Hindawi Publishing Corporation
Journal of Probability and Statistics
Volume 2013, Article ID 827048, 11 pages
http://dx.doi.org/10.1155/2013/827048

Research Article

A Survey Design for a Sensitive Binary Variable Correlated with Another Nonsensitive Binary Variable

Jun-Wu Yu,[1] Yang Lu,[2] and Guo-Liang Tian[3]

[1] *School of Mathematics and Computational Science, Hunan University of Science and Technology, Xiangtan, Hunan 411201, China*
[2] *School of Mathematics and Statistics, Central China Normal University, Wuhan, Hubei 430079, China*
[3] *Department of Statistics and Actuarial Science, The University of Hong Kong, Pokfulam Road, Hong Kong*

Correspondence should be addressed to Guo-Liang Tian; gltian@hku.hk

Received 30 August 2012; Accepted 19 May 2013

Academic Editor: Zhidong Bai

Tian et al. (2007) introduced a so-called hidden sensitivity model for evaluating the association of two sensitive questions with binary outcomes. However, in practice, we sometimes need to assess the association between one sensitive binary variable (e.g., whether or not a drug user, the number of sex partner being ≤1 or >1, and so on) and one nonsensitive binary variable (e.g., good or poor health status, with or without cervical cancer, and so on). To address this issue, by sufficiently utilizing the information contained in the non-sensitive binary variable, in this paper, we propose a new survey scheme, called combination questionnaire design/model, which consists of a main questionnaire and a supplemental questionnaire. The introduction of the supplemental questionnaire which is indeed a design of direct questioning can effectively reduce the noncompliance behavior since more respondents will not be faced with the sensitive question. Likelihood-based inferences including maximum likelihood estimates via the expectation-maximization algorithm, asymptotic confidence intervals, and bootstrap confidence intervals of parameters of interest are derived. A likelihood ratio test is provided to test the association between the two binary random variables. Bayesian inferences are also discussed. Simulation studies are performed, and a cervical cancer data set in Atlanta is used to illustrate the proposed methods.

1. Introduction

Warner [1] introduced a randomized response technique to obtain truthful answers to questions with sensitive attributes. Using the Warner design, Kraemer [2] derived a bivariate correlation between an attribute with polytomous responses and an attribute with normally distributed responses. Fox and Tracy [3] derived estimation of the Pearson product moment correlation coefficient between two sensitive questions by assuming that randomized response observations can be treated as individual-level scores that are contaminated by random measurement error. Edgell et al. [4] considered the correlation between two sensitive questions using the unrelated question design or the additive constants design. Christofides [5] presented a randomized response technique with two randomization devices to estimate the proportion of individuals having two sensitive characteristics at the same time. Kim and Warde [6] considered a multinomial randomized response model which can handle untruthful responses. They also derived the Pearson product moment correlation estimator which may be used to quantify the linear relationship between two variables when multinomial response data are observed according to a randomized response procedure. However, all these randomized response procedures make use of one or two randomizing devices which (i) entail extra costs in both efficiency and complexity, (ii) increase the cognitive load of randomized response techniques, and (iii) allow for new sources of error, such as misunderstanding the randomized response procedures or cheating on the procedures [7].

From the perspective of incomplete categorical data design, Tian et al. [8] proposed a nonrandomized response model (called the hidden sensitivity model) for assessing the association of two sensitive questions with binary outcomes. To protect respondents' privacy and to avoid the use of any randomization device, they utilized a non-sensitive question in the questionnaire to indirectly obtain respondents' answers to the two sensitive questions. In the hidden sensitivity model, they implicitly assumed that all respondents are willing to follow the design instructions. In other words, the noncompliance behavior will not occur. However, in practice, we sometimes need to assess the association between one sensitive binary variable (e.g., whether or not a drug user, the number of sex partner being $\leqslant 1$ or >1, and so on) and one nonsensitive binary variable (e.g., good or poor health status, with or without cervical cancer, and so on). To our knowledge, the survey design for addressing this issue and corresponding statistical analysis methods are not available. Although we could directly adopt the hidden sensitivity model, the information contained in the nonsensitive binary variable cannot be utilized in the design. Intuitively, such information can be used to enhance the degree of privacy protection, so that more respondents will not be faced with the sensitive question. The major objective of this paper is to propose a new survey design and to develop corresponding statistical methods for analyzing sensitive data collected by this technique.

The rest of the paper is organized as follows. In Section 2, without using any randomizing device, we propose a survey scheme, called combination questionnaire design/model, which consists of a main questionnaire and a supplemental questionnaire. Likelihood-based inferences including maximum likelihood estimates via an *expectation–maximization* (EM) algorithm, asymptotic confidence intervals, and bootstrap confidence intervals of parameters of interest are derived in Section 3. A likelihood ratio test is also provided to test the association between the two binary random variables. In Section 4, we discuss Bayesian inferences when prior information on parameters is available. In Section 5, two simulation studies are performed to compare the efficiency of the proposed combination questionnaire model with that of the existing hidden sensitivity model of Tianet al. [8] (i.e., the main questionnaire only). A cervical cancer data set in Atlanta is used in Section 6 to illustrate the proposed methods. A discussion and an appendix on the mode of a group Dirichlet density and a sampling method from it are also presented.

2. The Survey Design

Assume that X is a sensitive binary random variable, Y is a non-sensitive binary random variable, and they are correlated. Let $\{X = 1\}$ denote the sensitive class (e.g., $X = 1$ if a respondent is a drug user) and $\{X = 0\}$ denote the non-sensitive class (e.g., $X = 0$ if a respondent is not a drug user). Furthermore, let both $\{Y = 1\}$ (e.g., $Y = 1$ if a respondent receives at least some college training) and $\{Y = 0\}$ (e.g., $Y = 0$ if a respondent graduates at most from some high school) be non-sensitive classes. Define $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_4)^{\top}$, where $\theta_1 = \Pr(X = 0, Y = 0)$, $\theta_2 = \Pr(X = 1, Y = 0)$, $\theta_3 = \Pr(X = 1, Y = 1)$, and $\theta_4 = \Pr(X = 0, Y = 1)$, then $\boldsymbol{\theta} \in \mathbb{T}_4$, where

$$\mathbb{T}_n \mathrel{\widehat{=}} \left\{ (x_1, \ldots, x_n)^{\top} : \ x_i \geqslant 0, \ i = 1, \ldots, n, \ \sum_{i=1}^{n} x_i = 1 \right\} \quad (1)$$

denotes the n-dimensional closed simplex in $\mathbb{R}^n$. The objective is to make inferences on $\boldsymbol{\theta}$, $\theta_x \mathrel{\widehat{=}} \Pr\,(X = 1) = \theta_2 + \theta_3$, $\theta_y \mathrel{\widehat{=}} \Pr(Y = 1) = \theta_4 + \theta_3$ and the odds ratio $\psi = \theta_1\theta_3/(\theta_2\theta_4)$.

Table 1: The main questionnaire for the combination questionnaire model.

Category	$W = 1$	$W = 2$	$W = 3$
I: $\{X = 0,\ Y = 0\}$	Block 1: —	Block 2: —	Block 3: —
II: $\{X = 1,\ Y = 0\}$	Category II: Please put a tick in Block 2		
III: $\{X = 1,\ Y = 1\}$	Category III: Please put a tick in Block 3		
IV: $\{X = 0,\ Y = 1\}$	Block 4: —		

Note: Only $\{X = 1\}$ is a sensitive class, while $\{X = 0\}$, $\{Y = 0\}$, and $\{Y = 1\}$ are nonsensitive classes.

Table 2: The supplemental questionnaire for the combination questionnaire model.

Category	
$\{Y = 0\}$	Please put a tick in Block 5: — if you belong to $\{Y = 0\}$
$\{Y = 1\}$	Please put a tick in Block 6: — if you belong to $\{Y = 1\}$

Note: Both the $\{Y = 0\}$ and $\{Y = 1\}$ are nonsensitive classes.

The survey scheme consists of a main questionnaire and a supplemental questionnaire. To design the main questionnaire which is to be assigned to group 1 with n respondents (n is specified by the investigators), we first introduce a nonsensitive question (say, Q_W) with three possible answers. Assume that W is a non-sensitive variate with trichotomous outcomes associated with the Q_W and W is independent of both X and Y. Define $p_i = \Pr(W = i)$ for $i = 1, 2, 3$. For example, let $W = 1$ $(2, 3)$ denote that a respondent was born in January–April (May–August, September–December), and thus we could assume that $p_i \approx 1/3$. The main questionnaire is shown in Table 1, under which each respondent is asked to answer the non-sensitive question Q_W.

On the one hand, since Category I (i.e., $\{X = 0, Y = 0\}$) and Category IV (i.e., $\{X = 0, Y = 1\}$) are non-sensitive to each respondent, it is reasonable to assume that a respondent is willing to provide his/her truthful answer by putting a tick in Block i ($i = 1, \ldots, 4$) according to his/her true status. On the other hand, Category II (i.e., $\{X = 1, Y = 0\}$) and Category III (i.e., $\{X = 1, Y = 1\}$) are usually sensitive to respondents. In this case, if a respondent belongs to Category II (III), he/she is designed to put a tick in Block 2 (3).

The supplemental questionnaire is designed as shown in Table 2, under which m respondents (m is also specified by the investigators) in group 2 are asked to put a tick in Block 5 or Block 6 depending on their true status; that is, $\{Y = 0\}$ or $\{Y = 1\}$. Since both the $\{Y = 0\}$ and $\{Y = 1\}$ are nonsensitive classes, the supplemental questionnaire is in fact a design of direct questioning. Therefore, we call this design the combination questionnaire model.

Table 3 shows the cell probabilities $\{\theta_i\}_{i=1}^{4}$, the observed frequencies $\{n_i\}_{i=1}^{4}$ and the unobservable frequencies $\{Z_i\}_{i=1}^{3}$, for the main questionnaire. The observed frequency n_2 is the sum of the frequency of respondents belonging to Block 2 and

Table 3: Cell probabilities, observed and unobservable counts for the main questionnaire.

Category	$W = 1$	$W = 2$	$W = 3$	Total
I: $\{X = 0,\ Y = 0\}$	$p_1\theta_1$	$p_2\theta_1$	$p_3\theta_1$	$\theta_1(Z_1)$
II: $\{X = 1,\ Y = 0\}$				$\theta_2(Z_2)$
III: $\{X = 1,\ Y = 1\}$				$\theta_3(Z_3)$
IV: $\{X = 0,\ Y = 1\}$		$\theta_4(n_4)$		$\theta_4(n_4)$
Total	$p_1\theta_1(n_1)$	$p_2\theta_1 + \theta_2(n_2)$	$p_3\theta_1 + \theta_3(n_3)$	$1(n)$

Note: $n = \sum_{i=1}^{4} n_i$, $Z_1 = n - (Z_2 + Z_3) - n_4$, where (Z_2, Z_3) are unobservable.

Table 4: Cell probabilities and observed counts for the supplemental questionnaire.

Category		Total
$\{Y = 0\}$	Block 5: —	$\theta_1 + \theta_2(m_0)$
$\{Y = 1\}$	Block 6: —	$\theta_3 + \theta_4(m_1)$
Total		$1\ (m)$

Note: $m = m_0 + m_1$.

the frequency of those belonging to Category II. The observed frequency n_3 is the sum of the frequency of respondents belonging to Block 3 and the frequency of those belonging to Category III. Note that $n = \sum_{i=1}^{4} n_i = \sum_{i=1}^{3} Z_i + n_4$, we have $Z_1 = n-(Z_2+Z_3)-n_4$. Thus, only Z_2 and Z_3 are unobservable. Table 4 shows the cell probabilities and observed counts for the supplemental questionnaire.

Remark 1. The design of the main questionnaire is similar to that of the hidden sensitivity model of Tian et al. [8], while the design of the supplemental questionnaire is indeed a design of direct questioning since both the $\{Y = 0\}$ and $\{Y = 1\}$ are non-sensitive classes. Table 1 shows that Categories II and III are two sensitive subclasses. Therefore, putting a tick in Block 2 or Block 3 implies that the respondent could be suspected with the sensitive attribute. Let $\theta_1 = \cdots = \theta_4 \approx 0.25$ (see Table 3); then around half of the, say, $2n$ respondents will be suspected with the sensitive attribute if only the main questionnaire is employed. However, besides the main questionnaire (with n respondents), if the supplemental questionnaire (see Table 2) with $m = n$ respondents is also used, then only half of the n respondents will be suspected with the sensitive attribute. In other words, in the proposed combination questionnaire model, the information of the non-sensitive binary variable Y can be used to enhance the degree of privacy protection, so that more respondents will not be faced with the sensitive question. This is why we introduce the supplemental questionnaire besides the main questionnaire.

Remark 2. In practice, to simplify the design itself, we suggest that both the sample size n in the main questionnaire and the sample size m in the supplemental questionnaire should be fixed in advance rather than the total sample size $N = n + m$ being fixed. In this way, survey data can be collected in two independent groups, resulting in a relatively simpler statistical analysis. In addition, the interviewees are randomly assigned to either the group 1 or group 2.

3. Likelihood-Based Inferences

In this section, *maximum likelihood estimates* (MLEs) of the $\boldsymbol{\theta}$ and the odds ratio ψ are derived by using the *EM* algorithm. In addition, asymptotic confidence intervals and the bootstrap confidence intervals of an arbitrary function of $\boldsymbol{\theta}$ are also provided. Finally, a likelihood ratio test is presented for testing the association between the two binary random variables.

3.1. MLEs via the EM Algorithm. A total of $N = n + m$ respondents are classified into two groups by a randomization approach such that n respondents answer the questions in the main questionnaire and m respondents answer the questions in the supplemental questionnaire. Let $Y_{\text{obs,M}} = \{n;\ n_1, \ldots, n_4\}$ denote the observed counts collected in the main questionnaire (see Table 3), where $\sum_{i=1}^{4} n_i = n$. The likelihood function of $\boldsymbol{\theta}$ based on $Y_{\text{obs},M}$ is

$$L(\boldsymbol{\theta} \mid Y_{\text{obs},M}) = \binom{n}{n_1, n_2, n_3, n_4} \times (p_1\theta_1)^{n_1} \left\{ \prod_{i=2}^{3} (p_i\theta_1 + \theta_i)^{n_i} \right\} \theta_4^{n_4}. \tag{2}$$

Let $Y_{\text{obs},S} = \{m; m_0, m_1\}$ denote the observed counts gathered in the supplemental questionnaire (see Table 4), where $m_0 + m_1 = m$. The likelihood function of $\boldsymbol{\theta}$ based on $Y_{\text{obs},S}$ is

$$L(\boldsymbol{\theta} \mid Y_{\text{obs},S}) = \binom{m}{m_0} (\theta_1 + \theta_2)^{m_0} (\theta_3 + \theta_4)^{m_1}. \tag{3}$$

Let $Y_{\text{obs}} = \{Y_{\text{obs},M}, Y_{\text{obs},S}\}$. Since $Y_{\text{obs},M}$ and $Y_{\text{obs},S}$ are independent, the observed-data likelihood function of $\boldsymbol{\theta} \in \mathbb{T}_4$ is

$$L_{\text{CQ}}(\boldsymbol{\theta} \mid Y_{\text{obs},S}) \propto \theta_1^{n_1} \left\{ \prod_{i=2}^{3} (p_i\theta_1 + \theta_i)^{n_i} \right\} \theta_4^{n_4} \times (\theta_1 + \theta_2)^{m_0} (\theta_3 + \theta_4)^{m_1}, \tag{4}$$

where the subscript "CQ" denotes the "combination questionnaire" model.

Since the corresponding cell probabilities to the observed counts n_2 and n_3 in the group 1 are in the form of summation (i.e., $p_i\theta_1 + \theta_i$, $i = 2, 3$), we cannot obtain the explicit expressions of the MLEs of $\boldsymbol{\theta}$ from the score equations of (4). By treating the observed counts n_2 and n_3 as incomplete data, we use the EM algorithm [9] to find the MLE $\widehat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$. The counts $\{Z_i\}_{i=1}^{3}$ in Table 3 can be viewed as missing data. Briefly, Z_1, Z_2, Z_3, and n_4 represent the counts of the respondents belonging to Categories I, II, III, and IV, respectively. Thus, we denote the latent data by $Y_{\text{mis}} = \{z_2, z_3\}$ and the complete data by $Y_{\text{com}} = \{Y_{\text{obs}}, Y_{\text{mis}}\}$. Note that all $\{p_i\}$ are known. Consequently, the complete-data likelihood function for $\boldsymbol{\theta}$ is

$$L_{\text{CQ}}(\boldsymbol{\theta} \mid Y_{\text{COM}}) \propto \left(\prod_{i=1}^{3} \theta_i^{z_i} \right) \theta_4^{n_4} \times (\theta_1 + \theta_2)^{m_0} (\theta_3 + \theta_4)^{m_1}, \tag{5}$$

where $z_1 = n - (z_2 + z_3) - n_4$.

By treating $\{\theta_i\}_{i=1}^4$ as random variables, we note that the complete-data likelihood function (5) has the density form of a grouped Dirichlet distribution [10]. Ng et al. [11] derived the mode of a grouped Dirichlet density with explicit expressions (see the appendix). Hence, from (A.4) and (A.5), the complete-data MLEs for $\boldsymbol{\theta}$ are given by

$$\begin{aligned} \theta_1 &= 1 - \theta_2 - \theta_3 - \theta_4, \\ \theta_2 &= \frac{z_2}{N}\left(1 + \frac{m_0}{n - z_3 - n_4}\right), \\ \theta_3 &= \frac{z_3}{N}\left(1 + \frac{m_1}{z_3 + n_4}\right), \\ \theta_4 &= \frac{n_4}{N}\left(1 + \frac{m_1}{z_3 + n_4}\right). \end{aligned} \tag{6}$$

Given Y_{obs} and $\boldsymbol{\theta}$, Z_i follows the binomial distribution with parameters n_{i} and $\theta_i/(p_i\theta_1 + \theta_i)$; that is,

$$Z_i \mid (Y_{\text{obs}}, \boldsymbol{\theta}) \sim \text{Binomial}\left(n_i, \frac{\theta_i}{p_i\theta_1 + \theta_i}\right), \quad i = 2, 3. \tag{7}$$

Therefore, the E-step of the EM algorithm computes the following conditional expectations:

$$E\left(Z_i \mid Y_{\text{obs}}, \boldsymbol{\theta}\right) = \frac{n_i\theta_i}{p_i\theta_1 + \theta_{\text{i}}}, \quad i = 2, 3, \tag{8}$$

and the M-step updates (6) by replacing z_2 and z_3 with previous conditional expectations.

Remark 3. Based on the observed-data likelihood function (4), we could use the Newton-Raphson algorithm to find the MLEs of $\boldsymbol{\theta}$. However, it is well known that the Newton-Raphson algorithm does not necessarily increase the log likelihood, leading even to divergence sometimes [12, page 172]. In addition, the Newton–Raphson algorithm is sensitive to the initial values. One advantage of using the *EM* algorithm in the current situation is that both the E- and M-step have closed-form expressions. More importantly, the *EM* algorithm and the data augmentation algorithm of Tanner and Wong [13] share the same data augmentation structure in the Bayesian settings (see Section 4 for more details).

3.2. Asymptotic Confidence Intervals. Let $\boldsymbol{\theta}_{-4} = (\theta_1, \theta_2, \theta_3)^\top$. The asymptotic variance-covariance matrix of the MLE $\widehat{\boldsymbol{\theta}}_{-4}$ is then given by $\mathbf{I}_{\text{obs}}^{-1}(\widehat{\boldsymbol{\theta}}_{-4})$, where

$$\mathbf{I}_{\text{obs}}\left(\boldsymbol{\theta}_{-4}\right) = -\frac{\partial^2 \ell_{\text{CQ}}\left(\boldsymbol{\theta} \mid Y_{\text{obs}}\right)}{\partial\boldsymbol{\theta}_{-4}\partial\theta_{-4}^\top} \tag{9}$$

denotes the observed information matrix and $\ell_{\text{CQ}}(\boldsymbol{\theta} \mid Y_{\text{obs}}) = \log L_{\text{CQ}}(\boldsymbol{\theta} \mid Y_{\text{obs}})$ is the observed-data log-likelihood function. From (4), we have

$$\begin{aligned} \ell_{\text{CQ}}\left(\boldsymbol{\theta} \mid Y_{\text{obs}}\right) = {} & n_1 \log\theta_1 + n_2 \log\left(p_2\theta_1 + \theta_2\right) \\ & + n_3 \log\left(p_3\theta_1 + \theta_3\right) \\ & + n_4 \log\left(1 - \theta_1 - \theta_2 - \theta_3\right) \\ & + m_0 \log\left(\theta_1 + \theta_2\right) + m_1 \log\left(1 - \theta_1 - \theta_2\right). \end{aligned} \tag{10}$$

It is easy to show that

$$\begin{aligned} & \frac{\partial\ell_{\text{CQ}}\left(\boldsymbol{\theta} \mid Y_{\text{obs}}\right)}{\partial\theta_1} \\ & \quad = \frac{n_1}{\theta_1} - \frac{n_4}{\theta_4} + \sum_{\text{i}=2}^{3} \frac{n_{\text{i}}p_{\text{i}}}{p_{\text{i}}\theta_1 + \theta_{\text{i}}} + \frac{m_0}{\theta_1 + \theta_2} - \frac{m_1}{1 - \theta_1 - \theta_2}, \\ & \frac{\partial\ell_{\text{CQ}}\left(\boldsymbol{\theta} \mid Y_{\text{obs}}\right)}{\partial\theta_2} \\ & \quad = -\frac{n_4}{\theta_4} + \frac{n_2}{p_2\theta_1 + \theta_2} + \frac{m_0}{\theta_1 + \theta_2} - \frac{m_1}{1 - \theta_1 - \theta_2}, \\ & \frac{\partial\ell_{\text{CQ}}\left(\boldsymbol{\theta} \mid Y_{\text{obs}}\right)}{\partial\theta_3} = -\frac{n_4}{\theta_4} + \frac{n_3}{p_3\theta_1 + \theta_3}, \\ & -\frac{\partial^2\ell_{\text{CQ}}\left(\boldsymbol{\theta} \mid Y_{\text{obs}}\right)}{\partial\theta_1^2} = \frac{n_1}{\theta_1^2} + \frac{n_4}{\theta_4^2} + \sum_{i=2}^{3} \frac{n_i p_i^2}{\left(p_i\theta_1 + \theta_i\right)^2} + \phi \\ & -\frac{\partial^2\ell_{\text{CQ}}\left(\boldsymbol{\theta} \mid Y_{\text{obs}}\right)}{\partial\theta_2^2} = \frac{n_4}{\theta_4^2} + \frac{n_2}{\left(p_2\theta_1 + \theta_2\right)^2} + \phi \\ & -\frac{\partial^2\ell_{\text{CQ}}\left(\boldsymbol{\theta} \mid Y_{\text{obs}}\right)}{\partial\theta_3^2} = \frac{n_4}{\theta_4^2} + \frac{n_3}{\left(p_3\theta_1 + \theta_3\right)^2} \\ & -\frac{\partial^2\ell_{\text{CQ}}\left(\boldsymbol{\theta} \mid Y_{\text{obs}}\right)}{\partial\theta_1\partial\theta_2} = \frac{n_4}{\theta_4^2} + \frac{n_2 p_2}{\left(p_2\theta_1 + \theta_2\right)^2} + \phi \\ & -\frac{\partial^2\ell_{\text{CQ}}\left(\boldsymbol{\theta} \mid Y_{\text{obs}}\right)}{\partial\theta_1\partial\theta_3} = \frac{n_4}{\theta_4^2} + \frac{n_3 p_3}{\left(p_3\theta_1 + \theta_3\right)^2} \\ & -\frac{\partial^2\ell_{\text{CQ}}\left(\boldsymbol{\theta} \mid Y_{\text{obs}}\right)}{\partial\theta_2\partial\theta_3} = \frac{n_4}{\theta_4^2}, \end{aligned} \tag{11}$$

where

$$\phi = \frac{m_0}{\left(\theta_1 + \theta_2\right)^2} + \frac{m_1}{\left(1 - \theta_1 - \theta_2\right)^2}. \tag{12}$$

Hence, the observed information matrix can be expressed as

$$\mathbf{I}_{\text{obs}}\left(\boldsymbol{\theta}_{-4}\right) = \operatorname{diag}\left(\frac{n_1}{\theta_1^2}, 0, 0\right) + \frac{n_4}{\theta_4^2} \times \mathbf{1}_3\mathbf{1}_3^\top + \mathbf{A}, \tag{13}$$

where

$$\mathbf{A} = \begin{pmatrix} \sum_{i=2}^{3} \frac{n_i p_i^2}{(p_i\theta_1 + \theta_i)^2} + \phi, & \frac{n_2 p_2}{(p_2\theta_1 + \theta_2)^2} + \phi, & \frac{n_3 p_3}{(p_3\theta_1 + \theta_3)^2} \\ \frac{n_2 p_2}{(p_2\theta_1 + \theta_2)^2} + \phi, & \frac{n_2}{(p_2\theta_1 + \theta_2)^2} + \phi, & 0 \\ \frac{n_3 p_3}{(p_3\theta_1 + \theta_3)^2}, & 0, & \frac{n_3}{(p_3\theta_1 + \theta_3)^2} \end{pmatrix}. \tag{14}$$

Let $\text{se}(\widehat{\theta}_i)$ denote the standard error of $\widehat{\theta}_i$ for $i = 1, 2, 3$. Note that $\text{se}(\widehat{\theta}_i)$ can be estimated by the square root of the ith diagonal element of $\mathbf{I}_{\text{obs}}^{-1}(\widehat{\boldsymbol{\theta}}_{-4})$. We denote the estimated value of $\text{se}(\widehat{\theta}_i)$ by $\widehat{\text{se}}(\widehat{\theta}_i)$. Thus, a 95% normal-based asymptotic confidence interval for θ_{i} can be constructed as

$$\left[\widehat{\theta}_i - 1.96 \times \widehat{\text{se}}\left(\widehat{\theta}_i\right), \widehat{\theta}_i + 1.96 \times \widehat{\text{se}}\left(\widehat{\theta}_i\right)\right], \quad i = 1, 2, 3. \tag{15}$$

Let $\vartheta = h(\boldsymbol{\theta}_{-4})$ be an arbitrary differentiable function of $\boldsymbol{\theta}_{-4}$. For example, $\theta_4 = 1 - \sum_{i=1}^{3} \theta_{\text{i}}$ and the odds ratio $\psi = \theta_1\theta_3/(\theta_2\theta_4)$. The delta method (e.g., [14, page 34]) can be used to approximate the standard error of $\widehat{\vartheta} = h(\widehat{\boldsymbol{\theta}}_{-4})$, and a 95% normal-based asymptotic confidence interval for ϑ is given by

$$\left[\widehat{\vartheta} - 1.96 \times \widehat{\text{se}}\left(\widehat{\vartheta}\right), \widehat{\vartheta} + 1.96 \times \widehat{\text{se}}\left(\widehat{\vartheta}\right)\right], \tag{16}$$

where

$$\widehat{\text{se}}\left(\widehat{\vartheta}\right) = \left\{ \left(\frac{\partial\vartheta}{\partial\boldsymbol{\theta}_{-4}}\right)^{\top} \mathbf{I}_{\text{obs}}^{-1}(\boldsymbol{\theta}_{-4}) \left(\frac{\partial\vartheta}{\partial\boldsymbol{\theta}_{-4}}\right)\Bigg|_{\boldsymbol{\theta}_{-4} = \widehat{\boldsymbol{\theta}}_{-4}} \right\}^{1/2}. \tag{17}$$

3.3. Bootstrap Confidence Intervals. When the normal-based asymptotic confidence interval like (15) is beyond the low bound zero or the upper bound one, the bootstrap approach [15] can be used to construct the bootstrap confidence interval of $\vartheta = h(\boldsymbol{\theta}_{-4})$. Based on the obtained MLE $\widehat{\boldsymbol{\theta}}$, we independently generate

$$\begin{aligned} &(n_1^*, \ldots, n_4^*)^{\top} \\ &\quad \sim \text{Multinomial}\left(n; p_1\widehat{\theta}_1, p_2\widehat{\theta}_1 + \widehat{\theta}_2, p_3\widehat{\theta}_1 + \widehat{\theta}_3, \widehat{\theta}_4\right), \end{aligned} \tag{18}$$

$$m_0^* \sim \text{Binomial}\left(m, \widehat{\theta}_1 + \widehat{\theta}_2\right). \tag{19}$$

Having obtained $Y_{\text{obs},M}^* = \{n; n_1^*, \ldots, n_4^*\}$ and $Y_{\text{obs},S}^* = \{m; m_0^*, m_1^*\}$, where $m_1^* = m - m_0^*$, we can calculate the bootstrap replication $\widehat{\vartheta}^* = h(\boldsymbol{\theta}_{-4}^*)$ based on $Y_{\text{obs}}^* = \{Y_{\text{obs},M}^*, Y_{\text{obs},S}^*\}$ via the *EM* algorithm specified by (6) and (8). Independently repeating this process G times, we obtain G bootstrap replications $\{\widehat{\vartheta}_g^*\}_{g=1}^{G}$. Consequently, a $(1 - \alpha)100\%$ bootstrap confidence interval for ϑ is given by

$$\left[\widehat{\vartheta}_L, \widehat{\vartheta}_U\right], \tag{20}$$

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

3.4. The Likelihood Ratio Test for Testing Association. The likelihood ratio statistic can be used to test whether the two binary random variables X and Y are independent/correlated. The corresponding null and alternative hypotheses are [16, page 45]

$$H_0\colon \psi = 1 \text{ against } H_1\colon \psi \neq 1. \tag{21}$$

The likelihood ratio statistic is defined by

$$\Lambda = -2\left\{\ell_{\text{CQ}}\left(\widehat{\boldsymbol{\theta}}_R \mid Y_{\text{obs}}\right) - \ell_{\text{CQ}}\left(\widehat{\boldsymbol{\theta}} \mid Y_{\text{obs}}\right)\right\}, \tag{22}$$

where $\widehat{\boldsymbol{\theta}}_R$ denotes the restricted MLE of $\boldsymbol{\theta}$ under H_0, $\widehat{\boldsymbol{\theta}}$ denotes the MLE of $\boldsymbol{\theta}$, which can be obtained by the *EM* algorithm specified by (6) and (8), and $\ell_{\text{CQ}}(\boldsymbol{\theta} \mid Y_{\text{obs}}) = \log L_{\text{CQ}}(\boldsymbol{\theta} \mid Y_{\text{obs}})$.

To find the restricted MLE $\widehat{\boldsymbol{\theta}}_R$, we also employ the *EM* algorithm. Under H_0: $\theta_1\theta_3 = \theta_2\theta_4$, we have

$$\begin{aligned} \theta_1 &= (1 - \theta_x)\left(1 - \theta_y\right), \\ \theta_2 &= \theta_x\left(1 - \theta_y\right), \\ \theta_3 &= \theta_x\theta_y, \\ \theta_4 &= (1 - \theta_x)\,\theta_y. \end{aligned} \tag{23}$$

In other words, under H_0 we only have two free parameters θ_x and θ_y. Having obtained the restricted MLEs $\widehat{\theta}_{x,R}$ and $\widehat{\theta}_{y,R}$, we can compute the restricted MLE $\widehat{\boldsymbol{\theta}}_R = (\widehat{\theta}_{1,R}, \ldots, \widehat{\theta}_{4,R})^{\top}$ from (23) by

$$\begin{aligned} \widehat{\theta}_{1,R} &= \left(1 - \widehat{\theta}_{x,R}\right)\left(1 - \widehat{\theta}_{y,R}\right), \\ \widehat{\theta}_{2,R} &= \widehat{\theta}_{x,R}\left(1 - \widehat{\theta}_{y,R}\right), \\ \widehat{\theta}_{3,R} &= \widehat{\theta}_{x,R}\widehat{\theta}_{y,R}, \\ \widehat{\theta}_{4,R} &= \left(1 - \widehat{\theta}_{x,R}\right)\widehat{\theta}_{y,R}. \end{aligned} \tag{24}$$

In what follows, we consider the computation of the restricted MLEs $\widehat{\theta}_{x,R}$ and $\widehat{\theta}_{y,R}$. Now, the complete-data likelihood function (5) becomes

$$\begin{aligned} &L_{\text{CQ}}\left(\theta_x, \theta_y \mid Y_{\text{com}}, H_0\right) \\ &\quad \propto \left\{(1 - \theta_x)\left(1 - \theta_y\right)\right\}^{z_1}\left\{\theta_x\left(1 - \theta_y\right)\right\}^{z_2} \\ &\qquad \times \left(\theta_x\theta_y\right)^{z_3}\left\{(1 - \theta_x)\,\theta_y\right\}^{n_4}\left(1 - \theta_y\right)^{m_0}\theta_y^{m_1} \\ &\quad = \theta_x^{z_2+z_3}(1 - \theta_x)^{z_1+n_4}\theta_y^{z_3+n_4+m_1}\left(1 - \theta_y\right)^{z_1+z_2+m_0} \end{aligned} \tag{25}$$

so that the restricted MLEs of θ_x and θ_y based on the complete-data are given by

$$\theta_x = \frac{z_2 + z_3}{n}, \qquad \theta_y = \frac{z_3 + n_4 + m_1}{N}, \tag{26}$$

respectively. Thus, the M-step of the *EM* algorithm calculates (26), and the E-step computes the conditional expectations given in (8), where $\{\theta_i\}$ are defined in (23). Finally, under H_0, Λ asymptotically follows chi-squared distribution with one degree of freedom.

4. Bayesian Inferences

To derive the posterior mode of $\boldsymbol{\theta}$, we employ the *EM* algorithm again. The latent data $Y_{\text{mis}} = \{z_2, z_3\}$ are the same as those in Section 3.1. Based on the complete-data likelihood function (5), if the Dirichlet distribution Dirichlet $(a_1, \ldots, a_4)$ is adopted as the prior distribution of $\boldsymbol{\theta}$, then the complete-data posterior distribution is a *grouped Dirichlet* (GD) distribution with the formal definition given by (A.2); that is,

$$f(\boldsymbol{\theta} \mid Y_{\text{obs}}, Y_{\text{mis}}) = \text{GD}_{4,2,2}(\boldsymbol{\theta}_{-4} \mid \mathbf{a}, \mathbf{b}), \qquad (27)$$

where $\mathbf{a} = (a_1 + z_1, a_2 + z_2, a_3 + z_3, a_4 + n_4)^\top$, $\mathbf{b} = (m_0, m_1)^\top$, and $z_1 = n - (z_2 + z_3) - n_4$. The conditional predictive distribution is

$$f(Y_{\text{mis}} \mid Y_{\text{obs}}, \boldsymbol{\theta}) = \prod_{i=2}^{3} \text{Binomial}\left(z_i \,\middle|\, n_i, \frac{\theta_i}{p_i\theta_1 + \theta_i}\right). \qquad (28)$$

Therefore, the M-step of the *EM* algorithm is to calculate the complete-data posterior mode:

$$\begin{aligned}
\theta_1 &= 1 - \theta_2 - \theta_3 - \theta_4, \\
\theta_2 &= \frac{a_2 - 1 + z_2}{a_+ - 4 + N}\left(1 + \frac{m_0}{a_1 + a_2 - 2 + n - z_3 - n_4}\right), \\
\theta_3 &= \frac{a_3 - 1 + z_3}{a_+ - 4 + N}\left(1 + \frac{m_1}{a_3 + a_4 - 2 + z_3 + n_4}\right), \\
\theta_4 &= \frac{a_4 - 1 + n_4}{a_+ - 4 + N}\left(1 + \frac{m_1}{a_3 + a_4 - 2 + z_3 + n_4}\right),
\end{aligned} \qquad (29)$$

where $a_+ = \sum_{i=1}^{4} a_i$, and the E-step is to replace $\{z_i\}_{i=2}^{3}$ by the conditional expectations given by (8).

In addition, based on (27) and (28), the data augmentation algorithm of Tanner and Wong [13] can be used to generate posterior samples of $\boldsymbol{\theta}$. A sampling method from (27) is given in the appendix.

5. Simulation Studies

In this section, two simulation studies are conducted to compare the efficiency of the proposed combination questionnaire model with that of the hidden sensitivity model of Tian et al. [8] (i.e., the main questionnaire model only), where (p_1, p_2, p_3) are assumed to be $(1/3, 1/3, 1/3)$ and $p_i \widehat{=} \Pr(W = i)$ for $i = 1, 2, 3$. In the first simulated example, let the total sample size in the combination questionnaire model be the same as the sample size in the hidden sensitivity model. In the second example, the sample size for the main questionnaire in the combination questionnaire model is assumed to be equal to the sample size in the hidden sensitivity model.

In the first simulated example, let a total of $N = n + m = 50 + 50 = 100$ participants be interviewed by using the combination questionnaire model. The true values of $\{\theta_i\}_{i=1}^4$ are listed in the second column of Table 5. We first generate

$$(n_1, \ldots, n_4)^\top \sim \text{Multinomial}\ (n; p_1\theta_1, p_2\theta_1 + \theta_2, p_3\theta_1 + \theta_3, \theta_4), \qquad (30)$$

and $m_0 \sim$ Binomial $(m, \theta_1 + \theta_2)$ so that $m_1 = m - m_0$. The *EM* algorithm (6) and (8) is used to calculate the MLEs of $\{\theta_i\}_{i=1}^4$. We repeated this experiment 1000 times. The average MLEs of $\{\theta_i\}_{i=1}^4$ are reported in the third column of Table 5. The corresponding bias, variance, and *mean square error* (MSE) are displayed in the fourth, fifth, and sixth columns of Table 5. Next, let $N = n + m = 100 + 0 = 100$ participants be interviewed by using the hidden sensitivity model (i.e., the main questionnaire only). The corresponding results are reported in the last four columns of Table 5.

From Table 5, we can see that both the MLEs of $\{\theta_i\}_{i=1}^4$ in the combination questionnaire model and the hidden sensitivity model are very close to their true values, while the MSEs of $\{\widehat{\theta}_i\}_{i=1}^4$ in the combination questionnaire model are slightly larger than those in the hidden sensitivity model. These numerical results are not surprising since in the hidden sensitivity model, Tian et al. [8] implicitly assumed that all respondents must strictly follow the design instructions. In other words, the noncompliance behavior will not occur. The introduction of the supplemental questionnaire in the combination questionnaire model can effectively reduce the non-compliance behavior since more respondents will not be faced with the sensitive question, while the cost for introducing such a supplemental questionnaire is that we definitely lose a little of efficiency.

In the second simulated example, we assume that a total of $N = n + m = 100 + 100 = 200$ participants are interviewed by using the combination questionnaire model, while only $N = n + m = 100 + 0 = 100$ are interviewed by using the hidden sensitivity model. We repeat this experiment 1000 times. The corresponding results are reported in Table 6.

From Table 6, we can see that the MSEs of $\{\widehat{\theta}_i\}_{i=1}^4$ in the combination questionnaire model are smaller than those in the hidden sensitivity model. In addition, by comparing the fifth columns in Tables 5 and 6, we can see that the precisions of $\{\widehat{\theta}_i\}_{i=1}^4$ for the proposed combination questionnaire model in the second simulated example are significantly improved when compared with those in the first simulated example.

6. Analyzing Cervical Cancer Data in Atlanta

Williamson and Haber [17] reported a study which examined the relationship between disease status of cervical cancer and the number of sex partners and other risk factors. Cases were 20–79-year-old women of Fulton or DeKalb county in Atlanta, Georgia. They were diagnosed and were ascertained to have invasive cervical cancer. Controls were randomly chosen from the same counties and the same age ranges. Table 7 gives the cross-classification of number of sex partners ("few, 0–3" or "many, ⩾4", denoted by $X = 0$ or $X = 1$) and disease status (control or case, denoted by $Y = 0$ or $Y = 1$). Generally, a sizable proportion (13.5% in this example) of the responses would be missing because of the sensitive question about the number of sex partners in a telephone interview. The objective is to examine if association exists between the number of sex partners and disease status of cervical cancer.

For the purpose of illustration, we presume that $\{X = 0\}$ is non-sensitive although the number 0–3 of sex partners

Table 5: Comparison of the efficiency of the proposed combination questionnaire model with that of the hidden sensitivity model of Tian et al. [8] (i.e., the main questionnaire model only) with $(p_1, p_2, p_3) = (1/3, 1/3, 1/3)$.

Parameter	True value	The combination questionnaire model ($N = n + m = 50 + 50$)				The hidden sensitivity model ($N = n + m = 100 + 0$)			
		MLE	Bias	Variance	MSE	MLE	Bias	Variance	MSE
θ_1	0.10	0.1013	0.0013	0.0113	0.0113	0.1010	0.0010	0.0064	0.0064
θ_2	0.50	0.4985	−0.0015	0.0089	0.0089	0.4975	−0.0025	0.0041	0.0041
θ_3	0.20	0.2008	0.0008	0.0037	0.0037	0.2001	0.0001	0.0029	0.0029
θ_4	0.20	0.1994	−0.0006	0.0026	0.0026	0.2014	0.0014	0.0016	0.0016
θ_1	0.20	0.1957	−0.0043	0.0090	0.0090	0.1985	−0.0015	0.0055	0.0055
θ_2	0.30	0.2999	−0.0001	0.0070	0.0070	0.2999	−0.0001	0.0034	0.0034
θ_3	0.40	0.4038	0.0038	0.0040	0.0040	0.4025	0.0025	0.0037	0.0037
θ_4	0.10	0.1006	0.0006	0.0017	0.0017	0.0991	−0.0009	0.0009	0.0009
θ_1	0.30	0.3021	0.0021	0.0107	0.0107	0.3026	0.0026	0.0081	0.0081
θ_2	0.25	0.2483	−0.0017	0.0077	0.0077	0.2499	−0.0001	0.0039	0.0039
θ_3	0.15	0.1474	−0.0026	0.0042	0.0042	0.1476	−0.0024	0.0032	0.0032
θ_4	0.30	0.3021	0.0021	0.0033	0.0033	0.2999	−0.0001	0.0021	0.0021

Note: $\text{Bias}(\widehat{\theta}_i) = E(\widehat{\theta}_i) - \theta_i$ and $\text{MSE}(\widehat{\theta}_i) = [\text{Bias}(\widehat{\theta}_i)]^2 + \text{Var}(\widehat{\theta}_i)$.

Table 6: Comparison of the efficiency of the proposed combination questionnaire model with that of the hidden sensitivity model of Tian et al. [8] (i.e., the main questionnaire model only) with $(p_1, p_2, p_3) = (1/3, 1/3, 1/3)$.

Parameter	True value	The combination questionnaire model ($N = n + m = 100 + 100$)				The hidden sensitivity model ($N = n + m = 100 + 0$)			
		MLE	Bias	Variance	MSE	MLE	Bias	Variance	MSE
θ_1	0.10	0.1003	0.0003	0.0030	0.0030	0.1013	0.0013	0.0064	0.0064
θ_2	0.50	0.5015	0.0015	0.0029	0.0029	0.5019	0.0019	0.0041	0.0041
θ_3	0.20	0.1986	−0.0014	0.0017	0.0017	0.1969	−0.0031	0.0028	0.0028
θ_4	0.20	0.1995	−0.0005	0.0013	0.0013	0.1999	−0.0001	0.0016	0.0016
θ_1	0.20	0.2007	0.0007	0.0041	0.0041	0.2049	0.0049	0.0057	0.0057
θ_2	0.30	0.2989	−0.0011	0.0033	0.0033	0.2965	−0.0035	0.0034	0.0034
θ_3	0.40	0.4007	0.0007	0.0020	0.0020	0.3992	−0.0008	0.0037	0.0037
θ_4	0.10	0.0997	−0.0003	0.0009	0.0009	0.0994	−0.0006	0.0009	0.0009
θ_1	0.30	0.3032	0.0032	0.0054	0.0054	0.2965	−0.0036	0.0079	0.0079
θ_2	0.25	0.2488	−0.0012	0.0039	0.0039	0.2501	0.0001	0.0038	0.0038
θ_3	0.15	0.1481	−0.0019	0.0021	0.0021	0.1531	0.0031	0.0032	0.0032
θ_4	0.30	0.3000	0.0000	0.0017	0.0017	0.3003	0.0003	0.0021	0.0021

Note: $\text{Bias}(\widehat{\theta}_i) = E(\widehat{\theta}_i) - \theta_i$ and $\text{MSE}(\widehat{\theta}_i) = [\text{Bias}(\widehat{\theta}_i)]^2 + \text{Var}(\widehat{\theta}_i)$.

Table 7: Cervical cancer data from Williamson and Haber [17].

Number of sex partners	Disease status of cervical cancer	
	$Y = 0$ (control)	$Y = 1$ (case)
$X = 0$ (few, 0–3)	165 $(r_1,\ \theta_1)$	103 $(r_4,\ \theta_4)$
$X = 1$ (many, ⩾4)	164 $(r_2,\ \theta_2)$	221 $(r_3,\ \theta_3)$
Missing	43 $(r_{12},\ \theta_1 + \theta_2)$	59 $(r_{34},\ \theta_3 + \theta_4)$

Note: The observed counts and the corresponding cell probabilities are in parentheses. X is a sensitive binary variate and Y is a non-sensitive binary variate.

is somewhat sensitive for some respondents. To illustrate the proposed design and approaches, we let $W = 1\ (2, 3)$ if a respondent was born in January–April (May–August, September–December). It is then reasonable to assume that $p_k = \Pr(W = k) \approx 1/3$ for $k = 1, 2, 3$, and W is independent of the sensitive binary variate X and the the non-sensitive binary variate Y. For the ideal situation (i.e., no sampling errors), the observed counts from the main questionnaire as shown in Tables 1 and 3 would be $n_1 = r_1/3 = 55$, $n_2 = 55 + r_2 = 219$, $n_3 = 55 + r_3 = 276$, $n_4 = r_4 = 103$; that is,

$$Y_{\text{obs},M} = \{n; n_1, \ldots, n_4\} = \{653; 55, 219, 276, 103\}. \tag{31}$$

On the other hand, we can view the missing data in Table 7 as the observed counts from the supplemental questionnaire as shown in Table 2. From Table 4, we have $m_0 = r_{12} = 43$ and $m_1 = r_{34} = 59$; that is, $Y_{\text{obs},S} = \{m; m_0, m_1\} = \{102; 43, 59\}$. Therefore, we obtain the observed data $Y_{\text{obs}} = \{Y_{\text{obs},M}, Y_{\text{obs},S}\}$.

Table 8: MLEs and 95% confidence intervals of $\boldsymbol{\theta}$ and ψ.

Parameter	MLE	std	95% asymptotic CI	95% bootstrap CI
θ_1	0.23793	0.02937	[0.18036, 0.29551]	[0.18006, 0.29879]
θ_2	0.24916	0.02261	[0.20484, 0.29348]	[0.20375, 0.29359]
θ_3	0.35196	0.02247	[0.30790, 0.39601]	[0.30731, 0.39666]
θ_4	0.16095	0.01434	[0.13284, 0.18905]	[0.13400, 0.18908]
ψ	2.08830	0.43971	[1.22646, 2.95011]	[1.35361, 3.18593]

Table 9: Posterior modes and estimates of parameters for the cervical cancer data.

Parameter	Posterior mode	Bayesian mean	Bayesian std	95% Bayesian credible interval
θ_1	0.23793	0.24025	0.02934	[0.18574, 0.30050]
θ_2	0.24916	0.24789	0.02251	[0.20365, 0.29192]
θ_3	0.35196	0.35029	0.02246	[0.30621, 0.39432]
θ_4	0.16095	0.16155	0.01431	[0.13448, 0.19042]
ψ	2.08830	2.14538	0.45838	[1.39399, 3.18678]

6.1. Likelihood-Based Inferences. Using $\boldsymbol{\theta}^{(0)} = \mathbf{1}_4/4$ as the initial values, the *EM* algorithm in (6) and (8) converged in 29 iterations. The resultant MLEs of $\boldsymbol{\theta}$ and ψ are given in the second column of Table 8. From (13) and (14), the asymptotic variance-covariance matrix of the MLEs $\widehat{\boldsymbol{\theta}}_{-4}$ is

$$\mathbf{I}_{\text{obs}}^{-1}\left(\widehat{\boldsymbol{\theta}}_{-4}\right) = \begin{pmatrix} 0.00086302 & -0.000442447 & -0.000384452 \\ -0.00044245 & 0.000511253 & -0.000010026 \\ -0.00038445 & -0.000010026 & 0.000505241 \end{pmatrix}. \tag{32}$$

The estimated standard errors of $\widehat{\theta}_i$ $(i = 1, 2, 3)$ are square roots of the main diagonal elements of the above matrix. From (17), the estimated standard errors of $\widehat{\theta}_4 = 1 - \widehat{\theta}_1 - \widehat{\theta}_2 - \widehat{\theta}_3$ and $\widehat{\psi} = \widehat{\theta}_1\widehat{\theta}_3/(\widehat{\theta}_2\widehat{\theta}_4)$ are given by

$$\begin{aligned} \widehat{\text{se}}\left(\widehat{\theta}_4\right) &= \left\{\mathbf{1}_3^\top\, \mathbf{I}_{\text{obs}}^{-1}\left(\widehat{\boldsymbol{\theta}}_{-4}\right) \mathbf{1}_3\right\}^{1/2}, \\ \widehat{\text{se}}\left(\widehat{\psi}\right) &= \left\{\boldsymbol{\alpha}^\top \mathbf{I}_{\text{obs}}^{-1}\left(\widehat{\boldsymbol{\theta}}_{-4}\right) \boldsymbol{\alpha}\right\}^{1/2}, \end{aligned} \tag{33}$$

respectively, where

$$\boldsymbol{\alpha} = \left(\frac{\widehat{\theta}_3\left(1 - \widehat{\theta}_2 - \widehat{\theta}_3\right)}{\widehat{\theta}_2\widehat{\theta}_4^2}, \frac{\widehat{\theta}_1\widehat{\theta}_3\left(\widehat{\theta}_2 - \widehat{\theta}_4\right)}{\left(\widehat{\theta}_2\widehat{\theta}_4\right)^2}, \frac{\widehat{\theta}_1\left(1 - \widehat{\theta}_1 - \widehat{\theta}_2\right)}{\widehat{\theta}_2\widehat{\theta}_4^2}\right)^\top. \tag{34}$$

These estimated standard errors are listed in the third column of Table 8. From (15) and (16), we can obtain the 95% asymptotic confidence intervals of $\boldsymbol{\theta}$ and ψ, which are showed in the fourth column of Table 8.

Based on (18) and (19), we generate $G = 10{,}000$ bootstrap samples. The corresponding 95% bootstrap confidence intervals of $\boldsymbol{\theta}$ and ψ are displayed in the last column of Table 8.

To test the null hypothesis H_0: $\psi = 1$ against H_1: $\psi \neq 1$, we need to obtain the restricted MLE $\widehat{\boldsymbol{\theta}}_R$. Using

$$\left(\theta_x^{(0)}, \theta_y^{(0)}\right)^\top = (0.5, 0.5)^\top, \tag{35}$$

as the initial values, the *EM* algorithm in (26) converged to

$$\widehat{\theta}_{x,R} = 0.64807, \qquad \widehat{\theta}_{y,R} = 0.52948, \tag{36}$$

in 19 iterations. From (24), the restricted MLEs of $\boldsymbol{\theta}$ are obtained as

$$\widehat{\boldsymbol{\theta}}_R = (0.16559, 0.30493, 0.34314, 0.18634)^\top. \tag{37}$$

The log-likelihood ratio statistic Λ is equal to 12.469 and the P-value is 0.0004137. Since this P-value is far less than 0.05, the H_0 is rejected at the 0.05 level of significance. Thus, we can conclude that there is an association between sex partners and cervical cancer status based on the current data. This conclusion is identical to that from the two 95% confidence intervals of the odds ratio as shown in Table 8, where both confidence intervals exclude the value 1.

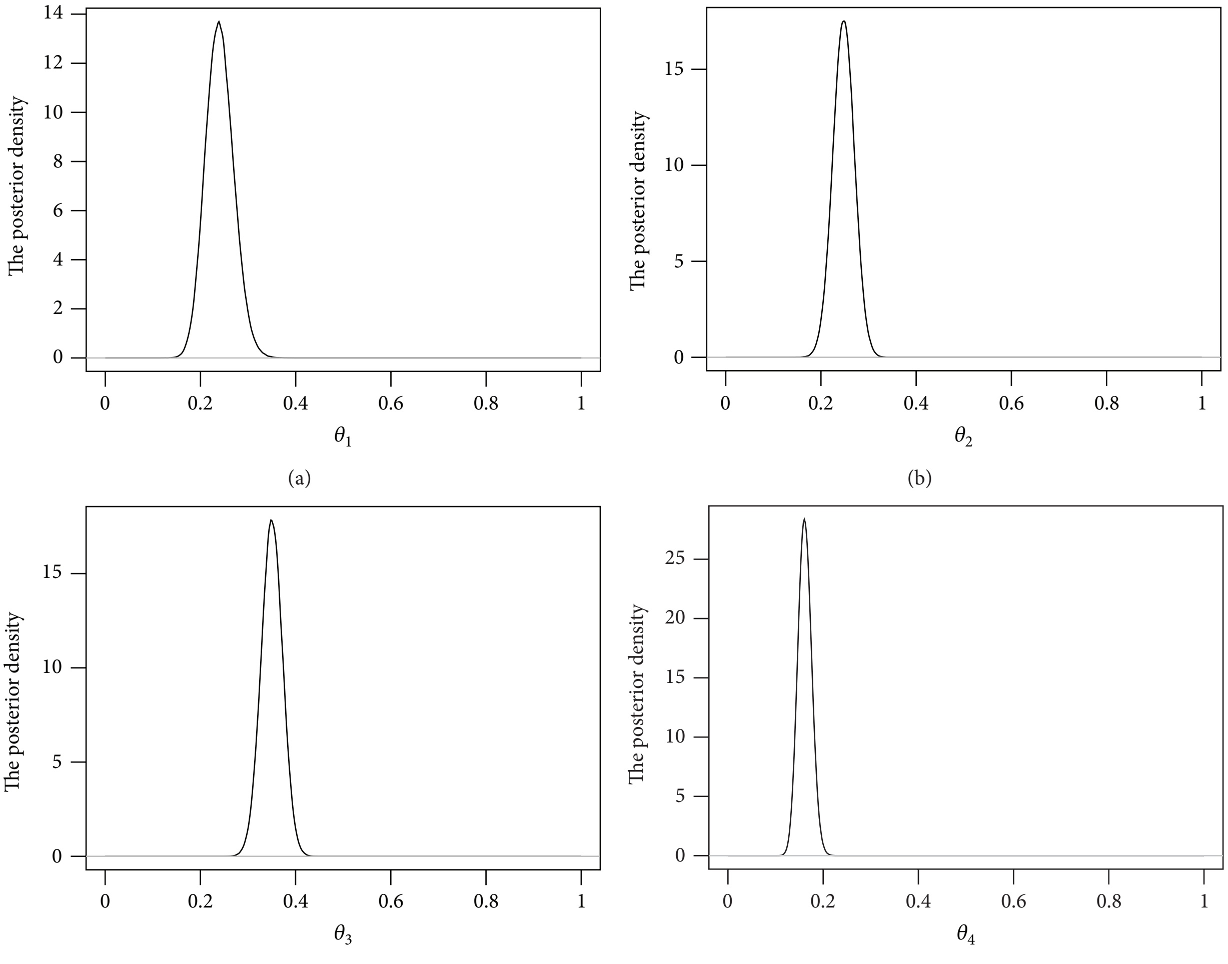
6.2. Bayesian Inferences. When Dirichlet $(1, 1, 1, 1)$ (i.e., the uniform distribution on $\mathbb{T}_4$) is adopted as the prior distribution of $\boldsymbol{\theta}$, the posterior modes of $\boldsymbol{\theta}$are equal to the corresponding MLEs. Using $\boldsymbol{\theta}^{(0)} = \mathbf{1}_4/4$ as the initial values, we employ the data augmentation algorithm to generate 1,000,000 posterior samples of $\boldsymbol{\theta}$ and discard the first half of the samples. The Bayesian estimates of $\boldsymbol{\theta}$ and ψ are given in Table 9. Since the lower bound of the Bayesian credible interval of the ψ is larger than 1, we believe that there is an association between the number of sexual partners and cervical cancer status.

The posterior densities of the $\{\theta_i\}_{i=1}^4$ and ψ estimated by a kernel density smoother are plotted in Figures 1 and 2.

7. Discussion

In this paper, we develop a general framework of design and analysis for the combination questionnaire model, which consists of a main questionnaire and a supplemental questionnaire. In fact, the main questionnaire (see Table 1) is a generalization of the nonrandomized triangular model [18] and is then a special case of the multicategory triangular model [19]. The supplemental questionnaire (see Table 2) is a design of direct questioning. The introduction of the supplemental questionnaire can effectively reduce the noncompliance behavior since more respondents will not be faced with the sensitive question, while the cost for introducing such a supplemental questionnaire is that we definitely lose a little of efficiency. The combination questionnaire model can be used to gather information to evaluate the association between one sensitive binary variable and one non-sensitive binary variable.

We note that the proposed combination questionnaire model has one limitation in applications; that is, it cannot be



(a)

(b)

(c)

(d)

Figure 1: The posterior densities of the $\{\theta_i\}_{i=1}^{4}$ estimated by a kernel density smoother based on the second 500,000 posterior samples of $\boldsymbol{\theta}$generated by the data augmentation algorithm when the prior distribution is Dirichlet $(1, 1, 1, 1)$. (a) The posterior density of θ_1; (b) The posterior density of θ_2; (c) The posterior density of θ_3; (d) The posterior density of θ_4.

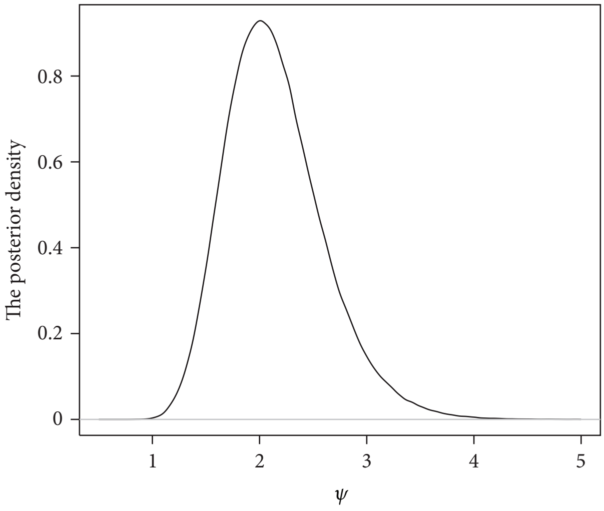


Figure 2: The posterior density of the odds ratio ψ estimated by a kernel density smoother based on the second 500,000 posterior samples of $\boldsymbol{\theta}$ generated by the data augmentation algorithm when the prior distribution is Dirichlet $(1, 1, 1, 1)$.

applied to situation where two categories $\{X = 0\}$ and $\{X = 1\}$ are sensitive like income. For example, let $X = 0$ if his/her annual income is \$25,000 or less and $X = 1$ if his/her annual income is more than \$25,000. For such cases, it is worthwhile to develop new designs to address this issue. One way is to replace the main questionnaire in Table 1 by a four-category parallel model (see [20, Section 4.1]). The other way is to employ the parallel model [21] to collect the information on X and to employ the design of direct questioning to collect information on Y then we could use the logistic regression to estimate the odds ratio, which is one of our further researches.

Appendix

The Mode of a Group Dirichlet Density and a Sampling Method from a GDD

Let

$$\mathbb{V}_{n-1} = \Bigg\{(x_1, \ldots, x_{n-1})^\top : x_i \geqslant 0, \quad i = 1, \ldots, n-1, \sum_{i=1}^{n-1} x_i \leqslant 1\Bigg\} \tag{A.1}$$

denote the $(n - 1)$-dimensional open simplex in $\mathbb{R}^{n-1}$. A random vector $\mathbf{x} = (\mathrm{X}_1, \ldots, \mathrm{X}_n)^\top \in \mathbb{T}_n$ is said to follow a *group Dirichlet distribution* (GDD) with two partitions, if the density of $\mathbf{x}_{-n} = (X_1, \ldots, X_{n-1})^\top \in \mathbb{V}_{n-1}$ is

$$\mathrm{GD}_{n,2,s}(x_{-n} \mid \mathbf{a}, \mathbf{b}) = c_{\mathrm{GD}}^{-1}\left(\prod_{i=1}^{n} x_i^{a_i - 1}\right)\left(\sum_{i=1}^{s} x_\mathrm{i}\right)^{b_1}\left(\sum_{i=s+1}^{n} x_\mathrm{i}\right)^{b_2}, \tag{A.2}$$

where $x_{-n} = (x_1, \ldots, x_{n-1})^\top$, $\mathbf{a} = (a_1, \ldots, a_n)^\top$ is a positive parameter vector, $\mathbf{b} = (b_1, b_2)^\top$ is a nonnegative parameter vector, s is a known positive integer less than n, and the normalizing constant is given by

$$c_{\mathrm{GD}} = B(a_1, \ldots, a_s)\, B(a_{s+1}, \ldots, a_n) \times B\left(\sum_{i=1}^{s} a_i + b_1, \sum_{i=s+1}^{n} a_i + b_2\right). \tag{A.3}$$

We write $\mathbf{x} \sim \mathrm{GD}_{n,2,s}(\mathbf{a}, \mathbf{b})$ on $\mathbb{T}_n$ or $\mathbf{x}_{-n} \sim \mathrm{GD}_{n,2,s}(\mathbf{a}, \mathbf{b})$ on $\mathbb{V}_{n-1}$ to distinguish the two equivalent representations.

If $a_i \geqslant 1$, then the mode of the grouped Dirichlet density (A.2) is given by [11, 22]

$$\hat{x}_i = \frac{a_i - 1}{\sum_{j=1}^{n}(a_j - 1) + b_1 + b_2}\left\{1 + \frac{b_1}{\sum_{j=1}^{s}(a_j - 1)}\right\}, \quad i = 1, \ldots, s, \tag{A.4}$$

$$\hat{x}_i = \frac{a_i - 1}{\sum_{j=1}^{n}(a_j - 1) + b_1 + b_2}\left\{1 + \frac{b_2}{\sum_{j=s+1}^{n}(a_j - 1)}\right\}, \quad i = s + 1, \ldots, n. \tag{A.5}$$

The following procedure can be used to generate i.i.d. samples from a GDD [11]. Let

(1) $\mathbf{y}^{(1)} \sim$ Dirichlet $(a_1, \ldots, a_s)$ on $\mathbb{T}_s$;

(2) $\mathbf{y}^{(2)} \sim$ Dirichlet $(a_{s+1}, \ldots, a_n)$ on $\mathbb{T}_{n-s}$;

(3) $R \sim \mathrm{Beta}(\sum_{i=1}^{s} a_i + b_1, \sum_{i=s+1}^{n} a_i + b_2)$; and

(4) $\mathbf{y}^{(1)}, \mathbf{y}^{(2)}$ and R are mutually independent.

Define

$$\mathbf{x}^{(1)} = R \times \mathbf{y}^{(1)}, \qquad \mathbf{x}^{(2)} = (1 - R) \times \mathbf{y}^{(2)}. \tag{A.6}$$

Then,

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}^{(1)} \\ \mathbf{x}^{(2)} \end{pmatrix} \sim \mathrm{GD}_{n,2,s}(\mathbf{a}, \mathbf{b}) \tag{A.7}$$

on $\mathbb{T}_n$, where $\mathbf{a} = (a_1, \ldots, a_n)^\top$ and $\mathbf{b} = (b_1, b_2)^\top$.

Acknowledgments

The authors are grateful to the Editor, an Associate Editor, and two anonymous referees for their helpful comments and suggestions which led to the improvement of the paper. They also would like to thank Miss Yin LIU of The University of Hong Kong for her help in the simulation studies. Jun-Wu Yu's research was partially supported by a Grant (no. 09BTJ012) from the National Social Science Foundation of China and a Grant from Hunan Provincial Science and Technology Department (no. 11JB1176). Partial work was done when Jun-Wu Yu visited The University of Hong Kong. Guo-Liang Tian's research was partially supported by a Grant (HKU 779210M) from the Research Grant Council of the Hong Kong Special Administrative Region.

References

[1] S. L. Warner, "Randomized response: a survey technique for eliminating evasive answer bias," *Journal of the American Statistical Association*, vol. 60, no. 309, pp. 63–69, 1965.

[2] H. C. Kraemer, "Estimation and testing of bivariate association using data generated by the randomized response technique," *Psychological Bulletin*, vol. 87, no. 2, pp. 304–308, 1980.

[3] J. A. Fox and P. E. Tracy, "Measuring associations with randomized response," *Social Science Research*, vol. 13, no. 2, pp. 188–197, 1984.

[4] S. E. Edgell, S. Himmelfarb, and D. J. Cira, "Statistical efficiency of using two quantitative randomized response techniques to estimate correlation," *Psychological Bulletin*, vol. 100, no. 2, pp. 251–256, 1986.

[5] T. C. Christofides, "Randomized response technique for two sensitive characteristics at the same time," *Metrika*, vol. 62, no. 1, pp. 53–63, 2005.

[6] J. M. Kim and W. D. Warde, "Some new results on the multinomial randomized response model," *Communications in Statistics: Theory and Methods*, vol. 34, no. 4, pp. 847–856, 2005.

[7] G. J. L. M. Lensvelt-Mulders, J. J. Hox, P. G. M. van der Heijden, and C. J. M. Maas, "Meta-analysis of randomized response research: thirty-five years of validation," *Sociological Methods & Research*, vol. 33, no. 3, pp. 319–348, 2005.

[8] G. L. Tian, J. W. Yu, M. L. Tang, and Z. Geng, "A new non-randomized model for analysing sensitive questions with binary outcomes," *Statistics in Medicine*, vol. 26, no. 23, pp. 4238–4252, 2007.

[9] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society B*, vol. 39, no. 1, pp. 1–38, 1977.

[10] G.-L. Tian, K. W. Ng, and Z. Geng, "Bayesian computation for contingency tables with incomplete cell-counts," *Statistica Sinica*, vol. 13, no. 1, pp. 189–206, 2003.

[11] K. W. Ng, M. L. Tang, M. Tan, and G. L. Tian, "Grouped Dirichlet distribution: a new tool for incomplete categorical data analysis," *Journal of Multivariate Analysis*, vol. 99, no. 3, pp. 490–509, 2008.

[12] D. R. Cox and D. Oakes, *Analysis of Survival Data*, Monographs on Statistics and Applied Probability, Chapman & Hall, London, UK, 1984.

[13] M. A. Tanner and W. H. Wong, "The calculation of posterior distributions by data augmentation," *Journal of the American Statistical Association*, vol. 82, no. 398, pp. 528–550, 1987.

[14] M. A. Tanner, *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions* , Springer Series in Statistics, Springer, New York, NY, USA, 3rd edition, 1996.

[15] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*, vol. 57 of *Monographs on Statistics and Applied Probability*, Chapman & Hall, New York, NY, USA, 1993.

[16] A. Agresti, *Categorical Data Analysis*, Wiley Series in Probability and Statistics, Wiley-Interscience, New York, NY, USA, 2nd edition, 2002.

[17] G. D. Williamson and M. Haber, "Models for three-dimensional contingency tables with completely and partially cross-classified data," *Biometrics*, vol. 50, no. 1, pp. 194–203, 1994.

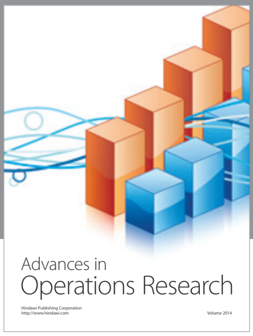
[18] J. W. Yu, G. L. Tian, and M. L. Tang, "Two new models for survey sampling with sensitive characteristic: design and analysis," *Metrika*, vol. 67, no. 3, pp. 251–263, 2008.

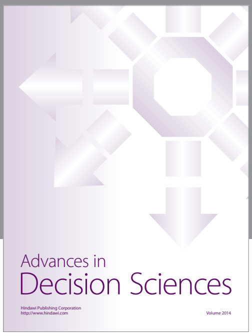
[19] M. L. Tang, G. L. Tian, N. S. Tang, and Z. Q. Liu, "A new non-randomized multi-category response model for surveys with a single sensitive question: design and analysis," *Journal of the Korean Statistical Society*, vol. 38, no. 4, pp. 339–349, 2009.

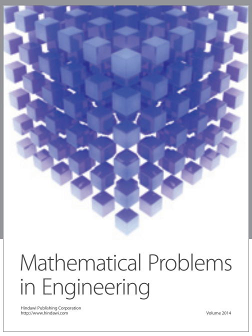
[20] Y. Liu and G. L. Tian, "Multi-category parallel models in the design of surveys with sensitive questions," *Statistics and Its Interface*, vol. 6, no. 1, pp. 137–149, 2013.

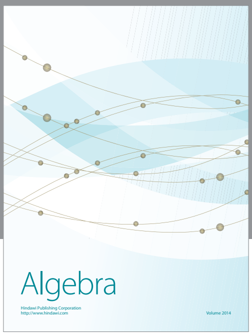
[21] G. L. Tian, "A new non-randomized response model: the parallel model," *Statistica Neerlandica*. In press.

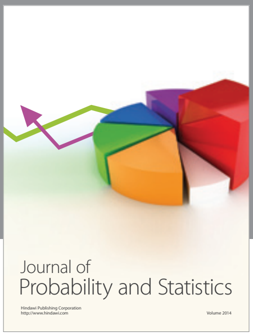
[22] K. W. Ng, G. L. Tian, and M. L. Tang, *Dirichlet and Related Distributions: Theory, Methods and Applications*, Wiley Series in Probability and Statistics, John Wiley & Sons, Chichester, UK, 2011.


Advances in
Operations Research

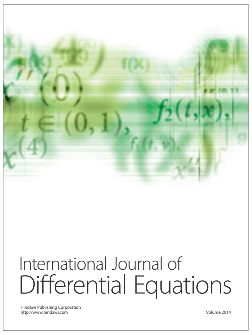

Advances in
Decision Sciences

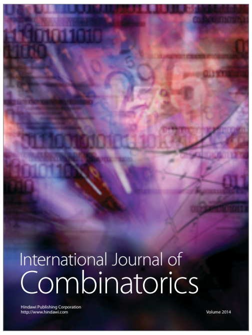

Mathematical Problems
in Engineering

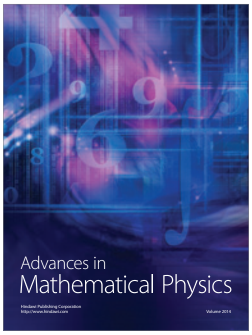

Algebra

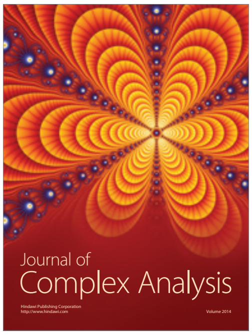

Journal of
Probability and Statistics

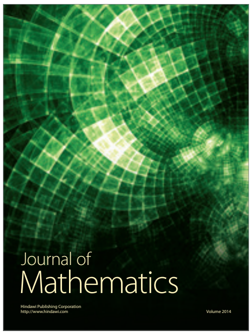
The Scientific
World Journal

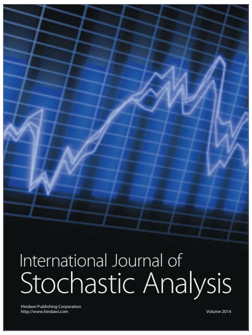

International Journal of
Differential Equations

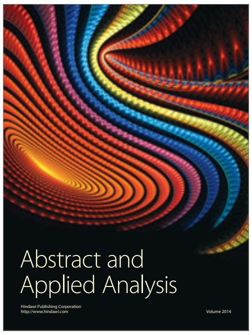

International Journal of
Combinatorics

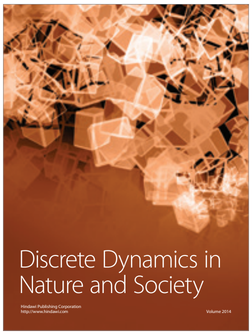

Advances in
Mathematical Physics

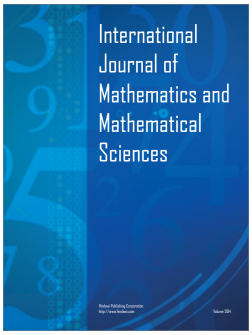

Journal of
Complex Analysis


Journal of
Mathematics


International Journal of
Stochastic Analysis

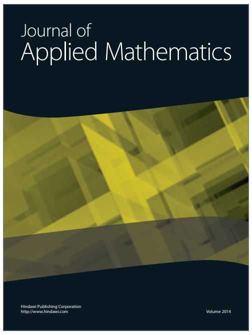

Abstract and
Applied Analysis


Discrete Dynamics in
Nature and Society


International
Journal of
Mathematics and
Mathematical
Sciences

Journal of
Discrete Mathematics

Journal of
Function Spaces


Journal of
Applied Mathematics

Journal of
Optimization