mathematics

MDPI

Article

Hierarchical Bayesian Modeling and Randomized Response Method for Inferring the Sensitive-Nature Proportion

Hua Xin [1,†], Jianping Zhu [2,†], Tzong-Ru Tsai [3,*,†] and Chieh-Yi Hung [3,†]

[1] School of Mathematics and Statistics, Northeast Petroleum University, Daqing 163318, China; xinhua@nepu.edu.cn
[2] School of Management and Data-Mining Research Center, Xiamen University, Xiamen 361005, China; xmjpzhu@xmu.edu.cn
[3] Department of Statistics, Tamkang University, New Taipei City 251301, Taiwan; 405651166@gms.tku.edu.tw
* Correspondence: tzongru@gms.tku.edu.tw; Tel.: +886-2-2621-5656 (ext. 2632)
† These authors contributed equally to this work.

Abstract: In this study, a new three-statement randomized response estimation method is proposed to improve the drawback that the maximum likelihood estimation method could generate a negative value to estimate the sensitive-nature proportion (SNP) when its true value is small. The Bayes estimator of the SNP is obtained via using a hierarchical Bayesian modeling procedure. Moreover, a hybrid algorithm using Gibbs sampling in Metropolis–Hastings algorithms is used to obtain the Bayes estimator of the SNP. The highest posterior density interval of the SNP is obtained based on the empirical distribution of Markov chains. We use the term 3RR-HB to denote the proposed method here. Monte Carlo simulations show that the quality of 3RR-HB procedure is good and that it can improve the drawback of the maximum likelihood estimation method. The proposed 3RR-HB procedure is simple for use. An example regarding the homosexual proportion of college freshmen is used for illustration.

Keywords: Bayesian estimation; Beta-Binominal model; maximum likelihood estimation; respondent protection; randomized response

Citation: Xin, H.; Zhu, J.; Tsai, T.-R.; Hung, C.-Y. Hierarchical Bayesian Modeling and Randomized Response Method for Inferring the Sensitive-Nature Proportion. *Mathematics* **2021**, *9*, 2518. https://doi.org/10.3390/math9192518

Academic Editors: Christophe Chesneau, María Del Mar Rueda and Andrés Cabrera-León

Received: 17 August 2021
Accepted: 3 October 2021
Published: 7 October 2021

1. Introduction

Warner [1] is the pioneer to propose the randomized response (RR) method for evaluating the sensitive-nature proportion (SNP). An individual in the RR method of [1] is required to answer "yes" or "no" to either the statement "I am a member of group A" or "I am not a member of group A" where group A is a sensitive-nature statement. Since the interviewees do not need to release the selected statement to the interviewer, the interviewees can be supposed to confide in the interviewer the true answer.

Let θ denote the population SNP and n be the sample size of interviewees. In the RR survey, Y interviewees answer "yes". It is trivial to show that Y follows a binomial distribution with the sample size n and proportion of $\omega = P(yes) = \theta p_s + (1-\theta)(1-p_s)$, where p_s denotes the proportion of selecting the sensitive-nature statement. Denote $Y \sim Bin(n, \omega)$. Warner [1] obtained the maximum likelihood estimator of θ, denoted by $\hat{\theta}$, and its variance is as follows:

$$\hat{\theta} = \frac{p_s - 1}{2p_s - 1} + \frac{1}{(2p_s - 1)} \times \frac{y}{n} \tag{1}$$

and

$$Var(\hat{\theta}) = \frac{\theta(1-\theta)}{n} + \frac{1}{4n}\left[\frac{1}{(2p_s-1)^2} - 1\right]. \tag{2}$$

We use the term MLE to denote maximum likelihood estimator/estimate here. Equations (1) and (2) can be used to construct the confidence interval of θ. The MLE proposed by [1] is valid only if $p_s \neq 0.5$ and $0 \leq \theta \leq 1$. When the value of θ is small, the MLE of θ could be negative. This fact makes the method of [1] invalid to infer θ, as its true value is small.

New RR methods have been proposed after [1]. Greenberg et al. [2] proposed an unrelated question RR methods. They proposed a theoretical framework to infer the model parameters for the design of two statements. A new RR procedure was proposed by Mangat and Singh [3]. Their method used two randomization devices to design the RR strategy. Mangat and Singh [3] demonstrated that their new strategy was more efficient than the usual strategy of [1].

Kuk [4] proposed an alternative method to perform an RR survey. The design method of [4] does not require direct answers from the interviewees, and such a design can enhance the confidence of interviewees to tell the true answer. The method of [4] can be applied to both qualitative and quantitative questions. Kuk [4] suggested to collect data for a mixture distribution, and the problem can reduce to the estimation of a mixture proportion.

Chaudhuri [5] emphasized the protection of the interviewee's privacy and also studied the impact of simple random sampling design on the final conclusions. Chaudhuri [5] illustrated two existing RR devices for indicating how an estimator along with an estimated measure of its error could be developed when the RR sample may be drawn adopting a complex survey design involving unequal selection probabilities with or without replacement.

Christofides [6] proposed a generalized RR (GRR) technique to eliminate a major bias in surveys of the population SNP resulting from an interviewee's refusal when using the RR method of [1]. Chang et al. [7] considered a simple generalization for some existing investigations and suggested suitable selection strategies for design parameters. They also discussed the superiority of their proposed strategies over the RR strategy of [1].

Hsieh et al. [8] proposed a modified GRR (MGRR) approach for a multi-level attribute using a single sensitive item. The MGRR approach has some merits over the other counterparts. Hsieh et al. [8] suggested using the Markov chain Monte Carlo (MCMC) method to obtain the Bayes estimator of the SNP instead of the maximum likelihood estimation method. We use the term BE to denote Bayes estimator/Bayes estimate here. Examples about using Bayesian methods for real applications can be found in the book of Gelman et al. [9].

Bar-Lev et al. [10] presented a Bayesian approach to four RR models. They used truncated beta distributions in a common conjugate prior structure to obtain the BE of the SNP. Barabesi and Marcheselli [11] proposed a Bayesian estimation procedure to obtain the BE of the SNP based on Frankin's RR procedure. They conducted a simulation study to evaluate the quality of their proposed method.

Barabesi and Marcheselli [12] proposed a Bayesian method to the joint estimation of the SNPs and sensitivity level of a stigmatizing attribute via applying a two-stage RR design. The MGRR method is designed for a multi-level attribute using a single sensitive-nature statement. Hsieh et al. [8] suggested using the MCMC approach to obtain the BE of the SNP. The MGRR method is effective to obtain a reliable BE of the SNP. However, the MGRR method could be too complicated to implement for users.

Bayesian estimation methods are useful for modeling multi-faceted or nonlinear practical phenomena other than the maximum likelihood estimation method. Among all popular Bayesian estimation methods, the hierarchical Bayesian (HB) modeling method can be run with multiple hierarchical levels for estimating the parameters of posterior distribution. If grouped observations are used in a survey, hierarchical modeling is a relevant design to obtain the reliable BEs of model parameters.

The example in Section 4 of this study is based on college students from different groups to study the homosexual proportion in a region during different years. Hence, the HB modeling method is helpful to obtain the reliable BEs of model parameters. The HB modeling method has been commonly applied in many different areas when the

information on several different levels of observational units is available; see [12] for comprehensive discussions. It is helpful to apply hierarchical analysis forms to understand multi-parameter problems and design computational strategies.

However, heavy computation loading is a problem to obtain BEs of the model parameters using the HB modeling method. Taking advantage of the recent advances of computer power, it becomes easier to reduce the impact of computation loading when using HB modeling methods for data analysis. Some applications using the HB modeling method other than the RR design can be found in [13–22]. The HB modeling method is not yet applied for RR design. The implementation of the proposed 3RR-HB modeling method is discussed in Section 3.

2. Motivation and Organization

The traditional RR methods of [1,2] are simple to use. However, the MLE of the SNP could be negative when the true value of the SNP is small. In order to make interviewees more confident to tell the true answer in the survey, it is helpful to use more than one non-sensitive-nature statement in the RR method. Hence, we extend the traditional RR design of [2] to a three-statement RR method, which contains one sensitive-nature and two non-sensitive-nature statements.

To escape the trap of obtaining a negative MLE of the SNP, the HB modeling method is adopted to improve the drawback of the maximum likelihood estimation method when the true value of the SNP is small. In order to overcome the complexity of numerical computation to obtain the BE of the SNP, the hybrid algorithm of using Gibbs sampling in the Metropolis–Hastings algorithm is proposed to implement the MCMC method to obtain the BEs of model parameters. The main contribution of this study is to propose a 3RR-HB procedure to obtain the BE of the SNP. Moreover, the highest posterior density interval (HPDI) of the SNP is constructed. The proposed 3RR-HB procedure can improve the drawback of using a negative MLE to estimate the population SNP when its true value is small.

The rest of this paper is organized as follows: In Section 3, we present the data structure and introduce the proposed 3RR-HB procedure. In Section 4, an example regarding the homosexual proportion of college freshmen is used to demonstrate the applications of the proposed 3RR-HB procedure. A Monte Carlo simulation study is also conducted in Section 4 to study the weakness of maximum likelihood estimation method and evaluate the quality of the proposed 3RR-HB procedure. Some concluding remarks are given in Section 5.

3. Materials and Methods

Conducting a RR survey with two non-sensitive-nature and one sensitive-nature statements as follows: (i) I am in Group A; (ii) I am in Group B; and (iii) I am in Group C. Group A is a sensitive-nature statement; Group B and Group C are non-sensitive-nature statements. Interviewees have probabilities p_s, p_1 and p_2 to randomly answers the Statements (i), (ii) and (iii). In this study, the values of p_s, p_1 and p_2 are pre-assigned.

The interviewees do not need to release which statement they have replied to the interviewer. Let θ, δ_1 and δ_2 denote the probabilities of individual answering "yes" under the Statement (i), (ii) and (iii), respectively and ω denote the probability of answering "yes" in the sample. It is trivial to shown that $\omega = P(yes) = p_s\theta + p_1\delta_1 + p_2\delta_2$. In this study, the values of δ_1 and δ_2 are known in the RR design.

Let the sample size be n, in which Y of them answer "yes". It can be shown that $Y \sim Bin(n, \omega)$. The log-likelihood function based on the data (n, y) can be presented as

$$\ell(\omega \mid n, y) = y \log(\omega) + (n - y) \log(1 - \omega). \tag{3}$$

The MLE of θ to maximize $\ell(\omega \mid n, y)$ can be presented by

$$\hat{\theta} = \frac{\bar{Y} - (p_1\delta_1 + p_2\delta_2)}{p_s}, \tag{4}$$

where $\bar{Y} = \frac{y}{n}$. We note that $\hat{\theta}$ is valid only if the working condition of

$$0 \leq \frac{\bar{Y} - (p_1\delta_1 + p_2\delta_2)}{p_s} \leq 1 \tag{5}$$

is true. Unfortunately, Equation (5) is often violated when the value of θ is small. The fact makes the maximum likelihood estimation method unreliable for the cases of small θ. We will show that the failure rate of $P(\hat{\theta} < 0)$ is high via using the Monte Carlo simulation method in Section 4.

The Bayesian inference method is used to obtain the BE of θ. Let ω be random and follow the prior distribution of Beta, $\omega \sim Beta(\alpha, \beta)$:

$$\pi(\omega \mid \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\omega^{\alpha-1}(1-\omega)^{\beta-1}, 0 < \omega < 1, \tag{6}$$

where $\alpha > 0$ and $\beta > 0$ are hyper-parameters. The posterior distribution can be presented by

$$\pi(\omega \mid n, y, \alpha, \beta) \propto \omega^{y+\alpha-1}(1-\omega)^{n-y+\beta-1}. \tag{7}$$

In some occasions, we may collect RR samples from $k(> 1)$ different regions or time periods. One example is to evaluate the homosexual proportions of the freshmen who enrolled in a university over different years. Since the enrolled freshmen come from different cities year by year, it is reasonable to assume that the proportion of homosexual freshmen varies year by year. Therefore, the SNPs are the proportions of homosexual freshmen in k years, denoted by $\theta_1, \theta_2, \cdots, \theta_k$.

These values of $\theta_1, \theta_2, \cdots, \theta_k$ are different. In this study, we are interested in studying the trend of the population proportion of the homosexual freshmen in a university over years. A 3RR-HB method is developed to obtain the BEs of SNPs along with $i = 1, 2, \cdots, k$. If all SNPs are same; that is, $\theta_i = \theta$ for $i = 1, 2, \cdots, k$.

Taking summation to the both sides of the equation $\omega_i = p_{s,i}\theta + p_{1,i}\delta_{1,i} + p_{2,i}\delta_{2,i}$ for $i = 1, 2, \cdots, k$, we can obtain $\sum_{i=1}^{k}\omega_i = (\sum_{i=1}^{k} p_{s,i}\theta) + (\sum_{i=1}^{k} p_{1,i}\delta_{1,i}) + (\sum_{i=1}^{k} p_{2,i}\delta_{2,i})$. Let $\omega_\cdot = \sum_{i=1}^{k} w_i$, $\delta_{1\cdot} = \sum_{i=1}^{k} p_{1,i}\delta_{1,i}$ and $\delta_{2\cdot} = \sum_{i=1}^{k} p_{2,i}\delta_{2,i}$, we can obtain

$$\begin{aligned} \theta &= \frac{\omega_\cdot - \delta_{1\cdot} - \delta_{2\cdot}}{\sum_{i=1}^{k} p_{s,i}} \\ &= \frac{\sum_{i=1}^{k}(\omega_i - p_{1,i}\delta_{1,i} - p_{2,i}\delta_{2,i})}{\sum_{i=1}^{k} p_{s,i}} \end{aligned} \tag{8}$$

Replacing $\omega_i - p_{1,i}\delta_{1,i} - p_{2,i}\delta_{2,i}$ in Equation (8) by the $p_{s,i}\tilde{\theta}_i$ based on the ith sub-sample, the BE of θ can be presented by

$$\tilde{\theta} = \frac{\sum_{i=1}^{k} p_{s,i}\tilde{\theta}_i}{\sum_{i=1}^{k} p_{s,i}}.$$

Assume that RR sub-samples were collected from k different regions or time periods and the values of $\theta_1, \theta_2, \cdots, \theta_k$ are different. Let (n_i, Y_i) denote the ith RR sample and $Y_i \sim Bin(n_i, \omega_i)$, where $\omega_i = p_{s,i}\theta_i + p_{1,i}\delta_{1,i} + p_{2,i}\delta_{2,i}$ for $i = 1, 2, \cdots, k$. Let $\boldsymbol{N} = (n_1, n_2, \cdots, n_k)$ and $\boldsymbol{Y} = (Y_1, Y_2, \cdots, Y_k)$, and the data structure can be simplified as $(\boldsymbol{N}, \boldsymbol{Y})$. To avoid subjectively setting up the values of hyper-parameters, the HB modeling method is used to

develop the proposed Bayesian inference procedure. Let $\omega_i \sim Beta(\alpha, \beta)$ for $i = 1, 2, \cdots, k$. The density function of ω_i can be denoted by

$$\pi(\omega_i \mid \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\omega_i^{\alpha-1}(1-\omega_i)^{\beta-1}, 0 < \omega_i < 1, i = 1, 2, \cdots, k. \tag{9}$$

Using the square loss function for the Bayesian inference, the BE of ω, denoted by $\tilde{\omega}$, is the posterior mean based on the $\pi(\omega \mid n, y, \alpha, \beta)$ in Equation (7), and we can present $\tilde{\omega}$ by

$$\tilde{\omega} = \frac{y+\alpha}{n+\alpha+\beta}. \tag{10}$$

Therefore, the BE of θ can be obtained by

$$\tilde{\theta} = \frac{\tilde{\omega} - p_1\delta_1 - p_2\delta_2}{p_s}. \tag{11}$$

It is trivial that $\tilde{\theta} > 0$ if $\tilde{\omega} > p_1\delta_1 + p_2\delta_2$. It could be subjective to select the values of α and β for Bayesian inference. Hence, the HB modeling method is used in the proposed Bayesian estimation procedure to obtain the BEs of model parameters.

To implement the HB modeling method, we need to assume the second layer of prior distribution. Let α and β follow a hyper-prior distribution with the structure of a product of two Gamma distributions:

$$\varphi(\alpha, \beta) = \varphi_1(\alpha) \times \varphi_2(\beta), \tag{12}$$

where

$$\varphi_1(\alpha) = \frac{\xi_1^{\eta_1}}{\Gamma(\eta_1)}\alpha^{\eta_1-1}e^{-\xi_1\alpha}, \alpha > 0, \tag{13}$$

and

$$\varphi_2(\beta) = \frac{\xi_2^{\eta_2}}{\Gamma(\eta_2)}\beta^{\eta_2-1}e^{-\xi_2\beta}, \beta > 0. \tag{14}$$

For simplification, let $\boldsymbol{\Theta} = (\alpha, \beta, \omega_1, \cdots, \omega_k)$. The full posterior distribution can be presented by

$$\begin{aligned}\pi(\boldsymbol{\Theta} \mid \boldsymbol{N}, \boldsymbol{y}) \propto &\left[\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\right]^k \left\{\prod_{i=1}^{k}\omega_i^{y_i+\alpha-1}(1-\omega_i)^{n_i-y_i+\beta-1}\right\} \\ &\times \alpha^{\eta_1-1}e^{-\xi_1\alpha} \times \beta^{\eta_2-1}e^{-\xi_2\beta}.\end{aligned} \tag{15}$$

Moreover, the conditional posterior of ω_i, given α and β can be presented by $\omega_i \sim Beta(y_i + \alpha, n_i - y_i + \beta)$, $i = 1, 2, \cdots, k$. The value of θ_i during the MCMC computation can be updated by

$$\tilde{\theta}_i^{(*)} = \frac{1}{p_{s,i}}\left[\frac{y_i+\alpha}{n_i+\alpha+\beta} - (p_{1,i}\delta_{1,i} + p_{2,i}\delta_{2,i})\right], i = 1, 2, ..., k. \tag{16}$$

Based on the condition of $\frac{y_i+\alpha}{n_i+\alpha+\beta} - (p_{1,i}\delta_{1,i} + p_{2,i}\delta_{2,i}) > 0$, we can obtain

$$\beta < \frac{y_i+\alpha}{p_{1,i}\delta_{1,i} + p_{2,i}\delta_{2,i}} - (n_i+\alpha).$$

Given the values of α and $(\boldsymbol{N}, \boldsymbol{Y})$, we can shown that

$$\Omega_{\beta|\alpha,\boldsymbol{N},\boldsymbol{Y}} = \{0 < \beta < c_m(\alpha, \boldsymbol{N}, \boldsymbol{Y})\}, \tag{17}$$

where

$$c_m(\alpha, \boldsymbol{N}, \boldsymbol{Y}) = \min\left\{ \frac{y_i + \alpha}{p_{1,i}\delta_{1,i} + p_{2,i}\delta_{2,i}} - (n_i + \alpha), i = 1, 2, ..., k \right\}.$$

The set $\Omega_{\beta|\alpha,\boldsymbol{N},\boldsymbol{Y}}$ can be used to guarantee $\tilde{\theta}_i^{(*)} > 0$ for $i = 1, 2, \cdots, k$. Hence, $\Omega_{\beta|\alpha,\boldsymbol{N},\boldsymbol{Y}}$ can be a reference set to select the hyper-parameter β when a value of α is generated and the data of $(\boldsymbol{N}, \boldsymbol{Y})$ are collected.

The proposed hybrid algorithm is constructed as follows: Let $\boldsymbol{\Theta_{-1}} = (\beta, \omega_1, \cdots, \omega_k)$ and $\boldsymbol{\Theta_{-2}} = (\alpha, \omega_1, \cdots, \omega_k)$ denote the vector of parameters by removing α and β from $\boldsymbol{\Theta}$, respectively. After algebraic computation, the marginal density distributions of α and β can be obtained, respectively, by

$$\pi(\alpha \mid \boldsymbol{\Theta_{-1}}, \boldsymbol{N}, \boldsymbol{y}) \propto \left[\frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)}\right]^k \left\{\prod_{i=1}^{k} \omega_i^{\alpha}\right\} \times \alpha^{\eta_1 - 1} e^{-\xi_1 \alpha} \tag{18}$$

and

$$\pi(\beta \mid \boldsymbol{\Theta_{-2}}, \boldsymbol{N}, \boldsymbol{y}) \propto \left[\frac{\Gamma(\alpha + \beta)}{\Gamma(\beta)}\right]^k \left\{\prod_{i=1}^{k} (1 - \omega_i)^{\beta}\right\} \times \beta^{\eta_2 - 1} e^{-\xi_2 \beta}, \tag{19}$$

In order to overcome the difficulty to update α and β via using Equations (18) and (19) in the Gibbs sampling procedure, the Metropolis and Hastings algorithm is used to update α and β. Hence, the proposed hybrid algorithm for implementing the HB modeling method can be followed based on the following steps:

Step 1: For $j \geq 1$, generate $\alpha^{(*)} \sim q_1\left(\alpha^{(*)} \mid \alpha^{(j)}\right)$ and $u \sim U(0,1)$, where $q_1(\cdot)$ is the proposal to generate α. Update $\alpha^{(j+1)}$ by $\alpha^{(*)}$ if $u \leq \Psi_1^{(j)}$, where

$$\Psi_1^{(j)} = min\left\{1, \frac{\pi\left(\alpha^{(*)} \mid \beta^{(j)}, \omega_1^{(j)}, \cdots, \omega_k^{(j)}, \boldsymbol{N}, \boldsymbol{y}\right) q_1\left(\alpha^{(j)} \mid \alpha^{(*)}\right)}{\pi\left(\alpha^{(j)} \mid \beta^{(j)}, \omega_1^{(j)}, \cdots, \omega_k^{(j)}, \boldsymbol{N}, \boldsymbol{y}\right) q_1\left(\alpha^{(*)} \mid \alpha^{(j)}\right)}\right\}; \tag{20}$$

otherwise, $\alpha^{(j+1)} = \alpha^{(j)}$.

Step 2: For $j \geq 1$, generate $\beta^{(*)} \sim q_2\left(\beta^{(*)} \mid \beta^{(j)}\right)$ and $u \sim U(0,1)$, where $q_2(\cdot)$ is the proposal to generate β. Update $\beta^{(j+1)}$ by $\beta^{(*)}$ if $u \leq \Psi_2^{(j)}$, where

$$\Psi_2^{(j)} = min\left\{1, \frac{\pi\left(\beta^{(*)} \mid \alpha^{(j+1)}, \omega_1^{(j)}, \cdots, \omega_k^{(j)}, \boldsymbol{N}, \boldsymbol{y}\right) q_2\left(\beta^{(j)} \mid \beta^{(*)}\right)}{\pi\left(\beta^{(j)} \mid \alpha^{(j+1)}, \omega_1^{(j)}, \cdots, \omega_k^{(j)}, \boldsymbol{N}, \boldsymbol{y}\right) q_2\left(\beta^{(*)} \mid \beta^{(j)}\right)}\right\}; \tag{21}$$

otherwise, $\beta^{(j+1)} = \beta^{(j)}$.

Step 3: Generate $\omega_i^{(*)} \sim Beta\left(y_i + \alpha^{(j+1)}, n_i - y_i + \beta^{(j+1)}\right)$ and evaluate $\theta_i^{(*)}$ by $\theta_i^{(*)} = \frac{1}{p_{s,i}}\left[\omega_i^{(*)} - (p_{1,i}\delta_{1,i} + p_{2,i}\delta_{2,i})\right]$ for $i = 1, 2, \cdots, k$. If $0 \leq \theta_i^{(*)} \leq 1$, update $\theta_i^{(j+1)} = \theta_i^{*}$; otherwise, $\theta_i^{(j+1)} = \theta_i^{(j)}$, $i = 1, 2, \cdots, k$.

Step 4: Repeat Step 1 to Step 3 B times, where B is a big positive integer. Perform the burn-in step by removing the leading B_1 Markov chains. The BE of parameter is obtained via using the remainder $(B - B_1)$ Markov chains. Since the square loss function is considered in this study, the BEs $\tilde{\alpha}$, $\tilde{\beta}$ and $\tilde{\theta}_i$, $i = 1, 2, \cdots, k$.

Some proposals with the property of $q(\tau^{(j)}|\tau^{(*)}) = q(\tau^{(*)}|\tau^{(j)})$, where τ is the target parameter for update, can be selected to reduce the computation loading of MCMC—for

example, the normal or uniform distribution. When such proposals are used to implement the MCMC approach, Equations (20) and (21) can reduce to

$$\begin{aligned}\Psi_1^{(j)} &= min\left\{1, \frac{\pi\left(\alpha^{(*)} \mid \beta^{(j)}, \omega_1^{(j)}, \cdots, \omega_k^{(j)}, \boldsymbol{N}, \boldsymbol{y}\right)}{\pi\left(\alpha^{(j)} \mid \beta^{(j)}, \omega_1^{(j)}, \cdots, \omega_k^{(j)}, \boldsymbol{N}, \boldsymbol{y}\right)}\right\} \\ &= min\left\{1, \left[\frac{\Gamma(\alpha^{(*)}+\beta^{(j)})\Gamma(\alpha^{(j)})}{\Gamma(\alpha^{(j)}+\beta^{(j)})\Gamma(\alpha^{(*)})}\right]\left\{\Pi_{i=1}^k\left(\omega_i^{(j)}\right)^{\alpha^{(*)}-\alpha^{(j)}}\right\}\right. \\ &\quad \left.\times\left(\frac{\alpha^{(*)}}{\alpha^{(j)}}\right)^{\eta_1-1} e^{\xi_1(\alpha^{(j)}-\alpha^{(*)})}\right\}\end{aligned}$$

and

$$\begin{aligned}\Psi_2^{(j)} &= min\left\{1, \frac{\pi\left(\beta^{(*)} \mid \alpha^{(j+1)}, \omega_1^{(j)}, \cdots, \omega_k^{(j)}, \boldsymbol{N}, \boldsymbol{y}\right)}{\pi\left(\beta^{(j)} \mid \alpha^{(j+1)}, \omega_1^{(j)}, \cdots, \omega_k^{(j)}, \boldsymbol{N}, \boldsymbol{y}\right)}\right\} \\ &= min\left\{1, \left[\frac{\Gamma(\alpha^{(j+1)}+\beta^{(*)})\Gamma(\beta^{(j)})}{\Gamma(\alpha^{(j+1)}+\beta^{(j)})\Gamma(\beta^{(*)})}\right]\left\{\Pi_{i=1}^k\left(1-\omega_i^{(j)}\right)^{\beta^{(*)}-\beta^{(j)}}\right\}\right. \\ &\quad \left.\times\left(\frac{\beta^{(*)}}{\beta^{(j)}}\right)^{\eta_2-1} e^{\xi_2(\beta^{(j)}-\beta^{(*)})}\right\},\end{aligned}$$

respectively. In this study, the normal distribution is considered as the proposal for MCMC approach. Generate $\alpha \sim N(\alpha^{(j)}, 1)$ and $\beta \sim N(\beta^{(j)}, 1)$. If $\alpha < 0$, we do not update α; if $\beta \notin \Omega_{\beta|\alpha,\boldsymbol{N},\boldsymbol{Y}}$, we do not update β. The obtained Markov chains based on the proposed 3RR-HB procedure can also be used to construct the empirical distribution of BE. Then, the HPDI of the model parameter can be obtained via using the empirical distribution of BE. The applications of the proposed 3RR-HB procedure and its quality will be studied in Section 4 via using a real example and Monte Carlo simulations.

4. Applications

4.1. Homosexual College Freshmen Survey Example

A RR survey was first conducted in October of 2019 and repeated in October of 2020 to evaluate the homosexual proportion of freshmen in a university, located in north Taiwan. College students from different majors were interviewed and asked to answer "yes" or "no" to one of the following three statements: (i) Is the last code of your ID card even? (ii) Is the last code of your student ID card even? (iii) Are you homosexual? Clearly, only Statement (iii) is a sensitive-nature statement.

Each student in the RR survey randomly draw a ball from a urn, which contains five white, five green and five red balls. If the drawn ball is white, answer Statement (i); if the drawn ball is green, answer Statement (ii); otherwise, answer Statement (iii). Students do not need to release the ball's color to the interviewer. The three-statement RR design can make students fully confide the true answer in the interviewer.

The samples are $n_1 = 283$, $y_1 = 101$ in 2019 and $n_2 = 178$, $y_2 = 60$ in 2020. It is trivial to obtain that $p_s = p_1 = p_2 = 1/3$ and $\delta_1 = \delta_2 = 1/2$. The data set is composed of $\boldsymbol{N} = (283, 178)$ and $\boldsymbol{Y} = (101, 60)$. The MLEs of θ_1 and θ_2 can be obtained by $\hat{\theta}_1 = 0.071$ and $\hat{\theta}_2 = 0.011$ via using Equation (4). We note that the values of $\hat{\theta}_1$ and $\hat{\theta}_2$ are small and that these two MLEs could be unreliable. The proposed 3RR-HB procedure in Section 3 is applied to characterize the data sets and obtain the BEs with $B = 100,000$ and $B_1 = 10,000$.

We need to set up the parameters of $\xi_1, \eta_1, \xi_2, \eta_2$ in the hyper-prior distribution. Based on the sample information of $y_1/n_1 = 101/283 = 0.357$ and $y_2/n_2 = 60/178 = 0.337$, we can assume that the expected values $E(\omega_1) = E(\omega_2) = \alpha/(\alpha+\beta)$ is in (0.33, 0.36). Since

we do not have sufficient information to set up the values of ξ_1, η_1, ξ_2, η_2 in Equations (13) and (14), we would select the proper vales of ξ_1, η_1, ξ_2, η_2 to generate large variances of α and β, and hence the prior distributions of α and β become non-informative to obtain BEs of the model parameters.

The combinations of ξ_1, η_1, ξ_2 and η_2 that satisfy the condition of $\chi \in (0.33, 0.36)$ can be selected for simulations where $\chi = \frac{E(\alpha)}{E(\alpha)+E(\beta)}$. A simulation study of sensitivity analysis for the selections of possible values of ξ_1, η_1, ξ_2 and η_2 was conducted to show that our proposed method is less sensitive to the selection of ξ_1, η_1, ξ_2 and η_2; that is, the obtained BEs of model parameters are less sensitive to the selection of ξ_1, η_1, ξ_2 and η_2. Five combinations of ξ_1, η_1, ξ_2 and η_2 were selected.

All these combinations can generate a wide range of large to extreme large variances of α and β. All selected parameters combinations can have similar values of χ; see Table 1. We label these five combinations as C-I: $(\xi_1, \eta_1, \xi_2, \eta_2) = (0.050, 1.5, 0.050, 2.8)$, C-II: $(\xi_1, \eta_1, \xi_2, \eta_2) = (0.040, 1.5, 0.040, 2.8)$, C-III: $(\xi_1, \eta_1, \xi_2, \eta_2) = (0.035, 1.5, 0.035, 2.8)$, C-IV: $(\xi_1, \eta_1, \xi_2, \eta_2) = (0.030, 1.5, 0.030, 2.8)$ and C-V: $(\xi_1, \eta_1, \xi_2, \eta_2) = (0.025, 1.5, 0.025, 2.8)$. Using the normal proposals for the proposed MCMC procedure to generate $\alpha \sim N(\alpha^{(*)}, 1)$ and $\beta \sim N(\beta^{(*)}, 1)$. If $\alpha < 0$, we do not update α; if $\beta \notin \Omega_{\beta|\alpha,\boldsymbol{N},\boldsymbol{Y}}$, we do not update β. All the estimation results are reported in Tables 2 and 3.

Table 1. The parameters in the hyper-prior for simulations.

	ξ_1	η_1	ξ_2	η_2	E(α)	Var(α)	E(β)	Var(β)	χ
C-I	0.050	1.5	0.050	2.8	30	600	56	1120	0.3489
C-II	0.040	1.5	0.040	2.8	38	938	70	1750	0.3989
C-III	0.035	1.5	0.035	2.8	43	1224	80	2286	0.3489
C-IV	0.030	1.5	0.030	2.8	50	1667	93	3111	0.3489
C-V	0.025	1.5	0.025	2.8	60	2400	112	4480	0.3489

Table 2. The obtained BEs and their standard errors (SEs) via using the proposed 3RR-HB method.

	$\tilde{\omega}_1$		$\tilde{\omega}_2$		$\tilde{\theta}_1$		$\tilde{\theta}_2$	
	BE	**SE**	**BE**	**SE**	**BE**	**SE**	**BE**	**SE**
C-I	0.355	0.027	0.341	0.032	0.092	0.061	0.082	0.061
C-II	0.355	0.026	0.341	0.032	0.090	0.060	0.080	0.061
C-III	0.355	0.026	0.342	0.031	0.091	0.060	0.081	0.061
C-IV	0.355	0.026	0.342	0.031	0.089	0.059	0.078	0.060
C-V	0.352	0.025	0.344	0.028	0.078	0.055	0.070	0.054

Table 3. The acceptance rates for α, β, θ_1 and θ_2.

	α	β	θ_1	θ_2
C-I	0.925	0.951	0.789	0.584
C-II	0.937	0.950	0.790	0.596
C-III	0.936	0.951	0.794	0.605
C-IV	0.943	0.948	0.790	0.605
C-V	0.964	0.870	0.771	0.638

In view of Tables 2 and 3, we can find the strength of the proposed 3RR-HB procedure. Table 2 shows that the obtained BEs based on the proposed 3RR-HB procedure are reliable. All bias and standard errors (SEs) of BEs based on the proposed 3RR-HB procedure with

10,000 iterations, where $B = 50,000$ and $B_1 = 10,000$, are small. Moreover, all acceptance rates in Table 3 are larger than 50%. Since $\omega_i^{(*)} \sim Beta(y_i + \alpha^{(j+1)}, n_i - y_i + \beta^{(j+1)})$, all the generated values are accepted. Therefore, the acceptance rates of ω_1 and ω_2 are 100%. Moreover, the obtained BE of the homosexual proportion of college freshmen in the region is about 9% in 2019 and 8% in 2020.

The empirical distribution can be constructed based on the obtained Markov chains of BEs. Using the values of ξ_1, η_1, ξ_2 and η_2 of C-III in Table 1, the 90% HPDI of θ_1 and θ_2 are (0, 0.174) and (0, 0.168), respectively. Overall, we can conclude that the point homosexual proportion of college freshmen in the region in 2019 and 2020 was about 9% and 8% and then up to 17.4% and 16.8%, respectively, under the considering of sampling error with 90% confidence. The length of HPDI is long as the sample size is small. If the confident coefficient 95% is used instead of 90%, the length of the HPDI of θ_i will be longer than the 90% HPDI for $i = 1, 2$.

4.2. Monte Carlo Simulations

A Monte Carlo simulation study was conducted to verify the quality of the proposed 3RR-HB procedure. Referring to the example information in Section 4.1, we let $p_s = p_1 = p_2 = 1/3$, $\delta_1 = \delta_2 = 1/2$ and $N = (n_1, n_2)$ with $n_1 = n_2 = n = 200, 300, 500, 800, 1000$ to generate random samples of y_1 and y_2. For each given value of θ_i, we generate $Y_i \sim Bin(n_i,\ \omega_i = \theta_i/3 + 1/3)$ for $i = 1, 2$.

Let $\theta_1 = \theta_2 = 0.05$ and 0.1. The proposed 3RR-HB procedure with B = 50,000, B_1 = 10,000 is implemented. Moreover, the BEs of θ_i, $i = 1, 2$ are obtained based on the values of ξ_1, η_1, ξ_2, η_2 of C-III in Table 1. Repeat each simulation procedure 10,000 times, and then the bias and mean squared error (MSE) of each BE are evaluated based on the 10,000 runs. All simulation results are reported in Figure 1 and Tables 4–6.

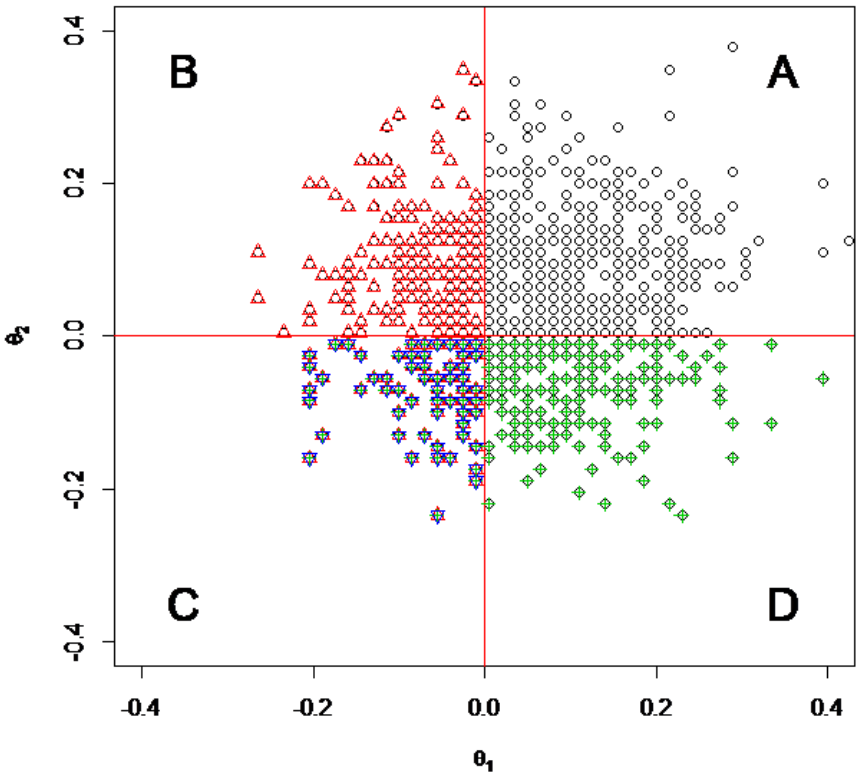

Figure 1. The scatter plot of 1000 MLEs of $\hat{\theta}_1$ and $\hat{\theta}_2$ for the case of $n = 200$ and $\theta_1 = \theta_2 = 0.05$.

Table 4. The sample proportions of the MLEs of $\hat{\theta}_1 > 0$ and $\hat{\theta}_2 > 0$ based on 10,000 runs.

	$\theta_1 = \theta_2 = 0.05$				
	$n = 200$	$n = 300$	$n = 500$	$n = 800$	$n = 1000$
$\hat{\theta}_1$	0.702	0.752	0.782	0.837	0.860
$\hat{\theta}_2$	0.693	0.751	0.784	0.842	0.863
	$\theta_1 = \theta_2 = 0.10$				
$\hat{\theta}_1$	0.848	0.896	0.940	0.978	0.986
$\hat{\theta}_2$	0.840	0.898	0.940	0.976	0.999

Table 5. The bias and MSEs of BEs for $\theta_1 = \theta_2 = 0.05$.

n	$\tilde{\theta}_1$		$\tilde{\theta}_2$	
	Bias	**MSE**	**Bias**	**MSE**
200	0.0537	0.0049	0.0574	0.0058
300	0.0409	0.0033	0.0440	0.0039
500	0.0287	0.0021	0.0299	0.0023
800	0.0194	0.0013	0.0200	0.0014
1000	0.0158	0.0011	0.0165	0.0012

Table 6. The bias and MSEs of BEs for $\theta_1 = \theta_2 = 0.10$.

n	$\tilde{\theta}_1$		$\tilde{\theta}_2$	
	Bias	**MSE**	**Bias**	**MSE**
200	0.0320	0.0043	0.0347	0.0047
300	0.0222	0.0032	0.0238	0.0034
500	0.0120	0.0022	0.0121	0.0022
800	0.0069	0.0017	0.0061	0.0017
1000	0.0043	0.0015	0.0048	0.0015

Figure 1 displays the scatter plot of the first 1,000 MLEs of $\hat{\theta}_1$ and $\hat{\theta}_2$ for $n = 200$ and $\theta_1 = \theta_2 = 0.05$. We can see that many pairs of $\hat{\theta}_1$ and $\hat{\theta}_2$ are in Zones B, C, or D, and those pairs are invalid MLEs due to $\hat{\theta}_1 < 0$ or $\hat{\theta}_2 < 0$. Only the pairs of $\hat{\theta}_1$ and $\hat{\theta}_2$ in Zone A are valid due to the required conditions $\theta_1 > 0$ and $\theta_2 > 0$.

We found that 7020 MLEs of θ_1 with the proportion of 0.702 and 6,930 MLEs of θ_2 with the proportion of 0.603 in Table 4 were positive when $n = 200$ and $\theta_1 = \theta_2 = 0.05$. In a scan of Table 4, we find that the maximum likelihood estimation method has a high risk to generate invalid values of $\hat{\theta}_1$ and $\hat{\theta}_2$ if the true values of θ_1 and θ_2 are closed to zero. The sample proportions of $P(\hat{\theta}_1 > 0)$ and $P(\hat{\theta}_2 > 0)$ are increased as n is increased. However, as the RR method goes through an interview process, it could be difficult to collect an large sample to obtain reliable MLEs of θ_1 and θ_2.

As many negative MLEs were found, the bias and MSE of MLE are not reliable. Tables 5 and 6 only report the bias and MSE of BE. In view of Tables 5 and 6, we found that the bias and MSE of BE is declined as n increased. The bias and MSEs of $\tilde{\theta}_1$ and $\tilde{\theta}_2$ were small. Hence, the simulation results indicate that the proposed 3RR-HB procedure can be a reliable method to evaluate SNP.

5. Conclusions

In this paper, we proposed a 3RR-HB procedure to infer the SNP by considering a hierarchical structure for the prior distribution in Bayesian modeling. Moreover, the Beta-Binomial distribution was applied to characterize the RR samples. In order to overcome

the computation complexity, the hybrid algorithm of using Gibbs sampling in Metropolis–Hastings algorithm was adopted to update the model parameters during MCMC computation. The proposed 3RR-HB procedure method is simple and minimally subjective for use.

A data set regarding the homosexual proportion of college freshmen was used to illustrate the applications of the proposed 3RR-HB procedure. We also conducted Monte Carlo simulations to study the performance of the proposed 3RR-HB procedure. The simulation results showed that the proposed 3RR-HB procedure was reliable to obtain the BEs of model parameters. Moreover, the 3RR-HB procedure can help users to escape the drawback of using invalid MLE to estimate the SNP.

The design of equal probabilities for the three statements was used to obtain RR samples. Such a design will reduce the chance of interviewees to select the sensitive-nature statement. However, such a design can enhance the willing of interviewees to confide in the interviewer the true answer. The equal-probability design is a trade-off. Practitioners can use unequal-probability design to obtain RR samples to implement the proposed 3RR-HB procedure based on their considerations.

We only used one sensitive-nature statement to obtain RR samples. It will be interesting to expand the proposed method for the RR method containing two or more sensitive-nature statements. How to establish the HB modeling inference procedure for the RR method with two or more sensitive-nature statements is an open question that can be studied in the future.

Author Contributions: Data curation, C.-Y.H.; Funding acquisition, T.-R.T.; Investigation, T.-R.T.; Methodology, H.X. and J.Z.; Project administration, T.-R.T.; Resources, J.Z.; Software, H.X. and C.-Y.H.; Supervision, T.-R.T.; Writing—original draft, T.-R.T.; Writing—review and editing, T.-R.T. All authors have read and agreed to the published version of the manuscript.

Funding: This study is supported by the grant of Ministry of Science and Technology, Taiwan MOST 108-2221-E-032-018-MY2.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study. The randomize response survey is based on the project of Ministry of Science and Technology, Taiwan MOST 108-2813-C-032-002-M for the authors Chieh-Yi Hung and under the supervision of the author Tzong-Ru Tsai.

Data Availability Statement: The data can be found in Section 4.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Warner, S.L. Randomized response: A survey technique for eliminating evasive answer bias. *J. Am. Stat. Assoc.* **1965**, *60*, 63–69. [CrossRef]
2. Greenberg, B.G.; Abul-Ela, E.L.A.; Simmons, W.R.; Horvitz, D.G. The unrelated question randomized response model: Theoretical framework. *J. Am. Stat. Assoc.* **1969**, *64*, 520–539. [CrossRef]
3. Mangat, N.S.; Singh, R. An alternative randomized response procedure. *Biometrika* **1990**, *77*, 439–442. [CrossRef]
4. Kuk, A.Y.C. Asking sensitive questions indirectly. *Biometrika* **1990**, *77*, 436–438. [CrossRef]
5. Chaudhuri, A. Using randomized response from a complex survey to estimate a sensitive proportion in a dichotomous finite population. *J. Stat. Plan. Inference* **2001**, *94*, 37–42. [CrossRef]
6. Christofides, T.C. A generalized randomized response technique. *Metrika* **2003**, *57*, 195–200. [CrossRef]
7. Chang, H.-J.; Wang, C.-L.; Huang, K.-C. On estimating the oroportion of a qualitative sensitive character using randomized response sampling. *Qual. Quant.* **2004**, *38*, 375–680.
8. Hsieh, S.H.; Lee, S.-M.; Tu, S.-H. Randomized response techniques for a multi-level attribute using a single sensitive question. *Stat. Pap.* **2018**, *59*, 291–306. [CrossRef]
9. Gelman, A.; Carlin, J.B.; Stern, H.S.; Rubin, D.B. *Bayesian Data Analysis*, 2nd ed.; CRC Press: Boca Raton, FL, USA, 2004; ISBN 1-58488-388-X.
10. Bar-Lev, S.K.; Bobovich, E.; Boukai, B. A common conjugate prior structure for several randomized response models. *Test* **2003**, *12*, 101–113. [CrossRef]
11. Barabesi, L.; Marcheselli, M. A practical implementation and Bayesian estimation in Franklin's randomized response procedures. *Commun. Stat.-Simul. Comput.* **2006**, *35*, 563–573. [CrossRef]

12. Barabesi, L.; Marcheselli, M. Bayesian estimation of proportion and sensitivity level in randomized response procedures. *Metrika* **2010**, *72*, 75–88. [CrossRef]
13. Arnab, R.; Singh, S. Randomized response techniques: An application to the Botswana AIDS impact survey. *J. Stat. Plan. Inference* **2010**, *140*, 941–953. [CrossRef]
14. Bae, K.; Mallick, B.K. Gene selection using a two-level hierarchical Bayesian model. *Bioinformatics* **2004**, *20*, 3423–3430. [CrossRef]
15. Dobigeon, N.; Tourneret, J.-Y.; Chang, C.-I. Semi-Supervised linear spectral unmixing using a hierarchical Bayesian model for hyperspectral imagery. *IEEE Trans. Signal Process.* **2008**, *56*, 2684–2695. [CrossRef]
16. Behmanesh, I.; Moaveni, B.; Lombaert, G.; Papadimitriou, C. Hierarchical Bayesian model updating for structural identification. *Mech. Syst. Signal Process.* **2015**, *64*, 360–376. [CrossRef]
17. Weber, S.A.; Insaf, T.Z.; Hall, E.S.; Talbot, T.O.; Huff, A.K. Assessing the impact of fine particulate matter (PM2.5) on respiratory-cardiovascular chronic diseases in the New York City Metropolitan area using Hierarchical Bayesian model estimates. *Environ. Res.* **2016**, *151*, 399–409. [CrossRef] [PubMed]
18. Gastelu, J.V.; Trujillo, J.D.M.; Padilha-Feltrin, A. Hierarchical Bayesian model for estimating spatial-temporal photovoltaic potential in residential areas. *IEEE Trans. Sustain. Energy* **2018**, *9*, 971–979. [CrossRef]
19. Xie, Z.; Wu, T.; Yang, X.; Zhang, L.; Wu, K. Jointly social grouping and identification in visual dynamics with causality-induced hierarchical Bayesian model. *J. Vis. Commun. Image Represent.* **2019**, *59*, 62–75. [CrossRef]
20. Alarcon A.; Sánchez, C.; Bernstein, G.M. Redshift inference from the combination of galaxy colours and clustering in a hierarchical Bayesian model. *Mon. Not. R. Astron. Soc.* **2019**, *498*, 2801–2813.
21. Zhang, X.; Dong, Q.; Costa, V.; Wang, X. A hierarchical Bayesian model for decomposing the impacts of human activities and climate change on water resources in China. *Sci. Total. Environ.* **2019**, *665*, 836–847. [CrossRef]
22. Tonkin-Hill, G.; Lees, J. A.; Bentley, S.D.; Frost, S.D.W.; Corander, J. Fast hierarchical Bayesian analysis of population structure. *Nucleic Acids Res.* **2019**, *47*, 5539–5549. [CrossRef] [PubMed]