Metrika (2008) 67:251–263
DOI 10.1007/s00184-007-0131-x

Two new models for survey sampling with sensitive characteristic: design and analysis

Jun-Wu Yu · Guo-Liang Tian · Man-Lai Tang

Received: 29 March 2006 / Published online: 18 April 2007

Abstract Sensitive topics or highly personal questions are often being asked in medical, psychological and sociological surveys. This paper proposes two new models (namely, the triangular and crosswise models) for survey sampling with the sensitive characteristics. We derive the maximum likelihood estimates (MLEs) and large-sample confidence intervals for the proportion of persons with sensitive characteristic. The modified MLEs and their asymptotic properties are developed. Under certain optimality criteria, the designs for the cooperative parameter are provided and the sample size formulas are given. We compare the efficiency of the two models based on the variance criterion. The proposed models have four advantages: neither model requires randomizing device, the models are easy to be implemented for both interviewer and interviewee, the interviewee does not face any sensitive questions, and both models can be applied to both face-to-face personal interviews and mail questionnaires.

Keywords Maximum likelihood estimate · Randomizing device · Randomized response technique · Sensitive questions · Warner's model

J.-W. Yu
School of Mathematics and Computational Science,
Hunan University of Science and Technology, Xiangtan,
Hunan 411201, People's Republic of China
e-mail: jwyu@265.com

G.-L. Tian (✉)
Division of Biostatistics, University of Maryland Greenebaum Cancer Center,
22 South Greene Street, Baltimore, MD 21201 USA
e-mail: gtian2@umm.edu

M.-L. Tang
Department of Mathematics, Hong Kong Baptist University, Kowloon Tong,
Hong Kong, People's Republic of China
e-mail: mltang@math.hkbu.edu.hk

1 Introduction

Sensitive topics or highly personal questions are often being asked in medical, psychological and sociological surveys. For questions related to abortion, illegitimate birth, AIDs, illegal betting, shoplifting, drug-taking, tax evasion, annual income and students' cheating behavior, some respondents may refuse to answer. Even worse, they may provide wrong answers to maintain privacy when these questions are being asked directly. The final data will then include refusal bias, response bias or both. As a result, it is difficult to make inferences based on these inaccurate data.

To overcome the aforementioned difficulty, Warner (1965) proposed a so-called randomized response (RR) technique that allows researchers to obtain sensitive information while protecting privacy of respondents. The objective is to encourage truthful answers. His model was to offer the respondent a choice of questions where one was the opposite of the other. For example,

(a) I have cheated in an examination (i.e., I am a member of class A);
(b) I have never cheated in an examination (i.e., I am not a member of class A).

The respondents are required to give a 'Yes' or 'No' answer to either the statement (a) or its opposite (b) depending on the outcome of a randomizing device not being revealed to the interviewer. Usually, the probability p of selecting statement (a) by the randomizing device is designed to be known. Suppose that we would like to estimate the proportion π of the population belonging to the sensitive class A. Let n' be the number of 'Yes' answers obtained from the n respondents selected by simple random sampling with replacement. Warner (1965) obtained the MLE $\hat{\pi}_W = (p - 1 + n'/n)/(2p - 1)$, where $p \neq 0.5$. The estimator $\hat{\pi}_W$ was shown to be unbiased and has variance

$$\mathrm{Var}(\hat{\pi}_W) = \frac{\pi(1-\pi)}{n} + \frac{p(1-p)}{n(2p-1)^2}. \tag{1.1}$$

Levy (1976) considered testing hypothesis about π.

It should be noted that the introduction of the randomizing device results in extra amount (i.e., the second term) in (1.1). One way to reduce variance is simply to increase the sample size n. Another way is to perform multiple trials on each respondent (Horvitz et al. 1967; Gould et al. 1969; Greenberg et al. 1974; Liu and Chow 1976). Flingner et al. (1977), Devore (1977) and Moors (1981) identified some technical problems in Warner's model and recommended the truncated estimator $\hat{\pi}_{TW} = \min\{\max(\hat{\pi}_W, 0), 1\}$ which is the true MLE of π. Abul-Ela et al. (1967) and Bourke (1982) extended Warner's model to the trichotomous case to estimate the proportions of three mutually exclusive groups with at least one sensitive group(s). Eriksson (1973) showed how multinomial proportions can be estimated with only one sample using a different randomizing device. Liu et al. (1975) developed a new randomizing device that can be used in the multi-proportions cases. Franklin (1989) considered a dichotomous population but used a randomizing device for continuous distributions.

Horvitz et al. (1967) and Greenberg et al. (1969) proposed an *unrelated question model*. They suggested that the respondents might be more cooperative if we replace statement (b) in Warner's original model by statement (c) I was born in the month of

April (i.e., I am a member of class U), which is non-sensitive and unrelated to statement (a). His/her privacy can be protected since the randomizing device is operated by the respondent and the interviewer does not know which question has been answered. Let π' denote the proportion of individuals in the population who would answer 'Yes' to statement (c). If π' is known, only one sample is required to estimate the proportion π of the population belonging to the sensitive class A. An unbiased estimator of π is $\hat{\pi}_U = [n'/n - (1-p)\pi']/p$ with variance

$$\mathrm{Var}(\hat{\pi}_U) = \frac{\pi(1-\pi)}{n} + \frac{(1-p)^2\pi'(1-\pi') + p(1-p)(\pi+\pi'-2\pi\pi')}{np^2}. \quad (1.2)$$

If π' is unknown in advance, we require two independent samples of size n_1 and n_2 with different probabilities p_1 and p_2 for two randomizing devices. Moors (1971) showed that with an optimal allocation of n_1 and n_2, and $p_2 = 0$, the unrelated question model would be more efficient than the Warner's model for $p_1 > 0.5$, regardless of the choice of π'. Dowling and Shachtman (1975) proved that $\mathrm{Var}(\hat{\pi}_U)$ less than $\mathrm{Var}(\hat{\pi}_W)$ for all π and π', provided that p (or the $\max(p_1, p_2)$ in the two-sample case) is greater than 0.339333. Folsom et al. (1973) evaluated an alternative two-sample design.

Kong (1997) proposed a *random-variable-sum response model*. His main idea is as follows. In survey sampling, interviewer may design a questionnaire in which two questions are designed in such a way that one is sensitive while the other is non-sensitive and unrelated to the sensitive question. For example, the respondent is offered two questions: (1) have you cheated in an examination? (2) Were you born in April? Please give your response according to the following rules: (i) if both of your answers are 'No', please say '0'; (ii) if one of your answer is 'No' and the other 'Yes', please say '1'; (iii) if both of your answers are 'Yes', please say '2'. Kong (1997) presented an extensive study on this model and extended it to the situation with a quantitative measure of the sensitive attribute.

Several papers provided thorough reviews on RR techniques (e.g., see, Greenberg et al. 1974, 1986; Horvitz et al. 1975, 1976; Daniel 1993; Tracy and Mangat 1996; Franklin 1998). For monographs about RR techniques, one can refer to Cochran (1977, 392–395), Chaudhuri and Mukerjee (1988), Hedayat and Sinha (1991, Chap. 11), and Chaudhuri and Stenger (1992, Chap. 10). For recent papers on RR techniques, one can refer to Kim and Warde (2004), Kim and Elam (2005) and Saha (2006).

We notice several drawbacks in Warner's model. Firstly, the interviewee must answer a sensitive question no matter which card he/she selects randomly. Note that question (b) is also sensitive because it is a complement of question (a). As pointed out by Franklin (1998), Warner's model implicitly makes an assumption that the respondent is sufficiently cognizant, informed, and educated to recognize and appreciate his or her anonymity. For an audience of poor education or low sophistication, whatever explanations, he/she might elect not to reply at all or to provide with incorrect answers when he/she is being asked a sensitive topic. Secondly, a randomizing device must be provided to the respondent. The device suggested by Warner (1965) is a spinner with an arrow pointer. Greenberg et al. (1986) pointed out the unsatisfactory aspects of this device and other devices designed by other authors. Finally, the application

of Warner's model has been limited almost exclusively to face-to-face personal interviews. It seems to be infeasible for mail questionnaire. For the unrelated question model, the respondents still need to answer a sensitive question with probability p. For the random-variable-sum response model, answer '2' indicates the explicit exposure to the sensitive attribute.

To overcome these drawbacks, we propose two new models (namely, the triangular and crosswise models) without using randomizing devices for survey sampling with sensitive characteristics. In Sect. 2, we describe the two models together with their advantages. In Sect. 3, we derive their maximum likelihood estimates (MLEs) and asymptotic properties. Section 4 considers the design of the cooperative parameter and gives the sample size formulas. We compare the efficiency of the two models based on the variance criterion in Sect. 5. Section 6 gives a discussion.

2 The proposed models

Let $X = 1$ denote the class of people who possess a sensitive characteristic (e.g., drug-taking) and $X = 0$ denote the complementary class. Y is also a dichotomous random variable. Suppose that Y is non-sensitive and independent of X. For example, $Y = 1$ may represent whether a person was born between August and December and $Y = 0$ represents the corresponding complementary class. The interviewer should select a suitable Y so that the proportion $p = \Pr(Y = 1)$ can be estimated easily. Without loss of generality, in this chapter we always assume that p is known. The purpose is to estimate the proportion $\pi = \Pr(X = 1)$.

2.1 The triangular model

For the face-to-face personal interviews, the interviewer may design an investigation format as the left-hand side of Table 1 and ask the interviewee to put a tick in the circle or in the triangle according to his/her truthful situation. We can see that $\{X = 0, Y = 0\}$ denotes the subclass of person who was not a drug user and was born between January and July. That is, $\{X = 0, Y = 0\}$ is a non-sensitive subclass. A tick put in the triangle indicates the interviewee either was a drug user, or was not a drug user but was born between August and December. Therefore $\{X = 1\} \cup \{X = 0, Y = 1\}$ is also a non-sensitive subclass. Such knowledge would presumably not only make respondents willing to participate in the survey but also persuade them to provide truthful responses.

Table 1 The triangular model and the corresponding cell probabilities

Categories	$Y = 0$	$Y = 1$	Categories	$Y = 0$	$Y = 1$	Total
$X = 0$	○	•	$X = 0$	$(1-\pi)(1-p)$	$(1-\pi)p$	$1-\pi$
$X = 1$	•	•	$X = 1$	$\pi(1-p)$	πp	π
			Total	$1-p$	p	1

Note Please truthfully put a tick in the circle or in the triangle

Table 2 The crosswise model and the corresponding cell probabilities

Categories	$Y=0$	$Y=1$	Categories	$Y=0$	$Y=1$	Total
$X=0$	○	•	$X=0$	$(1-\pi)(1-p)$	$(1-\pi)p$	$1-\pi$
$X=1$	•	○	$X=1$	$\pi(1-p)$	πp	π
			Total	$1-p$	p	1

Note Please truthfully put a tick in the main diagonal or in the antidiagonal

2.2 The crosswise model

The interviewer may also design another investigation format as the left-hand side of Table 2 and ask the interviewee to put a tick in the main diagonal or in the antidiagonal according to his/her truthful situation. Noting that both subclass $\{X=0, Y=0\}$ and $\{X=0, Y=1\}$ are non-sensitive, so are both $\{X=0, Y=0\}\cup\{X=1, Y=1\}$ and $\{X=0, Y=1\}\cup\{X=1, Y=0\}$. Thus the interviewer may record the response but never know whether the interviewee belongs to the sensitive class.

Obviously, the two models have the following advantages: neither model requires randomizing device, the models are easy to operate for both interviewer and interviewee, the interviewee does not face any sensitive questions, and both models can be applied to both face-to-face personal interviews and mail questionnaire.

3 MLEs and asymptotic properties

For the proposed models, we derive the MLEs of π and its large-sample confidence intervals, and investigate the modified MLEs and their asymptotic properties.

3.1 The triangular model

Suppose that a sample of size n with replacement or without replacement from a population results in s ticks in the circle and $n-s$ ticks in the triangle (see Table 1). By introducing a new parameter $\theta=(1-\pi)(1-p)$, we have $\pi=1-\theta/(1-p)$, where $p=\Pr(Y=1)$ is chosen by the interviewer and is therefore known. The likelihood function is proportional to $\theta^s(1-\theta)^{n-s}$ so that the MLE of θ is given by $\hat{\theta}=s/n$. Therefore, the MLE of π is

$$\hat{\pi}_T=1-\hat{\theta}/(1-p), \tag{3.1}$$

where the subscript 'T' represents the 'Triangular model'. Since $s\sim$ Binomial(n,θ), we have $E(s)=n\theta$ and $\mathrm{Var}(s)=n\theta(1-\theta)$. Hence, $E(\hat{\pi}_T)=\pi$, i.e., $\hat{\pi}_T$ is unbiased, and

$$\mathrm{Var}(\hat{\pi}_T)=\theta(1-\theta)/[n(1-p)^2]=\pi(1-\pi)/n+p(1-\pi)/[n(1-p)]. \tag{3.2}$$

Notice that the variance of $\hat{\pi}_T$ can be expressed as the sum of the variance due to sampling and the variance due to the introduction of non-sensitive variable Y. It is easy to show that an unbiased estimate of $\mathrm{Var}(\hat{\pi}_T)$ is given by

$$\widehat{\mathrm{Var}}(\hat{\pi}_T) = \hat{\theta}(1-\hat{\theta})/[(n-1)(1-p)^2]. \tag{3.3}$$

When $n \to \infty$, the central limit theorem implies that $\hat{\pi}_T$ is asymptotically normal, i.e.,

$$(\hat{\pi}_T - \pi) \Big/ \sqrt{\widehat{\mathrm{Var}}(\hat{\pi}_T)} \dot{\sim} N(0,1). \tag{3.4}$$

A $(1-\alpha)100\%$ confidence interval of π can be constructed as

$$\hat{\pi}_T \pm Z_{\alpha/2}\sqrt{\widehat{\mathrm{Var}}(\hat{\pi}_T)}, \tag{3.5}$$

where $Z_{\alpha/2}$ denotes the $1-\alpha/2$ percentage point of the standard normal variable Z such that $\Pr\{Z \le Z_{\alpha/2}\} = 1-\alpha/2$. From (3.4), the usual hypothesis testing about π can be easily established.

From (3.1), we know that $0 \le \hat{\pi}_T \le 1$ if and only if $0 \le \hat{\theta} \le 1-p$. Therefore, the MLE $\hat{\pi}_T$ can be modified as

$$\hat{\pi}_{TM} = \max(0, \hat{\pi}_T) = \begin{cases} \hat{\pi}_T, & \text{if } 0 \le s/n \le 1-p, \\ 0, & \text{if } 1-p < s/n \le 1. \end{cases}$$

Theorem 1 *If $0 < \pi < 1$, then $\sqrt{n}(\hat{\pi}_{TM} - \pi)$ and $\sqrt{n}(\hat{\pi}_T - \pi)$ have the same asymptotic distribution for sufficiently large n.*

The proof of this theorem is omitted since it is similar to that of Theorem 2 below. Theorem 1 states that $\hat{\pi}_{TM}$ and $\hat{\pi}_T$ are asymptotically equivalent.

3.2 The crosswise model

Suppose there are n respondents with r ticks being put in the main diagonal (see Table 2). The observed data are denoted by $\{r, n-r\}$. Defining a new parameter $\lambda = (1-\pi)(1-p)+\pi p$, we have $\pi = (\lambda+p-1)/(2p-1)$, where $p = \Pr(Y=1) \ne 0.5$ is known. The likelihood function is proportional to $\lambda^r(1-\lambda)^{n-r}$ so that the MLE of λ is given by $\hat{\lambda} = r/n$. Therefore, the unbiased MLE of π is

$$\hat{\pi}_C = (\hat{\lambda} + p - 1)/(2p-1), \tag{3.6}$$

where the subscript 'C' refers to 'Crosswise model'. The expression (3.6) shows that the crosswise model bears a formal resemblance to Warner's model. Similar to (3.2)–(3.5), we have respectively

$$\mathrm{Var}(\hat{\pi}_C) = \frac{\lambda(1-\lambda)}{n(2p-1)^2} = \frac{\pi(1-\pi)}{n} + \frac{p(1-p)}{n(2p-1)^2}, \tag{3.7}$$

$$\widehat{\mathrm{Var}}(\hat{\pi}_C) = \frac{\hat{\lambda}(1-\hat{\lambda})}{(n-1)(2p-1)^2} = \frac{\hat{\pi}_C(1-\hat{\pi}_C)}{n-1} + \frac{p(1-p)}{(n-1)(2p-1)^2},$$

$$(\hat{\pi}_C - \pi) \Big/ \sqrt{\widehat{\mathrm{Var}}(\hat{\pi}_C)} \dot{\sim} \mathrm{N}(0,1), \quad \text{as } n \to \infty,$$

$$\hat{\pi}_C \pm Z_{\alpha/2}\sqrt{\widehat{\mathrm{Var}}(\hat{\pi}_C)}.$$

From (3.6), we have $0 \le \hat{\pi}_C \le 1$ if and only if $\min(1-p, p) \le \hat{\lambda} \le \max(1-p, p)$. Therefore the modified MLE of π is

$$\hat{\pi}_{CM} = \min\{1, \max(0, \hat{\pi}_C)\} = \begin{cases} 0, & \text{if } 0 \le r/n < \min(1-p, p), \\ \hat{\pi}_C, & \text{if } \min(1-p, p) \le r/n \le \max(1-p, p), \\ 1, & \text{if } \max(1-p, p) < r/n \le 1. \end{cases} \tag{3.8}$$

The following theorem shows that $\hat{\pi}_{CM}$ and $\hat{\pi}_C$ are asymptotically equivalent.

Theorem 2 *If $0 < \pi < 1$, then $\sqrt{n}(\hat{\pi}_{CM} - \pi)$ and $\sqrt{n}(\hat{\pi}_C - \pi)$ have the same asymptotic distribution as $n \to \infty$.*

Proof It suffices to show that $\sqrt{n}(\hat{\pi}_{CM} - \pi) - \sqrt{n}(\hat{\pi}_C - \pi)$ converges to zero in probability as $n \to \infty$, i.e.,

$$\Pr\{|\sqrt{n}(\hat{\pi}_{CM} - \hat{\pi}_C)| > 0\} \to 0, \quad \text{as } n \to \infty. \tag{3.9}$$

Noting that $\hat{\lambda}$ is the MLE of $\lambda = (1-\pi)(1-p) + \pi p$ and that $\min(1-p, p) < \lambda < \max(1-p, p)$ since $0 < \pi < 1$, we naturally obtain $\Pr\{|\hat{\lambda} - \lambda| > \varepsilon\} \to 0$, as $n \to \infty$, for any $\varepsilon > 0$. We only need to prove

$$\Pr\{|\sqrt{n}(\hat{\pi}_{CM} - \hat{\pi}_C)| > 0\} \le \Pr\{|\hat{\lambda} - \lambda| > \varepsilon\},$$

or equivalently

$$\{|\sqrt{n}(\hat{\pi}_{CM} - \hat{\pi}_C)| > 0\} \subseteq \{|\hat{\lambda} - \lambda| > \varepsilon\}, \tag{3.10}$$

for any $\varepsilon < \min\{\max(1-p, p) - \lambda, \lambda - \min(1-p, p)\}$. Without loss of generality, we assume $p > 1/2$. We consider three cases.

Case 1 $1 - p \le \hat{\lambda} \le p$. From (3.8), we obtain $\hat{\pi}_{CM} = \hat{\pi}_C$. Therefore (3.9) follows immediately.

Case 2 $\hat{\lambda} < 1 - p$. Now $\hat{\pi}_{CM} = 0$. If

$$\begin{aligned} & |\sqrt{n}(\hat{\pi}_{CM} - \hat{\pi}_C)| > 0 \\ \Rightarrow\ & |\hat{\lambda} + p - 1| > 0 \\ \Rightarrow\ & 0 < |\hat{\lambda} + p - 1| = -(\hat{\lambda} + p - 1) = -(\hat{\lambda} - \lambda) - \{\lambda - (1-p)\} \\ \Rightarrow\ & |\hat{\lambda} - \lambda| \ge -(\hat{\lambda} - \lambda) > \lambda - (1-p). \end{aligned} \tag{3.11}$$

Noting that $\varepsilon < \min\{p - \lambda, \lambda - (1 - p)\}$, we have

$$\{\lambda - (1 - p)\} - \varepsilon > 0. \tag{3.12}$$

By combining (3.11) with (3.12), we obtain

$$|\hat{\lambda} - \lambda| - \varepsilon = |\hat{\lambda} - \lambda| - \{\lambda - (1 - p)\} + \{\lambda - (1 - p)\} - \varepsilon > 0$$

and hence (3.10) follows.

Case 3 $\hat{\lambda} > p$. Now $\hat{\pi}_{CM} = 1$. If

$$\begin{aligned} & |\sqrt{n}(\hat{\pi}_{CM} - \hat{\pi}_C)| > 0 \\ \Rightarrow\ & |p - \hat{\lambda}| > 0 \\ \Rightarrow\ & 0 < |p - \hat{\lambda}| = -(p - \hat{\lambda}) = -(\lambda - \hat{\lambda}) - (p - \lambda) \\ \Rightarrow\ & |\lambda - \hat{\lambda}| \geq -(\lambda - \hat{\lambda}) > p - \lambda. \end{aligned} \tag{3.13}$$

Noting that $\varepsilon < \min\{p - \lambda, \lambda - (1 - p)\}$, we have

$$(p - \lambda) - \varepsilon > 0. \tag{3.14}$$

By combining (3.13) and (3.14), we obtain

$$|\hat{\lambda} - \lambda| - \varepsilon = |\hat{\lambda} - \lambda| - (p - \lambda) + (p - \lambda) - \varepsilon > 0.$$

Hence, (3.10) follows. The proof is completed. □

4 Design consideration

For the proposed models, the unknown proportion π is required to be estimated, while the cooperative parameter $p = \Pr(Y = 1)$ and the sample size n are controllable. This section addresses the issues of determining p and n under certain optimality criteria. The value of p somehow indicates whether the respondents are cooperative, while the sample size n depends on the estimated degrees of precision.

4.1 Design of the cooperative parameter

We first consider the triangular model. From (3.2), we can see that $\mathrm{Var}(\hat{\pi}_T)$ is an increasing function of p $(0 \leq p \leq 1)$ for any fixed π and given sample size n. When $p = 1$, $\mathrm{Var}(\hat{\pi}_T) \to \infty$. When $p = 0$, $\mathrm{Var}(\hat{\pi}_T)$ attains its minimum value $\pi(1 - \pi)/n$, which reduces the triangular model to the direct inquiry model that requires the respondent to unreservedly state whether or not he/she belongs to the sensitive class $\{X = 1\}$.

Therefore, we may establish the following criterion:

$$\frac{p(1-\pi)}{n(1-p)} \Big/ \mathrm{Var}(\hat{\pi}_T) \le \alpha_0, \tag{4.1}$$

where $\alpha_0 \in (0, 1)$ is known and is decided by the interviewer. The inequality (4.1) states that the proportion of the variance due to the introduction of the non-sensitive variable Y to the total variance is less than or equal to $100\alpha_0\%$. From (4.1), we obtain

$$p \le p_{\alpha_0}(\pi) \hat{=} \alpha_0\pi/(1-\alpha_0+\alpha_0\pi).$$

When $0 \le \pi \le 1$, we have $0 \le p_{\alpha_0}(\pi) \le \alpha_0$. In particular, let $\alpha_0 = 0.5$, we have $p_{0.5}(\pi) = \pi/(1+\pi)$. Table 3 shows the values of π and $p_{\alpha_0}(\pi)$ for $\alpha_0 = 0.5$ and $\alpha_0 = 0.75$, respectively.

Now we consider the crosswise model. It is easy to verify from (3.7) that $\mathrm{Var}(\hat{\pi}_C)$ is an increasing (or a decreasing) function of p when $0 \le p < 0.5$ (or $0.5 < p \le 1$) for any fixed π and given n. When $p = 0.5$, $\mathrm{Var}(\hat{\pi}_C) \to \infty$. When $p = 0$ or $p = 1$, $\mathrm{Var}(\hat{\pi}_C)$ attains its minimum value $\pi(1-\pi)/n$, which reduces the crosswise model to the direct inquiry model. Similar to (4.1), we may construct the following criterion for any given $\alpha_0 \in (0, 1)$:

$$\frac{p(1-p)}{n(2p-1)^2} \Big/ \mathrm{Var}(\hat{\pi}_C) \le \alpha_0. \tag{4.2}$$

From (4.2), we have $p \le p^{(1)}_{\alpha_0}(\pi)$ or $p \ge p^{(2)}_{\alpha_0}(\pi)$, where

$$p^{(1)}_{\alpha_0}(\pi) = 0.5[1-(4\beta_0\pi(1-\pi)+1)^{-1/2}], \quad p^{(2)}_{\alpha_0}(\pi) = 1 - p^{(1)}_{\alpha_0}(\pi),$$

and $\beta_0 \hat{=} \alpha_0/(1-\alpha_0) > 0$. Thus, $p^{(1)}_{\alpha_0}(\pi)$ is an increasing (or a decreasing) function of π when $0 \le \pi < 0.5$ (or $0.5 < \pi \le 1$) and it reaches the maximum $0.5[1-(\beta_0+1)^{-1/2}]$ at $\pi = 0.5$. Similarly, $p^{(2)}_{\alpha_0}(\pi)$ is a decreasing (or an increasing) function of π when $0 \le \pi < 0.5$ (or $0.5 < \pi \le 1$) and it arrives the minimum $0.5[1+(\beta_0+1)^{-1/2}]$ at $\pi = 0.5$. Table 4 lists the values of π, $p^{(1)}_{\alpha_0}(\pi)$ and $p^{(2)}_{\alpha_0}(\pi)$ for $\alpha_0 = 0.5$ and $\alpha_0 = 0.75$.

Table 3 Values of π and $p_{\alpha_0}(\pi)$ for $\alpha_0 = 0.5$ and $\alpha_0 = 0.75$

π	0.01	0.03	0.05	0.07	0.09	0.10	0.20	0.30	0.40	0.50
$p_{0.50}(\pi)$	0.0099	0.0291	0.0476	0.0654	0.0826	0.0909	0.1667	0.2308	0.2857	0.3333
$p_{0.75}(\pi)$	0.0291	0.0826	0.1304	0.1736	0.2126	0.2308	0.3750	0.4737	0.5455	0.6000

Table 4 The values of π, $p_{\alpha_0}^{(1)}(\pi)$ and $p_{\alpha_0}^{(2)}(\pi)$ for $\alpha_0 = 0.5$ and 0.75

π	0.01	0.03	0.05	0.07	0.09	0.10	0.20	0.30	0.40	0.50
$p_{0.50}^{(1)}(\pi)$	0.0096	0.0267	0.0416	0.0546	0.0661	0.0712	0.1095	0.1313	0.1428	0.1464
$p_{0.50}^{(2)}(\pi)$	0.9904	0.9733	0.9584	0.9454	0.9339	0.9288	0.8904	0.8687	0.8572	0.8536
$p_{0.75}^{(1)}(\pi)$	0.0272	0.0695	0.1009	0.1253	0.1449	0.1533	0.2073	0.2334	0.2461	0.2500
$p_{0.75}^{(2)}(\pi)$	0.9728	0.9305	0.8991	0.8747	0.8551	0.8467	0.7927	0.7666	0.7539	0.7500

4.2 Determination of the sample sizes

To determine the sample size n for the triangular model, we assume that p is known. Differentiating (3.2) with respect to π and setting it to zero yield $\pi = (0.5 - p)/(1 - p) \cdot I_{(0<p<0.5)}$, where I_S denotes the indicator function for the event S. Noting that $\mathrm{Var}(\hat{\pi}_T)$ arrives its maximum at this π value, we have

$$\mathrm{Var}(\hat{\pi}_T) \leq \frac{1}{4n(1-p)^2} \cdot I_{(0<p<0.5)} + \frac{p}{n(1-p)} \cdot I_{(0.5\leq p<1)}.$$

Result 1 *Let V_0 denote the maximal tolerance variance. The desired sample size n can be determined by $n \geq [4V_0(1-p)^2]^{-1} \cdot I_{(0<p<0.5)} + p[V_0(1-p)]^{-1} \cdot I_{(0.5\leq p<1)}$.*

Result 2 *Let L_0 denote the maximal tolerance length of the confidence interval for π. The sample size n can be determined by*

$$n \geq \left(\frac{Z_{\alpha/2}}{L_0(1-p)}\right)^2 \cdot I_{(0<p<0.5)} + \left(\frac{2Z_{\alpha/2}}{L_0}\right)^2 \frac{p}{1-p} \cdot I_{(0.5\leq p<1)}.$$

For the crosswise model, we also assume that p is known. It is easy to verify from (3.7) that $\mathrm{Var}(\hat{\pi}_C)$ arrives its maximum at $\pi = 0.5$ and we have $\mathrm{Var}(\hat{\pi}_C) \leq 1/[4n(2p-1)^2]$. Similar to Results 1 and 2 obtained for the triangular model, the sample size n for the crosswise model are given by $n \geq 1/[4V_0(2p-1)^2]$ and $n \geq (Z_{\alpha/2}/\{L_0(2p-1)\})^2$, respectively.

5 Comparison of efficiency

In this section, we compare the efficiency between the triangular model and crosswise model by employing the variance criterion. From (3.2) and (3.7), we have

$$\mathrm{Var}(\hat{\pi}_C) - \mathrm{Var}(\hat{\pi}_T) = \frac{p}{n(1-p)(2p-1)^2} \cdot f(p), \tag{5.1}$$

where $f(p) = (4\pi - 3)p^2 + (2 - 4\pi)p + \pi$. We have the following result.

Theorem 3 (i) *If* $\pi = 3/4$, *then*

$$\begin{cases} Var(\hat{\pi}_C) \geq Var(\hat{\pi}_T), & when\ 0 < p \leq 3/4\ \ (p \neq 1/2), \\ Var(\hat{\pi}_C) < Var(\hat{\pi}_T), & when\ 3/4 < p < 1. \end{cases} \tag{5.2}$$

(ii) *If* $\pi \neq 3/4$, *then*

$$\begin{cases} Var(\hat{\pi}_C) \geq Var(\hat{\pi}_T), & when\ 0 < p \leq p_\pi\ \ (p \neq 1/2), \\ Var(\hat{\pi}_C) < Var(\hat{\pi}_T), & when\ p_\pi < p < 1, \end{cases} \tag{5.3}$$

where $p_\pi = (2\pi - 1 - \sqrt{1-\pi})/(4\pi - 3)$ *is an increasing function of* π.

Proof (i) If $\pi = 3/4$, then $f(p) = 3/4 - p$ is nonnegative (or negative) when $0 < p \leq 3/4$ (or $3/4 < p < 1$). From (5.1), we obtain (5.2) immediately.

(ii) If $\pi > 3/4$, then we obtain

$$\begin{cases} f(p) \geq 0, & \text{when } p \leq p_\pi \quad \text{or} \quad p \geq p_2, \\ f(p) < 0, & \text{when } p_\pi < p < p_2, \end{cases}$$

where p_π is defined in Theorem 3 and

$$p_2 = (2\pi - 1 + \sqrt{1-\pi})/(4\pi - 3).$$

It is easy to show that $0.5 < p_\pi < 1 < p_2$. (5.3) follows immediately.

(iii) If $\pi < 3/4$, similarly, we have

$$\begin{cases} f(p) \geq 0, & \text{when } p_1 \leq p \leq p_\pi, \\ f(p) < 0, & \text{when } p < p_1 \quad \text{or} \quad p > p_\pi, \end{cases}$$

where $p_1 = -(2\pi - 1 + \sqrt{1-\pi})/(3 - 4\pi)$. Now we have $p_1 < 0 < 1/2 < p_\pi < 1$. Hence, (5.3) follows. Note that

$$\frac{\partial p_\pi}{\partial \pi} = \frac{(\sqrt{1-\pi} - 2)^2}{2\sqrt{1-\pi}(4\pi - 3)^2} > 0.$$

Hence, p_π is an increasing function of π. The proof is completed. □

6 Discussion

In this paper, we proposed two new models: the triangular and crosswise models, for survey sampling with sensitive questions. The proposed models have four advantages: (i) neither model requires randomizing device; (ii) the models are easy to operate for both interviewer and interviewee; (iii) the interviewee does not face any sensitive

questions; (iv) both models can be applied to both face-to-face personal interviews and mail questionnaire. We mainly studied three problems: (1) which model is better (i.e., the problem of comparison); (2) how does one determine the design parameters (i.e., the problem of design); (3) how does one analyze the gathered data (i.e., the problem of analysis). In the frequentist framework, the unbiased maximum likelihood estimate and large-sample confidence interval for the proportion π of persons with sensitive characteristic are derived. The modified MLEs of π and their asymptotic properties are developed. Under certain optimality criteria, the designs for the cooperative parameter and the sample size formulas are given. A comparison of efficiency between the two models is presented.

Acknowledgments The authors would like to thank an Associate Editor and one referee for their comments and suggestions. Part of the researches in this paper was carried out when GL Tian worked at Peking University, P. R. China, as a postdoctor. The research of ML Tang was fully supported by a grant (CUHK4371/04M) from the Research Grant Council of the Hong Kong Special Administrative Region.

References

Abul-Ela AA, Greenberg BG, Horvitz DG (1967) A multi-proportions randomized response model. J Am Stat Assoc 62:990–1008

Bourke PD (1982) Randomized response multivariate designs for categorical data. Commun Stat A Theory Methods 11:2889–2901

Chaudhuri A, Mukerjee R (1988) Randomized response: theory and techniques. Marcel Dekker, New York

Chaudhuri A, Stenger H (1992) Survey sampling: theory and methods. Marcel Dekker, New York

Cochran WG (1977) Sampling techniques, 3rd edn. Wiley, New York

Daniel WW (1993) Collecting sensitive data by randomized response: an annotated bibliography, 2nd edn. Research Monograph No. 107. Georgia State University Business Press, Atlanta

Devore JL (1977) A note on the randomized response technique. Commun Stat A Theory Methods 6:1525–1529

Dowling TA, Shachtman RH (1975) On the relative efficiency of randomized response models. J Am Stat Assoc 70:84–87

Eriksson SA (1973) A new model for randomized response. Int Stat Rev 41:101–113

Flingner MA, Policello GE, Singh J (1977) A comparison of two randomized response survey methods with consideration for the level of respondent protection. Commun Stat A Theory Methods 6:1511–1524

Folsom RE, Greenberg BG, Horvitz DG, Abernathy JR (1973) The two alternate questions randomized response model for human surveys. J Am Stat Assoc 68:525–530

Franklin LA (1989) Randomized response sampling from dichotomous populations with continuous randomization. Surv Methodol 15:225–235

Franklin LA (1998) Randomized response techniques. In: Armitage P, Colton T (eds) Encyclopedia of biostatistics. Wiley, New York, pp 3696–3703

Gould AL, Shah BV, Abernathy JR (1969) Unrelated question randomized response techniques with two trials per respondent. In: 1969 Proceedings of the Social Statistics Section, American Statistical Association, pp 351–359

Greenberg BG, Abernathy JR, Horvitz DG (1986) Randomized response. In: Kotz S, Johnson NL (eds) Encyclopedia of statistical sciences, Vol. 7. Wiley, New York, pp 540–548

Greenberg BG, Abul-Ela AA, Simmons WR, Horvitz DG (1969) The unrelated question randomized response model: theoretical framework. J Am Stat Assoc 64:520–539

Greenberg BG, Horvitz DG, Abernathy JR (1974) Comparison of randomized response designs. In: Prochan F, Serfling RJ (eds) Reliability and biometry, statistical analysis of life length. Philadelphia, SIAM, pp 787–815

Hedayat AS, Sinha BK (1991) Design and inference in finite population sampling. Wiley, New York

Horvitz DG, Shah BV, Simmons WR (1967) The unrelated question randomized response model. In: 1967 Proceedings of the Social Statistics Section, American Statistical Association, pp 65–72

Horvitz DG, Greenberg BG, Abernathy JR (1975) Recent developments in randomized designs. In: Srivastava JN (ed) A survey of statistical design and linear models. North Holland / American Elsevier Publishing Co., New York, pp 271–285

Horvitz DG, Greenberg BG, Abernathy JR (1976) Randomized response: a data gathering device for sensitive questions. Int Stat Rev 44:181–196

Kim JM, Warde WD (2004) A stratified Warner's randomized response model. J Stat Plann Infer 120: 155–165

Kim JM, Elam ME (2005) A two-stage stratified Warner's randomized response model using optimal allocation. Metrika 61:1–7

Kong SY (1997) Survey sampling for sensitive questions. Unpublished Ph.D. dissertation, Renmin University, Beijing, P. R. China

Levy KJ (1976) Reducing the occurrence of omitted or untruthful responses when testing hypotheses concerning proportions. Psychol Bull 83:759–761

Liu PT, Chow LP (1976) The efficiency of the multiple trial randomized response technique. Biometrics 32:607–618

Liu PT, Chow LP, Mosley WH (1975) Use of the randomized response technique with a new randomizing device. J Am Stat Assoc 70:329–332

Moors JJA (1971) Optimization of the unrelated question randomized response model. J Am Stat Assoc 66:627–629

Moors JJA (1981) Inadmissibility of linearly invariant estimators in truncated parameter spaces. J Am Stat Assoc 76:910–915

Saha A (2006) Optimal randomized response in stratified unequal probability sampling—a simulation based numerical study with Kuk's method. Test (in press)

Tracy DS, Mangat NS (1996) Some developments in randomized response sampling during the last decade—a follow up of review by Chaudhuri and Mukerjee. J Appl Stat Sci 4:147–159

Warner SL (1965) Randomized response: a survey technique for eliminating evasive answer bias. J Am Stat Assoc 60:63–69