Calibration for Randomized Response Estimators
Chang-Kyoon Son\textsuperscript{a}; Jong-Min Kim\textsuperscript{b}; Ki-Hak Hong\textsuperscript{c}; Gi-Sung Lee\textsuperscript{d}
\textsuperscript{a} Korea Institute for Health and Social Affairs, Seoul, Republic of Korea \textsuperscript{b} Statistics, Division of Science and Mathematics, University of Minnesota, Morris, Minnesota, USA \textsuperscript{c} Department of Computer Science, Dongshin University, Jeonnam, Republic of Korea \textsuperscript{d} Department of E-Information Engineering, Woosuk University, Jeonbuk, Republic of Korea

Online publication date: 26 August 2010

To cite this Article Son, Chang-Kyoon , Kim, Jong-Min , Hong, Ki-Hak and Lee, Gi-Sung(2010) 'Calibration for Randomized Response Estimators', Communications in Statistics - Theory and Methods, 39: 17, 3163 — 3177

To link to this Article DOI: 10.1080/03610920903205206
URL: http://dx.doi.org/10.1080/03610920903205206
Calibration for Randomized Response Estimators

CHANG-KYON SON¹, JONG-MIN KIM², KI-HAK HONG³, AND GI-SUNG LEE⁴

¹Korea Institute for Health and Social Affairs, Seoul, Republic of Korea
²Statistics, Division of Science and Mathematics, University of Minnesota, Morris, Minnesota, USA
³Department of Computer Science, Dongshin University, Jeonnam, Republic of Korea
⁴Department of E-Information Engineering, Woosuk University, Jeonbuk, Republic of Korea

In the present article, we consider the calibration procedure for the Warner’s and Mangat–Singh’s (:M–S) randomized response survey estimators using auxiliary information associated with the variable of interest. In the calibration procedure, we can use auxiliary information such as age, gender, and income for the respondents of RR questions from an external source, and then the classical RR estimators can be improved with respect to the problems of noncoverage or nonresponse. From the efficiency comparison study, we show that the calibration estimators are more efficient than those of Warner’s and Mangat-Singh’s when the known population cell and marginal counts of auxiliary information are used for the calibration procedure.

Keywords Calibration estimator; Post-stratification; Randomized response model; Relative efficiency.

Mathematics Subject Classification 62D05.

1. Introduction

In a statistical survey, we often use auxiliary information to improve the precision of the estimator for the population total, mean, and ratio. For example, the generalized regression estimator for a finite population total requires the known population total of the auxiliary variable, where an available auxiliary variable has a strong correlation with the variable of interest.

The objectives of weighting adjustment including calibration, raking, and post-stratification are to reduce biases due to non coverage and non response in

Received October 7, 2008; Accepted July 23, 2009
Address correspondence to Jong-Min Kim, Statistics, Division of Science and Mathematics, University of Minnesota, Morris, MN 56267, USA; E-mail: jongmink@morris.umn.edu
the survey procedure. The calibration estimator uses the calibrated weighting, which is as close as possible, according to a given distance measure, to the original sampling weight while also respecting a set of constraints, calibration equations. Deville and Särndal (1993) and Deville et al. (1993) suggested the calibration estimator according to the distance functions. Dupont (1995) suggested an alternative calibration procedure using the two-phase sampling scheme, and Singh and Mohl (1996) considered the range restriction of calibration weight for handling the negative and extreme weights problems in calibration procedure. Singh et al. (1998) suggested modification of the calibration procedure according to the levels of auxiliary information.

Meanwhile, in the case of a survey for sensitive characteristics such as drug use and tax evasion, it is difficult to obtain a 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. 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. Mangat and Singh (1990) suggested the new RR survey model to improve Warner’s model.

In association with the calibration for RR survey, Tracy and Singh (1999) proposed the calibration procedure of the scrambled (quantitative) RR survey for the continuous sensitive characteristics, and they suggested the high-order calibration method using the population variance of the auxiliary variable. In contrast to their procedure, our procedure is specified by the calibration for the qualitative (binary) variable, and we consider the calibration procedure corresponding to the usage of either population cell counts or marginal counts. Also, we suggest the calibration method to improve the estimators of the typical Warner’s and M–S’s by using auxiliary information. However, Tracy and Singh (1999) did not consider these cases in estimation procedure.

For the RR survey, unlike typical surveys, research has access to limited information due to the privacy protection of respondents. Nevertheless, if we only have access to auxiliary information such as age, gender, and income group from external source, then the classical RR estimator can be improved by using this information. In Sec. 2, we introduce the Warner’s and M–S’s RR models. Section 3 is devoted to the calibration procedure for RR models, and Sec. 4 to the conditional and unconditional properties of the calibrated RR estimators. Section 5 is devoted to the empirical study in order to compare the efficiencies between the calibrated RR estimators and the original RR ones, and Sec. 6 concludes our discussion.

2. The Randomized Response Models

2.1. Warner’s RR Model

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

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

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

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

where \( \hat{Z} = \sum_{k=1}^{n} Z_k/n \) is the proportion of “Yes” answer in a sample selected by SRSWR.

Warner showed that the estimator \( \hat{\pi}_w \) is unbiased and its own variance is

\[ V(\hat{\pi}_w) = \frac{\pi(1 - \pi)}{n} + \frac{P(1 - P)}{n(2P - 1)^2}. \]

Kim and Flueck (1978) modified the Warner model to the case of SRSWOR. They showed that \( \hat{\pi}_w \) is unbiased and the variance is

\[ V(\hat{\pi}_w) = \left( \frac{N - n}{N - 1} \right) \frac{\pi(1 - \pi)}{n} + \frac{P(1 - P)}{n(2P - 1)^2}. \]

### 2.2. M–S’s Two-Stage Related RR Model

The Mangat and Singh’s (1990) two-stage related RR technique improved the efficiency using two randomization devices (R1 and R2) as follows. At the first stage, randomization device R1, 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 “Go to R2” with respective probabilities \( T \) and \( 1 - T \) (\( 0 < T < 1 \)) and answers “Yes” or “No” to the question chosen. At the second stage, a randomization device R2 which is composed of two questions: “Do you belong to sensitive group \( A \)?” or “Do you belong to non-sensitive group \( A^c \)” with respective probabilities \( P \) and \( 1 - P \) (\( 0 < P < 1 \)) and answers “Yes” or “No” to the question chosen.

Assuming truthful reporting, the probability of obtaining a “Yes” response is

\[ Z = T\pi + (1 - T)(1 - P + (2P - 1)\pi). \]

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

\[
\hat{\pi}_{ms} = \frac{(1 - T)(P - 1)}{T + (1 - T)(2P - 1)} + \frac{\hat{Z}}{T + (1 - T)(2P - 1)} \\
= \frac{(1 - T)(P - 1)}{T + (1 - T)(2P - 1)} + \frac{1}{T + (1 - T)(2P - 1)} \sum_{k=1}^{n} Z_k/n,
\]

where \( \hat{Z} = \sum_{k=1}^{n} Z_k/n \) is the proportion of “Yes” answer in a sample selected by SRSWR.
M–S estimator $\hat{\pi}_{ms}$ is unbiased for $\pi$ and its own variance is

$$V(\hat{\pi}_{ms}) = \frac{\pi(1 - \pi)}{n} + \frac{(1 - T)(1 - P)[1 - (1 - T)(1 - P)]}{n(T + (1 - T)(2P - 1))^2}.$$  

If the individual respondents are selected by SRSWOR, then the variance is

$$V(\hat{\pi}_{ms}) = \left(\frac{N - n}{N - 1}\right)\frac{\pi(1 - \pi)}{n} + \frac{(1 - T)(1 - P)[1 - (1 - T)(1 - P)]}{n(T + (1 - T)(2P - 1))^2}.$$  

### 3. Calibration for RR Estimators

#### 3.1. Known Population Cell Counts

Consider a finite population $U = \{1, 2, \ldots, k, \ldots, N\}$, and a probability sample $s$ of size $n$ with a given sampling design such that inclusion probabilities $v_k = \Pr(k \in s)$ and $v_{kl} = \Pr(k \& l \in s)$ are strictly positive. As defined in previous section, let $Z_k$ be the binary indicator taking value 1, if a unit reports “Yes” to RR question and 0 otherwise, and the sample respondents are selected by SRSWOR. Then the population total reporting “Yes” to RR question is defined by $\tau_z = \sum_U Z_k$ and the counterpart of the sample is $\hat{\tau}_z = \sum_s d_k Z_k$, where $d_k = 1/v_k$ is the sampling design weight, with $k$th element is also associated an auxiliary vector value $x_k = (x_{k1}, x_{k2}, \ldots, x_{kj}, \ldots, x_{kJ})'$. Unlike common statistical surveys, the RR survey is limited in the use of auxiliary information for the privacy protection of respondents. However, the RR survey data has some socio-demographical auxiliary information such as sex, age, and so on at the population level.

We assume that the population x total $\tau_x = \sum_{k \in U} x_k$ is known. Let us start with the Horvitz–Thompson estimator $\hat{\tau}_z = \sum_s d_k Z_k$ to develop the calibration procedure.

The objective of the calibration procedure is to adjust the original design weight $d_k$ using the known auxiliary information, which is as close as possible to $d_k$, according to the population distribution. That is, for a known auxiliary variable $x_k$, we can find a new weight $w_k$ subject to constraint $\sum_k w_k x_k = \tau_x$, which minimizes the distance measure $\sum_s G(w_k, d_k) = \sum_s (w_k - d_k)^2/d_k$. The calibration procedure suggested by Deville and Särndal (1993) is optimized as following Lagrange function

$$\min \left(\sum_s d_k G(w_k, d_k) - \lambda \left(\sum_s w_k x_k - \tau_x\right)\right),$$

where the vector of Lagrange multipliers is $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_J)'$.

Differentiate the Lagrange function with respect to $w_k$, and set equal to 0, then the new weight is

$$w_k = d_k (1 + x_k' \lambda) = d_k F(x_k' \lambda), \quad (3.1)$$

where $F = (\partial G/\partial w)^{-1}$, $\lambda$ is determined from calibration equation as follows:

$$\sum_s d_k (1 + x_k' \lambda)x_k = \tau_x.$$
We limit the two-way contingency table, and the auxiliary information $\tau_x = \sum_{k \in U} x_k$ is given in the form of known cell counts in contingency table with two dimensions as follows:

$$\tau_x = \sum_{k \in U} x_k = (N_{11}, N_{12}, \ldots, N_{ij}, \ldots, N_{n}).$$

For the Warner's RR model, the sample proportion of answer “Yes”, $\hat{Z}$, can be rewritten as follows:

$$\hat{Z} = \frac{1}{N} \sum_{k=1}^n Z_k = \frac{1}{N} \sum_{k=1}^n d_k Z_k,$$

where the original sampling weight is $d_k = N/n$ for SRWSON.

From (3.1), the original sampling weight $d_k$ is replaced by the new weight $w_k = d_k N_{ij}/\hat{N}_{ij}$, and then the calibrated sample proportion $\hat{Z}$ is given by

$$\hat{Z}_{\text{post}} = \frac{1}{N} \sum_{k=1}^n w_k Z_k = \frac{1}{N} \sum_{i=1}^r \sum_{j=1}^c N_{ij} \hat{Z}_{ij},$$

(3.2)

where $\hat{Z}_{ij} = \sum_{k=1}^n d_k Z_k / \hat{N}_{ij}$ is the weighted proportion in the sample cell with $\hat{N}_{ij} = \sum_{j=1}^c d_k$.

**Result 3.1.** Using (3.2), if the respondents are selected by SRWSON, $\hat{N}_{ij} = d_k n_{ij} = (N/n) n_{ij}$, then the post-stratified Warner's RR estimator is

$$\hat{\pi}_{\text{wpost}} = \frac{P - 1}{2P - 1} + \left( \frac{1}{2P - 1} \right) \sum_{i=1}^r \sum_{j=1}^c N_{ij} \sum_{k=1}^n \frac{n_{ij}}{n_{ij}} Z_k,$$

where $n_{ij}$ is the sample cell counts for $i$th row and $j$th column.

**Result 3.2.** The post-stratified M–S estimator of $\hat{\pi}_{\text{ms}}$ is given by

$$\hat{\pi}_{\text{mspost}} = \frac{(1 - T)(P - 1)}{T + (1 - T)(2P - 1)} + \left( \frac{1}{T + (1 - T)(2P - 1)} \right) \sum_{i=1}^r \sum_{j=1}^c \frac{N_{ij}}{N} \sum_{k=1}^n \frac{n_{ij}}{n_{ij}} Z_k.$$

### 3.2. Known Population Marginal Counts

We used the knowledge of population cell counts of the auxiliary variable in the previous calibration procedure. But if we only know the population marginal counts from auxiliary information, we can use the knowledge of marginal counts in calibration procedure as follows:

$$\sum_{k \in U} x_k = (N_{11}, N_{12}, \ldots, N_{r+1}, N_{r+2}, \ldots, N_{m+1}),$$

(3.3)

where $N_{r+1} = \sum_{j=1}^m N_{ij}, N_{r+2} = \sum_{i=1}^r N_{ij}$.

In (3.3), we defined the auxiliary variable vector $x_k = (\delta_{1k}, \ldots, \delta_{rk}, \delta_{1k}, \ldots, \delta_{rk})'$ where $\delta_{1k} = 1$, if the respondent $k$ is in row $i$ and 0 otherwise,
\( \delta_{jk} = 1 \) if the respondent \( k \) is in column \( j \) and 0 otherwise. Let the Lagrange multiplier \( \lambda = (u_1, \ldots, u_r, v_1, \ldots, v_c) \), then we can express \( x_k^j \lambda = u_i + v_j \), therefore \( F(x_k^j \lambda) = F(u_i + v_j) \) can be written, where \( F = (\partial G/\partial u)^{-1} \) is defined as in Deville et al. (1993). The calibration equations are

\[
\sum_{j=1}^c \hat{N}_{ij} F(u_i + v_j) = N_{i+}, \quad \text{for } i = 1, 2, \ldots, r,
\]

\[
\sum_{i=1}^r \hat{N}_{ij} F(u_i + v_j) = N_{+j}, \quad \text{for } j = 1, 2, \ldots, c,
\]

where \( u_i \) and \( v_j \) are determined by iterative computation.

We can obtain the calibrated cell counts estimates \( \hat{N}_{ij}^w = \hat{N}_{ij} F(u_i + v_j) \), and then the calibrated weight is \( u_k = d_k \hat{N}_{ij}^w / \hat{N}_{ij} \). Finally, the calibration estimator for \( \tilde{Z} \) is given by

\[
\tilde{Z}_{\text{cal}} = \frac{1}{N} \sum_{k=1}^n w_k Z_k = \frac{1}{N} \sum_{i=1}^r \sum_{j=1}^c \hat{N}_{ij}^w \tilde{Z}_{ij}, \tag{3.4}
\]

**Result 3.3.** From (3.4), if the respondents are selected by SRSWOR, then the calibrated Warner’s RR estimator is

\[
\hat{\pi}_{\text{wcal}} = \frac{P-1}{2P-1} + \left( \frac{1}{2P-1} \right) \sum_{i=1}^r \sum_{j=1}^c \hat{N}_{ij}^w \sum_{k=1}^n Z_k.
\]

**Result 3.4.** Also, the calibrated M–S’s RR estimator is

\[
\hat{\pi}_{\text{mscal}} = \frac{(1-T)(P-1)}{T + (1-T)(2P-1)} + \left( \frac{1}{T + (1-T)(2P-1)} \right) \sum_{i=1}^r \sum_{j=1}^c \hat{N}_{ij}^w \sum_{k=1}^n Z_k.
\]

### 4. Conditional and Unconditional Variances of Calibrated RR Estimators

We will investigate the conditional and unconditional properties of the calibrated Warner’s and M–S’s RR estimators. We can derive the unconditional variance for each estimator.

#### 4.1. Conditional Variances

Here, we try to consider a row effect and a column effect in a two-way contingency table for RR survey data. Let the two cross effect factors explain the population proportion reporting “Yes” for RR questions, then we parameterize the finite population using the ANOVA representing that for respondent \( k \) in population cell \( U_{ij} \), \( Z_k = \alpha_i + \beta_j + E_k \), where \( Z_k \) is the indicator taking value 1, if a respondent reports “Yes” to RR question and 0 otherwise. If \( \alpha_i \) is a row effect, \( \beta_j \) a column effect, and \( E_k \) is an error term, then \( \alpha_i \) and \( \beta_j \) are fixed unknown values defined by
normal equations
\[
\sum_{j=1}^{c} N_{ij}(x_i + \beta_j) = N_{ij}\bar{Z}_{ij}, \quad \text{for } i = 1, 2, \ldots, r,
\]
\[
\sum_{i=1}^{r} N_{ij}(x_i + \beta_j) = N_{ij}\bar{Z}_{ij}, \quad \text{for } j = 1, 2, \ldots, c.
\]

Let’s decompose the \(k\)th error term \(E_k = L_{ij} + R_k\), where \(L_{ij} = \bar{Z}_{ij} - (x_i + \beta_j)\) is an interaction term, and \(R_k = Z_k - \bar{Z}_{ij}\) is the deviation from \(\bar{Z}_{ij} = \sum_{ij} Z_{ij}/N_{ij}\), where \(\bar{Z}_{ij}\) represents the population cell proportions of “Yes” to the RR question in cell \(ij\).

The restrictions for the interaction term are
\[
\sum_{i=1}^{r} N_{ij}L_{ij} = \sum_{j=1}^{c} N_{ij}L_{ij} = 0.
\]

Now, the indicator value of \(Z_k\) of the RR question can be written as
\[
Z_k = \sum_{i=1}^{r} \sum_{j=1}^{c} \hat{N}_{ij}^w(x_i + \beta_j + L_{ij} + \tilde{R}_{ij}),
\]
where \(\tilde{R}_{ij} = \sum_{k=1}^{n_{ij}} d_k R_k / \hat{N}_{ij}\) is the deviation proportion of sample cells.

Also, we can express the calibration equation of \(Z\) as follows:
\[
\frac{1}{N} \sum_{i=1}^{r} \sum_{j=1}^{c} N_{ij}(x_i + \beta_j) = \frac{1}{N} \sum_{i=1}^{r} \sum_{j=1}^{c} \hat{N}_{ij}^w(x_i + \beta_j).
\] (4.1)

Let the population proportion answering “Yes” for RR question, \(Z = 1/N \sum_{ij} Z_{ij}\) be denoted by the left-hand side of Eq. (4.1), then we can express the error of \(\hat{Z}_{\text{cal}}\) as
\[
\hat{Z}_{\text{cal}} - Z = \frac{1}{N} \sum_{i=1}^{r} \sum_{j=1}^{c} (\hat{N}_{ij}^w - N_{ij})L_{ij} + \frac{1}{N} \sum_{i=1}^{r} \sum_{j=1}^{c} \hat{N}_{ij}^w \tilde{R}_{ij}.
\] (4.2)

Similar to (4.2), the error of the post-stratified estimator \(\hat{Z}_{\text{post}}\) is
\[
\hat{Z}_{\text{post}} - Z = \frac{1}{N} \sum_{i=1}^{r} \sum_{j=1}^{c} N_{ij} \tilde{R}_{ij}.
\]

In order to compare the conditional properties of each estimator, we consider the vector of cell count estimates as follows:
\[
\hat{N} = (\hat{N}_{11}, \hat{N}_{12}, \ldots, \hat{N}_{ij}, \ldots, \hat{N}_{nc})',
\]
where \(\hat{N}_{ij} = \sum_{k} d_k\).

Let the conditional bias and variance be denoted by \(B_c = B(\bullet | \hat{N})\), and \(V_c = V(\bullet | \hat{N})\), respectively.
The conditional biases of the estimators of population proportions, \( \hat{p}_{\text{post}} \) and \( \hat{p}_{\text{cal}} \), can be expressed by

\[
B_c(\hat{p}_{\text{post}}) = \left( \frac{1}{2P-1} \right) \frac{1}{N} \sum_{i=1}^{r} \sum_{j=1}^{c} N_{ij} E_c(\tilde{R}_{ij}),
\]

\[
B_c(\hat{p}_{\text{cal}}) = \left( \frac{1}{2P-1} \right) \left[ \frac{1}{N} \sum_{i=1}^{r} \sum_{j=1}^{c} \left( \tilde{N}_{ij}^w - N_{ij} \right) L_{ij} + \frac{1}{N} \sum_{i=1}^{r} \sum_{j=1}^{c} \tilde{N}_{ij}^w E_c(\tilde{R}_{ij}) \right],
\]

respectively.

From (4.3) and (4.4), the conditional expectation is \( E_c(\tilde{R}_{ij}) = 0 \) or nearly 0 for all \( i, j \), because the sampling design is SRSWOR and then the inclusion probability \( v_k \) is a constant. The conditional bias of post-stratified estimator \( B_c(\hat{Z}_{\text{post}}) \approx 0 \), whereas \( B_c(\hat{Z}_{\text{cal}}) = N^{-1} \sum_{i=1}^{r} \sum_{j=1}^{c} \left( \tilde{N}_{ij}^w - N_{ij} \right) L_{ij} \). For a large sample, \( \tilde{N}_{ij}^w \) is close to \( N_{ij} \), and then the conditional bias of \( \hat{Z}_{\text{cal}} \) is asymptotically equal to that of \( \hat{Z}_{\text{post}} \). The conditional variance of the calibrated Warner’s RR estimator can be rewritten by

\[
V_c(\hat{p}_{\text{post}}) = \sum_{i=1}^{r} \sum_{j=1}^{c} \left( \frac{N_{ij}}{N} \right)^2 \left( \frac{\pi_{ij}(1 - \pi_{ij})}{n_{ij}} \left( \frac{N_{ij} - n_{ij}}{N_{ij} - 1} \right) + \frac{P(1 - P)}{n_{ij}(2P - 1)^2} \right).
\]

The conditional variance of \( \hat{p}_{\text{cal}} \) is

\[
V_c(\hat{p}_{\text{cal}}) = \sum_{i=1}^{r} \sum_{j=1}^{c} \left( \frac{\tilde{N}_{ij}^w}{N} \right)^2 \left( \frac{\pi_{ij}(1 - \pi_{ij})}{n_{ij}} \left( \frac{N_{ij} - n_{ij}}{N_{ij} - 1} \right) + \frac{P(1 - P)}{n_{ij}(2P - 1)^2} \right).
\]

Meanwhile, the conditional biases of post-stratified and calibrated M–S RR estimators are

\[
B_c(\hat{\pi}_{\text{post}}) = \left( \frac{1}{T + (1 - T)(2P - 1)} \right) \frac{1}{N} \sum_{i=1}^{r} \sum_{j=1}^{c} N_{ij} E_c(\tilde{R}_{ij}),
\]

\[
B_c(\hat{\pi}_{\text{mscal}}) = \left( \frac{1}{T + (1 - T)(2P - 1)} \right) \left[ \frac{1}{N} \sum_{i=1}^{r} \sum_{j=1}^{c} \left( \tilde{N}_{ij}^w - N_{ij} \right) L_{ij} + \frac{1}{N} \sum_{i=1}^{r} \sum_{j=1}^{c} \tilde{N}_{ij}^w E_c(\tilde{R}_{ij}) \right],
\]

respectively.

The conditional variances of the post-stratified and calibrated M–S RR estimators are

\[
V_c(\hat{\pi}_{\text{post}}) = \sum_{i=1}^{r} \sum_{j=1}^{c} \left( \frac{N_{ij}}{N} \right)^2 \left( \frac{\pi_{ij}(1 - \pi_{ij})}{n_{ij}} \left( \frac{N_{ij} - n_{ij}}{N_{ij} - 1} \right) \right.
+ \left. \frac{(1 - T)(1 - P)(1 - (1 - T)(1 - P))}{n_{ij}(T + (1 - T)(2P - 1))^2} \right).
\]

\[
V_c(\hat{\pi}_{\text{mscal}}) = \sum_{i=1}^{r} \sum_{j=1}^{c} \left( \frac{\tilde{N}_{ij}^w}{N} \right)^2 \left( \frac{\pi_{ij}(1 - \pi_{ij})}{n_{ij}} \left( \frac{N_{ij} - n_{ij}}{N_{ij} - 1} \right) \right.
+ \left. \frac{(1 - T)(1 - P)(1 - (1 - T)(1 - P))}{n_{ij}(T + (1 - T)(2P - 1))^2} \right).
\]
If the interaction terms \( L_{ij} \) are negligible, then the conditional variances of \( \hat{N}_{wcal} \) and \( \hat{N}_{mscal} \) are equal to those of \( \hat{N}_{wpost} \) and \( \hat{N}_{mspost} \), replacing \( \hat{N}_{ij}^w \) by \( N_{ij} \). Ordinarily, it is reasonable to assume that the conditional variances of the calibration estimators are larger than the conditional variances of the post-stratified estimators. Also, we note the conditional biases of the post-stratified estimators are unaffected by interaction, whereas those of the calibration estimators depend on interaction.

4.2. Unconditional Variances

From the unconditional variances \( V = E(V_c) + V(B_c) \), we can derive the unconditional variances of Warner’s RR estimators \( \hat{N}_{wpost} \) and \( \hat{N}_{wcal} \) and those of the M–S’s RR estimators \( \hat{N}_{mspost} \) and \( \hat{N}_{mscal} \).

**Theorem 4.1.** The unconditional variance of the post-stratified Warner’s estimator can be expressed by

\[
V(\hat{N}_{wpost}) = \sum_{i=1}^{r} \sum_{j=1}^{c} \left( \frac{N_{ij}}{N} \right) \left( \frac{\pi_{ij}(1 - \pi_{ij})}{n} (1 - f) + \frac{P(1 - P)}{n(2P - 1)^2} \right) + \frac{1}{n} \sum_{i=1}^{r} \sum_{j=1}^{c} \left( 1 - \frac{N_{ij}}{N} \right) \left( \frac{\pi_{ij}(1 - \pi_{ij})}{n} (1 - f) + \frac{P(1 - P)}{n(2P - 1)^2} \right). \tag{4.11}
\]

**Proof.** From Cochran (1977, 5A.9), the size of sample cell \( n_{ij} \) is the random variable with \( E(n_{ij}) = n(N_{ij}/N) \), \( V(n_{ij}) = nN_{ij}/N(1 - N_{ij}/N) \) for the post-stratification. The \( n_{ij} \) can be expressed by

\[
n_{ij} = n \left( \frac{N_{ij}}{N} \right) \left( 1 - \frac{n(N_{ij}/N) - n_{ij}}{n(N_{ij}/N)} \right). \tag{4.12}
\]

Thus, the \( 1/n_{ij} \) can be written

\[
\frac{1}{n_{ij}} = \frac{1}{n(N_{ij}/N)} \left( 1 - \frac{n(N_{ij}/N) - n_{ij}}{n(N_{ij}/N)} + \frac{(nN_{ij}/N - n_{ij})^2}{(nN_{ij}/N)^2} - \ldots \right).
\]

Then the expectation of \( 1/n_{ij} \) is

\[
E \left( \frac{1}{n_{ij}} \right) \approx \frac{1}{n(N_{ij}/N)} + \frac{n(N_{ij}/N)(1 - N_{ij}/N)}{(nN_{ij}/N)^2} = \frac{1}{n(N_{ij}/N)} + \frac{(1 - N_{ij}/N)}{(nN_{ij}/N)^2}. \tag{4.13}
\]

Substituting \( E(n_{ij}) = n(N_{ij}/N) \) and (4.13) into the expectation of (4.5) and after some algebra, we obtain (4.12).

**Remark 4.1.** The unconditional variance represents the stratified Warner’s RR variance suggested by Hong et al. (1994) plus the increment due to the post-stratification, which agrees with the typical result given in Cochran (1977).
3172 Son et al.

**Theorem 4.2.** From (4.4) and (4.6), the unconditional variance is

\[
V(\hat{\pi}_{\text{scal}}) = \sum_{i=1}^{r} \sum_{j=1}^{c} \left( \frac{\hat{N}_{ij}}{N} \right) \left[ \pi_{ij}(1 - \pi_{ij}) \left( 1 - \frac{P(1 - P)}{n} \right) \right] + \frac{1}{n} \sum_{i=1}^{r} \sum_{j=1}^{c} \left( 1 - \frac{\hat{N}_{ij}}{N} \right) \left[ \pi_{ij}(1 - \pi_{ij}) \left( 1 - \frac{P(1 - P)}{n} \right) \right] + \left( \frac{1}{2P - 1} \right)^2 \left( 1 - \frac{f}{n} \right) \sum_{i=1}^{r} \sum_{j=1}^{c} \left( \frac{\hat{N}_{ij}}{N} \right)^2 L_{ij}^2
\]

(4.14)

**Proof.** First of all, we recall the proof of Theorem 4.1 from the conditional variance (4.6), then we can obtain the first and second terms of the right-hand side in (4.14). Then for the third term of (4.14), the variance of conditional bias leads to

\[
V(B_c) = \left( \frac{1}{2P - 1} \right)^2 \left( 1 - \frac{f}{n} \right) \sum_{i=1}^{r} \sum_{j=1}^{c} \left( \frac{\hat{N}_{ij}}{N} \right)^2 L_{ij}^2.
\]

Thus, we can obtain the unconditional variance of the calibration Warner’s estimator.

**Theorem 4.3.** The unconditional variance of the post-stratified M–S’s RR estimator, \( V(\hat{\pi}_{\text{mcal}}) \), can be expressed by

\[
V(\hat{\pi}_{\text{mcal}}) = \sum_{i=1}^{r} \sum_{j=1}^{c} \left( \frac{N_{ij}}{N} \right) \left[ \pi_{ij}(1 - \pi_{ij}) \left( 1 - \frac{(1 - T)(1 - P)(1 - (1 - T)(1 - P)}{n(1 - T)(2P - 1)} \right) \right] + \frac{1}{n} \sum_{i=1}^{r} \sum_{j=1}^{c} \left( 1 - \frac{N_{ij}}{N} \right) \left[ \pi_{ij}(1 - \pi_{ij}) \left( 1 - \frac{(1 - T)(1 - P)(1 - (1 - T)(1 - P)}{n(1 - T)(2P - 1)} \right) \right] + \left( \frac{1 - T}{n(1 - T)(2P - 1)} \right)^2 \frac{n}{T + (1 - T)(2P - 1)} \left( \frac{1 - f}{n} \right) \sum_{i=1}^{r} \sum_{j=1}^{c} \left( \frac{\hat{N}_{ij}}{N} \right)^2 L_{ij}^2.
\]

(4.15)

**Proof.** Refer to the proof of Theorem 4.1.

**Theorem 4.4.** From (4.13), the unconditional variance of the calibrated M–S’s RR estimator is

\[
V(\hat{\pi}_{\text{scal}}) = \sum_{i=1}^{r} \sum_{j=1}^{c} \left( \frac{\hat{N}_{ij}}{N} \right) \left[ \pi_{ij}(1 - \pi_{ij}) \left( 1 - \frac{(1 - T)(1 - P)(1 - (1 - T)(1 - P)}{n(1 - T)(2P - 1)} \right) \right] + \frac{1}{n} \sum_{i=1}^{r} \sum_{j=1}^{c} \left( 1 - \frac{\hat{N}_{ij}}{N} \right) \left[ \pi_{ij}(1 - \pi_{ij}) \left( 1 - \frac{(1 - T)(1 - P)(1 - (1 - T)(1 - P)}{n(1 - T)(2P - 1)} \right) \right] + \left( \frac{1 - T}{n(1 - T)(2P - 1)} \right)^2 \frac{n}{T + (1 - T)(2P - 1)} \left( \frac{1 - f}{n} \right) \sum_{i=1}^{r} \sum_{j=1}^{c} \left( \frac{\hat{N}_{ij}}{N} \right)^2 L_{ij}^2.
\]
Theorem 4.2. Refer to the proof of Theorem 4.2.

From the unconditional variances of the calibrated Warner’s and M–S’s RR estimators, (4.13) and (4.15), the first term of the unconditional variance equals the post-stratified variance replacing $\hat{N}_{ij}$ by $N_{ij}$. Therefore, if $E(\hat{N}_{ij}) \equiv N_{ij}$ for a large sample, then the last terms of the (4.13) and (4.15) are negligible, thus the unconditional variance of the calibrated RR estimator equals to that of the post-stratified RR estimator.

5. Empirical Comparison Study

To apply our proposed RR techniques to real data, we assume the sensitive characteristics, such as tax evasion, classify the sample data with rows and columns in a contingency table, according to auxiliary variables with $2 \times 2$ dimensions. As discussed in Deville et al. (1993), this dimension of the assumption of the population and sample contingency table can be extended to more than $2 \times 2$ dimensions. We assume the population and sample contingency tables describe the calibration procedure as follows. Table 1 shows the population distribution of the respondents, each cell count denoted by $N_{ij}$, which can be known from the socio-demographic information for respondents. Let $N_i^+$ and $N_j^+$ denote the row and column marginal counts, respectively. If the population cell counts $N_{ij}$ are known, then we can use the post-stratified estimator, and if these counts are unknown but the marginal counts $N_i^+$ and $N_j^+$ are known, then we can use the calibration estimator.

Table 2 describes the sample distribution of the respondent selected by SRSWOR with size of $n$, and each cell count $n_{ij}$ observed from the survey. Thus, $n_{11}$ includes the number of respondents for reporting "Yes" for the RR question. As a result, we will calibrate the proportions of respondents reporting "Yes" in sample cells according to the available information $N_{ij}$ or $N_i^+$ and $N_j^+$.

Table 1

<table>
<thead>
<tr>
<th>Income \ Gender</th>
<th>Male</th>
<th>Female</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>$N_{i1}$</td>
<td>$N_{i2}$</td>
<td>$N_{i+}$</td>
</tr>
<tr>
<td>Low</td>
<td>$N_{21}$</td>
<td>$N_{22}$</td>
<td>$N_{2+}$</td>
</tr>
<tr>
<td>Total</td>
<td>$N_{+1}$</td>
<td>$N_{+2}$</td>
<td>$N$</td>
</tr>
</tbody>
</table>
We did not select the sample repeatedly. To make the response set, we draw the response set with SRSWOR from the sample Table 2 according to a given sample proportion \( \hat{Z} \) of reporting “Yes” to a sensitive characteristic, where this value changed by the population proportion. We compute the unconditional variance of calibration and ordinary RR models changing the population proportion \( \pi \) and the selection probability \( p \). We compare the relative efficiencies (RE) between the unconditional variance of the calibration Warner’s and the ordinary Warner’s and the calibration M–S’s RR estimator and the ordinary M–S’s RR estimators.

\[
RE(\hat{\pi}_{\text{cal}}) = \frac{V(\hat{\pi}_{\text{RR}})}{V(\hat{\pi}_{\text{cal}})},
\]

where \( V(\cdot) \) represents the unconditional variance of the calibration estimator for the denominator and the suffixes “RR” and “cal” represent the ordinary RR estimators and the calibration estimators, respectively.

In the efficiency comparisons, we would expect that the RE of calibration RR estimators is greater than the RE of ordinary RR estimators.

### 5.1. Comparison of the Ordinary and the Calibrated Warner’s RR Estimator

We compare the efficiency between the ordinary Warner’s RR estimator and the calibrated Warner’s RR estimator. We would expect that the calibrated Warner’s RR estimators to be more efficient than the ordinary Warner’s.

Assuming the population parameters, the truthful population proportion is \( \pi = 0.1 \) to 0.4 by 0.1 increments and the selection probability of RR question is \( P = 0.6 \) to 0.9 by 0.1 increments, we compute the RE’s according to four different \( P \) and four different \( \pi \). To obtain the variance of the calibration estimator, the

<table>
<thead>
<tr>
<th>( \pi )</th>
<th>0.1</th>
<th>0.2</th>
<th>0.3</th>
<th>0.4</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.6</td>
<td>1.00595</td>
<td>1.00350</td>
<td>1.01218</td>
<td>1.00578</td>
</tr>
<tr>
<td>0.7</td>
<td>1.03727</td>
<td>1.03467</td>
<td>1.06432</td>
<td>1.05730</td>
</tr>
<tr>
<td>0.8</td>
<td>1.11143</td>
<td>1.10844</td>
<td>1.18207</td>
<td>1.17353</td>
</tr>
<tr>
<td>0.9</td>
<td>1.31895</td>
<td>1.31475</td>
<td>1.47415</td>
<td>1.46123</td>
</tr>
</tbody>
</table>

Table 3. Relative efficiencies for the calibration estimator and Warner’s estimator
population cell proportions $\pi_{ij}$ are assumed to be the weighted proportion of Warner’s for the post-stratification and the calibration. It is reasonable to compare the variances with the Warner’s model. The RE’s are shown in Table 3, where RE1 and RE2 are the relative efficiencies of the post-stratified and calibrated Warner’s variance with respect to ordinary Warner’s variance, respectively.

All of the RE’s except for the last column in Table 3 show that the proposed estimators are more efficient than Warner’s RR estimator, for given parameters. As discussed in Sec. 3, the variances of the calibration estimator are larger than those of post-stratification; this means that the interaction $L_{ij}$ may affect the estimator. Also, we can find that both RE1 and RE2 increase as the selection probability $P$ increases from 0.6 to 0.9 because the RR model approaches the direct question model. This result agrees with the typical RR estimator.

### 5.2. Comparison of the Ordinary and Calibrated M–S’s RR Estimator

Similar to Sec. 5.1, we compare the efficiency between the ordinary M–S’s RR estimator and the calibrated M–S’s. We would expect that the calibration M–S’s RR estimators are more efficient than the ordinary M–S’s. Let the population parameters be $\pi = 0.1$ to 0.4 and $T = 0.1$ to 0.7 by 0.2 increments, and $P = 0.6$ to 0.9 by 0.1 increments. We compute the RE’s according to four different $T$, $P$, and $\pi$.

To obtain the variance of the calibration estimator, assume the population cell proportions $\pi_{ij}$ to be the same as in Sec. 5.1. Now, the RE’s are shown in Table 4, where RE1 and RE2 are the relative efficiencies of the post-stratified and calibrated M–S’s variance to ordinary M–S’s variance, respectively.

#### Table 4

<table>
<thead>
<tr>
<th>$T$</th>
<th>$\pi$</th>
<th>RE1</th>
<th>RE2</th>
<th>RE1</th>
<th>RE2</th>
<th>RE1</th>
<th>RE2</th>
<th>RE1</th>
<th>RE2</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>0.6</td>
<td>1.01519</td>
<td>1.01270</td>
<td>1.02772</td>
<td>1.02113</td>
<td>1.03475</td>
<td>1.02126</td>
<td>1.03642</td>
<td>1.01344</td>
</tr>
<tr>
<td>0.7</td>
<td>1.05328</td>
<td>1.05060</td>
<td>1.09041</td>
<td>1.08306</td>
<td>1.11003</td>
<td>1.09470</td>
<td>1.11342</td>
<td>1.08726</td>
<td></td>
</tr>
<tr>
<td>0.8</td>
<td>1.13628</td>
<td>1.13316</td>
<td>1.21982</td>
<td>1.21077</td>
<td>1.25833</td>
<td>1.23902</td>
<td>1.25911</td>
<td>1.22640</td>
<td></td>
</tr>
<tr>
<td>0.9</td>
<td>1.35970</td>
<td>1.35523</td>
<td>1.52586</td>
<td>1.51209</td>
<td>1.57522</td>
<td>1.54590</td>
<td>1.54615</td>
<td>1.49847</td>
<td></td>
</tr>
<tr>
<td>0.3</td>
<td>0.6</td>
<td>1.04749</td>
<td>1.04484</td>
<td>1.08102</td>
<td>1.07379</td>
<td>1.09890</td>
<td>1.08385</td>
<td>1.10218</td>
<td>1.07649</td>
</tr>
<tr>
<td>0.7</td>
<td>1.10071</td>
<td>1.09778</td>
<td>1.16552</td>
<td>1.15721</td>
<td>1.19723</td>
<td>1.17961</td>
<td>1.20003</td>
<td>1.17006</td>
<td></td>
</tr>
<tr>
<td>0.8</td>
<td>1.20512</td>
<td>1.20161</td>
<td>1.32027</td>
<td>1.30978</td>
<td>1.36730</td>
<td>1.34478</td>
<td>1.36138</td>
<td>1.32365</td>
<td></td>
</tr>
<tr>
<td>0.9</td>
<td>1.46632</td>
<td>1.46112</td>
<td>1.65357</td>
<td>1.63752</td>
<td>1.69510</td>
<td>1.66147</td>
<td>1.64694</td>
<td>1.59337</td>
<td></td>
</tr>
<tr>
<td>0.5</td>
<td>0.6</td>
<td>1.11144</td>
<td>1.10845</td>
<td>1.18208</td>
<td>1.17354</td>
<td>1.21602</td>
<td>1.19789</td>
<td>1.21833</td>
<td>1.18753</td>
</tr>
<tr>
<td>0.7</td>
<td>1.18487</td>
<td>1.18147</td>
<td>1.29133</td>
<td>1.28126</td>
<td>1.33643</td>
<td>1.31484</td>
<td>1.33279</td>
<td>1.29651</td>
<td></td>
</tr>
<tr>
<td>0.8</td>
<td>1.31896</td>
<td>1.31475</td>
<td>1.47415</td>
<td>1.46124</td>
<td>1.52473</td>
<td>1.49714</td>
<td>1.50248</td>
<td>1.45725</td>
<td></td>
</tr>
<tr>
<td>0.9</td>
<td>1.62750</td>
<td>1.62111</td>
<td>1.82811</td>
<td>1.80869</td>
<td>1.84867</td>
<td>1.80909</td>
<td>1.77041</td>
<td>1.70915</td>
<td></td>
</tr>
<tr>
<td>0.7</td>
<td>0.6</td>
<td>1.25418</td>
<td>1.25034</td>
<td>1.38840</td>
<td>1.37682</td>
<td>1.43832</td>
<td>1.41358</td>
<td>1.42598</td>
<td>1.38489</td>
</tr>
<tr>
<td>0.8</td>
<td>1.35971</td>
<td>1.35524</td>
<td>1.52587</td>
<td>1.51208</td>
<td>1.57522</td>
<td>1.54591</td>
<td>1.54616</td>
<td>1.49848</td>
<td></td>
</tr>
<tr>
<td>0.9</td>
<td>1.53780</td>
<td>1.53209</td>
<td>1.73355</td>
<td>1.71600</td>
<td>1.76688</td>
<td>1.73053</td>
<td>1.70542</td>
<td>1.64827</td>
<td></td>
</tr>
<tr>
<td>0.9</td>
<td>1.89839</td>
<td>1.88970</td>
<td>2.08005</td>
<td>2.05520</td>
<td>2.05168</td>
<td>2.00352</td>
<td>1.92459</td>
<td>1.85306</td>
<td></td>
</tr>
</tbody>
</table>
As shown in Table 4, the RE’s increase as the selection probability increases from 0.6–0.9. This means that the proposed estimators are more efficient than the ordinary M–S’s RR estimator for given parameters.

6. Concluding Remarks

We incorporated the calibration procedure into Warner’s and M–S’s RR models, and then compared the relative efficiencies for different \( P, \pi \) and \( T \). If the cell counts are available for auxiliary variables at population level, then the ordinary Warner’s estimator becomes the post-stratified estimator, and if only the marginal counts are available, then it becomes the calibration estimator. So, we have derived the estimators and variances. Furthermore, we have investigated the conditionality for the variance and the bias for two calibration estimators. For a large sample size, we have shown that the calibration estimator is asymptotically equal to the post-stratified estimator and its bias is unaffected by variance. As the selection probability \( P \) increases from 0.6–0.9, then both \( \text{RE}_1 \) and \( \text{RE}_2 \) increase for both the Warner and M–S calibration estimators.

In this article, we use the same randomization device to keep equal protection for both males and females. In the further study, we will consider that different randomization devices could be used for collecting information from females and males.

References


