Journal of Korean
Data & Information Science Society
2002, Vol. 13, No.1 pp. 35 45

A Conditional Indirect Survey Method[1)]

Gi Sung Lee[2)] · Ki Hak Hong[3)]
Chang Kyoon Son[4)] · Ki Seong Nam[5)]

Abstract

For improving the quality of survey data of sensitive character, we suggest a conditional indirect survey method. In that method, only the respondents who answer directly to the less sensitive question respond indirectly to the more sensitive one by using the one sample unrelated question randomized response technique with the known π_y, the true proportion of unrelated group Y. We extend it to two sample method when π_y is unknown. We also consider the case that people who possess less sensitive character answer untruthfully. Finally we compare our method with the methods of Greenberg et al. and Carr et al..

KeyWords : a conditional indirect survey method, less sensitive character, sensitive character

1. Introduction

Most of marketing and opinion research companies have collected individual information through various surveys. They are continuously doing effort to improve the quality of survey to obtain more accurate data. But some respondents who are asked questions which related to privacy or anti-social problems may refuse to answer or may deliberately falsify their responses. It makes worse the quality of survey data and raises the problems of confidence. The randomized

---

1. This paper was supported by Woosuk University

2. Associate Professor, Division of Computer and Information Science, Woosuk University, 490 Hujung-ri, Samrye-up, Wanju-gun, Jeonbuk, 565-701, Korea
E-mail : gisung@woosuk.ac.kr
3. Associate Professor, Department of Computer Science, Dongshin University, 252 Daeho-dong, Naju, Chonnam, 520-714, Korea
4. Full-Time Lecturer, Department of Computer Science, Dongshin University, 252 Daeho-dong, Naju, Chonnam, 520-714, Korea
5. Lecturer, Department of Statistics, Changwon University, 9 Sarim-dong, Kyungnam, 641-773, Korea

response technique(RRT) which is one of indirect survey methods has been proposed as one means of obtaining an unbiased estimate of the proportion or quantity in a population group of persons possessing a particular trait or characteristic which the persons may be reluctant to acknowledge. This method was first suggested by Warner at 1965 as a related question form. He proposed a indirect survey method called RRT to procure trustworthy information about sensitive data from the respondents in sample survey, and estimated the sensitive population proportion by using the data collected from randomization device which was composed of sensitive and nonsensitive question with respective known probabilities, p and $1-p$.

Since then, many scientists have developed the method. Greenberg et al.(1969) improved the Warner method by replacing the nonsensitive question to the unrelated one. He also extended it to two-sample cases in case the unrelated proportion π_y is not known.

Loynes(1976) replaced the nonsensitive question to a forcible question that made respondents to answer only "yes". Carr et al.(1982) modified the Loynes' method to a kind of two stage method and proposed a conditional randomized response (CRR) method for reducing the standard error by asking questions conditional upon earlier answers. They borrowed a common sense that in the typical situation, not all subjects were not asked a subsequent question when their response could be deduced from their answer to a previous question. They asked the more sensitive question to respondents who answered "yes" for the earlier less sensitive question. But, in Carr et al.'s method the persons have to use the randomization device in twice and there is possibility that some people who said "yes" for the forcible question in the first stage also must say "yes" by the same type question in second stage. In that case it makes worse the quality of survey data and raises the problems of confidence.

We consider a method, a conditional indirect survey method, to improve quality of survey data by asking a direct question to the people of possessing less sensitive character. In that method, only the respondents who answer directly to the less sensitive question respond indirectly to the more sensitive one by using the one sample unrelated question randomized response technique with the known π_y, the true proportion of unrelated group Y. We extend it to two sample method when π_y is unknown. We also consider the case that people who possess less sensitive character respond untruthfully. Finally we compare our method with the methods of Greenberg et al. and Carr et al..

2. One sample conditional indirect survey method

2.1 truthful reporting

In this chapter, we will suggest a conditional one sample indirect survey method. According the method, only the respondents who answer directly to the less sensitive character B respond indirectly to the more sensitive one A by using the Greenberg et al.'s one sample unrelated question randomized response technique with the known π_y, the true proportion of unrelated group Y.

An SRSWR of size n is drawn from the population. In the first stage each interviewee who is asked directly the following question answers "yes" or "no".

Do you have possess the less sensitive character B?

If we assume the respondents respond truthfully, the value of λ_1 which is the probability of getting a "yes" is congruent corresponds to the value of π_1 which is the population proportion of B. Let n_1 be the number of say "yeses", then $\hat{\lambda}_1 = \frac{n_1}{n}$. The estimator $\hat{\pi}_1$ of π_1 is

$$\hat{\pi}_1 = \frac{n_1}{n} . \tag{2.1}$$

In the second stage, the n_1 respondents who said "yes" in the first stage, respond according to the results of randomization device, R which is composed of Greenberg et al.'s unrelated question technique.

<randomization device R>

	contents	selection probability
Question1	I am a member of Group A	p
Question2	I am a member of Group Y	$1-p$

Let λ_2 be the conditional probability of getting "yes" from the respondents who said "yes" in the first stage.

$$\lambda_2 = p\frac{\pi_2}{\pi_1} + (1-p)\pi_y , \tag{2.2}$$

where π_2 is the population proportion of sensitive group A, and π_y is the known population proportion of unrelated group Y.

Let n_2 be the number of "yeses" among n_1 respondents, then $\hat{\lambda}_2 = \frac{n_2}{n_1}$.

The estimator $\hat{\pi}_2$ of π_2 is

$$\hat{\pi}_2 = \frac{1}{np}[n_2 - (1-p)\pi_y n_1]. \tag{2.3}$$

Since $n_1 \sim b(n, \pi_1)$, and $n_2 \sim b(n, \pi_1\lambda_2)$, the expected value of $\hat{\pi}_2$ follows as

$$\begin{aligned} E(\hat{\pi}_2) &= \frac{1}{np}[E(n_2) - (1-p)\pi_y E(n_1)] \\ &= \frac{1}{np}[n\pi_1\lambda_2 - (1-p)\pi_y n\pi_1] \\ &= \pi_2 . \end{aligned}$$

$\hat{\pi}_2$ is unbiased estimator of π_2. The variance of $\hat{\pi}_2$ is

$$\begin{aligned} Var(\hat{\pi}_2) &= Var\left[\frac{n_2 - (1-p)\pi_y n_1}{np}\right] \\ &= \frac{Var(n_2) + (1-p)^2 \pi_y^2 Var(n_1) - 2(1-p)\pi_y Cov(n_1, n_2)}{(np)^2} \end{aligned} \tag{2.4}$$

Since $n_1 \sim b(n, \pi_1)$, $n_2 \sim b(n, \pi_1\lambda_2)$, and $n_2 \mid n_1 \sim b(n_1, \lambda_2)$

$$Var(n_1) = n\pi_1(1-\pi_1),$$

$$\begin{aligned} Var(n_2) &= E[Var(n_2 \mid n_1)] + Var[E(n_2 \mid n_1)] \\ &= E[n_1\lambda_2(1-\lambda_2)] + Var(n_1\lambda_2) \\ &= n[p\pi_2 + (1-p)\pi_1\pi_y][1 - p\pi_2 - (1-p)\pi_1\pi_y], \end{aligned}$$

$$\begin{aligned} Cov(n_1, n_2) &= E(n_1 n_2) - E(n_1)E(n_2) \\ &= E[n_1 E(n_2 \mid n_1)] - (n\pi_1)(n\pi_1\lambda_2) \\ &= n(1-\pi_1)[p\pi_2 + (1-p)\pi_1\pi_y]. \end{aligned}$$

If we apply the above three results to the equation (2.4), we can obtain the

variance of $\hat{\pi}_2$.

$$Var(\hat{\pi}_2) = \frac{\pi_1(1-p)\pi_y\{1-(1-p)\pi_y\} - p\pi_2\{2(1-p)\pi_y + p\pi_2 - 1\}}{np^2} . \quad (2.5)$$

2.2 less than completely truthful reporting

Let $\theta(0<\theta<1)$ denote the probability that respondents who belong to group B will tell the truth when confronted with a direct question concerning membership in the first stage. It is further postulated that respondents confronted with a question relating to membership in Group A will report truthfully because they use randomization device and respond indirectly. The probability of getting "yes" is

$$\lambda_2' = p\frac{\pi_2}{\pi_1\theta} + (1-p)\pi_y . \quad (2.6)$$

In this case the estimator $\hat{\pi}_2$ is biased estimator of π_2 and the bias is

$$B(\hat{\pi}_2) = \pi_2(1/\theta - 1) . \quad (2.7)$$

Hence, the mean squared error(MSE) of $\hat{\pi}_2$ is obtained as follows.

$$MSE(\hat{\pi}_2) = \frac{\pi_1\theta(1-p)\pi_y\{1-(1-p)\pi_y\} - p\pi_2\{2(1-p)\pi_y + p\pi_2 - 1\}}{np^2} + \{\pi_2(1/\theta - 1)\}^2 . \quad (2.8)$$

3. Two sample conditional indirect survey method

In this chapter we consider the case that the population proportion π_y of unrelated group Y is unknown, and will extend the method of chapter 2 to two sample case.

If the population proportion π_y of unrelated group Y is unknown in our suggested method, we need two independent sample to estimate π_y. We select two SRSWR of size $n_{1i}(i = 1, 2)$ from n_1 respondents who responded "yes" in the first stage.

The n_{1i} respondents respond according to the results of randomization device, $R(i)$ which is composed of Greenberg et al.'s two sample unrelated question

technique.

<randomization device $R(i)$>

	contents	selection probability
Question1	I am a member of Group A	p_i
Question2	I am a member of Group Y	$1-p_i$

Let $\lambda_{2i}(i=1,2)$ be the conditional probability of getting "yes" from the n_{1i} respondents.

$$\lambda_{2i} = p_i\frac{\pi_2}{\pi_1} + (1-p_i)\pi_y, \tag{3.1}$$

where π_1, π_2 are the population proportions of sensitive group B and A, and π_y is the unknown population proportion of unrelated group Y. Let n_{2i} be the number of "yeses" among n_{1i} respondents, then $\hat{\lambda}_{2i} = \frac{n_{2i}}{n_{1i}}$. The estimator $\hat{\pi}_2$ of π_2 is

$$\begin{aligned}\hat{\pi}_2 &= \hat{\pi}_1\left[\frac{(1-p_2)\hat{\lambda}_{21}-(1-p_1)\hat{\lambda}_{22}}{p_1-p_2}\right] \\ &= \frac{1}{n(p_1-p_2)}\left[\frac{n_1n_{21}}{n_{11}}(1-p_2)-\frac{n_1n_{22}}{n_{12}}(1-p_1)\right],\ p_1\neq p_2.\end{aligned} \tag{3.2}$$

Since $n_1 \sim b(n,\pi_1)$, $n_{2i} \mid n_{1i} \sim b(n_{1i},\lambda_{2i})$, the expected value of $\hat{\pi}_2$ is obtained as follows.

$$\begin{aligned}E(\hat{\pi}_2) &= \frac{1}{n(p_1-p_2)}\left[\frac{1-p_2}{n_{11}}E(n_1n_{21})-\frac{1-p_1}{n_{12}}E(n_1n_{22})\right] \\ &= \frac{1}{n(p_1-p_2)}\left[\frac{1-p_2}{n_{11}}E(n_1)E(n_{21}\mid n_{11})-\frac{1-p_1}{n_{12}}E(n_1)E(n_{22}\mid n_{12})\right]\end{aligned}$$

$$= \frac{1}{n(p_1 - p_2)}\left[\frac{(1-p_2)n_{11}\lambda_{21}}{n_{11}}E(n_1) - \frac{(1-p_1)n_{12}\lambda_{22}}{n_{12}}E(n_1)\right]$$

$$= \frac{1}{n(p_1 - p_2)}[(1-p_2)\lambda_{21}n\pi_1 - (1-p_1)\lambda_{22}n\pi_1]$$

$$= \pi_2 .$$

Therefore, $\hat{\pi}_2$ is unbiased estimator of π_2.

The variance of $\hat{\pi}_2$ is

$$Var(\hat{\pi}_2) = Var\left[\frac{1}{n(p_1 - p_2)}\left\{\frac{n_1 n_{21}}{n_{11}}(1-p_2) - \frac{n_1 n_{22}}{n_{12}}(1-p_1)\right\}\right]$$

$$= \frac{1}{n^2(p_1 - p_2)^2}\left[\frac{(1-p_2)^2 Var(n_1 n_{21})}{n_{11}^2} + \frac{(1-p_1)^2 Var(n_1 n_{22})}{n_{12}^2} - \frac{2(1-p_1)(1-p_2)Cov(n_1 n_{21}, n_1 n_{22})}{n_{11}n_{12}}\right]. \quad (3.3)$$

Since $n_1 \sim b(n, \pi_1)$, and $n_{2i} \mid n_{1i} \sim b(n_{1i}, \lambda_{2i})$, we can derive the following equations.

$$Var(n_1 n_{21}) = E[Var(n_1 n_{21} \mid n_{11})] + Var[E(n_1 n_{21} \mid n_{11})]$$

$$= E[n_1^2 n_{11}\lambda_{21}(1-\lambda_{21})] + Var(n_1 n_{11}\lambda_{21})$$

$$= n_{11}\lambda_{21}(1-\lambda_{21})[n\pi_1(1-\pi_1) + (n\pi_1)^2] + (n_{11}\lambda_{21})^2 n\pi_1(1-\pi_1),$$

$$Var(n_1 n_{22}) = n_{12}\lambda_{22}(1-\lambda_{22})[n\pi_1(1-\pi_1) + (n\pi_1)^2] + (n_{12}\lambda_{22})^2 n\pi_1(1-\pi_1),$$

$$Cov(n_1 n_{21}, n_1 n_{22}) = E(n_1 n_{21} n_1 n_{22}) - E(n_1 n_{21})E(n_1 n_{22})$$

$$= E(n_1^2)E(n_{21} n_{22} \mid n_{11} n_{12}) - E(n_1)E(n_{21} \mid n_{11})E(n_1)E(n_{22} \mid n_{12})$$

$$= n_{11} n_{12}\lambda_{21}\lambda_{22} n\pi_1(1-\pi_1) .$$

Hence, if we apply the above three results to the equation (3.3), we can obtain the variance of $\hat{\pi}_2$.

$$Var(\hat{\pi}_2) = \frac{\pi_1\left[(1-\pi_1+n\pi_1)\left\{(1-p_2)^2\frac{\lambda_{21}(1-\lambda_{21})}{n_{11}}+(1-p_1)^2\frac{\lambda_{22}(1-\lambda_{22})}{n_{12}}\right\}\right]}{n(p_1-p_2)^2}$$

$$+\frac{\pi_2^2(1-\pi_1)}{n\pi_1} \quad . \tag{3.4}$$

4. Efficiency comparison

We compare our one sample method with those of Greenberg et al. and Carr et al. and suggest the condition which our method is more efficiency than them.

In Greenberg et al.'s method, let $\hat{\pi}_g$ be the estimator of the population proportion π_2 of sensitive group A. The variance of $\hat{\pi}_g$ is

$$Var(\hat{\pi}_g) = \frac{\pi_2(1-\pi_2)}{n} + \frac{(1-p)[p\pi_2(1-2\pi_y)+\pi_y\{1-(1-p)\pi_y\}]}{np^2} \quad . \tag{4.1}$$

Now, the difference $Var(\hat{\pi}_g) - Var(\hat{\pi}_2)$ is

$$Var(\hat{\pi}_g) - Var(\hat{\pi}_2) = \frac{(1-\pi_1)(1-p)\pi_y\{1-(1-p)\pi_y\}}{np^2} \quad . \tag{4.2}$$

In equation (4.2), if $p \neq 1$, $\pi_1 \neq 1$, then $Var(\hat{\pi}_g) - Var(\hat{\pi}_2) > 0$. The conditions $p \neq 1$, $\pi_1 \neq 1$ are satisfied in general. We can see that the suggested one sample method is more efficiency than that of Greenberg et al.. Hence we can improve the quality of survey data even though our method is somewhat complicate to use.

Next, in Carr et al.'s method, let $\hat{\pi}_c$ be the estimator of the population proportion π_2 of sensitive group A. The variance of $\hat{\pi}_c$ is

$$Var(\hat{\pi}_c) = \frac{\{p\pi_1+(1-p)\}(1-p) - \pi_2(1-2p+p\pi_2)}{np} \quad . \tag{4.3}$$

The difference $Var(\hat{\pi}_c) - Var(\hat{\pi}_2)$ is

$$Var(\hat{\pi}_c) - Var(\hat{\pi}_2) = \frac{(1-p)[\pi_1\{p^2-\pi_y(1-(1-p)\pi_y)\}+p(1-p)-2p\pi_2(1-\pi_y)]}{np^2} \quad . \tag{4.4}$$

From the equation (4.4), we can obtain the condition region that satisfies $Var(\hat{\pi}_2) < Var(\hat{\pi}_c)$ as follows

$$\pi_2 < \frac{\pi_1[p^2 - \pi_y\{1 - (1-p)\pi_y\}] + p(1-p)}{2p(1-\pi_y)}, \quad \pi_y \neq 1. \tag{4.5}$$

Hence, our method is more efficiency than that of Carr et al. under the condition (4.5) and more simple to use than that because of using only one randomization device. But it is difficult to compare them analytically as we know in (4.5). So we do compare them numerically and find the conditions in which the suggested method achieves more efficiency than the corresponding method.

<Table 1> show the relative efficiency

$$RE = Var(\hat{\pi}_c) / Var(\hat{\pi}_2)$$

obtained under the conditions $n = 100$, $\pi_1 = 0.5$, π_2 of changing from 0.1 to 0.4 by 0.1, and p, π_y of changing from 0.1 to 0.9 by 0.2.

In <Table 1> the values greater than one demonstrate the gains of efficiency for the suggested method relative to Carr et al.'s method. We can see that the suggested method generally is effective when the values of π_2 is decreasing, and p and π_y are increasing.

5. Conclusions

For improving survey data quality, we suggest a conditional indirect survey method in which only the respondents who answer directly to the less sensitive question respond indirectly to the more sensitive one by using the unrelated question randomized response technique. We compare it with those of Greenberg et al. and Carr et al., generally our method is more efficiency than Greenberg et al's method even though it is some complicate in procedure. The suggested method can be reduced to Greenberg et al.'s one sample unrelated question technique if we let $\pi_1 = 1$. Hence, we can know that the suggested method is a generalized form of Greenberg et al.'s one sample unrelated question technique. So, the Greenberg et al.'s one sample unrelated question technique is a special case of the suggested method.

Comparing with Carr et al.'s method the suggested method is effective when the values of π_2 is decreasing, and p and π_y are increasing.

We also consider the case that the respondents who are confronted a direct question tell the truth with probability $\theta (0 < \theta < 1)$.

We extend our method to two sample conditional indirect survey method in case the true proportion of unrelated character Y is not known.

<Table 1> Efficiency comparison between the two methods, the suggested one sample conditional indirect survey method and Carr et al.'s method.

π_2	π_y / p	0.1	0.3	0.5	0.7	0.9
0.1	0.1	1.5779	0.7510	0.6209	0.6798	1.0955
	0.3	2.8825	1.6651	1.3590	1.3285	1.5354
	0.5	2.7924	1.9220	1.5913	1.4653	1.4653
	0.7	2.1153	1.7078	1.4798	1.3451	1.2681
	0.9	1.3333	1.2454	1.1733	1.1134	1.0632
0.2	0.1	1.2133	0.6436	0.5512	0.6228	1.0771
	0.3	1.8733	1.3219	1.1775	1.2313	1.5500
	0.5	1.7108	1.4343	1.3271	1.3271	1.4343
	0.7	1.3662	1.2699	1.2163	1.1961	1.2061
	0.9	1.0815	1.0644	1.0505	1.0396	1.0317
0.3	0.1	0.9373	0.5437	0.4815	0.5618	1.0548
	0.3	1.3195	1.0578	1.0131	1.1327	1.5802
	0.5	1.2110	1.1282	1.1282	1.2110	1.4193
	0.7	1.0509	1.0434	1.0585	1.0980	1.1678
	0.9	0.9794	0.9854	0.9935	1.0039	1.0166
0.4	0.1	0.7193	0.4499	0.4114	0.4964	1.0267
	0.3	0.9567	0.8403	0.8577	1.0278	1.6363
	0.5	0.9007	0.9007	0.9593	1.1028	1.4216
	0.7	0.8532	0.8852	0.9379	1.0188	1.1419
	0.9	0.9112	0.9308	0.9530	0.9781	1.0065

References

1. Carr, J. W. and Marascuilo, L. A. (1982). Optimal Randomized Response Techniques and Methods for Hypothesis Testing, *Journal of Educational Statistics*, 7, 295-310.
2. Chaudhuri, A. and Mukerjee, R. (1988). *Randomized Response : Theory and Techniques*, Marcel Dekker, Inc., New York.
3. Greenberg, B. G., Abul-Ela, Abdel-Latif A., Simmons, W. R., and Horvitz, D. G. (1969). The Unrelated Question Randomized Response Technique : Theoretical Framework, *Journal of the American Statistical Association*, 64,

520-539.
4. Loynes, R. M. (1976). Asymptotically Optimal Randomized Response Procedures, *Journal of the American Statistical Association*, 71, 924-928.
5. Ryu, J. B., Hong, K. H., and Lee, G. S. (1993). *Randomized Response Model*, Freedom Academy, Seoul.
6. Warner, S. L. (1965). Randomized Response ; A Survey Technique for Eliminating Evasive Answer Bias, *Journal of the American Statistical Association*, 60, 63-69.