J. Nonlinear Sci. Appl., 14 (2021), 8–14

ISSN: 2008-1898

Journal of Nonlinear Sciences and Applications

Journal Homepage: www.isr-publications.com/jnsa

Adaptive cluster sampling randomized response model with electronically application

Mahmoud M. Mansour[a,b,*], Enayat M. Abd Elrazik[a,b]

[a]*Department of MIS, Yanbu, Taibah University, Saudi Arabia.*

[b]*Department of Statistics, Mathematics and Insurance, Benha University, Egypt.*

Abstract

It is difficult to estimate sensitive matters (e.g., addiction, drunken driving, and abortion) in population distributed over a large geographical area by conventional designs of sampling because of the social, political and security conditions that usually lead to their concentration in certain areas. An adaptive sampling scheme extending the initial sample by appropriate 'network' formations dependent on well-defined 'neighborhoods' brings about dramatic improvements exploiting the clustering tendencies of people by different places. On another hand to reduce non-response and response bias was needed to make people comfortable and to encourage truthful answers. So also we introduce a new technique to apply a randomized response by tablets, computers, mobile phones and etc. The relative efficiency and protection of the respondents of the proposed randomization device have been investigated. We illustrate our methods using real data from a survey study on the spread of the addiction phenomenon among high school students.

Keywords: Randomized response method, sensitive attribute, rare unrelated attribute, adaptive cluster sampling.

2020 MSC: 62D05.

1. Introduction

Researchers often use sample survey methodology to obtain information about a large aggregate or population in a small or large area by selecting and measuring a sample from that population. Data obtained from surveys are affected by two main sources of error. The first is the sampling error that results from taking a sample instead of enumerating the whole population. The second type of error is the non-sampling error. The main sources of non-sampling error in any survey are non-response bias and response bias. Non-response bias arises from subjects refusal to respond and response bias arises from giving incorrect responses. When open or direct surveys are about sensitive matters (e.g., addiction, drunken driving, and abortion), non-response bias and response bias become serious problems because people do not often wish to give correct information. In order to reduce non-response and response bias, a survey technique different from open or direct surveys was needed to make people comfortable and to encourage truthful answers.

*Corresponding author

Email addresses: mmmansour@taibahu.edu.sa (Mahmoud M. Mansour), ekhalilabdelgawad@taibahu.edu.sa (Enayat M. Abd Elrazik)

doi: 10.22436/jnsa.014.01.02

Received: 2019-12-03 *Revised:* 2020-01-27 *Accepted:* 2020-02-10

Warner [16] developed such an alternative survey technique that is called "randomized response" technique. Warner's randomized response survey technique is designed to eliminate evasive answer bias and keep the respondents' confidentiality. Randomized response is a technique used to collect sensitive information from individuals in such a way that survey interviewers and those who process the data do not know which of two alternative questions the respondent has answered. It allows respondents to respond to sensitive issues (such as criminal behavior, sexuality, addiction and abortion) while remaining confidentiality. To apply the Warner model, a simple random sample of n people is drawn with replacement from the population. Before interviewing each person in the sample, each interviewer is furnished with an identical spinner which points to either of two statements: (i) statement 1 I belong to the sensitive trait group, with probability P and (ii) statement 2 I do not belong to the sensitive trait group, with probability $1-P$. The interviewee spins the spinner which is unobserved by the interviewer. Without reporting the outcome of the spinner to the interviewer, the interviewee only answers "Yes" or "No" depending on the outcome of the randomization device.

The probability of a "Yes" answer is given by

$$\delta_w = P\pi + (1+P)(1-\pi).$$

Let $n\hat{\delta}_w$ be the number of "Yes" answers in a random sample of n respondents, the estimator $\hat{\pi}_w$ and it's variance $V(\hat{\pi}_w)$ of sensitive proportion π are respectively,

$$\hat{\pi}_w = \frac{\hat{\delta}_w - (1-P)}{(2P1-)} \quad \text{for} \quad P \neq 0.5, \quad V(\hat{\pi}_w) = \frac{\pi(1-\pi)}{n} + \frac{P(1-P)}{n(2P-1)^2}. \tag{1.1}$$

Greenberg et al. [7] introduced the Theoretical framework for unrelated question randomized response model suggested, that is a variation of Warner's (1965) RR model. The unrelated question RR model has one question that asks about a very sensitive trait and a second question that asks about an innocuous (or non-sensitive) trait. Many researcher's have modified Warner [16] by developing new and more efficient randomized response techniques which includes; see, for example, Bourke and Dalenius [5], Liu and Chow [10], Mangat and Singh [12], Mangat [11], Christofides [6], Kim and Warde [9], Kim and Elam [8], Odumade and Singh [13], Abdelfatah et al [4], Abdelfatah and Mazloum [1], Abdelfatah and Mazloum [2], andAbdelfatah and Mazloum [3].

In the literature, several authors have tried to compare various existing randomized response models at equal protection of the respondents, but to our knowledge, no one has made an attempt to determine whether existing randomized response models can be adjusted for greater cooperation and efficiency. Thus, in this paper, in Section 2, we will introduce a new simple technique to apply randomized response by tablets, computers, mobile phones and etc. In Section 3, the relative efficiency and protection of the respondents of the proposed randomization device have been investigated. In Section 4, we will introduce adaptive cluster sampling for proposed randomized response model. In Section 5, we will illustrate our methods using real data from a survey study on the spread of the addiction phenomenon among High school students.

2. The proposed models

In the adjusted Warner's and Greenberg's models, a simple random sample of n people is drawn with replacement from the population. Each respondent has two screens respectively. The first contains explanatory information about the phenomenon and asks the person to choose one of a set of numbers in his or her mind (eg selecting a number from the set of numbers $1,2,3,\ldots,10$). After selecting a number in his mind, he goes to the second screen and shows all available questions. The second screen contains three statements randomly arranged,

1. I belong to the sensitive trait group, with probability P_1;
2. I am a member of the innocuous trait group with probability P_2;

3. I do not belong to the sensitive trait group, with probability $(1 - P_1 - P_2)$.

After reading the selected question, select Yes or No at the bottom of the screen. The respondent should answer the question with a "yes" or "no" without reporting which question he or she has in order to protect the respondent's privacy. Under the assumption that these reported, yes and no are made truthfully (see Figure 1).

- π_m: is the true proportional of the population with the sensitive characteristics.
- $X_i = 1$ if the i-th sample element reports a "yes" answer.
- $X_i = 0$ if the i-th sample element reports a "no" answer.

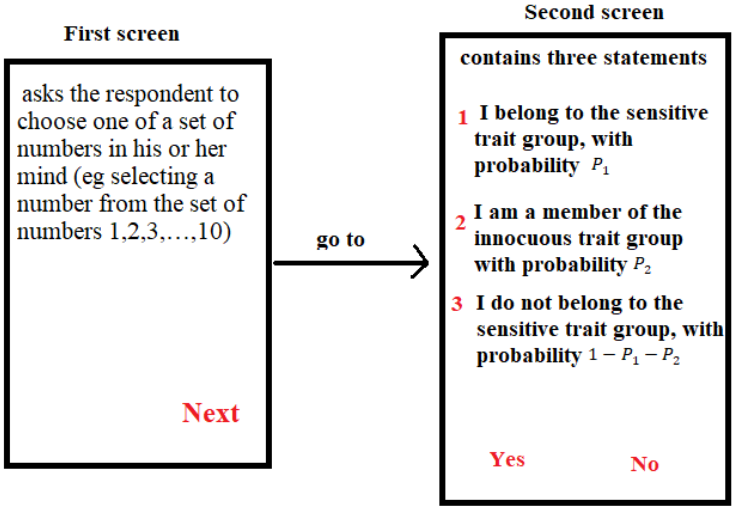

Figure 1: proposed randomized response model.

Then, the probability of a "Yes" answer is given by

$$\tau_m = P_1\pi_m + P_2\pi + (1 - P_1 - P_2)(1 - \pi_m), \quad \tau_m = \pi_m(2P_1 + P_2 - 1) + (1 - P_1),$$

where $(\pi = 1)$ is the proportion of "Yes" answers from the innocuous question, (e.g., the individual is asked to speak English and it is known in advance that all members of the sample speak English, the individual is asked about his nationality and it is known in advance that all members of the sample are of the same nationality) the moment as well as maximum likelihood estimate of π_m is easily shown to be

$$\hat{\pi}_m = \frac{\hat{\tau}_m - (1 - P_1)}{2P_1 + P_2 - 1}, \tag{2.1}$$

where $\hat{\tau}_m$ is the ML estimate of the proportion of "yes" answer in the sample. Since $n\hat{\tau}_m$ has a binomial distribution $B(n, \tau_m)$ the estimate is unbiased as follows.

Taking the expectation of (2.1) we get

$$E(\hat{\pi}_m) = \pi_m.$$

Then, $\hat{\pi}_m$ is an unbiased estimator of π_m with variance

$$V(\hat{\pi}_m) = \frac{\pi_m(1 - \pi_m)}{n} + \frac{P_1(1 - P_1)}{(2P_1 + P_2 - 1)}.$$

3. Efficiency comparison study

3.1. Relative efficiency

We compare the relative efficiencies (RE) between the modified model and Warner's randomized response as follows:

$$RE = \frac{V(\hat{\pi}_W)}{V(\hat{\pi}_m)},$$

let

$$G_W = V(\hat{\pi}_W) - \frac{P(1-P)}{n(2P-1)^2}, \quad F_m = V(\hat{\pi}_m) - \frac{P_1(1-P_1)}{(2P_1+P_2-1)^2},$$

We use the ratio G_W/F_mcompute the relative efficiency of, the proposed model based on the estimator $\hat{\pi}_m$, with respect to Warner's model, based on the estimator $\hat{\pi}_W$. It can be shown that $V_W > V_m$ for all π under condition $P_1 = P$.

3.2. Comparison for real survey data

To compare the confidence and protection of respondents in the traditional technique and proposed technique, a random sample of 272 respondents was drawn, 243 respondents selected the new approach in terms to make people comfortable and to encourage truthful answers. Thus, this study has shown the high confidence in the proposed technique to maintain privacy and thus encourage truthful answers.

4. The proposed randomized response model using adaptive cluster sampling

Adaptive cluster sampling (ACS) is an efficient design for uncommon and clustered populations (Thompson [14], Thompson and Seber [15]). ACS was presented for quadrat-based sampling, where the study area is typically divided into non-overlapping quadrats for sample choice. Depending on the situation, these are called "cells" or "secondary sampling units" (SSUs). In the first phase of the design, an initial sample is selected using one of the conventional designs, usually simple random sampling without replacement (SRSWOR). The term "conventional designs" (Thompson and Seber [15]) refers to designs in which the procedure for selecting the sample does not depend on any observation of the main variable, such as SRSWOR, stratified sampling and systematic sampling. If a rare event (a cell whose value is at least as large as the prespecified condition C) is found after the initial sample is obtained, then sampling continues in the neighborhood of that location with the hope of observing rarer events. The process of searching the neighborhood is continued until no rarer events are found. This design has been shown to be useful for estimating the parameters of highly clustered and rare populations.

It is difficult to estimate sensitive matters (e.g., addiction among high school students in one country) in population distributed over a large geographical area by conventional designs of sampling because of the social, political and security conditions that usually lead to their concentration in certain areas. An adaptive sampling scheme extending the initial sample by appropriate 'network' formations dependent on well-defined 'neighborhoods' brings about dramatic improvements exploiting the clustering tendencies of people by different places.

We will assume that the population is partitioned into N clusters, with unequal sizes $M_i, i = 1, 2, \ldots, N$. We will apply the adaptive cluster sampling procedure adopting the proposed randomized response strategy. In the first stage, the primary sampling units (PSUs) of n clusters are selected from the population of N clusters units, selected without replacement. To estimate the proportion π_i of units having a sensitive attribute in the i^{th} cluster, each respondent in the i^{th} cluster is provided with two screens, respectively.

- π_{mi}: is the true proportional of the population with the sensitive characteristics in the i^{th} cluster.
- $X_{im} = 1$ if the i-th respondent in the i^{th} cluster reports a "yes" answer.
- $X_{im} = 0$ if the i-th respondent in the i^{th} cluster reports a "no" answer.

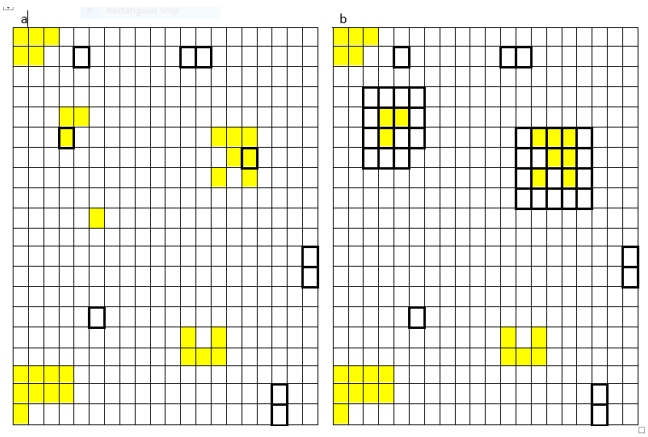

Figure 2: Adaptive cluster sampling to estimate the proportion of a sensitive attribute in a study region of 400 clusters. An initial random sample of 10 clusters is shown in (a), additional cluster in the neighborhood of that cluster is added to the sample, a cluster satisfies the condition if the proportion of interest $\hat{\pi}_k$ is greater than or equal c, that is, $C = \{\pi_k : \hat{\pi}_k \geqslant c\}$. The resulting sample is shown in (b).

The maximum likelihood estimate of π_{mi} in the i^{th} cluster is given by

$$\hat{\pi}_{mi} = \frac{\hat{\tau}_{mi} - (1 - P_1)}{2P_1 + P_2 - 1},$$

where $\hat{\tau}_{mi}$ is the ML estimate of the proportion of "yes" answer in the i^{th} cluster.

Let Ψ_k denote the network that includes unit k, and let ω_k, be the number of units in that network (note that a unit not fulfilling the criterion is considered as a network of size one.) With the adaptive, the inclusion probabilities are not known for all units included in the sample. Let $\hat{\pi}^*_{mk}$ represent the average of the $\hat{\pi}_{mk}$ in the network that includes the k^{th} unit of the initial sample, that is,

$$\hat{\pi}^*_{mk} = \frac{\sum_{j \in \Psi_k} \hat{\pi}_{mj}}{\omega_k}.$$

The estimator is

$$\hat{\pi}_m = \frac{1}{n} \sum_{k=1}^{n} \hat{\pi}^*_{mk}.$$

The variance of $\hat{\pi}_m$ is given by

$$V(\hat{\pi}_m) = \frac{N - n}{Nn} \sum_{k=1}^{n} \left[\frac{\hat{\pi}^*_{mk}(1 - \hat{\pi}^*_{mk})}{m_k} + \frac{P_1(1 - P_1)}{m_k(2P_1 + P_2 - 1)^2} \right],$$

if the initial sample is selected without replacement and

$$V(\hat{\pi}_m) = \frac{1}{n} \sum_{k=1}^{n} \left[\frac{\hat{\pi}^*_{mk}(1 - \hat{\pi}^*_{mk})}{m_k} + \frac{P_1(1 - P_1)}{m_k(2P_1 + P_2 - 1)^2} \right],$$

if the initial sample is selected without replacement.

5. Application

One of the highly sensitive matters is tackled and, thus, problems linked to refusals to respond or intentionally misleading replies are encountered. The study is aimed at getting better estimates of addiction among high school students by applying the proposed randomized response model.

This study is designed to estimate the proportions of the addiction for high school students, and the proportion of high school in the whole population of high school students in the government. The study population consists of 1726 classrooms distributed in adjacent residential neighborhoods within

the government and imagining them to be arranged in a circular way.

Since there is prior information on the proportion of addiction among teenagers, so we take $\pi_k = 0.08$ according to sample size, a sample of $n = 29$ classrooms was required. Previous field experience (see Warner [16]) has indicated that satisfactory results can be achieved with P equal $0.2(\pm 0.1)$ or $0.8(\pm 0.1)$. So we chose $P_1 = 0.7$ and $P_2 = 0.2$.

To estimate the proportion π_m of units having a sensitive attribute in the 29 clusters, electronic forms were sent to each respondent with two screens respectively. The first screen contains explanatory information about the model and how to maintain the privacy of the respondent and asks the respondent selecting a number from the set of numbers $1, 2, 3, \ldots, 10$ in his or her mind. After selecting a number in his or her mind, he or she goes to the second screen and shows all available questions, the second screen contains three statements randomly arranged (Figure 1)

1. I belong to the sensitive trait group, in seven questions;
2. I am a member of the innocuous trait group in two questions;
3. I do not belong to the sensitive trait group, in one question.

After reading the selected question, select Yes or No at the bottom of the screen. The respondent should answer the question with a "yes" or "no" without reporting which question he or she has in order to protect the respondent's privacy. Under the assumption that these reported, yes and no are made truthfully.

After getting the answers from each cluster we estimate $\hat{\pi}_{mk}$, whenever the proportion of addiction of a selected cluster satisfies a given criterion, additional cluster in the neighborhood of that cluster are added to the sample, a cluster satisfies the condition if the proportion of interest $\hat{\pi}_k$ is greater than or equal 0.10 that is, $C = \{\pi_k : \hat{\pi}_k \geqslant 0.10\}$. This condition is based on previous studies and the researcher's experience. Table 1 shows networks in the 29 clusters, the number of the respondent in each network and the $\hat{\pi}_{mk}$ and its variance in the network that includes the k^{th} unit of the initial sample.

Table 1: $\hat{\pi}_k^*$ and $V(\hat{\pi}_k^*)$ of the 29 clusters.

k	m_k	ω_k	$\hat{\pi}_k^*$	$V(\hat{\pi}_k^*)$
1	40	1	0.07	0.0035
2	35	1	0.02	0.0027
3	36	1	0.02	0.0026
4	36	1	0.01	0.0023
5	41	1	0.05	0.0030
6	45	1	0.03	0.0023
7	192	6	0.017	0.0011
8	38	1	0.08	0.0039
9	40	1	0.02	0.0024
10	40	1	0.07	0.0035
11	38	1	0.03	0.0027
12	42	1	0.04	0.0027
13	243	7	0.19	0.0009
14	48	1	0.04	0.0023
15	46	1	0.08	0.0032
16	50	1	0.06	0.0026
17	36	1	0.01	0.0024
18	136	4	0.14	0.0014
19	38	1	0.02	0.0025
20	37	1	0.01	0.0023
21	31	1	0.02	0.0031
22	41	1	0.03	0.0025
23	42	1	0.01	0.0020
24	191	6	0.15	0.0010
25	37	1	0.01	0.0023
26	39	1	0.01	0.0022
27	39	1	0.02	0.0024
28	164	5	0.14	0.0011
29	46	1	0.06	0.0069

From the previous results, we conclude the estimated proportion of addiction among high school students in the sampled population is $(\hat{\pi}_m) = 0.055$ with variance $V(\hat{\pi}_m) = 0.0038$.

References

[1] S. Abdelfatah, R. Mazloum, *Efficient estimation in a two-stage randomized response model*, Math. Popul. Stud., **22** (2015), 234–251. 1

[2] S. Abdelfatah, R. Mazloum, *An efficient two-stage randomized response model under stratified random sampling*, Math. Popul. Stud., **23** (2016), 222–238. 1

[3] S. Abdelfatah, R. Mazloum, *An improved cluster two-stage randomized response model*, Comm. Statist. Simulation Comput., **48** (2019), 58–72. 1

[4] S. Abdelfatah, R. Mazloum, S. Singh, *Efficient use of a two-stage randomized response procedure*, Braz. J. Probab. Stat., **27** (2013), 608–617. 1

[5] P. D. Bourke, T. Dalenius, *Some new ideas in the realm of randomized inquiries*, Int. statist. Review, **44** (1976), 219–221. 1

[6] T. C. Christofides, *A generalized randomized response technique*, Metrika, **57** (2003), 195–200. 1

[7] B. G. Greenberg, A.-L. A. Abul-Ela, W. R. Simmons, D. G. Horvitz, *The unrelated question randomized response model: Theoretical framework*, J. Amer. Statist. Assoc., **64** (1969), 520–539. 1

[8] J. M. Kim, M. E. Elam, *A stratified unrelated question randomized response model*, Statist. Papers, **48** (2007), 215–233. 1

[9] J. M. Kim, W. D. Warde, *A stratified Warner's randomized response model*, J. Statist. Plann. Inference, **120** (2004), 155–165. 1

[10] P. T. Liu, L. P. Chow, *The efficiency of the multiple trial randomized response technique*, Biometrics, **32** (1976), 607–618. 1

[11] N. S. Mangat, *An improved randomized response strategy*, J. Roy. Statist. Soc. Ser. B, **56** (1994), 93–95. 1

[12] N. S. Mangat, R. Singh, *An alternative randomized response procedure*, Biometrika, **77** (1990), 439–442. 1

[13] O. Odumade, S. Singh, *Efficient use of two decks of cards in randomized response sampling*, Comm. Statist. Theory Methods, **38** (2009), 439–446. 1

[14] S. K. Thompson, *Adaptive cluster sampling*, J. Amer. Statist. Assoc., **85** (1990), 1050–1059. 4

[15] J. N. Thompson, G. A. F. Seber, *Adaptive sampling*, John Wiley & Sons, New York, (1996). 4

[16] S. L. Warner, *Randomized response: A survey technique for eliminating evasive answer bias*, J. Am. Stat. Assoc., **60** (1965), 63–69. 1, 1, 5