

Multi-category parallel models in the design of surveys with sensitive questions

YIN LIU AND GUO-LIANG TIAN*

In the past few years, several *non-randomized response* (NRR) designs were introduced in sample surveys with sensitive questions. However, existing NRR models (e.g., the crosswise model, the triangular model, the hidden sensitive model and the multi-category triangular model) have certain limitations in applications, for example, they can only be applied to a situation where at least one of the population categories of interest is non-sensitive. In this paper, we propose a new NRR multi-category parallel model with a better degree of privacy protection and a wider application range, where all population categories of interest can be sensitive or one of them can be totally non-sensitive. Likelihood-based inferences for parameters of interest are developed. In addition, an important special case of the multi-category parallel model is studied to test the association of two sensitive binary variables. Furthermore, theoretic comparisons show that the multi-category parallel model is more efficient than the multi-category triangular model for some cases. An example on the study of association between the number of sex partners and annual income is used to illustrate the proposed method.

KEYWORDS AND PHRASES: Chi-squared test, Likelihood ratio test, Multi-category parallel model, Multi-category triangular model, Non-randomized response technique.

1. INTRODUCTION

In the past few decades, the sample survey technique has played an important role in epidemiological, psychological, medical and social studies and is indispensable in assisting researchers to make statistical inferences and in guiding them to establish a meaningful decision. However, in surveys involving sensitive information or highly private questions (e.g., sexual behavior, drug-taking, tax evasion, cheating on exams, gambling and so on), problems often arise when sensitive questions are asked directly. For example, some respondents may refuse to answer or provide false answers in order to protect their privacy. Statistical inferences based on these inaccurate survey data are in general unreliable.

For a single sensitive question with binary answers ('yes' or 'no'), Warner [15] proposed a *randomized response* (RR)







method for partially overcoming aforementioned problems while protecting respondents' privacy. Such a technique encourages interviewees to provide truthful responses, avoiding a non-response or false answer. Abul-Ela et al. [1] extended the Warner model from the dichotomous case to the multi-chotomous case. Another extension of the Warner model was made by Bourke and Dalenius [3], in which a Latin square design was suggested. In addition, Eriksson [4] proposed an unrelated question RR model, which could be used to estimate the proportions of m (> 2) mutually exclusive sensitive groups (up to $m - 1$ sensitive subclasses) only using one sample. Bourke [5] considered another unrelated question model to estimate the proportions of m mutually exclusive groups with k ($1 \leq k \leq m - 1$) groups containing sensitive information. If the distribution of the unrelated characteristic is known, only one sample is needed. Because of the use of one or two *randomized devices* (RDs), all RR models have some limitations including the lack of reproducibility, of trust, and of cost control.

Recently, without using any RDs some investigators proposed several *non-randomized response* (NRR) models [13, 14, 17, 11], which could overcome some of the limitations with RR designs. Despite greater advances, all NRR models including the multi-category triangular model [11] require that at least one of the population categories of interest be non-sensitive. For example, in some surveys, we may be interested in estimating the proportions of population groups associated with sensitive questions such as the number of sexual partners (≤ 3 , 4-6, >7), or days of illegal drug usage in the last month (≤ 1 , 2, or ≥ 3), and so on. First, as the unique NRR model dealing with the case of m ($m \geq 3$) groups, the existing multi-category triangular model cannot be applied to such situations where each subclass (denoted by $\{Y = i\}$ for $i = 1, \dots, m$) is sensitive. Second, the multi-category triangular model still has a lower efficiency for some cases. Third, the newly developed parallel model of Tian [12] can only deal with the case of $m = 2$ groups where both $\{Y = 0\}$ and $\{Y = 1\}$ could be sensitive. Therefore, these limitations with NRR models motivate us to further develop a new non-randomized multi-category parallel model, which is an extension of the NRR parallel model.

This article is organized as follows. In Section 2, we propose the survey design for the multi-category parallel model with a wider application range. Section 3 presents the *max-*

*Corresponding author.

Table 1. The multi-category parallel model and the corresponding cell probabilities

Category	$W = 0$	$W = 1$	Category	$W = 0$	$W = 1$	Marginal
$U = 1$			$U = 1$	$p_1(1 - q)$		p_1
$U = 2$			$U = 2$	$p_2(1 - q)$		p_2
\vdots	\vdots		\vdots	\vdots		\vdots
$U = m$			$U = m$	$p_m(1 - q)$		p_m
$Y = 1$			$Y = 1$		$\pi_1 q$	π_1
$Y = 2$			$Y = 2$		$\pi_2 q$	π_2
\vdots		\vdots	\vdots		\vdots	\vdots
$Y = m$			$Y = m$		$\pi_m q$	π_m
			Marginal	$1 - q$	q	1

Note: Please truthfully link the two circles by a straightline if you belong to $\{U = 1, W = 0\} \cup \{Y = 1, W = 1\}$, or link the two triangles by a straightline if you belong to $\{U = 2, W = 0\} \cup \{Y = 2, W = 1\}, \dots$, or link the two dots by a straightline if you belong to $\{U = m, W = 0\} \cup \{Y = m, W = 1\}$.

imum likelihood estimates (MLEs) and two bootstrap confidence intervals (CIs) of the parameters of interest for small to moderate sample sizes. In addition, two asymptotic CIs of parameters are also constructed for large sample sizes. An important special case ($m = 4$) for the multi-category parallel design is studied in Section 4. In Section 5, we compare the efficiency between the multi-category parallel model and the multi-category triangular model. In Section 6, an example on the study of association between the number of sex partners and annual income is used to illustrate the proposed method. Finally, we conclude with a discussion in Section 7.

2. THE SURVEY DESIGN FOR THE MULTI-CATEGORY PARALLEL MODEL

Consider a sensitive question Q_Y (e.g., how many sex partners do you have within a certain period?) with m possible answers (e.g., 0–3, 4–6 or ≥ 7), which classify the target population into m mutually exclusive categories and each category has a certain degree of a sensitive attribute. Let Y denote the categorical random variable associated with the question Q_Y and $\{Y = i\}$ denote that a person in the target population belongs to the i -th category ($i = 1, \dots, m$). The purpose here is to estimate the proportions $\pi_i = \Pr\{Y = i\}$ for $i = 1, \dots, m$. Let

$$(2.1) \quad \mathbb{T}_m \hat{=} \left\{ (x_1, \dots, x_m)^\top : x_i \geq 0, i = 1, \dots, m, \sum_{i=1}^m x_i = 1 \right\},$$

obviously, we have $\boldsymbol{\pi} = (\pi_1, \dots, \pi_m)^\top \in \mathbb{T}_m$.

To carry out the survey for which each category includes a sensitive attribute, we require an introduction of the non-sensitive dichotomous variate W and another non-sensitive multichotomous variate U so that the three variables W, U

and Y are mutually independent with known proportions $q = \Pr\{W = 1\}$ and

$$p_i = \Pr\{U = i\}, \quad i = 1, \dots, m.$$

For example, when $m = 4$, we may define $W = 0$ if the respondent's birthday is in the first half of a month and $W = 1$ otherwise. Similarly, we define $U = i$ if the respondent was born in the i -th quarter of a year ($i = 1, \dots, 4$). Hence, it is reasonable to assume that $q \approx 0.5$ and $p_i \approx 0.25$ for each i . Some practical guidelines in choosing the two non-sensitive variables W and U are given in Section 7. 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 link the two circles by a straight line if he/she belongs to one of the two circles (i.e., $\{U = 1, W = 0\}$ or $\{Y = 1, W = 1\}$); or to connect the two triangles by a straight line if he/she belongs to one of the two triangles (i.e., $\{U = 2, W = 0\}$ or $\{Y = 2, W = 1\}$); ...; or to connect the two dots by a straightline if he/she belongs to one of the two dots (i.e., $\{U = m, W = 0\}$ or $\{Y = m, W = 1\}$). Note that all $\{W = 0\}$, $\{W = 1\}$, and $\{U = i\}$ are non-sensitive subclasses, thus

$$\{U = i, W = 0\} \cup \{Y = i, W = 1\}, \quad i = 1, \dots, m,$$

are also non-sensitive subclasses. Therefore, the respondent's privacy is well protected, and the interviewer does not have information on whether the interviewee belongs to the sensitive class or not. We call this the multi-category parallel model. The right-hand side of Table 1 shows the corresponding cell probabilities. Since the three random variables W, U and Y are independent, the joint probability is the product of two corresponding marginal probabilities.

For those who may not completely understand the questionnaire shown in Table 1, we can formulate the survey design of the multi-category parallel model in another manner. For example, let $m = 4$ and define $Y = 1, 2, 3$ or 4 if the

number of days of the illegal drug usage in the last month for a respondent is 0–1, 2, 3 or ≥ 4 . Thus, the 4-category parallel model can be re-formulated in the following way:

- 1° If your birthday is in the first half of a month (i.e., $W = 0$), please answer ‘1’ (i.e., $U = 1$), or ‘2’ (i.e., $U = 2$), or ‘3’ (i.e., $U = 3$), or ‘4’ (i.e., $U = 4$) to the question: *In which quarter of the year is your birthday?*
- 2° If your birthday is in the second half of a month (i.e., $W = 1$), please answer ‘1’ (i.e., $Y = 1$), or ‘2’ (i.e., $Y = 2$), or ‘3’ (i.e., $Y = 3$), or ‘4’ (i.e., $Y = 4$) to the question: *How many days in the last month have you used illegal drugs?*

3. MAXIMUM LIKELIHOOD INFERENCE

3.1 MLEs of parameters via the EM algorithm

Suppose we conduct a sample survey with n questionnaires and observe n_1 respondents connecting the two circles, n_2 respondents connecting the two triangles, ..., and n_m respondents connecting the two dots (see, Table 1). Let $Y_{\text{obs}} = \{n; n_1, \dots, n_m\}$ denote the observed data, where $n = \sum_{i=1}^m n_i$. Hence, the observed-data likelihood function for $\boldsymbol{\pi} = (\pi_1, \dots, \pi_m)^\top$ is

$$(3.1) \quad L_{\text{MP}}(\boldsymbol{\pi}|Y_{\text{obs}}) = \binom{n}{n_1, \dots, n_m} \prod_{i=1}^m [p_i(1-q) + \pi_i q]^{n_i},$$

where the subscript ‘MP’ refers to the ‘multi-category parallel’ model. We employ the EM algorithm [7] to calculate the MLEs of $\{\pi_i\}_{i=1}^m$ by introducing the latent vector $\mathbf{z} = (z_1, \dots, z_m)^\top$, where z_i denotes the number of respondents belonging to the sensitive subclass $\{Y = i, W = 1\}$. We denote the complete data by $Y_{\text{com}} = \{Y_{\text{obs}}, \mathbf{z}\}$. Note that $\{p_i\}_{i=1}^m$ and q are known, the complete-data likelihood function for $\boldsymbol{\pi}$ is

$$(3.2) \quad L_{\text{MP}}(\boldsymbol{\pi}|Y_{\text{obs}}, \mathbf{z}) \propto \prod_{i=1}^m [p_i(1-q)]^{n_i - z_i} (\pi_i q)^{z_i} \propto \prod_{i=1}^m \pi_i^{z_i}.$$

Therefore, the M-step is to calculate the complete-data MLEs of $\{\pi_i\}_{i=1}^m$, which are given by

$$(3.3) \quad \hat{\pi}_i = \frac{z_i}{z_1 + \dots + z_m}, \quad i = 1, \dots, m.$$

Since the conditional predictive density is

$$(3.4) \quad \begin{aligned} f(\mathbf{z}|Y_{\text{obs}}, \boldsymbol{\pi}) &= \prod_{i=1}^m f(z_i|Y_{\text{obs}}, \pi_i) \\ &= \prod_{i=1}^m \text{Binomial}\left(z_i \middle| n_i, \frac{\pi_i q}{p_i(1-q) + \pi_i q}\right), \end{aligned}$$

the E-step is to replace z_i in (3.3) by its conditional expectation

$$(3.5) \quad E(z_i|Y_{\text{obs}}, \pi_i) = \frac{n_i \pi_i q}{p_i(1-q) + \pi_i q}, \quad i = 1, \dots, m.$$

3.2 Two bootstrap confidence intervals of parameters

We utilize the bootstrap method to derive the corresponding CIs of $\{\pi_i\}_{i=1}^m$. Based on the obtained MLE $\hat{\boldsymbol{\pi}}_{\text{MP}} = (\hat{\pi}_{\text{MP}1}, \dots, \hat{\pi}_{\text{MP}m})^\top$ of $\boldsymbol{\pi}$, we could generate

$$(n_1^*, \dots, n_m^*)^\top \sim \text{Multinomial}(n; p_1(1-q) + q\hat{\pi}_{\text{MP}1}, \dots, p_m(1-q) + q\hat{\pi}_{\text{MP}m}).$$

For the bootstrap sample $\{n_1^*, \dots, n_m^*\}$, we can compute the bootstrap replication $\hat{\pi}_{\text{MP}i}^*$ via the EM algorithm (3.3) and (3.5) by replacing $\{n_1, \dots, n_m\}$ with $\{n_1^*, \dots, n_m^*\}$. Independently repeating this process G times, we obtain G bootstrap replications $\{\hat{\pi}_{\text{MP}i}^*(g)\}_{g=1}^G$. Therefore, the standard error, $\text{Se}(\hat{\pi}_{\text{MP}i})$, of $\hat{\pi}_{\text{MP}i}$ can be estimated by the sample standard deviation of the G replications, i.e.

$$(3.6) \quad \begin{aligned} \widehat{\text{Se}}(\hat{\pi}_{\text{MP}i}) &= \left\{ \frac{1}{G-1} \sum_{g=1}^G \left[\hat{\pi}_{\text{MP}i}^*(g) - \frac{\hat{\pi}_{\text{MP}i}^*(1) + \dots + \hat{\pi}_{\text{MP}i}^*(G)}{G} \right]^2 \right\}^{\frac{1}{2}}. \end{aligned}$$

If $\{\hat{\pi}_{\text{MP}i}^*(g)\}_{g=1}^G$ is approximately normally distributed, a $(1-\alpha)100\%$ bootstrap CI for π_i is given by

$$(3.7) \quad [\hat{\pi}_{\text{MP}i} - z_{\alpha/2} \widehat{\text{Se}}(\hat{\pi}_{\text{MP}i}), \hat{\pi}_{\text{MP}i} + z_{\alpha/2} \widehat{\text{Se}}(\hat{\pi}_{\text{MP}i})],$$

where z_α is the upper α -th quantile of the standard normal distribution. Alternatively, if $\{\hat{\pi}_{\text{MP}i}^*(g)\}_{g=1}^G$ is non-normally distributed or the bootstrap CI (3.7) is beyond the unit interval $(0, 1)$, a $(1-\alpha)100\%$ bootstrap CI of π_i can be obtained by

$$(3.8) \quad [\hat{\pi}_{\text{MP}i, \text{BL}}, \hat{\pi}_{\text{MP}i, \text{BU}}],$$

where $\hat{\pi}_{\text{MP}i, \text{BL}}$ and $\hat{\pi}_{\text{MP}i, \text{BU}}$ are the $100(\alpha/2)$ and $100(1-\alpha/2)$ percentiles of $\{\hat{\pi}_{\text{MP}i}^*(g)\}_{g=1}^G$, respectively.

3.3 Explicit solutions to the valid estimators

Although the resulting MLE $\hat{\boldsymbol{\pi}}_{\text{MP}}$ via the EM algorithm (3.3) and (3.5) definitely belongs to \mathbb{T}_m , we can only obtain a numerical solution to $\hat{\boldsymbol{\pi}}_{\text{MP}}$. In addition, the variance-covariance matrix of $\hat{\boldsymbol{\pi}}_{\text{MP}}$ does not have a closed-form expression. However, for some cases, we can obtain explicit solutions to $\hat{\boldsymbol{\pi}}_{\text{MP}}$ and its variance-covariance matrix.

From (3.1), the log-likelihood function is given by

$$\ell_{\text{MP}}(\boldsymbol{\pi}|Y_{\text{obs}}) = c + \sum_{i=1}^m n_i \log[p_i(1-q) + \pi_i q],$$

where c is a constant not depending on $\boldsymbol{\pi}$. Let $\partial \ell_{\text{MP}}(\boldsymbol{\pi} | Y_{\text{obs}}) / \partial \pi_i = 0$, an alternative estimator of $\boldsymbol{\pi}$ is given by

$$(3.9) \quad \hat{\boldsymbol{\pi}}_v = (\hat{\pi}_{v1}, \dots, \hat{\pi}_{vm})^\top \\ = \left(\frac{n_1/n - p_1(1-q)}{q}, \dots, \frac{n_m/n - p_m(1-q)}{q} \right)^\top.$$

Although $\hat{\pi}_{vi}$ is an unbiased estimator of the true proportion π_i , $\hat{\boldsymbol{\pi}}_v$ may not belong to \mathbb{T}_m . For example, let $m = 4$, $p_1 = \dots = p_4 = 0.25$, $q = 1/3$ and $(n_1, \dots, n_4)^\top = (15, 19, 7, 9)^\top$, then

$$\hat{\boldsymbol{\pi}}_v = (0.40, 0.64, -0.08, 0.04)^\top \notin \mathbb{T}_4.$$

In this paper, the estimator $\hat{\boldsymbol{\pi}}_v$ given by (3.9) is said to be *valid* if $\hat{\boldsymbol{\pi}}_v \in \mathbb{T}_m$. Clearly, if $\hat{\boldsymbol{\pi}}_v$ specified by (3.9) is a valid estimator of $\boldsymbol{\pi}$, then $\hat{\boldsymbol{\pi}}_v = \hat{\boldsymbol{\pi}}_{\text{MP}}$. In the following discussion, we only consider the case of valid estimators.

Note that $(n_1, \dots, n_m)^\top \sim \text{Multinomial}_m(n, \boldsymbol{\lambda})$, where $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_m)^\top \in \mathbb{T}_m$,

$$(3.10) \quad \lambda_i = p_i(1-q) + \pi_i q, \quad i = 1, \dots, m.$$

Let $\hat{\boldsymbol{\lambda}} = (\hat{\lambda}_1, \dots, \hat{\lambda}_m)^\top = (n_1/n, \dots, n_m/n)^\top$ denote the MLE of $\boldsymbol{\lambda}$. Then (3.9) can be rewritten in the vector form as

$$(3.11) \quad \hat{\boldsymbol{\pi}}_{\text{MP}} = (\hat{\pi}_{\text{MP}1}, \dots, \hat{\pi}_{\text{MP}m})^\top = \frac{\hat{\boldsymbol{\lambda}} - (1-q)\mathbf{p}}{q},$$

where $\mathbf{p} = (p_1, \dots, p_m)^\top$. Thus, the variance-covariance matrix of $\hat{\boldsymbol{\pi}}_{\text{MP}}$ is given by

$$(3.12) \quad \text{Var}(\hat{\boldsymbol{\pi}}_{\text{MP}}) = \frac{1}{q^2} \text{Var}(\hat{\boldsymbol{\lambda}}) \\ = \frac{1}{nq^2} \begin{pmatrix} \lambda_1(1-\lambda_1) & -\lambda_1\lambda_2 & \cdots & -\lambda_1\lambda_m \\ -\lambda_2\lambda_1 & \lambda_2(1-\lambda_2) & \cdots & -\lambda_2\lambda_m \\ \vdots & \vdots & \ddots & \vdots \\ -\lambda_m\lambda_1 & -\lambda_m\lambda_2 & \cdots & \lambda_m(1-\lambda_m) \end{pmatrix}.$$

3.4 Three asymptotic confidence intervals of parameters for large sample sizes

From (3.12), it is not difficult to show the following result.

Theorem 1. *Let*

$$(3.13) \quad \widehat{\text{Var}}(\hat{\pi}_{\text{MP}i}) = \frac{\hat{\lambda}_i(1-\hat{\lambda}_i)}{(n-1)q^2}, \quad i = 1, \dots, m.$$

Then, we have

$$\widehat{\text{Var}}(\hat{\pi}_{\text{MP}i}) = \frac{\hat{\pi}_{\text{MP}i}(1-\hat{\pi}_{\text{MP}i})}{n-1} + \frac{(1-q)f(\hat{\pi}_{\text{MP}i}, p_i, q)}{(n-1)q^2},$$

where $f(\pi_i, p_i, q) \hat{=} q(1-2p_i)\pi_i + p_i(1-p_i+qp_i)$, and it is an unbiased estimator of $\text{Var}(\hat{\pi}_{\text{MP}i}) = \lambda_i(1-\lambda_i)/(nq^2)$, $i = 1, \dots, m$.

Based on the property of MLE for large sample sizes, we have

$$\frac{\hat{\pi}_{\text{MP}i} - \pi_i}{\sqrt{\widehat{\text{Var}}(\hat{\pi}_{\text{MP}i})}} \sim N(0, 1), \quad \text{as } n \rightarrow \infty, \quad i = 1, \dots, m.$$

Thus, an asymptotic $(1-\alpha)100\%$ Wald CI for π_i is given by

$$(3.14) \quad \left[\hat{\pi}_{\text{MP}i} - z_{\alpha/2} \sqrt{\widehat{\text{Var}}(\hat{\pi}_{\text{MP}i})}, \hat{\pi}_{\text{MP}i} + z_{\alpha/2} \sqrt{\widehat{\text{Var}}(\hat{\pi}_{\text{MP}i})} \right].$$

If the lower bound of the Wald CI in (3.14) is less than zero or the upper bound is larger than one, then the Wald CI is useless. For such cases, according to the Central Limit Theorem, we can establish an asymptotic $(1-\alpha)100\%$ Wilson (score) CI of π_i based on

$$(3.15) \quad 1 - \alpha = \Pr \left\{ \left| \frac{\hat{\pi}_{\text{MP}i} - \pi_i}{\sqrt{\widehat{\text{Var}}(\hat{\pi}_{\text{MP}i})}} \right| \leq z_{\alpha/2} \right\} \\ = \Pr \{ (\hat{\pi}_{\text{MP}i} - \pi_i)^2 \leq z_{\alpha/2}^2 \text{Var}(\hat{\pi}_{\text{MP}i}) \} \\ \stackrel{(3.12)}{=} \Pr \left\{ (\hat{\pi}_{\text{MP}i} - \pi_i)^2 \leq \frac{z_{\alpha/2}^2}{n} \left[\pi_i(1-\pi_i) \right. \right. \\ \left. \left. + \frac{(1-q)f(\pi_i, p_i, q)}{q^2} \right] \right\} \\ = \Pr \left\{ \hat{\pi}_{\text{MP}i}^2 - 2\hat{\pi}_{\text{MP}i}\pi_i + \pi_i^2 \right. \\ \left. \leq \frac{z_{\alpha/2}^2(-\pi_i^2 + \rho_1\pi_i + \rho_2)}{n} \right\} \\ = \Pr \{ (1+z_*)\pi_i^2 - (2\hat{\pi}_{\text{MP}i} + z_*\rho_1)\pi_i \\ + \hat{\pi}_{\text{MP}i}^2 - z_*\rho_2 \leq 0 \},$$

where $z_* \hat{=} z_{\alpha/2}^2/n$, $\rho_1 \hat{=} [1-2p_i(1-q)]/q$ and $\rho_2 \hat{=} p_i(1-q)(1-p_i+qp_i)/q^2$. Solving the quadratic inequality inside the probability in (3.15), we obtain the Wilson CI given by

$$(3.16) \quad \frac{2\hat{\pi}_{\text{MP}i} + z_*\rho_1 \pm \sqrt{(2\hat{\pi}_{\text{MP}i} + z_*\rho_1)^2 - 4(1+z_*)(\hat{\pi}_{\text{MP}i}^2 - z_*\rho_2)}}{2(1+z_*)},$$

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 [6, 2, 10, 5] for more detail.

For sensitive responses where some of the true values $\{\pi_i\}$ are often small, *likelihood ratio confidence intervals* (LRCIs) could provide better performance than other alternatives. To construct the LRCI of π_i ($i = 1, \dots, m$), we consider the null hypothesis $H_0: \pi_i = \pi_{i0}$ against the alternative hypothesis $H_1: H_0$ is not true. Let $\hat{\boldsymbol{\pi}}^R = (\hat{\pi}_1^R, \dots, \hat{\pi}_m^R)^\top$ denote the

Table 2. The four-category parallel model and the corresponding cell probabilities

Category	$W = 0$	$W = 1$	Category	$W = 0$	$W = 1$	Marginal
$U = 1$	○		$U = 1$	$p_1(1 - q)$		p_1
$U = 2$	△		$U = 2$	$p_2(1 - q)$		p_2
$U = 3$	□		$U = 3$	$p_3(1 - q)$		p_3
$U = 4$	●		$U = 4$	$p_4(1 - q)$		p_4
$X = 0, Y = 0$		○	$X = 0, Y = 0$		$\pi_1 q$	π_1
$X = 0, Y = 1$		△	$X = 0, Y = 1$		$\pi_2 q$	π_2
$X = 1, Y = 0$		□	$X = 1, Y = 0$		$\pi_3 q$	π_3
$X = 1, Y = 1$		●	$X = 1, Y = 1$		$\pi_4 q$	π_4
			Marginal	$1 - q$	q	1

restricted MLE of π under H_0 . It can be verified that

$$\begin{cases} \hat{\pi}_i^R = \pi_{i0}, \\ \hat{\pi}_j^R = \frac{[1 - p_i(1 - q) - \pi_{i0}q]n_j / (n - n_i) - p_j(1 - q)}{q}, \end{cases}$$

where $j = 1, \dots, m; j \neq i$.

When $n \rightarrow \infty$, it is well known that

$$\Lambda(\pi_{i0}) = -2\{\ell_{\text{MP}}(\hat{\pi}^R | Y_{\text{obs}}) - \ell_{\text{MP}}(\hat{\pi}_v | Y_{\text{obs}})\} \sim \chi^2(1),$$

where $\hat{\pi}_v$ denotes the unrestricted MLE of π specified by (3.9). Since

$$(3.17) \quad \Lambda(\pi_{i0}) = -2 \left\{ n_i \log[p_i(1 - q) + \pi_{i0}q] + \sum_{j=1, j \neq i}^m n_j \log[p_j(1 - q) + \hat{\pi}_j^R q] - \sum_{k=1}^m n_k \log[p_k(1 - q) + \hat{\pi}_{vk} q] \right\},$$

it is easy to verify that $\Lambda(\pi_{i0})$ is a decreasing function of π_{i0} when $\pi_{i0} \in [0, \frac{n_i/n - p_i(1 - q)}{q}]$ and an increasing function of π_{i0} when $\pi_{i0} \in [\frac{n_i/n - p_i(1 - q)}{q}, 1]$. Therefore, for a given significance level α , the $(1 - \alpha)100\%$ LRCI for π_i is given by

$$(3.18) \quad [\hat{\pi}_{\text{MP}i, \text{LRL}}, \hat{\pi}_{\text{MP}i, \text{LRU}}],$$

where $\hat{\pi}_{\text{MP}i, \text{LRL}}$ and $\hat{\pi}_{\text{MP}i, \text{LRU}}$ are two roots of π_{i0} to the following equation

$$(3.19) \quad \Lambda(\pi_{i0}) = \chi^2(\alpha, 1),$$

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

The asymptotic CIs (3.14), (3.16) and (3.18) are appropriate for the cases of large sample sizes. When n is small to moderate, we could use the bootstrap CIs (3.7) and/or (3.8).

4. A SPECIAL CASE FOR THE MULTI-CATEGORY PARALLEL MODEL

In this section, we consider a special case of the the multi-category parallel model with four categories, which can be utilized to investigate the association of two binary sensitive variates. Some simulation studies are conducted to assess the performances of the likelihood ratio test and the chi-squared statistic by comparing their empirical type I error rates (or the actual significance levels) and powers.

4.1 A four-category parallel model

Let X and Y be two dichotomous random variables associated with two sensitive questions. For example, X represents whether or not a respondent is an illegal drug user and Y denotes whether a respondent is with AIDS or not. Let $X = 1$ and $Y = 1$ denote the sensitive characteristics of a respondent (e.g., $X = 1$ if the respondent is a drug user), and $X = 0$ and $Y = 0$ denote the non-sensitive characteristics of a respondent (e.g., $Y = 0$ if a respondent is without AIDS). Define $\pi_1 = \Pr\{X = 0, Y = 0\}$, $\pi_2 = \Pr\{X = 0, Y = 1\}$, $\pi_3 = \Pr\{X = 1, Y = 0\}$ and $\pi_4 = \Pr\{X = 1, Y = 1\}$. Obviously, we have $\pi = (\pi_1, \dots, \pi_4)^\top \in \mathbb{T}_4$. From Table 1, the survey design for the four-category parallel model is displayed in Table 2. Two major objectives here are to collect sensitive data and to test whether or not the association exists between the two binary variates X and Y .

4.2 Testing hypothesis for association

A commonly used index for measuring the association of two binary variates is the odds ratio $\psi = \pi_1\pi_4/(\pi_2\pi_3)$. Assume that we want to test $H_0: \psi = 1$ against $H_1: \psi \neq 1$. The likelihood ratio statistic defined by

$$(4.1) \quad \Lambda_1 = -2\{\ell_{\text{MP}}(\hat{\pi}_0 | Y_{\text{obs}}) - \ell_{\text{MP}}(\hat{\pi}_{\text{MP}} | Y_{\text{obs}})\} \sim \chi^2(1), \quad \text{as } n \rightarrow \infty,$$

where $\hat{\pi}_0$ denotes the restricted MLE of π under H_0 and $\hat{\pi}_{\text{MP}}$ denotes the MLE of π given by (3.11). To calculate $\hat{\pi}_0$, let $\pi_x \hat{=} \Pr(X = 1) = \pi_3 + \pi_4$ and $\pi_y \hat{=} \Pr(Y = 1) = \pi_2 + \pi_4$.

If H_0 is true, i.e., X and Y are mutually independent, we have

$$(4.2) \quad \begin{cases} \pi_1 = (1 - \pi_x)(1 - \pi_y), \\ \pi_2 = (1 - \pi_x)\pi_y, \\ \pi_3 = \pi_x(1 - \pi_y), \\ \pi_4 = \pi_x\pi_y. \end{cases} \quad \text{and}$$

If we could obtain the restricted MLEs $\hat{\pi}_{0x}$ of π_x and $\hat{\pi}_{0y}$ of π_y , from (4.2) the restricted MLEs $\hat{\boldsymbol{\pi}}_0 = (\hat{\pi}_{01}, \dots, \hat{\pi}_{04})^\top$ can be calculated as

$$(4.3) \quad \begin{cases} \hat{\pi}_{01} = (1 - \hat{\pi}_{0x})(1 - \hat{\pi}_{0y}), \\ \hat{\pi}_{02} = (1 - \hat{\pi}_{0x})\hat{\pi}_{0y}, \\ \hat{\pi}_{03} = \hat{\pi}_{0x}(1 - \hat{\pi}_{0y}), \\ \hat{\pi}_{04} = \hat{\pi}_{0x}\hat{\pi}_{0y}. \end{cases} \quad \text{and}$$

Recall that the number of the respondents belonging to the subclass $\{Y = i, W = 1\}$ is denoted by z_i and the frequencies $\{z_i\}$ are unobservable. From (3.2), the complete-data likelihood function for $\boldsymbol{\pi}$ under H_0 becomes

$$\begin{aligned} L_{\text{MP}}(\pi_x, \pi_y | Y_{\text{obs}}, \mathbf{z}, H_0) &\propto [(1 - \pi_x)(1 - \pi_y)]^{z_1} [(1 - \pi_x)\pi_y]^{z_2} \\ &\quad \times [\pi_x(1 - \pi_y)]^{z_3} (\pi_x\pi_y)^{z_4} \\ &= \pi_x^{z_3+z_4} (1 - \pi_x)^{z_1+z_2} \times \pi_y^{z_2+z_4} (1 - \pi_y)^{z_1+z_3}. \end{aligned}$$

Thus, the M-step is to calculate the complete-data MLEs of π_x and π_y as follows:

$$(4.4) \quad \hat{\pi}_{0x} = \frac{z_3 + z_4}{z_+} \quad \text{and} \quad \hat{\pi}_{0y} = \frac{z_2 + z_4}{z_+},$$

respectively. From (3.4), the E-step is to find the conditional expectations:

$$(4.5) \quad E(z_i | Y_{\text{obs}}, \hat{\boldsymbol{\pi}}_0) = \frac{n_i \hat{\pi}_{0i} q}{p_i(1 - q) + \hat{\pi}_{0i} q}, \quad i = 1, \dots, m,$$

where $\{\hat{\pi}_{0i}\}_{i=1}^4$ are defined by (4.3). Alternatively, the chi-squared statistic can be utilized to test H_0 against H_1 . Let $\mathbf{p} = (p_1, \dots, p_4)^\top$ and $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_4)^\top$, where $\lambda_i = (1 - q)p_i + q\pi_i$ for $i = 1, \dots, 4$. Then, we have $\boldsymbol{\lambda} = (1 - q)\mathbf{p} + q\boldsymbol{\pi}$. Note that the restricted MLE $\hat{\boldsymbol{\pi}}_0 = (\hat{\pi}_{01}, \dots, \hat{\pi}_{04})^\top$ of $\boldsymbol{\pi}$ under H_0 is given by (4.3), then

$$\hat{\boldsymbol{\lambda}}_0 = (\hat{\lambda}_{01}, \dots, \hat{\lambda}_{04})^\top = (1 - q)\mathbf{p} + q\hat{\boldsymbol{\pi}}_0$$

is the restricted MLE of $\boldsymbol{\lambda}$ under H_0 . Therefore, under H_0 , the chi-squared statistic

$$(4.6) \quad \Lambda_2 = \sum_{i=1}^4 \frac{(n_i - n\hat{\lambda}_{0i})^2}{n\hat{\lambda}_{0i}} \sim \chi^2(1), \quad \text{as } n \rightarrow \infty.$$

Table 3. Various values of π_1 and ψ for the three scenarios specified by (4.7)

	π_1					
	0.200	0.300	0.500	0.700	0.800	0.900
Scenario 1: ψ	0.167	0.286	0.667	1.556	2.667	6.000
Scenario 2: ψ	0.094	0.161	0.375	0.875	1.500	3.375
Scenario 3: ψ	0.064	0.110	0.256	0.598	1.026	2.308

4.3 Comparison of the likelihood ratio test with the χ^2 test

For a given π_1 , we consider the following three combinations of π_2, π_3 and π_4 such that $\sum_{i=1}^4 \pi_i = 1$:

$$(4.7) \quad \begin{aligned} \text{Scenario 1: } (\pi_2, \pi_3, \pi_4) &= (3, 4, 1) \frac{1 - \pi_1}{8}, \quad \psi = \frac{2\pi_1}{3(1 - \pi_1)}; \\ \text{Scenario 2: } (\pi_2, \pi_3, \pi_4) &= (4, 10, 1) \frac{1 - \pi_1}{15}, \quad \psi = \frac{3\pi_1}{8(1 - \pi_1)}; \\ \text{Scenario 3: } (\pi_2, \pi_3, \pi_4) &= (6, 13, 1) \frac{1 - \pi_1}{20}, \quad \psi = \frac{10\pi_1}{39(1 - \pi_1)}. \end{aligned}$$

The sample sizes in simulations are designed by $n = 50(50)500$. To compare the type I error rates (i.e., $\pi_1\pi_4/(\pi_2\pi_3) = \psi = 1$), we take $\pi_1 = \frac{3}{5}$ for scenario 1, $\pi_1 = \frac{8}{11}$ for scenario 2, and $\pi_1 = \frac{39}{49}$ for scenario 3. For the comparison of powers (i.e., $\psi \neq 1$), the chosen π_1 and the corresponding ψ are listed in Table 3. For a given pair (n, π_1) , we independently generate

$$(4.8) \quad \begin{aligned} (n_1^{(l)}, \dots, n_4^{(l)}) &\sim \\ \text{Multinomial} &\left(n; \frac{1}{8} + \frac{1}{2}\pi_1, \frac{1}{8} + \frac{1}{2}\pi_2, \frac{1}{8} + \frac{1}{2}\pi_3, \frac{1}{8} + \frac{1}{2}\pi_4 \right) \end{aligned}$$

for $l = 1, \dots, L$ ($L = 1,000$), where only $p_i = \frac{1}{4}$ ($i = 1, \dots, 4$) and $q = \frac{1}{2}$ are considered. All hypothesis testings are conducted at level 0.05. Let r_j denote the number rejecting the null hypothesis (i.e., $H_0: \psi = 1$) by the statistics Λ_j ($j = 1, 2$). Hence, the actual significance level can be estimated by r_j/L with $\psi = 1$ and the power of the test statistic Λ_j ($j = 1, 2$) can be estimated by r_j/L with $\psi \neq 1$.

Figure 1 shows that some comparisons of type I error rates between the likelihood ratio test and the χ^2 test for the three scenarios. In general, the chi-squared test has a better performance in controlling its Type I error rates around the pre-chosen nominal level than the likelihood ratio test, which can be seen in the three scenarios.

Figure 2 gives the comparisons of powers between the likelihood ratio test and the chi-squared test for different cases with $\psi \neq 1$. It is not difficult to find that there is no significant difference between the powers of the two test when ψ is small (i.e., < 0.40). When $0.60 < \psi < 1$, always

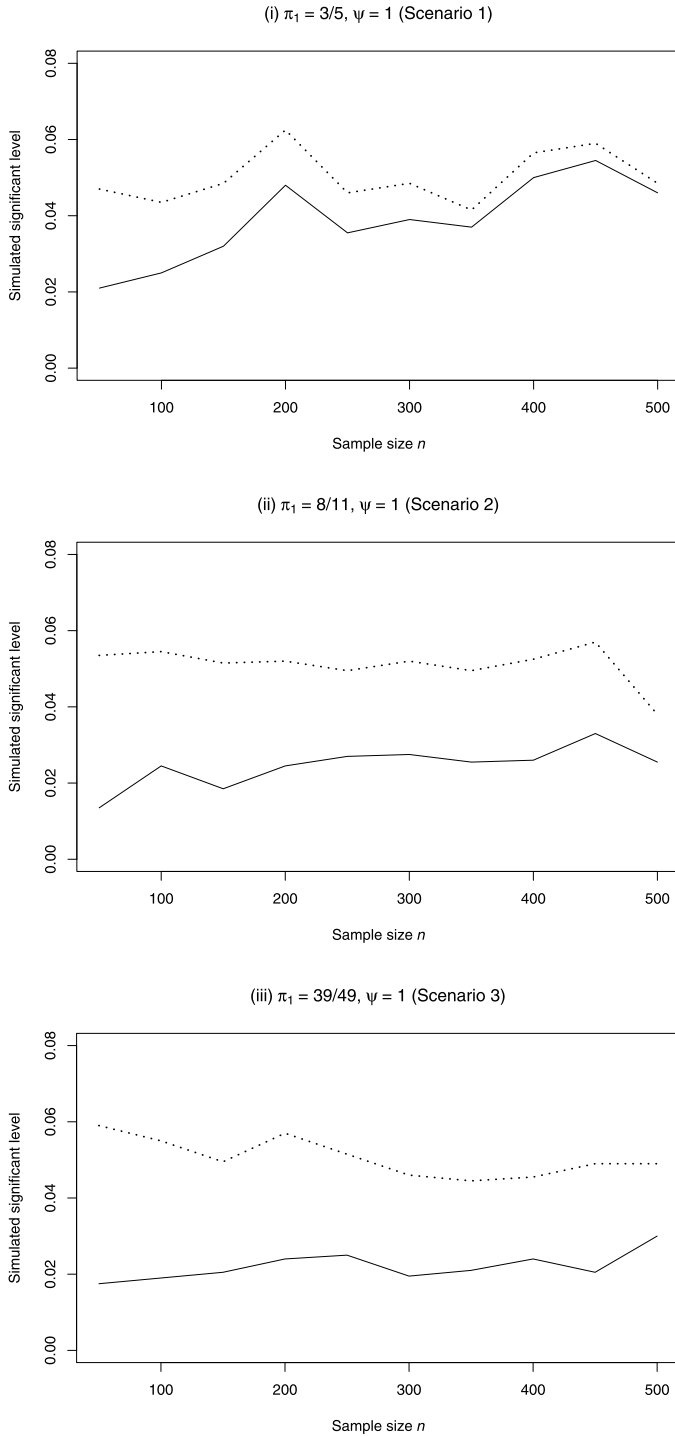


Figure 1. Comparisons of type I error rates between the likelihood ratio test (solid line) and the χ^2 test (dotted line): (a) $\pi_1 = 3/5$, $\psi = 1$ (Scenario 1); (b) $\pi_1 = 8/11$, $\psi = 1$ (Scenario 2); (c) $\pi_1 = 39/49$, $\psi = 1$ (Scenario 3).

the likelihood ratio test is slightly less powerful than the chi-squared test, no matter whether the sample size is large or small.

Table 4. Questionnaire for the multi-category triangular model

Category	$U = 1$	$U = 2$	\dots	$U = m$
1: $\{Y = 1\}$	Block 1: <input type="checkbox"/>	Block 2: <input type="checkbox"/>	\dots	Block m : <input type="checkbox"/>
2: $\{Y = 2\}$	Category 2: please put a tick in Block 2			
\vdots	\dots			
m : $\{Y = m\}$	Category m : please put a tick in Block m			

Note: $\{Y = 1\}$ represents the non-sensitive class.

5. COMPARISON OF THE MULTI-CATEGORY PARALLEL MODEL WITH THE MULTI-CATEGORY TRIANGULAR MODEL

In this section, we first briefly introduce the multi-category triangular model [11], and we then theoretically compare the efficiency of the multi-category parallel model with the multi-category triangular model by comparing the two variance-covariance matrices of the MLEs of parameters based on the trace criterion. Finally, we also consider the comparison of the degree of privacy protection for the two models.

5.1 The survey design for the multi-category triangular model

Tang et al. [11] proposed the survey design of the multi-category triangular model and developed the corresponding methods of statistical inference. Let U and Y be two independent categorical random variables as defined in Section 2, $\{Y = 1\}$ denote the non-sensitive class and $\{Y = j\}$ the sensitive classes for $j = 2, \dots, m$. Define $p_j = \Pr\{U = j\}$ and $\pi_j = \Pr\{Y = j\}$, $j = 1, \dots, m$. Assume that $\{p_j\}_{j=1}^m$ are known. The objective is to estimate $\boldsymbol{\pi} = (\pi_1, \dots, \pi_m)^\top \in \mathbb{T}_m$. The survey design shown in Table 4 can be described as following: Since the category 1 (i.e., $\{Y = 1\}$) is a non-sensitive class, it is supposed that the respondents belonging to this class can provide correct answers (i.e., putting a tick in Block j for $j = 1, \dots, m$) according to their true status. In addition, the respondents belonging to the category j ($j = 2, \dots, m$) will be asked to put a tick in Block j . The cell probabilities $\{\pi_j\}$, the observed frequencies $\{n_j\}$ and the unobservable frequencies $\{z_j\}$ are shown in Table 5.

5.2 The difference between two traces of variance-covariance matrix of the MLEs of parameters

(a) The variance-covariance matrix of $\hat{\boldsymbol{\pi}}_{\text{MT}}$

For the multi-category triangular model, let $\boldsymbol{\theta} = (\theta_1, \dots, \theta_m)^\top$, where $\theta_1 = p_1\pi_1$ and $\theta_j = p_j\pi_1 + \pi_j$ ($j = 2, \dots, m$) represent the proportions that the respondents belonging to Block j . In matrix notation, we have

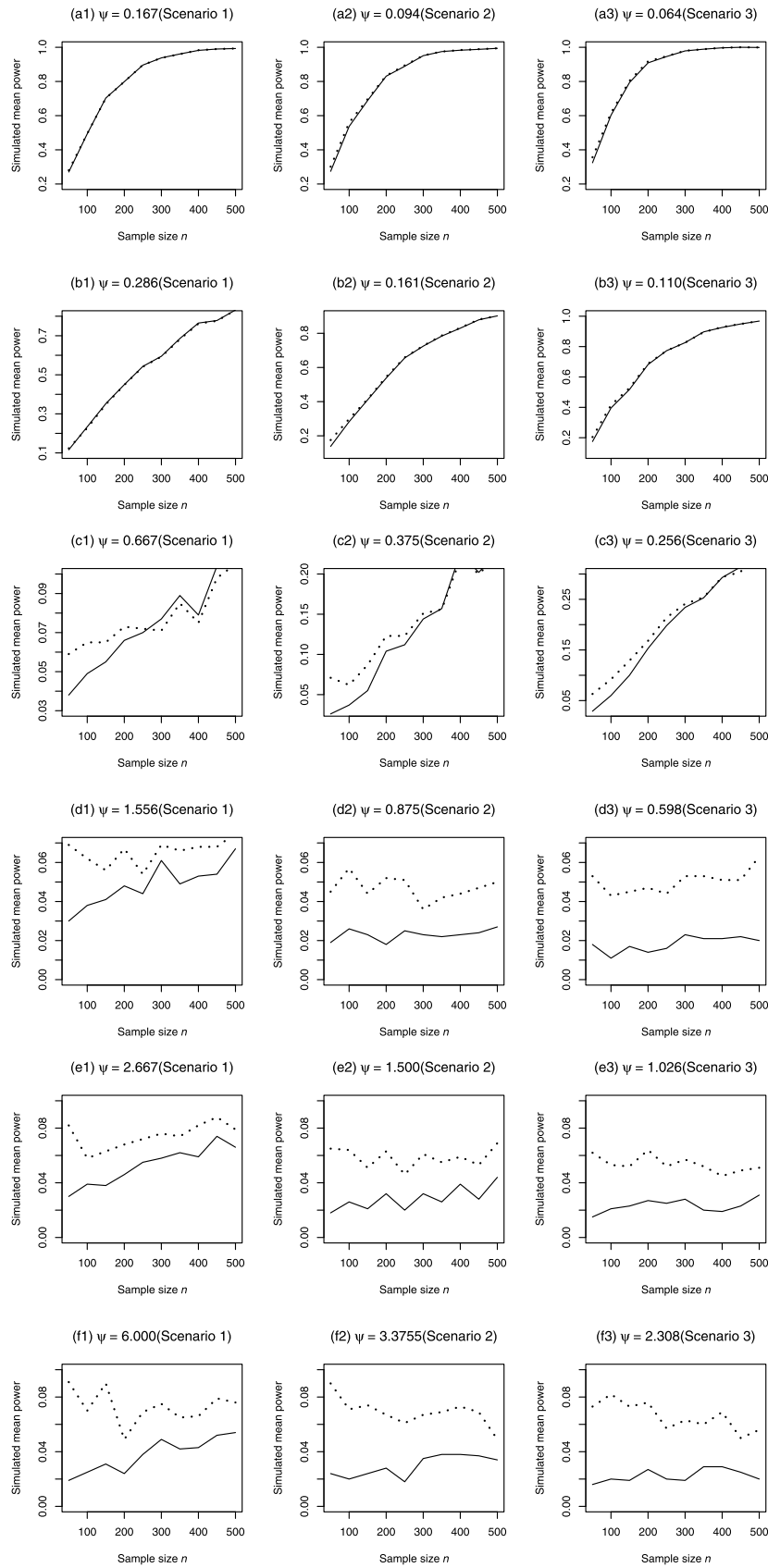


Figure 2. Comparisons of powers between the likelihood ratio test (solid line) and the χ^2 test (dotted line).

Table 5. Questionnaire for the multi-category triangular model

Category	$U = 1$	$U = 2$	\dots	$U = m$	Total
1: $Y = 1$	$p_1\pi_1$	$p_2\pi_1$	\dots	$p_m\pi_1$	$\pi_1(z_1)$
2: $Y = 2$					$\pi_2(z_2)$
\vdots					\vdots
m : $Y = m$					$\pi_m(z_m)$
Total	$p_1\pi_1(n_1)$	$p_2\pi_1 + \pi_2(n_2)$	\dots	$p_m\pi_1 + \pi_m(n_m)$	$1(n)$

Note: $n = \sum_{j=1}^m n_j$, $z_1 = n - \sum_{j=2}^m z_j$, where $\{z_2, \dots, z_m\}$ are unobservable.

$$(5.1) \quad \boldsymbol{\theta} = \mathbf{P}\boldsymbol{\pi} = \begin{pmatrix} p_1 & \mathbf{0}_{m-1}^\top \\ \mathbf{p}_{-1} & \mathbf{I}_{m-1} \end{pmatrix} \begin{pmatrix} \pi_1 \\ \vdots \\ \pi_m \end{pmatrix},$$

where $\mathbf{p}_{-1} = (p_2, \dots, p_m)^\top$, $\mathbf{0}_{m-1}$ is the $(m-1) \times 1$ vector of zeros and \mathbf{I}_{m-1} denotes the $(m-1) \times (m-1)$ identity matrix. Since $(n_1, \dots, n_m)^\top \sim \text{Multinomial}(n; \theta_1, \dots, \theta_m)$, the MLE of θ_j is $\hat{\theta}_j = n_j/n$. Thus, the variance-covariance matrix of $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \dots, \hat{\theta}_m)^\top$ is

$$(5.2) \quad \text{Var}(\hat{\boldsymbol{\theta}}) = \frac{1}{n} [\text{diag}(\boldsymbol{\theta}) - \boldsymbol{\theta}\boldsymbol{\theta}^\top].$$

In general, the MLE $\hat{\boldsymbol{\pi}}_{\text{MT}}$ of $\boldsymbol{\pi}$ for the multi-category triangular model can be obtained by using the EM algorithm [11]. However, for some cases, we can obtain a closed-form solution to $\hat{\boldsymbol{\pi}}_{\text{MT}}$. In fact, from (5.1), we have $\boldsymbol{\pi} = \mathbf{P}^{-1}\boldsymbol{\theta}$. Since the MLE of $\boldsymbol{\theta}$ is $\hat{\boldsymbol{\theta}} = (n_1/n, \dots, n_m/n)^\top$, an alternative estimator of $\boldsymbol{\pi}$ for the multi-category triangular model is given by

$$(5.3) \quad \hat{\boldsymbol{\pi}}_v = \mathbf{P}^{-1}\hat{\boldsymbol{\theta}} = \begin{pmatrix} 1/p_1 & \mathbf{0}_{m-1}^\top \\ -\mathbf{p}_{-1}/p_1 & \mathbf{I}_{m-1} \end{pmatrix} \begin{pmatrix} n_1/n \\ \vdots \\ n_m/n \end{pmatrix}.$$

It should be noted that it is possible that $\hat{\boldsymbol{\pi}}_v \notin \mathbb{T}_m$. For example, let $m = 4$, $p_1 = \dots = p_4 = 0.25$ and $(n_1, \dots, n_4)^\top = (12, 8, 6, 19)^\top$, then

$$\hat{\boldsymbol{\pi}}_v = (1.066667, -0.088889, -0.133333, 0.155556)^\top \notin \mathbb{T}_4.$$

In this paper, the estimator $\hat{\boldsymbol{\pi}}_v$ given by (5.3) is said to be *valid* if $\hat{\boldsymbol{\pi}}_v \in \mathbb{T}_m$. Clearly, if $\hat{\boldsymbol{\pi}}_v$ specified by (5.3) is a valid estimator of $\boldsymbol{\pi}$ then $\hat{\boldsymbol{\pi}}_v = \hat{\boldsymbol{\pi}}_{\text{MT}}$. In the following discussion, we only consider the case of valid estimators.

Hence, from (5.3), (5.2) and (5.1), the variance-covariance matrix of $\hat{\boldsymbol{\pi}}_{\text{MT}}$ is

$$(5.4) \quad \begin{aligned} \text{Var}(\hat{\boldsymbol{\pi}}_{\text{MT}}) &= \text{Var}(\hat{\boldsymbol{\pi}}_v) \\ &= \mathbf{P}^{-1} \text{Var}(\hat{\boldsymbol{\theta}}) (\mathbf{P}^{-1})^\top \\ &= \frac{1}{n} [\mathbf{P}^{-1} \text{diag}(\mathbf{P}\boldsymbol{\pi}) (\mathbf{P}^{-1})^\top - \boldsymbol{\pi}\boldsymbol{\pi}^\top], \end{aligned}$$

or equivalently

$$(5.5) \quad \begin{aligned} \text{Var}(\hat{\pi}_{\text{MT}1}) &= \frac{1}{n} \left(\frac{\pi_1}{p_1} - \pi_1^2 \right), \\ \text{Var}(\hat{\pi}_{\text{MT}j}) &= \frac{1}{n} \left(\frac{p_j^2}{p_1} \pi_1 + p_j \pi_1 + \pi_j - \pi_j^2 \right), \\ \text{Cov}(\hat{\pi}_{\text{MT}1}, \hat{\pi}_{\text{MT}j}) &= \frac{1}{n} \left(-\frac{p_j}{p_1} \pi_1 - \pi_1 \pi_j \right), \quad \text{and} \\ \text{Cov}(\hat{\pi}_{\text{MT}i}, \hat{\pi}_{\text{MT}j}) &= \frac{1}{n} \left(\frac{p_i p_j}{p_1} \pi_1 - \pi_i \pi_j \right), \end{aligned}$$

where $i \neq j$, $i, j = 2, \dots, m$.

(b) The comparison between $\text{Var}(\hat{\boldsymbol{\pi}}_{\text{MT}})$ and $\text{Var}(\hat{\boldsymbol{\pi}}_{\text{MP}})$

In the multi-category triangular model, there are only two parameter vectors (i.e., $\boldsymbol{\pi}$ and \mathbf{p}), while in the multi-category parallel model, besides $\boldsymbol{\pi}$ and \mathbf{p} , there is an additional parameter q . By controlling q within a certain subset of the unit interval, we may have $\text{Var}(\hat{\boldsymbol{\pi}}_{\text{MP}})$ being ‘smaller’ than $\text{Var}(\hat{\boldsymbol{\pi}}_{\text{MT}})$. In the follows, we only apply the trace criterion in the comparison between $\text{Var}(\hat{\boldsymbol{\pi}}_{\text{MT}})$ and $\text{Var}(\hat{\boldsymbol{\pi}}_{\text{MP}})$.

First, from (5.4) and (5.5), we have

$$\begin{aligned} \text{tr}[\text{Var}(\hat{\boldsymbol{\pi}}_{\text{MT}})] &= \frac{1}{n} \left[\frac{\pi_1}{p_1} - \pi_1 + \pi_1(1 - \pi_1) \right. \\ &\quad \left. + \sum_{j=2}^m \left(\frac{p_j^2}{p_1} \pi_1 + p_j \pi_1 + \pi_j(1 - \pi_j) \right) \right] \\ &= \frac{\pi_1}{n} \left(\frac{1}{p_1} - p_1 + \sum_{j=2}^m \frac{p_j^2}{p_1} \right) + \frac{1}{n} \sum_{j=1}^m \pi_j(1 - \pi_j). \end{aligned}$$

Next, from (3.12) and (3.10), we obtain

$$\begin{aligned} \text{tr}[\text{Var}(\hat{\boldsymbol{\pi}}_{\text{MP}})] &= \frac{1}{nq^2} \sum_{j=1}^m \lambda_j(1 - \lambda_j) \\ &= \frac{1}{nq^2} \sum_{j=1}^m [p_j(1 - q) + \pi_j q] [1 - p_j(1 - q) - \pi_j q] \\ &= \frac{1 - q}{nq^2} \left(1 + q - 2q \sum_{j=1}^m \pi_j p_j - (1 - q) \sum_{j=1}^m p_j^2 \right) \\ &\quad + \frac{1}{n} \sum_{j=1}^m \pi_j(1 - \pi_j). \end{aligned}$$

Thus, the difference of them is

$$\begin{aligned} \text{tr}[\text{Var}(\hat{\boldsymbol{\pi}}_{\text{MT}})] - \text{tr}[\text{Var}(\hat{\boldsymbol{\pi}}_{\text{MP}})] &= \frac{\pi_1}{n} \left(\frac{1}{p_1} - p_1 + \sum_{j=2}^m \frac{p_j^2}{p_1} \right) \end{aligned}$$

$$\begin{aligned}
& -\frac{1-q}{nq^2} \left(1+q-2q \sum_{j=1}^m \pi_j p_j - (1-q) \sum_{j=1}^m p_j^2 \right) \\
& = \frac{1}{nq^2} h(q|\boldsymbol{\pi}, \mathbf{p}),
\end{aligned}$$

where

$$\begin{aligned}
(5.6) \quad h(q|\boldsymbol{\pi}, \mathbf{p}) & = \left[\pi_1 \left(\frac{1}{p_1} - p_1 + \sum_{j=2}^m \frac{p_j^2}{p_1} \right) \right. \\
& \quad \left. + \left(1 - 2 \sum_{j=1}^m \pi_j p_j + \sum_{j=1}^m p_j^2 \right) \right] q^2 \\
& \quad + 2 \left(\sum_{j=1}^m \pi_j p_j - \sum_{j=1}^m p_j^2 \right) q - 1 + \sum_{j=1}^m p_j^2 \\
& \doteq aq^2 + bq + c
\end{aligned}$$

is a quadratic function of q for given $\boldsymbol{\pi}$ and \mathbf{p} . In both survey designs (see Table 1 and Table 4), we require $p_1 \in (0, 1)$ so that $1 - p_1^2 > 0$. In addition, $0 \leq \sum_{j=1}^m \pi_j^2 \leq \sum_{j=1}^m \pi_j = 1$. Thus,

$$\begin{aligned}
a & = \pi_1 \left(\frac{1}{p_1} - p_1 + \sum_{j=2}^m \frac{p_j^2}{p_1} \right) + \left(1 - 2 \sum_{j=1}^m \pi_j p_j + \sum_{j=1}^m p_j^2 \right) \\
& = \frac{\pi_1}{p_1} \left(1 - p_1^2 + \sum_{j=2}^m p_j^2 \right) + 1 - \sum_{j=1}^m \pi_j^2 + \sum_{j=1}^m (\pi_j - p_j)^2 \\
& > 0.
\end{aligned}$$

Now, the discriminant of the quadratic function $h(q|\boldsymbol{\pi}, \mathbf{p})$ is

$$\begin{aligned}
D(h) & = b^2 - 4ac \\
& = 4 \left(\sum_{j=1}^m \pi_j p_j - \sum_{j=1}^m p_j^2 \right)^2 \\
& \quad + 4\pi_1 \left(\frac{1}{p_1} - p_1 + \sum_{j=2}^m \frac{p_j^2}{p_1} \right) \left(1 - \sum_{j=1}^m p_j^2 \right) \\
& \quad + 4 \left(1 - 2 \sum_{j=1}^m \pi_j p_j + \sum_{j=1}^m p_j^2 \right) \left(1 - \sum_{j=1}^m p_j^2 \right) \\
& = 4 \left(1 - \sum_{j=1}^m \pi_j p_j \right)^2 \\
& \quad + \frac{4\pi_1}{p_1} \left(1 - p_1^2 + \sum_{j=2}^m p_j^2 \right) \left(1 - \sum_{j=1}^m p_j^2 \right) \\
& > 0.
\end{aligned}$$

By applying Result (iii) of the following lemma,

Lemma 1. *Let $a > 0$ and $D(f) = b^2 - 4ac$ denote the discriminant of a parabola $f(x) = ax^2 + bx + c$. We have*

(i) *If $D(f) < 0$, then $f(x) > 0$ for all $x \in (-\infty, \infty)$.*

- (ii) *If $D(f) = 0$, then $f(x) \geq 0$ for all $x \in (-\infty, \infty)$, and $f(x)$ reaches its minimum zero at $x = -b/(2a)$.*
(iii) *If $D(f) > 0$, then $f(x) > 0$ for all $x \in (-\infty, x_1) \cup (x_2, \infty)$ and $f(x) \leq 0$ for all $x \in [x_1, x_2]$, where $x_1 = [-b - \sqrt{D(f)}]/(2a)$ and $x_2 = [-b + \sqrt{D(f)}]/(2a)$.*

We immediately obtain the following theorem.

Theorem 2. *Let $\boldsymbol{\pi} \in \mathbb{T}_m$ and $\mathbf{p} \in \mathbb{T}_m$, we always have $\text{athrmtr}[\text{Var}(\hat{\boldsymbol{\pi}}_{\text{MT}})] > \text{tr}[\text{Var}(\hat{\boldsymbol{\pi}}_{\text{MP}})]$ for any $q \in (0, q_L) \cup (q_U, 1)$, where*

$$\begin{aligned}
q_L & = \max \left\{ 0, \frac{-b - \sqrt{b^2 - 4ac}}{2a} \right\} \quad \text{and} \\
q_U & = \min \left\{ 1, \frac{-b + \sqrt{b^2 - 4ac}}{2a} \right\},
\end{aligned}$$

where a , b and c are defined in (5.6).

5.3 Degree of privacy protection

In this subsection, we compare *degrees of privacy protection* (DPP) of the multi-category parallel model with those of the multi-category triangular model. For the multi-category parallel model (see Table 1), we define

$$\begin{aligned}
\text{DPP}_{\text{MP}}(\pi_1, p_1, q) & = \Pr(Y = 1 | \text{two circles are connected}), \\
\text{DPP}_{\text{MP}}(\pi_2, p_2, q) & = \Pr(Y = 2 | \text{two triangles are connected}), \\
& \quad \vdots \\
\text{DPP}_{\text{MP}}(\pi_m, p_m, q) & = \Pr(Y = m | \text{two dots are connected}),
\end{aligned}$$

where, for example, $\text{DPP}_{\text{MP}}(\pi_1, p_1, q)$ denotes the conditional probability that the respondent belongs to the subclass $\{Y = 1\}$ given that he/she connected the two circles. For the multi-category triangular model (see Table 4), we can similarly define

$$\begin{aligned}
\text{DPP}_{\text{MT}}(\pi_1, p_1) & = \Pr(Y = 1 | \text{a tick is put in Block 1}), \\
\text{DPP}_{\text{MT}}(\pi_2, p_2) & = \Pr(Y = 2 | \text{a tick is put in Block 2}), \\
& \quad \vdots \\
\text{DPP}_{\text{MT}}(\pi_m, p_m) & = \Pr(Y = m | \text{a tick is put in Block } m).
\end{aligned}$$

First, for any $q \in (0, 1)$, $\boldsymbol{\pi} \in \mathbb{T}_m$ and $\mathbf{p} \in \mathbb{T}_m$, we always have

$$(5.7) \quad \text{DPP}_{\text{MT}}(\pi_1, p_1) = 1 > \frac{\pi_1 q}{p_1(1-q) + \pi_1 q} = \text{DPP}_{\text{MP}}(\pi_1, p_1, q).$$

Next, when $0 < q < \frac{1}{1+\pi_1}$, we obtain

$$\begin{aligned}
(5.8) \quad \text{DPP}_{\text{MT}}(\pi_j, p_j) & = \frac{\pi_j}{p_j \pi_1 + \pi_j} \\
& > \frac{\pi_j q}{p_j(1-q) + \pi_j q} \\
& = \text{DPP}_{\text{MP}}(\pi_j, p_j, q), \quad j = 2, \dots, m,
\end{aligned}$$

Table 6. Survey data from Williamson and Haber (1994)

Number of sex partners	Income		
	$Y = 0$ (low)	$Y = 1$ (high)	Missing
$X = 0$ (0-3)	144 (m_1, π_1)	123 (m_2, π_2)	17
$X = 1$ (≥ 4)	237 (m_3, π_3)	148 (m_4, π_4)	17

for any $\boldsymbol{\pi} \in \mathbb{T}_m$ and $\boldsymbol{p} \in \mathbb{T}_m$. Inequalities (5.7) and (5.8) show that if we choose q within the open interval $(0, \frac{1}{1+\pi_1})$, the multi-category parallel model is more efficient than the multi-category triangular model in protecting the individual's privacy for any $\boldsymbol{\pi} \in \mathbb{T}_m$ and $\boldsymbol{p} \in \mathbb{T}_m$.

Williamson and Haber [16] reported a study aimed to examine the relationship among disease status of cervical cancer, the number of sexual partners and income. Respondents were women of 20–79 year old in Fulton or Dekalb County in Atlanta, Georgia. Table 6 displays the cross-classification of income (low or high, denoted by $Y = 0$ or $Y = 1$) and number of sex partners ('few' (0-3) or 'many' (≥ 4), denoted by $X = 0$ or $X = 1$). Since all four questions (i.e., the number of sex partners and income status) are highly sensitive to respondents, a sizable proportion (19.9% in this example) of the responses would be missing because of 'unknown' or 'refused to answer' in a telephone interview. The major objective is to examine if association exists between the number of sex partners and income. The existing multi-category triangular model and the corresponding statistical methods [11] cannot be applied to such studies because each of the four subclasses $\{X = 0, Y = 0\}$, $\{X = 0, Y = 1\}$, $\{X = 1, Y = 0\}$ and $\{X = 1, Y = 1\}$ is sensitive to respondents. To demonstrate the proposed multi-category parallel design in Tables 1 and 2 and the developed estimation methods in Sections 3 and 4, we let $m = 4$ and define $W = 0$ if the respondent's birthday is in the first half of a month and $W = 1$ otherwise. Similarly, we define $U = i$ if the respondent was born in the i -th quarter of a year ($i = 1, \dots, 4$). Thus, it is reasonable to assume that $q = \Pr(W = 1) = 0.5$ and $p_i = \Pr(U = i) = 0.25$ for each i .

6. AN EXAMPLE

To obtain the observed data $Y_{\text{obs}} = \{n; n_1, \dots, n_4\}$ in the four-category parallel model (see Table 2), we only consider the complete observations in Table 6 and discard the associated missing data and obtain $n = m_1 + \dots + m_4 = 144 + 123 + 237 + 148 = 652$. Note that n_1 denotes the number of respondents connecting the two circles in Table 2, n_2 is the number of respondents connecting the two triangles, n_3 is the number of respondents connecting the two rectangles, and n_4 is the number of respondents connecting the two dots. Let z_1, z_2, z_3 and z_4 denote the number of respondents belonging to $\{X = 0, Y = 0\} \cap \{W = 1\}$, $\{X = 0, Y = 1\} \cap \{W = 1\}$, $\{X = 1, Y = 0\} \cap \{W = 1\}$ and $\{X = 1, Y = 1\} \cap \{W = 1\}$ in Table 2 respectively. Since $q = 1/2$, we have $z_i = m_i/2$ for the ideal

situation, i.e., $(z_1, z_2, z_3, z_4)^\top \approx (72, 62, 118, 74)^\top$. Furthermore, let n'_i denote the number of respondents belonging to $\{U = i\} \cap \{W = 0\}$ for $i = 1, \dots, 4$ in Table 2. To obtain these $\{n'_i\}_{i=1}^4$ by considering the sampling error, we first generate 50 i.i.d. samples from

$$\begin{aligned} & \text{Multinomial} \left(n - \sum_{i=1}^4 z_i; p_1, p_2, p_3, p_4 \right) \\ & = \text{Multinomial}(326; 0.25 \times \mathbf{1}_4) \end{aligned}$$

and then average these counts for each component, yielding $(n'_1, n'_2, n'_3, n'_4)^\top = (81, 82, 81, 82)^\top$. Therefore, we obtain the following observed counts

$$\begin{aligned} & (n_1, n_2, n_3, n_4)^\top \\ & = (z_1 + n'_1, z_2 + n'_2, z_3 + n'_3, z_4 + n'_4)^\top \\ & = (153, 144, 199, 156)^\top. \end{aligned}$$

Using $\boldsymbol{\pi}^{(0)} = 0.25 \times \mathbf{1}_4$ as the initial values, the EM algorithm in (3.3) and (3.5) converged in 25 iterations. The resultant MLEs for $\boldsymbol{\pi} = (\pi_1, \dots, \pi_4)^\top$ and the odds ratio ψ are listed in the second column of Table 7. Based on (3.6), we generate $G = 10,000$ bootstrap samples to estimate the standard errors of $\{\hat{\pi}_{\text{MP}i}\}_{i=1}^4$ and $\hat{\psi}$, which are given in the third column of Table 7. The corresponding 95% normal-based bootstrap CIs and non-normal-based bootstrap CIs are displayed in the fourth and the fifth columns of Table 7. Since the two bootstrap CIs of ψ include 1, we do not have reason to believe that there exists association between the number of sex partners and income. According to (3.9), we obtain $\hat{\boldsymbol{\pi}}_v = (0.2193252, 0.1917178, 0.3604294, 0.2285276)^\top$. Since $\hat{\boldsymbol{\pi}}_v \in \mathbb{T}_4$, we know that $\hat{\boldsymbol{\pi}}_v$ is a valid estimator of $\boldsymbol{\pi}$ and $\hat{\boldsymbol{\pi}}_v = \hat{\boldsymbol{\pi}}_{\text{MP}}$. Based on (3.12), the estimated variance-covariance matrix of $\hat{\boldsymbol{\pi}}_{\text{MP}}$ is

$$\begin{aligned} & \widehat{\text{Var}}(\hat{\boldsymbol{\pi}}_{\text{MP}}) \\ & = \begin{pmatrix} 0.00110 & -0.00032 & -0.00044 & -0.00034 \\ -0.00032 & 0.00106 & -0.00041 & -0.00032 \\ -0.00044 & -0.00041 & 0.00130 & -0.00045 \\ -0.00034 & -0.00032 & -0.00045 & 0.00112 \end{pmatrix} \end{aligned}$$

so that the unbiased estimates of $\{\text{Var}(\hat{\pi}_{\text{MP}i})\}_{i=1}^4$, from (3.13), are given by

$$\begin{aligned} & (\widehat{\text{Var}}(\hat{\pi}_{\text{MP}1}), \dots, \widehat{\text{Var}}(\hat{\pi}_{\text{MP}4}))^\top \\ & = (0.00110, 0.00106, 0.00130, 0.00112)^\top. \end{aligned}$$

Therefore, from (3.14), (3.16) and (3.18), the 95% Wald, Wilson and likelihood ratio CIs of $\{\pi_i\}_{i=1}^4$ can be calculated and are given in the second, the fourth and sixth columns of Table 8. We noted that the width of the 95% Wilson CI of π_i is slightly shorter than those of the 95% Wald CIs and LRCIs of π_i .

Table 7. MLEs and two bootstrap CIs of parameters for the observed counts $(n_1, n_2, n_3, n_4)^T = (153, 144, 199, 156)^T$

Parameter	std		95% bootstrap CI [†]	95% bootstrap CI [‡]
π_1	0.2196	0.0328	[0.1553, 0.2839]	[0.1549, 0.2837]
π_2	0.1919	0.0327	[0.1277, 0.2561]	[0.1273, 0.2561]
π_3	0.3604	0.0359	[0.2900, 0.4308]	[0.2929, 0.4310]
π_4	0.2281	0.0338	[0.1619, 0.2943]	[0.1641, 0.2960]
ψ	0.7661	0.2712	[0.2344, 1.2977]	[0.3690, 1.4099]

CI[†]: Normal-based bootstrap CIs, cf. (3.7). CI[‡]: Non-normal-based bootstrap CIs, cf. (3.8).

Table 8. Three asymptotic 95% CIs of parameters for large sample sizes

Parameter	Wald CI	Width	Wilson CI	Width	LRCI	Width
π_1	[0.1542, 0.2844]	0.1302	[0.1575, 0.2874]	0.1299	[0.1564, 0.2864]	0.1300
π_2	[0.1280, 0.2554]	0.1274	[0.1314, 0.2586]	0.1272	[0.1303, 0.2575]	0.1272
π_3	[0.2897, 0.4312]	0.1415	[0.2922, 0.4332]	0.1410	[0.2914, 0.4326]	0.1412
π_4	[0.1630, 0.2941]	0.1311	[0.1662, 0.2970]	0.1308	[0.1652, 0.2960]	0.1308

7. DISCUSSION

As a natural generalization of the NRR parallel model of Tian [12], we develop an NRR multi-category parallel model for a single sensitive question with multiple answers/outcomes. When comparing with the existing NRR multi-category triangular model, the newly developed model has several significant advantages: (i) *A wider applicability.* The multi-category parallel model can be applied to such situations where all population subclasses could be sensitive, while the former is only applicable in the case that at least one of the population subclasses is non-sensitive. (ii) *A higher efficiency.* Because of the introduction of additional parameter $q = \Pr(W = 1) \in (0, 1)$, the multi-category parallel model is more efficient than the multi-category triangular model for a certain range of q (see Theorem 2 in Section 5.2). (iii) *A better degree of privacy protection.* The comparisons in Section 5.3 show that if $0 < q < \frac{1}{1+\pi_1}$, the multi-category parallel model is more efficient than the multi-category triangular model in protecting the individual's privacy for any $\boldsymbol{\pi} \in \mathbb{T}_m$ and $\boldsymbol{p} \in \mathbb{T}_m$.

How to choose the two non-sensitive variables W and U in Table 1 is an important issue in practice. On the one hand, since W is a binary variable, we could define $W = 0$ if the respondent was born between January and June; or the respondent was born in an odd numbered month; or the respondent's birthday is in the first half of the month; or the respondent's age is odd numbered; or the respondent's house/apartment number is even. On the other hand, since U is an m -category variable, for example, when $m = 3$, we may let

- $U = 1$ if the respondent's mother was born in January–April;
- $U = 2$ if the respondent's mother was born in May–August; and
- $U = 3$ if the respondent's mother was born in September–December.

In this case, it is reasonable to assume that each $p_i = \Pr(U = i)$ is approximately equal to $1/3$. Similarly, when $m = 5$, we may define

- $U = 1$ if the last digit of the respondent's ID card/phone number 1 or 2;
- $U = 2$ if the last digit of the respondent's ID card/phone number 3 or 4;
- $U = 3$ if the last digit of the respondent's ID card/phone number 5 or 6;
- $U = 4$ if the last digit of the respondent's ID card/phone number 7 or 8; and
- $U = 5$ if the last digit of the respondent's ID card/phone number 9 or 0.

In Section 2, we assumed that $q = \Pr(W = 1)$ and all $p_i = \Pr(U = i)$ for $i = 1, \dots, m$ are known. When $m = 2$ and p_1 is unknown, Liu and Tian [9] further developed a variant of the parallel model [12] for sample surveys with sensitive characteristics. When $m \geq 3$ and q or/and $\{p_i\}_{i=1}^m$ is unknown, it is worthwhile to investigate the corresponding multi-category parallel model.

ACKNOWLEDGEMENTS

The authors would like to thank the Editor, an AE and two referees for their comments and valuable suggestions. GL Tian's research was fully supported by a grant (HKU 779210M) from the Research Grant Council of the Hong Kong Special Administrative Region.

Received 20 September 2011

REFERENCES

- [1] ABUL-ELA, A. A., GREENBERG, B. G. AND HORVITZ, D.G. (1967). A multi-proportions randomized response model. *Journal of the American Statistical Association* **62** 990–1008. [MR0217982](#)

- [2] AGRESTI, A. AND COULL, B. A. (1998). Approximate is better than “exact” for interval estimation of binomial proportions. *The American Statistician* **52** 119–126. [MR1628435](#)
- [3] BOURKE, P. D. (1974). Multi-proportions randomized response using the unrelated question technique. Technical Report No. 74 of the Errors on Surveys Research Project. Institute of Statistics, University of Stockholm (Mimeo).
- [4] BOURKE, P. D. AND DALENIUS, T. (1973). Multi-proportions randomized response using a single sample. Report No. 68 of the Errors on Surveys Research Project. Institute of Statistics, University of Stockholm (Mimeo).
- [5] BROWN, L. D., CAI, T. T. AND DASGUPTA, A. (2001). Interval estimation for a binomial proportion. *Statistical Science* **16** 101–133. [MR1861069](#)
- [6] CLOPPER, C. J. AND PEARSON, E. S. (1934). The Use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika* **26**(4) 404–413.
- [7] DEMPSTER, A. P., LAIRD, N. M. AND RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B* **39**(1) 1–38. [MR0501537](#)
- [8] ERIKSSON, S. A. (1973). A new model for randomized response. *International Statistical Review* **41** 101–113.
- [9] LIU, Y. AND TIAN, G. L. (2012). A variant of the parallel model for sample surveys with sensitive characteristics. Technical Report of the Department of Statistics and Actuarial Science, The University of Hong Kong.
- [10] NEWCOMBE, R. G. (1998). Two-sided confidence intervals for the single proportion: comparison of seven methods. *Statistics in Medicine* **17**(8) 857–872.
- [11] TANG, M. L., TIAN, G. L., TANG, N. S. AND 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. [MR2750777](#)
- [12] TIAN, G. L. (2012). A new non-randomized response model: The parallel model. Technical Report of the Department of Statistics and Actuarial Science, The University of Hong Kong.
- [13] TIAN, G. L., YU, J. W., TANG, M. L. AND GENG, Z. (2007). A new non-randomized model for analyzing sensitive questions with binary outcomes. *Statistics in Medicine* **26**(23) 4238–4252. [MR2405351](#)
- [14] TIAN, G. L., TANG, M. L., LIU, Z. Q., TAN, M. AND 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**(3) 159–173. [MR2828973](#)
- [15] WARNER, S. L. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association* **60** 63–69.
- [16] WILLIAMSON, G. D. AND HABER, M. (1994). Models for three-dimensional contingency tables with completely and partially cross-classified data. *Biometrics* **50** 194–203. [MR1279436](#)
- [17] YU, J. W., TIAN, G. L. AND TANG, M. L. (2008). Two new models for survey sampling with sensitive characteristic: Design and analysis. *Metrika* **67** 251–263. [MR2390876](#)

Yin Liu
 Department of Statistics and Actuarial Science
 The University of Hong Kong, Pokfulam Road
 Hong Kong
 P. R. China
 E-mail address: liuyin31@hku.hk

Guo-Liang Tian
 Rm 520, Meng Wah Complex, Pokfulam Road
 Hong Kong
 P. R. China
 E-mail address: gltian@hku.hk