Bayes linear estimator for two-stage and stratified randomized response models

Chang Kyoon Son\textsuperscript{a} and Jong Min Kim\textsuperscript{b,∗}
\textsuperscript{a}Department of Applied Statistics, Dongguk University-Gyeongju, Gyeongju, Gyeongbuk, Korea
\textsuperscript{b}Statistics Discipline, Division of Science and Mathematics, University of Minnesota-Morris, Morris, MN, USA

Abstract. In this paper, we suggest the Bayes linear estimator (BLE) for randomized response model (RRM) to improve the efficiency of RR estimators, only using the first and second prior moments. The randomized response model is an indirect questioning technique used to protect the privacy of respondents in a survey regarding a sensitive characteristic. Meanwhile Bayes linear estimation is useful for parameter estimation compared to the typical Bayesian method because it only uses the first and second prior knowledge of the variable of interest. Also, it has an advantage of robustness with the distribution.

We suggest the Bayes linear estimators for the two-stage and the stratified RRM and find the optimal sample size to minimize the Bayes risk for the stratified RRM. Also, we show the difference in efficiency between the Bayes linear estimators and the typical non-Bayesian RR estimators by simulation study.

Keywords: Bayes linear estimator, two-stage randomized response model, stratified randomized response model, optimal sampling design

1. Introduction

In a social survey of sensitive characteristics such as tax evasion, drug abuse, or HIV infection, the direct question method leads to an occurrence of response bias due to an untruthful or a refusal to answer because they are concerned about disclosing their privacy.

Warner \cite{21} first suggested the randomized response model as an indirect question method to reduce the response bias by using a randomization device. After his work, Horvitz, Shah and Simmons \cite{11} suggested the unrelated question model to improve upon the Warner model by using the unrelated question, and Greenberg et al. \cite{9} completed the framework of the unrelated question model.

Drane \cite{5} suggested the forced answer model which use both a sensitive and forced answer question. Mangat and Singh \cite{15} suggested the two-stage RRM which uses two randomization devices, and Kim et al. \cite{1} extended their RRM which uses both the unrelated question and the forced question.

Recently Kim and Warde \cite{12} suggested the stratified RRM to increase the efficiency of estimators using the Warner model for each stratum.

Meanwhile Lesyseiffer and Warner \cite{13} emphasized that the RRM needs to have a suitable compromise between privacy protection and the limitation of information loss, so they considered Bayesian estimation for the RRM.

In the Bayesian context, Winkler and Franklin \cite{23} applied the Bayesian estimation method to Warner’s RRM, and Pitz \cite{17} suggested applying the Bayes estimator to the unrelated RRM.

O’Hagan \cite{16} considered the midway between a super-population model and a randomization approach. He assumed only second order exchangeability, which in practice means the need of stating first and second order

∗Corresponding author: Jong Min Kim, Statistics Discipline, Division of Science and Mathematics, University of Minnesota-Morris, Morris, MN 56267, USA. E-mail: jongmink@morris.umn.edu.
moments only, describing prior information about the structures present the population. Also in this approach O’Hagan [16] suggested applying the Bayes linear estimator to the RRM.

Later, Yeum and Son [4] adapted the Bayes linear estimator to the two-stage RRM suggested by Mangat and Singh [15].

Now, we suggest the Bayes linear estimators for the two-stage and the stratified RRM’s and we find the optimal sampling size of the Bayes linear estimator for the stratified RRM.

The paper is organized as follows: In Section 2, we review the Bayes linear estimator. Section 3 is devoted to the Bayes linear estimators for several RRMs. In Section 4 we apply the BLE to the stratified RRM and find the optimal sample size under a constraint minimization of Bayes risk. Section 5 is devoted to the simulation study by assumption of a prior knowledge. Section 6 offers some concluding remarks and suggestions for further study.

2. Bayes linear estimator

The Bayes linear estimator worked by Whittle [22], Stone [20], Hartigan [10], Goldstein [8], and Smouse [19] suggested the Bayes linear estimator for the super-population model.

O’Hagan [16] considered it for the simple marginal probability distribution. In the super-population approach, the BLUE is mathematically equivalent to the Bayes linear estimator.

In this respect, Royall [18] adapted the BLUE to sampling theory which satisfies the following theorem:

Theorem 1. For any random variables \( Y \) and random n-vector \( X \) whose joint distribution has the finite second moments with \( E(Y) = \mu_y, V(Y) = \sigma^2_y, V(X) = \sigma^2_x, \) and \( \text{Cov}(X,Y) = \sigma_{xy} \), if for any constants \( a, b \), \( E(Y|X) = a + b'X \) holds, then we can find \( a^*, b^* \) subject to minimize following Eq. (1)

\[
\begin{align*}
\min_{a,b} Z(a,b) &= E \left[ E(Y|X) - (a + b'X) \right]^2, \\
a^* &= \mu_y + b^*\mu_X, \\
b^* &= \sigma_{xy}/\sigma^2_x
\end{align*}
\]

The estimates \( a^*, b^* \) substitute to Eq. (1), then

\[
Z(a^*, b^*) = \sigma^2_y - \sigma^2_{xy}/\sigma^2_x.
\]

From this theorem, the following estimator among all linear estimators of \( Y \) has a minimal expected squared error loss.

\[
\hat{Y}(x) = a^* + b^*x.
\]  

(2)

For given observed vector value \( X = x \), the estimate is given by

\[
\hat{Y}(x) = a^* + b^*(x - \mu_x)
\]

(3)

From the estimator Eq. (3), the Bayes linear estimator can be expressed with first and second moments of \( X \) and \( Y \). The expected squared error of \( Y(x) \) (the precision of it) becomes the asymptotic variance

\[
V(\hat{Y}(x)) = \sigma^2_y - \sigma^2_{xy}/\sigma^2_x.
\]

(4)

If \( Y \) has a normal distribution, then \( \hat{Y}(x) \) is the posterior mean of \( Y \) given \( X = x \), and \( V(\hat{Y}(x)) \) is the posterior variance.

3. Bayes linear estimator for RRMs

In this section we first review the Bayes linear estimators of the Warner model, suggested by O’Hagan [16], and then we suggest the BLE for the two-stage unrelated model and forced answer models.
3.1. Warner RRM

In the Warner model, the respondents answer the question for a sensitive attribute with selection probability \( P \) or for a non-sensitive attribute with \((1 - P)\) as follows;

**Question 1:** Do you have a sensitive attribute \( A \)? (with probability \( P \))

**Question 2:** Do you have a non-sensitive attribute \( A^c \)? (with probability \( 1 - P \))

Let \( A_i = 1 \) if individual \( i \) has a sensitive attribute, \( A_i = 0 \) otherwise. Then the population proportion that has the sensitive attribute is \( Y = \frac{\sum_{i=1}^{N} A_i}{N} \). The simple random sample from the population consists of \( n \) responses. \( X_j = 1 \) if the \( j \)th response is “Yes”, otherwise \( X_j = 0 \).

We assume the second-order exchangeability that every \( A_i \) has the same prior mean and variance and every pair \((A_i, A_j)\) has the same prior covariance. The expectation, variance and covariance of \( A_i \) are then given by

\[
E(A_i) = m, \quad V(A_i) = m(1 - m), \quad Cov(A_i, A_j) = rm(1 - m), \quad i, j = 1, 2, \ldots, N,
\]

where \( r \in [-1, 1] \), the correlation of any pair \((A_i, A_j)\).

From these assumptions, we can derive the prior expectations of the variables of interest \( X \) and \( Y \) as follows;

\[
E(Y) = m, \\
V(Y) = m(1 - m) \frac{[1 + r(n - 1)]}{N}, \\
E(X_i) = Pm + (1 - P)(1 - m) = (1 - P) + (2P - 1)m, \\
V(X_i) = P(1 - P) + m(1 - m)(2P - 1)^2, \\
Cov(X_i, X_j) = (2P - 1)^2 rm(1 - m), \\
Cov(Y, X_i) = (2P - 1)m(1 - m) \frac{[1 + r(n - 1)]}{N},
\]

where \( P \) is the selection probability of a sensitive question in a randomization device.

Let the sample proportion of a simple random sample with size \( n \) drawn from a population be denoted \( \bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} \). Then the first and second order moments are given by, respectively

\[
E(\bar{X}) = (1 - P) + m(1 - 2P), \\
V(\bar{X}) = \frac{n^{-1} [P(1 - P) + m(1 - m)(2P - 1)^2] + r(n - 1)]}{N} = v_W, \\
Cov(Y, \bar{X}) = \frac{(2P - 1)m(1 - m) [1 + r(n - 1)]}{N} = c_W.
\]

where the subscript \( W \) of variance and covariance denotes the Warner’s RRM.

Then the Bayes linear estimate is given by

\[
\hat{Y}(\bar{x}) = m + \left( \frac{c_W}{v_W} \right) [\bar{x} - (1 - P) - (2P - 1)m] \\
= \alpha_W \bar{y}_W + (1 - \alpha_W)m, \tag{5}
\]

where \( \bar{y}_W = (2P - 1)^{-1}[\bar{x} - (1 - P)] \) is the original non-Bayesian Warner RR estimate, and \( \alpha_W = (2P - 1)\left( \frac{c_W}{v_W} \right) \) is the weight between the Bayes and non-Bayesian estimate.

Using the prior variance and covariance, the posterior variance is given by

\[
V(\hat{Y}(\bar{x})) = (1 - \alpha_W)m(1 - m) \frac{[1 + r(n - 1)]}{N} \tag{6}
\]

3.2. Mangat-Singh RRM

Mangat and Singh [15] considered the two-stage RRM, where the respondents should answer “Yes” or “No” in the following two randomization devices, \( R_1 \) and \( R_2 \), respectively.
3.3. Two-stage unrelated RRM

In these randomization devices the respondents should answer to the first-stage randomization device $R_1$ with probability $T$ and $1-T$, and then they answer to the second-stage randomization device $R_2$ with probability $P$ and $1-P$.

Similar to Section 3.1, we can derive that the expectation and variance of $X$ for the two-stage RRM are given by

$$E(X_i) = Tm + (1 - T) [(1 - P) + (2P - 1)m] ,$$
$$V(X_i) = [(2P - 1) + 2T(1 - P)]^2 m(1 - m) + (1 - P)(1 - T) [T + P(1 - T)].$$

And the covariances are given by

$$Cov(X_i, X_j) = [(2P - 1) + 2T(1 - P)]^2 rm(1 - m) ,$$
$$Cov(Y, X_i) = [(2P - 1) + 2T(1 - P)] m(1 - m) [1 + r(N - 1)] / N .$$

Let the sample proportion of a simple random sample with size $n$ drawn from a population be denoted $\bar{X} = n^{-1} \sum_{i=1}^{n} X_i$. Then the first and second order moments for the two-stage RRM are given, respectively

$$E(\bar{X}) = Tm + (1 - T) [(1 - P) + (2P - 1)m] ,$$
$$V(\bar{X}) = n^{-1} [(2P - 1) + 2T(1 - P)]^2 m(1 - m) + (1 - P)(1 - T) [T + P(1 - T)] + (n - 1) [(2P - 1) + 2T(1 - P)]^2 rm(1 - m)] = v_{MS} ,$$
$$Cov(\bar{Y}, \bar{X}) = N^{-1} [(2P - 1) + 2T(1 - P)] m(1 - m) [1 + r(N - 1)] = c_{MS} ,$$

where $T$ is the selection probability of a sensitive question in the first stage random device of the M-S model, and $P$ is the selection probability of a sensitive question in the second stage random device of the model.

Thus the Bayes linear estimator of the Mangat-Singh (M-S) RR model, given observation $x$ for a simple random sample of size $n$, is defined by

$$\hat{Y}(\bar{x}) = m + (c_{MS}/v_{MS}) [\bar{x} - Tm - (1 - T) [(1 - P) + (2P - 1)m]]$$
$$= \alpha_{MS} \hat{y}_{MS} + (1 - \alpha_{MS})m ,$$

where $\hat{y}_{MS} = [\bar{x} - (1 - T) [(1 - P) + (2P - 1)]] [2P - 1] + 2T(1 - P) \{c_{MS}/v_{MS}\}$ is the typical non-Bayesian M-S RR estimate, and $\alpha_{MS} = \{[(2P - 1) + 2T(1 - P)](c_{MS}/v_{MS})\}$ be the weight of between of Bayes and non-Bayesian estimate.

Then the posterior variance is given by

$$V(\hat{Y}(\bar{x})) = (1 - \alpha_{MS})m(1 - m) [1 + r(N - 1)] / N .$$

3.3. Two-stage unrelated RRM

Kim et al. [1] suggested the two-stage unrelated RR model. In this model, the first randomization device $R_1$ is similar to the M-S model, whereas the Second $R_2$ is consistent to the unrelated question suggested by Horvitz et al. [1].

Then the respondents answer “Yes” or “No” with $R_1$ at first stage, and then they answer to the unrelated question at second-stage as follows;

$$R_2 \ \text{Question 1: Do you have a sensitive attribute } A? (P)$$
$$R_2 \ \text{Question 2: Do you have a non-sensitive attribute } B? (1 - P)$$
Now, we assume two characteristics, $A_i$ and $B_i$, for an individual $i (i = 1, 2, \ldots, N)$, where $A_i$ is a sensitive attribute but $B_i$ is not. Also, we assume that they are exchangeable and independent. Then the prior moments of $A_i$ and $B_i$ are defined as

$$E(A_i) = m_A, \quad E(B_i) = m_B,$$

$$V(A_i) = v_A, \quad V(B_i) = v_B,$$

$$\text{Cov}(A_i, A_j) = c_A, \quad \text{Cov}(B_i, B_j) = c_B$$

Let the sample proportion of a simple random sample with size $n$ drawn from a population be denoted $\bar{X} = \sum_{i=1}^{n} X_i$. Then the expectation and variance are given by

$$E(X_i) = Tm_A + (1-T)(Pm_A + (1-P)m_B) = E(\bar{X}),$$

$$V(X_i) = [T+p(1-T)]^2v_A + (1-T)^2(1-P)^2v_B + [T+p(1-T)](1-T)(1-P)(m_A-m_B)^2,$$

$$\text{Cov}(X_i, X_j) = [T+(1-T)p^2]c_A + (1-T)^2(1-P)^2c_B,$$

$$V(\bar{X}) = n^{-1} [(T+p(1-T))^2v_A + (1-T)^2v_B] + (n-1)(T+p(1-T))(1-T)(1-P)(m_A-m_B)^2 = v_{2UR}.$$

Also, the first and second moments of $Y$ are given by

$$E(Y) = m_A,$$

$$\text{Cov}(Y, \bar{X}) = [T + P(1-T)][v_A + (N-1)c_A]/N = c_{2UR}.$$

As before, the Bayes linear estimator for the two-stage unrelated RR model, given observation $x$, is the following:

$$\hat{Y}(x) = m_A + (c_{2UR}/v_{2UR})[\hat{x} - Tm_A - (1-T)p^2m_A - (1-T)(1-P)m_B$$

$$= \alpha_{2UR}\hat{y}_{2UR} + (1-\alpha_{2UR})m_A, \quad (9)$$

where $\alpha_{2UR} = [T + P(1-T)](c_{2UR}/v_{2UR})$ and $\hat{y}_{2UR} = [\hat{x} - (1-T)(1-P)m_B][T + P(1-T)]^{-1}$ is a typical non-Bayesian two-stage unrelated RR estimator.

The corresponding variance of it is given by

$$V(\hat{Y}(x)) = (1-\alpha_{2UR})[v_A + (N-1)c_A]/N. \quad (10)$$

### 3.4. Two-stage forced RRM

Similar to Section 3.3, Kim et al. [11] suggested the two-stage forced answer RR model which is composed of two randomization devices, $R_1$ and $R_2$, where $R_1$ is the first randomization device of M-S, and $R_2$ is the forced answer, defined as follows:

- **Question 1**: Do you have a sensitive attribute $A$? ($P$)
- **Question 2**: Answer “Yes”, ($1-P$)

Then, using the assumptions of the prior moment in Section 3.2, we can derive the first and second moments of $X$ for the two-stage forced answer RRM response for a simple random sample of size $n$ as follows:

$$E(X_i) = Tm + (1-T)[Pm + (1-P)] = E(\bar{X}),$$
The first and second moments of a sensitive population proportion $Y$ are given by
\[
E(Y) = m, \\
V(Y) = m(1-m) \left[1 + r(N-1)\right]/N, \\
\text{Cov}(Y, \bar{X}) = \left[T + P(1-T)\right] m(1-m) \left[1 + r(N-1)\right]/N = c_{2F}.
\]

Then, the Bayes linear estimator for two-stage forced answer RRM is given by
\[
\hat{Y}(\bar{x}) = m + \left[c_{2F}/v_{2F}\right] \left[\bar{x} - Tm - (1-T)Pm - (1-T)(1-P)\right] \\
= \alpha_{2F} \tilde{y}_{2F} + (1 - \alpha_{2F})m,
\]
where $\alpha_{2F} = c_{2F}/v_{2F}$, and $\tilde{y}_{2F} = \left[\bar{x} - (1-T)(1-P)\right] \left[T + P(1-T)\right]^{-1}$ is the typical non-Bayesian estimator for the two-stage forced answer RR model.

The posterior variance is
\[
V(\hat{Y}(\bar{x})) = (1 - \alpha_{2F})m(1-m) \left[1 + (N-1)\right]/N.
\]

4. Bayes linear estimator for the stratified RRM

4.1. Stratified RRM

Kim and Warde [12] considered that the population is partitioned into several strata and a sample is selected by simple random sampling with replacement in each stratum. To construct the stratified RR model, they assumed that an individual respondent in the sample of stratum $h$ is instructed to use the randomization device $R_{th}$, which consists of a sensitive question ($S$) card with probability $P_h$ and a negative question ($S^c$) card with probability $1-P_h$. The respondent should answer the question by “Yes” or “No”. A respondent belonging to the sample in different strata will perform different randomization devices, each having different pre-assigned probabilities. Let $n_h$ denote the number of units in the sample from stratum $h$ and $n = \sum_{h=1}^{H} n_h$ is the total sample size over all strata. Then the probability of a “Yes” answer in stratum $h$ is
\[
Z_h = P_h\theta_h + (1 - P_h)(1 - \theta_h), \quad h = 1, 2, \ldots, H,
\]
where $Z_h$ is the proportion of “Yes” answers in stratum $h$, $\theta_h$ is the proportion of respondents with the sensitive attribute in stratum $h$, and $P_h$ is the probability that a respondent in the sample stratum $h$ has a sensitive question ($S$) card.

For a stratum $h$, the MLE of the population proportion with a sensitive attribute $\theta_h$ is given by
\[
\hat{\theta}_h = \frac{\hat{Z}_h - (1 - P_h)}{2P_h - 1}, \quad h = 1, 2, \ldots, H,
\]
where $\hat{Z}_h$ is the sample proportion of “Yes” answer in a sample stratum $h$. 

\[
V(X_i) = \{T + P(1-T)\}^2 m(1-m) + (1-T)(1-P)(T + P(1-T)), \\
\text{Cov}(X_i, X_j) = \{T + P(1-T)\}^2 rm(1-m), \\
\text{Cov}(Y, X_i) = \{T + P(1-T)\} m(1-m) \left[1 + r(N-1)\right]/N, \\
V(\bar{X}) = n^{-1} \left\{\{T + P(1-T)\}^2 m(1-m) + (1-T)(1-P)(T + P(1-T)) \right. \\
\left.+ \{T + P(1-T)\}^2 rm(1-m)(n-1)\right\} = v_{2F}.
\]
Thus, the stratified estimator of the population proportion with a sensitive attribute $\theta$ is

$$
\hat{\theta}_{st} = \sum_{h=1}^{H} W_h \hat{\theta}_h = \sum_{h=1}^{H} W_h \hat{Z}_h - \frac{(1 - P_h)}{(2P_h - 1)},
$$

(15)

where $W_h = N_h / N$ is the stratum weight of stratum $h$, where the whole population has size $N$ and the population stratum $h$ has size $N_h$ for $h = 1, 2, \ldots, H$. The variance of $\hat{\theta}_{st}$ is

$$
V(\hat{\theta}_{st}) = \sum_{h=1}^{H} W_h \left[ \frac{\theta_h(1 - \theta_h)}{n} \right] (1 - f_h) + \frac{P_h(1 - P_h)}{n} \left(1 - \frac{1}{2P_h - 1}\right)^2.
$$

(16)

If a sample is drawn by simple random sampling without replacement, then Eq. (16) can be written as

$$
V(\hat{\theta}_{st}) = \sum_{h=1}^{H} W_h \left[ \frac{\theta_h(1 - \theta_h)}{n} \right] (1 - f_h) + \frac{P_h(1 - P_h)}{n(2P_h - 1)^2}.
$$

(17)

4.2. Bayes linear estimator for stratified RRM

To construct the Bayes linear estimator for the stratified RR model, we consider that the population consists of $H$ strata. The sample is selected by simple random sampling with size $n_h$. Let $X_{hi} = 1$ if an individual $i$ among the sample of $n_h$ gives a response of “Yes”, otherwise $X_{hi} = 0$ using the randomization device $R_h$ for each stratum.

Also, we assume the exchangeability of $A_{hi}$ and $A_{hj}$ so that the $A_{hi}$’s have the same prior mean, variance and covariance and the $A_{hi}$’s are uncorrelated in different strata for each stratum $h (h = 1, 2, \ldots, H)$.

Since $A_{hi}$ is a binary variable, we can define the first and second order prior moments of $A_{hi}$ as

$$
E(A_{hi}) = m_h,
$$

$$
V(A_{hi}) = m_h(1 - m_h),
$$

$$
Cov(A_{hi}, A_{hj}) = r_h m_h (1 - m_h),
$$

where $r_h \in [-\left(N_h - 1\right)^{-1}, 1]$, the correlation of any pair $(A_{hi}, A_{hj})$ for any stratum $h$.

For stratum $h$, the prior expectations are

$$
E(Y_h) = m_h,
$$

$$
V(Y_h) = m_h(1 - m_h) \left[1 + r_h(N_h - 1)\right] / N_h,
$$

$$
E(X_{hi}) = P_h m_h + (1 - P_h)(1 - m_h) = (1 - P_h) + (2P_h - 1)m_h,
$$

$$
V(X_{hi}) = P_h(1 - P_h) + m_h(1 - m_h)(2P_h - 1)^2,
$$

$$
Cov(X_{hi}, X_{hj}) = (2P_h - 1)^2r_h m_h (1 - m_h),
$$

$$
Cov(Y_h, X_{hi}) = (2P_h - 1)m_h(1 - m_h) \left[1 + r_h(N_h - 1)\right] / N_h.
$$

Then we can observe the sample proportion of response “Yes” to be $\bar{X}_h = n_h^{-1} \sum_{i=1}^{n_h} X_{hi}$. Furthermore, the expectation, variance and covariance are

$$
E(\bar{X}_h) = (1 - P_h) + (2P_h - 1)m_h,
$$

$$
V(\bar{X}_h) = n_h^{-1} \left[P_h(1 - P_h) + m_h(1 - m_h)(2P_h - 1)^2 \{1 + (n_h - 1)r_h\}\right] = v_h,
$$

$$
Cov(Y_h, \bar{X}_h) = (2P_h - 1)m_h(1 - m_h) \left[1 + (N_h - 1)r_h\right] / N_h = c_h.$$
The Bayes linear estimator for the stratified RR model is, for stratum $h$:

$$
\hat{Y}_h(x_h) = m_h + (c_h/v_h) [x_h - (1 - P_h) - (2P_h - 1)m_h] \\
= \alpha_h \bar{y}_h + (1 - \alpha_h)m_h,
$$

(18)

where $\alpha_h = (2P_h - 1)(c_h/v_h)$, and $\bar{y}_h = (2P_h - 1)^{-1} [x_h - (1 - P_h)]$ is the typical non-Bayesian estimator of the stratified RR model for each stratum $h$ ($h = 1, 2, \ldots, H$).

The posterior variance is, for stratum $h$:

$$
V(\hat{Y}_h(x_h)) = (1 - \alpha_h)m_h(1 - m_h)[1 + (N_h - 1)r_h]/N_h.
$$

(19)

Thus we can find that the stratified Bayes linear estimator and its variance for the population proportion $Y = N^{-1} \sum_{h=1}^{H} \sum_{i=1}^{N_h} A_{hi}$ are given by:

$$
\hat{Y}_{st}(\bar{x}) = \sum_{h=1}^{H} W_h \hat{Y}_h(\bar{x}_h),
$$

(20)

and

$$
V(\hat{Y}_{st}(\bar{x})) = \sum_{h=1}^{H} W_h^2 V(\hat{Y}_h(\bar{x}_h)),
$$

(21)

where $W_h = N_h/N$.

4.3. The optimal sampling design for stratified Bayes linear estimator

Because of the assumption of within stratum exchangeability in Section 4.2, from the subjective Bayesian perspective one can ignore which of the $n_h$ units are sampled from the $h$th stratum. However, the stratum size of $n_h$ should be regarded, as it is related to the prior information about the $A_{hi}$’s and the sampling cost.

We derived the Bayes linear estimator for the population proportion $Y = N^{-1} \sum_{h=1}^{H} \sum_{i=1}^{N_h} A_{hi}$ subject to minimize the Eq. (1). It can also make the optimal design subject to minimize a Bayes risk for any particular cost function $C$.

To make the optimum design, let a convex function $R(n)$ with $n = (n_1, n_2, \ldots, n_H)$, be defined as follows (Erickson [17]):

$$
R(n) = \sum_{h=1}^{H} Q_h(N_h - n_h)/n_h + k_h,
$$

(22)

where $Q_h = k_h [v_h' + (N_h - 1)c_h']/N_h^2$, $k_h = (2P_h - 1)(v_h' - c_h')/c_h' = (2P_h - 1)(1 - r_h)/r_h$ and $v_h'$, $c_h'$ are the prior variance and covariance for the stratified RRM respectively.

Now, if the optimum value of $n_h$ is $n_h^*$, then under the conditions $0 \leq n_h \leq N_h$ and $\sum_{h=1}^{H} q_hn_h \leq C$ where $q_h$ is the sampling cost for stratum $h$, the optimum value $n_h^*$ for stratum $h$, subject to minimize Eq. (22), is given by

$$
n_h^* = n \times \left[ \frac{C + \sum_{h=1}^{H} q_hk_h}{\sum_{h=1}^{H} q_hz_h} \right] - k_h,
$$

(23)

where $z_h = \sqrt{Q_h(N_h + k_h)/q_h}$.
If a prior is a diffuse prior, then the optimum sample size is proportional to the stratum size and standard deviation within stratum and inversely proportional to the square root of the per unit sampling cost.

**Theorem 2.** The optimum allocation of the stratified Bayes linear estimator for stratified RR model is given by

\[
 n_h^o = n \times \frac{C \sqrt{Q_h(N_h + k_h)/q_h}}{\sum_{h=1}^{H} q_h \sqrt{Q_h(N_h + k_h)/q_h}} \tag{24}
\]

**Proof** From Eqs (22) and (23), we can obtain the optimum sample size \( n_h^o \) of stratum size \( n_h \).

From Theorem 1, the optimum allocation of the stratified Bayes linear estimator is similar to the typical allocation method.

**Corollary 1.** For the stratified RRM, if \( k_h \to 0 \) as \( c_h' \to \infty \) and \( c_h'' \to \infty \) in Eq. (24) and the sampling cost \( q_h \) is equal to \( q \) for each stratum, then the optimum allocation Eq. (24) is given by

\[
 n_h^o = n \times \frac{\sqrt{Q_hN_h}}{\sum_{h=1}^{H} \sqrt{Q_hN_h}} = n \times \frac{\sqrt{(2P_n - 1)m_h(1 - m_h)(1 + r_h(N_h - 1))N_h}}{\sum_{h=1}^{H} \sqrt{(2P_n - 1)m_h(1 - m_h)(1 + r_h(N_h - 1))N_h}} \tag{25}
\]

From Eq. (25) in the diffuse prior, the optimum sample size can be rewritten to the power allocation with \( a = 1/2 \).

5. Simulation study

In the simulation study, O’Hagan [16] used the Winiker and Franklin’s [23] analysis which assumed the Beta prior with parameters \( \alpha \) and \( \beta \) for a prior specification by Warner’s RRM. He considered that the prior knowledge is only the first and second order moments with the Beta distribution. As pointed out by O’Hagan [16], the assumption of full prior information is much stronger than that of only the first and second moments. He used the data on abortion in Taiwan that was reported by Liu and Chow [14] and analyzed by Winiker and Franklin [23] to estimate.

Meanwhile, for the forced RRM, Pitz [17] assumed a uniform prior for \( Y \) over \([0, 1]\) in case of a binary response and infinite population. O’Hagan [16] used Pitz’s [17] analysis to estimate the variable of interest and he pointed out that a uniform prior density is different from Pitz’s [17] figure, since the BLE is exact in case of normality. He showed that the computations and his suggestions provided good approximations to exact Bayesian techniques unless the distributions are very non-normal.

5.1. Assumptions of population and prior knowledge

Now, we assume that the prior distribution for each RRM is the Beta prior with parameters \( \alpha \) and \( \beta \). Also, in order to determine the optimal sample size for the Bayes linear estimator of the stratified RRM, without loss of generality, we assume that the population is partitioned into three strata of different sizes, \( N_h \) \((h = 1, 2, 3)\), for a convenient comparisons between the BLE and non-Bayesian estimator. For the stratified RRM we assume that the prior density has a Beta distribution with parameters \( \alpha_h \) and \( \beta_h \) for each stratum \( h \) \((h = 1, 2, 3)\).

To use the Bayes estimator for the given RRM’s, the prior parameters \( m \) and \( r \) need to be specified. For the variable of interest \( Y \), \( m \) is a prior estimate of \( Y \), and the correlation \( r \) is the strength of prior knowledge, so that a higher value of \( r \) means weaker prior information.

We generate an artificial population of size \( N = 10,000 \) and it is partitioned into 3 strata with arbitrary size \( N_1 = 3000, N_2 = 3000, N_3 = 4000 \) for the stratified RRM. Also, the sample is selected by simple random sampling without replacement (SRSWOR) of size \( n = 1000 \), and for the stratified RRM, \( n_1 = 300, n_2 = 300, n_3 = 400 \) by proportional allocation. We summarize the population and sample for each situation to simulate in the following Table 1.
By the Beta prior density we can define the first and second moments of the variable of interest with the parameters \( \alpha, \beta \) as follows;

\[
E(Y) = \frac{\alpha}{\alpha + \beta}, \quad V(Y) = \frac{(\alpha + \beta)^2}{(\alpha + \beta + 1)}.
\]  

(26)

And the prior mean and correlation are given by

\[
m = \frac{\alpha}{\alpha + \beta} = E(Y), \quad r = \frac{1}{(\alpha + \beta + 1)^2}.
\]  

(27)

For the stratification, we can express the mean, variance and the prior knowledge by adding the subscript \( h \) to Eqs (26) and (27).

5.2. Simulation results for BLES’

We will describe some results of the proposed BLE for two-stage RRM’s and the stratified RRM. In the simulation study, the population proportion of a sensitive attribute, \( Y = \frac{1}{N} \sum_{i=1}^{N} A_i (= \pi_y) \), is fixed as 0.4 and the selection probabilities of the randomization device increases from 0.6 to 0.8 by 0.1 for \( P \) and \( T \) at the first and second stage respectively. Tables 2 to 5 show the results of the BLE and non-Bayesian estimate with their standard deviations (SD).

In the case of the stratified RRM, we consider that the prior parameters \( \alpha_h, \beta_h \) of a Beta density are different between each stratum, for example \((\alpha_1, \beta_1) = (1, 2)\) for stratum 1, \((\alpha_2, \beta_2) = (1, 10)\) for stratum 2 and \((\alpha_3, \beta_3) = (10, 20)\) for stratum 3. Also, similar to the previous simulation, the selection probabilities of the question are varying from 0.6 to 0.8 by 0.1 for each stratum.

In Figs 1 to 4, we compare the standard deviations between the proposed BLE and the typical MLE of two-stage RRM’s. Also, we compare the standard deviations between the typical MLE of a stratified RRM and the proposed BLE of a stratified RRM.
Table 5
Standard deviations of the BLE for the Stratified RRM

<table>
<thead>
<tr>
<th>$P_3$</th>
<th>$P_1 = 0.6$</th>
<th>$P_2 = 0.6$</th>
<th>$P_2 = 0.7$</th>
<th>$P_2 = 0.8$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.6</td>
<td>0.0577</td>
<td>0.0566</td>
<td>0.0557</td>
<td></td>
</tr>
<tr>
<td>0.7</td>
<td>0.0551</td>
<td>0.0539</td>
<td>0.0530</td>
<td></td>
</tr>
<tr>
<td>0.8</td>
<td>0.0535</td>
<td>0.0522</td>
<td>0.0513</td>
<td></td>
</tr>
<tr>
<td>$P_3$</td>
<td>$P_1 = 0.7$</td>
<td>0.0471</td>
<td>0.0457</td>
<td>0.0446</td>
</tr>
<tr>
<td>0.6</td>
<td>0.0438</td>
<td>0.0423</td>
<td>0.0411</td>
<td></td>
</tr>
<tr>
<td>0.7</td>
<td>0.0417</td>
<td>0.0402</td>
<td>0.0389</td>
<td></td>
</tr>
<tr>
<td>0.8</td>
<td>0.0434</td>
<td>0.0419</td>
<td>0.0407</td>
<td></td>
</tr>
<tr>
<td>$P_3$</td>
<td>$P_1 = 0.8$</td>
<td>0.0398</td>
<td>0.0382</td>
<td>0.0368</td>
</tr>
<tr>
<td>0.6</td>
<td>0.0376</td>
<td>0.0358</td>
<td>0.0344</td>
<td></td>
</tr>
<tr>
<td>0.7</td>
<td>0.0355</td>
<td>0.0339</td>
<td>0.0330</td>
<td></td>
</tr>
<tr>
<td>0.8</td>
<td>0.0335</td>
<td>0.0322</td>
<td>0.0313</td>
<td></td>
</tr>
</tbody>
</table>

Figure 1 shows that the BLE of the two-stage M-S RRM(s_bms) has a smaller dispersion measure than that of the typical RRM(s_ms). Figure 2 shows that the BLE of the two-stage unrelated RRM(s_bus) has a smaller dispersion measure than that of the typical RRM(s_us). In this case, the BLE is sensitive to the prior information of the unrelated question. Figure 3 shows that the BLE of the two-stage forced RRM(s_bfs) has a smaller dispersion measure than that of the typical RRM(s_fs).

Finally, Fig. 4 shows the BLE of the stratified RRM(s_bay_st) has a smaller dispersion measure than that of the typical stratified RRM(s_st).

5.3. Optimum stratum size for the BLE of the stratified RRM

By Kim and Warde [12], the optimum allocation sample size is given by

$$n^*_h = n \times \frac{N_h \sqrt{\theta_h(1 - \theta_h) + P_h(1 - P_h)/(2P_h - 1)^2}}{\sum_{h=1}^{L} N_h \sqrt{\theta_h(1 - \theta_h) + P_h(1 - P_h)/(2P_h - 1)^2}}.$$  (28)
We can compare the optimum sample size between Eq. (25) for the BLE and Eq. (26) for the MLE under the fixed sampling cost. In Figs 5(a)–(c), the optimum sample size of the BLE is represented by the solid lines and that of the MLE is represented by the dotted lines for strata 1, 2 and 3 under the fixed sampling cost respectively. We can find that the optimum sample size of the BLE is more stable over varying selection probabilities than that of the typical MLE. Also, the optimum sample size of the BLE is likely to converge to the MLE by increasing the selection probabilities of the question.

6. Concluding remarks

We suggest the Bayes linear estimator for several RRM’s to improve the efficiency of an RR estimator using the prior knowledge of a population. Our study was motivated by O’Hagan [16], and he mentioned the availability of his works to extend any other RR design. However, he did not consider the two-stage RR and an optimal design for the stratified RRM.

In this point view, we suggested using the BLE for the two-stage RRM and the stratified RRM, and a method of determining the optimal sample size for the stratified Bayes RRM. Also, we compared the efficiency of the BLE and the typical RR estimator by simulation study.

We summarize some results of the proposed BLE for the RRM as follows;

Firstly, the proposed BLE’s for the two-stage RRM is more efficient than that of the typical two-stage RRM. Secondly, the dispersion measure (SD) of the two-stage RRM is decreasing according to change in the selection probabilities \( P \) and \( T \), but that of the BLE is decreasing according to the prior information. Thirdly, for the unrelated question RRM, the BLE is more sensitive than the typical RRM to a change of the unrelated question proportion. Finally, for the stratified RRM, we can find that the optimum sample size of the BLE converges to the sample size of the MLE by increasing the selection probabilities of the question.

Acknowledgment

This work was supported by the Dongguk University Research Fund of 2012.

References


