STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **12**, Month 2024, pp 0–9.
Published online in International Academic Press (www.IAPress.org)

Improved estimation of the sensitive proportion using a new randomization technique and the Horvitz–Thompson type estimator

Hadi Farokhinia*, Rahim Chinipardaz, Golamali Parham

Department of Statistics, Shahid Chamran University of Ahvaz, Iran

Abstract Randomized response techniques efficiently collect data on sensitive subjects to protect individual privacy. This paper aims to introduce a new randomizing technique in the additive scrambled model so that privacy is well preserved and the estimator's efficiency for the sensitive population proportion is improved. Also, a Horvitz–Thompson type estimator is presented as an unbiased estimator of the sensitive proportion of the population, then convergence to the normal distribution for the Horvitz–Thompson type estimator is considered by the entropy of the inclusion indicators in the Poisson sampling. Eventually, using the new additive scrambled model, the ratio of taking addictive drugs is estimated among students of the University.

Keywords Additive scrambled model, Entropy, Horvitz–Thompson type estimator, Lindeberg condition, Poisson sampling, randomized response

AMS 2010 subject classifications 62D05, 62E20, 62P25

DOI: 10.19139/soic-2310-5070-1807

1. Introduction

The researchers are usually interested in sensitive subjects in survey sampling. Sensitive issues may be too personal or upsetting and some cases, illegal. So if the interviewer asks the interviewees directly the sensitive question, the estimation of the sensitive proportion would be biased largely due to non-responses or incorrect answers. Warner [26] proposed the randomized response technique as an alternative to the direct questioning method to increase cooperation among respondents and actual data in the survey. In this manner, using a randomizing technique, each respondent randomly selects to answer the sensitive question or its complimentary with known probability$p$ or $1-p$, in order. Suppose a random sample of size n is chosen from the finite population that the proportion of people with the sensitive characteristic is θ. So the interviewer would received "Yes" or "No" from n interviewees in the sample without understanding whether "Yes" or "No" is related to the sensitive question or complementary. Since randomizing technique for each respondent is a Bernoulli trial, it is expected that np and $n(1-p)$ times, the sensitive question, and its complementary will be selected respectively by respondents. Let X_j and Y_j be the actual and randomized responses for interviewee jth, respectively. Then the randomized response in Warner's model will be as follows.

$$Y_j = X_j R_j + (1-X_j)(1-R_j), j = 1, \ldots, n. \tag{1}$$

where X_j and R_j have Bernoulli distribution with probability θ and p, respectively. Therefore, the randomized response is distributed as Bernoulli with probability $\eta = \theta p + (1-\theta)(1-p)$, and an unbiased estimator of θ is

$$\hat{\theta}_W = \frac{\hat{Y} - (1-p)}{2p-1} \qquad for \qquad p \neq \frac{1}{2}$$

*Correspondence to: Hadi Farokhinia (Email: hadi-farokhinia@stu.scu.ac.ir). Department of Statistics, Shahid Chamran University of Ahvaz, Iran.

ISSN 2310-5070 (online) ISSN 2311-004X (print)

where $\bar{Y}$ is the sample mean of the randomized responses. Greenberg et al. [7] proposed the use of unrelated or non-sensitive questions to increase the protection of privacy. In this way, each individual selected for the interview answers "Yes" or "No" to the sensitive or non-sensitive question. Consequently, it is not known which question was answered by the interviewee. Another version of Warner's model and the unrelated-question model is based on repeated random trials. In Warner's model, Mangat and Singh[14] proposed to expand a two-stage randomized response technique and Singh and Mathur[23] studied the anonymous replication of the randomized trial in the unrelated-question randomized response model. Also, a new procedure based on replicating Warner's model or the unrelated-question model is introduced [1]. Several improvements in Warner and URL models have been studied by many authors, such as Christofides[4], Kim and Ward [12], Gupta et al [9], etc. Some of the more recently developed randomized response techniques (RRT) can be found in [8], [13], [15] and [17]. But these methods have trouble protecting privacy because it is possible to observe the result of the random trial (such as tossing a coin) by the interviewer. Therefore, the repeated randomized response technique is time-consuming, and the interviewee's answer to the sensitive or unrelated question may be exposed. Hence, it was investigated how to hide the respondent's response to a sensitive question by scrambling it through additive or multiplicative value from a known distribution [20], [21]. Therefore, the randomized response of the jth respondent in the additive scrambled model is

$$Y_j = Z_j + X_j, \qquad j = 1, \ldots, n. \tag{2}$$

where Z_j is the true response with mean φ_Z and variance σ_Z^2. Also, the random value X_j is from a known distribution with mean φ_X and variance σ_X^2.. So the estimation of $\varphi_Z = E(Z_j)$ and the variance of the estimator can be obtained as

$$\hat{\varphi}_Z = \bar{Y} - \varphi_X, var(\hat{\varphi}_Z) = \frac{\sigma_X^2 + \sigma_Z^2}{n}. \tag{3}$$

Such that the estimator of the variance is $\hat{var}(\hat{\varphi}_Z) = \frac{\hat{\sigma}_Y^2}{n}$, where $\hat{\sigma}_Y^2 = \frac{\sum_{j=1}^{n}(Y_j - \bar{Y})^2}{n}$. Accordingly, several suggestions have been created to improve the estimation of the population proportion of sensitive variables in the scrambled model [5], [6], [19], [22]. In the randomized response methods that are mentioned, the estimate of the expected value of the randomized response variable is first calculated because the values of the true responses to the sensitive question are not revealed. Consequently, an estimate of the expected value of the sensitive variable is obtained indirectly. Further, the random number generated from a known distribution may not well mask the actual response in the scrambled model. In this paper, an improved additive scrambled model is proposed with a new randomizing technique to solve these drawbacks. So, the interviewee scrambles the answer to the sensitive question with a random sample from a randomization bag containing responses "Yes" and "No", where the number of "Yes" in the random sample has a hypergeometric distribution. Hence privacy is preserved without the problems raised in the previous methods. Also, the researcher obtains the exact number of answers "Yes" to the sensitive question in the sample. In addition to estimating the sensitive proportion of the population with the mean of randomized responses, we intend to investigate the Horvitz–Thompson type estimator and its asymptotic distribution for the improved additive scrambled model in real-time sampling. Therefore, we will prove that the central limit theorem holds for the Horvitz–Thompson type estimator by putting conditions on the entropy of the inclusion indicators in the Poisson sampling. The Horvitz–Thompson type estimator was suggested to estimate the mean of the sensitive proportion in the item sum technique by Rueda et al. [18]. Also, the asymptotic normality of the Horvitz–Thompson estimator for simple random sampling without replacement is proved [2], [10].

The materials of this article are categorized into seven sections. In Section 2, the new additive scrambled model is explained. Section 3 is planned to study the Horvitz–Thompson type estimator, and Section 4 is for its asymptotic distribution. The simulation study is presented in Section 5. The application of the new additive scrambled model for estimating the proportion of taking addictive drugs at Shahid Chamran University of Ahvaz is performed in Section 6. Finally, in Section 7, the results are summarized.

2. The new additive scrambled model

Suppose a simple random sampling without replacement of size n is taken from the finite population $\mathcal{W} = \{1, \ldots, N\}$. Then the selected individuals scramble the answer to the sensitive question with a simple random sample from a randomization bag containing responses "Yes" and "No" to the sensitive question. The randomization bag can be considered a randomizing technique in which the interviewer prepares the $\mathcal{M}$ number of answers to the sensitive question for each selected individual such that the number of "Yes" is equal to β, and the number of "No" is equal $\mathcal{M} - \beta$. Let T_j be the random variable of the number of "Yes" responses selected in a random sample of size n_j from the randomization bag for the jth interviewee. Hence, the random variable T_j has the hypergeometric distribution with parameters $\mathcal{M}$, β, and n_j, and its expected value is $n_j \frac{\beta}{M}$; it is written as $T_j \sim Hyperg(M, \beta, n_j)$. If Z_j and Y_j are the true and the randomized response for interviewee jth in order, then the randomized response under the new additive scrambled model is obtained as follows.

$$Y_j = Z_j + T_j \tag{4}$$

where$Z_j$is a Bernoulli random variable with success probability θ because the population proportion of individuals with the sensitive feature is assumed to equal θ. From (4), we have

$$\phi = E(Y_j) = \theta + n_j \frac{\beta}{M} \tag{5}$$

Since $\bar{Y}$ is an unbiased estimator for $\theta + \frac{\frac{\beta}{\mathcal{M}} \sum_{j=1}^{N} n_j}{n}$, an unbiased estimator of θ is

$$\hat{\theta} = \bar{Y} - \bar{T}. \tag{6}$$

Note that $\bar{T} = \frac{\sum_{j=1}^{n} T_j}{n}$ can be calculated by subtracting the number of "Yes" answers remaining in the randomization bags from the total "Yes" answers prepared by the interviewer. Also, the variance of $\hat{\theta}$ is

$$var(\hat{\theta}) = var(\bar{Z}) = \frac{\sigma_Z^2}{n} = \frac{\theta(1-\theta)}{n}. \tag{7}$$

where $\sigma_Z^2 = var(Z_j)$ for$j = 1, \ldots, N$. Then an estimator of the variance is

$$\hat{var}(\hat{\theta}) = \frac{\hat{\sigma}_Z^2}{n} = \frac{\hat{\theta}(1-\hat{\theta})}{n}. \tag{8}$$

where $\hat{\sigma}_Z^2 = \frac{\sum_{j=1}^{n}(Z_j - \bar{Z})^2}{n}$. From (3), it can conclude that $var(\hat{\theta}) \leq var(\hat{\varphi}_Z)$, so $\hat{\theta}$ is more efficient than $\hat{\varphi}_Z$. The proof of equations (7) and (8) are given in Appendix A.

3. The Horvitz–Thompson type estimator (HT-type estimator)

Let $\mathcal{W}$ be a finite population consisting of N units such that the population units sequentially over time are observed. Also, it is possible that the sampling frame (set of all population units) is unavailable, and the population size would be determined after sampling. In this situation, the real-time sampling is suitable [16]. We use Poisson sampling in the article because it adapts to real-time sampling, so the sample size is a random variable with considerable variation. If the aim is to estimate the total of the variables, $\sum_{j=1}^{N} y_j$, the HT estimator for a sample with unequal probability can be used that is introduced by Horvitz and Thompson [11], i.e., $\hat{Y}_{HT} = \sum_{j=1}^{N} \frac{Y_j}{\pi_j} I_j$. Therefore, when the unit j enters the sample with inclusion probability π_j, the inclusion indicator I_j equals one and zero otherwise. Suppose the purpose is to estimate the mean of the variables, $\bar{Y}_N = \sum_{j=1}^{N} \frac{Y_j}{N}$, and the operators $\varepsilon_P(.)$ and $\nu_P(.)$ are expectation and variance under the Poisson sampling design. Hence, the HT-type estimator is

obtained as follows

$$\dot{\hat{Y}}_{HT} = \frac{\sum_{j=1}^{N} \frac{Y_j}{\pi_j} I_j}{N} \tag{9}$$

According to (4) and (9), we have

$$\dot{\hat{Y}}_{HT} = \frac{\sum_{j=1}^{N} \frac{Z_j}{\pi_j} I_j}{N} + \frac{\sum_{j=1}^{N} \frac{T_j}{\pi_j} I_j}{N} = \dot{\hat{Z}}_{HT} + \dot{\hat{T}}_{HT} \tag{10}$$

As a result, the unbiased estimator of θ is obtained using the HT- type estimator as follows

$$\dot{\hat{\theta}} = \dot{\hat{Y}}_{HT} - \dot{\hat{T}}_{HT} \tag{11}$$

Considering the expectation operator under the sampling design, we get

$$\varepsilon_P(\dot{\hat{\theta}}) = \bar{Y}_N - \bar{T}_N = \bar{Z}_N \tag{12}$$

where$\bar{Y}_N = \sum_{j=1}^{N} \frac{Y_j}{N}$ and $\bar{T}_N = \sum_{j=1}^{N} \frac{T_j}{N}$. Also, the variance of $\dot{\hat{\theta}}$ under the sampling design can be calculated as

$$\nu_P(\dot{\hat{\theta}}) = \nu_P(\dot{\hat{Z}}_{HT}) = \frac{\sum_{j=1}^{N} \frac{1-\pi_j}{\pi_j} Z_j^2}{N^2} + \frac{2\sum_{j=1}^{N-1} \sum_{i=j+1}^{N} \frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j} Z_j Z_i}{N^2} \tag{13}$$

In Poisson sampling, the covariance between inclusion indicator units i and j is zero, so the variance of $\dot{\hat{\theta}}$ under the sampling design is

$$\nu_P(\dot{\hat{Z}}_{HT}) = \frac{\sum_{j=1}^{N} \frac{1-\pi_j}{\pi_j} Z_j^2}{N^2}. \tag{14}$$

So the variance of $\dot{\hat{Z}}_{HT}$ can be obtained as follows

$$var(\dot{\hat{Z}}_{HT}) = \frac{\sum_{j=1}^{N} \frac{1-\pi_j}{\pi_j} \zeta_Z^2}{N^2} + \sum_{j=1}^{N} \frac{\sigma_Z^2}{\pi_j N^2} \tag{15}$$

where $\zeta = E(Z_j)$ for $j = 1, \ldots, N$. Therefore, an estimator of the variance of $\dot{\hat{Z}}_{HT}$ is

$$\hat{var}(\dot{\hat{\theta}}) = \dot{\hat{\theta}} \frac{\sum_{j=1}^{N} \frac{1-\pi_j}{\pi_j}}{N^2} + \frac{\dot{\hat{\theta}}(1-\dot{\hat{\theta}})}{N}. \tag{16}$$

The proofs of equations (12), (14), (15), and (16) are given in Appendix C.

4. Asymptotic Normality

In the real-time sampling, the Lindeberg condition gives the necessary and sufficient condition for the asymptotic normality of the HT-type estimator. Hence, we first define an index of the divergence of the sum entropy of the inclusion indicators as follows [24]

$$H_N(I) = -\sum_{j=1}^{N} \pi_j log(\pi_j). \tag{17}$$

Let $log(.)$ be the natural logarithm, and $0log(0) = 0$. Therefore, the sum entropy of the inclusion indicators is divergent, if $H_N(I)$ is divergent while N tends to infinity. Also, in the next theorem, it will be proved that the Lindeberg condition holds provide that $H_N(I)$ is divergent. Berger[2] studied a rate of convergence to normal

Table 1. Calculation of AE, AV, COP and ALCI when N is equal to 200, 400, and 1000.

The sensitive ratio		$\theta = 0.2$			$\theta = 0.5$			$\theta = 0.8$		
Models		$\dot{\hat{\theta}}$	$\hat{\theta}$	$\hat{\varphi}_Z$	$\dot{\hat{\theta}}$	$\hat{\theta}$	$\hat{\varphi}_Z$	$\dot{\hat{\theta}}$	$\hat{\theta}$	$\hat{\varphi}_Z$
$N = 200$	n	20.11	20	20	20.16	20	20	120.14	20	20
	$\sigma(n)$	4.39	0	0	4.38	0	0	4.36	0	0
	AE	0.21	0.22	0.26	0.51	0.51	0.57	0.83	0.82	0.85
	AV	0.008	0.007	0.05	0.02	0.011	0.05	0.03	0.007	0.03
	COP	0.94	0.95	0.79	0.93	0.96	0.8	0.91	0.92	0.81
	ALCI	0.35	0.36	0.77	0.55	0.42	0.75	0.69	0.33	0.74
$N = 400$	n	40.25	40	40	40.21	40	40	40.24	40	40
	$\sigma(n)$	6.15	0	0	6.18	0	0	6.16	0	0
	AE	0.197	0.193	0.23	0.497	0.489	0.51	0.809	0.806	0.77
	AV	0.004	0.003	0.01	0.01	0.006	0.29	0.018	0.003	0.02
	COP	0.91	0.90	0.92	0.93	0.93	0.79	0.92	0.91	0.76
	ALCI	0.25	0.23	0.59	0.39	0.30	0.57	0.49	0.24	0.55
$N = 1000$	n	100.02	100	100	100.01	100	100	100.02	100	100
	$\sigma(n)$	9.57	0	0	9.55	0	0	9.56	0	0
	AE	0.21	0.21	0.22	0.501	0.508	0.53	0.79	0.78	0.84
	AV	0.001	0.001	0.009	0.004	0.002	0.008	0.007	0.001	0.011
	COP	0.93	0.95	0.89	0.94	0.95	0.88	0.94	0.95	0.90
	ALCI	0.16	0.15	0.34	0.26	0.19	0.44	0.33	0.16	0.43

for the Horwitz-Thompson estimator by comparing the entropy of the sampling designs with the entropy of the rejective sampling. Under more comfortable conditions on the entropy of the inclusion indicators, we prove the Central Limit Theorem for the HT-type estimator in the Poisson sampling.

Theorem 1. The Horvitz–Thompson type estimator has an asymptotic normal distribution if $H_N(I) \to \infty$ as $N \to \infty$, i.e.

$$\frac{\dot{\hat{Z}}_{HT} - \bar{Z}}{\sqrt{\nu_P(\dot{\hat{Z}})}} \Rightarrow N(0,1) \tag{18}$$

Proof
where $\Rightarrow$ stands for convergence in distribution. The full proof of this theorem is in Appendix D. □

5. Simulation Examination

In the previous Sections, it was shown that the estimator of the additive scrambled model $\hat{\varphi}_Z$ and the estimators of the new additive scrambled model $\hat{\theta}$ and $\dot{\hat{\theta}}$ are unbiased for the sensitive ratio θ. Hence, to compare the efficiency of the estimators, we generated the population of individuals with the sensitive feature from the Bernoulli distribution for θ=0.2,0.5, and 0.8 using the simulation by R software where the population sizes are N=200,400, and 1000.

Then the average estimate (AE) of the parameter θ, the average estimate of variance (AV), coverage probability (COP), and the average length of the confidence interval (ALCI) are calculated in 10,000 repetitions by the estimators $\dot{\hat{\theta}}$, $\hat{\theta}$ and $\hat{\varphi}_Z$. The simulation results can be seen in Table 1. In the first step to getting $\dot{\hat{\theta}}$, Poisson sampling

was used with the inclusion probability $\pi_j = \frac{1}{10}$ for each unit j.In Poisson sampling, the sample size is a random variable as $n = \sum_{j=1}^{N} I_j$, so the expected value of the sample size is $\varepsilon_P(n) = \sum_{j=1}^{N} \pi_j$. The average sample sizes, $\hat{n}$, and their standard deviation,$\sigma(\hat{n})$, are shown in Table(**??**). Also, it was assumed that $T_j \sim Hyperg(20, 5, n_j)$ with expected value $\frac{1}{4}n_j$, where n_j be generated from the discrete uniform distribution on the integers $1, \ldots, 20$, for $j = 1, \ldots, n$. For simulation, the true responses,Z_j's, are generated from the Bernoulli distribution with success probability θ. Therefore, the randomized response Y_j is obtained from the sum of T_j and Z_j for each unit j entered in the sample. In the second step to estimate θ by the estimator $\hat{\theta}$, 10000 samples were chosen from the population according to simple random sampling without replacement with the constant sample sizes of $n = 20, 40, and 100$ for population sizes $N = 200, 400, and 1000$, respectively. Then Y_j's are calculated for each j unit as in the first step. The last step is related to the scrambled model estimator $\hat{\varphi}_Z$, assuming that $X_j \sim Hyperg(20, 5, 4)$. So the randomized response Y_j is obtained for each unit j entered in the simple random sample, and the simulation was repeated 10000 times for the sample sizes of $n = 20, 40, and 100$. Since sample sizes are constant, the standard deviation of the sample size is zero. The $(1-\alpha)\%$ confidence interval (CI) of θ represented by $CI = (L, U)$ is constructed based on Theorem 1. Also, using $\hat{\theta}$ and $\hat{\varphi}_Z$, and the central limit theorem, two other $(1-\alpha)\%$ CI's for θ can be constructed [25]. To evaluate the success of a confidence interval in catching the population parameter, COP is calculated by counting the number of times the actual parameter falls between the lower (L) and upper limits (U). Having a smaller length is a good feature for the CI because the true parameter falls within a smaller interval and the results are more accurate. So the length of the confidence interval is computed by subtracting the lower limit from the upper limit.COP and the ALCI are estimated, respectively, from CI's including parameter θ in all simulations using the following two formulas:

$$COP = \frac{\#(L \leq \theta \leq U)}{10000}, \tag{19}$$

$$ALCI = \frac{\sum_{j=1}^{10000}(U_j - L_j)}{10000} \tag{20}$$

where $\#(L \leq \theta \leq U)$ indicates the number of simulations in which the population parameter falls within the confidence interval.

From the results obtained in Table (**??**), it can be concluded that the new additive scrambled model is more efficient than the additive scrambled model because the estimates of the new additive scrambled model are closer to the actual value and have less AV. Also, the COP for estimators $\hat{\theta}$ and $\hat{\theta}$ obtains values closer to 95% than estimator $\hat{\varphi}_Z$ in the simulation, except when θ is equivalent to zero or 1. The larger COP values of the two proposed estimators compared to estimator $\hat{\varphi}_Z$ can be seen in Figure (1). According to Figure (2), it can be concluded that the ALCI made by the estimators $\hat{\theta}$ and $\hat{\theta}$ is smaller than the ALCI of estimator $\hat{\varphi}_Z$ for $\theta = 0.2, 0.5 and 0.8$. Also, with the increase in population size N, the ALCI decreases. However, the ALCI made by the estimators $\hat{\theta}$ and $\hat{\theta}$ decreases more than $\hat{\varphi}_Z$ when N increases.

6. A study example

Unfortunately, investigations show that many people take drugs without a doctor's prescription, which can have adverse or dangerous effects, especially if these drugs are addictive. The use of addictive drugs has been reported among young people, especially students. In this research, we intend to obtain the proportion of students from the Shahid Chamran University of Ahvaz who have used these drugs without a doctor's prescription. To take a random sample without replacement from the students, we must have access to the list of students as a sampling frame. Since there may be two problems with accessing the list of students, the first problem is getting permission from the University to access the list of students, which may be time-consuming or even impossible. The second problem occurs when the interviewee wants to remain anonymous. In order not to face these problems, Poisson sampling and the Horvitz–Thompson type estimator were used in this study. The data collection process was carried out on June 1, 2022, with the new additive scrambled model for 20 days. Therefore, the students were observed when

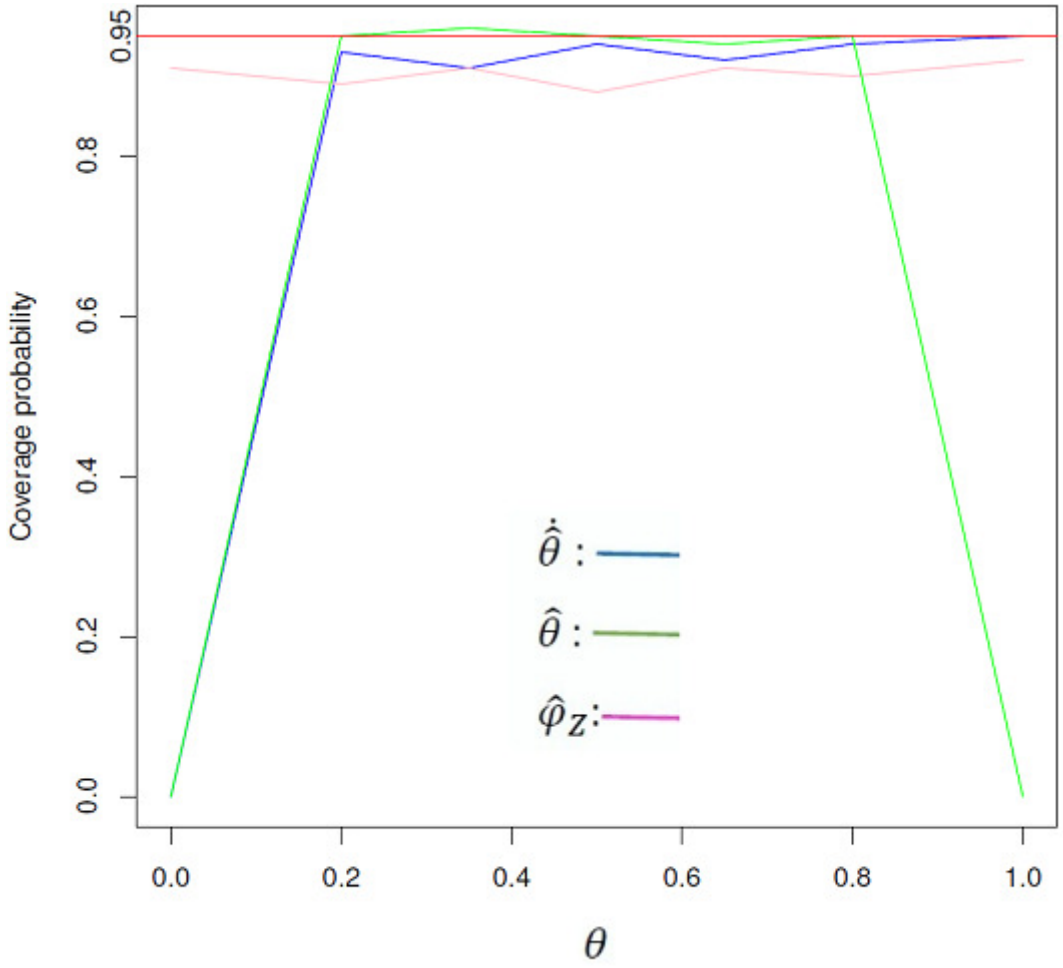

Figure 1. Calculated COP of the 95% confidence interval in 10,000 repetitions for by the estimators $\dot{\hat{\theta}}$, $\hat{\theta}$ and $\hat{\varphi}_Z$.

entering or exiting the dormitory. The Bernoulli trial was performed for each unit with the inclusion probability $\pi_j = \frac{1}{12}$ (e.g., observing a sum of 10 in throwing two dice). Since we estimated the number of students living or guests in the dormitory to be approximately 3000 people, the sampling was stopped when the population of size $N \cong 3000$ was observed. Accordingly, the expected value of the sample size is $\varepsilon_P(n) = sum_{j=1}^{3000}\pi_j \cong 250$. For the simulation study, it was assumed that $T_j \sim Hyperg(50, 10, n_j)$, and its expected value is $\frac{1}{5}n_j$, for $j = 1, \ldots, n$. Also, selected students were requested to perform the following steps secretly:
1. Choosing a random sample of arbitrary size from the randomization bag.
2. Answering the question: Have you ever taken an addictive drug without a doctor's permission?
The results are presented in Table (2).

Considering that Theorem1, the $100(1-\alpha)\%$ confidence interval (CI) for $\bar{Z}_N$ can be obtained. Also, according to Table (2), the difference in taking addictive drugs in males and females is significant at the 1% level.

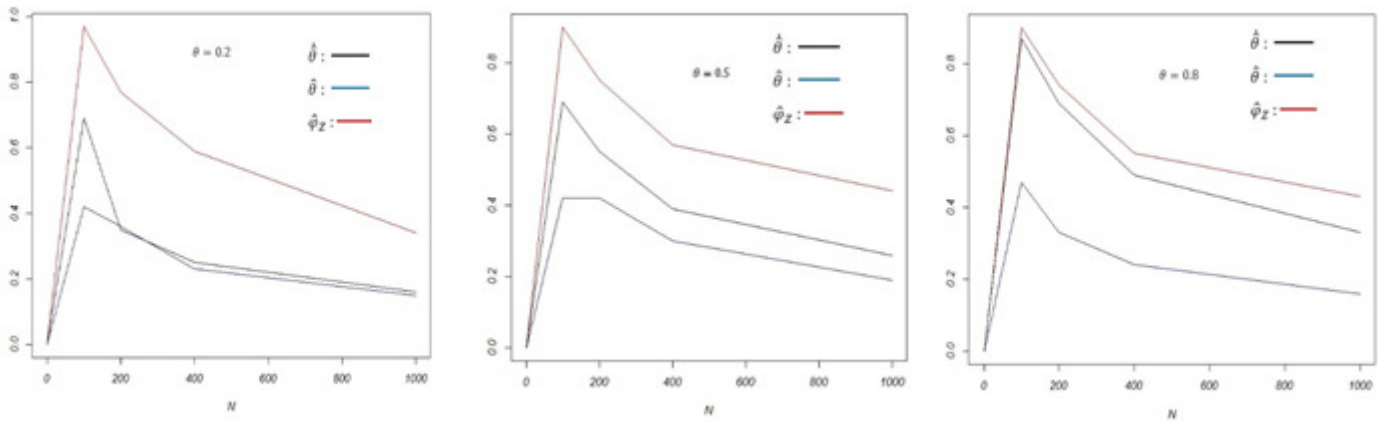

Figure 2. The average length of the confidence interval

Table 2. Summary of data and the estimation of taking addictive drugs among students.

Gender	Sample size	Population size	Number of "yes"	$\hat{\dot{\theta}}$	$\nu_P(\hat{\dot{Z}}_{HT})$	$\hat{var}(\hat{\dot{\theta}})$	95%CI
Female	151	1801	23	0.15	0.00007	0.00006	(0.133,0.166)
Male	102	1205	22	0.21	0.00016	0.00013	(0.185,0.234)
Total	253	3006	45	0.17	0.000054	0.000056	(0.166,0.177)

7. Conclusion

The new additive scrambled model introduced in this article can be a suitable alternative to the additive scrambled model. Therefore, privacy is preserved without repeating a random experiment (such as throwing a coin or dice) or combining the answer to a sensitive question with the value generated from a known random variable, which is time-consuming and can be disclosed. In this article, the variance of $\hat{\dot{\theta}}$ is calculated assuming that covariance between inclusion indicator units i and j is zero; otherwise, it can be the subject of further studies.

Appendix

A. The variance of $\hat{\theta}$ is obtained as follows.
Given that Z_j's and T_j's are independent, and $\hat{\theta} = \bar{Y} - \bar{T} = \bar{Z} + \bar{T} - \bar{T} = \bar{Z}$, so

$$var(\hat{\theta}) = var((\bar{Y} + \bar{T}) - \bar{T}) = var(\bar{Z} + \bar{T}) + var(\bar{T}) - 2cov(\bar{Z} + \bar{T}, \bar{T}) = var(\bar{Z}) = \frac{{\sigma^2}_Z}{n}$$

Since Z_j has the Bernoulli distribution, we have

$$\hat{\sigma}^2_Z = \frac{\sum_{j=1}^{n}(Z_j - \bar{Z})^2}{n} = \bar{Z} - \bar{Z}^2 = \hat{\theta} - \hat{\theta}^2.$$

B. Given that $Y_j = Z_j + T_j$ and the inclusion indicator I_j has Bernoulli distribution with parameter π_j, so we get

$$\varepsilon_P(\dot{\hat{\theta}}) = \varepsilon_P(\frac{\sum_{j=1}^N \frac{Y_j}{\pi_j} I_j}{N} - \frac{\sum_{j=1}^N \frac{T_j}{\pi_j} I_j}{N}) = \varepsilon_P(\frac{\sum_{j=1}^N \frac{Y_j}{\pi_j} I_j}{N}) - \varepsilon_P(\frac{\sum_{j=1}^N \frac{T_j}{\pi_j} I_j}{N}) = \tag{21}$$

$$\frac{\sum_{j=1}^N \frac{Y_j}{\pi_j} \varepsilon_P(I_j)}{N} - \frac{\sum_{j=1}^N \frac{T_j}{\pi_j} \varepsilon_P(I_j)}{N} = \bar{Y} - \bar{T} = \bar{Z} + \bar{T} - \bar{T} = \bar{Z}. \tag{22}$$

So, we have

$$\varepsilon_P(\dot{\hat{\theta}}) = \varepsilon_P(\bar{Z}) = \theta.$$

C. Since the inclusion indicator, I_j, has the Bernoulli distribution with parameter π_j, the probability of occurrence $\{I_j = 1\}$ is equal toπ_j. Also, the event of the presence of units i and j in the sample and its probability is shown by $\pi_{ij} = P\{I_i j = 1\}$, then the covariance is getting as follows.

$$\varepsilon_P(I_j) = \pi_j, \qquad cov_p(I_i, I_j) = \pi_{ij} - \pi_i \pi_j.$$

the variance of $\dot{\hat{\theta}}$, under the sampling design, can be represented as

$$\nu_P(\dot{\hat{\theta}}) = \nu_P(\dot{\hat{Z}}_{HT}) = \frac{\sum_{j=1}^N \sum_{i=1}^N \frac{Z_i Z_j}{\pi_i \pi_j} cov_P(I_i, I_j)}{N^2}$$

$$= \frac{\sum_{j=1}^N \frac{1-\pi_j}{\pi_j} Z_j^2}{N^2} + \frac{2\sum_{j=1}^{N-1} \sum_{i=j+1}^N \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} Z_j Z_i}{N^2}.$$

From (10), (11) we have

$$var(\dot{\hat{\theta}}) = E(var(\dot{\hat{\theta}}|P)) + var(E(\dot{\hat{\theta}}|P)) =$$

$$E(\frac{\sum_{j=1}^N \frac{1-\pi_j}{\pi_j} Z_j^2}{N^2}) + var(\frac{\sum_{j=1}^N Z_j}{N}) =$$

$$\frac{\sum_{j=1}^N \frac{1-\pi_j}{\pi_j} E(Z_j^2)}{N^2} + \frac{\sum_{j=1}^N var(Z_j)}{N} =$$

$$\frac{\theta \sum_{j=1}^N \frac{1-\pi_j}{\pi_j}}{N^2} + \frac{\theta(1-\theta)}{N}.$$

D. Let $R_j = (\frac{Z_j}{N\pi_j} I_j - \frac{Z_j}{N})$, so $\varepsilon_P({R_j}^2) = \frac{1-\pi_j}{N^2 \pi_j} Z_j^2$ and we get s

$$\delta^2{}_N = \sum_{j=1}^N \varepsilon_P({R_j}^2), \qquad S_N = \dot{\hat{Z}}_H T - \bar{Z}.$$

According to$H_N(I) \to \infty$ as $N \to \infty$, and the Lindeberg condition in Billingsley (1995, Theorem 27.2) , we have

$$\lim_N \frac{1}{\delta^2{}_N} \sum_{j:|R_j| \geq \epsilon \delta_N} \frac{1-\pi_j}{N^2 \pi_j} Z_j^2.$$

which holds because $\lim_N P(|\frac{Z_j}{N\pi_j} I_j - \frac{Z_j}{N}| \geq \epsilon \delta_N) = 0$.We can write, $|Z_j| \leq a$, where $1 \leq a$ for all j. So, it is obtained that

$$P(|\frac{Z_j}{N\pi_j} I_j - \frac{Z_j}{N}| \geq \epsilon \delta_N) \leq \frac{\frac{1-\pi_j}{\pi_j}}{\epsilon \sum_{j=1}^N \frac{1-\pi_j}{\pi_j}}.$$

Given that, $log(x) \leq x - 1$, and $-log(x) \leq \frac{1}{x} - 1$, we get

$$-\sum_{j=1}^{N} \pi_j log(\pi_j) \leq -\sum_{j=1}^{N} log(\pi_j) \leq \sum_{j=1}^{N} \frac{1-\pi_j}{\pi_j},$$

where $0 < \pi_j \leq 1$.

REFERENCES

1. S. M. R. Alavi, and M. Tajodini, *Maximum Likelihood Estimation of Sensitive Proportion Using Repeated Randomized Response Techniques*, Journal of Applied Statistics, vol. 43, pp. 563–571, 2016.
2. Y. G. Berger, *Rate of convergence to normal distribution for the Horvitz–Thompson estimator*, Journal of Statistical Planning and Inference, vol. 67, pp. 209–226, 1998.
3. H. Cardot, D. Degras, and E. Josserand, *Confidence bands for Horvitz-Thompson estimators using sampled noisy functional data*, Bernoulli, vol. 19, pp. 2067–2097, 2013.
4. T.C. Christofides, *Randomized response in stratified sampling*, Journal of Statistical Planning and Inference, vol. 128, pp.303–310, 2005.
5. G. Diana, and P. F. Perri, *Estimating a sensitive proportion through randomized response procedures based on auxiliary information*, Statistical Papers, vol. 50, pp.661–672, 2009.
6. G. Diana, and P. F. Perri, *New scrambled response models for estimating the mean of a sensitive quantitative character*, Journal of Applied Statistics, vol. 37, pp.1875–1890, 2010.
7. B. G. Greenberg, R. R. Kuebler, J. R. Abernathy, and D. G. Horvitz, *The Unrelated Question Randomized Response Model: Theoretical Framework*, Journal of the American Statistical Association, vol. 64, pp.520–539, 1996.
8. S. Gupta, J. Zhang, S. Khalil, and P. Sapra, *Mitigating lack of trust in quantitative randomized response technique models*, Communications in Statistics - Simulation and Computation, pp.1–9, 2022.
9. S. Gupta, J. Shabbir, and S. Sehra, *Mean and sensitivity estimation in optional randomized response models*, Journal of Statistical Planning and Inference, vol. 140, pp.2870–2874, 2010.
10. J. Hájek, *Limiting distributions in simple random sampling from a finite population*, Publ. Math. Inst. Hungar. Acad. Sci, vol. 5, pp. 361–374, 1960.
11. D. G. Horvitz, and D. J. Thompson, *A generalization of sampling without replacement from a finite universe*, Journal of the American Statistical Association, vol. 47, pp. 663–685, 1952.
12. J. M. Kim, and D. W. Warde, *A stratified Warner's randomized response model*, Journal of Statistical Planning and Inference, vol. 120, pp. 155–165, 2004.
13. M. Lovig, S. Khalil, S. Rahman, P. Sapra, and S. Gupta, *A mixture binary rrt model with a unified measure of privacy and efficiency*, Communications in Statistics - Simulation and Computation, vol. 47, no. 16, pp. 1–12, 2021.
14. N. S. Mangat, and R. Singh, *An alternative randomized response procedure*, Biometrika, vol. 77, pp. 439–442, 1990.
15. S. Mehta, and P. Aggarwal, *Bayesian estimation of sensitivity level and population proportion of a sensitive characteristic in a binary optional unrelated question rrt model*, Communications in Statistics - Theory and Methods, vol. 47, no. 16, pp. 4021–4028, 2018.
16. K. Meister, *On methods for real time sampling and distributions in sampling*, Ph.D. thesis, Department of Mathematical Statistics, Umeå University, Umeå, 2004.
17. G. Narjis, and J. Shabbir, *An efficient partial randomized response model for estimating a rare sensitive attribute using Poisson distribution*, Communications in Statistics - Theory and Methods, vol. 50, no. 1, pp. 1–17, 2021.
18. M. M. Rude, B. Coba, and P. F. Perri, *Advances in estimation by the item sum technique using auxiliary information in complex surveys*, Advances in Statistical Analysis, vol. 102, pp. 455–478, 2017.
19. V. R. Padmawar, and K. Vijayan, *Randomized response revisited*, Journal of Statistical Planning and Inference, vol. 90, pp. 293–304, 2000.
20. KH. Pollock, and Y. Bek, *A comparison of three randomized response models for quantitative data*, Journal of the American Statistical Association, vol. 71, pp. 884–886, 1976.
21. WK. Poole, *Estimation of the distribution function of a continuous type random variable through randomized response*, Journal of the American Statistical Association, vol. 69, pp. 1002–1005, 1974.
22. G. N. Singh, C. Singh, and A. Kumar, *A modified randomized device for estimation of population mean of quantitative sensitive variable with measure of privacy protection*, Communications in Statistics - Theory and Methods, vol. 51, pp. 1867–1890, 2019.
23. H. P. Singh, and N. Mathur, *Unknown repeated trials in the unrelated question randomized response model*, Biometrical Journal, vol. 46, pp. 375–378, 2004.
24. S. Treanta, *Optimization on the distribution of population densities and the arrangement of urban activities*, Statistics, Optimization and Information Computingl, vol. 6, no. 2, pp. 208–218, 2018.
25. G. L. Tian, M. L. Tang, Q. Wu, Y. Liu, *Poisson and negative binomial item count techniques for surveys with sensitive question*, Stat Methods Med Res, vol. 26, no. 2, pp. 931–947, 2017.
26. S. L. Warner, *Randomized response: A survey technique for eliminating evasive answer bias*, Journal of the American Statistical Association, vol. 60, pp. 63–69, 1965.