TWO STAGE IMPROVED RANDOMIZED RESPONSE MODEL FOR QUANTITATIVE DATA

Rawan Arafa*, Reda Mazloum
Faculty of Economical and Political Science, Cairo University, Egypt.

ABSTRACT
Researchers often have problems getting truthful responses when asking questions related to sensitive personal, financial, or societal topics. The randomized response technique was developed to guarantee the respondents privacy by concealing their true response, thus ensuring more truthful cooperation. This paper introduces a two stage randomized response model for sensitive quantitative data. The model and its estimator are developed for use with simple and stratified random sampling. Under each sampling scheme, the efficiency of the estimator is investigated with respect to various estimators and it is found to be more efficient. A Simulation study using data on age of first alcohol consumption is conducted to showcase the gains in efficiency when using the proposed estimator compared to other competing estimators.

KEYWORDS: Estimation of mean, Quantitative data, Randomized response technique, Sensitive questions, Two stage, Scrambling variable.

MSC: 62D05.

RESUMEN
Los investigadores a menudo tienen problemas para obtener respuestas veraces cuando hacen preguntas relacionadas con temas personales, financieros o sociales delicados. La técnica de respuesta aleatoria se desarrolló para garantizar la privacidad de los encuestados ocultando su verdadera respuesta, asegurando así una cooperación más veraz. Este artículo presenta un modelo de respuesta aleatoria en dos etapas para datos cuantitativos sensibles.El modelo y su estimador están desarrollados para su uso con muestreo aleatorio simple y estratificado. En cada esquema de muestreo, se investiga la eficiencia del estimador con respecto a varios estimadores y se encuentra que es más eficiente. Se lleva a cabo un estudio de simulación utilizando datos sobre la edad del primer consumo de alcohol para mostrar las ganancias en eficiencia cuando se utiliza el estimador propuesto en comparación con otros estimadores de la competencia.

PALABRAS CLAVE: Estimación de la media, Datos cuantitativos, Técnica de respuesta aleatoria, Preguntas sensibles, Dos etapas, Variable de aleatorización.

*Rawan.Ebrahim2012@feps.edu.eg

1. INTRODUCTION

Researchers often have problems getting truthful responses when asking questions related to sensitive personal, financial, or societal topics. Questions related to tax evasions, sexual tendencies, or drug usage causes respondents to either avoid getting questioned altogether, or provide incorrect information. Warner (1965) was the first to suggest a technique, called the randomized response technique (RRT), to induce more cooperation from respondents. Warner's model was designed for dichotomous variables, where the respondent may belong to either a stigmatizing group "A", or it's complement "A^c". The respondent, based on the outcome of the randomization device, provides an answer to one of the two questions "Do you belong to group A?", or "Do you belong to A^c?", with probabilities p and $1-p$, respectively, with $p \in (0,1)$. A randomization device, which could be a spinner, a deck of cards, or a box with different colored balls, is used to determine which question the respondent answers. This process is carried out without the involvement of the interviewer and without his knowledge of the outcome, therefore encouraging more cooperation from the respondents.

Since then, the RRT has been used in many surveys and proved to provide better response rates and more reasonable estimates of the parameters under study. For example, Goodstadt and Gruson (1975) used an RR model to inquire about drug usage and reported that using RRT resulted in a significantly reduced non-response figure compared to the direct question method. Lensvelt-Mulders et al. (2005) conducted meta-analyses that compared the performance of RR models and direct questioning and found that RRT elicited more socially undesirable answers. Krumpal (2013) has a review on many comparative studies that show that the added cost of RRT is balanced by a better estimate of the parameter under study.

To estimate the mean of a quantitative random variable, Greenberg et al. (1971) provided an RR model using an unrelated question for the randomization. The problem with the unrelated question method is the need to find another variable with known parameters that can be used in the randomization. The unrelated question should have a matching range to the sensitive study question. If no such variable is available, two samples would need to be drawn in order to estimate the averages of the sensitive and the unrelated variables. For this reason, the scrambled response technique was considered and developed by many researchers. According to this technique, the respondent multiplies (or adds) an independently generated value to his/her answer, discretely, before reporting to the interviewer.

The first to study the usage of a scrambling variable to obtain the mean of a sensitive random variable in extensive details were Eichhorn and Hayre (1983). According to Eichhorn and Hayre (1983) multiplicative model, each respondent in a simple random sample of size n provides the interviewer with the answer $Z = XS$, where X is the answer to the sensitive question, S is a random variable independent of X. This means that the respondent never has to reveal his answer to the sensitive question as $P(S = 1) = 0$. Assume that $X \geq 0$ and $S > 0$, and let μ_X and σ_X^2 be the mean and variance of X, respectively, while, μ_S and σ_S^2 be the known mean and variance of S. Based on a simple

random sample with replacement (SRSWR) of size n, an unbiased estimator for μ_X is given by:

$$\hat{\mu}_E = \frac{\bar{Z}}{\mu_S}$$

where $\bar{Z}$ is the moment estimator of the mean of the responses.
The variance of $\hat{\mu}_E$ is given by:

$$V(\hat{\mu}_E) = \frac{\mu_X^2}{n}[C_X^2 + C_S^2(1 + C_X^2)]$$

where C_X and C_S are the coefficients of variation of the sensitive variable X and the scrambling variable S, respectively.

Bar-Lev et al. (2004) proposed a quantitative RR model that combines a randomization mechanism with the scrambling technique used by Eichhorn and Hayre (1983). According to this model, the interviewee's response is as follows:

$$Z = \begin{cases} X, & \text{with probability } p \\ XS, & \text{with probability } 1-p \end{cases}$$

where $p \in (0, 1)$. The respondents conduct a simple Bernoulli experiment and based on the outcome either report the true value of X, or report the scrambled response XS. The experiment is conducted away from the interviewer and its result is not disclosed.
Based on a SRSWR of size n, an unbiased estimator of μ_X is given by:

$$\hat{\mu}_B = \frac{\bar{Z}}{p + \mu_S(1-p)}$$

with variance:

$$V(\hat{\mu}_B) = \frac{\mu_X^2}{n}[C_X^2 + C_S^*(p)(1 + C_X^2)] \tag{1.1}$$

where

$$C_S^*(p) = \frac{p + (1-p)(\mu_S^2 + \sigma_S^2)}{(p + \mu_S(1-p))^2} - 1$$

They showed that their estimator, $\hat{\mu}_B$, is more efficient than Eichhorn and Hayre (1983) estimator, $\hat{\mu}_E$ if the distribution of the scrambling variable S satisfies the following condition

$$0 < \mu_S < \frac{2E(S^2)}{1 + E(S^2)} \tag{1.2}$$

They also suggested the exponential distribution with mean $\mu_S = 1/\lambda$, where $2 - \sqrt{2} < \lambda < 2 + \sqrt{2}$, to be used for the variable S since it satisfies the aforementioned condition in Equation 1.2 for uniform efficiency.

Ryu et al. (2005) suggested the following two-stage RR model.

$$Z = \begin{cases} X, & \text{with probability } p \\ \text{go to } R_2, & \text{with probability } 1-p \end{cases} \quad \begin{cases} X, & \text{with probability } t \\ XS, & \text{with probability } 1-t \end{cases}$$

In their analysis they set the mean of the scrambling variable, μ_S, equal to 1 and obtained the following unbiased estimator of μ_X:

$$\hat{\mu}_R = \bar{Z}$$

with variance given by:

$$V(\hat{\mu}_R) = \frac{\mu_X^2}{n}[C_X^2 + (1-p)(1-t)\sigma_S^2(1+C_X^2)] \tag{1.3}$$

They showed that their estimator is more efficient than that of Greenberg et al. (1971), and Gupta et al. (2002) optional RR model's estimator. Unlike the aforementioned compulsory RR models, the optional randomized response technique (ORRT) proposed in Gupta et al. (2002) suggests that instead of conducting a Bernoulli trial to determine whether the respondent answers X or the scrambled response XS, it is up to the respondents themselves whether they answer truthfully or use the scrambling method. The respondents make the choice without informing the researcher and report their answers. Thus, if ORRT is to be used, the question sensitivity which determines the proportion that would report the scrambled response needs to be estimated along with μ_X to determine the variance of the estimator.

They also extended their model to the stratified random sampling. Suppose a population of size N is divided to k strata each having N_h individuals, where $h = 1, 2, ..., k$, and k independent SRSWRs are selected, each of size n_h, with $n = \sum_{h=1}^{k} n_h$ as the total sample size. Each respondent in a SRSWR of size n_h from stratum h is provided with two random devices R_{1h} and R_{2h}. The first randomization device, R_{1h}, has two statements (i) report your true response X_h for the sensitive question, and (ii) go to R_{2h}, with probabilities p_h and $1-p_h$, respectively. The second randomization device, R_{2h}, has two statements (i) report your true response X_h for the sensitive question, and (ii) report the scrambled response X_hS_h, with probabilities t_h and $1-t_h$, respectively.
Assuming that $\mu_{S_h} = 1$, for $h = 1, \ldots, k$, they obtained the following unbiased estimator of μ_X:

$$\hat{\mu}_R^S = \sum_{h=1}^{k} w_h \bar{Z}_h$$

with variance given by:

$$V(\hat{\mu}_R^S) = \sum_{h=1}^{k} \frac{w_h^2 \mu_{X_h}^2}{n_h}\Big[C_{X_h}^2 + (1-p_h)(1-t_h)\sigma_{S_h}^2(1+C_{X_h}^2)\Big] \tag{1.4}$$

where $w_h = N_h/N$ is the weight of the h^{th} stratum in the population.
They also obtained the variance of their estimator in case of Neyman's optimal allocation

$$NeyV(\hat{\mu}_R^S) = \frac{1}{n}\Big(\sum_{h=1}^{k} w_h \mu_{X_h}\sqrt{C_{X_h}^2 + (1-p_h)(1-t_h)\sigma_{S_h}^2(1+C_{X_h}^2)}\ \Big)^2 \tag{1.5}$$

Afterwards, Tarray and Singh (2017) suggested a modification on Bar-Lev et al. (2004) model, which

made it more efficient than the original. According to their model the response is given by:

$$Z = \begin{cases} X, & \text{with probability } p \\ XS^*, & \text{with probability } 1-p \end{cases}$$

where $S^* = \frac{aS+b\mu_S}{a+b}$ and a and b are positive real numbers. The estimator of μ_X is given by:

$$\hat{\mu}_T = \frac{\bar{Z}}{[p+\mu_S(1-p)]}$$

with variance equal to:

$$V(\hat{\mu}_T) = \frac{\mu_X^2}{n}\left[C_X^2 + C_{S^*}^*(p)(1+C_X^2)\right] \tag{1.6}$$

where

$$C_{S^*}^*(p) = \frac{p+(1-p)(\mu_S^2+\eta^2\sigma_S^2)}{[p+(1-p)\mu_S]^2} - 1 \quad \text{and} \quad \eta = \frac{a}{(a+b)}$$

Comparing $V(\hat{\mu}_T)$ and $V(\hat{\mu}_B)$ in (1.6) and (1.1), it is easy to observe that $V(\hat{\mu}_T)$ is always less than $V(\hat{\mu}_B)$.

Most recently, Bouza-Herrera et al. (2022) proposed an RRT design with two scrambling procedures. The respondent performs a Bernoulli experiment which decides which procedure to be used. The respondent then gives an answer without disclosing which procedure was used. The response is as follows:

$$Z = \begin{cases} X + A, & \text{with probability } p \\ X + AB, & \text{with probability } 1-p \end{cases}$$

where A and B are two independent scrambling variables, independent from the sensitive variable X. The estimator of μ_X is given by:

$$\hat{\mu}_{CB} = \bar{Z} - \mu_A(p+(1-p)\mu_B)$$

with variance equal to:

$$V(\hat{\mu}_{CB}) = \frac{\sigma_X^2 + \sigma_A^2(p+(1-p)\sigma_B^2)}{n} \tag{1.7}$$

where μ_A, μ_B, σ_A^2 and σ_B^2 are the means and variances of the scrambling variables A and B, respectively.
The authors extended their model to the case of stratified random sampling and obtained an unbiased estimator for μ_X.

In the next section, we introduce our proposed model for the estimation of the mean of a sensitive quantitative random variable in case of simple random sampling. It is shown, in section 3, that the proposed estimator is more efficient than all the previously discussed estimators. In section 4, the proposed model is extended to stratified random sampling, and the stratified estimator is shown to be more efficient than that of Ryu et al. (2005). In section 5, a simulation study is conducted using real data to assess the efficiency of all the competing estimators. The last section contains a summary of the results presented in the paper.

2. PROPOSED MODEL IN CASE OF SIMPLE RANDOM SAMPLING

The proposed model suggests a modification on Ryu et al. (2005) model using an alternative form of the scrambled variable, S^*, suggested by Tarray and Singh (2017). According to the proposed model, each respondent in a SRSWR of size n, from a population of size N, is provided with two randomization devices R_1 and R_2. Using R_1, the respondent either reports the true value of X, or goes to R_2, with probabilities p, and $1-p$, respectively. Based on the outcome of R_2, the respondent either reports X, or the scrambled response XS^*, with probabilities t and $1-t$, respectively. The respondent's response, Z, can be represented as follows:

$$Z = \begin{cases} X, & \text{with probability } p \\ \text{go to } R_2, & \text{with probability } 1-p \end{cases} \quad \begin{cases} X, & \text{with probability } t \\ XS^*, & \text{with probability } 1-t \end{cases}$$

where $S^* = \eta S + (1-\eta)\mu_S$, and $\eta \in (0,1)$.

At $\eta = 1$, the model reduces to that of Ryu et al. (2005). It should also be noted that at $t = 0$ the proposed model reduces to Tarray and Singh (2017) model, and at $t = 1$ it reduces to a direct response question.

It is clear that the expected value of S^*, and its variance are given by:

$$E(S^*) = \mu_S$$
$$V(S^*) = \eta^2 \sigma_S^2$$

It is easy to show for the responses, Z, that

$$E(Z) = \mu_X(p + (1-p)t + (1-p)(1-t)\mu_S) \tag{2.1}$$

$$\text{and} \quad E(Z^2) = (\mu_X^2 + \sigma_X^2)[p + (1-p)t + (1-p)(1-t)(\mu_S^2 + \eta^2\sigma_S^2)] \tag{2.2}$$

Therefore, an unbiased estimator for the mean of the sensitive random variable, μ_X, is given by:

$$\hat{\mu}_R^* = \frac{\bar{Z}}{p + (1-p)t + (1-p)(1-t)\mu_S} \tag{2.3}$$

Theorem 1: $\hat{\mu}_R^*$ is an unbiased estimator for μ_X with variance equal to:

$$V(\hat{\mu}_R^*) = \frac{\mu_X^2}{n}[C_X^2 + C_{S^*}^*(p,t)(1 + C_X^2)] \tag{2.4}$$

where

$$C_{S^*}^*(p,t) = \frac{p + (1-p)t + (1-p)(1-t)(\mu_S^2 + \eta^2\sigma_S^2)}{[p + (1-p)t + (1-p)(1-t)\mu_S]^2} - 1 \tag{2.5}$$

Proof:

The unbiasedness of $\hat{\mu}_R^*$ follows by taking the expected values of both sides of equation 2.3 and

substituting for $E(Z)$ from equation 2.1.

$$\begin{aligned} E(\hat{\mu}_R^*) &= \frac{E(Z)}{p+(1-p)t+(1-p)(1-t)\mu_S} \\ &= \frac{\mu_X[p+(1-p)t+(1-p)(1-t)\mu_S]}{p+(1-p)t+(1-p)(1-t)\mu_S} \\ &= \mu_X \end{aligned}$$

$$\begin{aligned} V(\hat{\mu}_R^*) &= \frac{V(Z)}{n[p+(1-p)t+(1-p)(1-t)\mu_S]^2} \\ &= \frac{E(Z^2)-E^2(Z)}{n[p+(1-p)t+(1-p)(1-t)\mu_S]^2} \end{aligned} \tag{2.6}$$

substituting 2.1 and 2.2 into 2.6 gives:

$$\begin{aligned} V(\hat{\mu}_R^*) &= \frac{(\mu_X^2+\sigma_X^2)[p+(1-p)t+(1-p)(1-t)(\mu_S^2+\eta^2\sigma_S^2)]}{n[p+(1-p)t+(1-p)(1-t)\mu_S]^2} - \frac{\mu_X^2}{n} \\ &= \frac{\mu_X^2(1+C_X^2)}{n}(C_{S^*}^*(p,t)+1) - \frac{\mu_X^2}{n} \\ &= \frac{\mu_X^2}{n}\Big[C_X^2 + C_{S^*}^*(p,t)(1+C_X^2)\Big]. \end{aligned}$$

where $C_{S^*}^*(p,t)$ is as defined in 2.5.

3. EFFICIENCY COMPARISON

In what follows, the efficiency of our proposed estimator is investigated relative to that of Ryu et al. (2005), Tarray and Singh (2017) and Bouza-Herrera et al. (2022) estimators.

3.1. Efficiency comparison with Ryu et al. (2005) estimator

The efficiency of the proposed estimator $\hat{\mu}_R^*$ relative to Ryu et al. (2005) estimator, $\hat{\mu}_R$, is given by:

$$R.E = \frac{V(\hat{\mu}_R)}{V(\hat{\mu}_R^*)},$$

where $V(\hat{\mu}_R)$ and $V(\hat{\mu}_R^*)$ are as given in equations 1.3 and 2.4, respectively.
To compare the proposed estimator to that of Ryu et al. (2005), we have to set $\mu_S = 1$ as is done in their model. In this case, the variance of the proposed model, given by equation 2.4, reduces to the following:

$$V(\hat{\mu}_R^*) = \frac{\mu_X^2}{n}[C_X^2 + (1-p)(1-t)\eta^2\sigma_S^2(1+C_X^2)]$$

Consequently,

$$R.E = \frac{V(\hat{\mu}_R)}{V(\hat{\mu}_R^*)} = \frac{C_X^2 + (1-p)(1-t)\sigma_S^2(1+C_X^2)}{C_X^2 + (1-p)(1-t)\eta^2\sigma_S^2(1+C_X^2)}$$

It is clear that the relative efficiency of the proposed estimator to that of Ryu et al. (2005) will always be greater than 1, for all values of $\eta \in [0, 1)$, and the lower the value of η, the higher the relative efficiency. Knowing that Ryu et al. (2005) already proved that their estimator is more efficient than Greenberg et al. (1971) and Gupta et al. (2002) estimators, then so is the proposed estimator.

3.2. Efficiency comparison with Tarray and Singh (2017) estimator

As noted before, the proposed model reduces to the Tarray and Singh (2017) model at $t = 0$, while at $t = 1$ it reduces to a direct response question. Therefore, to compare the efficiency of the proposed estimator to that of Tarray and Singh (2017) it is sufficient to prove that $V(\hat{\mu}_R^*)$ is a decreasing function in t for $t \in (0, 1)$. To do so we show that the first derivative of $V(\hat{\mu}_R^*)$ with respect to t is negative for all $t \in (0, 1)$.

Theorem 2: The proposed estimator is more efficient than that of Tarray and Singh (2017) provided that μ_S satisfies the following condition:

$$0 \leq \mu_S < \frac{2E(S^{*2})}{1 + E(S^{*2})} \tag{3.1}$$

Proof: By looking at $V(\hat{\mu}_R^*)$ in 2.4, we can see that the first derivative of $V(\hat{\mu}_R^*)$ with respect to t reduces to the first derivative of $C_{S^*}^*(p, t)$, given in equation 2.5, with respect to t.
Let $a = E(S^{*2}) = \eta^2\sigma_S^2 + \mu_S^2$

$$\begin{aligned}
&\frac{\delta C_{S^*}^*(p,t)}{\delta t} \\
&= \frac{[p + (1-p)t + (1-p)(1-t)\mu_S]^2[(1-p) - a(1-p)]}{[p + (1-p)t + (1-p)(1-t)\mu_S]^4} \\
&\quad - \frac{[p + (1-p)t + (1-p)(1-t)a]\Big[2(p + (1-p)t + (1-p)(1-t)\mu_S)((1-p) - (1-p)\mu_S)\Big]}{[p + (1-p)t + (1-p)(1-t)\mu_S]^4} \\
&= \frac{[p + (1-p)t + (1-p)(1-t)\mu_S][(1-p) - a(1-p)]}{[p + (1-p)t + (1-p)(1-t)\mu_S]^3} \\
&\quad - \frac{[p + (1-p)t + (1-p)(1-t)a]\Big[2((1-p) - (1-p)\mu_S)\Big]}{[p + (1-p)t + (1-p)(1-t)\mu_S]^3} \\
&= \frac{(1-p)\Big[(1-a)[p + (1-p)t + (1-p)(1-t)\mu_S] - 2(1-\mu_S)[p + (1-p)t + (1-p)(1-t)a]\Big]}{[p + (1-p)t + (1-p)(1-t)\mu_S]^3} \\
&\equiv \frac{N(t)}{D(t)}
\end{aligned}$$

Assuming that $\mu_S > 0$, then $D(t) > 0 \; \forall \; t \in (0,1)$. Therefore, we need $N(t)$ to be negative.

$$\begin{aligned}
N(t) &= (1-p)\Big[(1-a)(1-p)t - (1-a)(1-p)\mu_S t - 2(1-\mu_S)(1-p)t + 2(1-\mu_S)(1-p)at \\
&\quad + (1-a)p + (1-a)(1-p)\mu_S - 2(1-\mu_S)p - 2(1-\mu_S)(1-p)a\Big] \\
&= (1-p)\Big[(1-a)(1-p)(1-\mu_S)t - 2(1-\mu_S)(1-p)(1-a)t + (1-a)p \\
&\quad - (1-a)\mu_S p - 2(1-\mu_S)p + 2(1-\mu_S)ap + (1-a)\mu_S - 2(1-\mu_S)a\Big] \\
&= (1-p)\Big[-(1-a)(1-p)(1-\mu_S)t + \mu_S - a\mu_S - 2a \\
&\quad + 2\mu_S a + (1 - a - \mu_S + a\mu_S - 2 + 2\mu_S + 2a - 2\mu_S a)p\Big] \\
&= (1-p)\Big[-(1-a)(1-p)(1-\mu_S)t + (a + \mu_S - \mu_S a - 1)p + \mu_S(1+a) - 2a\Big] \\
&= (1-p)\Big[-(1-a)(1-p)(1-\mu_S)t - (1-a)(1-\mu_S)p - [2a - \mu_S(1+a)]\Big] \\
&= (1-p)\Big[-(1-a)(1-\mu_S)[(1-p)t + p] - [2a - \mu_S(1+a)]\Big]
\end{aligned}$$

For the derivative to be negative, we need $N(t) < 0$. We have two terms, for the second term to be negative, we need $2a > \mu_S(1+a)$, which means that we need $\mu_S < \frac{2a}{1+a}$. Meanwhile, the first term attains its greatest value at $t = 1$, where $p + t(1-p) = 1$. So,

$$\begin{aligned}
N(t) &< (1-p)\Big[-(1-a)(1-\mu_S) - [2a - \mu_S(1+a)]\Big] \\
&< (1-p)[-(1+a) + 2\mu_S]
\end{aligned}$$

Therefore, μ_S should be less than $(1+a)/2$, however this term is larger than $2a/(1+a)$.

The following theorem gives a good candidate for the probability distribution of the scrambling variable, S, that satisfies the condition given in Equation 3.1.

Theorem 3: The exponential distribution, with mean $1/\lambda$, where

$$1 + \eta^2 - \eta\sqrt{1+\eta^2} < \lambda \leq 1 + \eta^2 + \eta\sqrt{1+\eta^2}, \tag{3.2}$$

satisfies the condition in Equation 3.1 for S.

Proof: First we need $\mu_S < \frac{2E(S^{*2})}{1+E(S^*)^2}$, then under the exponential distribution:

$$\begin{aligned}
&\frac{1}{\lambda} < \frac{2(\frac{1}{\lambda^2} + \frac{\eta^2}{\lambda^2})}{1 + (\frac{1}{\lambda^2} + \frac{\eta^2}{\lambda^2})} \\
&\lambda^2 - 2(1+\eta^2)\lambda + (1+\eta^2) < 0
\end{aligned}$$

Solving the corresponding quadratic equation, we get:

$$\lambda = (1+\eta^2) \pm \sqrt{\eta^2(1+\eta^2)}$$

Under the condition in 3.2, the proposed estimator is more efficient than that of Tarray and Singh (2017). Moreover, Tarray and Singh (2017) showed that their estimator is more efficient than that of Bar-Lev et al. (2004), and Bar-Lev et al. (2004) showed that their estimator is more efficient than

Eichhorn and Hayre (1983) estimator. Therefore, the proposed estimator is more efficient than that of Eichhorn and Hayre (1983), Bar-Lev et al. (2004) and Tarray and Singh (2017) under condition 3.1.

3.3. Efficiency comparison with Bouza-Herrera et al. (2022) estimator

A theoretical comparison between the variances of the proposed estimator and Bouza-Herrera et al. (2022) estimator is hard due to both models having different parameters. Therefore, in this section, a simulation study is conducted to compare the two estimators. Investigation of the possible distributions of the parameters and the sensitive variable is necessary for the simulation study of relative efficiency. In the additive model of Bouza-Herrera et al. (2022) the scrambling variable A should have similar range and variance to the sensitive variable X in order to adequately conceal the response. On the other hand, the multiplicative scrambling variables S and B should be centered around 1 regardless of the range of X. Thus, the Exp (1) distribution was used for both S and B.

Next, There are two possible forms of quantitative sensitive variables, either count data or continuous data. The Poisson (θ) and the exponential (λ) distributions can be used for count and continuous data, respectively. The following values of θ were considered in the comparison $\{2, (2), 10\}$. For the exponential distribution, we considered the mean $1/\lambda = \{20000, (10000), 60000\}$. Banerjee et al. (2006) discusses the use of the exponential distribution for modelling personal income in Australia. In all cases, we let the additive scrambling variable A assume the same distribution as X.

Finally, we restrict the range of possible values of p, t, and η in the model to $\{0.3, (0.1), 0.7\}$. The considered range takes into account the privacy of the respondents and the efficiency of the estimators.

Approximate relative efficiency $\widehat{R.E.}(\hat{\mu}_R^*, \hat{\mu}_{CB})$ was calculated by generating 10,000 samples of size $n = 100$ for X, A, B, and S from the above mentioned distributions. In addition, two Bernoulli random variables were generated with probabilities p and t for the randomization process. The results for the first case where X is count data is summarized in Table 1. As we can see, the proposed estimator is more efficient than that of Bouza-Herrera et al. (2022) with minimum relative efficiency of 1.348 which occurs at $\theta = 2$, $p = \eta = 0.7$, and $t = 0.3$. Meanwhile, the maximum efficiency of 3.545 occurs at $\theta = 10$, $p = 0.5$, $t = 0.7$, and $\eta = 0.3$. This shows that we do not need to maximize the values of p or t to achieve much better efficiency than Bouza-Herrera et al. (2022) estimator.

Table 1: Summary of $\widehat{R.E.}(\hat{\mu}_R^*, \hat{\mu}_{CB})$ for $X \sim$ Poisson (θ)

Statistic	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
$\widehat{R.E.}(\hat{\mu}_R^*, \hat{\mu}_{CB})$	1.348	1.775	2.055	2.140	2.445	3.545

For the case where X is a continuous variable, Table 2 shows the results of the simulation. Once more we find that the proposed estimator is more efficient than that of Bouza-Herrera et al. (2022) under

the considered parameters. The minimum efficiency is 1.202 and occurs at $1/\lambda = 60,000$, $p = \eta = 0.7$, and $t = 0.7$. The maximum efficiency is 1.9 and occurs at $1/\lambda = 30000$, $p = \eta = 0.3$, and $t = 0.7$. Again this shows that much higher efficiency can be achieved even with low value of p. The simulation was done using R and the code is available upon request from the corresponding author.

Table 2: Summary of $\widehat{R.E.}(\hat{\mu}_R^*, \hat{\mu}_{CB})$ for $X \sim$ Exp (λ)

Statistic	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
$\widehat{R.E.}(\hat{\mu}_R^*, \hat{\mu}_{CB})$	1.202	1.453	1.552	1.552	1.650	1.906

4. PROPOSED MODEL IN CASE OF STRATIFIED RANDOM SAMPLING

If the population, of size N, is divided into k non-overlapping strata, each of size N_h, where $h = 1, 2, ..., k$, and $\sum_{h=1}^{k} N_h = N$, then a simple random sample with replacement of size n_h, is selected from each stratum, and $\sum_{h=1}^{k} n_h = n$ is the total sample size. The selections in different strata are made independently. The respondent's response from the h^{th} stratum is given by:

$$Z_h = \begin{cases} X_h, & \text{with probability } p_h \\ \text{go to } R_2, & \text{with probability } 1 - p_h \end{cases} \quad \begin{cases} X_h, & \text{with probability } t_h \\ X_h S_h^*, & \text{with probability } 1 - t_h \end{cases}$$

where $S_h^* = \eta_h S_h + (1 - \eta_h) S_h$, $\eta_h \in (0, 1)$, $p_h \in (0, 1)$, $t_h \in (0, 1)$, and $h = 1, 2, ..., k$.
The expected value, and the variance of the response from the h^{th} stratum is given by:

$$\begin{aligned} E(Z_h) &= \mu_{X_h}(p_h + (1 - p_h)t_h + (1 - p_h)(1 - t_h)\mu_{S_h}) \\ E(Z_h^2) &= (\mu_{X_h}^2 + \sigma_{X_h}^2)[p_h + (1 - p_h)t_h + (1 - p_h)(1 - t_h)(\mu_{S_h}^2 + \eta_h^2 \sigma_{S_h}^2)] \\ V(Z_h) &= E(Z_h^2) - E^2(Z_h) \end{aligned}$$

Therefore an unbiased estimator for the mean of the sensitive variable in stratum h, μ_{X_h}, is given by:

$$\hat{\mu}_{X_h} = \frac{\bar{Z}_h}{p_h + (1 - p_h)t_h + (1 - p_h)(1 - t_h)\mu_{S_h}} \tag{4.1}$$

and its variance is given by:

$$V(\hat{\mu}_{X_h}) = \frac{\mu_{X_h}^2}{n_h}[C_{X_h}^2 + C_{S_h^*}^*(p_h, t_h)(1 + C_{X_h}^2)]$$

where

$$C_{S_h^*}^*(p_h, t_h) = \frac{p_h + (1 - p_h)t_h + (1 - p_h)(1 - t_h)(\mu_{S_h}^2 + \eta_h^2 \sigma_{S_h}^2)}{[p_h + (1 - p_h)t_h + (1 - p_h)(1 - t_h)\mu_{S_h}]^2} - 1$$

Then, an unbiased estimator for the population mean of the sensitive attribute, μ_X, is:

$$\hat{\mu}_X^s = \sum_{h=1}^{k} w_h \hat{\mu}_{X_h}$$

where $w_h = N_h/N$ is the weight of stratum h, such that $\sum_{h=1}^{k} w_h = 1$, and $\hat{\mu}_{X_h}$ is given in 4.1.
The variance of the proposed estimator is given by:

$$V(\hat{\mu}_X^s) = \sum_{h=1}^{k} w_h^2 V(\hat{\mu}_{X_h})$$

Using Neyman's optimum allocation minimizes the variance under the stratified random sample by taking $n_h = n\frac{w_h s_h}{\sum w_h s_h}$, where s_h is the standard deviation of the responses in the h^{th} stratum. On doing so, the variance of the estimator of the mean of the sensitive variable becomes as follows:

$$\begin{aligned} NeyV(\hat{\mu}_X^S) &= \sum_{h=1}^{k} \frac{w_h^2 V(Z_h)}{n[p_h + (1-p_h)t_h + (1-p_h)(1-t_h)\mu_{S_h}]} \\ &= \frac{1}{n}\sum_{h=1}^{k} \frac{w_h\sqrt{V(Z_h)}}{[p_h + (1-p_h)t_h + (1-p_h)(1-t_h)\mu_{S_h}]} \sum_{h=1}^{k} w_h\sqrt{V(Z_h)} \end{aligned}$$

Neyman's method requires prior estimates on the mean and variance of the sensitive variable which can be obtained from a pilot study or prior research.

4.1. Efficiency comparison with Ryu et al. (2005) estimator

As mentioned in section 1, Ryu et al. (2005) obtained their estimator at $\mu_{S_h} = 1$, for $h = 1, \ldots, k$. Therefore to compare the efficiency of our proposed estimator to that of Ryu et al. (2005), we set $\mu_{S_h} = 1$, for $h = 1, \ldots, k$. In this case, the variance of our proposed estimator reduces to the following:

$$V(\hat{\mu}_X^s) = \sum_{h=1}^{k} \frac{w_h^2 \mu_{X_h}^2}{n_h}\Big[C_{X_h}^2 + (1-p_h)(1-t_h)\eta_h^2\sigma_{S_h}^2(1+C_{X_h}^2)\Big]$$

Comparing the variance above with that in 1.4, we can see that ours is always smaller due to the fact that η_h^2 is less than 1, $h = 1, 2, ..., k$.
In addition, if we also let $\mu_{S_h} = 1$ under Neyman's optimal allocation, the minimum variance reduces to

$$minV(\hat{\mu}_X^S) = \frac{1}{n}\Big[\sum_{h=1}^{k} w_h\sqrt{V(Z_h)}\Big]^2$$

where

$$V(Z_h) = \mu_{X_h}^2\Big[C_{X_h}^2 + (1-p_h)(1-t_h)\eta_h^2\sigma_{S_h}^2(1+C_{X_h}^2)\Big]$$

Therefore the minimum variance of our proposed estimator is smaller than that of Ryu et al. (2005) in 1.5.

5. APPLICATION ON ALCOHOL EXPOSURE

In this section, to compare the efficiency of all the discussed estimators, we use data from The 2021 National Survey on Drug Use and Health (NSDUH). The NSDUH is annually conducted by the Substance Abuse and Mental Health Services Administration (SAMHSA) which is an agency within the U.S. Department of Health and Human Services (HHS). More on the survey can be found in Center

for Behavioral Health Statistics and Quality (2022). Our variable of interest is the age, in years, of first alcoholic beverage consumption, denoted X, for those who have consumed alcohol at any point in their lives. This variable is very important in assessing risk of alcohol dependence, see Sartor et al. (2009). While the question itself might not be sensitive, unusually young responses can be socially undesirable for the respondents.

The data of the 41,354 respondent in the sample are considered the population. To approximate the relative efficiency of the different estimators using the data, we drew 10,000 times a SRSWR, each of size $n = 100$. Similar to the procedure in subsection 3.3., for each sample, we generated 100 values for the scrambling variables S and B from Exp (1) distribution. For the additive variable A, we used the Poisson ($\theta = 17$) distribution. This distribution provides a reasonable range for A and its average is within the expected average of the study variable ($16 - 18$). For p, t, and η, we considered the reasonable range of $\{0.3, (0.1), 0.7\}$.

Table 3 presents a summary of the simulation results. It is clear that the proposed estimator is more efficient relative to the rest of the estimators under study. In particular, the percent efficiency of the proposed estimator is at least 377%, 239%, and 207% relative to Eichhorn and Hayre (1983), Bouza-Herrera et al. (2022), and Bar-Lev et al. (2004) estimators, respectively. Relative to Ryu et al. (2005) and Tarray and Singh (2017), respectively, the percent efficiency of the proposed estimator is at least 141% and 107%. Tarray and Singh (2017) can be considered as having the second best efficiency among all the considered estimators. Of course these numbers are subject to the data and the distributions used in the simulation.

Table 3: Summary of $\widehat{R.E.}(\hat{\mu}_R^*, \hat{\mu}_.)$ using 2021 NSDUH data

Statistic	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
$\widehat{R.E.}(\hat{\mu}_R^*, \hat{\mu}_E)$	3.777	7.418	10.042	10.204	12.619	17.364
$\widehat{R.E.}(\hat{\mu}_R^*, \hat{\mu}_B)$	2.075	3.572	4.579	4.890	5.773	10.476
$\widehat{R.E.}(\hat{\mu}_R^*, \hat{\mu}_R)$	1.409	1.829	2.315	2.603	3.050	5.374
$\widehat{R.E.}(\hat{\mu}_R^*, \hat{\mu}_T)$	1.066	1.275	1.443	1.536	1.681	3.260
$\widehat{R.E.}(\hat{\mu}_R^*, \hat{\mu}_{CB})$	2.392	3.903	4.936	5.328	6.416	10.962

6. SUMMARY AND CONCLUSION

This paper introduces an improved two-stage RR model for estimating the mean of a sensitive quantitative random variable. The proposed estimator is derived under both simple and stratified random sampling. In case of simple random sampling, the proposed estimator is more efficient compared to Ryu et al. (2005) estimator, and the relative efficiency increases as η decreases. Consequently, the proposed estimator is also more efficient than Gupta et al. (2002) estimator for ORRT.

As for the comparison with Tarray and Singh (2017) and subsequently Bar-Lev et al. (2004) and Eichhorn and Hayre (1983) estimators, the proposed estimator is more efficient under an achievable condition discussed in Theorem 2. The efficiency of the proposed estimator relative to Bouza-Herrera et al. (2022) is also explored numerically using simulation and the results show that the proposed estimator is more efficient under all the cases considered.

In reality most surveys have a more complex structure than SRSWR. In addition, the stratified random sample allows us to retain some information on the respondents that the randomization technique obstructs. Therefore, the proposed estimator is derived under stratified random sampling with replacement in general, and also under Neyman's optimum allocation. The proposed estimator is shown to be more efficient than that of Ryu et al. (2005) under both cases.

To explore the relative efficiency of the proposed estimator compared to its competitors, a simulation study is carried out using the NSDUH data on age of first alcoholic beverage consumption. The variable is considered a sensitive variable for those that have been exposed to alcohol from an abnormally young age. The results of the simulation show that the proposed estimator is more efficient than the rest of the estimators under different settings of the parameters. On average, the percent efficiency of the proposed estimator is 1020%, 489%, 260%, 153%, and 532% relative to Eichhorn and Hayre (1983), Bar-Lev et al. (2004), Ryu et al. (2005), Tarray and Singh (2017), and Bouza-Herrera et al. (2022) estimators, respectively.

To conclude, we recommend using the proposed RR model when the sensitive study variable is quantitative. The proposed estimator offers great gains in efficiency and its implementation should not be difficult. As Rueda et al. (2016) explores in their review, there are many software programs designed to generate random values from different distributions that can be used in RRT studies. These programs make it easier for the respondent to trust the randomization process and to ensure correct calculations of the final response. The proposed model can also be extended to other sampling schemes such as stratified sampling.

RECEIVED: OCTOBER, 2023.
REVISED: JANUARY, 2024.

REFERENCES

[1] BOUZA-HERRERA, C. N. and Pablo O. JUÁREZ-MORENO and Agustín SANTIAGO-MORENO and José M. SAUTTO-VALLEJO (2022): A two-stage scrambling procedure: Simple and stratified random sampling. an evaluation of covid19's data in mexico **Revista Investigación Operacional**, 43(4):421–430.

[2] BANERJEE, A., YAKOVENKO, V. M., and DI MATTEO, T. (2006): A study of the personal income distribution in australia **Physica A: Statistical Mechanics and its Applications**, 370(1):54–59 Econophysics Colloquium.

[3] BAR-LEV, S. K., BOBOVITCH, E., and BOUKAI, B. (2004): A note on randomized response models for quantitative data **Metrika**, 60(3):255–260.

[4] Center for Behavioral Health Statistics and Quality (2022): **2021 National Survey on Drug Use and Health Public Use File Codebook** Substance Abuse and Mental Health Services Administration, Rockville, MD.

[5] EICHHORN, B. H. and L. S. HAYRE (1983): Scrambled randomized response methods for obtaining sensitive quantitative data **Journal of Statistical Planning and Inference**, 7(4):307 – 316.

[6] GOODSTADT, M. S. and GRUSON, V. (1975): The randomized response technique: A test on drug use **Journal of the American Statistical Association**, 70(352):814–818.

[7] GREENBERG, B. G. and R. R. KUEBLER JR. and J. R. ABERNATHY and D. G. HORVITZ (1971): Application of the randomized response technique in obtaining quantitative data **Journal of the American Statistical Association**, 66(334):243–250.

[8] GUPTA, S., GUPTA, B., and SINGH, S. (2002): Estimation of sensitivity level of personal interview survey questions **Journal of Statistical Planning and Inference**, 100(2):239 – 247.

[9] KRUMPAL, I. (2013): Determinants of social desirability bias in sensitive surveys: a literature review **Quality & Quantity**, 47(4):2025–2047.

[10] LENSVELT-MULDERS, G. J. L. M., HOX, J. J., van der HEIJDEN, P. G. M., and MAAS, C. J. M. (2005): Meta-analysis of randomized response research: Thirty-five years of validation **Sociological Methods & Research**, 33(3):319–348.

[11] RUEDA, M., COBO, B., ARCOS, A., and ARNAB, R. (2016): Chapter 10 - software for randomized response techniques In CHAUDHURI, A., CHRISTOFIDES, T. C., and RAO, C., editors, **Data Gathering, Analysis and Protection of Privacy Through Randomized Response Techniques: Qualitative and Quantitative Human Traits**, volume 34 of **Handbook of Statistics**, pages 155–167. Elsevier.

[12] RYU, J., KIM, J., HEO, T., and PARK, C. (2005): On stratified randomized response sampling **Model Assisted Statistics and Applications**, 1(1):31–36.

[13] SARTOR, C. E., LYNSKEY, M. T., BUCHOLZ, K. K., MADDEN, P. A. F., MARTIN, N. G., and HEATH, A. C. (2009): Timing of first alcohol use and alcohol dependence: evidence of common genetic influences **Addiction**, 104(9):1512–1518.

[14] TARRAY, T. and SINGH, P. H. (2017): A simple way of improving the Bar-Lev, Bobovitch and Boukai randomized response model **Kuwait Journal of Science**, 44(4):83–90.

[15] WARNER, S. (1965): Randomized response: A survey technique for eliminating evasive answer bias **Journal of the American Statistical Association**, 60(309):63–69 PMID: 12261830.