A variant of the parallel model for sample surveys with sensitive characteristics

Yin LIU[a], Guo-Liang TIAN[a,1]

[a]*Department of Statistics and Actuarial Science, The University of Hong Kong, Pokfulam Road, Hong Kong, P. R. China*

Abstract

A new *non-randomized response* (NRR) model (called a variant of the parallel model) is proposed. The survey design and corresponding statistical inferences including likelihood-based methods, Bayesian methods and bootstrap methods are provided. Theoretical and numerical comparisons showed that the proposed variant of the parallel model over-performs two existing NRR crosswise and triangular models for most of the possible parameter ranges. An outline for handling the possible non-compliance behavior in the proposed model is provided. An illustrative example from an existing survey on 'sexual practices' in San Francisco, Las Vegas and Portland is used to demonstrate the proposed statistical analysis methods. Two real surveys on the cheating behavior in examinations at the University of Hong Kong are conducted and are used to illustrate the proposed design and analysis methods.

Keywords: Asymptotic properties; Bayesian methods; Non-compliance behavior; Non-randomized response technique; The parallel model; Unmatched count technique.

1. Introduction

Consider a target population which can be divided into two mutually exclusive groups: one with a sensitive attribute and the other without. Statistically, let Y be a sensitive binary variable, $\{Y = 1\}$ denote the population group that has the sensitive attribute and $\{Y = 0\}$ denote the complementary group. Usually, a well-designed survey is conducted for collecting sensitive data, which are used to estimate the proportion (denoted by $\pi = \Pr(Y = 1)$) of persons with the sensitive characteristic. Several techniques are developed to encourage truthful responses while protecting the privacy of respondents (or minimizing the interviewee's feeling of jeopardy). The first one is the *randomized response technique*

[1]Corresponding author. Tel.: +(852)28591984.
E-mail address: gltian@hku.hk

(RRT), which includes Warner's design (Warner, 1965) and its improvement versions such as the unrelated question RR design (Horvitz et al., 1967; Greenberg et al., 1969). For a comprehensive review on RR designs, one is referred to Fox and Tracy (1986), Chaudhuri and Mukerjee (1988), and Chaudhuri (2011). One difficulty in implementing the RRT is the choice of an appropriate randomizing device in a self-administered setting. Another challenge in using RR models is the possible non-compliance behavior because of respondents' mistrust. A complicated or novel randomizing device may lead the interviewee to doubt the method itself or, even worst, to feel that they are being tricked by the interviewer into providing information under false pretenses. To handle non-compliance to RRT instructions, many developments on RR models were proposed by some researchers, e.g., Lakshmi and Raghavarao (1992), Clark and Desharnais (1998), Böckenholt and van der Heijden (2007), Van Den Hout and Klugkist (2009), Ostapczuk et al. (2009a), Ostapczuk et al. (2009b, 2011), Moshagen (2010), Van Den Hout et al. (2010), Moshagen et al. (2012), and so on.

The second one is called the *unmatched count technique* (UCT), which provides absolute anonymity and confidentiality. Under UCT, two forms are needed: Form 1 contains a number of innocuous or neutral questions with answer 'yes' or 'no' and Form 2 is identical to Form 1, except for the addition of one embarrassing question of interest (Dalton et al., 1994, 1997; Coutts and Jann, 2011). The respondents of the survey are randomly assigned to one of two groups. Participants in group 1 (or group 2) are asked to reveal only the number of 'yes'-answer to all items listed in Form 1 (or Form 2). Since the interviewer does not know how they arrived at that number, it is safe to answer the sensitive question truthfully. One advantage of the UCT over the RRT is that no randomized device is required. The UCT is also called the item count technique (Droitcour et al., 1991; Tsuchiya et al., 2007), the unmatched block design, or block total response (Raghavarao et al., 1979). For more detailed description on the UCT, see Dalton et al. (1994).

The third one is called the *non-randomized response* (NRR) technique, which utilizes one or two independent non-sensitive random variates (e.g., respondent's birth date/month or the last digit of a respondent's ID card/phone number) combined with one or two sensitive random variables to form an incomplete contingency table and to indirectly obtain respon-

dents' sensitive answers (Takahasi and Sakasegawa, 1977; Tian et al., 2007b, 2011; Yu et al., 2008; Tan et al., 2009; Tang et al., 2009). Like the UCT, the NRR designs don't require any randomizing devices.

One basic distinction between a randomized response model and a non-randomized response model is that the former usually requires a randomization device such as a coin or a die which is related to a random variable without reproducibility, while the latter requires an independent non-sensitive variate such as birth date combined with the sensitive response variable to form an incomplete contingency table, resulting in a reproducibility. That is, the same respondent may yield different answers depending on the outcome of the randomization device in repeated experiments (e.g., repeatedly flip a coin). For example, in the unrelated question design with a coin as the randomization device, if the outcome is head, the first question is answered; if the outcome is tail, the second question is answered. Suppose that the result of the first (second) flip is a tail (head), the answer is a 'yes' ('no'). As a result, interviewers do not know which answer should be collected.

It is not true that any randomized response model can be easily transformed to a non-randomized response model. Up to now, only the Warner model and the unrelated question model were successfully transformed to the non-randomized crosswise model (Yu et al., 2008) and the non-randomized parallel model (Tian, 2012), respectively. Next, although some randomized response models can be transformed to non-randomized versions, the resulting statistical analysis methods are totaly different. For example, for the randomized unrelated question model with an unknown $\theta = \Pr(U = 1)$, two independent samples of sizes n_1 and n_2 and two randomization devices are required, while for its non-randomized version, i.e., the proposed variant of the parallel model in this paper, only one sample is needed without using any randomization devices and the corresponding statistical analysis methods are developed based on a trinomial distribution with two complete observations and one incomplete observation. The second example is as follows. To assess the association of two sensitive questions with binary outcomes, a randomized response model in general requires two randomization devices (Christofides, 2005), while in the non-randomized hidden sensitivity model (Tian et al., 2007b), respondents only need to answer a non-sensitive ques-

tion instead of the original two sensitive questions and the corresponding analysis methods are developed based on an incomplete 4×4 contingency table. Finally, for other randomized response models (e.g., Kuk, 1990), the corresponding non-randomized partners are not yet available up to now.

Recently, Tian (2012) proposed a new NRR model, called the parallel model, to estimate the unknown proportion, $\pi = \Pr(Y = 1)$, of individuals with a sensitive characteristic. By introducing two non-sensitive dichotomous variates U and W such that Y, U and W are mutually independent, Tian (2012) developed a general framework of design and analysis for the NRR parallel model. Theoretical comparison showed that the parallel model over-performs two existing NRR crosswise and triangular models for most of the possible parameter ranges. It was noted that all these findings are based on the assumption of known proportions $\theta = \Pr(U = 1)$ and $p = \Pr(W = 1)$. However, in survey practice, it is usually difficult to choose an appropriate non-sensitive dichotomous variate U with known $\theta = \Pr(U = 1)$. Even such a binary variable U can be found and a constant θ_0 is assumed to be equal to the true value of the θ, how to test the hypothesis H_0: $\theta = \theta_0$ is still not available for the parallel design. The main goal of this paper is to propose a variant of the parallel model with unknown $\theta = \Pr(U = 1)$.

The rest of the paper is organized as follows. In Section 2, we propose the survey design for the variant of the parallel model, and discuss the estimation of the parameters, relative efficiency and the degree of privacy protection. In Section 3, three asymptotic *confidence intervals* (CIs) and the exact CI of π are derived. In addition, a modified *maximum likelihood estimate* (MLE) of π is provided and the corresponding asymptotic property is investigated. Statistical inferences on θ and two bootstrap CIs of the parameters are given in Sections 4 and 5, respectively. Bayesian inferences are discussed in Section 6. Comparisons with the NRR crosswise and triangular models are conducted theoretically and numerically in Section 7. An outline for handling the possible non-compliance behavior in the proposed model is presented in Section 8. In Section 9, an illustrative example from an existing survey on 'sexual practices' in San Francisco, Las Vegas and Portland is used to demonstrate the proposed statistical analysis methods. Two real surveys on the cheating behavior in examinations at

the University of Hong Kong are conducted and are used to illustrate the proposed design and analysis methods. A discussion is given in Section 10. The exact *inversion Bayesian formulae* (IBF) sampling is provided in Appendix A.

2. A new non-randomized response model: A variant of the parallel model

2.1. The survey design for the variant of the parallel model

Let $\{Y = 1\}$ denote the population class with a sensitive characteristic and $\{Y = 0\}$ denote the complementary class. The objective is to estimate the proportion $\pi = \Pr(Y = 1)$. Suppose that U and W are two non-sensitive dichotomous variates, and Y, U and W are mutually independent with unknown $\theta = \Pr(U = 1)$ and known $p = \Pr(W = 1)$. For example, we may define $U = 1$ if the respondent lives in Hong Kong Island (or likes watching football/soccer on TV, or likes fishing/singing/shopping/traveling, or is educated above the level of high school) and $U = 0$ otherwise. Similarly, we could define $W = 1$ if the last digit of the respondent's ID/cell phone number is odd (or the respondent's birthday is in the second half of a year/month) and $W = 0$ otherwise. Hence, it is reasonable to assume that $p \approx 0.5$.

The interviewer may design the questionnaire in the format as shown at the left-hand side of Table 1 and ask the interviewee to truthfully put a tick in the circle if he/she belongs to $\{U = 0, W = 0\}$ or put a tick in the triangle if he/she belongs to $\{Y = 0, W = 1\}$ or put a tick in the upper square if he/she belongs to $\{U = 1, W = 0\} \cup \{Y = 1, W = 1\}$. Note that all $\{W = 0\}$, $\{W = 1\}$, $\{U = 0\}$, $\{U = 1\}$ and $\{Y = 0\}$ are non-sensitive classes, thus $\{U = 1,\ W = 0\} \cup \{Y = 1,\ W = 1\}$ is also a non-sensitive subclass. Therefore, whether the interviewee belongs to the sensitive class $\{Y = 1,\ W = 1\}$ is not on record. Since θ is unknown, we call this a variant of the parallel model. The corresponding cell probabilities are displayed at the right-hand side of Table 1. Since the three binary variables U, Y and W are independent, the joint probability is the product of two corresponding marginal probabilities.

Table 1. The survey design for the variant of parallel model with unknown $\theta = \Pr(U = 1)$

Category	$W = 0$	$W = 1$
$U = 0$	◯	
$U = 1$	□	
$Y = 0$		△
$Y = 1$		□

Category	$W = 0$	$W = 1$	Marginal
$U = 0$	$(1-\theta)(1-p)$		$1-\theta$
$U = 1$	$\theta(1-p)$		θ
$Y = 0$		$(1-\pi)p$	$1-\pi$
$Y = 1$		πp	π
Marginal	$1-p$	p	1

Note:

- Please truthfully put a tick in the circle if you belong to $\{U = 0, W = 0\}$ or put a tick in the triangle if you belong to $\{Y = 0, W = 1\}$ or put a tick in the upper square if you belong to $\{U = 1, W = 0\} \cup \{Y = 1, W = 1\}$.
- For those respondents not completely understanding the questionnaire shown in Table 1, investigators can formulate the questionnaire of the variant of the parallel model as follows:

  Let $Y = 1$ if a respondent is a drug user and $Y = 0$ otherwise.

  (1) If your birthday is in the first half of a year (i.e., $W = 0$), please answer '0' (i.e., $U = 0$), or '2' (i.e., $U = 1$) to the question: *Do you like shopping?*

  (2) If your birthday is in the second half of a year (i.e., $W = 1$), please answer '1' (i.e., $Y = 0$), or '2' (i.e., $Y = 1$) to the question: *Are you a drug user?*

  Answering '0' is equivalent to putting a tick in the circle in Table 1, answering '1' is equivalent to putting a tick in the triangle in Table 1 and answering '2' is equivalent to putting a tick in the upper square in Table 1.

2.2. Estimation

Suppose that a sample survey with n questionnaires is conducted. Let $Y_{\text{obs}} = \{n;\ n_1, n_2, n_3\}$ denote the observed data, where $n = \sum_{i=1}^{3} n_i$, n_1 represents the number of respondents putting a tick in the circle, n_2 represents the number of respondents putting a tick in the triangle, and n_3 represents the number of individuals who put a tick in the upper square (see Table 1). The likelihood function of the two unknown parameters π and θ for the observed data Y_{obs} is

$$L_{\text{v}}(\pi, \theta | Y_{\text{obs}}) = \binom{n}{n_1, n_2, n_3} [(1-\theta)(1-p)]^{n_1} [(1-\pi)p]^{n_2} [\theta(1-p) + \pi p]^{n_3}, \tag{2.1}$$

where the subscript 'V' refers to the 'variant' of the parallel model. Hence, the corresponding log-likelihood function is given by

$$\ell_{\text{V}}(\pi, \theta|Y_{\text{obs}}) = c + n_1 \log(1-\theta) + n_2 \log(1-\pi) + n_3 \log\{\theta(1-p) + \pi p\},$$

where c is a constant not depending on π and θ. Let

$$\frac{\partial \ell_{\text{V}}(\pi, \theta|Y_{\text{obs}})}{\partial \pi} = 0 \quad \text{and} \quad \frac{\partial \ell_{\text{V}}(\pi, \theta|Y_{\text{obs}})}{\partial \theta} = 0,$$

we obtain

$$\begin{aligned} \frac{-n_2}{1-\pi} + \frac{n_3 p}{\theta(1-p) + \pi p} &= 0 \quad \text{and} \\ \frac{-n_1}{1-\theta} + \frac{n_3(1-p)}{\theta(1-p) + \pi p} &= 0. \end{aligned}$$

Hence, the MLEs of π and θ are given by

$$\hat{\pi}_{\text{V}} = 1 - \frac{n_2}{np} \quad \text{and} \quad \hat{\theta} = 1 - \frac{n_1}{n(1-p)}. \tag{2.2}$$

To derive the expectation and variance of the $\hat{\pi}_{\text{V}}$, we define

$$\begin{aligned} \lambda_1 &= \Pr\{U = 0, W = 0\} = (1-\theta)(1-p), \\ \lambda_2 &= \Pr\{Y = 0, W = 1\} = (1-\pi)p \quad \text{and} \\ \lambda_3 &= \Pr\{U = 1, W = 0\} + \Pr\{Y = 1, W = 1\} = \theta(1-p) + \pi p. \end{aligned} \tag{2.3}$$

Obviously, we have $(n_1, n_2, n_3)^\top \sim \text{Multinomial}(n; \lambda_1, \lambda_2, \lambda_3)$. Note that the MLEs of $\{\lambda_i\}_{i=1}^3$ are given by $\hat{\lambda}_i = n_i/n$ and $E(n_i) = n\lambda_i$, $i = 1, 2, 3$. It is easy to verify that $\hat{\pi}_{\text{V}}$ is an unbiased estimator of π and the variance of $\hat{\pi}_{\text{V}}$ is given by

$$\text{Var}(\hat{\pi}_{\text{V}}) = \frac{\lambda_2(1-\lambda_2)}{np^2} \overset{(2.3)}{=} \text{Var}(\hat{\pi}_{\text{D}}) + \frac{(1-p)(1-\pi)}{np}, \tag{2.4}$$

where $\text{Var}(\hat{\pi}_{\text{D}}) \hat{=} \pi(1-\pi)/n$ denotes the variance of $\hat{\pi}_{\text{D}}$ in *design of direct questioning* (DDQ). It is clear that when $p = 1$ the variant of the parallel design will reduce to the DDQ. Furthermore, we observed that the $\text{Var}(\hat{\pi}_{\text{V}})$ does not depend on the unknown parameter θ. Hence, for any fixed π,

$$n\text{Var}(\hat{\pi}_{\text{V}}) = \pi(1-\pi) + \frac{(1-p)(1-\pi)}{p} \tag{2.5}$$

is a decreasing function of p as shown in Figure 1. We can see that $n\text{Var}(\hat{\pi}_{\text{V}}) \to \infty$ as $p \to 0$.

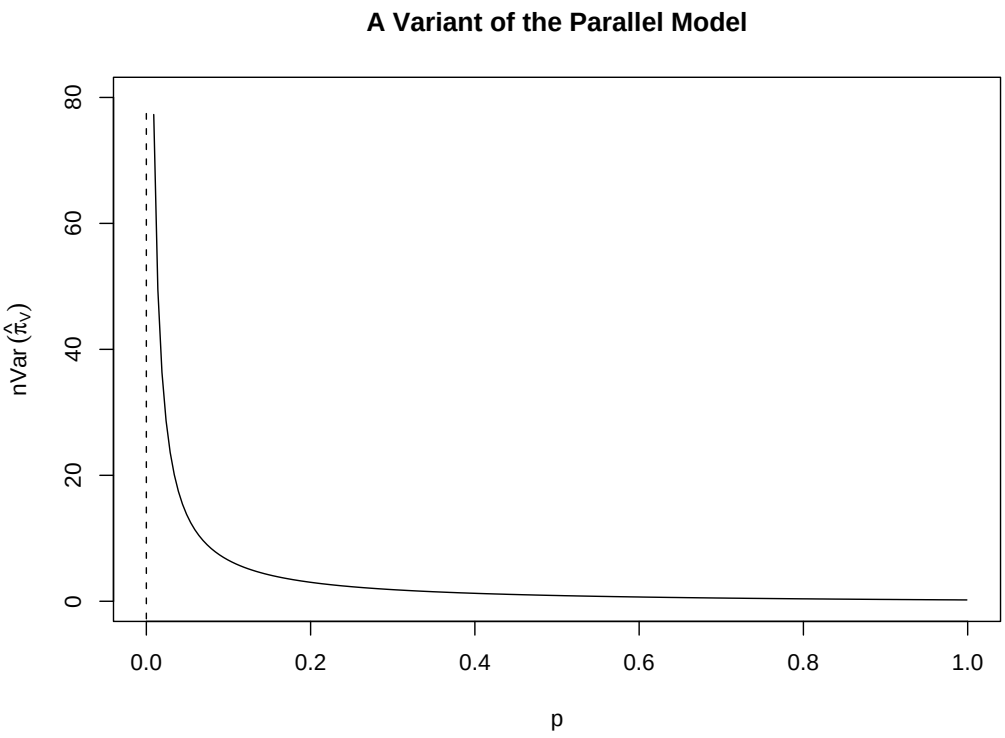

Figure 1 Plot of $n\mathrm{Var}(\hat{\pi}_{\mathrm{V}})$ defined by (2.5) against p with $\pi = 0.3$ for the variant of the parallel model.

2.3. Relative efficiency

The *relative efficiency* (RE) is a useful tool to compare two survey designs. The RE of the the variant of the parallel design to the DDQ is defined by

$$\mathrm{RE}_{\mathrm{V}\to\mathrm{D}}(\pi, p) = \frac{\mathrm{Var}(\hat{\pi}_{\mathrm{V}})}{\mathrm{Var}(\hat{\pi}_{\mathrm{D}})} = 1 + \frac{1-p}{\pi p}.$$

It is noted that $\mathrm{RE}_{\mathrm{V}\to\mathrm{D}}(\pi, p)$ does not depend on the unknown parameter θ and the sample size n. When p is fixed, $\mathrm{RE}_{\mathrm{V}\to\mathrm{D}}(\pi, p)$ is a decreasing function of π. Similarly, when π is fixed, $\mathrm{RE}_{\mathrm{V}\to\mathrm{D}}(\pi, p)$ is also a decreasing function of p. Table 2 lists the values of $\mathrm{RE}_{\mathrm{V}\to\mathrm{D}}(\pi, p)$ for various combinations of π and p. For example, when $\pi = 0.10$ and $p = 2/3$, we have $\mathrm{RE}_{\mathrm{V}\to\mathrm{D}}(0.10, 2/3) = 6$, which implies that the sample size needed for the variant of the parallel design is about 6 times of that needed for the DDQ in order to achieve the same estimation precision. When $\pi = 0.10$ and $p = 0.50$, we have $\mathrm{RE}_{\mathrm{V}\to\mathrm{D}}(0.10, 0.50) = 11$. This might be a drawback for a social researcher who is willing to investigate a sensitive

topic being forced to interview 1100 respondents via the proposed model instead of only 100 respondents using a direct questioning technique. Although a direct questioning technique requires a relatively smaller sample size, the respondents are, in general, not willing to cooperate because of highly sensitive topics. Therefore, in order to smooth the research, some acceptable sacrifice is worthwhile. In other words, with a larger sample size is the cost when an investigator uses an RRT/UCT or an NRR model to implement a survey with sensitive questions.

Table 2. Relative efficiency $\mathrm{RE}_{\mathrm{V}\to\mathrm{D}}(\pi, p)$ for various combinations of π and p

π	p				
	1/3	0.40	0.50	0.60	2/3
0.05	41.000	31.000	21.000	14.333	11.000
0.10	21.000	16.000	11.000	7.6667	6.0000
0.20	11.000	8.5000	6.0000	4.3333	3.5000
0.30	7.6667	6.0000	4.3333	3.2222	2.6667
0.40	6.0000	4.7500	3.5000	2.6667	2.2500
0.50	5.0000	4.0000	3.0000	2.3333	2.0000
0.60	4.3333	3.5000	2.6667	2.1111	1.8333
0.70	3.8571	3.1429	2.4286	1.9524	1.7143
0.80	3.5000	2.8750	2.2500	1.8333	1.6250
0.90	3.2222	2.6667	2.1111	1.7407	1.5556
0.95	3.1053	2.5789	2.0526	1.7018	1.5263

2.4. Degree of privacy protection

To evaluate how the respondent's privacy is protected, we investigate the *degree of privacy protection* (DDP) for the variant of the parallel model. Define

$$Y^{\mathrm{V}} = \begin{cases} -1, & \text{if a tick is put in the circle,} \\ 0, & \text{if a tick is put in the triangle,} \\ 1, & \text{if a tick is put in the upper square.} \end{cases}$$

Let $\mathrm{DPP}_{\bigcirc}(\pi, \theta, p)$ (or $\mathrm{DPP}_{\triangle}(\pi, \theta, p)$) denote the conditional probability of a respondent belonging to the sensitive class $\{Y = 1\}$ given that a tick is put in the circle (or triangle) in

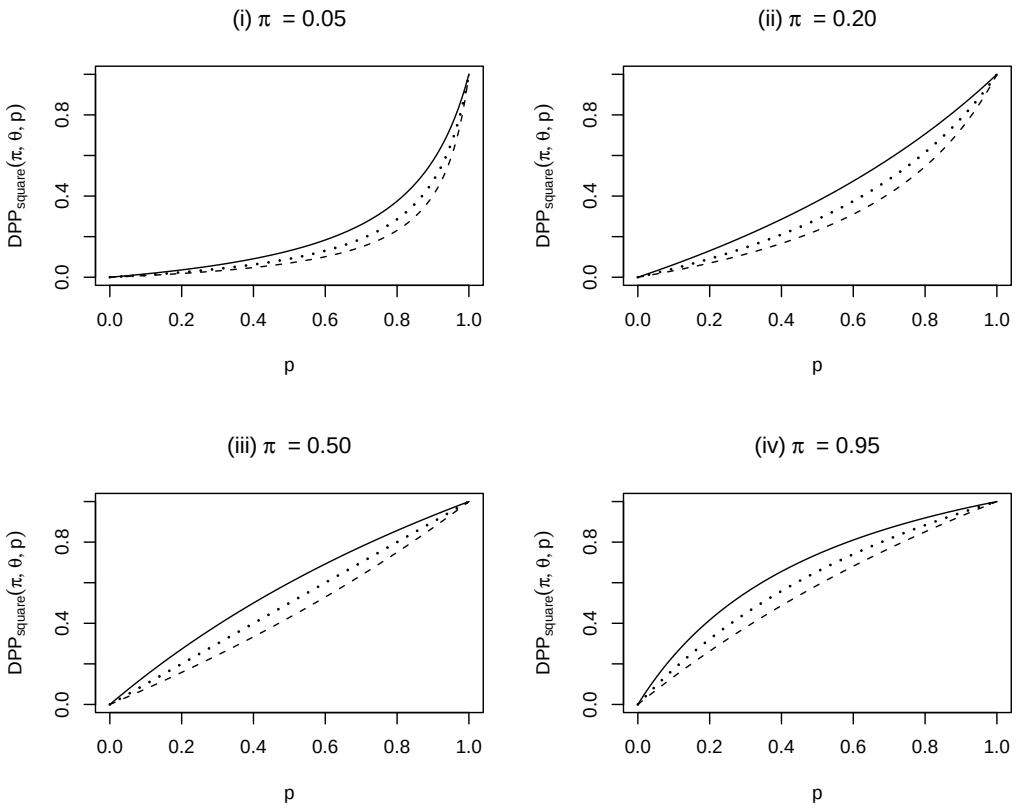

Figure 2 Plots of $\mathrm{DPP}_{\square}(\pi, \theta, p)$ defined by (2.6) against p for the variant of the parallel model with a fixed π and three different values of θ, where the solid line is corresponding to $\theta = 1/3$; the dashed line is corresponding to $\theta = 0.5$; and the dotted line is corresponding to $\theta = 2/3$. (i) $\pi = 0.05$; (ii) $\pi = 0.20$; (iii) $\pi = 0.50$; (iv) $\pi = 0.95$.

Table 1. Clearly, we have

$$\begin{aligned} \mathrm{DPP}_{\bigcirc}(\pi, \theta, p) &= \Pr(Y = 1 | Y^{\mathrm{V}} = -1) = 0 \quad \text{and} \\ \mathrm{DPP}_{\triangle}(\pi, \theta, p) &= \Pr(Y = 1 | Y^{\mathrm{V}} = 0) = 0. \end{aligned}$$

Similarly, let $\mathrm{DPP}_{\square}(\pi, \theta, p)$ represent the conditional probability of a respondent belonging to the sensitive class when a tick is put in the upper square in Table 1, we obtain

$$\mathrm{DPP}_{\square}(\pi, \theta, p) = \Pr(Y = 1 | Y^{\mathrm{V}} = 1) = \frac{\pi p}{\pi p + \theta(1 - p)}. \tag{2.6}$$

In particular, when $p = 1$, we have $\mathrm{DPP}_{\square}(\pi, \theta, 1) = 1$, which equals to the DPP for the DDQ. For any fixed π and θ, $\mathrm{DPP}_{\square}(\pi, \theta, p)$ is a monotonically increasing function of p. Each plot in Figure 2 shows three curves (corresponding to $\theta = 1/3$, 0.5 and 2/3) of $\mathrm{DPP}_{\square}(\pi, \theta, p)$ against p with a fixed π, where $\pi = 0.05$, 0.20, 0.50 and 0.95, respectively.

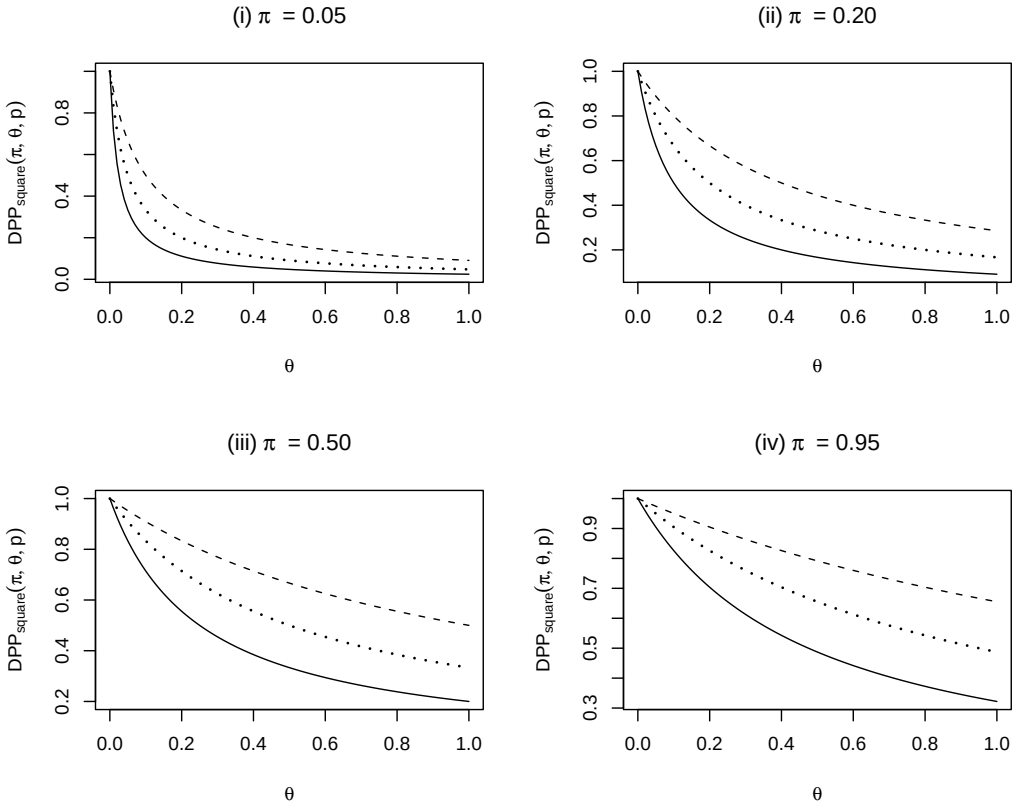

Figure 3 Plots of $\mathrm{DPP}_{\square}(\pi, \theta, p)$ defined by (2.6) against θ for the variant of the parallel model with a fixed π and three different values of p, where the solid line is corresponding to $p = 1/3$; the dashed line is corresponding to $p = 0.5$; and the dotted line is corresponding to $p = 2/3$. (i) $\pi = 0.05$; (ii) $\pi = 0.20$; (iii) $\pi = 0.50$; (iv) $\pi = 0.95$.

In addition, for any fixed π and p, $\mathrm{DPP}_{\square}(\pi, \theta, p)$ is a monotonically decreasing function of θ. Each plot in Figure 3 shows three curves (corresponding to $p = 1/3$, 0.5 and 2/3) of $\mathrm{DPP}_{\square}(\pi, \theta, p)$ against θ with a fixed π, where $\pi = 0.05$, 0.20, 0.50 and 0.95, respectively.

3. Statistical inferences on π

First, we provide an unbiased estimator of the variance of $\hat{\pi}_{\mathrm{V}}$ in Theorem 1 below. Second, we construct three asymptotic confidence intervals (i.e., Wald, Wilson and likelihood ratio CIs) of π by using this unbiased estimator. Third, the exact or Clopper–Pearson CI of π is also derived. Finally, a modified MLE of π is presented and the corresponding asymptotic property is investigated.

3.1. An unbiased estimator of the variance of $\hat{\pi}_{\rm V}$

Theorem 1. Let $\widehat{\rm Var}(\hat{\pi}_{\rm V}) = \hat{\lambda}_2(1-\hat{\lambda}_2)/[(n-1)p^2]$. Then, we have

$$\widehat{\rm Var}(\hat{\pi}_{\rm V}) = \frac{\hat{\pi}_{\rm V}(1-\hat{\pi}_{\rm V})}{n-1} + \frac{(1-\hat{\pi}_{\rm V})(1-p)}{(n-1)p} \tag{3.1}$$

and it is an unbiased estimator of ${\rm Var}(\hat{\pi}_{\rm V}) = \lambda_2(1-\lambda_2)/(np^2)$. ¶

PROOF. From (2.3), we know that $\hat{\lambda}_2 = p(1-\hat{\pi}_{\rm V})$, where $\hat{\pi}_{\rm V}$ is given by (2.2). Hence,

$$\begin{aligned}
\widehat{\rm Var}(\hat{\pi}_{\rm V}) &= \frac{\hat{\lambda}_2(1-\hat{\lambda}_2)}{(n-1)p^2} \\
&= \frac{p(1-\hat{\pi}_{\rm V})(1-p+p\hat{\pi}_{\rm V})}{(n-1)p^2} \\
&= \frac{\hat{\pi}_{\rm V}(1-\hat{\pi}_{\rm V})}{n-1} + \frac{(1-\hat{\pi}_{\rm V})(1-p)}{(n-1)p},
\end{aligned}$$

which implies (3.1). Next, we prove the second part. Since $n_2 \sim {\rm Binomial}(n;\lambda_2)$, we have

$$E(\hat{\lambda}_2) = E(n_2/n) = \lambda_2 \quad \text{and} \quad {\rm Var}(\hat{\lambda}_2) = \frac{{\rm Var}(n_2)}{n^2} = \frac{\lambda_2(1-\lambda_2)}{n},$$

so that

$$E[\hat{\lambda}_2(1-\hat{\lambda}_2)] = E(\hat{\lambda}_2) - [E(\hat{\lambda}_2)]^2 - {\rm Var}(\hat{\lambda}_2) = \frac{(n-1)\lambda_2(1-\lambda_2)}{n}.$$

Thus, we obtain

$$E\Big[\widehat{\rm Var}(\hat{\pi}_{\rm V})\Big] = \frac{E[\hat{\lambda}_2(1-\hat{\lambda}_2)]}{(n-1)p^2} = \frac{\lambda_2(1-\lambda_2)}{np^2},$$

i.e., $\widehat{\rm Var}(\hat{\pi}_{\rm V})$ is an unbiased estimator of ${\rm Var}(\hat{\pi}_{\rm V})$. □

3.2. Three asymptotic confidence intervals of π for large sample sizes

Let z_α denote the upper α-th quantile of the standard normal distribution. From the Central Limit Theorem, as $n \to \infty$, the $(1-\alpha)100\%$ Wald CI of π based on the unbiased estimate $\widehat{\rm Var}(\hat{\pi}_{\rm V})$ is given by

$$[\hat{\pi}_{\rm V,WL},\ \hat{\pi}_{\rm V,WU}] = \left[\hat{\pi}_{\rm V} - z_{\alpha/2}\sqrt{\widehat{\rm Var}(\hat{\pi}_{\rm V})},\ \ \hat{\pi}_{\rm V} + z_{\alpha/2}\sqrt{\widehat{\rm Var}(\hat{\pi}_{\rm V})}\right]. \tag{3.2}$$

One drawback for the Wald CI (3.2) is that the lower bound may be less than zero when the true value of π is close to zero while the upper bound may be beyond one when the true

value of π is near to one. For this situation, we can construct the $(1-\alpha)100\%$ Wilson CI of π based on

$$\begin{aligned}
1-\alpha &= \Pr\left\{\left|\frac{\hat{\pi}_{\mathrm{V}}-\pi}{\sqrt{\mathrm{Var}(\hat{\pi}_{\mathrm{V}})}}\right| \leqslant z_{\alpha/2}\right\} \\
&= \Pr\{(\hat{\pi}_{\mathrm{V}}-\pi)^2 \leqslant z_{\alpha/2}^2 \mathrm{Var}(\hat{\pi}_{\mathrm{V}})\} \\
&\overset{(2.4)}{=} \Pr\left\{(\hat{\pi}_{\mathrm{V}}-\pi)^2 \leqslant \frac{z_{\alpha/2}^2}{n}\left[\pi(1-\pi)+\frac{(1-p)(1-\pi)}{p}\right]\right\} \\
&= \Pr\left\{\hat{\pi}_{\mathrm{V}}^2 - 2\hat{\pi}_{\mathrm{V}}\pi + \pi^2 \leqslant \frac{z_{\alpha/2}^2(-\pi^2+\rho_1\pi+\rho_2)}{n}\right\} \\
&= \Pr\Big\{(1+z_*)\pi^2 - (2\hat{\pi}_{\mathrm{V}}+z_*\rho_1)\pi + \hat{\pi}_{\mathrm{V}}^2 - z_*\rho_2 \leqslant 0\Big\}, \qquad (3.3)
\end{aligned}$$

where $z_* \hat{=} z_{\alpha/2}^2/n$, $\rho_1 \hat{=} 1-\rho_2$ and

$$\rho_2 \hat{=} \frac{1-p}{p}. \qquad (3.4)$$

Solving the quadratic inequality inside the probability in (3.3), we obtain the Wilson (or score) CI of π as follows:

$$[\hat{\pi}_{\mathrm{V,WSL}},\ \hat{\pi}_{\mathrm{V,WSU}}] = \frac{2\hat{\pi}_{\mathrm{V}}+z_*\rho_1 \pm \sqrt{(2\hat{\pi}_{\mathrm{V}}+z_*\rho_1)^2 - 4(1+z_*)(\hat{\pi}_{\mathrm{V}}^2 - z_*\rho_2)}}{2(1+z_*)}, \qquad (3.5)$$

which is, in general, within $[0,1]$. The Wilson CI has been shown to have better performance than the Wald CI and the exact (Clopper–Pearson) CI. See Agresti and Coull (1998), Brown et al. (2001), and Newcombe (1998) for more detail.

When the true value of π is small, the *likelihood ratio confidence interval* (LRCI) could provide better performance than other alternatives. To construct the LRCI of π, we consider the null hypothesis H_0: $\pi = \pi_0$ against the alternative hypothesis H_1: $\pi \neq \pi_0$. Let $\hat{\theta}^R$ denote the restricted MLE of θ under H_0. Then $\hat{\theta}^R = [n_3(1-p) - n_1\pi_0 p]/[(n_1+n_3)(1-p)]$. When $n \to \infty$, it is well known that

$$\Lambda(\pi_0) = -2\{\ell_{\mathrm{V}}(\pi_0, \hat{\theta}^R | Y_{\mathrm{obs}}) - \ell_{\mathrm{V}}(\hat{\pi}_{\mathrm{V}}, \hat{\theta} | Y_{\mathrm{obs}})\} \dot{\sim} \chi^2(1),$$

where $\hat{\pi}_{\mathrm{V}}$ and $\hat{\theta}$ denote the unrestricted MLEs of π and θ specified by (2.2), respectively. Since

$$\Lambda(\pi_0) = -2\Big\{n_1\log(1-\hat{\theta}^R) + n_2\log(1-\pi_0) + n_3\log[\hat{\theta}^R(1-p)+\pi_0 p]$$

$$-n_1 \log(1-\hat{\theta}) - n_2 \log(1-\hat{\pi}_{\rm V}) - n_3 \log[\hat{\theta}(1-p)+\hat{\pi}_{\rm V} p]\Big\}, \tag{3.6}$$

it is easy to verify that $\Lambda(\pi_0)$ is a increasing function of π_0 when $\pi_0 \in \left[0, 1-\frac{n_2}{np}\right]$ and an decreasing function of π_0 when $\pi_0 \in \left[1-\frac{n_2}{np}, 1\right]$. Therefore, for a given significance level α, the $(1-\alpha)100\%$ LRCI for π is given by

$$[\hat{\pi}_{\rm V,LRL},\ \hat{\pi}_{\rm V,LRU}], \tag{3.7}$$

where $\hat{\pi}_{\rm V,LRL}$ and $\hat{\pi}_{\rm V,LRU}$ are two roots of π_0 to the following equation

$$\Lambda(\pi_0) = \chi^2(\alpha, 1), \tag{3.8}$$

where $\chi^2(\alpha, 1)$ denotes the upper α-th quantile of χ^2 distribution with one degree of freedom.

The asymptotic CIs (3.2), (3.5) and (3.7) are appropriate for the cases of large sample sizes. When n is small to moderate, we could use the bootstrap CIs (5.2) and/or (5.3).

3.3. The exact or Clopper–Pearson confidence interval

When the sample size is small to moderate, Clopper and Pearson (1934) proposed a method to calculate the exact confidence limits for the binomial proportion by inverting the equal-tailed test based on the binomial distribution. In this subsection, we employ this method to compute the CI of $\pi = 1-\lambda_2/p$, see (2.3). Note that $n_2 \sim \text{Binomial}(n; \lambda_2)$, the $(1-\alpha)100\%$ exact (or Clopper–Pearson) CI $[\hat{\lambda}_{2,\rm EL}, \hat{\lambda}_{2,\rm EU}]$ of λ_2 satisfy the following equations:

$$\begin{aligned}
\hat{\lambda}_{2,\rm EL} &= 0, \qquad \text{when } n_2 = 0, \\
\sum_{x=n_2}^{n} \binom{n}{x} \hat{\lambda}_{2,\rm EL}^x (1-\hat{\lambda}_{2,\rm EL})^{n-x} &= \frac{\alpha}{2}, \qquad n_2 = 1, \ldots, n-1, \qquad (3.9) \\
\sum_{x=0}^{n_2} \binom{n}{x} \hat{\lambda}_{2,\rm EU}^x (1-\hat{\lambda}_{2,\rm EU})^{n-x} &= \frac{\alpha}{2}, \qquad n_2 = 1, \ldots, n-1 \quad \text{and} \qquad (3.10) \\
\hat{\lambda}_{2,\rm EU} &= 1, \qquad \text{when } n_2 = n.
\end{aligned}$$

By solving (3.9) and (3.10), we obtain

$$\begin{aligned}
\hat{\lambda}_{2,\rm EL} &= \left[1 + \frac{n-n_2+1}{n_2 F(1-\alpha/2; 2n_2, 2(n-n_2+1))}\right]^{-1} \quad \text{and} \\
\hat{\lambda}_{2,\rm EU} &= \left[1 + \frac{n-n_2}{(n_2+1) F(\alpha/2; 2(n_2+1), 2(n-n_2))}\right]^{-1},
\end{aligned}$$

where $F(\alpha; k_1, k_2)$ denotes the upper α-th quantile of the F distribution $F(k_1, k_2)$. Thus, the $(1-\alpha)100\%$ exact CI of π is given by

$$\hat{\pi}_{\text{V,EL}} = 1 - \frac{\hat{\lambda}_{2,\text{EU}}}{p} \quad \text{and} \quad \hat{\pi}_{\text{V,EU}} = 1 - \frac{\hat{\lambda}_{2,\text{EL}}}{p}. \tag{3.11}$$

Because this is a discrete problem, the confidence coefficient (or coverage probability) of the exact CI is not exactly $1-\alpha$ but is at least $1-\alpha$. Thus, this exact CI is conservative.

3.4. A modified MLE of π and its asymptotic property

The MLE of π specified by (2.2) may be beyond the unit interval $[0, 1]$. For example, let $(n_1, n_2, n_3)^\top = (15, 20, 35)^\top$ and $p = 1/4$. From (2.2), we obtain $\hat{\pi}_{\text{V}} = -0.1429 < 0$ and $\hat{\theta} = 0.7143$. For such cases, we can apply an *expectation and maximization* (EM) algorithm (Dempster, Laird and Rubin, 1977) to calculate the MLEs of π and θ. In Section 6.2, we introduced an EM algorithm to find the posterior modes for both π and θ by using two independent beta prior distributions. Especially, when two independent uniform distributions on $[0, 1]$ are adopted as the priors, the posterior modes of π and θ are identical to their MLEs. In (6.7) and (6.8) let $a_1 = b_1 = a_2 = b_2 = 1$, and let $\pi^{(0)} = \theta^{(0)} = 0.5$ be the initial values of π and θ, the EM algorithm converged to $\hat{\pi}_{\text{V}} = 2.22 \times 10^{-17} \approx 0$ and $\hat{\theta} = 0.70$ (< 0.7143) in 197 iterations.

From (2.2), it can be seen that $0 \leqslant \hat{\pi}_{\text{V}} \leqslant 1$ if and only if $0 \leqslant n_2 \leqslant np$. Therefore, a modified MLE of π is

$$\hat{\pi}_{\text{VM}} = \max\{0, \hat{\pi}_{\text{V}}\} = \begin{cases} 0, & \text{if } n_2 > np, \\ \hat{\pi}_{\text{V}}, & \text{if } n_2 \leqslant np. \end{cases} \tag{3.12}$$

The following result shows that the $\hat{\pi}_{\text{VM}}$ and $\hat{\pi}_{\text{V}}$ are asymptotically equivalent.

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

PROOF. It suffices to show that $\sqrt{n}\,(\hat{\pi}_{\text{VM}} - \pi) - \sqrt{n}\,(\hat{\pi}_{\text{V}} - \pi)$ converges to zero in probability as $n \to \infty$, i.e.,

$$\Pr\{|\sqrt{n}\,(\hat{\pi}_{\text{VM}} - \hat{\pi}_{\text{V}})| > 0\} \to 0, \quad \text{as} \quad n \to \infty. \tag{3.13}$$

When $n_2 \leqslant np$, from (3.12), we have $\hat{\pi}_{\text{VM}} = \hat{\pi}_{\text{V}}$. Hence, (3.13) follows immediately.

Now, we consider the case of $n_2 > np$, i.e.,

$$\hat{\lambda}_2 > p. \tag{3.14}$$

Note that $\hat{\lambda}_2$ is the MLE of $\lambda_2 = (1-\pi)p$, it is natural to have $\Pr\{|\hat{\lambda}_2 - \lambda_2| > \varepsilon\} \to 0$ for any given $\varepsilon > 0$ as $n \to \infty$. Thus, we need only to prove

$$\Pr\{|\sqrt{n}\,(\hat{\pi}_{\rm VM} - \hat{\pi}_{\rm V})| > 0\} \leqslant \Pr\{|\hat{\lambda}_2 - \lambda_2| > \varepsilon\} \tag{3.15}$$

for any $\varepsilon < \pi p = p - \lambda_2$. Since $\hat{\pi}_{\rm VM} = 0$, we have

$$\begin{aligned} |\sqrt{n}\,(\hat{\pi}_{\rm VM} - \hat{\pi}_{\rm V})| > 0 \;&\Rightarrow\; |\sqrt{n}\,[0 - (1 - \hat{\lambda}_2/p)]| > 0 \Rightarrow |\hat{\lambda}_2 - p| > 0 \\ &\Rightarrow\; 0 < |\hat{\lambda}_2 - p| \overset{(3.14)}{=} \hat{\lambda}_2 - p = (\hat{\lambda}_2 - \lambda_2) - (p - \lambda_2) \\ &\Rightarrow\; |\hat{\lambda}_2 - \lambda_2| \geqslant \hat{\lambda}_2 - \lambda_2 > p - \lambda_2 > \varepsilon. \end{aligned}$$

Consequently, (3.15) follows immediately. □

4. Statistical inferences on $\boldsymbol{\theta}$

4.1. Three asymptotic confidence intervals of θ for large sample sizes

From (2.2), the variance of $\hat{\theta}$ is

$$\mathrm{Var}(\hat{\theta}) = \frac{\mathrm{Var}(n_1)}{n^2(1-p)^2} = \frac{\lambda_1(1-\lambda_1)}{n(1-p)^2}. \tag{4.1}$$

Similar to Theorem 1, it is easy to verify that

$$\widehat{\mathrm{Var}}(\hat{\theta}) = \frac{\hat{\lambda}_1(1-\hat{\lambda}_1)}{(n-1)(1-p)^2}$$

is an unbiased estimator of $\mathrm{Var}(\hat{\theta})$. Based on this unbiased estimator, the $(1-\alpha)100\%$ Wald CI of θ is

$$[\hat{\theta}_{\rm WL},\ \hat{\theta}_{\rm WU}] = \left[\hat{\theta} - z_{\alpha/2}\sqrt{\widehat{\mathrm{Var}}(\hat{\theta})},\ \ \hat{\theta} + z_{\alpha/2}\sqrt{\widehat{\mathrm{Var}}(\hat{\theta})}\right]. \tag{4.2}$$

The $(1-\alpha)100\%$ Wilson CI of θ can be constructed based on

$$\begin{aligned} 1-\alpha \;&=\; \Pr\left\{\left|\frac{\hat{\theta}-\theta}{\sqrt{\mathrm{Var}(\hat{\theta})}}\right| \leqslant z_{\alpha/2}\right\} \\ &\overset{(4.1)}{=}\; \Pr\left\{(\hat{\theta}-\theta)^2 \leqslant \frac{z^2_{\alpha/2}(1-\theta)(1-p)[1-(1-\theta)(1-p)]}{n(1-p)^2}\right\} \\ &=\; \Pr\Big\{(1+z_*)\theta^2 - (2\hat{\theta} + 2z_* - z_*\rho_3)\theta + \hat{\theta}^2 + z_* - z_*\rho_3 \leqslant 0\Big\}, \end{aligned} \tag{4.3}$$

where $z_* \hat{=} z^2_{\alpha/2}/n$ and $\rho_3 \hat{=} 1/(1-p)$. Solving the quadratic inequality inside the probability in (4.3), we obtain the Wilson (or score) CI of π as follows:

$$[\hat{\theta}_{\text{WSL}},\ \hat{\theta}_{\text{WSU}}] = \frac{2\hat{\theta} + 2z_* - z_*\rho_3 \pm \sqrt{(2\hat{\theta} + 2z_* - z_*\rho_3)^2 - 4(1+z_*)(\hat{\theta}^2 + z_* - z_*\rho_3)}}{2(1+z_*)}. \quad (4.4)$$

which is, in general, within $[0, 1]$.

To construct the LRCI of θ, we consider the null hypothesis H_0: $\theta = \theta_0$ against the alternative hypothesis H_1: $\theta \neq \theta_0$. Let $\hat{\pi}^R$ denote the restricted MLE of π under H_0. Then $\hat{\pi}^R = [n_3 p - n_2\theta_0(1-p)]/[(n_2+n_3)p]$. When $n \to \infty$, it is well known that

$$\Lambda(\theta_0) = -2\{\ell_{\text{V}}(\hat{\pi}^R, \theta_0 | Y_{\text{obs}}) - \ell_{\text{V}}(\hat{\pi}_{\text{V}}, \hat{\theta} | Y_{\text{obs}})\} \dot{\sim} \chi^2(1),$$

where $\hat{\pi}_{\text{V}}$ and $\hat{\theta}$ denote the unrestricted MLEs of π and θ specified by (2.2). Since

$$\begin{aligned} \Lambda(\theta_0) \ &= \ -2\Big\{ n_1 \log(1-\theta_0) + n_2 \log(1-\hat{\pi}^R) + n_3 \log[\theta_0(1-p) + \hat{\pi}^R p] \\ &\quad -n_1 \log(1-\hat{\theta}) - n_2 \log(1-\hat{\pi}_{\text{V}}) - n_3 \log[\hat{\theta}(1-p) + \hat{\pi}_{\text{V}} p] \Big\}, \end{aligned} \quad (4.5)$$

it is easy to verify that $\Lambda(\theta_0)$ is a increasing function of θ_0 when $\theta_0 \in \left[0, 1 - \frac{n_1}{n(1-p)}\right]$ and an decreasing function of θ_0 when $\theta_0 \in \left[1 - \frac{n_1}{n(1-p)}, 1\right]$. Therefore, for a given significance level α, the $(1-\alpha)100\%$ LRCI for θ is given by

$$[\hat{\theta}_{\text{LRL}},\ \hat{\theta}_{\text{LRU}}], \quad (4.6)$$

where $\hat{\theta}_{\text{LRL}}$ and $\hat{\theta}_{\text{LRU}}$ are two roots of θ_0 to the following equation

$$\Lambda(\theta_0) = \chi^2(\alpha, 1). \quad (4.7)$$

4.2. The exact or Clopper–Pearson confidence interval

Similar to Section 3.3, the $(1-\alpha)100\%$ exact (or Clopper–Pearson) CI of θ is given by

$$\hat{\theta}_{\text{EL}} = 1 - \frac{\hat{\lambda}_{1,\text{EU}}}{1-p} \quad \text{and} \quad \hat{\theta}_{\text{EU}} = 1 - \frac{\hat{\lambda}_{1,\text{EL}}}{1-p}, \quad (4.8)$$

where

$$\begin{aligned} \hat{\lambda}_{1,\text{EL}} \ &= \ \left[1 + \frac{n - n_1 + 1}{n_1 F(1-\alpha/2; 2n_1, 2(n-n_1+1))}\right]^{-1} \quad \text{and} \\ \hat{\lambda}_{1,\text{EU}} \ &= \ \left[1 + \frac{n - n_1}{(n_1+1) F(\alpha/2; 2(n_1+1), 2(n-n_1))}\right]^{-1}. \end{aligned}$$

4.3. Testing Hypotheses

Sometimes, we may have a certain knowledge on the unknown parameter $\theta = \Pr(U = 1)$ before our investigation. For example, we may define $U = 1$ if the respondent's birthday is in the second half of a month and $U = 0$ otherwise. Usually, we assume that $\theta \approx 0.5$. To test whether or not this assumption is valid, in this subsection, we focus on testing the following hypotheses:

$$H_0\colon\ \theta = \theta_0 \quad \text{against} \quad H_1\colon\ \theta \neq \theta_0. \tag{4.9}$$

4.3.1. Hypothesis test for large sample sizes

Let n_1 represent the number of respondents putting a tick in the circle in Table 1 and X be the corresponding random variable, then $X \sim \text{Binomial}(n; \lambda_1)$. Since $\lambda_1 = (1-\theta)(1-p)$, the null and alternative hypotheses in (4.9) are reduced to

$$H_0^*\colon\ \lambda_1 = \lambda_{10} \quad \text{against} \quad H_1^*\colon\ \lambda_1 \neq \lambda_{10},$$

where $\lambda_{10} = (1-\theta_0)(1-p)$. For large sample sizes, we can use the normal distribution to approximate the binomial distribution. The test statistic and the corresponding z value are given by

$$Z = \frac{X - n\lambda_{10}}{\sqrt{n\lambda_{10}(1-\lambda_{10})}} \quad \text{and} \quad z = \frac{n_1 - n\lambda_{10}}{\sqrt{n\lambda_{10}(1-\lambda_{10})}}.$$

Under H_0^*, we have $Z \dot{\sim} N(0,1)$. Hence, the corresponding p-value is given by

$$p_{\text{v1}} = 2\Pr\{Z > |z|\} = \Pr\{Z^2 > z^2\} = \Pr\{\chi^2(1) > z^2\}, \tag{4.10}$$

where $\chi^2(\nu)$ denotes the chi-square distribution with ν degrees of freedom. When $p_{\text{v1}} \geq \alpha$, we cannot reject the null hypothesis H_0^* (equivalently, H_0) at the α level of significance.

4.3.2. Hypothesis test for small to moderate sample sizes

When the sample size is not too large, we need to compute the exact p-value for testing H_0 against H_1. Note that $X|H_0^* \sim \text{Binomial}(n; \lambda_{10})$, we define

$$\beta_x \hat{=} \Pr(X = x|H_0^*) = \binom{n}{x}\lambda_{10}^x(1-\lambda_{10})^{n-x}, \quad x = 0, 1, \ldots, n.$$

Thus, the exact two-sided p-value is calculated by

$$p_{\text{v2}} = \sum_{x=0}^{n} \beta_x I_{(\beta_x \leqslant \beta_{n_1})}, \tag{4.11}$$

where $I_{(\cdot)}$ denote the indicate function.

5. Bootstrap confidence intervals of the parameters

In Section 3.2, we provided two asymptotic CIs (3.2) and (3.5) of π, which are available only for large sample sizes. Although the exact CI (3.11) is available for small to moderate sample sizes, its performance was shown (Agresti and Coull, 1998) to be even inferior to that of the Wilson CI specified by (3.5). Alternatively, we could employ the bootstrap method to find bootstrap CIs of π for the cases of small to moderate sample sizes. Next, in the beginning of Section 3.4, we mentioned that if the MLE of π calculated by (2.2) is less than zero, then the EM algorithm (6.7) and (6.8) with $a_1 = b_1 = a_2 = b_2 = 1$ can be used to compute the MLEs of π and θ. For such situations, the bootstrap method is also a useful tool to find CIs for an arbitrary function of π and θ, say, $\vartheta = h(\pi, \theta)$.

Let $\hat{\vartheta} = h(\hat{\pi}_{\text{v}}, \hat{\theta})$ denote the MLE of ϑ, where $\hat{\pi}_{\text{v}}$ and $\hat{\theta}$ represent the respective MLEs of π and θ calculated by means of either (2.2) or the EM algorithm (6.7) and (6.8) with $a_1 = b_1 = a_2 = b_2 = 1$. Based on the obtained MLEs $\hat{\pi}_{\text{v}}$ and $\hat{\theta}$, we can generate

$$(n_1^*, n_2^*, n_3^*)^\top \sim \text{Multinomial}(n;\ (1-\hat{\theta})(1-p), (1-\hat{\pi}_{\text{v}})p, \hat{\theta}(1-p) + \hat{\pi}_{\text{v}}p).$$

Having obtained $Y_{\text{obs}}^* = \{n;\, n_1^*, n_2^*, n_3^*\}$, we can calculate a bootstrap replication $\hat{\pi}_{\text{v}}^*$ and $\hat{\theta}^*$ and calculate $\hat{\vartheta}^* = h(\hat{\pi}_{\text{v}}^*, \hat{\theta}^*)$. Independently repeating this process G times, we obtain G bootstrap replications $\{\hat{\vartheta}_g^*\}_{g=1}^G$. Consequently, the standard error, $\text{se}(\hat{\vartheta})$, of $\hat{\vartheta}$ can be estimated by the sample standard deviation of the G replications, i.e.

$$\widehat{\text{se}}(\hat{\vartheta}) = \left\{ \frac{1}{G-1} \sum_{g=1}^{G} [\hat{\vartheta}_g^* - (\hat{\vartheta}_1^* + \cdots + \hat{\vartheta}_G^*)/G]^2 \right\}^{1/2}. \tag{5.1}$$

If $\{\hat{\vartheta}_g^*\}_{g=1}^G$ is approximately normally distributed, a $(1-\alpha)100\%$ bootstrap CI for ϑ is

$$\left[\hat{\vartheta} - z_{\alpha/2} \cdot \widehat{\text{se}}(\hat{\vartheta}),\ \hat{\vartheta} + z_{\alpha/2} \cdot \widehat{\text{se}}(\hat{\vartheta}) \right]. \tag{5.2}$$

Alternatively, if $\{\hat{\vartheta}_g^*\}_{g=1}^G$ is non-normally distributed, a $(1-\alpha)100\%$ bootstrap CI of ϑ can be obtained as

$$[\hat{\vartheta}_{\mathrm{L}},\ \hat{\vartheta}_{\mathrm{U}}], \tag{5.3}$$

where $\hat{\vartheta}_{\mathrm{L}}$ and $\hat{\vartheta}_{\mathrm{U}}$ are the $100(\alpha/2)$ and $100(1-\alpha/2)$ percentiles of $\{\hat{\vartheta}_g^*\}_{g=1}^G$, respectively.

6. Bayesian inferences

In this section, we first derive the joint posterior distribution of π and θ when a certain prior information is available and obtain their posterior moments which have explicit expressions. Second, we utilize the EM algorithm to calculate the posterior modes of π and θ when their posterior distributions are highly skewed. Finally, we generate i.i.d. posterior samples of π and θ via the exact IBF sampling.

6.1. Posterior moments with explicit expressions

By ignoring the normalizing constant and the known factor $(1-p)^{n_1}p^{n_2+n_3}$, we write the kernel of (2.1) as

$$L_{\mathrm{V}}(\pi,\theta|Y_{\mathrm{obs}}) = (1-\theta)^{n_1}(1-\pi)^{n_2}(\theta\rho_2+\pi)^{n_3}, \tag{6.1}$$

where $0<\pi<1$, $0<\theta<1$ and ρ_2 is defined in (3.4). If two independent beta distributions $\mathrm{Beta}(a_1,b_1)$ and $\mathrm{Beta}(a_2,b_2)$ are adopted as the prior distributions of π and θ, respectively, then the joint posterior distribution of π and θ is

$$f(\pi,\theta|Y_{\mathrm{obs}}) = \frac{\pi^{a_1-1}(1-\pi)^{b_1-1}\theta^{a_2-1}(1-\theta)^{b_2-1}\cdot l_{\mathrm{V}}(\pi,\theta|Y_{\mathrm{obs}})}{c_{\mathrm{V}}(a_1,b_1,a_2,b_2;n_1,n_2,n_3)}, \tag{6.2}$$

where the normalizing constant is given by

$$\begin{aligned}
& c_{\mathrm{V}}(a_1,b_1,a_2,b_2;n_1,n_2,n_3)\\
= & \int_0^1\int_0^1 \pi^{a_1-1}(1-\pi)^{b_1-1}\theta^{a_2-1}(1-\theta)^{b_2-1}\cdot l_{\mathrm{V}}(\theta,\pi|Y_{\mathrm{obs}})\,\mathrm{d}\pi\mathrm{d}\theta\\
= & \sum_{i=0}^{n_3}\binom{n_3}{i}\rho_2^i\int_0^1 \pi^{a_1+n_3-i-1}(1-\pi)^{b_1+n_2-1}\mathrm{d}\pi\int_0^1\theta^{a_2+i-1}(1-\theta)^{b_2+n_1-1}\mathrm{d}\theta\\
= & \sum_{i=0}^{n_3}\binom{n_3}{i}\rho_2^i B(a_1+n_3-i,\ b_1+n_2)B(a_2+i,\ b_2+n_1).
\end{aligned} \tag{6.3}$$

Therefore, the r-th posterior moments of π and θ are given by

$$\begin{aligned} E(\pi^r|Y_{\rm obs}) &= \frac{c_{\rm V}(a_1+r,b_1,a_2,b_2;n_1,n_2,n_3)}{c_{\rm V}(a_1,b_1,a_2,b_2;n_1,n_2,n_3)} \quad \text{and} \\ E(\theta^r|Y_{\rm obs}) &= \frac{c_{\rm V}(a_1,b_1,a_2+r,b_2;n_1,n_2,n_3)}{c_{\rm V}(a_1,b_1,a_2,b_2;n_1,n_2,n_3)}, \end{aligned} \tag{6.4}$$

respectively.

6.2. Calculation of the posterior modes via the EM algorithm

The EM algorithm is a useful tool for computing MLEs in the presence of missing or latent data. Let Z denote the number of respondents belonging to the sensitive subclass $\{Y=1, W=1\}$ in Table 1. Since Z is unobservable, it is natural to treat Z as the latent variable. Thus, the likelihood function of π and θ for the complete data $\{Y_{\rm obs}, Z\}$ is

$$\begin{aligned} L_{\rm V}(\pi,\theta|Y_{\rm obs},Z) &= \binom{n}{n_1,n_2,n_3-Z,Z}[(1-\theta)(1-p)]^{n_1}[(1-\pi)p]^{n_2}[\theta(1-p)]^{n_3-Z}(\pi p)^Z, \\ &\propto \pi^Z(1-\pi)^{n_2}\theta^{n_3-Z}(1-\theta)^{n_1}. \end{aligned}$$

Again, the product of two independent beta densities $\text{Beta}(\pi|a_1,b_1)$ and $\text{Beta}(\theta|a_2,b_2)$ is adopted as the joint prior density of π and θ. Hence, the complete-data posterior distribution and the conditional predictive distribution are

$$f(\pi,\theta|Y_{\rm obs},Z) = \text{Beta}(\pi|a_1+Z,b_1+n_2)\times\text{Beta}(\theta|a_2+n_3-Z,b_2+n_1) \quad \text{and} \tag{6.5}$$

$$f(Z|Y_{\rm obs},\pi,\theta) = \text{Binomial}\left(Z\Big|n_3,\ \frac{\pi p}{\theta(1-p)+\pi p}\right), \tag{6.6}$$

respectively. The M-step of the EM algorithm is to calculate the complete-data posterior modes of π and θ as

$$\tilde{\pi}_{\rm V} = \frac{a_1+Z-1}{a_1+b_1+n_2+Z-2} \quad \text{and} \quad \tilde{\theta}_{\rm V} = \frac{a_2+n_3-Z-1}{a_2+b_2+n_1+n_3-Z-2}, \tag{6.7}$$

and the E-step is to replace Z in (6.7) by its conditional expectation

$$E(Z|Y_{\rm obs},\pi,\theta) = \frac{n_3\pi p}{\theta(1-p)+\pi p}. \tag{6.8}$$

6.3. Generation of i.i.d. posterior samples via the exact IBF sampling

In this subsection, we use the exact IBF sampling (Tian et al., 2007a) to generate i.i.d. posterior samples of π and θ. We simply need to identify the conditional support of $Z|(Y_{\text{obs}}, \pi, \theta)$, denoted by $\mathcal{S}_{(Z|Y_{\text{obs}},\pi,\theta)}$, and calculate the weights $\{\omega_k\}_{k=1}^K$ (see Appendix A). From (6.6), we have

$$\mathcal{S}_{(Z|Y_{\text{obs}})} = \mathcal{S}_{(Z|Y_{\text{obs}},\pi,\theta)} = \{z_1, \ldots, z_K\} = \{0, 1, \ldots, n_3\},$$

where $K = n_3 + 1$. Setting $\pi_0 = \theta_0 = 0.5$, from (A.2) and (A.3), we obtain

$$q_k(\pi_0, \theta_0) = \frac{f(Z = z_k|Y_{\text{obs}}, \pi_0, \theta_0)}{f(\pi_0, \theta_0|Y_{\text{obs}}, z_k)}, \tag{6.9}$$

and $\omega_k = q_k(0.5, 0.5)/\sum_{k'=1}^K q_{k'}(0.5, 0.5)$ for $k = 1, \ldots, K$.

7. Comparison with the non-randomized crosswise and triangular models

In this section, we will compare the variant of the parallel model with the non-randomized crosswise and triangular models. The criteria of the difference of variances and the ratio of variances are considered. Theoretical and numerical results are provided.

7.1. Comparison with the crosswise model

7.1.1. The difference of variances

Let $\hat{\pi}_{\text{C}}$ denote the MLE of $\pi = \Pr(Y = 1)$ under the crosswise model with $p = \Pr(W = 1) \neq 0.5$. From (3.7) of Yu et al. (2008) and (2.4), we have

$$\begin{aligned} \text{Var}(\hat{\pi}_{\text{C}}) - \text{Var}(\hat{\pi}_{\text{V}}) &= \frac{p(1-p)}{n(2p-1)^2} - \frac{(1-p)(1-\pi)}{np} \\ &= \frac{1-p}{np(2p-1)^2} h_{\text{CV}}(p|\pi), \quad p \neq 0.5, \end{aligned} \tag{7.1}$$

where $h_{\text{CV}}(p|\pi) \hat{=} (4\pi - 3)p^2 + 4(1-\pi)p + \pi - 1$ is a quadratic function of p for any fixed π ($\pi \neq 3/4$). The discriminant of the $h_{\text{CV}}(p|\pi)$ is given by

$$D(h_{\text{CV}}) = 16(1-\pi)^2 - 4(4\pi - 3)(\pi - 1) = 4(1-\pi) > 0.$$

We then have the following results.

Theorem 3. Let $\pi \in (0,1)$ and $p \in (0,1)$.

(i) When $\pi = 3/4$, the variant of the parallel model is always more efficient than the crosswise model for any $p > 1/4$.

(ii) When $\pi > 3/4$, the variant of the parallel model is more efficient than the crosswise model for any $p \in (p_\pi, 1)$, where

$$p_\pi = \frac{-2(1-\pi) + \sqrt{1-\pi}}{4\pi - 3} \tag{7.2}$$

is a monotonic decreasing function of $\pi \in (3/4, 1)$ and $0 < p_\pi < 0.25$.

(iii) When $\pi < 3/4$, the variant of the parallel model is always more efficient than the crosswise model for any $p \in (p_\pi, 1)$, where p_π defined by (7.2) is a monotonic decreasing function of $\pi \in (0, 3/4)$ and $0.25 < p_\pi < 1/3$. ¶

PROOF. (i) When $\pi = 3/4$, we have $h_{\text{CV}}(p|\pi) = p - 1/4$. Hence, $h_{\text{CV}}(p|\pi) > 0$ if and only if $p > 1/4$. From (7.1), we obtain $\text{Var}(\hat{\pi}_{\text{C}}) > \text{Var}(\hat{\pi}_{\text{V}})$ for $p > 1/4$.

(ii) When $3/4 < \pi < 1$, it can be shown that the equation $h_{\text{CV}}(p|\pi) = 0$ has two roots

$$p_{_L} = \frac{-2(1-\pi) - \sqrt{1-\pi}}{4\pi - 3}$$

and p_π, which is defined by (7.2). It is clear that $p_{_L} < 0$. Since

$$\frac{\mathrm{d}p_\pi}{\mathrm{d}\pi} = \frac{-(2\sqrt{1-\pi} - 1)^2}{2\sqrt{1-\pi}\,(4\pi - 3)^2} < 0, \tag{7.3}$$

p_π is a monotonic decreasing function of π. The infimum of p_π equals to $\lim_{\pi \to 1} p_\pi = 0$ and the supremum of p_π is equal to

$$\lim_{\pi \to 0.75} p_\pi = \lim_{\pi \to 0.75} \frac{2 - \frac{1}{2\sqrt{1-\pi}}}{4} = \frac{1}{4},$$

so that $0 < p_\pi < 0.25$. Thus, $h_{\text{CV}}(p|\pi) > 0$ if and only if $p_\pi < p < 1$.

(iii) When $0 < \pi < 3/4$, it can see that the equation $h_{\text{CV}}(p|\pi) = 0$ has two roots p_π defined by (7.2) and

$$p_{_U}(\pi) = \frac{-2(1-\pi) - \sqrt{1-\pi}}{4\pi - 3} = \frac{2(1-\pi) + \sqrt{1-\pi}}{3 - 4\pi}.$$

Note that

$$\frac{\mathrm{d}p_{U}(\pi)}{\mathrm{d}\pi} = \frac{(2\sqrt{1-\pi}+1)^2}{2\sqrt{1-\pi}\,(4\pi-3)^2} > 0,$$

then $p_{U}(\pi)$ is a monotonic increasing function of $\pi \in (0, 3/4)$, so that $p_{U}(\pi) > p_{U}(0) = 1$. From (7.3), we know that p_{π} is also a monotonic decreasing function of $\pi \in (0, 3/4)$. The infimum of p_{π} is $\lim_{\pi\to 0.75} p_{\pi} = 0.25$ and the supremum of p_{π} is $\lim_{\pi\to 0} p_{\pi} = 1/3$. In other words, we have $0.25 < p_{\pi} < 1/3$. Thus, $h_{\mathrm{CV}}(p|\pi) > 0$ if and only if $p_{\pi} < p < 1$. □

From Theorem 3, we have immediately the following result.

Corollary 1. The variant of the parallel model is always more efficient than the non-randomized crosswise model for any $\pi \in (0, 1)$ and $p > 1/3$. ¶

7.1.2. Relative efficiency of the crosswise model to the variant of the parallel model

The RE of the crosswise model ($p \neq 0.5$) to the variant of the parallel model is

$$\mathrm{RE}_{\mathrm{C}\to\mathrm{V}}(\pi, p) = \frac{\mathrm{Var}(\hat{\pi}_{\mathrm{C}})}{\mathrm{Var}(\hat{\pi}_{\mathrm{V}})} = \frac{\pi(1-\pi) + p(1-p)/(2p-1)^2}{\pi(1-\pi) + (1-\pi)(1-p)/p},$$

which is independent of the sample size n.

Table 3. Relative efficiency $\mathrm{RE}_{\mathrm{C}\to\mathrm{V}}(\pi, p)$ for various combinations of π and p

π	p					
	1/3	0.40	0.45	0.55	0.60	2/3
0.05	1.0513	4.1070	20.5174	30.0659	8.8825	3.9187
0.10	1.1058	4.2292	20.8739	30.0594	8.8261	3.8704
0.20	1.2273	4.5294	21.8936	30.5815	8.8846	3.8571
0.30	1.3727	4.9286	23.4244	31.8885	9.1773	3.9464
0.40	1.5556	5.4737	25.6747	34.1903	9.7500	4.1481
0.50	1.8000	6.2500	29.0320	37.9310	10.714	4.5000
0.60	2.1538	7.4286	34.2851	44.0529	12.316	5.0909
0.70	2.7284	9.4091	43.2832	54.8024	15.146	6.1389
0.80	3.8571	13.391	61.5907	76.9691	21.000	8.3077
0.90	7.2069	25.375	117.047	144.571	38.872	14.929
0.95	13.881	49.367	228.315	280.486	74.814	28.241

Table 3 reports some values of $\mathrm{RE}_{\mathrm{C}\to\mathrm{V}}(\pi, p)$ for various combinations of π and p. For example, when $\pi = 0.95$ and $p = 0.55$, we have $\mathrm{RE}_{\mathrm{C}\to\mathrm{V}}(0.95, 0.55) = 280.486$, which implies that the efficiency of the variant of the parallel model greatly outweighs that of the crosswise model. When $\pi = 0.80$ and $p = 0.60$, we have $\mathrm{RE}_{\mathrm{C}\to\mathrm{V}}(0.80, 0.60) = 21.000$, implying that the efficiency of the variant of the parallel model is 21 times of that of the crosswise model.

7.2. Comparison with the triangular model

7.2.1. The difference of variances

Let $\hat{\pi}_{\mathrm{T}}$ denote the MLE of $\pi = \Pr(Y = 1)$ under the triangular model with $p = \Pr(W = 1)$. From (3.2) of Tan, Tian and Tang (2009) and (2.4), we have

$$\mathrm{Var}(\hat{\pi}_{\mathrm{T}}) - \mathrm{Var}(\hat{\pi}_{\mathrm{V}}) = \frac{(1-\pi)p}{n(1-p)} - \frac{(1-\pi)(1-p)}{np} = \frac{(1-\pi)(2p-1)}{np(1-p)}, \quad p \in (0,1). \tag{7.4}$$

Theorem 4. For any $\pi \in (0,1)$ and $p \in (0.5,1)$, the variant of the parallel model is always more efficient than the triangular model, i.e., $\mathrm{Var}(\hat{\pi}_{\mathrm{T}}) > \mathrm{Var}(\hat{\pi}_{\mathrm{V}})$. ¶

7.2.2. Relative efficiency of the triangular model to the variant of the parallel model

The RE of the triangular model to the variant of the parallel model is

$$\mathrm{RE}_{\mathrm{T}\to\mathrm{V}}(\pi, p) = \frac{\mathrm{Var}(\hat{\pi}_{\mathrm{T}})}{\mathrm{Var}(\hat{\pi}_{\mathrm{V}})} = \frac{\pi + p/(1-p)}{\pi + (1-p)/p},$$

which is independent with the sample size n.

Table 4. Relative efficiency $\mathrm{RE}_{\mathrm{T}\to\mathrm{V}}(\pi, p)$ for various combinations of π and p

π	p						
	1/3	0.40	0.45	0.5	0.55	0.60	2/3
0.05	0.2683	0.4624	0.6824	1	1.4654	2.1628	3.7273
0.10	0.2857	0.4792	0.6944	1	1.4400	2.0870	3.5000
0.20	0.3182	0.5098	0.7159	1	1.3968	1.9615	3.1429
0.30	0.3478	0.5370	0.7346	1	1.3613	1.8621	2.8750
0.40	0.3750	0.5614	0.7509	1	1.3317	1.7813	2.6667
0.50	0.4000	0.5833	0.7654	1	1.3065	1.7143	2.5000
0.60	0.4231	0.6032	0.7783	1	1.2849	1.6579	2.3636
0.70	0.4444	0.6212	0.7898	1	1.2661	1.6098	2.2500
0.80	0.4643	0.6377	0.8002	1	1.2497	1.5682	2.1538
0.90	0.4828	0.6528	0.8096	1	1.2352	1.5319	2.0714
0.95	0.4915	0.6599	0.8140	1	1.2285	1.5155	2.0345

Table 4 reports some values of $\mathrm{RE}_{\mathrm{T}\to\mathrm{V}}(\pi, p)$ for various combinations of π and p. We have observed from Table 4 that, for any $\pi \in (0, 1)$, $\mathrm{RE}_{\mathrm{T}\to\mathrm{V}}(\pi, p) > 1$ if $p > 0.5$, while $\mathrm{RE}_{\mathrm{T}\to\mathrm{V}}(\pi, p) < 1$ if $p < 0.5$. In other words, when $p > 0.5$, the efficiency of the variant of the parallel model is superior to that of the triangular model and when $p < 0.5$ the efficiency of the variant of the parallel model is inferior to that of the triangular model. In particular, when $p = 0.5$, the efficiency of the two models is equivalent.

8. The non-compliance behavior

The non-compliance behavior encountered in randomized response practice is that some respondents are not willing to follow the design instructions even if interviewers provide them with secret answer sheets, sealed envelopes, and sincere promises of confidentiality Mangat (1994). However, in our opinion, a possible/partial reason for such non-compliance behaviors may be caused by the use of randomizing devices which are, in general, controlled by interviewers. One aim for developing non-randomized response techniques is trying to alleviate the non-compliance behavior. For example, for the crosswise model and the parallel model with two sensitive categories (i.e., both $\{Y = 0\}$ and $\{Y = 1\}$ are sensitive), we in general believe that for those respondents not refusing, they are willing to follow the design instruction since their privacy is well protected. However, for the triangular model, a tick put in the triangle indicates that the respondent may belong to the sensitive class. Thus, the non-compliance behavior may occur in the triangular model. Tang and Wu (2013) proposed two design techniques which incorporate the non-compliance into the non-randomized triangular model.

And actually, the non-compliance behavior can also occur in the proposed variant of the parallel model. We note that only the sub-category $\{Y = 1, W = 1\}$ (i.e. the lower square in Table 1) contains sensitive information. Respondents belonging to this sub-category and having no sufficient confidence on such a survey may put a tick in the triangle in Table 1, resulting in the non-compliance. Taking the non-compliance into consideration, we denote the probability of the respondents who have the sensitive characteristic and belong to $\{W = 1\}$ following the design instruction in Table 1 by ω. Because the new parameter

ω is added, the respondents need be randomly assigned into one of two groups. For the first group, we utilize the variant of parallel model with two non-sensitive binary variates W and U and the sensitive binary variate Y. However, for the second group, we employ the parallel model (Tian, 2012) with the same W, U and Y to estimate the sensitive proportion $\pi = \Pr\{Y = 1\}$.

Suppose that in the first group, we observed n_{11}, n_{12} and $n_{13}\,(n_1 = n_{11} + n_{12} + n_{13})$ respondents put ticks in the circle, triangle and upper square, respectively. Thus, the cell probabilities for the three categories are given by

$$\begin{aligned}
\lambda_1^* &= \Pr\{U = 0, W = 0\} = (1-\theta)(1-p), \\
\lambda_2^* &= \Pr\{Y = 0, W = 1\} = (1-\pi\omega)p \quad \text{and} \\
\lambda_3^* &= \Pr\{U = 1, W = 0\} + \Pr\{Y = 1, W = 1\} = \theta(1-p) + \pi\omega p.
\end{aligned}$$

From the first equation, the MLE of θ is

$$\hat{\theta} = 1 - \frac{n_{11}}{(1-p)n_1}. \tag{8.1}$$

From the second/third equation, it is clear that only $\pi\omega$ is estimable. The corresponding estimate is $1 - n_{12}/(pn_1)$. This is why we need the second group. Assume that in the second group, we observed n_{21} and $n_{22}\,(n_2 = n_{21} + n_{22})$ individuals put ticks in the upper circle and upper square, respectively. Then, the MLEs of π and ω are given by

$$\hat{\pi}_{\text{P}} = \frac{n_{22}/n_2 - \hat{\theta}(1-p)}{p} \quad \text{and} \quad \hat{\omega} = \frac{1}{\hat{\pi}_{\text{P}}}\left(1 - \frac{n_{12}}{pn_1}\right), \tag{8.2}$$

respectively. If at least one of the values of $\hat{\theta}$, $\hat{\pi}_{\text{P}}$ and $\hat{\omega}$ are beyond the unit interval $[0, 1]$, we need employ the EM algorithm to calculate the corresponding MLEs.

9. Illustrative and real examples

9.1. An illustrative example of sexual practices

As a sensitive topic, talking about individual sexual practices is still embarrassing even in countries with open minds. Consequently, it is very difficult to estimate the average numbers of sexual partners or the cell probabilities of having x ($x \leqslant 1$ or $x \geqslant 2$) sexual partners in a

targeting population based on survey data from direct questionnaires. However, gathering information from this kind of sensitive topic plays a crucial role in assisting researchers to investigate the relationship between sexual behaviors and some diseases such as cervical cancer or AIDS. Consider a subset of the sexual practice data from the study of Monto (2001), in which participants were all men arrested for trying to hire prostitutes in three Western cities (San Francisco, Las Vegas and Portland, Oregon) of the United States. From participants' background characteristics shown in Table 1 of Monto (2001), we can see that 343 individuals graduated at most from some high school and 927 individuals received at least some college training. Also, there are 593 respondents having no more than one sexual partner and 668 respondents having no less than two sexual partners.

Table 5. Survey data from Monto (2001)

Level of education	The number of sexual partners		Total
	$Y = 0$ ($\leqslant 1$)	$Y = 1$ (> 2)	
$U = 0$	160 (m_1)	180 (m_2)	340
$U = 1$	433 (m_3)	488 (m_4)	921
Total	593	668	1261

To demonstrate the proposed design for the variant of the parallel model presented in Table 1, we define $Y = 1$ if the respondent has at least two sexual partners and $Y = 0$ otherwise. To estimate the unknown proportion $\pi = \Pr\{Y = 1\}$, we employ two non-sensitive binary variables U and W, where $U = 1$ if the respondent received at least some college training and $U = 0$ otherwise; and $W = 1$ if the respondent's birthday is from September to December and $W = 0$ otherwise. Thus, it is reasonable to assume that $p = \Pr\{W = 1\} \approx \frac{1}{3}$.

First, we need to verify the independence between the level of education and the number of sexual partners. Table 5 displays the survey data of Monto (2001). The MLE of the odds ratio is given by

$$\hat{\psi} = \frac{m_1 m_4}{m_2 m_3} = 1.001796.$$

We would like to test the null hypothesis H_0: $\psi = 1$ against the alternative hypothesis H_1: $\psi \neq 1$. The corresponding p-value is

$$p\text{-value} = 2\,\Pr\left(Z < \frac{-|L|}{\text{se}}\right) = 0.9887377,$$

where Z denote the standard normal random variable, $L = \log[m_1 m_4/(m_2 m_3)]$ and $\text{se} = \sqrt{1/m_1 + 1/m_2 + 1/m_3 + 1/m_4}$. Since the p-value $= 0.9887377 \gg 0.05$, we strongly believe that there is no association between the level of education and the number of sexual partners.

As a result, the observed data can be constructed as

$$\begin{aligned}(n_1, n_2, n_3)^\top &= (343 \times (1-p),\ 593 \times p,\ 927 \times (1-p) + 668 \times p)^\top \\ &\approx (229,\ 198,\ 841)^\top,\end{aligned}$$

where $n = n_1 + n_2 + n_3 = 1268$.

Table 6. Six 95% confidence intervals of π

Type of CIs	95% CI	Width
Wald CI (3.2)	[0.4715823, 0.5915091]	0.1199268
Wilson CI (3.5)	[0.4684999, 0.5883603]	0.1198604
Likelihood ratio CI (3.7)	[0.4695520, 0.5893770]	0.1198250
Exact CI (3.11)	[0.4680426, 0.5902233]	0.1221806
Normal-based bootstrap CI (5.2)	[0.4716088, 0.5914962]	0.1198874
Nonnormal-based bootstrap CI (5.3)	[0.4700315, 0.5906940]	0.1206625

Table 7. Six 95% confidence intervals of θ

Type of CIs	95% CI	Width
Wald CI (4.2)	[0.6973280, 0.7608739]	0.06354587
Wilson CI (4.4)	[0.6959085, 0.7593993]	0.06349080
Likelihood ratio CI (4.6)	[0.6963906, 0.7598780]	0.06348745
Exact CI (4.8)	[0.6956505, 0.7603133]	0.06466284
Normal-based bootstrap CI (5.2)	[0.6972551, 0.7609398]	0.06368466
Nonnormal-based bootstrap CI (5.3)	[0.6971609, 0.7610410]	0.06388013

According to (2.2), the MLEs of π and θ are given by $\hat{\pi}_{\rm V} = 0.5315$ and $\hat{\theta} = 0.7291$. Six 95% confidence intervals of π based on (3.2), (3.5), (3.7), (3.11), (5.2) and (5.3) are shown in Table 6. Similarly, six 95% confidence intervals of θ based on (4.2), (4.4), (4.6), (4.8), (5.2) and (5.3) are shown in Table 7.

Suppose that we want to test the null hypothesis H_0: $\theta = \theta_0 = 0.73$ against the alternative hypothesis H_1: $\theta \neq 0.73$. Let $\alpha = 0.05$, from (4.10) and (4.11), we have $p_{\rm v1} = 0.9557$ and $p_{\rm v2} = 0.9418$. Since both p-values are larger than 0.05, we cannot reject H_0. If we set $\theta_0 = 0.69$, then $p_{\rm v1} = 0.0219$ and $p_{\rm v2} = 0.0220$. As a result, the H_0 should be rejected at the level of $\alpha = 0.025$.

In the setting of Bayesian analysis, we adopt two independent uniform distributions (i.e., $a_1 = b_1 = a_2 = b_2 = 1$) as the prior distributions of π and θ, respectively. Using $\pi^{(0)} = \theta^{(0)} = 0.5$ as the initial values, the EM algorithm specified by (6.7) and (6.8) converged to the posterior modes $\tilde{\pi} = 0.5315$ and $\tilde{\theta} = 0.7291$ in 19 iterations.

Based on (6.9), we employ the IBF sampling to generate L = 20,000 i.i.d. posterior samples of π and θ. The posterior means, the posterior standard deviations and 95% Bayesian credible intervals of π and θ are given in the third, fourth and fifth columns of Table 8. Figure 4 shows the posterior densities of π and θ and their histograms.

Table 8. Posterior estimates of parameters for the data of sexual practices

Parameter	Posterior mode	Posterior mean	Posterior std	95% Bayesian credible interval
π	0.5315	0.5302	0.0303	[0.4688, 0.5881]
θ	0.7291	0.7285	0.0163	[0.6959, 0.7596]

9.2. A real example of cheat in examinations at HKU

9.2.1. Design and analysis under the assumption of complete compliance

Cheating behavior in examinations in universities and colleges around the world definitely result in unfairness and it has been considered as a sensitive issue in which we can hardly obtain reliable answer via direct asking. To investigate the proportion of undergraduates who have ever cheated in examinations, we used the variant of parallel model to conduct a

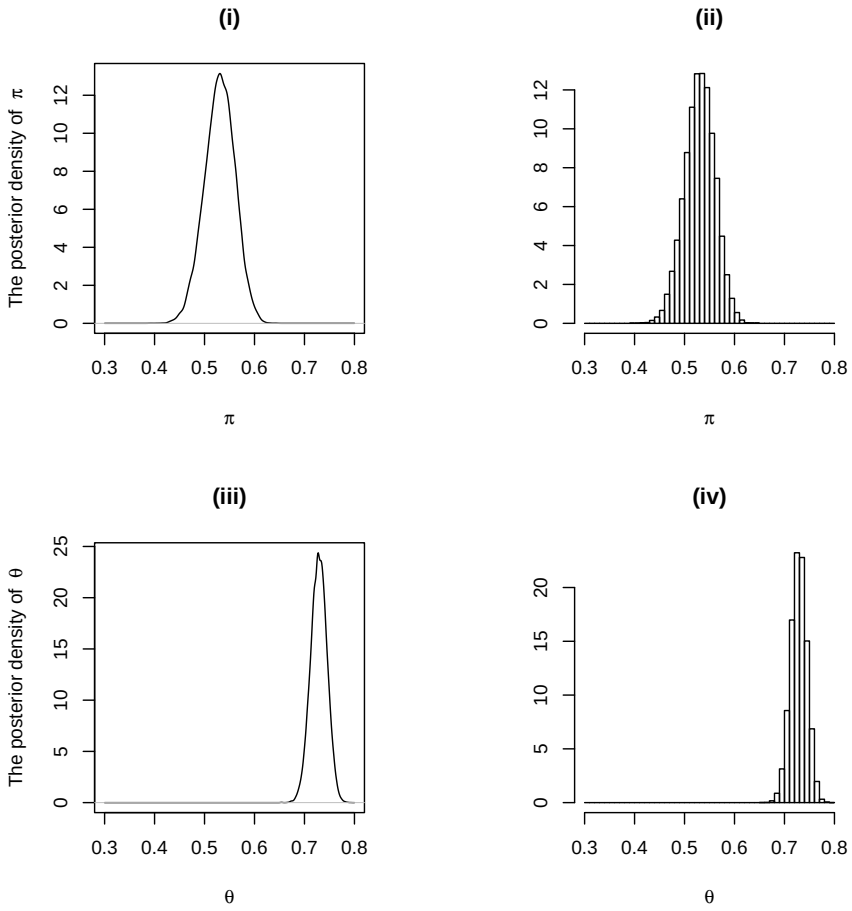

Figure 4 Posterior densities of π and θ via a kernel density smoother based on $L = 20,000$ i.i.d. posterior samples generated by the IBF sampling with two independent uniform distributions on $(0, 1)$ as the prior distributions of π and θ. (i) The posterior density of π; (ii) the histogram of π; (iii) the posterior density of θ; (iv) the histogram of θ.

survey in March 2013 among 150 undergraduates at the University of Hong Kong (HKU) in Hong Kong, P. R. China. The questions in the questionnaire are as follows:

- If your birthday is in the first half of the year and you are not a Hong Kong permanent resident, please circle 1;
- If your birthday is in the second half of the year and you had never cheated in examinations at HKU, please circle 2;
- If your birthday is in the first half of the year and you are a Hong Kong permanent resident OR if your birthday is in the second half of the year and you had ever cheated in examinations at HKU, please circle 3.

At the end of the data collection, 115 students (52 female and 63 male) returned the completed questionnaire, where 1 student was from the Faculty of Arts, 22 were from the Faculty of Business and Economics, 2 were from the Faculty of Engineering, 89 are from the Faculty of Science and 1 student did not tell us the name of his/her faculty. Among these students, 99 were Year 1 students, 2 were Year 2 student, 13 were Year 3 students and 1 was Year 4 student. It was observed that 22 circles on 1, 54 circles on 2 and 39 circles on 3. Let $\pi = \Pr(Y = 1)$ denote the unknown proportion of undergraduates with cheating behavior in examinations at HKU and $\theta = \Pr(U = 1)$ denote the unknown proportion of undergraduates with Hong Kong permanent residents. The observed data can be represented by

$$Y_{\text{obs}} = \{n;\ n_1,\ n_2,\ n_3\} = \{115; 22, 54, 39\}.$$

According to (2.2), the MLEs of π and θ are given by $\hat{\pi}_{\text{V}} = 0.0609$ and $\hat{\theta} = 0.6174$. Six 95% confidence intervals of π based on (3.2), (3.5), (3.7), (3.11), (5.2) and (5.3) are shown in Table 9. Similarly, six 95% confidence intervals of θ based on (4.2), (4.4), (4.6), (4.8), (5.2) and (5.3) are shown in Table 10.

Table 9. Six 95% confidence intervals of π

Type of CIs	95% CI	Width
Wald CI (3.2)	$[-0.1223575,\ 0.2440966]$	0.3664541
Wilson CI (3.5)	$[-0.1205648,\ 0.2383688]$	0.3589336
Likelihood ratio CI (3.7)	$[-0.1213907,\ 0.2404458]$	0.3618365
Exact CI (3.11)	$[-0.1297411,\ 0.2482693]$	0.3780104
Normal-based bootstrap CI (5.2)	$[-0.0676406,\ 0.2186684]$	0.2863091
Nonnormal-based bootstrap CI (5.3)	$[6.832142\times10^{-18},\ 0.2347826]$	0.2347826

Table 10. Six 95% confidence intervals of θ

Type of CIs	95% CI	Width
Wald CI (4.2)	$[0.4729868,\ 0.7617958]$	0.2888089
Wilson CI (4.4)	$[0.4546011,\ 0.7402681]$	0.2856670
Likelihood ratio CI (4.6)	$[0.4608817,\ 0.7467020]$	0.2858203
Exact CI (4.8)	$[0.4496279,\ 0.7521093]$	0.3024814
Normal-based bootstrap CI (5.2)	$[0.4725176,\ 0.7512286]$	0.2787111
Nonnormal-based bootstrap CI (5.3)	$[0.4608696,\ 0.7407407]$	0.2798712

Suppose that we want to test the null hypothesis H_0: $\theta = \theta_0 = 0.35$ against the alternative hypothesis H_1: $\theta \neq 0.35$. Let $\alpha = 0.05$, from (4.10) and (4.11), we have $p_{\text{v1}} = 0.0022$ and $p_{\text{v2}} = 0.0019$. As a result, the H_0 should be rejected at the level of $\alpha = 0.01$. If we set $\theta_0 = 0.60$, then $p_{\text{v1}} = 0.8157$ and $p_{\text{v2}} = 0.9073$. Since both p-values are larger than 0.05, we cannot reject H_0.

For the Bayesian analysis, we adopt two independent uniform distributions (i.e., $a_1 = b_1 = a_2 = b_2 = 1$) as the prior distributions of π and θ, respectively. Using $\pi^{(0)} = \theta^{(0)} = 0.5$ as the initial values, the EM algorithm specified by (6.7) and (6.8) converged to the posterior modes $\tilde{\pi} = 0.0609$ and $\tilde{\theta} = 0.6174$ in 133 iterations.

Based on (6.9), we employ the IBF sampling to generate $L = 20{,}000$ i.i.d. posterior samples of π and θ. The posterior means, the posterior standard deviations and 95% Bayesian credible intervals of π and θ are given in the third, fourth and fifth columns of Table 11. Figure 5 shows the posterior densities of π and θ and their histograms.

Table 11. Posterior estimates of parameters for the data of cheating behaviors

Parameter	Posterior mode	Posterior mean	Posterior std	95% Bayesian credible interval
π	0.0609	0.1040	0.0678	[0.0061, 0.2256]
θ	0.6174	0.5977	0.0704	[0.4503, 0.7261]

9.2.2. Design and analysis under the consideration of non-compliance

To account for the non-compliance behavior in the questionnaire, we conducted the second survey by using the parallel model in March 2013 among 100 undergraduates at HKU. The questions in the questionnaire are as follows:

- If your birthday is in the first half of the year and you are not a Hong Kong permanent resident OR if your birthday is in the second half of the year and you had NEVER cheated in examinations at HKU, please circle 'No';

- If your birthday is in the first half of the year and you are a Hong Kong permanent resident OR if your birthday is in the second half of the year and you had EVER cheated in examinations at HKU, please circle 'Yes'.

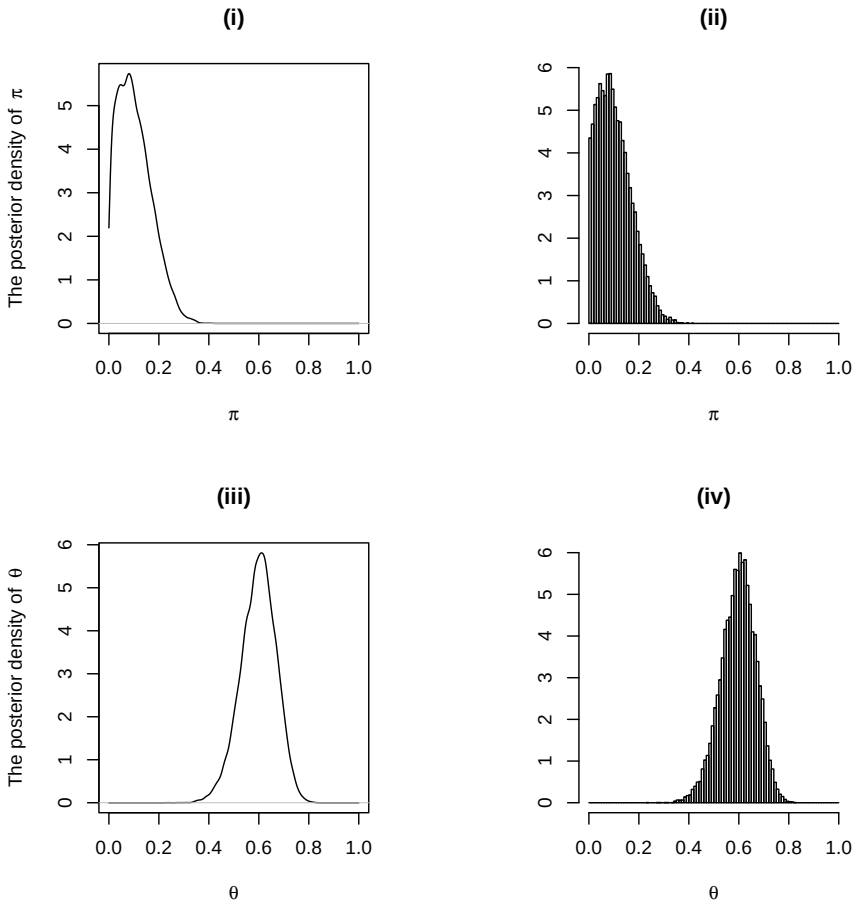

Figure 5 Posterior densities of π and θ via a kernel density smoother based on $L = 20,000$ i.i.d. posterior samples generated by the IBF sampling with two independent uniform distributions on $(0, 1)$ as the prior distributions of π and θ. (i) The posterior density of π; (ii) the histogram of π; (iii) the posterior density of θ; (iv) the histogram of θ.

At the end of the data collection, 77 students (27 female and 50 male) returned the completed questionnaire, where 2 were from the Faculty of Law, 7 were from the Faculty of Business and Economics, 67 are from the Faculty of Science and 1 was from an unknown faculty. Among these students, 43 were Year 1 students, 10 were Year 2 students, 20 were Year 3 students, and 4 were Year 4 students. It was observed that 40 circles on 'No' and 37 circles on 'Yes'. Let $\pi = \Pr(Y = 1)$ denote the unknown proportion of undergraduates with cheating behavior in examinations at HKU, $\theta = \Pr(U = 1)$ denote the unknown proportion of undergraduates being Hong Kong permanent resident, and ω denote the unknown proportion of undergraduates with cheating behavior following the design instruction in Table 1. The observed data can be denoted by

$$Y_{\text{obs}} = \{n_2;\ n_{21},\ n_{22}\} = \{77; 40, 37\}.$$

Then, according to (8.1) and (8.2), we have $\hat{\pi} = 0.3436$, $\hat{\theta} = 0.6174$ and $\hat{\omega} = 0.1771$. The MLE of π obtained from the combined data of two groups is significantly higher than that obtained only from the first sample. Since $\hat{\omega} = 0.1771$, we can see that about $\hat{\pi}(1-\hat{\omega})p = 0.3436 * (1 - 0.1771) * 0.5 = 14.14\%$ students did not follow the instruction of the design for the variant of parallel model in our surveys.

10. Discussion

The paper presents a new development for the parallel model originally proposed by Tian (2012) in sample surveys with sensitive questions. The basic idea is to use two additional non-sensitive binary variates in conjunction with the sensitive binary response variable to create a scenario under which confidentiality of the respondent is preserved and partial information on the sensitive response variable is also obtained. The proposed model assumes that the population mean (proportion) of one of the non-sensitive variates is known but the other one is unknown. This last fact is new and provides certain flexibility in chosen the non-sensitive binary variate. Point and variance estimates of the population proportion of the sensitive response are derived, and several asymptotic confidence intervals are provided. Theoretical and numerical comparisons showed that the proposed variant of the parallel model over-performs two existing NRR crosswise and triangular models for most of the possible parameter ranges as shown in Corollary 1, Theorem 4, Table 3 and Table 4. A possible reason for these conclusions is that the variant of the parallel design can gather exact information (instead of mixing information) for two cells (i.e., the circle and the triangle in Table 1) because of the introduction of an additional non-sensitive binary variate U when comparing with the crosswise and triangular models. Finally, we provide a simple way to handle the possible non-compliance behavior in the proposed model.

One referee pointed out that from the analysis viewpoint (rather than the design viewpoint), the proposed model in this paper is a member of the family of multinomial processing tree models (see, e.g., Erdfelder et al., 2009). Hu and Batchelder (1994) obtained point estimates of parameters by using the EM algorithm and the corresponding standard errors from the Fisher information matrix in multinomial processing tree models. However, we noted

that the resulting interval estimates (based on derivatives from Hu and Batchelder, 1994) in the form of point estimate plus/minus 1.96 times standard error may be beyond the unit interval [0, 1] when the true value of the proportion with the sensitive characteristic is close to zero or one. In fact, in Section 6.2 of this paper we have given the EM algorithm to calculate the posterior modes which are identical to the MLEs of the corresponding parameters π and θ if two independent uniform priors are adopted. In addition, our bootstrap CIs for π and θ in the form of (5.3) are always within the unit interval [0,1].

Acknowledgments

The authors would like to thank the Editor, an Associate Editor and four referees for their helpful comments and useful suggestions. The manuscript has benefitted a lot from the three rounds of revision, especially, the ideas brought up by Reviewer 3. GL Tian's research was fully supported by a grant (HKU 779210M) from the Research Grant Council of the Hong Kong Special Administrative Region.

References

Agresti, A., Coull, B.A., 1998. Approximate is better than "exact" for interval estimation of binomial proportions. The American Statistician 52, 119–126.

Böckenholt, U., van der Heijden, P.G.M., 2007. Item randomized-response models for measuring noncompliance: Risk-return perceptions, social influences, and self-protective responses. Psychometrika 72, 245–262.

Brown, L.D., Cai, T.T., DasGupta, A., 2001. Interval estimation for a binomial proportion. Statistical Science 16, 101–133.

Chaudhuri, A., 2011. Randomized Response and Indirect Questioning Techniques in Surveys. CRC/Chapman & Hall, Boca Raton.

Chaudhuri, A., Mukerjee, R., 1988. Randomized Response: Theory and Technique. Marcel Dekker, New York.

Christofides, T.C., 2005. Randomized response technique for two sensitive characteristics at the same time. Metrika 62, 53–63.

Clark, S.J., Desharnais, R.A., 1998. Honest answers to embarrassing questions: Detecting cheating in the randomized response model. Psychological Methods 3, 160–168.

Clopper, C.J., Pearson, E.S., 1934. The use of confidence or fiducial limits illustrated in the case of binomial. Biometrika 26, 404–413.

Coutts, E., Jann, B., 2011. Sensitive questions in online surveys: Experimental results for the randomized response technique and the unmatched count technique (uct). Sociological Methods & Research 40, 169–193.

Dalton, D.R., Daily, C.M., Wimbush, J.C., 1997. Collecting sensitive data in business ethics research: A case for the unmatched count technique (uct). Journal of Business Ethics 16, 1049–1057.

Dalton, D.R., Wimbush, J.C., Daily, C.M., 1994. Using the unmatched count technique (uct) to estimate base rates for sensitive behavior. Personnel Psychology 47, 817–829.

Droitcour, J., Caspar, R.A., Hubbard, M.L., Parsley, T.L., Visscher, W., Ezzati, T.M., 1991. The item count technique as a method of indirect questioning: A review of its development and a case study application, in: P. P. Biemer, R. M. Groves, L.E.L.N.A.M., Sudman, S. (Eds.), Measurement Errors in Surveys, Wiley, New York. pp. 185–210.

Erdfelder, E., Auer, T.S., Hilbig, B.E., Aßfalg, A., Moshagen, M., Nadarevic, L., 2009. Multinomial processing tree models: A review of the literature. Zeitschrift für Psychologie/Journal of Psychology 217, 108–124.

Fox, J.A., Tracy, P.E., 1986. Randomized Response: A Method for Sensitive Surveys. Quantitative Applications in the Social Sciences, SAGE Publications, Inc., California.

Greenberg, B.G., Abul-Ela, A.A., Simmons, W.R., Horvitz, D.G., 1969. The unrelated

question randomized response model: Theoretical framework. Journal of the American Statistical Association 64, 520–539.

Horvitz, D.G., Shah, B.V., Simmons, W.R., 1967. The unrelated question randomized response model, in: 1967 Social Statistics Section Proceedings of the American Statistical Association, pp. 65–72.

Hu, X., Batchelder, W.H., 1994. The statistical analysis of general processing tree models with the em algorithm. Psychometrika 59, 21–47.

Kuk, A.Y.C., 1990. Asking sensitive questions indirectly. Biometrika 77, 436–438.

Lakshmi, D.V., Raghavarao, D., 1992. A test for detecting untruthful answering in randomized response procedures. Journal of Statistical Planning and Inference 31, 387–390.

Mangat, N.S., 1994. An improved randomized response strategy. Journal of the Royal Statistical Society, B 56, 93–95.

Monto, M.A., 2001. Prostitution and fellatio. The Journal of Sex Research 38, 140–145.

Moshagen, M., 2010. Multitree: A computer program for the analysis of multinomial processing tree models. Behavior Research Methods 42, 42–54.

Moshagen, M., Musch, J., Erdfelder, E., 2012. A stochastic lie detector. Behavior Research Methods 44, 222–231.

Newcombe, R.G., 1998. Improved confidence intervals for the difference between binomial proportions based on paired data. Statistics in Medicine 17, 2635–2650.

Ostapczuk, M., Moshagen, M., Zhao, Z., Musch, J., 2009a. Assessing sensitive attributes using the randomized response technique: Evidence for the importance of response symmetry. Journal of Educational and Behavioral Statistics 34, 267–287.

Ostapczuk, M., Musch, J., Moshagen, M., 2009b. A randomized-response investigation of the education effect in attitudes towards foreigners. European Journal of Social Psychology 39, 920–931.

Ostapczuk, M., Musch, J., Moshagen, M., 2011. Improving self-report measures of medication non-adherence using a cheating detection extension of the randomized-response-technique. Statistical Methods in Medical Research 20, 489–503.

Raghavarao, D., , Federer, W.T., 1979. Block total response as an alternative to the randomized response method in surveys. Journal of the Royal Statistical Society, B 41, 40–45.

Takahasi, K., Sakasegawa, H., 1977. A randomized response technique without making use of any randomizing device. Ann. Inst. Statist. Math. 29, 1–8.

Tan, M., Tian, G.L., Tang, M.L., 2009. Sample surveys with sensitive questions: A non-randomized response approach. The American Statistician 63, 9–16.

Tang, M.L., Tian, G.L., Tang, N.S., Liu, Z.Q., 2009. A new non-randomized multi-category response model for surveys with a single sensitive question: Design and analysis. Journal of the Korean Statistical Society 38, 339–349.

Tang, M.L., Wu, Q., 2013. Non-randomized response model for sensitive survey with non-compliance. Biometrika In press.

Tian, G.L., 2012. A new non-randomized response model: the parallel model. Technical Report. Department of Statistics and and Actuarial Science, The University of Hong Kong.

Tian, G.L., Tan, M., Ng, K.W., 2007a. An exact non-iterative sampling procedure for discrete missing data problems. Statistica Neerlandica 61, 232–242.

Tian, G.L., Tang, M.L., Liu, Z.Q., Tan, M., Tang, N.S., 2011. Sample size determination for the non-randomized triangular model for sensitive questions in a survey. Statistical Methods in Medicine Research 20, 159–173.

Tian, G.L., Yu, J.W., Tang, M.L., Geng, Z., 2007b. A new non-randomized model for analyzing sensitive questions with binary outcomes. Statistics in Medicine 26, 4238–4252.

Tsuchiya, T., Hirai, Y., Ono, S., 2007. A study of the properties of the item count technique. Public Opinion Quarterly 71, 253–272.

Van Den Hout, A., Böckenholt, U., Van Der Heijden, P.G.M., 2010. Estimating the prevalence of sensitive behavior and cheating with a dual design for direct questioning and randomized response. Appl. Statist. 59, 723–736.

Van Den Hout, A., Klugkist, I., 2009. Accounting for non-compliance in the analysis of randomized response data. Australian & New Zealand Journal of Statistics 51, 353–372.

Warner, S.L., 1965. Randomized response: A survey technique for eliminating evasive answer bias. Journal of the American Statistical Association 60, 63–69.

Yu, J.W., Tian, G.L., Tang, M.L., 2008. Two new models for survey sampling with sensitive characteristic: Design and analysis. Metrika 67, 251–263.

Appendix A: The exact IBF sampling

Suppose that both the complete-data posterior distribution $f(\pi, \theta|Y_{\text{obs}}, Z)$ and the conditional predictive distribution $f(Z|Y_{\text{obs}}, \pi, \theta)$ are available. The fundamental conditional sampling principle states that: If we could obtain independent samples $\{Z^{(l)}\}_{l=1}^{L}$ from $f(Z|Y_{\text{obs}})$ and generate $(\pi^{(l)}, \theta^{(l)}) \sim f(\pi, \theta|Y_{\text{obs}}, Z^{(l)})$ for $l = 1, \ldots, L$, then $\{\pi^{(l)}, \theta^{(l)}\}_{l=1}^{L}$ are i.i.d. sammples from the observed posterior distribution $f(\pi, \theta|Y_{\text{obs}})$. In other words, the key issue is to generate independent samples from $f(Z|Y_{\text{obs}})$.

Let $\mathcal{S}_{(\pi,\theta|Y_{\text{obs}})}$ and $\mathcal{S}_{(Z|Y_{\text{obs}})}$ denote the conditional supports of $\pi, \theta|Y_{\text{obs}}$ and $Z|Y_{\text{obs}}$, respectively. The sampling-wise IBF states that

$$f(Z|Y_{\text{obs}}) \propto \frac{f(Z|Y_{\text{obs}}, \pi_0, \theta_0)}{f(\pi_0, \theta_0|Y_{\text{obs}}, Z)}, \tag{A.1}$$

for any arbitrary $(\pi_0, \theta_0) \in \mathcal{S}_{(\pi,\theta|Y_{\text{obs}})}$ and all $Z \in \mathcal{S}_{(Z|Y_{\text{obs}})}$. When Z is a discrete random variable/vector taking finite values on the domain, we denote the conditional support of $Z|(Y_{\text{obs}}, \pi, \theta)$ by $\mathcal{S}_{(Z|Y_{\text{obs}},\pi,\theta)} = \{z_1, \ldots, z_K\}$. Since $f(Z|Y_{\text{obs}}, \pi)$ is available, we can first directly identify $\{z_k\}_{k=1}^{K}$ from the model specification and all $\{z_k\}_{k=1}^{K}$ become known. Noting that $\{z_k\}_{k=1}^{K}$ generally do not depend on π and θ, we have $\mathcal{S}_{(Z|Y_{\text{obs}})} = \mathcal{S}_{(Z|Y_{\text{obs}},\pi,\theta)} = \{z_1, \ldots, z_K\}$. Due to the discreteness of Z, the notation $f(z_k|Y_{\text{obs}})$ will be used to denote

the probability mass function, i.e., $f(z_k|Y_{\text{obs}}) = \Pr\{Z = z_k|Y_{\text{obs}}\}$. Therefore, it suffices to find $\omega_k = f(z_k|Y_{\text{obs}})$ for $k = 1, \ldots, K$. For any $(\pi_0, \theta_0) \in \mathcal{S}_{(\pi,\theta|Y_{\text{obs}})}$, let

$$q_k(\pi_0, \theta_0) = \frac{\Pr\{Z = z_k|Y_{\text{obs}}, \pi_0, \theta_0\}}{f(\pi_0, \theta_0|Y_{\text{obs}}, z_k)}, \quad k = 1, \ldots, K. \tag{A.2}$$

From the sampling-wise IBF (A.1), we immediately obtain

$$\omega_k = \frac{q_k(\pi_0, \theta_0)}{\sum_{k'=1}^{K} q_{k'}(\pi_0, \theta_0)}, \quad k = 1, \ldots, K. \tag{A.3}$$

and $\{\omega_k\}_{k=1}^{K}$ are independent of π_0 and θ_0. Thus, it is easy to sample from $f(Z|Y_{\text{obs}})$ since it is a discrete distribution with probability ω_k on z_k for $k = 1, \ldots, K$. We summarize the algorithm as follows

The exact ibf sampling:

Step 1. Identify $\mathcal{S}_{(Z|Y_{\text{obs}})} = \mathcal{S}_{(Z|Y_{\text{obs}},\pi,\theta)} = \{z_1, \ldots, z_K\}$ from $f(Z|Y_{\text{obs},\pi,\theta})$ and calculate $\{\omega_k\}_{k=1}^{K}$ according to (A.2) and (A.3).

Step 2. Generate i.i.d. samples $\{Z^{(l)}\}_{l=1}^{L}$ of Z from the probability mass function $f(Z|Y_{\text{obs}})$ with probabilities $\{\omega_k\}_{k=1}^{K}$ on $\{z_k\}_{k=1}^{K}$.

Step 3. Generate $(\pi^{(l)}, \theta^{(l)}) \sim f(\pi, \theta|Y_{\text{obs}}, Z^{(l)})$ for $l = 1, \ldots, L$, then $\{\pi^{(l)}, \theta^{(l)}\}_{l=1}^{L}$ are i.i.d. samples from the observed posterior distribution $f(\pi, \theta|Y_{\text{obs}})$.