The Korean Communications
in Statistics Vol. 10 No. 1, 2003
pp. 247-256

The Three-Stage Cluster Randomized Response Model for Obtaining Sensitive Information[1]

Gi Sung Lee[2], Ki Hak Hong[3], Chang Kyoon Son[4], Young Mee Jung[5]

Abstract

In this study, we systemize the theoretical validity for applying RRM to three-stage cluster sampling method and derive the estimate and it's variance of sensitive parameter. We derive the minimum variance form under the optimal values of the subsample sizes when the costs are fixed. Under the some given precision, we obtain the optimal values of the subsample sizes and derive the minimum cost form by using them. We apply the three-stage cluster RRM to field survey and suggest some necessary points for practical use.

Keywords : Sensitive question, Randomized response model, Three-stage cluster sampling, Field survey.

1. Introduction

In socioeconomic investigations we sometimes need facts about highly personal matters which people usually like to hide from others. Also, an inquirer often feels a delicacy in asking direct questions about private and confidential subjects, especially if the subjects carry any social stigma. For example, he or she would feel uncomfortable asking a person about gambling, drug taking, tax evasion, or the extent of any illegal income, accumulated assets, history of induced abortion, and many similar items.

Attempted direct questions about such sensitive issues often result in high non-response rates or high response biases arising out of willful misstatements or blatant lies. So intelligent devices are needed to reduce rates of non-response and biased response in order that fruitful inferences may be drawn from survey data when the issues involved demand protection of privacy.

---

1) This work was supported by grant No. (R05-2001-000-00068-0) from the Basic Research Program of the Korea Science & Engineering Foundation.

2) Associate Professor, Division of Computer and Information Science, Woosuk University, Jeonbuk, 565-701, Korea.
E-mail : gisung@woosuk.ac.kr

3) Professor, Department of Computer Science, Dongshin University, Naju, 520-714, Korea.

4) Full-Time Lecturer, Department of Computer Science, Dongshin University, Naju, 520-714, Korea.

5) Graduate Student, Department of Statistics, Dongguk University, Seoul, 100-715, Korea.

A simple technique involving the use of a randomized response rather than a direct one was introduced by Warner(1965). He has proposed a indirect survey method called randomized response model(RRM) to procure trustworthy information about sensitive data from the respondents in sample survey. He has 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 improved the method and developed new ones. In the Warner model, the two questions relate to groups that are perfectly negatively associated. Greenberg et al.(1969) extended two related groups to unrelated ones and suggested the unrelated question model. Recently, Ryu, Lee, and Lee(1995) considered practical aspects of RRM by analysing the cases adapted RRM. Since Warner, the RRMs which have been suggested and applied in field survey so far underlied sample selected by SRS(simple random sampling) from simple population. In general the populations considered are large and have complex structure. Lee and Hong(1998) extended the simple RRM to complex one by suggesting the two-stage cluster RRM. For a discussion of subject and further references, we refer to Fox and Tracy(1986), Chaudhuri and Mukerjee(1988), and Ryu, Hong and Lee(1993).

In this article a three-stage randomized response model is considered to estimate the sensitive population proportion and variance from a complex population which is composed of several clusters. We only consider the population which is composed of clusters of equal size in dealing with the three-stage cluster RRM, and assume that each sample is selected by SRSWOR(simple random sampling without replacement). We derive both optimal values of primary sampling unit(psu), secondary sampling unit(ssu), and thirdly sampling nuit(tsu) to minimize variance for a specified cost, and ones to minimize cost function for a specified precision. We apply the three-stage cluster RRM to field survey and suggest some necessary points for practical use.

2. Three-Stage Cluster Randomized Response Model

In this section we use RRM to estimate the sensitive proportion of population when it is composed of several clusters of containing a sensitive characteristic.

The population contains N primary sampling units(psu), each psu is composed of M secondary sampling units(ssu), and each ssu is composed of K thirdly sampling units(tsu). The corresponding numbers for the sample are n, m and k, respectively.

Each respondent selected by three-stage cluster sampling is provided with a randomization device by which he/she chooses one of the two questions with respective probabilities $p(\neq 0.5)$ and $1-p$ and then replies "yes" or "no" to the question chosen.

question 1 : Do you belong to the sensitive group A?

question 2 : Do you belong to the non-sensitive group $\overline{A}$?

Since the ultimate responses via the suggested RRM are obtained from k tsu's selected from ssu of size K, the population proportion of sensitive character of each tsu becomes π_{ij}. Where π_{ij} is unknown sensitive proportion of population members in jth ssu group within i th psu group.

Assuming truthful reporting, the probability of getting a "yes" response from the lth respondent ($l = 1, 2, \cdots, K$) of jth ssu ($j = 1, 2, \cdots, M$) within ith psu ($i = 1, 2, \cdots, N$) is

$$\lambda_{ij} = (2p-1)\pi_{ij} + (1-p) \ . \tag{2.1}$$

Define z_{ijl} as 1 or 0 according to the response of respondent saying "yes" or "no". Denoting the number of "yes" responses in lth tsu group by Z_{ij}, an unbiased estimator of π_{ij} is

$$\hat{\pi}_{ij} = \frac{\hat{\lambda}_{ij} - (1-p)}{2p-1} \ , \ p \neq \frac{1}{2} \ , \tag{2.2}$$

$$V(\hat{\pi}_{ij}) = \frac{\pi_{ij}(1-\pi_{ij})}{k} + \frac{p(1-p)}{k(2p-1)^2} \ , \tag{2.3}$$

where $\hat{\lambda}_{ij} = \frac{Z_{ij}}{k}$.

The problem we have to solve is to estimate π, the unknown proportion members in sensitive group of population which is composed of several clusters of containing a sensitive characteristic.

$$\pi = \frac{1}{N}\sum_{i=1}^{N}\pi_i \ , \tag{2.4}$$

where π_i is the unknown population proportion members in sensitive group of ith psu,

$$\pi_i = \frac{1}{M}\sum_{j=1}^{M}\pi_{ij} \ . \tag{2.5}$$

We can define the estimator $\hat{\pi}$ of π as follows

$$\hat{\pi} = \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\hat{\pi}_{ij} \ . \tag{2.6}$$

Theorem 1. If we assume SRSWOR at each stage, $\hat{\pi}$ is an unbiased estimator of π.

Proof.

$$
\begin{aligned}
E(\hat{\pi}) &= E_1E_2\left(\frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\pi_{ij}\right)\\
&= E_1\left(\frac{1}{n}\sum_{i=1}^{n}\pi_i\right)\\
&= \pi .
\end{aligned}
$$

■

Theorem 2. If we assume SRSWOR at each stage, the variance of $\hat{\pi}$ is

$$
\begin{aligned}
V(\hat{\pi}) &= \frac{N-n}{N(N-1)n}\sum_{i=1}^{N}(\pi_i-\pi)^2+\frac{M-m}{NM(M-1)nm}\sum_{i=1}^{N}\sum_{j=1}^{M}(\pi_{ij}-\pi_i)^2\\
&\quad+\frac{K-k}{NMKnmk}\sum_{i=1}^{N}\sum_{j=1}^{M}\left[\pi_{ij}(1-\pi_{ij})+\frac{p(1-p)}{(2p-1)^2}\right].
\end{aligned}
\tag{2.7}
$$

Proof.

$$
V(\hat{\pi}) = V_1E_2E_3(\hat{\pi})+E_1V_2E_3(\hat{\pi})+E_1E_2V_3(\hat{\pi})
$$

where

$$
\begin{aligned}
V_1E_2E_3(\hat{\pi}) &= V_1E_2\left(\frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\pi_{ij}\right)\\
&= V_1\left(\frac{1}{n}\sum_{i=1}^{n}\pi_i\right)\\
&= \frac{N-n}{N(N-1)n}\sum_{i=1}^{N}(\pi_i-\pi)^2 ,\\
E_1V_2E_3(\hat{\pi}) &= E_1V_2\left(\frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\pi_{ij}\right)\\
&= E_1\left[\frac{M-m}{M(M-1)n^2m}\sum_{i=1}^{n}\sum_{j=1}^{M}(\pi_{ij}-\pi_i)^2\right]\\
&= \frac{M-m}{NM(M-1)nm}\sum_{i=1}^{N}\sum_{j=1}^{M}(\pi_{ij}-\pi_i)^2
\end{aligned}
$$

and

$$
\begin{aligned}
E_1E_2V_3(\hat{\pi}) &= E_1E_2\left[\frac{K-k}{K(nm)^2k}\sum_{i=1}^{n}\sum_{j=1}^{m}\left\{\pi_{ij}(1-\pi_{ij})+\frac{p(1-p)}{(2p-1)^2}\right\}\right]\\
&= E_1\left[\frac{K-k}{MKn^2mk}\sum_{i=1}^{n}\sum_{j=1}^{M}\left\{\pi_{ij}(1-\pi_{ij})+\frac{p(1-p)}{(2p-1)^2}\right\}\right]\\
&= \frac{K-k}{NMKnmk}\sum_{i=1}^{N}\sum_{j=1}^{M}\left[\pi_{ij}(1-\pi_{ij})+\frac{p(1-p)}{(2p-1)^2}\right].
\end{aligned}
$$

These complete the proof.

■

In (2.7), $f_1 = \frac{n}{N}$, $f_2 = \frac{m}{M}$, $f_3 = \frac{k}{K}$, and let

$$S_1^{\,2} = \frac{1}{N-1}\sum_{i=1}^{N}(\pi_i - \pi)^2 ,$$

$$S_2^{\,2} = \frac{1}{N(M-1)}\sum_{i=1}^{N}\sum_{j=1}^{M}(\pi_{ij} - \pi_i)^2 ,$$

$$S_3^{\,2} = \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}\left[\pi_{ij}(1-\pi_{ij}) + \frac{p(1-p)}{(2p-1)^2}\right].$$

We can simplify (2.7) to

$$V(\hat{\pi}) = (1-f_1)\frac{S_1^{\,2}}{n} + (1-f_2)\frac{S_2^{\,2}}{nm} + (1-f_3)\frac{S_3^{\,2}}{nmk} . \tag{2.8}$$

The variance estimator $\hat{V}(\hat{\pi})$ is verified as an unbiased estimator of $V(\hat{\pi})$ by the Theorem 3.

Theorem 3. An unbiased estimator of $V(\hat{\pi})$ is

$$\hat{V}(\hat{\pi}) = (1-f_1)\frac{s_1^{\,2}}{n} + f_1(1-f_2)\frac{s_2^{\,2}}{nm} + f_1 f_2(1-f_3)\frac{s_3^{\,2}}{nmk} , \tag{2.9}$$

where $s_1^{\,2}$, $s_2^{\,2}$, $s_3^{\,2}$ are sample variances corresponding to $S_1^{\,2}$, $S_2^{\,2}$, $S_3^{\,2}$,

$$s_1^{\,2} = \frac{1}{n-1}\sum_{i=1}^{n}(\hat{\pi}_i - \hat{\pi})^2 ,$$

$$s_2^{\,2} = \frac{1}{n(m-1)}\sum_{i=1}^{n}\sum_{j=1}^{m}(\hat{\pi}_{ij} - \hat{\pi}_i)^2 ,$$

$$s_3^{\,2} = \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\left[\hat{\pi}_{ij}(1-\hat{\pi}_{ij}) + \frac{p(1-p)}{(2p-1)^2}\right].$$

Proof.

By the methods of two-stage sampling, we can show that

$$E(s_1^{\,2}) = S_1^{\,2} + (1-f_2)\frac{S_2^{\,2}}{m} + (1-f_3)\frac{S_3^{\,2}}{mk} . \tag{2.10}$$

We define $\hat{\pi}_{iK}$ as the sample proportion of sensitive characteristic over m ssu's in the i th psu given that all K tsu's were enumerated. Let $\hat{\pi}_K$ be the mean of n values of $\hat{\pi}_{iK}$.

Hence

$$\hat{\pi}_{iK} = \frac{1}{mK}\sum_{j=1}^{m}\sum_{l=1}^{K}\pi_{ijl} = \frac{1}{m}\sum_{j=1}^{m}\pi_{ij} , \quad \hat{\pi}_K = \frac{1}{n}\sum_{i=1}^{n}\hat{\pi}_{iK} .$$

Then from the two-stage sampling, it follows that

$$E\left[\frac{1}{n-1}\sum_{i=1}^{n}(\hat{\pi}_{iK}-\hat{\pi}_K)^2\right]= S_1^2+(1-f_2)\frac{S_2^2}{m}\ . \tag{2.11}$$

Now, if $\hat{\pi}_i$ is the sample proportion of sensitive members for the ith psu,

$$(\hat{\pi}_i-\hat{\pi})=(\hat{\pi}_{iK}-\hat{\pi}_K)+[(\hat{\pi}_i-\hat{\pi}_{iK})-(\hat{\pi}-\hat{\pi}_K)]\ . \tag{2.12}$$

By first averaging over samples in which the psu's and ssu's are fixed, it may be shown that

$$\begin{aligned} &E\left[\frac{1}{n-1}\sum_{i=1}^{n}\{(\hat{\pi}_i-\hat{\pi}_{iK})-(\hat{\pi}-\hat{\pi}_K)\}^2\right] \\ &=\frac{n}{n-1}\left(\frac{1}{mk}-\frac{1}{nmk}\right)(1-f_3)\,S_3^2 \\ &=(1-f_3)\frac{S_3^2}{mk}\ , \end{aligned} \tag{2.13}$$

and that the cross product terms from (2.12) contribute nothing. This establishes the result for (2.10). Those for $E(s_2)$ and $E(s_3)$ are found similarly. Hence

$$\begin{aligned} E[\hat{V}(\hat{\pi})] &=\frac{(1-f_1)}{n}\left[S_1^2+(1-f_2)\frac{S_2^2}{m}+(1-f_3)\frac{S_3^2}{mk}\right] \\ &\quad+\frac{f_1(1-f_2)}{nm}\left[S_2^2+(1-f_3)\frac{S_3^2}{k}\right]+\frac{f_1f_2(1-f_3)}{nmk}\,S_3^2 \\ &=(1-f_1)\frac{S_1^2}{n}+f_1(1-f_2)\frac{S_2^2}{nm}+f_1f_2(1-f_3)\frac{S_3^2}{nmk} \\ &=V(\hat{\pi})\ . \end{aligned}$$

■

3. Optimum values of sub-sample sizes of m and k

3.1. The optimum values of m and k given a specified cost

The simplest cost function of three-stage sampling is of the form

$$C'=C-c_0=c_1n+c_2nm+c_3nmk\ . \tag{3.1}$$

The terms C and c_0 represent a total and an overhead cost respectively. The terms c_1, c_2, and c_3 represent required costs for obtaining psu, ssu, and tsu respectively.

We can rewrite the variance of (2.8) as follows

$$V(\hat{\pi})=\frac{1}{n}\left(S_1^2-\frac{S_2^2}{M}\right)+\frac{1}{nm}\left(S_2^2-\frac{S_3^2}{K}\right)+\frac{S_3^2}{nmk}-\frac{S_1^2}{N}\ . \tag{3.2}$$

Our problem is to choose m and k so as to minimize $V(\hat{\pi})$ for specified C. Consider the Lagrange function

$$\varnothing = \frac{1}{n}\left(S_1^2 - \frac{S_2^2}{M} \right) + \frac{1}{nm}\left(S_2^2 - \frac{S_3^2}{K} \right) + \frac{S_3^2}{nmk} - \frac{S_1^2}{N} + \lambda(c_1 n + c_2 nm + c_3 nmk - C'), \tag{3.3}$$

where λ is the Lagrange multiplier.

Differentiating $\varnothing$ with respect to n, m and k and equating them to zero, and using (3.1) we obtain the optimum values m_{opt}, k_{opt} and n_{opt} as follows.

$$m_{opt} = \sqrt{\frac{\left(S_2^2 - \frac{S_3^2}{K} \right)c_1}{\left(S_1^2 - \frac{S_2^2}{M} \right)c_2}}, \tag{3.4}$$

$$k_{opt} = \sqrt{\frac{S_3^2 c_2}{\left(S_2^2 - \frac{S_3^2}{K} \right)c_3}}, \tag{3.5}$$

$$n_{opt} = \frac{(C - c_0)\sqrt{\left(S_1^2 - \frac{S_2^2}{M} \right)/c_1}}{\sqrt{\left(S_1^2 - \frac{S_2^2}{M} \right)c_1} + \sqrt{\left(S_2^2 - \frac{S_3^2}{K} \right)c_2} + S_3\sqrt{c_3}}. \tag{3.6}$$

A formula for the minimum variance with fixed cost is obtained by substituting the values of m_{opt} , k_{opt} and n_{opt} in (3.4), (3.5) and (3.6) into (3.2). The result is

$$V_{\min}(\hat{\pi}) = \frac{\left[\sqrt{\left(S_1^2 - \frac{S_2^2}{M} \right)c_1} + \sqrt{\left(S_2^2 - \frac{S_3^2}{K} \right)c_2} + S_3\sqrt{c_3} \right]^2}{C - c_0} - \frac{S_1^2}{N}. \tag{3.7}$$

3.2. The optimum values of m and k given a specified variance

We determine the optimum values m_{opt} , k_{opt} and n_{opt} given a specified variance by using the same method of section 3.1. The Lagrange function $\varnothing$ is

$$\varnothing = c_0 + c_1 n + c_2 nm + c_3 nmk + \lambda\left[\frac{1}{n}\left(S_1^2 - \frac{S_2^2}{M} \right) + \frac{1}{nm}\left(S_2^2 - \frac{S_3^2}{K} \right) + \frac{S_3^2}{nmk} - \frac{S_1^2}{N} - V_0 \right], \tag{3.8}$$

the optimum values m_{opt}, k_{opt} and n_{opt} as follows.

$$m_{opt} = \sqrt{\frac{\left(S_2^2 - \frac{S_3^2}{K} \right)c_1}{\left(S_1^2 - \frac{S_2^2}{M} \right)c_2}}, \tag{3.9}$$

$$k_{opt} = \sqrt{\frac{S_3^{\,2} c_2}{\left(S_2^{\,2} - \frac{S_3^{\,2}}{K} \right) c_3}}\,, \tag{3.10}$$

$$n_{opt} = \frac{\sqrt{\left(S_1^{\,2} - \frac{S_2^{\,2}}{M} \right) c_1} + \sqrt{\left(S_2^{\,2} - \frac{S_3^{\,2}}{K} \right) c_2} + S_3 \sqrt{c_3}}{\left(V_0 + \frac{S_1^{\,2}}{N} \right) \sqrt{c_1 / \left(S_1^{\,2} - \frac{S_2^{\,2}}{M} \right)}}\,. \tag{3.11}$$

A formula for the cost function with fixed variance is obtained by substituting the values of m_{opt}, k_{opt} and n_{opt} in (3.9), (3.10) and (3.11) into (3.1). The result is

$$C = c_0 + \frac{\left[\sqrt{\left(S_1^{\,2} - \frac{S_2^{\,2}}{M} \right) c_1} + \sqrt{\left(S_2^{\,2} - \frac{S_3^{\,2}}{K} \right) c_2} + S_3 \sqrt{c_3} \right]^2}{V_0 + \frac{S_1^{\,2}}{N}}\,. \tag{3.12}$$

4. Empirical survey

We applied the suggested method to estimate the proportion of feeling sexual drive to coed. The structure of survey population and sample of each stage are showed in Table 1.

Table 1. Survey Population and Sample

survey population	10 Universities 6,000 and over. (Assume that each university has 3 clusters of size 2,000. The clusters are College of Natural Science, College of Humanities and Social Science, and College of Arts and Physical Training.)
1st stage sample	Woosuk University, Dongshin University
2nd stage sample	two colleges per university
3rd stage sample	50 students per college

The ultimate sample selected by three-stage cluster sampling is composed of 200 students.

We asked first them directed question and then explained to each selected student how to use the randomization device, how their responses were kept more confidential. After that we asked them the same question via the suggested RRM to compare two methods. For directed method, each student was provided a questionnaire including a question about sexual drive to coed, and answer directly about his/her status. For the suggested RRM, each respondent was provided with the following randomization device by which he/she chooses one of the two questions in Table 2.

Table 2.

Question 1	"Have you ever felt sexual drive to coed?"
Question 2	"Have you never felt sexual drive to coed?"

We could see from pretest results that the question about sexual drive to coed was very sensitive to students. The survey was continued from 19. Nov. 2001 to 30. Nov. 2001. The data were collected indirectly by using the randomization device.

The randomization device consists of a box containing balls of two different colors, say, red and yellow with respective selection probabilities 0.6 and 0.4. The student is requested to report his/her true status according to the sensitive question if the selected ball is red, or report the nonsensitive question if the selected ball is yellow.

Table 3. The resulting statistics for RRM and directed method.

Statistic / Question type	The sample proportion of say "yes"	The proportion of population	sampling error
Directed question	35/200	0.175	0.000287
Indirected question	107/200	0.325	0.00309

We can notice that the estimate from RRM is larger than that of directed method by 0.15, and this result is statistically significant.

5. Conclusions and Discussions

We systemize the theoretical validity for applying RRM to three-stage cluster sampling method to estimate the sensitive population proportion and variance from a complex population which is composed of several clusters. We derive both optimal values of primary sampling unit(psu), secondary sampling unit(ssu), and thirdly sampling unit(tsu) to minimize variance for a specified cost, and ones to minimize cost function for a specified precision.

We apply the suggested three-stage cluster RRM to field survey and compare it with directed method. We can find that the estimate from the suggested three-stage cluster RRM is larger than that of directed method. This means that our model is statistically significant.

Choosing a method for collecting survey data is a complex decision involving considerations of expense, response rates, the sorts of question being asked, and the amount of information needed. The suggested RRM is one of useful methods although extra effort is necessary to get an acceptable information from randomization device.

We suggest some points which are necessary and useful for applying the suggested three-stage cluster RRM to field survey.

First, to study whether or not the matter to estimate is sensitive through pretest.

Second, to ascertain if the population to survey is appropriate to apply three-stage cluster sampling.

Third, to make sure that interviewers fully understand about the randomization device.

Fourth, to choose proper randomization device according to survey method.

Finally, to effort to reduce non-sampling errors.

We expect the RRM suggested in this paper is helpful to researchers of various fields of study such as sociology, economy, medicine, business administration, and so on.

We only consider clusters of equal size in dealing with the three-stage cluster RRM. It is possible to apply RRM based on probability proportion to size sample to field survey in the case of unequal clusters. We leave it for further study.

References

[1] Chaudhuri, A. and Mukerjee, R. (1988). *Randomized Response : Theory and Techniques,* Marcel Dekker, Inc., New York.

[2] Fox, J.A. and Tracy, P.E. (1986). *Randomized Response : A Method for Sensitive Survey,* Sage Publications.

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

[4] Lee, G.S., Hong, K.H. (1998). Two-Stage Cluster Randomized Response Model, *The Korean Communications in Statistics*, Vol. 5, No. 1, 99-105.

[5] Ryu, J.B., Hong, K.H. and Lee, G.S. (1993). *Randomized Response Model*, Freedom Academy, Seoul.

[6] Ryu, J.B., Lee, K.O., Lee, G.S. (1995). A Practical Method of Randomized Response Technique, *The Korean Journal of Applied Statistics*, Vol. 8, No. 1, 9-26.

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

[Received December 2002, Accepted April 2003]