A Two-Stage Sampling Model for the Estimation of Population Proportion and Cheating with Randomized Response and Direct Questioning

Evrim Oral, Husam I. Ardah
[1]LSUHSC School of Public Health, Biostatistics Program
2020 Gravier Street, New Orleans, LA 70112

Abstract

Randomized response methods (RRMs) are proposed in survey sampling as a solution to the problem of social desirability bias (SDB) when dealing with sensitive questions. RRMs reduce the SDB by providing privacy protection for respondents. However, their variances are inflated with respect to the direct questioning method (DQM); in other words, the RRMs provide unbiased estimators in exchange for less precision with respect to the DQM. The success of an RRM heavily depends on the assumption that the question under study is in fact sensitive. However, the question of interest may not be considered as really sensitive by some of the respondents, in which case using an RRM instead of the DQM inflates the variance of the estimates. In this study, we propose a two-stage sampling design where one can accurately estimate the prevalence of the sensitive characteristic under study without paying the price of the inflated variance by choosing between the proposed model and the DQM. With the proposed model one can also estimate the probability of cheating in the population.

Key Words: Randomized Response Methods, Warner Model, Social Desirability Bias, Mean Square Error.

1. Introduction

When administering surveys, researchers might be interested in asking questions that could be considered as being personal in nature by participants. If sensitive questions are asked directly, it is highly possible to get false responses in which participants report a more socially desirable answer instead of the true one. Participants might do this for many reasons: they might feel embarrassed, they might question the confidentiality of the survey or they might think they will get into trouble with the law. For example, suppose the question of interest is “Have you ever hit your children?” Then, a respondent is more likely to answer this question with a “no” even if the true answer is a “yes”, because although there is no federal law generally governing how parents must conduct themselves with regard to their children, hitting one’s own children is a sensitive moral topic for many parents.

Warner (1965) was the first researcher who built a model to counteract the problem of social desirability bias (SDB). He proposed a randomized response method (RRM) to

deal with stigmatizing questions. Basically, he suggested to use a spinner with its circle divided into two mutually exclusive areas as such as A and $\overline{A}$, with known probabilities θ and $1-\theta$, respectively $(0 \le \theta \le 1)$. Let A be the sensitive characteristic that we are interested in studying. In Warner's RRM, the area A corresponds to the statement "*I belong to group A*", and the area $\overline{A}$ corresponds to the statement "*I do not belong to group A"* (or equivalently "*I belong to group* $\overline{A}$"). Unobserved by the interviewer, the participant is asked to spin the spinner. If the spinner lands on *A*, the respondent has to answer the statement: "*I belong to group A*" with a "yes" or "no". If the spinner lands on $\overline{A}$, then the respondent has to answer the contrary statement "*I do not belong to group A"* with a "yes" or "no". An example to such statements above can be given as:

"*I am a drug user*" (i.e. "*I belong to group A*")
"*I am not a drug user*" (i.e. "*I do not belong to group A*").

Since the interviewer does not know where the spinner lands and just records a "yes" or "no" response without knowing which statement the participant is answering, this method gives privacy protection to the respondents by reducing privacy concerns and thus reduces the number of refusals or evasive answers.

Since the publication of Warner's model in 1965, a great deal of research has been done on RRMs such as Greenberg et al. (1969, 1971), Gupta (2001), Gupta et al. (2002, 2004, 2007, 2013), Yu et al. (2015) and many more. However, despite all the advances in the area, RRMs are known to have some limitations; see Chaudhuri and Mukerjee (1988), Chaudhuri (2011) and Tian and Tang (2014) for a comprehensive review on RRMs.

A common feature of all RRMs is that they lead to more accurate estimates of the sensitive characteristic of interest compared to the direct-questioning method (DQM); however, while providing estimates with smaller biases, they all inflate the variance of the estimators which is due to the randomization process. In fact, there is a direct relationship between the privacy level and the variance of the estimates from RRMs: as the level of privacy increases, the variance of the estimates also increases (Chaudhuri and Mukerjee, 1988). As a result, the number of participants surveyed using an RRM has to be larger than the number of participants surveyed with DQM in order to get an estimate of the true mean response with the same confidence margin. More importantly, the sensitive question of interest may not be considered as truly sensitive by most of the respondents in particular populations, in which case using an RRM instead of the DQM inflates the variance of the estimates unnecessarily. As an example, in a research study where the surveyed population is the patients in an HIV clinic, questions regarding with "HIV status" might not be considered as sensitive at all by the respondents.

When a question's sensitivity level is low in the population of interest, using an RRM instead of DQM inflates the variance of the estimates unnecessarily, thus, in this study we propose a two-stage sampling model for a binary response where one is able to choose between the RRM and DQM. With the proposed model, one can both estimate the prevalence of the sensitive characteristic under study and also the probability of cheating in the population simultaneously, and thus, the proposed model enables one to obtain more accurate estimates by avoiding the unnecessary penalty if the question is not in fact highly sensitive. For simplicity we consider the well-known Warner's RRM in the proposed model for the randomization process.

2. Warner's RRM

In Warner's model, which is the first and simplest RRM, the respondents are provided a randomization device by which they randomly chose one of the two questions "*Do you belong to group A*?" or "*Do you belong to group* $\overline{A}$?" with known probabilities θ and $1-\theta$ respectively $(0 \le \theta \le 1)$, and reply truthfully as "*yes*" or "*no*" to the question chosen. If we denote the unknown proportion of population members belong to group A with π_A where $0 \le \pi_A \le 1$ and let

$$\lambda = \begin{cases} 1, & \text{if the respondent says } "yes" \\ 0, & \text{if the respondent says } "no" \end{cases}$$

the probability of getting a *"yes"* response is $\pi_\lambda = \pi_A\theta + (1-\pi_A)(1-\theta)$. An unbiased estimator of π_λ $(0 \le \pi_\lambda \le 1)$ from the sample is $\hat{\pi}_\lambda = n_1/n$, where $n = n_0 + n_1$, n_0 and n_1 are the numbers of *"no"* and *"yes"* responses respectively. Then, the unbiased estimator of π_A follows as (Warner, 1967)

$$\hat{\pi}_{AW} = (\hat{\pi}_\lambda + \theta - 1)/(2\theta - 1), \tag{1}$$

where $\theta \neq 0.5$, and the variance of this estimator can be derived as

$$Var(\hat{\pi}_{AW}) = \frac{1}{n}\left[\pi_A(1-\pi_A) + \frac{\theta(1-\theta)}{(2\theta-1)^2}\right], \tag{2}$$

which can be estimated by replacing π_A by its unbiased estimate $\hat{\pi}_{AW}$. Realize that taking θ=1 or θ=0 in Warner's model corresponds to the DQM. Also, realize that the second term in (2) corresponds to the excess variance resulting from using Warner's model instead of using DQM. Thus, although considering very high ($\cong 1$) or very low ($\cong 0$) θ values would decrease the total variance in (2), considering such θ values will violate privacy protection and bring in SDB. Particularly, considering θ=0.5 would give the respondents maximum privacy protection; however, it would also make π_{AW} non-estimable and blow up the variance given in equation (2). In fact, it is a well-known fact that efficiency and privacy protection are generally in conflict within the context of RRMs; see Chaudhuri and Mukerjee (1988) for details. As a result, although the RRMs reduce the response bias in surveys with sensitive questions, there is a price paid for using them instead of DQM, which is the inflated variance.

Thus we propose a procedure below that will allow one to choose between an RRM and DQM, by estimating the cheating proportion in the target population for a sensitive question. With the proposed two-stage design, we can classify a question as being either sensitive or not-sensitive, and use the estimates from RRM only if the question is categorized as sensitive.

2. Proposed Model for Binary Data

In the proposed model, we assume that the characteristic under study is socially unacceptable in nature. We define "cheating" as not telling the truth in the DQM and we

assume that if a respondent cheats, he/she always answers in favor of the least stigmatizing category. This assumption is called Self-Protective no saying (Hout et al., 2010) when the characteristic under study is socially unacceptable in nature.

Assume that we apply a survey via DQM and the unknown proportion of population members belong to group A is denoted by π_A where $0 \le \pi_A \le 1$. Let

$$X_i = \begin{cases} 1, & \text{if the respondent belongs to Group } A \\ 0, & \text{if the respondent belongs to Group } \overline{A} \end{cases}$$

with probabilities $P(X_i = 1) = \pi_A$ and $P(X_i = 0) = 1 - \pi_A$, respectively, where $1 \le i \le n$, and let

$$X_{D_i} = \begin{cases} 1, & \text{if the respondent says } "yes" \text{ in the DQ stage} \\ 0, & \text{if the respondent says } "no" \text{ in the DQ stage} \end{cases}$$

with probabilities $P(X_{D_i} = 1) = \pi_D$ and $P(X_{D_i} = 0) = 1 - \pi_D$, respectively. Similarly, let

$$T_i = \begin{cases} 1, & \text{if the respondent answers truthfully in the DQ stage} \\ 0, & \text{if the respondent answers untruthfully in the DQ stage} \end{cases}$$

with probabilities $P(T_i = 1) = \pi_T$ and $P(T_i = 0) = 1 - \pi_T$, respectively. We can write the joint probability mass function (pmf) of X_D and T as given below

$X_{Di} \mid T_i$	0	1	pmf
0	$1-\pi_T$	$1-\pi_A$	$1-\pi_D$
1	0	π_D	π_D
pmf	$1-\pi_T$	π_T	1

Realize that if we define a new random variable for $0 \le j \le 2$ such that

$$Y_{ij} = \begin{cases} 0, & \text{if } (X_{D_i}, T_i) = (0,0) \text{ with probability } (1-\pi_T) \\ 1, & \text{if } (X_{D_i}, T_i) = (0,1) \text{ with probability } (1-\pi_A) \\ 2, & \text{if } (X_{D_i}, T_i) = (1,1) \text{ with probability } \pi_D \end{cases}$$

then the joint pmf $P(X_{D_i}, T_i) = P(Y_{i0} = y_{i0}, Y_{i1} = y_{i1}, Y_{i2} = y_{i2})$ can be considered as a multinomial distribution with three categories:

$$P(Y_{i0} = y_{i0}, Y_{i1} = y_{i1}, Y_{i2} = y_{i2}) = (1-\pi_T)^{y_{i0}} (1-\pi_A)^{y_{i1}} \pi_D^{\,y_{i2}}$$

when $\sum_{j=0}^{2} y_{ij} = 1$ for $1 \le i \le n$.

In a Self-Protective no saying model, when we use DQM, we can write the probability of belonging to group A as

$$P(X_i = 1) = P(X_{D_i} = 1 \setminus T_i = 1)P(T_i = 1) + P(X_{D_i} = 0 \setminus T_i = 0)P(T_i = 0)$$

which reduces to

$$\pi_A = \pi_D + (1 - \pi_T) \tag{3}$$

from the joint pmf given above. When the characteristic under study is socially unacceptable in nature and if some respondents are known to answer untruthfully, then the estimator from DQM will underestimate the true proportion π_A, thus the estimate of bias $1-\pi_T$ needs to be added to the estimate of π_D in order to attain an unbiased estimate of π_A. Since the SDB is not observable in a DQM, we propose applying a two-stage model, which allows one to obtain an estimator for $1-\pi_T$.

In the proposed model, we first ask the respondents the sensitive question "*Do you belong to group A*?" with DQM and record their answers, then we provide a randomization device by which they randomly chose one of the two questions "Did you answer the previous question truthfully?" or "Did you answer the previous question untruthfully?" with known probabilities θ and $1-\theta$, respectively $(0 \le \theta \le 1)$. Thus, by combining DQM and Warner's RRM in the same model, one can estimate the SDB, choose between the DQM or the combined model, and hence avoid paying the price of the increased variance. To illustrate the proposed two-stage design, we provide the following chart below:

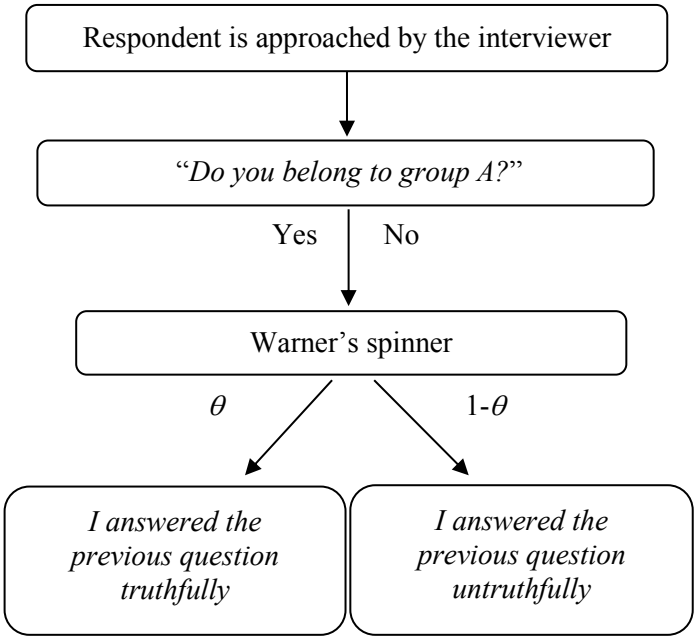

Figure 1: Proposed two-stage design for binary data

Now, let

$$X_{Ri} = \begin{cases} 1, \text{if the respondent says } "yes" \text{ to the first question} \\ 0, \text{if the respondent says } "no" \text{ to the first question} \end{cases}$$

then, the probability of getting a *"yes"* response from first stage is $\pi_R = \pi_T\theta + (1-\pi_T)(1-\theta)$. Solving this equation with respect to $1-\pi_T$ (the proportion of the cheaters) for $\theta \neq 0.5$ yields the unbiased estimator

$$1-\hat{\pi}_T = (\hat{\pi}_R - \theta)/(1-2\theta) \tag{4}$$

where $\hat{\pi}_R = \sum_{i=1}^{n} X_{Ri} / n$. Realize that the unbiased estimator in (4) is the maximum likelihood estimator (MLE). From equations (3) and (4), the proposed estimator becomes:

$$\hat{\pi}_{Ap} = \frac{n_1^D}{n} + \left(\frac{1}{1-2\theta}\right)\left(\frac{n_1^R}{n} - \theta\right) \tag{5}$$

where $n_1^D = \sum_{i=1}^{n} X_{Di}$ is the number of *"yes"* responses from the first stage, and $n_1^R = \sum_{i=1}^{n} X_{Ri}$ is the number of *"yes"* responses from the second stage. Since $\sum_{i=1}^{n} X_{Di} \sim Binom(n, \pi_D)$ and $\sum_{i=1}^{n} X_{R_i} \sim Binom(n, \pi_R)$, it follows from (5) that the proposed estimator is unbiased, and the variance of the proposed estimator can easily derived as

$$Var(\hat{\pi}_{Ap}) = \frac{1}{n}\left[\pi_D(1-\pi_D) - 2\pi_D(1-\pi_T) + \pi_T(1-\pi_T) + \frac{\theta(1-\theta)}{(2\theta-1)^2}\right] \tag{6}$$

Details of the proposed model, as well as the derivations of the formulas given above are provided in a recently submitted paper by Ardah and Oral (2015). The variance in (6) can be estimated by replacing π_D with

$$\hat{\pi}_D = \frac{n_1^D}{n} \tag{7}$$

and by replacing $1-\pi_T$ with its estimate (4). Clearly, the proposed estimator's variance also includes a penalty for using a randomization process, namely the Warner's RRM. However, realize that the proposed framework enables one to estimate π_A both from (5) and also from (7) using DQM; furthermore it allows one to estimate the cheating proportion in the study population from (4). Thus, although the proposed two-stage model's variance equals to the Warner's model's variance, it has a clear advantage: it lets one to choose between the estimators (5) and (7) by estimating the cheating proportion in the study population; more details are given in Ardah and Oral (2015).

3. Simulation Study

In order to study the behavior of the proposed design, we performed the following simulation study. For the sample size n=100, assuming that $\pi_A = 0.3$ (without loss of generality), we changed the values of $1-\pi_T$, i.e. the proportion of the cheaters, from 0 to 0.25 and obtained the estimates and MSE values from both DQM and proposed model for

several different θ values, specifically for θ=0.1, 0.3, 0.35, and 0.4. We also calculated the relative efficiency (RE) values, which is defined as the MSE of the proposed model over MSE of the DQM. The results from 10,000 runs are given in the table below.

Table 1: Simulation results for n=100. T represents the results from the proposed two-stage model; DQM represents the results from the DQM.

	$1-\pi_T$	0	0.01	0.025	0.05	0.1	0.2	0.25
θ								
0.1	Bias(T)	0.00040	0.00076	0.00049	0.00011	0.00104	0.00027	0.00004
	Bias(DQM)	0.00057	0.01030	0.02514	0.05039	0.10041	0.20001	0.25000
	Theoretical Var(T)	0.00351	0.00350	0.00350	0.00350	0.00350	0.00350	0.00350
	Empirical Var(T)	0.00348	0.00362	0.00361	0.00370	0.00379	0.00396	0.00370
	MSE(T)	0.00350	0.00351	0.00350	0.00351	0.00350	0.00350	0.00350
	MSE(DQM)	0.00210	0.00216	0.00267	0.00437	0.01163	0.04106	0.06304
	R.E.	1.66660	1.62830	1.30880	0.80400	0.30100	0.08520	0.05550
0.3	Bias(T)	0.00106	0.00086	0.00029	0.00211	0.00021	0.00113	0.00017
	Bias(DQM)	0.00014	0.00993	0.02464	0.05031	0.10022	0.19992	0.25003
	Theoretical Var(T)	0.01523	0.01523	0.01523	0.01521	0.01522	0.01522	0.01523
	Empirical Var(T)	0.01521	0.01536	0.01532	0.01545	0.01537	0.01564	0.01528
	MSE(T)	0.01523	0.01523	0.01522	0.01522	0.01522	0.01522	0.01523
	MSE(DQM)	0.00210	0.00216	0.00264	0.00435	0.01162	0.04074	0.06304
	R.E.	7.25190	7.06640	5.75900	3.49810	1.31040	0.37360	0.24160
0.35	Bias(T)	0.00070	0.00103	0.00036	0.00148	0.00018	0.00083	0.00189
	Bias(DQM)	0.00088	0.01003	0.02469	0.05018	0.10024	0.20007	0.25023
	Theoretical Var(T)	0.02738	0.02736	0.02738	0.02737	0.02737	0.02738	0.02738
	Empirical Var(T)	0.02733	0.02825	0.02736	0.02762	0.02779	0.02766	0.02767
	MSE(T)	0.02737	0.02739	0.02739	0.02737	0.02737	0.02737	0.02736
	MSE(DQM)	0.00210	0.00216	0.00262	0.00443	0.01153	0.04100	0.06295
	R.E.	13.0317	12.6549	10.4719	6.17940	2.37380	0.66760	0.43470
0.4	Bias(T)	0.00019	0.00166	0.00188	0.00215	0.00420	0.00229	0.00330
	Bias(DQM)	0.00015	0.00993	0.02497	0.04955	0.10008	0.20002	0.24986
	Theoretical Var(T)	0.06211	0.06211	0.06209	0.06210	0.06208	0.06211	0.06209
	Empirical Var(T)	0.06075	0.06186	0.06351	0.06086	0.06270	0.06050	0.06208
	MSE(T)	0.06210	0.06210	0.06214	0.06209	0.06210	0.06209	0.06212
	MSE(DQM)	0.00210	0.00215	0.00259	0.00434	0.01177	0.04097	0.06286
	R.E.	29.5708	28.8351	23.9536	14.2930	5.27550	1.51550	0.98830

From the table above, it may be seen that the bias of the proposed model is always smaller than the bias of the DQM, excluding the case when there is no cheating (i.e. when $1-\pi_T = 0$) , as expected. We also observe that theoretical and empirical variances are consistent with each other, which is also expected. An important result from these simulations is, when the cheating amount is small in the population, using DQM is always better than using the proposed two-stage model (RE>1); however, when the cheating proportion slightly increases then the proposed model becomes more efficient (RE<1). We also provided an alternative method of choosing between these two models in Ardah and Oral (2015).

References

Ardah I. H. and Oral, E., 2015. Improving the Efficiency of Randomized Response Designs by a Two-Stage Model, *Under Review.*

Chaudhuri A., 2011. Randomized Response and Indirect Questioning Techniques in Surveys. CRC Press,NY.

Chaudhuri A. and Mukerjee R., 1988. Randomized Response. Marcel Dekker, NY.

Hout A. V., Bockenholt U. and Heijden P. G. M., 2010. Estimating the Prevalence of Sensitive Behaviour and Cheating with a Dual Design for Direct Questioning and Randomized Response, applied Statistics, 59, 723-736.

Greenberg, R. G., Abul-Ela, A. L. A., Simmons, W. R. & Horvitz, D. G., 1969. The Unrelated Question Randomized Response Model- Theoretical Framework. *Journal of the American Statistical Association*. 520-539.

Greenberg, R. G., Keubler, R. T., Abernathy, J. R. & Horvtiz, D. G., 1971. Application of Randomized Response Technique in Obtaining Quantitative Dta. *Journal of the American Statistical Association*. 243-250.

Gupta, S. N., 2001. Qualifying the Sensitivity Level of Binary Response Personal Interview Survey Questions. *Journal of Combinatorics, Information and System Sciences*. 101-109.

Gupta, S. N. a. T. B., 2002. Circumventing Social Desirability Response Bias in Personal Interview Surveys. *American Journal of Mathematical and Management Sciences*. 369-383.

Gupta, S. N. & Shabbir, J., 2004. Sensitivity Estimation for Personal Interview Survey Questions. *Statistica*. 643-653.

Gupta, S. N. & Shabbir, J., 2007. Mean and Sensitivity Estimation in Optional Randomized Response Models. *Journal of Statistical Planning and Inference*. 2870-2874.

Gupta, S. N., Tuke, S., Spears Gill, T. & Crowe, M., 2013. Optional Unrelated Question Randomized Response Models.. *Involve.,* pp. 483-492.

Tian G. L. and Tang, M. L. 2014. Incomplete Categorical Data Design: Nonrandomized Response Techniques for Sensitive Questions in Surveys. CRC Press, NY.

Warner, S. L., 1965. Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias.. *Journal of the American Statistical Association.* 63-69.

Yu, B, Jin, Z., Tian, J. and Gao, G. (2015) Estimation of Sensitive Proportion by Randomized Response Data in Successive Sampling, Comput Math Methods Med., Volume 2015, Article ID 172918.