Journal of the Korean Data & Information Science Society
2019, **30**(4), 943–949

http://dx.doi.org/10.7465/jkdi.2019.30.4.943
한국데이터정보과학회지

Bayesian analysis of a sensitive proportion

WonYoung Yun[1] · Balgobin Nandram[2] · Dal Ho Kim[3]

[13]Department of Statistics, Kyungpook National University
[2]Department of Mathematical Sciences, Worcester Polytechnic Institute

Received 26 June 2019, revised 13 July 2019, accepted 13 July 2019

Abstract

Respondents tend to answer untruthfully when they are asked to response sensitive questions, as they are reluctant to expose their identity. In order to reduce the response bias that can be generated through this, Warner (1965) proposed a randomized design which uses a randomization device that conceals individual response and protects the respondent. Thereafter, various survey designs to reduce the response bias were proposed, and Bayesian estimation and simulation methods have also been studied. This study proposes an analysis method to reduce posterior standard deviations of the sensitive proportions in a survey with sensitive questions and compare the results with the existing analyzes through the simulation. In addition, this study applies the proposed analysis method to the actual survey with the sensitive questionnaires related to the organizational commitment to the experienced employees.

Keywords: Blocked Gibbs sampler, Gibbs sampler, grid method, latent variables, mirrored question design.

1. Introduction

When respondents are asked to respond sensitive question related to the particular society or organization they belonged, they are more likely to lie as they are reluctant to expose their identity. Thereby, a bias for response occurs. Warner (1965) proposed a randomized design to reduce nonresponse and potential bias by social needs for investigating sensitive behaviors and beliefs. This survey design uses a randomization device, such as flipping a coin or pulling a "yes"/ "no" card, to respondents to prevent the investigator from observing the results, thereby protecting the privacy of respondents. As a result, it creates an environment for respondents to be more honest and more responsive, which enables to obtain more accurate information. For example, in order to ask a sensitive question such as "have you ever been cheating on this test," to university students, the options for questionnaires are designed as below:

[1] Ph.D. candidate, Department of Statistics, Kyungpook National University, Daegu 41566, Korea.
[2] Professor, Department of Mathematical Sciences, Worcester Polytechnic Institute, 100 Institute Road, Worcester, MA 01609, USA.
[3] Corresponding author: Professor, Department of Statistics, Kyungpook National University, Daegu 41566, Korea. E-mail: dalkim@knu.ac.kr

Option 1: "I have been cheating on this test" ["Yes"]
Option 2: "I have never cheated on this test" ["No"]

The n respondents with simple random sampling with replacement select the corresponding card through randomization device, and they make Bernoulli's trials by either answering "Yes" or "No" from the selected option. When the probability of selecting option 1 is p, and the other is $1-p$. The probability of the respondents answering "Yes" is as follows:

$$P(yes) = p\pi + (1-p)(1-\pi). \tag{1.1}$$

In this equation, the sensitive proportion, which is the object to be estimated, is π.

Warner (1965) defines either case of Bernoulli's trial as mirrored question design, that the respondents are asked to answer a sensitive question in the case of success, or asked to answer an opposite response of the sensitive question. Mirrored question design has been extended to three areas; the unrelated question design (Greenberg *et al.*, 1969), the forced response design (Fox and Tracy, 1986), and the disguised design (Kuk, 1990). After that, researchers have also worked on many other extensions of Warner's randomized response techniques (RRT).

The Bayesian approach of Warner's randomized response has also been actively researched. Winkler and Franklin (1979) provided an approximate Bayesian analysis of Warner's mirrored design, O'Hagan (1987) derived Bayesian linear estimators for the unrelated question design 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. Blair *et al.* (2015) published a review paper that classified randomized response technique into four types; mirrored question, forced response, disguised response, and unrelated question, and Nandram and Yu (2019) reviewed randomized response design in detail, respectively.

This paper attempts to investigate a better method by composing the Bayesian model of the mirrored question in Warner's randomized response through various simulation method. The outline of the remaining section is as follows. In section 2, the general model for Bayesian estimation of a sensitive proportion is reviewed. In Section 3, the Bayesian methodology is discussed to analyze data from a randomized response application. Especially for the mirrored question, Gibbs sampler with grid method, Gibbs sampler using latent variables, and blocked Gibbs sampler is proposed as a method of Bayesian estimation of a sensitive proportion. In Section 4, under the various simulation conditions, posterior mean (PM) and posterior standard deviations (PSD) are computed regarding the three types of a general model that are proposed in Section 3. In Section 5, among the survey questionnaires related to organizational commitment that has been dealt with in an actual company, convert the sensitive section as a mirrored question, and explore the result. Section 6 concludes.

2. Bayes estimation of a sensitive proportion

Let p denote the probability of selecting "yes" with the randomization device, and denote n the number of respondents. Note that p is a known value in the mirror design ($p \neq 1/2$). Let π denote the probability of selecting "yes" for the sensitive question. Then the probability of selecting the "yes" is $p\pi + (1-p)(1-\pi)$ and the "no" is $p(1-\pi) + (1-p)\pi$. Under random sampling, letting y denote the number of "yeses" obtained, then:

$$y|\pi \overset{ind}{\sim} \text{Binomial}\{n, p\pi + (1-p)(1-\pi)\}, y = 0, 1, ..., n. \tag{2.1}$$

Under the assumption of a priori $\pi \overset{iid}{\sim} \text{Uniform}(0, 1)$, the posterior density is the same as the likelihood function. Using Bayes' theorem, the joint posterior density is as follows:

$$\pi(\pi|y) \propto \{p\pi + (1-p)(1-\pi)\}^y \{p(1-\pi) + (1-p)\pi\}^{n-y}. \tag{2.2}$$

3. Computations

Through the joint posterior density (2.2), Gibbs sampler based on grid method and Gibbs sampler and blocked Gibbs sampler with latent variables are calculated, and the estimations of a sensitive proportion are compared, respectively.

3.1. Gibbs sampler with grid method

A Gibbs sampler is used to fit the joint posterior density. Through a Gibbs samplers with grid method, a computational difficulty may able to be resolved, and more sophisticated simulation could be derived. A grid method is processed as follows:

Step 1) Divide 100 intervals between 0 and 1 and take the point values $I_k, k = 1, ..., 100$ (i.e $I_0 = 0$ and $I_{100} = 1$).
Step 2) Calculate 100 mid-points $(M_k, k = 1, ..., 100)$ for each interval.
Step 3) Input the mid-points to the conditional posterior density and calculate the values $(a_k, k = 1, ..., 100)$ according to the mid-points.
Step 4) Calculate $b_k = a_k/A$, where $A = \sum_{k=1}^{100} a_k$.
Step 5) Calculate $c_k = \sum_{i=1}^{k} b_i$.
Step 6) Generate $u_1 \overset{iid}{\sim} \text{Uniform}(0, 1)$.
Step 7) Select the kth interval, which satisfies $c_k \leq u_1 < c_{k+1}$ made in Step 1.
Step 8) Generate $u_2 \overset{iid}{\sim} \text{Uniform}(I_k, I_{k+1})$ and set $\pi = u_2$.

3.2. Gibbs sampler using appropriate latent variables

It is difficult to treat the joint posterior density (2.2), and in the posterior analysis of the parameters of interest, computational difficulties are triggered. Thereby, it is convenient to introduce latent variables, z, and w, to obtain the augmented joint posterior density. Using the Gibbs sampler, the estimation of a sensitive proportion is obtained.

$$\pi(\pi, z, w|y) \propto \binom{y}{z} (p\pi)^z \{(1-p)(1-\pi)\}^{y-z} \binom{n-y}{w} \{p(1-\pi)\}^w \{(1-p)\pi\}^{n-y-w}. \tag{3.1}$$

Through this, the full conditionals are obtained as follows:

(i) $[z|\pi, w, y] \overset{ind}{\sim} \text{Binomial}\Big\{y, \dfrac{p\pi}{p\pi + (1-p)(1-\pi)}\Big\}$;

(ii) $[w|\pi, z, y] \overset{ind}{\sim} \text{Binomial}\Big\{n-y, \dfrac{p(1-\pi)}{p(1-\pi) + (1-p)\pi}\Big\}$;

(iii) $[\pi|z, w, y] \overset{ind}{\sim} \text{Beta}(z + n - y - w + 1, y - z + w + 1)$.

3.3. Blocked Gibbs sampler using appropriate latent variables

From the joint posterior density (3.1) with latent variables, another way to draw samples from $\pi(\pi.z.w|y)$ is to use blocked Gibbs sampler. In the case of the joint posterior density (3.1), integral calculus could be applied as (3.2):

$$\int_0^1 \pi(\pi, z, w|y)d\pi \propto \int_0^1 \binom{y}{z}\binom{n-y}{w} p^{z+w}(1-p)^{n-z-w}\pi^{n-y+z-w}(1-\pi)^{y-z+w}d\pi$$
$$\pi(z, w|y) \propto \binom{y}{z}\binom{n-y}{w} p^{z+w}(1-p)^{n-z-w}\frac{\Gamma(n-y+z-w+1)\Gamma(y-z+w+1)}{\Gamma(n+2)}. \quad (3.2)$$

Here, two blocks, $\pi(z, w|y)$ and $\pi(\pi|z, w, y)$, are obtained. Through (3.2), two full conditionals are obtained. By applying the derived value of Gibbs sampling through the two full conditionals to the beta distribution of $\pi(\pi|z, w, y)$, the value of blocked Gibbs samplers is obtained. Through (3.2), the full conditionals are obtained as follows:

(i) $[z|y, w] \propto \binom{y}{z} p^z(1-p)^{-z}\Gamma(n-y+z-w+1)\Gamma(y-z+w+1)$;

(ii) $[w|y, z] \propto \binom{n-y}{w} p^w(1-p)^{-w}\Gamma(n-y+z-w+1)\Gamma(y-z+w+1)$.

4. Simulation study

The simulation is processed to explore the difference among the performances of Gibbs sampler with grid method, and Gibbs sampler and blocked Gibbs sampler with latent variables. The design plan is as follows. Let the sample number as $n = 50$, and the probabilities driven by a randomization device as $\boldsymbol{p} = (0.8, 0.6, 0.3)$. Set the sensitive proportion as $\pi \overset{iid}{\sim} \text{Uniform}(0, 1)$. The condition for the simulation is as follows. First, perform $1,000$ simulated runs for Gibbs sampler with grid method, and $15,000$ simulated runs for Gibbs sampler and blocked Gibbs sampler. Then, burn in first $5,000$ and thin in one value in tenth units. As a result, total $1,000$ simulation results are derived. Though this, the following values are obtained: $\boldsymbol{\pi} = (0.336, 0.336, 0.336)$, $\mathbf{y} = (20, 23, 29)$.

PM and PSD with the 95% highest posterior density (HPD) intervals, are computed. Table 4.1 presents the result values regarding the three analysis methods. All three methods present estimation in the proximity with π. The most of the PSD values for blocked Gibbs sampler are shown out to be the lowest but when $p = 0.6$, the PSD value for Gibbs sampler is lower than the PSD value for blocked Gibbs sampler with a minute difference. Thereby, this study has additionally explored the HPD value and found out that the range of HPD is the lowest for blocked Gibbs sampler for all three given probability setting.

The convergence diagnostics of Gibbs sampler and blocked Gibbs sampler with latent variable have been processed. Table 4.2 shows Geweke's statistics and effective size of latent variables, respectively. Geweke's statistics show stable convergence, but effective size shows that the latent variables of blocked Gibbs sampler are more effective.

Table 4.1 Summaries under the Gibbs sampler with grid method and Gibbs /blocked Gibbs sampler with appropriate latent variables

		π		
		PM	PSD	HPD
$p = 0.8$				
	Grid method	0.3303	0.1137	(0.1074, 0.5441)
	Gibbs sampler	0.3410	0.1131	(0.1151, 0.5474)
	Blocked Gibbs sampler	0.3417	0.1111	(0.1420, 0.5606)
$p = 0.6$				
	Grid method	0.3921	0.2447	(0.0003, 0.8335)
	Gibbs sampler	0.3807	0.2363	(0.0004, 0.8262)
	Blocked Gibbs sampler	0.4086	0.2399	(0.0024, 0.8171)
$p = 0.3$				
	Grid method	0.3406	0.1869	(0.0003, 0.6635)
	Gibbs sampler	0.3757	0.1898	(0.0241, 0.7335)
	Blocked Gibbs sampler	0.3450	0.1807	(0.0233, 0.6837)

Table 4.2 Convergence diagnostics for Gibbs sampler and blocked Gibbs sampler

		Gibbs sampler		Blocked Gibbs sampler	
		Geweke's statistics	Effective size	Geweke's statistics	Effective size
$p = 0.8$					
	z	Z-score:-0.1318 P-value:0.8951	902	Z-score:-0.2127 P-value:0.8316	881
	w	Z-score:0.7099 P-value:0.7099	1224	Z-score:1.4964 P-value:0.1345	875
	π	Z-score:0.2716 P-value:0.7859	1000	Z-score:-1.0862 P-value:0.2774	772
$p = 0.6$					
	z	Z-score:-0.9121 P-value:0.3617	336	Z-score:-0.7689 P-value:0.442	1203
	w	Z-score:0.4816 P-value:0.6301	307	Z-score:0.7263 P-value:0.4677	873
	π	Z-score:-1.1894 P-value:0.2343	302	Z-score:-1.0689 P-value:0.2851	781
$p = 0.3$					
	z	Z-score:-0.104 P-value:0.9172	498	Z-score:0.135 P-value:0.8926	1000
	w	Z-score:0.6123 P-value:0.5403	512	Z-score:-0.9639 P-value:0.3351	833
	π	Z-score:-0.3967 P-value:0.6916	458	Z-score:0.5786 P-value:0.5628	1000

5. Real data analysis

After employing experienced employees, corporate human resource (HR) managers observe closely their movements between departments and internal organizations. This is an indicator to predict the productivity and service quality of the organization, and furthermore, it has a great influence on the overall operation of the company. Therefore, HR managers conduct a continuous analysis on the experienced employees through the survey. Since truth is required from the sensitive question, in this case, the RRT has been processed with the concept of drawing cards for the sensitive question. Twenty-one experienced employees who have

recently been employed in the last three months were asked to answer the following two questions. Then, their answers were converted into data, and the results were derived and compared based on the three analysis methods.

[Case 1]

- Q : The IoT/Data division merely employ the experienced employees, but does not show concerns.
- ratio : Yes : No=2:1

[Case 2]

- Q : If I have a chance, I would like to move to another team /project.
- ratio : Yes : No=3:1

For case 1, $p = 15/21$, $y = 9$. The number of running simulations was the same as that of the simulation study case in Section 4. For case 2, $p = 13/21$, $y = 7$. Table 5.1 and Table 5.2 present the results.

Table 5.1 Summaries under the Gibbs sampler with grid method and Gibbs/blocked Gibbs sampler with appropriate latent variables in the Question 1

	π		
	PM	PSD	HPD
Grid method	0.3747	0.2155	(0.0008, 0.7685)
Gibbs sampler	0.3809	0.2124	(0.0051, 0.7546)
Blocked Gibbs sampler	0.3754	0.2073	(0.0117, 0.7914)

Table 5.2 Summaries under the Gibbs sampler with grid method and Gibbs/blocked Gibbs sampler with appropriate latent variables in the Question 2

	π		
	PM	PSD	HPD
Grid method	0.2876	0.2304	(0.0008, 0.7635)
Gibbs sampler	0.2967	0.2286	(0, 0.7444)
Blocked Gibbs sampler	0.2888	0.2179	(0, 0.7358)

6. Discussion

In this paper, Warner's RRT was reviewed, and mirrored question design case has been explored among the various survey designs of RRT from the Bayesian model perspective. The three cases of Bayesian models, Gibbs sampler with grid method and Gibbs/ Blocked Gibbs sampler with appropriate latent variables, were examined and the values of PM, PSD, and HPD for each case were compared. According to the simulation, it has been found out that the PSD and HPD internal are the lowest for blocked Gibbs sampler. This study further applied the sensitive questionnaire relative to the organizational commitment to an actual survey to the experienced employees who are belonged to an organization and applied the three research methods. As a result, this study confirmed that the results for the real data analysis are similar to the simulation results.

References

Blair, G., Imai, K. and Zhou, Y-Y. (2015). Design and analysis of the randomized response technique. *Journal of the American Statistical Association*, **110**, 1304-1319.

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.

Kuk, A. Y. (1990). Asking sensitive questions indirectly. *Biometrika*, **77**, 436-438.

Nandram, B. and Yu, Y. (2019). Bayesian analysis of a sensitive proportion for a small area. *International Statistical Review*, **87**, 104-120.

O'Hagan, A. (1987). Bayes linear estimators for randomized response models. *Journal of the American Statistical Association*, **82**, 580-585.

Oh, M. (1994). Bayesian analysis of randomized response models: A Gibbs sampling approach. *Journal of the Korean Statistical Society*, **23**, 463-482.

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.