Journal of Korean
Data & Information Science Society
2003, Vol. 14, No.4 pp. 1067~1082

Randomized Response Model with Discrete Quantitative Attribute by Three-Stage Cluster Sampling[1)]

Gi Sung Lee[2)] · Ki Hak Hong[3)]

Abstract

In this paper, we propose a randomized response model with discrete quantitative attribute by three-stage cluster sampling for obtaining discrete quantitative data by using the Liu & Chow model(1976), when the population was made up of sensitive discrete quantitative clusters. We obtain the minimum variance by calculating the optimum number of fsu, ssu, tsu under the some given constant cost. And we obtain the minimum cost under the some given accuracy.

KeyWords : Randomized response model; Discrete quantitative attribute; Three-stage cluster sampling; Subsample size.

1. Introduction

One of the most interest in socioeconomic investigations is to reduce non-sampling errors which can be arisen due to the evasive or untruthful answers. These errors are increasing when the more respondents are to be asked the sensitive questions.

A simple technique involving the use of a randomized response rather than a direct one was introduced by Warner(1965). He has proposed an 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

---

1) This paper was supported by Woosuk University

2) Associate Professor, Department of Computer Science & Statistics, Woosuk University, Wanju-gun, Jeonbuk, 565-701, Korea.
E-mail : gisung@woosuk.ac.kr

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

randomization device which was composed of sensitive and nonsensitive question.

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.(1971) suggested a quantitative unrelated question model by modifying the unrelated question model(1969) which had extended the Warner model by extending the two related groups to unrelated ones. But his method has several difficulties in choosing the unrelated question that has the same mean and variance to those for sensitive question.

Liu & Chow(1976) suggested a randomized response model to deal with discrete quantitative cases.

Since Warner, the RRMs which have been suggested and applied in field survey based on sample selected by SRS(simple random sampling) from simple population. However those RRM methods require more efforts and cost than those of direct methods, especially the populations considered in field are usually large and have complex structure.

To solve those difficulties and problems, Lee and Hong(1998) suggested a two-stage cluster randomized response model for estimating the proportion of people with a sensitive characteristic when the population was composed of several clusters. Recently, a field survey for the sensitive character such as the feeling of sexual impulse has been executed through a three-stage cluster randomized response technique by Lee et al.(2003). The question was that "Have you ever felt sexual drive to coed?". They assumed 10 medium-sized universities of 6,000 students each other in Cholla province, each university has three colleges of Natural Science, Humanities and Social Science, and Arts and Physical Training. They also assumed that each college equally has 2,000 students. Since the number of students of each college was too large to survey by two-stage cluster sampling they applied three-stage cluster sampling to select ultimate 50 students from each college of 2,000 students and obtained responses by applying randomized response model to them.

We can see that three-stage cluster sampling may have more practical applications than two-stage cluster sampling in field work. While Lee et al.(2003) dealt with only qualitative questions for sensitive characters it is necessary to study discrete quantitative ones for them.

In this article a three-stage discrete quantitative randomized response model which apply Liu & Chow's technique to the ultimate sampling unit, the third sampling unit(tsu), is considered to estimate the sensitive population proportion and variance from a complex population which is composed of several clusters. We assume that each sample is selected by SRSWOR(simple random sampling without replacement). We derive both optimal values of first sampling unit(fsu), second sampling unit(ssu), and third sampling unit(tsu) to minimize variance for a specified cost, and ones to minimize cost function for a specified precision.

2. Randomized Response Model with Discrete Quantitative Attribute by Three-Stage Cluster Sampling

2.1. proposed model

In this section we are to estimate the sensitive proportion of population when it is composed of several clusters of containing a sensitive attribute by using three stage RRM which apply Liu & Chow's technique to the ultimate sampling unit.

Let the number of fsu be N, the number of ssu in the $i(i=1,2,\cdots,N)$th fsu be M, and the number of tsu in the $j(j=1,2,\cdots,M)$th ssu of the $i(i=1,2,\cdots,N)$th fsu be K. The corresponding numbers for the sample are n, m and k, respectively.

The ultimate respondents selected by three-stage cluster sampling answer to the result of the Liu & Chow's randomization device. Each respondent is asked to turn the device upside down, shake the device throughly, and turn it right side up to allow one of the balls to appear in the window of the device. The ball in the window will either be red or white. If it is a red ball, the respondent will be asked to answer the sensitive question. If the ball is white, there will be a number marked on its surface, and the respondent simply tells the number. The answers will again be $0,1,\cdots,s,$ depending on the number marked on the surface of the white ball. Interviewers standing opposite the respondents do not know which color appeared in the window of the device, and, therefore, do not know if the respondents have experienced the sensitive event in the question.

Let w_t represent the number of white balls marked $t(t=0,1,2,\cdots,s)$, and r represents the number of unmarked red balls, then the total number of balls in the device is $r+w$ where $w=\sum_{t=0}^{s} w_t$.

The probability $\lambda_{ij(t)}$ of response $t(t=0,1,2,\cdots,s)$ of the $l(l=1,2,\cdots,K)$th respondent in the $j(j=1,2,\cdots,M)$th ssu drawn from the $i(i=1,2,\cdots,N)$th fsu, is given by

$$\lambda_{ij(t)}=\pi_{ij(t)}\left(\frac{r}{r+w}\right)+\frac{w_t}{r+w}\ , \tag{2.1}$$

where $\pi_{ij(t)}$ is the population proportion of respondents who possess t quantitative measure in the jth ssu drawn from the ith fsu, and can be written as

$$\pi_{ij(t)}=\frac{1}{K}\sum_{l=1}^{K} x_{ijl(t)}\ .$$

If $z_{ij(t)} = \sum_{l=1}^{k} z_{ijl(t)}$ represent the number of response $t(t=0,1,2,\cdots,s)$ of the $l(l=1,2,\cdots,K)$th respondent in the $j(j=1,2,\cdots,M)$th ssu drawn from the $i(i=1,2,\cdots,N)$th fsu, $\hat{\lambda}_{ij(t)}$ is given by $\hat{\lambda}_{ij(t)} = \frac{z_{ij(t)}}{k}$, and from the equation (2.1) the estimator of $\pi_{ij(t)}$, $\hat{\pi}_{ij(t)}$ can be written as

$$\hat{\pi}_{ij(t)} = \frac{(r+w)\ \hat{\lambda}_{ij(t)}}{r} - \frac{w_t}{r} \ . \tag{2.2}$$

The variance and covariance of $\hat{\pi}_{ij(t)}$ are

$$V(\hat{\pi}_{ij(t)}) = ((r+w)/r)^2\, \lambda_{ij(t)}(1-\lambda_{ij(t)})\,/\,k\,, \tag{2.3}$$

$$Cov(\hat{\pi}_{ij(t)},\ \hat{\pi}_{ij(u)}) = -((r+w)/r)^2 \lambda_{ij(t)}\lambda_{ij(u)}/\,k\,. \tag{2.4}$$

The population proportion $\pi_{(t)}$ of respondents who possess t quantitative measure is

$$\pi_{(t)} = \frac{1}{N}\sum_{i=1}^{N}\pi_{i(t)}\,, \tag{2.5}$$

where $\pi_{i(t)}$ is the population proportion of respondents who possess t quantitative measure in the ith fsu, and can be written as

$$\pi_{i(t)} = \frac{1}{M}\sum_{j=1}^{M}\pi_{ij(t)}\,. \tag{2.6}$$

The estimator $\hat{\pi}_{(t)}$ of $\pi_{(t)}$ is given by

$$\hat{\pi}_{(t)} = \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\hat{\pi}_{ij(t)}\,. \tag{2.7}$$

Theorem 1. The estimator $\hat{\pi}_{(t)}$ is unbiased estimator of $\pi_{(t)}$.

Proof.

$$\begin{aligned} E(\hat{\pi}_{(t)}) &= E_1E_2E_3\left(\frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\hat{\pi}_{ij(t)}\right) \\ &= E_1E_2\left(\frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\pi_{ij(t)}\right) \\ &= E_1\left(\frac{1}{n}\sum_{i=1}^{n}\pi_{i(t)}\right) \\ &= \pi_{(t)}\,. \end{aligned}$$

■

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

$$
\begin{aligned}
V(\hat{\pi}_{(t)}) &= (1-f_1)\frac{1}{n(N-1)}\sum_{i=1}^{N}(\pi_{i(t)}-\pi_{(t)})^2 \\
&\quad + (1-f_2)\frac{1}{nmN(M-1)}\sum_{i=1}^{N}\sum_{j=1}^{M}(\pi_{ij(t)}-\pi_{i(t)})^2 \\
&\quad + (1-f_3)\frac{((r+w)/r)^2}{nmkNM}\sum_{i=1}^{N}\sum_{j=1}^{M}\lambda_{ij(t)}(1-\lambda_{ij(t)}) ,
\end{aligned} \tag{2.8}
$$

where $f_1 = \frac{n}{N}$, $f_2 = \frac{m}{M}$, $f_3 = \frac{k}{K}$.

Proof.

Since $V(\hat{\pi}_{(t)}) = V_1E_2E_3(\hat{\pi}_{(t)}) + E_1V_2E_3(\hat{\pi}_{(t)}) + E_1E_2V_3(\hat{\pi}_{(t)})$,

$$
\begin{aligned}
V_1E_2E_3(\hat{\pi}_{(t)}) &= V_1E_2\left(\frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\pi_{ij(t)}\right) \\
&= (1-f_1)\frac{1}{n(N-1)}\sum_{i=1}^{N}(\pi_{i(t)}-\pi_{(t)})^2 , \\
E_1V_2E_3(\hat{\pi}_{(t)}) &= E_1V_2\left(\frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\pi_{ij(t)}\right) \\
&= (1-f_2)\frac{1}{nmN(M-1)}\sum_{i=1}^{N}\sum_{j=1}^{M}(\pi_{ij(t)}-\pi_{i(t)})^2
\end{aligned}
$$

and

$$
\begin{aligned}
E_1E_2V_3(\hat{\pi}_{(t)}) &= E_1E_2V_3\left(\frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\hat{\pi}_{ij(t)}\right) \\
&= (1-f_3)\frac{((r+w)/r)^2}{nmkNM}\sum_{i=1}^{N}\sum_{j=1}^{M}\lambda_{ij(t)}(1-\lambda_{ij(t)}) .
\end{aligned}
$$

We can see that the variance of $\hat{\pi}_{(t)}$ is given as (2.8).

■

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

$$
\begin{aligned}
\widehat{V}(\hat{\pi}_{(t)}) &= \frac{1-f_1}{n}\frac{1}{n-1}\sum_{i=1}^{n}(\hat{\pi}_{i(t)}-\hat{\pi}_{(t)})^2 \\
&\quad + \frac{f_1(1-f_2)}{nm}\frac{1}{n(m-1)}\sum_{i=1}^{n}\sum_{j=1}^{m}(\hat{\pi}_{ij(t)}-\hat{\pi}_{i(t)})^2 \\
&\quad + \frac{f_1f_2(1-f_3)}{nmk}\frac{((r+w)/r)^2}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\hat{\lambda}_{ij(t)}(1-\hat{\lambda}_{ij(t)}) ,
\end{aligned} \tag{2.9}
$$

where $\hat{\pi}_{i(t)} = \frac{1}{m}\sum_{j=1}^{m}\hat{\pi}_{ij(t)}$.

Proof.

Before complete proof, we first show that $E\left[\frac{1}{n-1}\sum_{i=1}^{n}(\hat{\pi}_{i(t)}-\hat{\pi}_{(t)})^2\right]$ is described as follows.

$$E\left[\frac{1}{n-1}\sum_{i=1}^{n}(\hat{\pi}_{i(t)}-\hat{\pi}_{(t)})^2\right]=\frac{1}{N-1}\sum_{i=1}^{N}(\pi_{i(t)}-\pi_{(t)})^2 + \frac{1-f_2}{m}\frac{1}{N(M-1)}\sum_{i=1}^{N}\sum_{j=1}^{M}(\pi_{ij(t)}-\pi_{i(t)})^2 + \frac{1-f_3}{mk}\frac{((r+w)/r)^2}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}\lambda_{ij(t)}(1-\lambda_{ij(t)}) . \tag{2.10}$$

If we define $\hat{\pi}_{iK(t)}$ as the sample proportion of sensitive attribute t over m ssu's in the ith fsu given that all K tsu's were enumerated, then $\hat{\pi}_{iK(t)}=\frac{1}{m}\sum_{j=1}^{m}\pi_{ij(t)}$, and $\hat{\pi}_{K(t)}=\frac{1}{n}\sum_{i=1}^{n}\hat{\pi}_{iK(t)}$.

Then from the two-stage sampling, we can show that

$$E\left[\frac{1}{n-1}\sum_{i=1}^{n}(\hat{\pi}_{iK(t)}-\hat{\pi}_{K(t)})^2\right]=\frac{1}{N-1}\sum_{i=1}^{N}(\pi_{i(t)}-\pi_{(t)})^2 + \frac{1-f_2}{m}\frac{1}{N(M-1)}\sum_{i=1}^{N}\sum_{j=1}^{M}(\pi_{ij(t)}-\pi_{i(t)})^2 . \tag{2.11}$$

Now, if $\hat{\pi}_{i(t)}$ is the sample proportion of sensitive attribute t for the ith fsu,

$$(\hat{\pi}_{i(t)}-\hat{\pi}_{(t)})=(\hat{\pi}_{iK(t)}-\hat{\pi}_{K(t)})+[(\hat{\pi}_{i(t)}-\hat{\pi}_{iK(t)})-(\hat{\pi}_{(t)}-\hat{\pi}_{K(t)})] . \tag{2.12}$$

We can describe the expected value of the sum of squares of second term of right side of (2.12) as follows.

$$E\left[\frac{1}{n-1}\sum_{i=1}^{n}\{(\hat{\pi}_{i(t)}-\hat{\pi}_{iK(t)})-(\hat{\pi}_{(t)}-\hat{\pi}_{K(t)})\}^2\right] = \frac{1-f_3}{mk}\frac{((r+w)/r)^2}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}\lambda_{ij(t)}(1-\lambda_{ij(t)}) . \tag{2.13}$$

From (2.11) and (2.13), $E\left[\frac{1}{n-1}\sum_{i=1}^{n}(\hat{\pi}_{i(t)}-\hat{\pi}_{(t)})^2\right]$ can be showed as (2.10).

Similarly, we can deliver the following equations

$$E\left[\frac{1}{n(m-1)}\sum_{i=1}^{n}\sum_{j=1}^{m}(\hat{\pi}_{ij(t)}-\hat{\pi}_{i(t)})^2\right]$$
$$=\frac{1}{N(M-1)}\sum_{i=1}^{N}\sum_{j=1}^{M}(\pi_{ij(t)}-\pi_{i(t)})^2 \tag{2.14}$$
$$+\frac{1-f_3}{k}\frac{((r+w)/r)^2}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}\lambda_{ij(t)}(1-\lambda_{ij(t)}),$$

$$E\left[\frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\hat{\lambda}_{ij(t)}(1-\hat{\lambda}_{ij(t)})\right] \tag{2.15}$$
$$=\frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}\lambda_{ij(t)}(1-\lambda_{ij(t)}).$$

By using the equations (2.10), (2.14) and (2.15), we can obtain (2.9). ■

2.2. numerical example

For example, the office of education in a city of name A wants to know the quantities of smoking that it's high school students did smoke in the school. We further suppose for simplicity that a sample of $n=2$ schools is selected from $N=3$ schools, a sample of $m=2$ classes is selected from $M=3$ classes for each selected school, and finally a sample of $k=4$ students is selected from $K=9$ students for each selected class. Each students in ultimate sample is asked to answer the question selected from the device. Let's assume that students usually smoke between 1 to 3 in the school, and 6 balls of two different colors, 3 red balls and 3 white balls marked on 1, 2 or 3 each, be placed in the device.

red ball : How many smoking did you have in the school?
white ball : Read the number marked on its surface.

Suppose the population and response structure are given as below. Those numbers represent sensitive values of respondents in population, and the numbers shadowed are represent sensitive values of respondents in sample.

<Population and sample>

School (N)	Class (M)	Student (K)	x_{i1l}	x_{i2l}	x_{i3l}
1	$M=3$	$K=9$	$x_{111}=1$	$x_{121}=1$	$x_{131}=1$
			$x_{112}=3$	$x_{122}=1$	$x_{132}=2$
			$x_{113}=3$	$x_{123}=2$	$x_{133}=2$
			$x_{114}=1$	$x_{124}=3$	$x_{134}=3$
			$x_{115}=3$	$x_{125}=3$	$x_{135}=2$
			$x_{116}=2$	$x_{126}=1$	$x_{136}=3$
			$x_{117}=1$	$x_{127}=2$	$x_{137}=3$
			$x_{118}=3$	$x_{128}=2$	$x_{138}=1$
			$x_{119}=2$	$x_{129}=3$	$x_{139}=1$
2	$M=3$	$K=9$	$x_{211}=2$	$x_{221}=1$	$x_{231}=2$
			$x_{212}=2$	$x_{222}=2$	$x_{232}=1$
			$x_{213}=3$	$x_{223}=3$	$x_{233}=2$
			$x_{214}=1$	$x_{224}=1$	$x_{234}=3$
			$x_{215}=2$	$x_{225}=1$	$x_{235}=2$
			$x_{216}=3$	$x_{226}=2$	$x_{236}=1$
			$x_{217}=1$	$x_{227}=1$	$x_{237}=2$
			$x_{218}=3$	$x_{228}=3$	$x_{238}=3$
			$x_{219}=2$	$x_{229}=2$	$x_{239}=1$
3	$M=3$	$K=9$	$x_{311}=1$	$x_{321}=3$	$x_{331}=3$
			$x_{312}=2$	$x_{322}=1$	$x_{332}=1$
			$x_{313}=3$	$x_{323}=2$	$x_{333}=1$
			$x_{314}=1$	$x_{324}=2$	$x_{334}=3$
			$x_{315}=3$	$x_{325}=2$	$x_{335}=3$
			$x_{316}=2$	$x_{326}=1$	$x_{336}=2$
			$x_{317}=2$	$x_{327}=3$	$x_{337}=1$
			$x_{318}=3$	$x_{328}=1$	$x_{338}=2$
			$x_{319}=1$	$x_{329}=2$	$x_{339}=3$

The results of our survey on the sample selected from the above population can be constructed as below.

<The selected questions and the responses>

sample	question selected from the device	randomized response	sample	question selected from the device	randomized response
$x_{112}=3$	red ball	$z_{111}=3$	$x_{212}=2$	red ball	$z_{211}=2$
$x_{114}=1$	white ball marked 1	$z_{112}=1$	$x_{214}=1$	white ball marked 1	$z_{212}=1$
$x_{116}=2$	red ball	$z_{113}=2$	$x_{216}=3$	red ball	$z_{213}=3$
$x_{118}=3$	white ball marked 2	$z_{114}=2$	$x_{218}=3$	white ball marked 2	$z_{214}=2$
$x_{121}=1$	red ball	$z_{121}=1$	$x_{232}=1$	red ball	$z_{221}=1$
$x_{123}=2$	white ball marked 3	$z_{122}=3$	$x_{234}=3$	white ball marked 3	$z_{222}=3$
$x_{125}=3$	red ball	$z_{123}=3$	$x_{236}=1$	red ball	$z_{223}=1$
$x_{127}=2$	white ball marked 2	$z_{124}=2$	$x_{238}=3$	white ball marked 2	$z_{224}=2$

We can calculate parameters and estimators that dealt in previous chapter.

<The population proportion and the variance of $\hat{\pi}_{(t)}$>

The population proportion				The population variance
$\lambda_{11(1)}=0.333$	$\pi_{11(1)}=0.333$	$\pi_{1(1)}=0.333$	$\pi_{(1)}=0.333$	$V(\hat{\pi}_{(1)})=0.0311$
$\lambda_{12(1)}=0.333$	$\pi_{12(1)}=0.333$			
$\lambda_{13(1)}=0.333$	$\pi_{13(1)}=0.333$			
$\lambda_{21(1)}=0.278$	$\pi_{21(1)}=0.222$	$\pi_{2(1)}=0.333$		
$\lambda_{22(1)}=0.389$	$\pi_{22(1)}=0.444$			
$\lambda_{23(1)}=0.333$	$\pi_{23(1)}=0.333$			
$\lambda_{31(1)}=0.333$	$\pi_{31(1)}=0.333$	$\pi_{3(1)}=0.333$		
$\lambda_{32(1)}=0.333$	$\pi_{32(1)}=0.333$			
$\lambda_{33(1)}=0.333$	$\pi_{33(1)}=0.333$			
$\lambda_{11(2)}=0.278$	$\pi_{11(2)}=0.222$	$\pi_{1(2)}=0.296$	$\pi_{(2)}=0.346$	$V(\hat{\pi}_{(2)})=0.0320$
$\lambda_{12(2)}=0.333$	$\pi_{12(2)}=0.333$			
$\lambda_{13(2)}=0.333$	$\pi_{13(2)}=0.333$			
$\lambda_{21(2)}=0.389$	$\pi_{21(2)}=0.444$	$\pi_{2(2)}=0.407$		
$\lambda_{22(2)}=0.333$	$\pi_{22(2)}=0.333$			
$\lambda_{23(2)}=0.389$	$\pi_{23(2)}=0.444$			
$\lambda_{31(2)}=0.333$	$\pi_{31(2)}=0.333$	$\pi_{3(2)}=0.333$		
$\lambda_{32(2)}=0.389$	$\pi_{32(2)}=0.444$			
$\lambda_{33(2)}=0.278$	$\pi_{33(2)}=0.222$			
$\lambda_{11(3)}=0.389$	$\pi_{11(3)}=0.444$	$\pi_{1(3)}=0.370$	$\pi_{(3)}=0.321$	$V(\hat{\pi}_{(3)})=0.0314$
$\lambda_{12(3)}=0.333$	$\pi_{12(3)}=0.333$			
$\lambda_{13(3)}=0.333$	$\pi_{13(3)}=0.333$			
$\lambda_{21(3)}=0.333$	$\pi_{21(3)}=0.333$	$\pi_{2(3)}=0.259$		
$\lambda_{22(3)}=0.278$	$\pi_{22(3)}=0.222$			
$\lambda_{23(3)}=0.278$	$\pi_{23(3)}=0.222$			
$\lambda_{31(3)}=0.333$	$\pi_{31(3)}=0.333$	$\pi_{3(3)}=0.333$		
$\lambda_{32(3)}=0.278$	$\pi_{32(3)}=0.222$			
$\lambda_{31(3)}=0.389$	$\pi_{31(3)}=0.444$			

<The sample proportion and the variance estimator of $\hat{\pi}_{(t)}$>

The sample proportion				The variance estimator
$\hat{\lambda}_{11(1)}=0.25$	$\hat{\pi}_{11(1)}=0.167$	$\hat{\pi}_{1(1)}=0.167$	$\hat{\pi}_{(1)}=0.292$	$\widehat{V}(\hat{\pi}_{(1)})=0.0212$
$\hat{\lambda}_{12(1)}=0.25$	$\hat{\pi}_{12(1)}=0.167$			
$\hat{\lambda}_{21(1)}=0.25$	$\hat{\pi}_{21(1)}=0.167$	$\hat{\pi}_{2(1)}=0.416$		
$\hat{\lambda}_{22(1)}=0.50$	$\hat{\pi}_{22(1)}=0.667$			
$\hat{\lambda}_{11(2)}=0.50$	$\hat{\pi}_{11(2)}=0.667$	$\hat{\pi}_{1(2)}=0.416$	$\hat{\pi}_{(2)}=0.416$	$\widehat{V}(\hat{\pi}_{(2)})=0.0204$
$\hat{\lambda}_{12(2)}=0.25$	$\hat{\pi}_{12(2)}=0.167$			
$\hat{\lambda}_{21(2)}=0.50$	$\hat{\pi}_{21(2)}=0.667$	$\hat{\pi}_{2(2)}=0.416$		
$\hat{\lambda}_{22(2)}=0.25$	$\hat{\pi}_{22(2)}=0.167$			
$\hat{\lambda}_{11(3)}=0.25$	$\hat{\pi}_{11(3)}=0.167$	$\hat{\pi}_{1(3)}=0.416$	$\hat{\pi}_{(3)}=0.292$	$\widehat{V}(\hat{\pi}_{(3)})=0.0212$
$\hat{\lambda}_{12(3)}=0.50$	$\hat{\pi}_{12(3)}=0.667$			
$\hat{\lambda}_{21(3)}=0.25$	$\hat{\pi}_{21(3)}=0.167$	$\hat{\pi}_{2(3)}=0.167$		
$\hat{\lambda}_{22(3)}=0.25$	$\hat{\pi}_{22(3)}=0.167$			

From the above table, we can see that each value of $\pi_{(t)}$ is estimated as $\hat{\pi}_{(t)}$ that is, $\pi_{(1)}=0.333$ as $\hat{\pi}_{(1)}=0.292$, $\pi_{(2)}=0.346$ as $\hat{\pi}_{(2)}=0.416$, $\pi_{(3)}=0.321$ as $\hat{\pi}_{(3)}=0.292$. In this example the variance estimators are appeared to underestimate the population variances. But it is not general case because our situation is only one of them.

3. The optimum values of sub-sample sizes of m and k

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

In this section we are to determine the optimum values of m and k given a specified cost.

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

$$C' = C - c_0 = c_1 n + c_2 nm + c_3 nmk . \tag{3.1}$$

Where C and c_0 represent a total and an overhead cost respectively, and c_1,

c_2, and c_3 represent required costs for obtaining fsu, ssu, and tsu respectively.

We can rewrite the variance of (2.8) as follows

$$V(\hat{\pi}_{(t)}) = \frac{S_a^2}{n} + \frac{S_b^2}{nm} + \frac{S_c^2}{nmk} - \frac{S_1^2}{N}, \tag{3.2}$$

where

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

$$S_a^2 = S_1^2 - \frac{1}{NM(M-1)}\sum_{i=1}^{N}\sum_{j=1}^{M}(\pi_{ij(t)} - \pi_{i(t)})^2,$$

$$S_b^2 = \frac{1}{N(M-1)}\sum_{i=1}^{N}\sum_{j=1}^{M}(\pi_{ij(t)} - \pi_{i(t)})^2 - \frac{S_c^2}{K},$$

$$S_c^2 = \frac{((r+w)/r)^2}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}\lambda_{ij(t)}(1-\lambda_{ij(t)}).$$

Consider the following equation to obtain the optimum value of minimizing variance under the specified cost.

$$\left(V(\hat{\pi}_{(t)}) + \frac{S_1^2}{N}\right)(C - c_0) = \left(S_a^2 + \frac{S_b^2}{m} + \frac{S_c^2}{mk}\right)(c_1 + c_2 m + c_3 mk). \tag{3.3}$$

The optimum values of m and k that minimize (3.3) can be obtained via Cauchy-Schwartz inequality as follows

$$k_{opt} = \frac{S_c}{S_b}\sqrt{\frac{c_2}{c_3}}, \tag{3.4}$$

$$m_{opt} = \frac{S_b}{S_a}\sqrt{\frac{c_1}{c_2}}. \tag{3.5}$$

We can see that the optimum values k_{opt} and m_{opt} depend not only on variance ratio but also on cost ratio. Although the optimum values k_{opt} and m_{opt} can be obtained only when we know the exact values of the cost and variance ratios, we can alternatively obtain them by using the optimal method suggested by Mohammad(1986) when the cost and variance ratios are bounded to some intervals.

The optimum value n_{opt} is obtained by substituting (3.4) and (3.5) into (3.1)

$$n_{opt} = \frac{(C - c_0)\sqrt{S_a^2 / c_1}}{S_a\sqrt{c_1} + S_b\sqrt{c_2} + S_c\sqrt{c_3}} \quad . \tag{3.6}$$

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

$$V_{\min}(\hat{\pi}_{(t)}) = \frac{(S_a\sqrt{c_1} + S_b\sqrt{c_2} + S_c\sqrt{c_3})^2}{C - c_0} - \frac{S_1^2}{N} \quad . \tag{3.7}$$

When N is so large that $\frac{1}{N}$ is ignored the relative efficiency(RE) of m, k under a specified cost is given by

$$RE(m, k \mid m_{opt}, k_{opt}) = \frac{(S_a\sqrt{c_1} + S_b\sqrt{c_2} + S_c\sqrt{c_3})^2}{(c_1 + c_2 m + c_3 mk)\left(S_a^2 + \frac{S_b^2}{m} + \frac{S_c^2}{mk}\right)} \quad . \tag{3.8}$$

Table 1 below gives optimal values for selected values of the parameters. From the equations (3.4), (3.5), and (3.6) we can see that optimal values are determined by several parameters. While it is desirable to induce optimal values by considering the varieties of all the parameters, we observe optimal values according as the number of red ball r varies under the given values.

$$N = 10,\ M = 10,\ K = 50,\ w = 3,$$
$$C = 100,\ c_0 = 10,\ c_1 = 10,\ c_2 = 5,\ c_3 = 1,$$
$$\sum_{i=1}^{N}(\pi_{i(t)} - \pi_{(t)})^2 = 2,\ \sum_{i=1}^{N}\sum_{j=1}^{M}(\pi_{ij(t)} - \pi_{i(t)})^2 = 20,$$
$$\sum_{i=1}^{N}\sum_{j=1}^{M}\lambda_{ij(t)}(1 - \lambda_{ij(t)}) = 10 .$$

<Table 1> Optimal values under various values of r.

r	$\frac{S_c}{S_b}$	$\frac{c_2}{c_3}$	$\frac{S_b}{S_a}$	$\frac{c_1}{c_2}$	n_{opt}	m_{opt}	k_{opt}
3	1.37	5	1.03	2	4.13	1.46	3.06
6	1.02	5	1.04	2	4.34	1.48	2.27
9	0.90	5	1.05	2	4.42	1.48	2.02
12	0.84	5	1.05	2	4.45	1.48	1.89

As can be seen from Table 1, the ultimate sampling unit k_{opt}s are decreasing steadily as the values of r are increasing. This is consistent with the fact that the size of ultimate sampling units by direct question method without randomization device is less than that of indirect method with randomization device.

Optimal values also depend on cost ratios. Table 2 below gives them for various cost ratios. Where $w = r = 3$, $\frac{S_c}{S_b} = 1.37$, and $\frac{S_b}{S_a} = 1.03$.

<Table 2> Optimal values under various values of cost ratio.

c_1	c_2	c_3	$\frac{c_2}{c_3}$	$\frac{c_1}{c_2}$	n_{opt}	m_{opt}	k_{opt}
5	4	3	1.33	1.25	5.96	1.15	1.58
5	4	2	2	1.25	6.38	1.16	1.93
5	4	1	4	1.25	7.04	1.16	2.73
5	2	1	2	2.5	7.87	1.64	1.93
5	1	2	0.5	5	7.63	2.31	1.00

Some problems of choosing integer values may be arise from Table 1 and Table 2. One way of solving those problems is to maintain a fixed standard efficiency by using the Mohammad's optimal method.

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

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 optimum values of m_{opt} and k_{opt} which minimize the cost function of (3.1) under the condition that $V(\hat{\pi}_{(t)}) = V_0$ are

$$k_{opt} = \frac{S_c}{S_b}\sqrt{\frac{c_2}{c_3}}, \tag{3.9}$$

$$m_{opt} = \frac{S_b}{S_a}\sqrt{\frac{c_1}{c_2}}. \tag{3.10}$$

The optimum value n_{opt} is obtained by substituting (3.9) and (3.10) into (3.2)

$$n_{opt} = \frac{S_a\sqrt{c_1} + S_b\sqrt{c_2} + S_c\sqrt{c_3}}{\left(V_0 + \frac{S_1^2}{N}\right)\sqrt{c_1/S_a^2}} \quad . \tag{3.11}$$

The minimum cost function with fixed variance $V(\hat{\pi}_{(t)}) = V_0$ can be 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

$$C = c_0 + \frac{(S_a\sqrt{c_1} + S_b\sqrt{c_2} + S_c\sqrt{c_3})^2}{V_0 + \frac{S_1^2}{N}} \quad . \tag{3.12}$$

4. Conclusions and Discussions

We consider and systemize the theoretical validity for applying three-stage cluster RRM which employ Liu & Chow's technique to the ultimate sampling unit to estimate the sensitive population proportion and variance from a complex population which is composed of several clusters. We derive both optimal values of first sampling unit, second sampling unit, and third sampling unit to minimize variance for a specified cost, and ones to minimize cost function for a specified precision.

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 to estimating sensitive proportion from a complex population including sensitive attributes although extra effort is necessary to get an acceptable information from randomization device.

In a way, 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.

Reference

1. Lee, G. S. and Hong, K. H. (1998). Two-Stage Cluster Randomized Response Model, *The Korean Communications in Statistics*, Vol. 5 No. 1, 99-105.
2. Lee, G. S., Hong, K. H., Son, C. K. and Jung, Y. M. (2003). The Three-Stage Cluster Randomized Response Model for Obtaining

Sensitive Information, *The Korean Communications in Statistics*, Vol. 10 No. 1, 247-256.
3. Chaudhuri, A. and Mukerjee, R. (1988). *Randomized Response : Theory and Techniques*, Marcel Dekker, Inc., New York.
4. Cochran, W. G. (1977). *Sampling Techniques*, 3rd ed. John Wiley and Sons, New York.
5. Greenberg, B. G., Kubler, R. R., Abernathy, J. R., and Horvitz, D. G. (1971). Applications of the RR Technique in Obtaining Quantitative Data, *Journal of the American Statistical Association*, Vol. 66, 243-250.
6. Liu, P. T. and Chow, L. P. (1976). A New Discrete Quantitative Randomized Model, *Journal of the American Statistical Association*, Vol. 71, 72-73.
7. Mohammad, S. A. (1986). The Choice of Subsampling Size in Two-stage Sampling, *Journal of the American Statistical Association*, Vol. 81, 555-558.
8. 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 date : May. 2003, accepted date : Nov. 2003]