Bayesian Analysis of Sparse Counts Under the Unrelated Question Design

Balgobin Nandram* Yuan Yu†

Abstract

In sample surveys with sensitive items, sampled units may not respond or they respond untruthfully. Usually a negative answer is given when it is actually positive, thereby leading to an estimate of the population proportion of positives (sensitive proportion) that is too small. In our study, we have binary data obtained from the unrelated-question design, and the sensitive proportion is of interest. A respondent answers the sensitive item with a known probability, and to avoid non-identifiable parameters, at least two different random mechanisms are used, but only one for each cluster of respondents. The key point here is that the counts are sparse (very small sample sizes), and we show how to overcome some of the problems associated with the unrelated question design. We have presented an example with sparse data on college cheating and a simulation study to illustrate the properties of our procedure. Finally, we discuss two extensions to accommodate finite population sampling and optional responses.

Key Words: Latent variables, Data augmentation, Gibbs sampler, Non-identifiable parameters, Proportion, Rao-Blackwellized estimates.

1. Introduction

When people are asked sensitive (stigmatizing) questions, there is a tendency for them not to respond or to tell lies if they do. One way to reduce these effects is to use the technique of randomized response. In this approach to survey sampling, the randomization is not only in drawing the sample but also in obtaining the response, and there is an enormous literature. One possible design (Greenberg, Abu-Ela, Simmons and Horvitz 1969) is to ask an unrelated nonsensitive (innocuous) question in addition to the sensitive question. The respondents are asked to give a honest answer to one of the two questions selected according to a random mechanism, the essential features of the random mechanism being known to the investigator. This is an extension of the mirrored question design (Warner 1965) in which the respondents are asked the opposite question instead of the unrelated question. When randomized response techniques are used, a respondent's individual answer is not of interest, rather inference is needed for the population. 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.

For example, one tosses a die and if one or six comes up, the respondent must give a honest answer to the sensitive question, and if two, three, four or five comes up, the respondent must give a honest answer to the nonsensitive question. 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. We do not rule out the situations where both questions might be sensitive; one of them may be much less sensitive (most respondents do not care). For example, two possible questions are stated below. Students at a university are asked to circle the true answer to the question selected.

*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 1: Have you ever cheated on an examination anywhere?
Question 2: Do you spend more than twenty hours per week on
all courses studying outside the classroom?

Circle your response. [Yes, No]

Here Question 1 is sensitive and Question 2 may be sensitive to some respondents, but much less sensitive than Question 1. We need inference of the proportions of students answering 'yes' to respectively Question 1 and Question 2, the first being the sensitive proportion of greater interest.

Warner (1965) proposed the randomized response method as a survey technique to reduce potential bias due to nonresponse and social desirability when asking questions about sensitive behaviors and beliefs. The method asks respondents to use a randomization device, such as a coin flip, whose outcome is unobserved by the interviewer. By introducing random noise, the randomized response method conceals individual responses and protects respondent privacy. As a result, respondents may be more inclined to answer truthfully.

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. We will call the model for the unrelated question design the individual-area model, and in this paper, we are mainly concerned with this model.

The plan of the rest of the paper is as follows. In Section 2, we present a review of the literature of the unrelated question design. In Section 3, we show some difficulties associated with the analysis of the unrelated question design. In Section 4, we present the Bayesian methodology for individual-area model. In Section 5, we discuss empirical studies, where we use an illustrative example on college cheating, and we describe a small simulation study to show the frequentist properties of the individual-area model. In Section 6, we present concluding remarks and a discussion of two possible extensions.

2. Unrelated Question Design and Extensions

Blair, Imai, and Zhou (2015) gave an excellent review paper on randomized response techniques (RRT). They classified many of the designs into four types: mirrored question, forced response, disguised response, and unrelated question. For each type, they provide a brief explanation, an example, and a discussion about identification. All four designs make two general assumptions: (1) the randomization distribution is fully known to researchers, and (2) respondents comply with the instructions and answer the sensitive question truthfully.

In the mirrored question design (Warner 1965), a respondent is asked to perform a Bernoulli trial. If a success occurs, the respondent is asked to answer the sensitive question, otherwise the respondent is asked to answer the opposite of the sensitive question. In the unrelated question design (Greenberg, Abul-Ela, Simmons and Horvitz 1969), a Bernoulli trial is performed, and if it is a success, the respondent is asked to answer the sensitive question, otherwise the respondent is asked to answer the unrelated question. The forced response design (Boruch 1971, Fox and Tracy 1986), is like the unrelated question design, but there is no unrelated question. Two Bernoulli trials are performed. In the first trial, if a success occurs, the respondent is asked to answer the sensitive question ('yes' or 'no'); otherwise a different Bernoulli trial is performed. If it is a success, the respondent must answer 'yes'; otherwise the respondent must answer 'no'. The unrelated question design is a natural extension of the mirrored question design. See Kuk (1990) for the disguised design.

Researchers have worked on many extensions of Warner's randomized response techniques (RRT). Greenberg et al. (1969), Folsom et al. (1973), Christofides (2005), Odumade and Singh (2009), Mangat (1992), Perri (2008), Mahmood, Singh and Horn (1998), Kim and Warde (2004) are some of them. The interested readers may refer to Chaudhuri and Mukerjee (1988), Fox and Tracy (1986) and Tracy and Mangat (1996) and more recently Chaudhuri (2011) and Chaudhuri and Christofides (2013) for a comprehensive discussion on RRT. Finally, see Johnson, Sedory and Singh (2016).

It is important to use an optimal design. For the unrelated question design Greenberg, Abul-Ela, Simons and Horvitz (1969) used a heuristic argument to suggest a choice of p_1 as $.2 \pm .1$ or $.8 \pm .1$ and $p_2 = 1 - p_1$. Moors (1971) provided a more systematic study of optimality and showed that $p_2 = 0$. That is, the randomized experiment should be performed on a sample of n_1 individuals and a sample of n_2 individuals should only be asked the nonsensitive question and none of these should be asked the sensitive question; see also Lanke (1975). Of course, the unrelated question design is more efficient than the mirrored design. There are further increases in efficiency with mild addition of response burden. Mangat, Singh and Singh (1992), Mangat and Singh (1990), Mangat (1992) and Mangat (1994) introduced the two-stage designs. One design is an extension of the mirrored question design and another is an extension of the unrelated question design. For example, two different Bernoulli trials are performed. If the first Bernoulli trial is a success, the respondent answers the sensitive question. If the first trial is a failure, the respondent performs the second Bernoulli trial. If the second Bernoulli trial is a success, the respondents answer the sensitive question; otherwise the respondent answers the unrelated or opposite question depending on whether the mirrored question design or the unrelated question design is used. The two-stage design with the unrelated question is more efficient than the corresponding one for the mirrored question design and each is more efficient than the corresponding one-stage design (Mangat, Singh and Singh 1992, Mangat and Singh 1990, Mangat 1992, Mangat 1994).

We also mention nonrandomized designs that do not use a random mechanism as in randomized design. The crosswise and triangular designs (e.g., Tan, Tian and Tang 2009, Tian, Yuen, Tang and Tan 2009) can be viewed as extensions of the unrelated question design because each of them has a sensitive (stigmatizing) question and an unrelated nonsensitive question. Let X be a binary sensitive variable and W denote a nonsensitive variable $X = 0$ or $W = 0$ are 'nos' and $X = 1$ or $W = 1$ are 'yeses'. In the crosswise design, a respondent is asked to answer 'yes' to $X = 1, W = 1$ or $X = 0, W = 0$ and 'no' otherwise. In the triangular design, a respondent is asked to answer 'no' to $X = 0, W = 0$ and 'yes' to $X = 0, W = 1$ or $X = 1, W = 1$ or $X = 1, W = 1$. As in the unrelated question design, if both probabilities are unknown, two samples are needed. There are similar difficulties for estimation and multiple answers are proposed (e.g., Groenitz 2017); the key gain is a random mechanism is not needed.

Greenberg, Kuebler, Abernathy and Horvitz (1971) and Eriksson (1973) extended the unrelated question model of Greenberg, Abul-Ela, Simmons and Horvitz (1969) to the case in which the response is quantitative.

Prior information about the unknown parameters is sometimes obtainable and can be used along with the sample information for estimation of these unknown parameters. This is the Bayesian approach of estimation. There are not many works within the Bayesian paradigm of randomized response models. Nonetheless, attempts have been made on the Bayesian analysis of randomized response techniques. 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. Also, Tian, Yuen, Tang and Tan (2009) proposed Bayesian approaches to non-randomized response models without using randomized mechanisms; non-randomized designs (crosswise and triangular) may be more efficient than their randomized response counterparts; see, for example, Tan, Tian and Tang (2009). Most recently, Song and Kim (2017) gave a Bayesian analysis of two rare unrelated questions (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; MLEs are difficult to obtain.

Finally, we note that in 2017, there was a special issue on randomized response techniques in the journal of Statistics and Applications in honor of Warner. Many of these fifteen papers cover extensions of the mirrored question design and the unrelated question design. Randomized designs for both qualitative and quantitative data were discussed. There are extensions of the unrelated question design to optional response (e.g., Dass and Chhabra, 2017) and to inverse unrelated question design used for additional privacy protection (Dihidar and Basu, 2017). But most of the papers are on design-based analyzes, generally true in randomized response analyzes.

3. Difficulties of the Unrelated Question Design

We discuss some difficulties arising in the analysis of the unrelated question design that has two unknown parameters. We consider the situation where two Bernoulli trials are performed with success probabilities p_1 and p_2. Both proportions, π_1 and π_2, of people with the sensitive character and the nonsensitive character respectively are of interest.

We consider the first problem of the design-based estimator. The standard model is

$$y_s \overset{ind}{\sim} \text{Binomial}\{n_s, p_s\pi_1 + (1-p_s)\pi_2\}, s = 1, 2.$$

Let $a_s = y_s/n_s, s = 1, 2,$ be the MLE of $p_s\pi_1 + (1-p_s)\pi_2$. Then, one can find the MLEs of π_1 and π_2 by solving the two equations, $a_s = p_s\hat{\pi}_1 + (1-p_s)\hat{\pi}_2, s = 1, 2,$ where $\hat{\pi}_1$ and $\hat{\pi}_2$ are respectively MLEs of π_1 and π_2 provided that $\hat{\pi}_1$ and $\hat{\pi}_2$ lie in $(0, 1)$ (may not happen). It can be shown that for $p_1 < p_2$ that $\hat{\pi}_1$ and $\hat{\pi}_2$ lie in $(0, 1)$ provided that

$$\frac{1-p_2}{1-p_1} < \min\{\frac{a_2}{a_1}, \frac{1-a_2}{1-a_1}\}, \quad \max\{\frac{a_2}{a_1}, \frac{1-a_2}{1-a_1}\} < \frac{p_2}{p_1}.$$

This is a tight condition especially for small sample sizes that are of interest here. Therefore, $\hat{\pi}_1$ and $\hat{\pi}_2$ may not lie in $(0, 1)$, and this is, indeed, disturbing. The expectation-maximization (EM) algorithm can be used to obtain MLEs, but now there is an underestimation of the standard error. We note that the EM algorithm was applied to randomized response (e.g., Bourke and Moran 1988). Lee, Sedory and Singh (2013) considered the problem of developing minimum sample size requirements for different randomized response models based on guessed values of the parameters of interest and on the randomization device parameters being used in collecting the datasets. They found that very large sample sizes are needed to get an admissible estimate of the required proportion of a sensitive attribute. Of course, the Bayesian method can overcome this difficulty.

The second problem is that $\hat{\pi}_1$ and $\hat{\pi}_2$ can be highly correlated. This is also disturbing because π_1 and π_2 are completely unrelated. It is true that

$$\text{var}(\hat{\pi}_1) = \frac{(1-p_2)^2 v_1 + (1-p_1)^2 v_2}{(p_2-p_1)^2},$$

$$\text{var}(\hat{\pi}_2) = \frac{p_2^2 v_1 + p_1^2 v_2}{(p_2-p_1)^2},$$

$$\text{cov}(\hat{\pi}_1, \hat{\pi}_2) = -\{\frac{p_2(1-p_2)v_1 + p_1(1-p_1)v_2}{(p_2-p_1)^2}\},$$

where

$$v_s = (p_s\pi_1 + (1-p_s)\pi_2)\{p_s(1-\pi_1) + (1-p_s)(1-\pi_2)\}/n_s, s = 1, 2.$$

One way to reduce the correlation is to increase the sample size, but this may be costly and prohibitive (or infeasible). Nevertheless, this is a useful result because it warns us to be cautious in running a Gibbs sampler as there will be the weak mixing phenomenon. Again, the Bayesian method can reduce this correlation, but it cannot be eliminated completely.

The third problem with the design-based estimators is that while they are theoretically unbiased and consistent, in small samples they may be practically biased. Again, one solution is to increase the sample size but this may be costly in many applications.

4. Bayesian Methodology

In this section, we discuss the Bayesian methodology to analyze data from a randomized response application. First, in Section 3.1, we discuss the analysis for a single area with sparse data; this is the individual-area model.

We assume that there is a single area and in each area there are g clusters of individuals, and the respondents within a cluster toss a possible different random mechanism. Because we are studying two items, we need at least two distinct random mechanisms (i.e., at least two clusters, more the better). In this paper, we are not interested in combining data from a number of small areas, but rather we are interested in the individual-area model.

Let p_j denote the probability that the sensitive item is selected for the j^{th} cluster of respondents interviewed, where n_j is the number of respondents in the j^{th} cluster, $j = 1, \ldots, g$. Note that p_j is known for the j^{th} cluster interviewed. Let π_1 and π_2 denote respectively the probabilities of a 'yes' for the sensitive question and the nonsensitive question. Then the probability of getting a 'yes' answer from a respondent in the j^{th} cluster is $p_j\pi_1 + (1-p_j)\pi_2$ and a 'no' answer is $p_j(1-\pi_1) + (1-p_j)(1-\pi_2)$. Under random sampling, letting y_j denote the number of 'yeses' obtained, we have

$$y_j \mid \pi_1, \pi_2 \overset{ind}{\sim} \text{Binomial}\{n_j, p_j\pi_1 + (1-p_j)\pi_2\}, j = 1, \ldots, g, g \geq 2,$$

where the number of different random mechanisms can be g. Then, the joint probability mass function of $\boldsymbol{y} = (y_1, \ldots, y_g)$ is

$$p(\boldsymbol{y} \mid \pi_1, \pi_2) = \prod_{j=1}^{g} \{p_j\pi_1 + (1-p_j)\pi_2\}^{y_j} \{p_j(1-\pi_1) + (1-p_j)(1-\pi_2)\}^{n_j - y_j}. \quad (1)$$

Among the g clusters, we assume that there at least two distinct p_j. While the design-based analysis via maximum likelihood estimation is defective, the Bayesian analysis is attractive here.

It is worth noting that this is a more general model than the one discussed in the literature with just two random mechanisms. By introducing more than two random mechanisms (samples), we can improve the efficiency of the unrelated question design.

For a full Bayesian analysis, we assume a priori $\pi_s \overset{iid}{\sim}$ Uniform$(0,1), s = 1,2$; the posterior density is the same as the likelihood function. However, because of the sparseness of the data, posterior inference may be sensitive to this assumption, and one may need to use a more informative prior (subject to availability). Other possibilities are Jeffreys' prior (not much different from the uniform prior) and Haldane's prior that may cause posterior impropriety with zero counts of 'yeses' or 'nos'.

There are at least two methods to get estimators of π_1 and π_2 within the Bayesian paradigm. First, one can use a grid method to draw π_1 and π_2 separately. This can be done by numerically integrating out one of them, say π_2, and then draw π_1 using a grid. Then, draw π_2 conditional on π_1 using a grid again. Or, to avoid the numerical integration, one can draw from a bivariate grid on (π_1, π_2) (more computer storage and time are required). Second, one can introduce latent variables as a data augmentation and use the Gibbs sampler; e.g., see Oh (1994). It is convenient that this latter method allows implementation of the more efficient Rao-Blackwellized estimators, not discussed in Oh (1994).

Using the latent variables $(z_j, w_j), j = 1, \ldots, g$, the joint posterior density is

$$\pi(\pi_1, \pi_2, \boldsymbol{z}, \boldsymbol{w} \mid \boldsymbol{y}) \propto \prod_{j=1}^{g} \left(\begin{array}{c} y_j \\ z_j \end{array} \right) (p_j \pi_1)^{z_j} ((1-p_j)\pi_2)^{y_j - z_j}$$

$$\times \prod_{j=1}^{g} \left(\begin{array}{c} n_j - y_j \\ w_j \end{array} \right) (p_j(1-\pi_1))^{w_j} ((1-p_j)(1-\pi_2))^{n_j - y_j - w_j}.$$

For any proper priors on π_1 and π_2, the joint posterior density of $\pi_1, \pi_2, \boldsymbol{z}, \boldsymbol{w} \mid \boldsymbol{y}$ is proper because it is proportional to a product of binomial probability mass functions that are all bounded by unity.

The Gibbs sampler is easy to run because the conditional posterior densities are all in simple forms,

$$z_j \mid \pi_1, \pi_2, y_j \overset{ind}{\sim} \text{Binomial}\{y_j, \frac{p_j \pi_1}{p_j \pi_1 + (1-p_j)\pi_2}\}, j = 1, \ldots, g,$$

$$w_j \mid \pi_1, \pi_2, y_j \overset{ind}{\sim} \text{Binomial}\{n_j - y_j, \frac{p_j(1-\pi_1)}{p_j(1-\pi_1) + (1-p_j)(1-\pi_2)}\}, j = 1, \ldots, g.$$

Note that given π_1, π_2, y_j, z_j and w_j are independent. More importantly, letting $y. = \sum_{j=1}^{g} y_j$, $z. = \sum_{j=1}^{g} z_j$, $w. = \sum_{j=1}^{g} w_j$,

$$\pi_1 \mid z., w., y. \sim \text{Beta}(z. + 1, w. + 1),$$

$$\pi_2 \mid z., w., y. \sim \text{Beta}(y. - z. + 1, n. - y. - w. + 1).$$

Again, it is convenient that, given $z.$, $w.$, $y.$, π_1 and π_2 are independent and they are beta random variables. This independence is important because it provides a better mixing Gibbs sampler than if they were correlated. One needs to be careful in running any Gibbs sampler. However, it is unfortunate that the conditional density function of π_1 is a function of the missing data, $z.$, $w.$, but not a function of the observed data, $y.$.

We can obtain Rao-Blackwellized estimators of π_1 and π_2 because

$$\pi(\pi_1, \pi_2 \mid \boldsymbol{y}) = \sum_{z_1=0}^{y_1} \sum_{w_1=0}^{n_1-y_1} \cdots \sum_{z_g=0}^{y_g} \sum_{w_g=0}^{n_g-y_g} \pi(\pi_1, \pi_2 \mid \boldsymbol{z}, \boldsymbol{w}, \boldsymbol{y}) \pi(\boldsymbol{z}, \boldsymbol{w} \mid \boldsymbol{y})$$

$$= \sum_{z_1=0}^{y_1} \sum_{w_1=0}^{n_1-y_1} \cdots \sum_{z_g=0}^{y_g} \sum_{w_g=0}^{n_g-y_g} \pi(\pi_1 \mid \boldsymbol{z}, \boldsymbol{w}, \boldsymbol{y}) \pi(\pi_2 \mid \boldsymbol{z}, \boldsymbol{w}, \boldsymbol{y}) \pi(\boldsymbol{z}, \boldsymbol{w} \mid \boldsymbol{y})$$

as follows. Let $(z_j^{(h)}, w_j^{(h)}), j = 1, \ldots, g, h = 1, \ldots, M,$ denote a random sample of size M from the posterior density, $\pi(\boldsymbol{z}, \boldsymbol{w} \mid \boldsymbol{y})$, obtained from the Gibbs sampler. Then, the Rao-Blackwellized density estimator of $\pi(\pi_1, \pi_2 \mid \boldsymbol{y})$ is

$$\pi(\widehat{\pi_1, \pi_2} \mid \boldsymbol{y}) = \frac{1}{M} \sum_{h=1}^{M} \pi(\pi_1 \mid \boldsymbol{z}^{(h)}, \boldsymbol{w}^{(h)}, \boldsymbol{y}) \pi(\pi_2 \mid \boldsymbol{z}^{(h)}, \boldsymbol{w}^{(h)}, \boldsymbol{y}).$$

The first method does not allow Rao-Blackwellization, and this was not discussed in the Bayesian approach of the analysis of data from the unrelated question design before.

5. Empirical Studies

In Section 5.1, we describe an application on college cheating and, in Section 5.2, we perform a simulation study to assess the performance of the individual-area model.

5.1 Illustrative Example

The data on college cheating were collected by a social science professor at a University. This is an important practical issue because college cheating has become a serious problem in the US (e.g., Shon 2006 and Jones 2011).

The professor asked the students in the class to carry out the experiment and find 20-30 different students on campus to ask these questions. He told them to be certain that this is the first time the respondent is doing this questionnaire. The students in the class were divided into teams of 2-3 to make 11 teams and they were asked to visit various locations (campus center, library, fitness center, food courts, etc.) on campus to carry out the survey.

The students were given specific instructions as follows. “When each of you collect the data, you would need to use at least two different random mechanisms. This has to be done because both the probabilities for a ‘yes’ answer of the sensitive question and the nonsensitive questions are assumed to be unknown. For example, if you use a six-sided die, you may take 1 or 2 for the sensitive question and 3, 4, 5, or 6 for the nonsensitive question for some students and 1,2, 3, or 4 for the sensitive question and 5 or 6 for the non-sensitive question. You must not use the same random mechanism for all the students you interviewed; the example just discussed has two random mechanisms. However, you must record which mechanism you use for a particular student. You would ask each student to toss the die, and answer honestly, ‘yes’ or ‘no’, to the question that turns up, but the student must not tell you which question she/he is answering. When you interview the students, you must ensure that they are giving you independent answers. If you enjoy collecting the data, you may try a third random mechanism.” Note that the nonsensitive question is a bit sensitive and the optimal design was not used. The experiment was performed at 11 locations on campus. However, the data are confidential and cannot be presented. There is a single location with three random mechanisms.

For each location, when the individual-area model is fit using the Gibbs sampler, we ran $20,000$ iterations, used 5000 as a “burn-in” and picked every 15^{th} thereafter. For this sample all the autocorrelations are nonsignificant, effective sample sizes are 1000 for all locations except two of them, which were nearly 800 for π_1 and π_2. The Geweke test of stationarity did not reject stationarity. The computations for the 11 locations took just less than one minute.

In Table 1 and Table 2, we present some results for the individual-area model. The design-based estimators are not very good and some of them are out of range. Note that the Bayesian method provides much more reasonable estimates, standard errors and coefficient of variations although they are large especially for π_1, not so bad for π_2 (CVs are a bit misleading here). Also, we can see that the correlations between π_1 and π_2 are smaller for the Bayesian model.

5.2 Simulation Study

We run a simulation study to assess the performance of the individual-area model. Our objective is modest and we want to study the frequentist properties of the individual-area model.

The design plan is as follows. First, we have used three random mechanisms with $p_1 = .30, p_2 = .60, p_3 = .80$. Second, we obtain the sample sizes by taking $n \overset{ind}{\sim}$ Uniform$(24, 48)$, (i.e., $8 - 16$ for each random mechanism). The numbers of respondents to the three mechanisms are drawn from a multinomial distribution [$(n_1, n_2, n_3) \overset{ind}{\sim}$ Multinomial$(n, (.30, .40, .30)$]. We have taken the true values of $\pi_1 \overset{ind}{\sim}$ Beta$(\mu_1\tau, (1 - \mu_1)\tau)$ and $\pi_2 \overset{ind}{\sim}$ Beta$(\mu_2\tau, (1 - \mu_2)\tau)$, where $\mu_1 = .279$, $\mu_2 = .798$ (posterior means of π_1 and π_2 when all the 11 locations are combined into a single location). We perform a two-factor experiment. The first factor is the number of locations that we select as $\ell = 10, 18, 25, 50, 75, 100$. The second factor is τ that we select as $\tau = 10, 100, 1000$ to have different degrees of closeness of the locations. Of course, if we increase the sample sizes in each location, the individual-area model will perform well, but this is not our interest. When $\tau = 10$, the locations are very different and when $\tau = 1000$, the areas are very similar. In fact, as τ gets larger, the locations get internally more consistent.

We have generated the data as follows. We explain the procedure for the first random mechanism (i.e., $p_1 = .30$) with n_1 respondents. For each respondent at a location, we select the question to be answered with probability p_1. If the first question is selected, we select a “yes” answer with probability π_1, and if the second question is selected, we select a “yes” answer with probability π_2. This procedure is run in the same manner for all respondents in all locations. We have performed 1000 simulated runs.

We fit each of the 1000 simulated runs using the individual-area model in exactly the same manner as described for the data. For each of the 1000 simulated runs, we performed the Geweke test of stationarity, calculated the effective sample sizes and the autocorrelations. We found that the performance of the Gibbs sampler was satisfactory in almost all cases. These computations were done using parallel computing on our Solar Cluster.

We computed the posterior means (PM), posterior standard deviations (PSD), the numerical standard errors (NSE), the 95% HPD intervals and its width, and the correlation between π_1 and π_2 for all locations. Then, at a macro level, we take the averages and standard errors for all quantities over all simulated runs and all locations. Specifically, 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, π_1 or π_2 (known by simulation). With respect to the intervals, we computed their average width (WID) and the coverage C, which is the proportion of intervals containing the true in the 1000 simulated runs. We also computed the average of the correlations between π_1 and π_2. The NSEs are mostly smaller than $.001$.

In Table 3, we present the bias (B) and the posterior standard deviation (PSD) of π_1 and π_2 averaged over all simulation runs and locations, number of locations (ℓ) and τ. There are minor changes in PSDs over ℓ and τ. Also for both models the bias in π_1 is positive and π_2 is negative; this is expected because π_2 are large and π_1 are small. In addition, the biases are virtually the same over the number of locations for the individual-area model. There are also some changes in the biases as τ increases.

In Table 4, we have presented these summaries by ℓ and the individual-area model for π_1 and π_2 for $\tau = 10,\ 100,\ 1000$ respectively. The RAB and PRMSE are quite large. As expected, the model gets better as τ increases. The coverage C is reasonably closed to the nominal value of 95% but Wid is relatively large. However, the changes in these quantities with ℓ are not important.

6. Concluding Remarks

We have shown how to analyze data from the unrelated question design using the individual-area model via a Bayesian method. We have pointed out several problems associated with randomization-based method, and we have shown how the Bayesian methods can overcome or reduce their effects. This is evident in our example on college cheating.

It is unfortunate that we did not find very good frequentist properties of the individual-area model that is used to analyze data from the unrelated question design. However, we have seen some improvement in accuracy and precision as a location becomes more internally consistent. This effect should be similar to the one when the sample size increases. But we have been primarily concerned with sparse data as in our example on college cheating.

It is possible to improve the individual-area estimates by pooling the data from several areas. We are reporting this work in elsewhere.

However, here we consider two extensions of our method to accommodate covariates, to prediction of the finite population proportion, and to optional response. Extension to stratification is trivial, but it is not so trivial for cluster sampling. These extensions have not been discussed in the literature on randomized response under a model-based analysis.

The first extension is how to do prediction in a finite population under simple random sampling. Once our individual-area model is fit, we will obtain π_1 and π_2. Therefore,

$$X_s \mid \pi_1 \overset{ind}{\sim} \text{Binomial}(N, \pi_s), s = 1, 2.$$

Then, the finite population proportion $P_s = X_s/N, s = 1, 2$, and inference about the P_s can be made in a straight forward manner under the Bayesian model. It is worth noting that we need to sample both the sample part and the non-sample part. Of course, we are assuming that there is no selection bias.

Our second extension is to optional responses (e.g., Gupta, Gupta and Singh 2002, Gupta, Javid and Supriti 2010) for quantitative data with the unrelated question design. A coin is tossed with probability heads, $p \neq 1/2$. If the coin comes up tails, the respondent is asked to answer the nonsensitive question. If the coin comes up heads, answer the sensitive question. If the respondent has the sensitive characteristic, answer 'yes' to the sensitive question if you are comfortable in doing so. Otherwise, answer the unrelated question. This option gives the respondents an incentive to answer truthfully, thereby improving the efficiency of the design.

APPENDIX A: EM Algorithm

We obtain the EM algorithm for a single area, where

$$y_j \mid \pi_1, \pi_2 \overset{ind}{\sim} \text{Binomial}(n_j, p_j\pi_1 + (1-p_j)\pi_2), j = 1, \ldots, g.$$

We introduce latent variables $(z_j, w_j), j = 1, \ldots, g$, as in the Gibbs sampler.

Then, for the expectation step,

$$z_j \mid \pi_1, \pi_2, y_j \overset{ind}{\sim} \text{Binomial}\{y_j, \frac{p_j\pi_1}{p_j\pi_1 + (1-p_j)\pi_2}\}, j = 1, \ldots, g,$$

$$w_j \mid \pi_1, \pi_2, y_j \overset{ind}{\sim} \text{Binomial}\{n_j - y_j, \frac{p_j(1-\pi_1)}{p_j(1-\pi_1) + (1-p_j)(1-\pi_2)}\}, j = 1, \ldots, g.$$

Therefore,

$$E(z_j \mid \pi_1, \pi_2, \boldsymbol{y}) = \frac{y_j p_j \pi_1}{p_j\pi_1 + (1-p_j)\pi_2}, \; E(w_j \mid \pi_1, \pi_2, \boldsymbol{y}) = \frac{(n_j - y_j)p_j(1-\pi_1)}{p_j(1-\pi_1) + (1-p_j)(1-\pi_2)},$$

$j = 1, \ldots, g$.

For the maximization step, it is convenient that, given π_1, π_2, y_j, z_j and w_j are independent. It is worth noting that the z_j and w_j are replaced by their expectations. More importantly, letting $y. = \sum_{j=1}^g y_j$, $z. = \sum_{j=1}^g z_j$, $w. = \sum_{j=1}^g w_j$,

$$\hat{\pi}_1 = \frac{z.}{z. + w.},$$

$$\hat{\pi}_2 = \frac{y. - z.}{n. - z. - w.}.$$

It is convenient that again, given $z., w., y.$, π_1 and π_2 are independent and they are beta random variables.

The estimated covariance matrix of $\hat{\pi}_1$ and $\hat{\pi}_2$, based on the negative inverse Hessian matrix, is

$$C = \begin{pmatrix} 1/a & -c/ab \\ -c/ab & 1/b \end{pmatrix} /(1 - c^2/ab),$$

where

$$a = \sum_{j=1}^g p_j^2 \left\{ \frac{y_j}{(p_j\hat{\pi}_1 + (1-p_j)\hat{\pi}_2)^2} + \frac{(n_j - y_j)}{(p_j(1-\hat{\pi}_1) + (1-p_j)(1-\hat{\pi}_2))^2} \right\},$$

$$b = \sum_{j=1}^g (1-p_j)^2 \left\{ \frac{y_j}{(p_j\hat{\pi}_1 + (1-p_j)\hat{\pi}_2)^2} + \frac{(n_j - y_j)}{(p_j(1-\hat{\pi}_1) + (1-p_j)(1-\hat{\pi}_2))^2} \right\},$$

$$c = \sum_{j=1}^g p_j(1-p_j) \left\{ \frac{y_j}{(p_j\hat{\pi}_1 + (1-p_j)\hat{\pi}_2)^2} + \frac{(n_j - y_j)}{(p_j(1-\hat{\pi}_1) + (1-p_j)(1-\hat{\pi}_2))^2} \right\}.$$

This covariance matrix, C, gives an underestimate of the true variability of $\hat{\pi}_1$ and $\hat{\pi}_2$ because $\hat{\pi}_1$ and $\hat{\pi}_2$ themselves are substituted into the population covariance matrix.

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.

Bourke, P. D., and Moran, M. A. (1988), "Estimating Proportions From Randomized Response Data Using the EM Algorithm," in *Journal of the American Statistical Association*, 83, 964-968.

Boruch, R. F. (1971), "Assuring Confidentiality of Responses in Social Research: A Note on Strategies," *American Sociologist*, 6, 308-311.

Chaudhuri, A. and Christofides, T. C. (2013), *Indirect Questioning Techniques in Surveys*, Springer: New York.

Chaudhuri, A. and Mukerjee, R. (1988) *Randomized Response: Theory and Techniques*, Dekker: New York.

Christofides, T. C. (2005), "Randomized Response Techniques for Two Sensitive Characteristics at the Same Time," *Metrika*, 62, 53-63.

Dass, B. K. and Chhabra, A. (2017), "Generalized Multi-Stage Optional Unrelated Question RRT Models," *Statistics and Applications*, 15, 1 & 2, 7-18.

Dihidar, K. and Basu, L. (2017) "Privacy Protection in Estimating Sensitive Population Proportion by a Modified Unrelated Question Model," *Statistics and Applications*, 15 Nos. 1 & 2, 19-25.

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

Folsom, R. E., Greenberg, B. G. and Horvitz, D. G. (1973), "The Two Alternate Questions RR Model for Human Surveys," *Journal of the American Statistical Association*, 68, 525-530.

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.

Groenitz, H. (2017), "Improving Estimation Accuracy in Nonrandomized Response Questioning Methods by Multiple Answers," *International Journal of Statistics and Probability* 6 (5), 101-109.

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.

Jones, D. L. R. (2011), "Academic Dishonesty: Are More Students Cheating?" *Business Communication Quarterly*, 74 (2), 141-150.

Johnson, M. L., Sedory, S. A. and Singh, S. (2016), "Alternative Methods to Make Efficient Use of Two Decks of Cards in Randomized Response Sampling," *Sociological Methods and Research*, 1-30.

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

Kuk, A. Y. (1990), "Asking Sensitive Questions Indirectly," *Biometrika*, 77, 436-438.

Lanke, J. (1975), "On the Choice of the Unrelated Question in Simmons' Version of Randomized Response," *Journal of the American Statistical Association* 70, 80-83.

Lee, C-S., Sedory, S. A. and Singh, S. (2013), "Simulated Minimum Sample Size Requirements in Various Randomized Response Models," *Communications in Statistics - Simulation and Computation*, 42, 4, 771-789.

Mangat, N. S. (1994), "An Improved Randomized Response Strategy," *Journal of the Royal Statistical Society* Series B, 56 (1), 93-95.

Mangat, N. S. (1992), "Two Stage Randomized Response Sampling Procedure Using Unrelated Question," *Journal of the Indian Society of Agricultural Statistics*, 44 (1), 82-87.

Mangat, N. S. and Singh, R. (1990) "An alternative randomized response procedure," *Biometrika*, 77 (2), 439-442.

Mangat, N. S., Singh, R. and Singh, S. (1992) "An Alternative Randomized Response Procedure," *Calcutta Statistical Association Bulletin*, 42, 277-281.

Mahmood, M., Singh, S. and Horn, S. (1998), "On the Confidentiality Guaranteed Under Randomized Response Sampling: A Comparison with Several New Techniques," *Biometrical Journal*, 40, 237-242.

Moors, J. (1971), "Optimization of the Unrelated Question Randomized Response Model," *Journal of the American Statistical Association*, 66, 627-629.

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.

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

plications, 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.

Shaw, P. and Chaudhuri, A. (2017), “Empirical Bayes Estimation Method in Some Randomized Response Techniques,” *Statistics and Applications*, 15, 1 & 2, 101-116.

Shon, P. C. H. (2006), “How College Students Cheat on In-Class Examinations: Creativity, Strain, and Techniques of Innovation,” *Plagiary: Cross Disciplinary Studies in Plagiarism, Fabrication, and Falsification*, 130-148.

Song, J. J. and Kim, J-M. (2017), “Bayesian Estimation of Rare Sensitive Attribute,” *Communications in Statistics - Simulation and Computation*, DOI: 10.1080/03610918.2015.1109655.

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.

Table 1: Summaries under the design-based procedure - estimates ($\hat{\pi}_1, \hat{\pi}_2$), standard errors (SE), coefficient of variation (CV) and correlation (Cor)

		π_1			π_2		
	$\hat{\pi}_1$	SE	CV	$\hat{\pi}_2$	SE	CV	Cor
1	.050	.325	6.492	1.100	.248	.225	-.829
2	.450	.335	.745	.600	.273	.455	-.817
3	5.700	4.210	.739	-.300	.990	-3.301	-.995
4	.731	.510	.698	.038	.692	17.991	-.944
5	.535	.195	.364	.823	.243	.295	-.591
6	-.150	.548	-3.654	1.050	.390	.371	-.943
7	1.291	.821	.636	.582	.312	.537	-.919
8	.071	.286	4.009	.786	.202	.258	-.707
9	.115	.188	1.633	.577	.215	.373	-.605
10	.889	.355	.400	.222	.367	1.650	-.800
11	-.556	.251	-.452	1.444	.296	.205	-.811

NOTE: Using the EM algorithm, the overall estimates for π_1 and π_2 are respectively $.279\ (.067)$ and $.798\ (.037)$; see Appendix A for the EM algorithm.

Table 2: Summaries under the individual-area Bayesian model - posterior mean (PM), posterior standard deviation (PSD), coefficient of variation (CV) and correlation (Cor)

	π_1			π_2			
	PM	PSD	CV	PM	PSD	CV	Cor
1	.353	.198	.562	.810	.138	.171	-.604
2	.482	.239	.495	.567	.209	.369	-.730
3	.555	.279	.502	.808	.110	.136	-.395
4	.479	.225	.469	.411	.260	.633	-.736
5	.568	.163	.286	.710	.187	.264	-.469
6	.387	.249	.643	.661	.195	.295	-.785
7	.601	.201	.335	.753	.134	.178	-.532
8	.265	.186	.700	.679	.157	.232	-.554
9	.216	.146	.677	.530	.173	.327	-.447
10	.667	.222	.333	.427	.233	.545	-.612
11	.191	.152	.796	.736	.180	.245	-.414

NOTE: Under the individual-area Bayesian model, the overall estimates for π_1 and π_2 are respectively $.279\ (.082)$ and $.793\ (.070)$.

Table 3: Bias (B) and posterior standard deviation (PSD) of π_1 and π_2 averaged over all locations and the 1000 simulation runs, number of locations, ℓ, and τ.

	$\tau = 10$				$\tau = 100$				$\tau = 1000$			
	π_1		π_2		π_1		π_2		π_1		π_2	
ℓ	B	PSD	B	PSD	B	PSD	B	PSD	B	PSD	B	PSD
10	.129	.183	-.189	.214	.128	.183	-.189	.213	.132	.178	-.194	.214
18	.130	.183	-.188	.214	.129	.183	-.188	.213	.132	.178	-.192	.214
25	.131	.183	-.189	.214	.130	.183	-.189	.213	.133	.178	-.192	.214
50	.130	.183	-.188	.214	.130	.183	-.188	.214	.134	.178	-.192	.214
75	.130	.184	-.188	.214	.131	.183	-.188	.214	.135	.178	-.193	.214
100	.130	.183	-.188	.214	.131	.183	-.188	.214	.135	.178	-.193	.214

Table 4: 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 the number locations, ℓ, and τ

		π_1				π_2			
τ	ℓ	RAB	PRMSE	C	Wid	RAB	PRMSE	C	Wid
10	10	.980	.264	.883	.638	.258	.307	.924	.721
	100	1.01	.265	.880	.637	.258	.306	.924	.722
100	10	.571	.252	.929	.658	.247	.301	.945	.734
	100	.578	.253	.926	.659	.246	.301	.946	.735
1000	10	.543	.251	.935	.660	.246	.301	.948	.735
	100	.548	.252	.931	.661	.245	.300	.950	.736

NOTE: The correlations between π_1 and π_2 for all locations at $\tau = 10, 100, 1000$ are respectively and approximately $-.550$, $-.565$, $-.567$.