e_{2l})),
\]

(2.25)

where the sample residuals are \( e_{1k} = y_k - x_{1k}\hat{B}_1 \) for \( k \in s_1 \) and \( e_{2k} = y_k - x_k\hat{B}_2 \) for \( k \in s_2 \), \( \hat{B}_1 \) and \( \hat{B}_2 \) are given by

\[
\hat{B}_1 = \left( \sum_{k \in s_1} d_{1k}x_{1k}^2 x_{1k} \right)^{-1} \left[ \sum_{k \in s_1} d_{1k}x_{1k}\hat{y}_{2k} + \sum_{k \in s_2} d_{k}x_{1k}(y_k - \hat{y}_{2k}) \right], \quad \text{with} \quad \hat{y}_{2k} = x_k\hat{B}_2, \quad \text{and}
\]

\[
\hat{B}_2 = \left( \sum_{k \in s_2} d_kF(x'_{1k}x_{1k})x_{1k} \right)^{-1} \left( \sum_{k \in s_2} d_kF(x'_{1k}x_{1k})x_k \right), \quad \text{respectively.}
\]

### 3. The Two-Phase Randomized Response Technique

#### 3.1. Two-Phase RR Estimators

The RR model to obtain truthful data for estimating the true proportion \( \pi \) of the population possessing the sensitive characteristic was first suggested by Warner (1965). In this survey technique, each individual respondent is provided with a randomization device by which he/she chooses one of the two questions: “Do you belong to sensitive group \( A \)” or “Do you belong to non sensitive group \( A' \)” with respective probabilities \( P \) and \( 1 - P \) \( (0 < P < 1, P \neq 0.5) \) and answers “Yes” or “No” to the question chosen. Assuming truthful reporting, the probability of obtaining “Yes” response is

\[
Z = P\pi + (1 - P)(1 - \pi).
\]

(3.1)

The maximum likelihood estimate of \( \pi \) is shown to be

\[
\hat{\pi}_w = \frac{P - 1}{2P - 1} + \frac{\widehat{Z}}{2P - 1} = \frac{P - 1}{2P - 1} + \frac{1}{2P - 1} \sum_{k=1}^{n} Z_k / n,
\]

(3.2)

where \( \widehat{Z} = \sum_{k=1}^{n} Z_k / n \) is the proportion of “Yes” answer in a sample selected by SRSWR.
We can extend the Warner RR model to a two-phase RR technique with a general sampling scheme for each phase. With the first-phase sample drawn from a population, we investigate to get the population information from a selected sample of size $n'$. Then the second-phase sample is selected from the first-phase sample by incorporating auxiliary information. The final sampling units are provided with a randomization device and he (or she) chooses one of the two questions: “Do you belong to sensitive group $A$?” or “Do you belong to sensitive group $A'$?” with selection probabilities $P$ and $(1 - P)$, and replies “Yes” or “No” to the question.

Let $Z_k$ be the binary indicator taking value 1, if a unit response belongs to second phase “Yes” to RR question and 0 otherwise. Then the population total of reporting “Yes” to RR question is defined by \( \sum Z_k \) and the sample proportion is \( \hat{Z} = \frac{1}{N} \sum d_k Z_k \). By the assumptions of population and sample in Sec. 2.1, $Z_k$ is only observed in the second-phase of sample. Then the estimator for a true population proportion $\pi$ of possessing sensitive characteristic is given by

\[
\hat{\pi}_{\text{two}} = \hat{Z} - \frac{(1 - P)}{2P - 1} = \frac{P - 1}{2P - 1} + \frac{1}{2P - 1} \frac{1}{N} \sum d_k Z_k,
\]

where $\hat{Z} = \frac{1}{N} \sum d_k Z_k$ is the sample proportion of “Yes” reporting to sensitive question and $d_k = d_k^1 d_k^2$ is the original sampling weight corresponding to the first- and second-phase sampling.

**Theorem 3.1.** The estimator (3.3) is unbiased for the population proportion $\pi$ of a sensitive characteristic, i.e., $E(\hat{\pi}_{\text{two}}) = \pi$.

**Proof.** From Sarndal et al. (1992, Ch. 9), the expectation of the two-phase estimator defined as (3.3) is as follows:

\[
E(\hat{\pi}_{\text{two}}) = E_{s_1} E_{s_2}(\hat{\pi}_{\text{two}} | s_1) = E_{s_1} \left[ \frac{P - 1}{2P - 1} + \frac{1}{2P - 1} \frac{1}{N} \sum d_k Z_k \right]
\]

\[
= \frac{P - 1}{2P - 1} + \frac{1}{2P - 1} \frac{1}{N} \sum Z_k = \pi.
\]

To simplify, we can specify the two cases corresponding to the first- and second-phase sampling design as follows: (i) the first-phase calls for drawing as SRSWOR of $n'$ units from $N$, and the second-phase calls for the stratified sampling; (ii) the first-phase calls for drawing as SRSWOR of $n'$ units from $N$, and the second-phase calls for the arbitrary sampling.

**Case 1.** If the first-phase sampling design considers that a large sample of fixed size $n'$ is SRSWOR, and the second-phase design is stratified sampling, then the phase-one sampling fraction is $f_1 = n'/N$, the relative size of stratum $h$ is $w_h = n'_h/n'$, and the phase-two sampling fraction $f_h = n_h/n'_h$ is. So, the estimator of the population proportion for a sensitive character can be reduced as follows:

\[
\hat{\pi}_{s_1} = \frac{P - 1}{2P - 1} + \frac{1}{2P - 1} \frac{1}{n} \sum_{h=1}^{H_{s_1}} \frac{n'_h}{n_h} \sum_{k=1}^{n_h} Z_k,
\]

where $H_{s_1}$ denotes the number of strata of first phase sample $s_1$. 


Case 2. If the first-phase sampling design considers that a large sample of fixed
size \(n'\) is SRSWOR, and a second-phase sample \(s_2\) of size \(n\) is selected by a general
sampling design \(d_{2k} = 1/\upsilon_{2k}\), then the estimator is rewritten as follows:

\[
\hat{\pi}_2 = \frac{P - 1}{2P - 1} + \frac{1}{2P - 1} \frac{1}{n'} \sum_{s_2} d_{2k} Z_k.
\]

(3.5)

Remark 3.1. Two-phase RR estimators (3.4) and (3.5) are the type of two-phase
estimators of Sarndal et al. (1992) and Wu and Luan (2003), respectively.

3.2. Variance and Variance Estimator of Two-Phase RR Technique

In general, the sampling design with the first and the second sampling weights are
\(d_{1k} = 1/\upsilon_{1k}\), \(d_{2k} = 1/\upsilon_{2k}\), \(d_{3kl} = 1/\upsilon_{3kl}\), and \(d_{2kl} = 1/\upsilon_{2kl}\); then the variance and its
estimator for true proportion of possessing sensitive character are given by

\[
V(\hat{\pi}_{two}) = \left(\frac{1}{2P - 1}\right)^2 \frac{1}{N^2} \left[ \sum_{h} \sum_{i} \frac{(d_{1k} d_{1l} - d_{1kl})}{d_{3kl}} Z_k Z_l \right.
\]

\[
+ E_1 \left( \sum_{h} \sum_{i} \frac{(d_{2k} d_{2l} - d_{2kl}) (d_{1k} Z_k)(d_{1l} Z_l)}{d_{3kl}} \right) \right],
\]

(3.6)

and

\[
\hat{V}(\hat{\pi}_{two}) = \left(\frac{1}{2P - 1}\right)^2 \frac{1}{N^2} \left[ \sum_{s_2} d_{2kl}(d_{1k} d_{1l} - d_{1kl}) Z_k Z_l 
\]

\[
+ \sum_{s_2} (d_{2k} d_{2l} - d_{2kl})(d_{1k} Z_k)(d_{1l} Z_l) \right],
\]

(3.7)

respectively.

We can obtain the population variance of the true proportion of possessing a
sensitive character with respect to sampling design for each phase. The variance (3.6)
can be reduced to the specific form as follows.

Theorem 3.2. If the first-phase is SRSWOR and the second phase is a stratified
sampling for each phase, then the variance is given by

\[
V(\hat{\pi}_{two}) = \left(\frac{1}{2P - 1}\right)^2 \left[ \left(\frac{1}{n'} - \frac{1}{N}\right) V_1 + E_1 \sum_{h=1}^{H} \frac{n_h}{n'} \left(\frac{1}{n_p} - \frac{1}{n_h}\right) V_{2h} \right],
\]

(3.8)

where \(V_1 = Z(1 - Z)\) with \(Z = (1 - P) + (2P - 1)\pi\) is a population proportion of
reporting “Yes”, and \(V_{2h} = Z_h(1 - Z_h)\) with \(Z_h = (1 - P) + (2P - 1)\pi_h\) is a sample
proportion of reporting “Yes” to sensitive question in \(h\) strata.
Proof. From Sarndal et al. (1992, Ch. 9), the variance of the estimator of a population total is

\[
V(\hat{\pi}) = N^2 \left( \frac{1}{n^*} - \frac{1}{N} \right) S^2_{SU} + E_1 \left( N^2 \sum_{h=1}^{H} \left( \frac{n'_h}{n^*} - \frac{1}{n_h'} \right) S^2_{sys} \right).
\]

Let’s replace the estimator of population total by that of population proportion and \( S^2_{yU} \) and \( S^2_{ys1h} \) into \( V_1 = Z/l(1 - Z) \) and \( V_2 = Z_h(1 - Z_h) \), respectively. Then we obtain the variance of two-phase RR estimator for the population proportion of a sensitive character. It is reasonable that the two-phase RR variance is the type of typical two-phase sampling variance derived by Sarndal et al. (1992). □

Corollary 3.1. From (3.6), if \( E_1(n'_h/n^*) = N_h/N \) and \( n_h = \kappa_h n'_h = \kappa_h (n'_h/n^*) n' \), with a priori fixed constant \( \kappa_h \), and the second-phase is carried out by the stratified random sampling, then the variance can be rewritten as follows:

\[
V(\tilde{\pi}) = \left( \frac{1}{n^*} - \frac{1}{N} \right) \left[ \pi(1 - \pi) + \frac{P(1 - P)}{2P - 1} \right]
+ \sum_{h=1}^{H} \left( \frac{N_h}{N} \right) \left[ \frac{\pi_h(1 - \pi_h)}{n'} + \frac{P(1 - P)}{n'(2P - 1)^2} \right] \left( \frac{1}{\kappa_h} - 1 \right).
\]  

(3.9)

Proof. From (3.8), we define the population proportion of reporting “Yes” as \( Z = (1 - P) + (2P - 1)\pi \) and the variance component is \( V_1 = Z(1 - Z) \). The second component of variance is \( V_2 = Z_h(1 - Z_h) \) and a sample proportion of reporting “Yes” to sensitive question in \( h \) strata is \( Z_h = (1 - P) + (2P - 1)\pi_h \). Thus, we can obtain the variance of two-phase RR estimator given by (3.9) after some algebra. □

4. Calibration for Two-Phase RR Estimator

In this section, we consider the calibration procedure for two-phase RR estimator according to the levels of auxiliary variable. To do this, we classify the two cases. The first case is one-way classification such as the two-phase stratified RR model, and the second case is the two-way classification.

The two-phase RR estimator for a true population proportion \( \pi \) defined by (3.3) is calibrated in the first- and second-phase using auxiliary information. From the ordinary two-phase calibration procedure described in Sec. 2, the calibrated RR estimator substituted by the new weight \( w_h \) is given by

\[
\hat{\pi}_{cal} = \frac{P - 1}{2P - 1} + \frac{1}{2P - 1} \frac{1}{N} \sum_s w_k Z_k
= \frac{P - 1}{2P - 1} + \frac{1}{2P - 1} \frac{1}{N} \sum_s d_i F(x'_i, \lambda_1) F(x'_i, \lambda_2) Z_k,
\]  

(4.1)

where the Lagrange multipliers \( \lambda_1 \) and \( \lambda_2 \) are computed by (2.16) and (2.21), respectively.
4.1. One-Way Classification

We consider the structure of the population and sample as one-way classification, so that the auxiliary variable vector in first-phase is \( x_{1k} = \Gamma_k \) with \( \Gamma_k = (\delta_{1k}, \delta_{2k}, \ldots, \delta_{hk}, \ldots, \delta_{ik}) \), where for \( h = 1, \ldots, H \), \( \delta_{ik} = 1 \), if the respondent \( k \) is in stratum \( h \) and 0 otherwise. The calibration equation (2.6) for the first-phase sample becomes

\[
\sum_{k \in s_1} w_{1k} x_{1k} = \sum_{k \in U} x_{1k} = (N_1, N_2, \ldots, N_h, \ldots, N_H)'.
\] (4.2)

In this calibration procedure, for every \( k \) in stratum \( h \), \( x_{1k} \hat{\lambda}_1 \) is constant and equal to \( \hat{\lambda}_1 \). Then \( F(x_{1k} \hat{\lambda}_1) = F(\hat{\lambda}_1) = \hat{N}_h/\tilde{N}_{1h} \), where \( \hat{N}_h = \sum_{i=1}^{n_h} d_{ik} = (Nn_h)/n' \) and \( n' = \sum_{h=1}^{H} n_h' \) is the size of first-phase sample and \( \sum_{i=1}^{n_h} x_{1k} = (n_1', n_2', \ldots, n_h', \ldots, n_H') \). Thus, the calibrated weights (2.7) in the first-phase become \( \hat{w}_{1k} = d_{1k} \hat{N}_h/\tilde{N}_h \) for all \( k \) in stratum \( h \).

Next, in order to calibrate from the second-phase to the first-phase, we define the available auxiliary vector as \( x_k = (x'_{1k}, x'_{2k})' = x'_{1k} \); then the calibration Eq. (2.10) for the second-phase sample becomes

\[
\sum_{s_2} w_{2k} x_{2k} = \sum_{s_1} w_{1k} x_{1k} = (\tilde{N}_{11}, \tilde{N}_{12}, \ldots, \tilde{N}_{1h}, \ldots, \tilde{N}_{1H})',
\] (4.3)

where \( \tilde{N}_{1h} = \sum_{s_{1h}} d_{1k} = (Nn_h)/n' \).

From Sec. 2.3, we can obtain the final weight \( w_k \) as follows:

\[
w_k = d_k N_h \left( \frac{1}{\tilde{N}_{1h}} + \frac{1}{\tilde{N}_{2h}} - \frac{1}{N_h} \right),
\] (4.4)

where \( \tilde{N}_{2h} = \sum_{s_{2h}} d_k = n_h(N/n) \).

Remark 4.1. If \( 1/N_h \) is negligible in (4.4), then the final weight \( w_k \) reduced to the additive form of the first- and second-phase sampling weight is as follows:

\[
w_k = d_k N_h \left( \frac{1}{\tilde{N}_{1h}} + \frac{1}{\tilde{N}_{2h}} \right).
\]

Result 4.1. The calibration estimator of \( \pi \) is given by

\[
\hat{\pi}_{cal} = \frac{P - 1}{2P - 1} + \frac{1}{2P - 1} \frac{1}{N} \sum_{s_2} w_k Z_k
\]

\[
= \frac{P - 1}{2P - 1} + \left( \frac{1}{2P - 1} \right) \left[ \frac{1}{N} \sum_{h=1}^{H} \sum_{k=1}^{n_h} d_k N_h \left( \frac{1}{\tilde{N}_{1h}} + \frac{1}{\tilde{N}_{2h}} - \frac{1}{N_h} \right) Z_k \right]
\]

\[
= \frac{P - 1}{2P - 1} + \left( \frac{1}{2P - 1} \right) \left[ \frac{1}{N} \sum_{h=1}^{H} N_h \tilde{Z}_{1h} + \frac{1}{N} \sum_{h=1}^{H} \sum_{k=1}^{n_h} N_h \tilde{Z}_{2h} - \frac{1}{N} \sum_{h=1}^{H} \sum_{k=1}^{n_h} d_k Z_k \right],
\] (4.5)

where \( \tilde{Z}_{1h} = \sum_{k=1}^{n_h} d_k Z_k/\tilde{N}_{1h} \) and \( \tilde{Z}_{2h} = \sum_{k=1}^{n_h} d_k Z_k/\tilde{N}_{2h} \).
Theorem 4.1. From Result 4.1, if the first-phase sample size of \( n' \) is as large as the population size of \( N \), then the calibration estimator with two-phase sampling reduces to the single-phase calibration estimator. That calibration estimator is well known as the post-stratified estimator for the stratified RR design

\[
\hat{\pi}_{\text{cal}} = \frac{P - 1}{2P - 1} + \left( \frac{1}{2P - 1} \right) \frac{1}{N} \sum_{k=1}^{H} \sum_{l=1}^{n_h} d_k N_h \left( \frac{1}{N_{2h}} - \frac{1}{N_h} \right) Z_k
\]

where \( Z_k = \sum_{l=1}^{n_h} d_l Z_l/N_h \).

Proof. If the first-phase sample of size \( n' \to N \), then it means \( n'_h \to N_h \). The calibration estimator (4.4) can be rewritten as follows:

\[
\hat{\pi}_{\text{cal}} = \frac{P - 1}{2P - 1} + \left( \frac{1}{2P - 1} \right) \frac{1}{N} \sum_{k=1}^{H} \sum_{l=1}^{n_h} d_k N_h \left( \frac{1}{N_{2h}} + \frac{1}{N_{1h}} - \frac{1}{N_h} \right) Z_k
\]

Hence, the calibration estimator for the true population proportion of a sensitive trait reduces to the post-stratified estimator as \( n' \to N \). \( \square \)

Remark 4.2. Theorem 4.1 verifies the properties of two-phase sampling design, and the calibration estimator on two-phase sampling reduces to that of single-phase sampling. This fact is consistent with the ordinary single-phase calibration procedure.

4.2. Two-Way Classification

Different from Sec. 4.1, here we consider the calibration procedure with two-way classification with known cell counts. Let’s define the auxiliary variable vector \( x_{1k} = (1) \), with \( \Gamma_k = (\delta_{11k}, \delta_{12k}, \ldots, \delta_{j1k}, \ldots, \delta_{ek})' \), where \( \delta_{jk} = 1 \), if the first-phase sample \( k \) is in \((i, j)\) cell and 0 otherwise.

The population total of \( x_{1k} \), \( \tau_{1k} = \sum_{k \in U} x_{1k} \) is given in the form of known cell counts in a frequency table in two dimensions as follows:

\[
\sum_{k \in U} x_{1k} = (N_{11}, N_{12}, \ldots, N_{ij}, \ldots, N_n).
\]

In this calibration procedure, for every \( k \) in given cell \((i, j)\), \( x'_{1k} \hat{\lambda}_1 \) is constant and equals to \( \hat{\lambda}_{i'j'} \). Then \( F(x'_{1k} \hat{\lambda}_1) = F(\hat{\lambda}_{i'j'}) = N_{i'j'}/\tilde{N}_{i'j'} \), where \( \tilde{N}_{i'j'} = \sum_{i'j' \in k} d_{1k} = (n'_{i'j'}/n' \to \). \( n' = \sum_{i'j'=1}^{n'} \sum_{j'=1}^{n'_{i'}} n'_{i'j'} \) is the size of the first-phase sample and \( \sum_{i'j' \in k} x'_{1k} = (n'_{i'1}, n'_{i'2}, \ldots, n'_{i'n'}, n'_{n'}) \). Thus, the calibrated weights for the first-phase become \( w_{1k} = d_{1k} N_{i'j'}/\tilde{N}_{i'j'} \) for all \( k \) in cell \((i, j)\).
Next, we can calibrate the final weight in the second-phase using auxiliary variable defined by 
\[ x_k = (x'_{1k}, x'_{2k})' = x_{1k} = \Gamma_k. \] 
The final weight \( w_k \) can be obtained by
\[ w_k = d_k N_{ij} \left( \frac{1}{N_{ij}} + \frac{1}{N_{2ij}} - \frac{1}{N_{ij}} \right). \] 
(4.8)

where \( \hat{N}_{1ij} = \sum_{n_{ij}} d_{1k} \) and \( \hat{N}_{2ij} = \sum_{n_{ij}} d_{k} \).

**Result 4.2.** The calibration estimator of \( \pi \) is given by
\[ \hat{\pi}_{cal} = \frac{P - 1}{2P - 1} + \frac{1}{2P - 1} \frac{1}{N} \sum_{s_2} w_i Z_k \]
\[ = \frac{P - 1}{2P - 1} + \left( \frac{1}{2P - 1} \right) \left[ \frac{1}{N} \sum_{i=1}^{c} \sum_{j=1}^{r} N_{ij} \hat{Z}_{1ij} + \frac{1}{N} \sum_{i=1}^{c} \sum_{j=1}^{r} N_{ij} \hat{Z}_{2ij} - \frac{1}{N} \sum_{i=1}^{c} \sum_{j=1}^{r} \sum_{k=1}^{n_{ij}} d_k Z_k \right], \] 
(4.9)

where \( \hat{Z}_{1ij} = \sum_{k=1}^{n_{ij}} d_k Z_k / \hat{N}_{1ij} \) and \( \hat{Z}_{2ij} = \sum_{k=1}^{n_{ij}} d_k Z_k / \hat{N}_{2ij} \).

## 5. Variances of the Two-Phase Calibration RR Estimators

In this section, we derive the variance of two-phase calibration RR estimators (4.5) and (4.9) for each case. To begin with, we consider the one-way classification from Sec. 4.1, and then two-way classification from Sec. 4.2. In order to derive the variance of the calibration estimator, we let the leading term of the calibration RR estimator be \( \hat{Z}_{cal} = N^{-1} \sum_{s_2} w_i Z_k \), then the estimation errors are decomposed as follows:
\[ \hat{Z}_{cal} - Z = \left( \frac{1}{N} \sum_{s_1} w_{1k} Z_k - \frac{1}{N} \sum_{U} Z_k \right) + \left( \frac{1}{N} \sum_{s_2} w_i Z_k - \frac{1}{N} \sum_{s_1} w_{1k} Z_k \right). \]
(5.1)

The first term of the right-hand side of Eq. (5.1) is called the error due to the first-phase of sampling and the second term is called the error due to the second phase. According to the two-phase sampling design, we can obtain the variance of the calibration estimator using the two-phase unconditional variance formulation
\[ V(\hat{Z}_{cal}) = V_1 E_2(\hat{Z}_{cal}) + E_1 V_2(\hat{Z}_{cal}), \]

where \( V_1 \) and \( E_1 \) denote the first-phase variance and expectation, and \( V_2 \) and \( E_2 \) are the second-phase variance and expectation.

### 5.1. One-Way Classification

We consider that the estimation error due to the two-phase calibration procedure for RR estimator is given by
\[ \hat{Z}_{cal} - Z = \left( \frac{1}{N} \sum_{h=1}^{H} \sum_{k=1}^{n_{1h}} \frac{N_{1h}}{N_{1h}} Z_k - \frac{1}{N} \sum_{k=1}^{N} Z_k \right) \]
\[ + \left( \frac{1}{N} \sum_{h=1}^{H} \sum_{k=1}^{n_{1h}} d_{1k} N_{1h} \left( \frac{1}{N_{1h}} + \frac{1}{N_{2h}} - \frac{1}{N_{1h}} \right) Z_k - \frac{1}{N} \sum_{h=1}^{H} \sum_{k=1}^{n_{1h}} d_{1k} \frac{N_{1h}}{N_{1h}} Z_k \right). \]
where

\[ V = \left( \frac{1}{N} \sum_{h=1}^{H} \sum_{k=1}^{n_h} d_{ik} \frac{N_h}{N_{1h}} Z_k - \frac{1}{N} \sum_{k=1}^{N} Z_k \right) \]

\[ + \left( \frac{1}{N} \sum_{h=1}^{H} \sum_{k=1}^{n_h} d_k \frac{N_h}{N_{2h}} Z_k - \frac{1}{N} \sum_{k=1}^{N} d_k Z_k \right) , \quad (5.2) \]

The last term of the right-hand side of Eq. (5.2) becomes 0 or nearly 0 because \( n' \to N \), so that the estimation error of estimator can be expressed as follows:

\[ \tilde{Z}_{cal} - Z = \left( \frac{1}{N} \sum_{h=1}^{H} \sum_{k=1}^{n_h} d_{ik} \frac{N_h}{N_{1h}} Z_k - \frac{1}{N} \sum_{k=1}^{N} Z_k \right) \]

\[ + \left( \frac{1}{N} \sum_{h=1}^{H} \sum_{k=1}^{n_h} d_k \frac{N_h}{N_{2h}} Z_k - \frac{1}{N} \sum_{k=1}^{N} d_k Z_k \right) . \quad (5.3) \]

We can obtain the variance by having the calibrated RR estimator substituted into the new weight for the first term in (5.1) as follows:

\[ V_1 E_2(\tilde{Z}_{cal}) = V_1 \left( \frac{1}{N} \sum_{h=1}^{H} \sum_{k=1}^{n_h} d_{ik} \frac{N_h}{N_{1h}} Z_k - \frac{1}{N} \sum_{k=1}^{N} Z_k \right) = \sum_{h} \left( \frac{N_h}{N} \right)^2 \left( \frac{1}{n_h} - \frac{1}{N_h} \right) V_{1h} , \]

where \( V_{1h} = Z_{1h}(1 - Z_{1h}) \) with \( Z_{1h} = \hat{N}_{1h}^{-1} \sum_{k=1}^{n_h} d_{ik} Z_k \).

The variance of second term is written as follows:

\[ E_1 V_2(\tilde{Z}_{cal}) = E_1 V_2 \left[ \frac{1}{N} \sum_{h=1}^{H} \sum_{k=1}^{n_h} d_k \frac{N_h}{N_{2h}} Z_k - \frac{1}{N} \sum_{h=1}^{H} \sum_{k=1}^{n_h} d_k \frac{N_h}{N_{1h}} Z_k \right] \]

\[ = E_1 V_2 \left( \frac{1}{N} \sum_{h=1}^{H} \sum_{k=1}^{n_h} d_k \frac{N_h}{N_{2h}} Z_k - \frac{1}{N} \sum_{h=1}^{H} \sum_{k=1}^{n_h} d_k \frac{N_h}{N_{1h}} Z_k \right) \]

\[ = \sum_{h} \left( \frac{N_h}{N} \right)^2 \left( \frac{1}{n_h} - \frac{1}{N_h} \right) V_{2h} , \]

where \( V_{2h} = Z_{2h}(1 - Z_{2h}) \) with \( Z_{2h} = \hat{N}_{2h}^{-1} \sum_{k=1}^{n_h} d_k Z_k \).

Thus, the variance components of the leading term of the calibration RR estimator for two-phase sampling is

\[ AV(\tilde{Z}_{cal}) = \sum_{h} \left( \frac{N_h}{N} \right)^2 \left( \frac{1}{n_h} - \frac{1}{N_h} \right) V_{1h} + \sum_{h} \left( \frac{N_h}{N} \right)^2 \left( \frac{1}{n_h} - \frac{1}{N_h} \right) V_{2h} . \quad (5.4) \]

**Result 5.1.** From (5.4), we can obtain the asymptotic variance of the calibration RR estimator for the population proportion in two-phase sampling design as
obtain the first component of variance as follows:

\[ \text{Var}(\hat{\pi}_{\text{cal}}) = \sum_h \left( \frac{N_h}{N} \right)^2 \left[ \frac{\pi_h (1 - \pi_h)}{n_h} \left( \frac{N_h - n_h}{N_h - 1} \right) + \frac{P(1 - P)}{n_h^2(2P - 1)^2} \right] + \sum_h \left( \frac{N_h}{N} \right)^2 \left[ \frac{\pi_h (1 - \pi_h)}{N_h^2} \left( \frac{n_h' - n_h}{n_h'^2 - 1} \right) + \frac{P(1 - P)}{n_h(2P - 1)^2} \right]. \]  \hspace{1cm} (5.5)

Meanwhile, as discussed in Deville et al. (1993), the asymptotic calibration variance (5.5) is the conditional variance given \( \hat{N}_h \), since the calibrated weight is obtained from \( \hat{N}_h \). Thus, we have to consider the unconditional variance formula of (5.5). The unconditional variance, \( V(\hat{\pi}_{\text{cal}}) = E(V(\hat{\pi}_{\text{cal}}) | \hat{N}_h) \) is given by the following one.

**Result 5.2.** The unconditional asymptotic variance by post-stratification is

\[ V(\hat{\pi}_{\text{cal}}) = \sum_h \left( \frac{N_h}{N} \right)^2 \left[ \frac{\pi_h (1 - \pi_h)}{n_h} (1 - f_i) + \frac{P(1 - P)}{n_h^2(2P - 1)^2} \right] + \sum_h \left( 1 - \frac{N_h}{N} \right) \left[ \frac{\pi_h (1 - \pi_h)}{n_h} (1 - f_i) + \frac{P(1 - P)}{n_h^2(2P - 1)^2} \right] + \sum_h \left( \frac{N_h}{N} \right)^2 \left[ \frac{\pi_h (1 - \pi_h)}{n_h} (n_h' - n_h) + \frac{P(1 - P)}{n_h(2P - 1)^2} \right], \]  \hspace{1cm} (5.6)

where \( f_i = n'/N \).

Result 5.2 has a form of the post-stratified variance in the first two components and the last term has a stratified variance.

### 5.2. Two-Way Classification

Similar to Sec. 5.1, we can extend the one-way classification to two-way calibration of two-phase RR estimator. By the unconditional variance formulation, we can obtain the first component of variance as follows:

\[ V_1E_2(\hat{Z}_{\text{cal}}) = V_1 \left( \frac{1}{N} \sum_{i,j} w_{1i} Z_k - \frac{1}{N} \sum_{i,j} Z_k \right) = \sum_{i,j} \left( \frac{N_{ij}}{N} \right)^2 \left( \frac{1}{n_{ij}} - \frac{1}{N_{ij}} \right) V_{1ij}, \]

where \( V_{1ij} = Z_{1ij}(1 - Z_{1ij}) \) with \( Z_{1ij} = \hat{N}_{1ij}^{-1} \sum_{k=1}^{n_{ij}} d_{1i} Z_k \).

And the second term of variance is

\[ E_1V_2(\hat{Z}_{\text{cal}}) = E_1V_2 \left( \frac{1}{N} \sum_{i,j} d_{i,j} \frac{N_{ij}}{N_{2ij}} Z_k - \frac{1}{N} \sum_{i,j} d_{i,j} \frac{N_{ij}}{N_{1ij}} Z_k \right) \]

\[ = \sum_{i,j} \left( \frac{N_{ij}}{N} \right)^2 \left( \frac{1}{n_{ij}} - \frac{1}{N_{ij}} \right) V_{2ij}, \]

where \( V_{2ij} = Z_{2ij}(1 - Z_{2ij}) \) with \( Z_{2ij} = \hat{N}_{2ij}^{-1} \sum_{k=1}^{n_{ij}} d_{1i} Z_k \).
Thus, the variance components of the leading term of the calibration RR estimator for two-phase sampling is

\[
\text{AV}(\hat{Z}_{\text{cal}}) = \sum_{i,j} \left( \frac{N_{ij}}{N} \right)^2 \left( \frac{1}{n_{ij}} - \frac{1}{N_{ij}} \right) V_{iij} + \sum_{i,j} \left( \frac{N_{ij}}{N} \right)^2 \left( \frac{1}{n_{ij}} - \frac{1}{n'_{ij}} \right) V_{2ij}
\] (5.7)

**Result 5.3.** The variance of the two-way calibration RR estimator for two-phase sampling is

\[
\text{AV}(\hat{\pi}_{\text{cal}}) = \sum_{i,j} \left( \frac{N_{ij}}{N} \right)^2 \left[ \frac{\pi_{ij}(1 - \pi_{ij})}{n_{ij}} \left( \frac{N_{ij} - n'_{ij}}{N_{ij} - 1} \right) + \frac{P(1 - P)}{n'_{ij}(2P - 1)^2} \right]
\]

\[
+ \sum_{i,j} \left( \frac{N_{ij}}{N} \right)^2 \left[ \frac{\pi_{ij}(1 - \pi_{ij})}{n_{ij}} \left( \frac{n'_{ij} - n_{ij}}{n'_{ij} - 1} \right) + \frac{P(1 - P)}{n_{ij}(2P - 1)^2} \right].
\] (5.8)

Similar to Result 5.2, we have to compute the unconditional variance of the calibration variance, since the asymptotic variance is the conditional variance given \( \hat{N}_{ij} \). Thus, the unconditional variance is given by the following result.

**Result 5.4.** The unconditional asymptotic variance by post-stratification at first-phase is

\[
\text{V}(\hat{\pi}_{\text{cal}}) = \sum_{i,j} \left( \frac{N_{ij}}{N} \right) \left[ \frac{\pi_{ij}(1 - \pi_{ij})}{n'} (1 - f_i) + \frac{P(1 - P)}{n'(2P - 1)^2} \right]
\]

\[
+ \sum_{i,j} \left( 1 - \frac{N_{ij}}{N} \right) \left[ \frac{\pi_{ij}(1 - \pi_{ij})}{n'} (1 - f_i) + \frac{P(1 - P)}{n'(2P - 1)^2} \right]
\]

\[
+ \sum_{i,j} \left( \frac{N_{ij}}{N} \right) \left[ \frac{\pi_{ij}(1 - \pi_{ij})}{n_{ij}} \left( \frac{n'_{ij} - n_{ij}}{n_{ij} - 1} \right) + \frac{P(1 - P)}{n_{ij}(2P - 1)^2} \right].
\] (5.9)

where \( f_i = n'/N \).

In Result 5.4, we know that the variance of the two-phase calibration estimator is composed of the post-stratified variance due to estimated unknown population strata, cells information, and stratified variance of the second phase of the sample.

### 6. Efficiency Comparisons

To apply the two-phase RR technique to a real example, we selected a tax evasion as the sensitive characteristic in this article. So we obtain the first-phase sample with relatively large size to get auxiliary information, then the second-phase of the sample is selected from the first-phase of sample. The final sample units are provided with a randomization device with \( P \) for sensitive question cards and \( 1 - P \) for negative question cards:

**Question A:** Did you do tax evasion in last year?

**Question A’:** Didn’t you do tax evasion in last year?
By using a randomization device, the respondents will get one of the above two questions and should answer to the questions with a “Yes” or “No” response. After collecting this RR survey data, we can estimate the true population proportion about the tax evasion.

We compare the efficiency of the proposed calibration estimator for two-phase RR estimator with the ordinary two-phase RR estimator. The relative efficiency is

\[
RB(\hat{\pi}_{cal} | \hat{\pi}_{two}) = \frac{V(\hat{\pi}_{two})}{V(\hat{\pi}_{cal})} \times 100\%.
\] (6.1)

We use the variance of ordinary RR estimator given by (3.9) and the variance of the calibrated RR estimator given by (5.6) and (5.9).

### 6.1. One-Way Classification

We assume that the household income level of the respondent used in the first phase of the sample unit is available as the auxiliary information for one-way classification. In this case, we assume the three strata according to the level of household income “High”, “Medium”, and “Low” and the size of strata are assumed as follows.

Assuming the population parameters, the truthful population proportion increases \( \pi = 0.1 \) to \( 0.4 \) by 0.05 and the selection probability of RR questions increases \( P = 0.6 \) to \( 0.9 \) by 0.1. For the variance of calibration RR estimator, we also assume that the population proportions of strata \( \pi_h \)'s are the weighted proportions of population proportions. Based on those assumptions and with Tables 1–3, we can make Table 4 which shows that the RE's are greater than 100%. This means that our proposed estimator is better than the ordinary RR estimator. Also, if we assume other selection probability of RR questions increases \( P = 0.6 \) to \( 0.9 \) by 0.1, then the RE's have symmetric patterns; this fact follows the properties of the ordinary RR estimator.

### Table 1
Population distribution

<table>
<thead>
<tr>
<th>Household-income level</th>
<th>( N_1 = 250 )</th>
<th>( N_2 = 250 )</th>
<th>( N_3 = 500 )</th>
<th>( N = 1,000 )</th>
</tr>
</thead>
</table>

### Table 2
First-phase sample distribution

<table>
<thead>
<tr>
<th>Household-income level</th>
<th>( n'_1 = 100 )</th>
<th>( n'_2 = 100 )</th>
<th>( n'_3 = 150 )</th>
<th>( n' = 350 )</th>
</tr>
</thead>
</table>
Table 3
Second-phase sample distribution

<table>
<thead>
<tr>
<th>Household-income level</th>
<th>High</th>
<th>Medium</th>
<th>Low</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>(n_1 = 30)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(n_2 = 30)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(n_3 = 40)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(n = 100)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4
Relative efficiencies between \(\hat{\pi}_{\text{two}}\) and \(\hat{\pi}_{\text{cal}}\)

<table>
<thead>
<tr>
<th>(P)</th>
<th>(\pi)</th>
<th>(\pi_1)</th>
<th>(\pi_2)</th>
<th>(\pi_3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.6</td>
<td>0.10</td>
<td>0.025</td>
<td>0.025</td>
<td>0.05</td>
</tr>
<tr>
<td></td>
<td>0.15</td>
<td>0.0375</td>
<td>0.0375</td>
<td>0.075</td>
</tr>
<tr>
<td></td>
<td>0.20</td>
<td>0.050</td>
<td>0.050</td>
<td>0.10</td>
</tr>
<tr>
<td></td>
<td>0.25</td>
<td>0.0625</td>
<td>0.0625</td>
<td>0.125</td>
</tr>
<tr>
<td></td>
<td>0.30</td>
<td>0.075</td>
<td>0.075</td>
<td>0.15</td>
</tr>
<tr>
<td></td>
<td>0.35</td>
<td>0.0875</td>
<td>0.0875</td>
<td>0.175</td>
</tr>
<tr>
<td></td>
<td>0.40</td>
<td>0.10</td>
<td>0.10</td>
<td>0.20</td>
</tr>
<tr>
<td>0.7</td>
<td>134.593</td>
<td>135.884</td>
<td>138.968</td>
<td>147.797</td>
</tr>
<tr>
<td>0.8</td>
<td>134.755</td>
<td>136.601</td>
<td>140.919</td>
<td>152.620</td>
</tr>
<tr>
<td>0.9</td>
<td>134.903</td>
<td>137.247</td>
<td>142.623</td>
<td>156.490</td>
</tr>
<tr>
<td></td>
<td>135.037</td>
<td>137.822</td>
<td>144.098</td>
<td>159.584</td>
</tr>
<tr>
<td></td>
<td>135.156</td>
<td>138.331</td>
<td>145.362</td>
<td>162.035</td>
</tr>
<tr>
<td></td>
<td>135.262</td>
<td>138.773</td>
<td>146.430</td>
<td>163.943</td>
</tr>
<tr>
<td></td>
<td>135.353</td>
<td>139.152</td>
<td>147.315</td>
<td>165.384</td>
</tr>
</tbody>
</table>

6.2. Two-Way Classification

As similar to one-way classification, we make the two-way tables with the assumption of auxiliary information. We denote that H-tax means the household tax, and type is the type of household earning in the following tables.

To describe the structure of population and the first-phase and the second-phase of a sample for two-way classification, see Tables 5–7, respectively. Of course, the population cell counts are unknown, and these values can be known from the auxiliary information in the first-phase of sample.

We set the truthful population proportion \(\pi = 0.1\) to 0.4 by 0.1 increments and the selection probabilities of RR questions \(P = 0.6\) to 0.9 by 0.1 increments. With the assumptions of the population, the first- and second-phase of the sample, we can get the RE between the calibration and ordinary two-phase RR estimators. From Table 8, we can conclude that our proposed estimator is more efficient than the ordinary two-phase RR estimator.

Table 5
Population distribution

<table>
<thead>
<tr>
<th>H-tax \ Type</th>
<th>Self-employ</th>
<th>Salary</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>(N_{11} = 100)</td>
<td>(N_{12} = 200)</td>
<td>(N_{1+} = 300)</td>
</tr>
<tr>
<td>Low</td>
<td>(N_{21} = 300)</td>
<td>(N_{22} = 400)</td>
<td>(N_{2+} = 700)</td>
</tr>
<tr>
<td>Total</td>
<td>(N_{+1} = 400)</td>
<td>(N_{+2} = 600)</td>
<td>(N = 1,000)</td>
</tr>
</tbody>
</table>
Table 6
First-phase sample distribution

<table>
<thead>
<tr>
<th>H-tax \ Type</th>
<th>Self-employ</th>
<th>Salary</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>$n'_{11} = 80$</td>
<td>$n'_{12} = 160$</td>
<td>$n'_{1+} = 240$</td>
</tr>
<tr>
<td>Low</td>
<td>$n'_{21} = 240$</td>
<td>$n'_{22} = 320$</td>
<td>$n'_{2+} = 560$</td>
</tr>
<tr>
<td>Total</td>
<td>$n'_{+1} = 320$</td>
<td>$n'_{+2} = 480$</td>
<td>$n' = 800$</td>
</tr>
</tbody>
</table>

Table 7
Second-phase sample distribution

<table>
<thead>
<tr>
<th>H-tax \ Type</th>
<th>Self-employ</th>
<th>Salary</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>$n_{11} = 40$</td>
<td>$n_{12} = 60$</td>
<td>$n_{1+} = 100$</td>
</tr>
<tr>
<td>Low</td>
<td>$n_{21} = 100$</td>
<td>$n_{22} = 120$</td>
<td>$n_{2+} = 220$</td>
</tr>
<tr>
<td>Total</td>
<td>$n_{+1} = 140$</td>
<td>$n_{+2} = 180$</td>
<td>$n = 320$</td>
</tr>
</tbody>
</table>

Table 8
Relative efficiencies between $\hat{\pi}_{two}$ and $\hat{\pi}_{cal}$

<table>
<thead>
<tr>
<th>$\pi$</th>
<th>$\pi_{11}$</th>
<th>$\pi_{12}$</th>
<th>$\pi_{21}$</th>
<th>$\pi_{22}$</th>
<th>0.6</th>
<th>0.7</th>
<th>0.8</th>
<th>0.9</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.10</td>
<td>0.01</td>
<td>0.02</td>
<td>0.03</td>
<td>0.04</td>
<td>130.848</td>
<td>132.856</td>
<td>137.744</td>
<td>152.514</td>
</tr>
<tr>
<td>0.15</td>
<td>0.015</td>
<td>0.03</td>
<td>0.045</td>
<td>0.06</td>
<td>131.103</td>
<td>134.000</td>
<td>140.966</td>
<td>161.303</td>
</tr>
<tr>
<td>0.20</td>
<td>0.02</td>
<td>0.04</td>
<td>0.06</td>
<td>0.08</td>
<td>131.339</td>
<td>135.050</td>
<td>143.873</td>
<td>168.821</td>
</tr>
<tr>
<td>0.25</td>
<td>0.025</td>
<td>0.05</td>
<td>0.075</td>
<td>0.1</td>
<td>131.556</td>
<td>136.010</td>
<td>146.480</td>
<td>175.226</td>
</tr>
<tr>
<td>0.30</td>
<td>0.03</td>
<td>0.06</td>
<td>0.09</td>
<td>0.12</td>
<td>131.755</td>
<td>136.879</td>
<td>148.801</td>
<td>180.650</td>
</tr>
<tr>
<td>0.35</td>
<td>0.035</td>
<td>0.07</td>
<td>0.105</td>
<td>0.14</td>
<td>131.934</td>
<td>137.661</td>
<td>150.850</td>
<td>185.201</td>
</tr>
<tr>
<td>0.40</td>
<td>0.04</td>
<td>0.08</td>
<td>0.12</td>
<td>0.16</td>
<td>132.095</td>
<td>138.355</td>
<td>152.637</td>
<td>188.969</td>
</tr>
</tbody>
</table>

All simulations for this section were performed on a personal computer using programs written in SAS IML (Version 9.2). The software is available through the corresponding author upon request.

7. Concluding Remarks

Different from a general survey study, the RR survey has a limitation to the use of auxiliary information of survey respondents because of privacy protection. If auxiliary information for a survey respondent is available in the estimation procedure, then we can improve the survey estimator by incorporating this information. So we proposed the calibration procedure to improve the two-phase RR estimator which is an extension of the Warner RR model. We showed that the proposed two-phase calibration procedure is more useful than the two-phase RR estimator. The proposed estimator can be applied to real situations such as the examples of Secs. 6.1 and 6.2. Therefore, we can conclude that the calibration procedure for two-phase RR estimator has the advantage compared to...
the two-phase RR estimator with respect to the usage of auxiliary information and improvement of estimators.

Acknowledgments

The authors are thankful to the Editor, Associate Editor, and two referees for the valuable comments on the original version of this manuscript which led to substantial improvement.

References


