Bayesian Analysis of Unrelated Question Design for Correlated Sensitive Questions from Small Areas

Yuan Yu*　　Balgobin Nandram†

Abstract

In sample surveys with sensitive questions, randomized response techniques, like the unrelated question methodology, have a huge advantage in estimating population proportions by adjusting for non-response or untruthful response. In reality, multiple sensitive proportions from small areas could be of great interest. Therefore, we consider using the unrelated question design with multiple sensitive questions and single random mechanism. Given combined binary response data, we can construct a hierarchical Bayesian model with latent variables to get more accurate estimates. There is also a computing challenge since we need to estimate parameters from all stages of the model. Thus, Markov chain Monte Carlo methods are applied to predict the finite population proportions of sensitive attributes. We compare the estimation error under different area size and correlation between the two sensitive questions during the simulation study. In the end, an application on body mass index data from the Third National Health and Nutrition Examination Survey is provided to verify our procedures.

Key Words: Gibbs sampler, Hierarchical Bayesian, Latent variables, MCMC, Unrelated question design

1. Introduction

In survey sampling, more curious and problematic issue is to collect information about sensitive questions that a respondent have tendency to refuse to answer, or answer untruthfully. Some cases of sensitive and stigmatizing items in surveys like habitual tax evasion, drunken driving, gambling, consuming drugs, etc. S. L. Warner(JASA, 1965) gave a standard procedure for estimating the proportion of people bearing the sensitive character A, on adopting a suitable randomization device. This randomized response technique also called mirror's question design. In Warner's design, each individual will be required to play a random game, like throwing a die. Thus with probability p, the respondent will give his/her answer ('Yes' or 'No') of the sensitive question "Do you belong to the group A?"; with the probability $1 - p$, he/she will give the answer of the opposite question "Do you belong to the group A^C?". Notice here A^C represent the compliment set of A. Each sample unit is required to play the game unobserved by the interviewer and provide there responses in the end.

In fact, when randomized response techniques are used, a respondent's individual answer is not of interest, rather inference is needed for the population. the respondent does not provide a direct answer to the sensitive question, thus the identity of the respondent is protected while the true answer to the sensitive question is elicited. In this approach to survey sampling, the randomization is not only in drawing the sample but also in obtaining the response.

An important extension of the Warner's method is the unrelated question design, proposed by Greenberg, Abu-Ela, Simmons and Horvitz in 1969. Instead of asking an opposite

*Department of Mathematical Sciences, Worcester Polytechnic Institute, 100 Institute Road, Worcester, MA 01609

†Department of Mathematical Sciences, Worcester Polytechnic Institute, 100 Institute Road, Worcester, MA 01609

question, we ask an unrelated nonsensitive (innocuous) question. For example, each sample unit will flip an unfair coin, and choose one of the following questions to answer according to the result of head or tail.

Question 1: Have you ever cheated on an exam anywhere? (A)
Question 2: Are you born in Massachusetts? (U)

Circle your response. [Yes, No]

In this way the respondents should be more comfortable to answer the question because the investigator can never know which question the respondents are answering. Instead, we can only list the following probability relationships

$$\lambda_1 = p_1\pi_A + (1 - p_1)\pi_U$$
$$\lambda_2 = p_2\pi_A + (1 - p_2)\pi_U$$

where
λ_i= the probability that a "Yes" answer will be reported in the i^{th} sample;
p_i= the probability of answering the sensitive question in the first sample;
π_A = true proportion with the sensitive attribute A;
π_U = true proportion with the unrelated attribute U.

One needs to strike a compromise between efficiency and response burden but respondents' protection is paramount (US Privacy Act of 1974), currently a hotly debated issue in the US Congress especially in connection with the use of the Internet.

Direct questioning exposes a respondent's privacy that is obviously unacceptable. Any design, which adds noise to the response, will be less efficient than a direct questioning design. One cannot compromise respondents' privacy, but one can compromise respondents' burden and efficiency. However, it has been argued that socially desirable answers and refusals are expected when sensitive questions are asked directly (e.g., see Tourangeau, Rips and Rasinski 2000 and Tourangeau and Yan 2007). Evidently, as supported by many psychologists, sensitive questions should not be asked directly.

We assume that respondents respond truthfully. It should be obvious that this assumption is more easily attained under indirect questioning than direct question. In direct questioning, it is more likely that there will be nonresponse that may be nonignorable, and we need to develop nonignorable nonresponse models (Nandram and Choi 2002, 2010) to handle them. So at least on two fronts, indirect questioning is advantageous.

There is also a amount of detailed literature review of varies randomized response techniques, including many extensions of the unrelated question design. The forced response design (Fox & Tracy, 1986) is like the unrelated question design, but using the results coming fro two Bernoulli trials instead. The disguised design (Kuk, 1990) is like the other way around, the respondent need to report the results from two Bernoulli trials based on their answer to the sensitive questions. Blair, Imai, and Zhou (2015) gave an excellent review paper on these four method. Other nonrandomised designs like crosswise and triangular designs (e.g. Tan *et al*., 2009) can be viewed as extensions of the unrelated question design, which get rid of the random mechanism.

For continuous response, Greenberg *et al*. (1971) and Eriksson (1973) extended the unrelated question model of Greenberg *et al*. (1969) to the case in which the response is quantitative. Pollock and Bek (1976) described the additive/multiplicative models, which involve the respondent adding/multiplying the answer to the sensitive question by a random number from a known distribution. More recently, see Gupta et al. (2002), Gupta et al. (2010) for optional designs for quantitative data.

However, there are some concerns for the traditional method of getting the maximum likelihood estimation of the proportion. First, direct estimates from solving the equation systems cannot guarantee a reasonable solution between (0,1). Also large sample sizes are needed to get an admissible estimate. In the unrelated question design, $\hat{\pi}_A$ and $\hat{\pi}_U$ can be highly correlated, reducing the correlation by increasing the sample size will be costly. The design-based estimator may be practically biased in small samples. Especially for small areas, more discussion about the individual Bayesian model in Nandram and Yu, 2017.

There are not many works within the Bayesian paradigm of randomized response models. Nonetheless, attempts have been made on the Bayesian analysis of RRTs. For example, Winkler and Franklin (1979) gave an approximate Bayesian analysis of Warner's mirrored design, O'Hagan (1987) derived Bayesian linear estimators for the unrelated question design, and Oh (1994) used data augmentation to introduce latent variables to Gibbs sampling of the mirrored design, the unrelated question design and the two-stage design with binary and polychotomous responses. van den Hout and Klugkist (2009) proposed Bayesian inference that takes into account assumptions with respect to non-compliance under simple random sampling. Also, Tian, Yuen, Tang and Tan (2009) proposed Bayesian approaches to non-randomized response models without using random mechanisms. Avetisyan and Fox (2012) used beta-binomial and multinomial-Dirichlet models for empirical Bayes analysis and to a less extent Bayesian analysis for a small sample from a single population. They considered multiple items with multiple categories. But this latter work is not within the small area context although each person may be considered a small area. Most recently, Song and Kim (2017) gave a Bayesian analysis of two unrelated questions with rare outcomes (i.e., Poisson modeling rather Binomial modeling). Bayesian methods, with useful prior information, deserve much more attention because it is easy to obtain proper estimates. Of course, hierarchical Bayesian models can be used to study data arising from sample surveys with randomized responses.

When it comes to the case of the multiple sensitive questions, lots of work focusing on estimating the correlations, see Bellhouse (1995). Edgell *et al.* (1986) provides a further statistical efficiency study about the correlation. More recent paper of Chung *et al.* (2018) interest in making causal inference among the sensitive attributes through the Bayesian RRT which compare with the multivariate analysis extension of RRT of Kwan (2010). However, we are more interest in how the correlation between the sensitive questions will influence the proportion estimation.

The main purpose of this paper serves as an multi-question extension to the work of Nandram and Yu (2018). The plan of the rest of the paper is as follows. In Section 2, we propose an hierarchical bayesian model for unrelated question design of two sensitive questions and the computational methodology. In Section 3, we provided a simulation study in comparison of three different models and also in area size and correlation effect. In Section 4, we present an application on the TNHNES data to evaluate our Bayesian model.

2. Unrelated question design for multi sensitive questions

In this section, we discuss the Bayesian methodology to analyse the binary response data from combined sensitive questions. If there are more than one sensitive question from small areas. For each area, at least two groups of people are sampled and required to flip an unfair coin. Depending on different results (head or tail), which is unobserved by the instructor, they will provide their true answer of the combined questions either sensitive or non-sensitive. So with the probability p_{ij} (the success probability of jth group from ith area), the respondents will get the chance to answer two sensitive questions; otherwise they should answer the other two non-sensitive questions. In the end, the instructor will collect the binary response data of four types: (No, No), (No, Yes), (Yes, No), (Yes, Yes), where

$\underset{\sim}{y}_{ij} = (y_{ij1}, y_{ij2}, y_{ij3}, y_{ij4})$ represent the counts, where $i = 1..l, j = 1..g_i$.

Since the survey conductor cannot know the data from which branch, the respondent will feel more comfortable to give the true response to the sensitive questions which will lead to a more accurate estimation of the sensitive population proportion $\underset{\sim}{\pi}_{i1}$.

We may think of a straightforward probability relationship below:

$$\frac{y_{ij1}}{n_{ij}} = p_{ij}\pi_{i11} + (1 - p_{ij})\pi_{i21}$$
$$\frac{y_{ij2}}{n_{ij}} = p_{ij}\pi_{i12} + (1 - p_{ij})\pi_{i22}$$
$$\frac{y_{ij3}}{n_{ij}} = p_{ij}\pi_{i13} + (1 - p_{ij})\pi_{i23}$$
$$\frac{y_{ij4}}{n_{ij}} = p_{ij}\pi_{i14} + (1 - p_{ij})\pi_{i24}$$

,where $n_{ij} = \sum_{k=1}^{4} y_{ijk}$.

In order to estimate the population proportion for each area, instead of combining the equation system which may not guarantee a solution, we propose a three-stage bayesian model with latent variables. It is natural to think of the counts data from the four type of responses for each area follows multinomial distribution with the four cell probabilities given above. Based on that, we developed the hierarchical bayesian model to solve the problem.

2.1 Gibbs Sampler

Step 1. Draw ω_{ijk} from Binomial distribution.

$$\omega_{ijk} \mid \underset{\sim}{\pi}_{i1}, \underset{\sim}{\pi}_{i2}, \underset{\sim}{y} \overset{ind}{\sim} \text{Binomial}\{y_{ijk}, \frac{p_{ij}\pi_{i1k}}{p_{ij}\pi_{i1k} + (1 - p_{ij})\pi_{i2k}}\},$$
$$i = 1, \ldots, l; j = 1, \ldots, g_i; k = 1, 2, 3, 4.$$

Step 2. Draw $\underset{\sim}{\pi}_{i1}, \underset{\sim}{\pi}_{i2}$ from the Dirichlet distribution.

$$\underset{\sim}{\pi}_{i1} \mid \underset{\sim}{\mu}_1, \underset{\sim}{\mu}_2, \tau, \underset{\sim}{\omega}, \underset{\sim}{y} \overset{ind}{\sim} \text{Dirichlet}(\omega_{i1\cdot} + \mu_{11}\tau, \omega_{i2\cdot} + \mu_{12}\tau, \omega_{i3\cdot} + \mu_{13}\tau, \omega_{i4\cdot} + (1 - \sum_{k=1}^{3} \mu_{1k})\tau)$$
$$\underset{\sim}{\pi}_{i2} \mid \underset{\sim}{\mu}_1, \underset{\sim}{\mu}_2, \tau, \underset{\sim}{\omega}, \underset{\sim}{y} \overset{ind}{\sim} \text{Dirichlet}(y_{i1\cdot} - \omega_{i1\cdot} + \mu_{21}\tau, y_{i2\cdot} - \omega_{i2\cdot} + \mu_{22}\tau, y_{i3\cdot} - \omega_{i3\cdot} + \mu_{23}\tau,$$
$$y_{i4\cdot} - \omega_{i4\cdot} + (1 - \sum_{k=1}^{3} \mu_{2k})\tau)$$

where $\omega_{ik\cdot} = \sum_{j=1}^{g_i} \omega_{ijk}, y_{ik\cdot} = \sum_{j=1}^{g_i} y_{ijk}$.

The first two steps are easy to run because the conditional posterior densities are all in simple forms. However, drawing $(\underset{\sim}{\mu}_1, \underset{\sim}{\mu}_2, \tau)$ jointly require more work using the grid method from the joint distribution of $(\underset{\sim}{\mu}_1, \underset{\sim}{\mu}_2, \tau)$ after integrating out $\underset{\sim}{\pi}_{i1}, \underset{\sim}{\pi}_{i2}$.

Step 3. Draw $\underset{\sim}{\mu_1}, \underset{\sim}{\mu_2}, \tau$ using grid method.

$$\pi(\underset{\sim}{\mu_1}, \underset{\sim}{\mu_2}, \tau \mid \underset{\sim}{\omega}, \underset{\sim}{y}) \propto$$
$$\prod_{i=1}^{l} \left(\frac{D(\omega_{i1\cdot} + \mu_{11}\tau, \omega_{i2\cdot} + \mu_{12}\tau, \omega_{i3\cdot} + \mu_{13}\tau, \omega_{i4\cdot} + (1 - \sum_{k=1}^{3} \mu_{1k})\tau)}{D(\mu_{11}\tau, \mu_{12}\tau, \mu_{13}\tau, (1 - \sum_{k=1}^{3} \mu_{1k})\tau)} \cdot \right.$$
$$\left. \frac{D(y_{i1\cdot} - \omega_{i1\cdot} + \mu_{21}\tau, y_{i2\cdot} - \omega_{i2\cdot} + \mu_{22}\tau, y_{i3\cdot} - \omega_{i3\cdot} + \mu_{23}\tau, y_{i4\cdot} - \omega_{i4\cdot} + (1 - \sum_{k=1}^{3} \mu_{2k})\tau)}{D(\mu_{21}\tau, \mu_{22}\tau, \mu_{23}\tau, (1 - \sum_{k=1}^{3} \mu_{2k})\tau)} \right)$$
$$\cdot \frac{1}{(1+\tau)^2}$$

Then we can draw $\boldsymbol{\omega}, \boldsymbol{\pi}_{i1}, \boldsymbol{\pi}_{i2}$, and (μ_1, μ_2, τ) successively from the conditional density. We run 11000 iterates and burn in the first 2000 iterates, taking every 9th and get the final 1000 gibbs sampler. We can obtain Rao-Blackwellized estimators of $\boldsymbol{\pi}_{i1}$ and $\boldsymbol{\pi}_{i2}$ because

$$\pi(\boldsymbol{\pi}_{i1}, \boldsymbol{\pi}_{i2} \mid \boldsymbol{y}) = \sum_{\boldsymbol{\omega_{i1}=0}}^{\boldsymbol{y_{i1}}} \cdots \sum_{\boldsymbol{\omega_{ig_i}=0}}^{\boldsymbol{y_{ig_i}}} \pi(\boldsymbol{\pi}_{i1}, \boldsymbol{\pi}_{i2} \mid \boldsymbol{\omega}, \boldsymbol{y}) \pi(\boldsymbol{\omega} \mid \boldsymbol{y})$$

$$= \sum_{\boldsymbol{\omega_{i1}=0}}^{\boldsymbol{y_{i1}}} \cdots \sum_{\boldsymbol{\omega_{ig_i}=0}}^{\boldsymbol{y_{ig_i}}} \pi(\boldsymbol{\pi}_{i1} \mid \boldsymbol{\omega}, \boldsymbol{y}) \pi(\boldsymbol{\pi}_{i2} \mid \boldsymbol{\omega}, \boldsymbol{y}) \pi(\boldsymbol{\omega} \mid \boldsymbol{y})$$

as follows. Let $\boldsymbol{\omega}^{(h)} = (\omega_{ijk}^{(h)}), j = 1, \ldots, g_i, k = 1, \ldots, 4, h = 1, \ldots, M$, denote a random sample of size M from the posterior density, $\pi(\boldsymbol{\omega} \mid \boldsymbol{y})$, obtained from the Gibbs sampler. Then, the Rao-Blackwellized density estimator of $\pi(\boldsymbol{\pi}_{i1}, \boldsymbol{\pi}_{i2} \mid \boldsymbol{y})$ is

$$\pi(\widehat{\boldsymbol{\pi}_{i1}, \boldsymbol{\pi}_{i2}} \mid \boldsymbol{y}) = \frac{1}{M} \sum_{h=1}^{M} \pi(\boldsymbol{\pi}_{i1} \mid \boldsymbol{\omega}^{(h)}, \boldsymbol{y}) \pi(\boldsymbol{\pi}_{i2} \mid \boldsymbol{\omega}^{(h)}, \boldsymbol{y}), i = 1, \ldots, \ell.$$

Then we can get the Rao-Blackawellized estimator of the sensitive proportion $\phi_{i11} = \pi_{i13} + \pi_{i14}$ and $\phi_{i12} = \pi_{i12} + \pi_{i14}, i = 1 \ldots \ell$ for each area.

Next we are able to do the prediction in a finite population under simple random sampling. Assume that each sample unit is drawn from an finite population of size N, let X_s denote the total counts of yeses from s^{th} sensitive question. Therefore,

$$X_s \mid \pi_s \overset{ind}{\sim} \text{Binomial}(N, \pi_s), s = 1, 2.$$

Then, the finite population proportion $P_s = X_s/N, s = 1, 2$, and inference about the P_s can be made in a straightforward manner under the Bayesian model.

3. Simulation Study

In section 3.1, we preform a simulation study to assess the the performance of the combined area model compared with individual-area model and separate question model. In section 3.2, we adjust the parameter setting to increase the area size and the correlation within the sensitive and non-sensitive questions to see the possible gains in estimation accuracy.

3.1 Comparison of three models

In this section, we are going to test our model by using a 10 area simulated data. Based on the 1000 simulated runs, we prepare a comparison of our combined model with separate question model and individual area model. Also, we try to explore whether the correlation between the two questions will gain in estimation accuracy.

Assuming that the probability of answering ”Yes” to the first and second sensitive questions are ϕ_{11} and ϕ_{12}; the probability of answering ”Yes” to the first and second sensitive questions are ϕ_{21} and ϕ_{22}. Then we simulated the correlated response data with correlation ρ_1= 0.2 and ρ_2= 0.25 with respect to the sensitive and nonsensitive questions correspondingly among 10 areas. In other words, the four type of response (”No”, ”No”), (”No”, ”Yes”),(”Yes”, ”No”),(”Yes”, ”Yes”) for the sensitive questions are generated with the probability $(\pi_{11}, \pi_{12}, \pi_{13}, \pi_{14})$, where

$$\begin{aligned}
\pi_{11} &= (1-\phi_{11})(1-\phi_{12}) + dum \\
\pi_{12} &= (1-\phi_{11})\phi_{12} - dum \\
\pi_{13} &= \phi_{11}(1-\phi_{12}) - dum \\
\pi_{14} &= \phi_{11}\phi_{12} + dum
\end{aligned}$$

where $dum = \rho_1\sqrt{\phi_{11}(1-\phi_{11})\phi_{12}(1-\phi_{12})}$ with true value set as $\phi_{11} = 0.25, \phi_{11} = 0.35$. The correlated probability construction come from Yu, Bhadra, Nandram (2017). And we can generate the response from correlated nonsensitive questions in the same way with probability $(\pi_{21}, \pi_{22}, \pi_{23}, \pi_{24})$, with the true value set as $\phi_{21} = 0.35, \phi_{22} = 0.45$. Now we want to mock the sampling process. For each individual from i^{th} area and j^{th} group ($i = 1..\ell = 10, j = 1..g_i \geq 2$). At first, we generate the number of groups uniformly from 2 to 5. For each coming individual from i^{th} area and j^{th} group with size n_{g_i}= (20, 25, 30, 35, 40), we choose an random mechanism with probability p_{ij}= (.25, .75, .2, .7, .3) to answer the sensitive questions and (1-p_{ij}) to answer the nonsensitive questions. Following the simulation process, we are able to collect the combined binary response data sourcing from both sensitive and nonsensitive questions, without knowing which exact question the respondent answer. Our interest is to find the finite population estimation of the probability of answering ”Yes” to the sensitive questions $\phi_{11} = \pi_{13} + \pi_{14}$ and $\phi_{12} = \pi_{12} + \pi_{14}$. Then we can fit the three stage bayesian model to get the Rao-Blackwellized estimates of $(\pi_{11}, \pi_{12}, \pi_{13}, \pi_{14})$ first, and finally get the corresponding finite population estimation. To compare with the individual area model and separate question model, we can construct a 95% HPD interval from the 1000 simulations of the estimated probability $\widehat{\phi_{11}}$ and $\widehat{\phi_{12}}$ from 10 areas. In Figure 1, we can observe that within each area, the width of the 95% HPD interval increases from the combined-area model to the separate question model, and to the individual area model. And the case are the same for both sensitive proportions across 10 areas. Also in Table 1, for the case of area size 10,

3.2 Discussion of the area size and correlation effect

To further compare the three models and the the area size effect. We change the area size to 25. We calculated the relative absolute bias, $RAB = (PM - T)/T$, and the posterior root mean squared error, $PRMSE = \sqrt{(PM-T)^2 + PSD^2}$, where T denotes the true proportions, ϕ_{11} or ϕ_{12} (known by simulation). We also computed their average width (Wid) of the 95% HPD intervals and the coverage C, which is the proportion of intervals containing the true in the 1000 simulated runs. We simulated the correlated data with correlation $\rho_1 = 0.2$ and $\rho_2 = 0.25$ with respect to the sensitive and nonsensitive questions correspondingly among 10 areas.

In Table 1, we provided the simulation results from different area sizes ($\ell = 10, 25$). And we observed that for the combined bayesian model (cb), the average ralative absolute bias and the root mean squared error get smaller as the area size increase from 10 to 25 for both π_1 and π_2; the average width of 95% HPD interval is 0.216, shorter than 0.283 for π_1. Even though we obtained a smaller coverage of 0.951 when the area size equal to 25, it's still greater than the expected 95%. Conclusions are similar for π_2.

In Table 2, we showed the comparison results from different correlations 0.2 and 0.8 when area size is fixed to be 10. For the estimation of π_1 from the combined model (cb), we observed that the relative absolute bias (RAB) is 0.096 for the highly correlated data ($\rho = 0.8$), which is a little bit smaller than 0.101 of the less correlated data ($\rho = 0.2$). The PRMSE is .088, which is also smaller than 0.093 for the less correlated one. However, the change is not quite significant. Even though we benefit from combining questions together, there still need future discussion for the correlation effect based on this simulation result.

4. Application on NHANES III data

In this chapter, we apply the body mass index (BMI) and bone mineral density (BMD) data from the third National Health and Nutrition Examination Survey (NHANES III). The survey is a program of studies designed runned by CDC (Center of Disease Control and Prevention) to assess the health and nutritional status of adults and children in the United States, which is conducted during the period October 1988 through September 1994. Due to confidentiality reasons, the final data set for this study uses only 6557 samples of the 35 largest counties with a population at least 500,000. Also, This survey contains BMI and BMD data together with covariates of age, sex and race.

We have 4 levels of BMI (=1,2,3,4) and 3 levels of BMD (=1,2,3), where BMI= 3&4 represents the BMI value greater than 25 which can be considered as overweight; and BMD= 2&3 means the BMD value smaller than 0.82 which indicates osteopenia or osteoporosis. In our case, the interest is finding out the proportion of people from the targeted population who are overweight or have osteoporosis, which both can be considered as sensitive. Instead of asking questions directly, we are thinking of applying randomized response technique to obtain the sensitive proportion, even though, the designed survey is not implemented in reality. Thus this example serves more like an evaluation of our bayesian unrelated question model. As compared to our survey design, the two sensitive questions would be 'Are you overweight ?' and 'Do you have osteopenia or osteoporosis ?'. Thus to simulate our design we need to construct other two unrelated non-sensitive questions. We set the race and sex to be the non-sensitive question as 'Are you white? ' and 'Are you male?', which are less sensitive.

In Figure 2, we provide an comparison plot of the direct estimates and the Bayesian estimates for 35 areas. The direct estimates are calculated directly from dividing the sensitive response counts by the sample size within each area. We treat direct estimates as a close value to the true proportion since it obtained based on all the samples from each area. We can observe that the Bayesian estimates almost follow the trend of the direct estimates showing that we are able to get a reasonable estimation for the overweight proportion through the Bayesian random response model while applying the randomized response technique.

5. Concluding Remarks

We provided a bayesian method to solve for the sensitive proportions though the unrelated question design when we have more than one sensitive questions. The simulated results shown that for data from small areas, the Bayesian combined model (cb) will give a more accurate estimation, in both sense of relative absolute bias and posterior root mean square, compared to the individual area model (ind) and separate question model. It is also examined by simulate the process on the TNHNES data.

This might give us a clue of how to design the survey in the situation that more than one sensitive question are available, and also they are likely to be correlated. It could be the case that even if we are only interested in one sensitive question, we can include other correlated sensitive questions into the design to get a better proportion estimation. Of course, the same number of unrelated questions should be constructed since the answer should have the same dimension. Even though there might be concern of the extra cost by asking more questions, the availability of online survey tools will make the survey collector get the response more easily.

The design can be easily generalized to multiple questions case. If there are k sensitive questions, that means we have k dimension of cell probabilities for both sensitive and nonsenstive questions from each area. And also there will be 2^k latent variables for each group within areas. Even in our case when $k = 2$, we need to run at least 11000 iterates to pass the Geweke test and other convergence diagnostics. As k get larger, the model will be more complicated and the mixing effect will also grow when we run the gibbs sampler, which means more iterates is needed to get the useful draws in the end.

For the above concerns, we try to construct a normal approximation model to increase the computation efficiency in the future.

REFERENCES

Blair, G., Imai, K. and Zhou, Y-Y. (2015), "Design and Analysis of the Randomized Response Technique," *Journal of the American Statistical Association, Review*, 110, 1304-1319.

Bellhouse, D. R. (1995), "Estimation of Correlation in Randomized Response," *Survey Methodology*, 21, 1, 13-19.

Chung, Ray S. W., Chu, Amanda M. Y. and So, Mike K. P. (1995), "Bayesian randomized response technique with multiple sensitive attributes: The case of information systems resource misuse," *The Annals of Applied Statistics*, 12, 3, 1969-1992.

Eriksson, S. A. (1973), "A New Model for Randomized Response," *International Statistical Review*, 41. 101-113.

Edgell, S. E., Himmelfarb, S., and Cira, D. J. (1986), "Statistical efficiency of using two quantitative randomized response techniques to estimate correlation," *Psychological Bulletin*, 100, 2, 251-256.

Fox, J. A. and Tracy, P. E. (1986), *Randomized Response: A Method for Sensitive Surveys*, Sage: London.

Greenberg, B. G., Abul-Ela, A.-L. A., Simmons, W. R. and Horvitz, D. G. (1969), "The Unrelated Question Randomized Response Model: Theoretical Framework," *Journal of the American Statistical Association*, 64, 520-539.

Greenberg, B. G., Kuebler, R. R., Abernathy, J. R. and Horvitz, D. G. (1971) "Application of the Randomized Response Technique in Obtaining Quantitative Data," *Journal of the American Statistical Association* 66, 243-250.

Gupta, S., Gupta, B. and Singh, S. (2002), "Estimation of Sensitivity Level of Personal Interview Survey Questions," *Journal of Statistical Planning and Inference*, 100, 239-247.

Gupta, S., Javid, S. and Supriti, S. (2010), "Mean and Sensitivity Estimation in Optional Randomized Response Models," *Journal of Statistical Planning and Inference*, 140, 2870-2874.

Kim, J. M. and Warde, W. D. (2004), "A Stratified Warner Randomized Response Model," *Journal of Statistical Planning and Inference*, 120, 155-165.

Kwan, S. S. K., So, M. K. P. and Tam, K. Y. (2010), "Research note - Applying the randomized response technique to elicit truthful responses to sensitive questions in IS research: The case of software piracy behavior," *Information Systems Research*, 21, 941-959.

Nandram, B. and Choi, J. W. (2002), "Hierarchical Bayesian Nonresponse Models for Binary Data from small areas with Uncertainty about Ignorability," *Journal of the American Statistical Association*, 97, 381-388.

Nandram, B. and Choi, J. W. (2002), "A Bayesian Analysis of a Proportion under Nonignorable Nonresponse," *Statistics in Medicine*, 21, 1189-1212.

Nandram, B. and Yu, Y. (2017), "Bayesian Analysis of Sparse Counts Under the Unrelated Question Design," In *JSM Proc., Survey Res. Meth. Sec.* American Statistical Association, Alexandria, VA, pp. 11621175.

Nandram, B. and Yu, Y. (2018), "Bayesian Analysis of a Sensitive Proportion for a Small Area," *International Statistical Review*, 0, 1-17.

Perri, P. F. (2008), "Modified Randomized Devices for Simmons' Model," *Model Assisted Statistics and Applications*, 3, 233-239.

Oh, M. (1994), "Bayesian Analysis of Randomized Response Models: A Gibbs Sampling Approach," *Journal of the Korean Statistical Society*, 23, 463-482.

Odumade, O. and Singh, S. (2009), "Efficient Use of Two Deck of Cards in Randomized Response Sampling," *Communications in Statistics - Theory and Methods*, 38, 439-446.

O'Hagan, A. (1987), "Bayes Linear Estimators for Randomized Response Models," *Journal of the American Statistical Association*, 82, 580-585.

Tan, M. T., Tian, G.-L. and Tang, M-L. (2009), "Sample Surveys with Sensitive Questions: A Randomized Response Approach," *The American Statistician*, 63, 9-16.

Tian, G.-L., Yuen, K. C., Tang,M.-L. and Tan, M. T. (2009), "Bayesian Non-randomized Response Models for Surveys with Sensitive Questions," *Statistics and Its Interface*, 2, 13-25.

Tourangeau, R., Rips, L. J. and Rasinski, K. (2000), *The Psychology of Survey Response*, Cambridge University Press: Cambridge, England.

Tourangeau, R. and Yan, T. (2007), "Sensitive Questions in Surveys," *Psychological Bulletin*, 133, 859-883.

Warner, S. L. (1965), "Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias," *Journal of the American Statistical Association*, 60, 63-69.

Winkler, R. L., and Franklin, L. A. (1979), "Warner's Randomized Response Model: A Bayesian Approach," *Journal of the American Statistical Association*, 74, 207-214.

Yu, Y., Bhadra, D., Nandram, B. (2017), "Tests of Independence for a Two-by-two Contingency Table with Random Margins," *International Journal of Statistics and Probability*, 6, 2, 106.

Figure 1: Comparison of the direct estimates and the Bayesian estimates for 35 counties

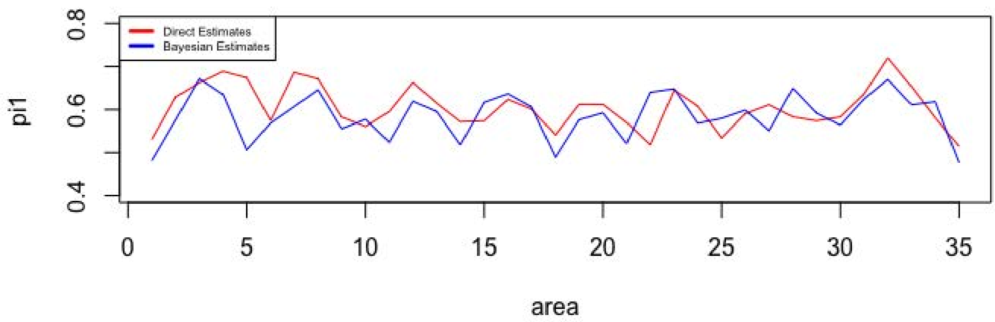

Table 1: Relative absolute bias, posterior root mean squared error, coverage of 95% credible intervals and width of 95% credible interval averaged over the 1000 runs and different area sizes (ℓ=10, 25) for combined model (cb), individual area model (ind) and separate question model (sep).

			$\widehat{\phi_{11}}$				$\widehat{\phi_{12}}$		
ℓ	$Mol.$	RAB	PRMSE	C	Wid	RAB	PRMSE	C	Wid
10	cb	0.101	0.093	0.971	0.283	0.101	0.094	0.972	0.283
	Ind	0.188	0.168	0.950	0.490	0.187	0.167	0.951	0.490
	Sep	0.181	0.142	0.896	0.375	0.179	0.141	0.899	0.374
25	cb	0.087	0.074	0.951	0.216	0.087	0.074	0.952	0.216
	Ind	0.188	0.168	0.950	0.491	0.186	0.167	0.952	0.491
	Sep	0.175	0.138	0.910	0.403	0.176	0.166	0.913	0.401

Table 2: Relative absolute bias, posterior root mean squared error, coverage of 95% credible intervals and width of 95% credible interval averaged over the 1000 runs and two different levels of correlations $(\rho_1, \rho_2) = (0.2, 0.25)$ and $(0.8, 0.8)$.

		π_1				π_2			
ρ_1, ρ_2	$Mol.$	RAB	PRMSE	C	Wid	RAB	PRMSE	C	Wid
0.2, 0.25	cb	0.101	0.093	0.971	0.283	0.101	0.094	0.972	0.283
	Ind	0.188	0.168	0.950	0.490	0.187	0.167	0.951	0.490
	Sep	0.181	0.142	0.896	0.375	0.179	0.141	0.899	0.374
0.8, 0.8	cb	0.096	0.088	0.956	0.267	0.097	0.091	0.957	0.267
	Ind	0.199	0.173	0.940	0.499	0.198	0.173	0.941	0.499
	Sep	0.173	0.136	0.901	0.363	0.173	0.137	0.899	0.363