Journal of the Korean
Statistical Society
Vol. 25, No. 1, 1996

A Mixed Randomized Response Technique

You Sung Park and Jae Choul Lee[1]

Abstract

A new method that uses information obtained not only from randomization device but also from direct question is introduced. The new maximum likelihood estimator is compared with those of Warner (1965) and Mangat(1994). For a choice of randomized device, we propose a choice depending on the sample size n and show that our estimator is more efficient than that of Mangat under the randomization device. The proposed procedure is extended to more general one which can be easily applied to some specific cases. Under the specified conditions, it is shown that the variance of this generalized estimator is smaller than that of Warner.

Key Words : Randomization device ; Maximum likelihood estimator.

[1] Department of Statistics, Korea University, Seoul, 136-701, Korea.

1. INTRODUCTION

In sample surveys which study certain sensitive problems, many interviewees are likely to provide untruthful answers or to refuse to respond. To reduce evasive biases for estimating the proportion π of the population that has the sensitive attribute A, many authors, including Warner(1965) who made the seminal work known as the randomized response technique, have studied what is called randomization devices. Lately, Mangat(1994) proposed an improved randomized response strategy that is more efficient than Warner's procedure if

$$\pi > 1 - \left\{ \frac{p}{(2p-1)} \right\}^2$$

which always holds for $p > 1/3$.

Provided that all the people answered truthfully to a direct question, one weak point of the randomized response technique is that its estimator has larger variance than that of simple random sampling. In the situation where some people are bold enough to answer frankly the sensitive question, the randomized response can not use valuable information about π that is obtainable from asking direct question. Why not ask directly an embarrassing question? This leads us to consider a new procedure which mixes information obtained from the randomized response and the direct question.

2. DESCRIPTION OF A MIXED TECHNIQUE

A simple random sample of n respondents is drawn with replacement from the population. Each people is required to reply only 'yes' or 'no' to the sensitive question

$$\text{"I have the attribute A".} \tag{A}$$

Since the attribute A is likely to stigmatize him, it is reasonable to assume that there is no interviewee who gives a 'yes' answer without that characteristic A. If a person says 'yes' to the question (A), there is no more procedure for him. Otherwise, he has to give a 'yes' or 'no' answer to the statement that is selected by him with the aid of a randomization device consisting of two mutually exclusive statements

$$\text{"I have the attribute A"} \tag{B}$$

"I do not have the attribute A", (C)

where the chance that the question (B) is chosen equals p. The interviewer does not know which question any interviewee has answered, but know the probability with which the statement (B) is presented.

Assuming x is the unknown proportion of individuals in a population that belong to the characteristic A and reply 'no' to the direct question, we can represent the probabilities of 'yes' answer for this procedure as follows:

$$\begin{aligned} \lambda_1 &= \pi - x \\ \lambda_2 &= xp + (1-\pi)(1-p), \end{aligned}$$

where λ_1 is the probability that a 'yes' answer will be given to the direct question (A) and λ_2 the probability that the interviewee will say 'no' at the direct question (A) and 'yes' at the second question (B) or (C). We further assume that all the individuals required to reply to the outcome of the randomization device tell the truth.

Let n_1 and n_2 denote the number of 'yes' answers to the direct question and the number of 'no' answers to the direct question and 'yes' answers to randomized response, respectively. Since n_1 has a binomial $b(n, \lambda_1)$ and the conditional distribution of n_2 given n_1 is $b(n - n_1, \frac{\lambda_2}{1-\lambda_1})$, $(n_1, n_2, n - n_1 - n_2)$ follows the multinomial distribution with parameters n, λ_1, λ_2, and $1-\lambda_1-\lambda_2$. Thus the ML estimators of π and x can be obtained by

$$\begin{aligned} \hat{\pi} &= \frac{n_2/n + n_1 p/n - (1-p)}{2p-1} \\ \hat{x} &= \frac{n_2/n + n_1(1-p)/n - (1-p)}{2p-1} \end{aligned}$$

which are unbiased, with variances

$$Var(\hat{\pi}) = \frac{\pi(1-\pi)}{n} + \frac{p(1-p)(1-\pi+x)}{n(2p-1)^2}$$

$$Var(\hat{x}) = \frac{x(1-x)}{n} + \frac{p(1-p)(1-\pi+x)}{n(2p-1)^2},$$

respectively, where $p \neq 1/2$. This results in the following theorems.

Theorem 1. The estimator $\hat{\pi}$ of the proposed method has the smaller variance than Warner estimator $\hat{\pi}_W$.

Proof. Since $0 \le \pi - x < 1$,

$$\begin{aligned} Var(\hat{\pi}) &= \frac{\pi(1-\pi)}{n} + \frac{p(1-p)(1-\pi+x)}{n(2p-1)^2} \\ &\le \frac{\pi(1-\pi)}{n} + \frac{p(1-p)}{n(2p-1)^2} = Var(\hat{\pi}_W). \end{aligned}$$

Theorem 2. The variance of $\hat{\pi}$ is smaller than that of Mangat estimator $\hat{\pi}_M$ if

$$\pi < 1 - \frac{xp^2}{(1-p)(1-3p)} \quad \text{for} \quad p < \frac{1}{3}.$$

Proof. The variance of Mangat estimator $\hat{\pi}_M$ given by

$$Var(\hat{\pi}_M) = \frac{\pi(1-\pi)}{n} + \frac{(1-p)(1-\pi)}{np}$$

is larger than that of $\hat{\pi}$ if

$$\pi(1-p)(1-3p) < (1-p)(1-3p) - xp^2.$$

This implies

$$\begin{aligned} \pi &< 1 - \frac{xp^2}{(1-p)(1-3p)} \quad \text{if} \quad p < \frac{1}{3} \\ \pi &> 1 - \frac{xp^2}{(1-p)(1-3p)} \quad \text{if} \quad p > \frac{1}{3}. \end{aligned}$$

But the last inequality always holds, since the right hand side of it is larger than 1 regardless of x. This completes the proof.

3. A CHOICE OF P DEPENDING ON THE SAMPLE SIZE

Starting Warner's work, the probability, p, that the question (B) is chosen in a randomization device has been fixed. But we may consider a choice criteria of p.

Theorem 3. Let p be a function of sample size n, that is, $p \equiv p_n = n^{-\alpha}$, for $0 < \alpha \leq 1$. Then the variance of this estimator $\hat{\pi}$ more rapidly converges to zero with order $O(n^{-1})$ than that of $\hat{\pi}_M$ does with order $O(n^{-(1-\alpha)})$.

Proof. It can be easily shown by (2.1) and (2.2) that

$$\begin{aligned} Var(\hat{\pi}) &= O(n^{-1}) + o(n^{-1}) &= O(n^{-1}) \\ Var(\hat{\pi}_M) &= O(n^{-1}) + O(n^{-(1-\alpha)}) &= O(n^{-(1-\alpha)}). \end{aligned}$$

This theorem tell us that the variance of $\hat{\pi}$ is dominated by sampling error but that of $\hat{\pi}_M$ is done by randomizing device error. In other words, if we choice p proportional to a reciprocal of sample size n, our estimator of π is more efficient than Mangat's estimator. What we want is that the randomizing device error has less influence on variance as possible. These suggested that whenever one takes p_n with an appropriate α instead of a fixed number p, our estimator $\hat{\pi}$ is better than $\hat{\pi}_M$.

4. EXTENSION AND SOME SPECIFIC CASES

There may be people who don't have the attribute A and reply 'yes' to direct question. If y is the unknown proportion of these respondents, the preceding method can be extended to the following one. Suppose that a simple random sample of n respondents is required to be replied by a 'yes' or 'no' to a direct question and to the randomized device with the device probability p. Then it can be shown that

λ_1 = the probability that a 'yes' answer will be reported to a direct question

$= \pi - x + y$

λ_2 = the probability that the interviewee will say 'no' at the direct and give a 'yes' answer to the randomized response

$$
\begin{aligned}
&= xp + (1-\pi-y)(1-p) \\
\lambda_3 &= \text{the probability that the interviewee will say `yes' at the direct} \\
&\quad \text{and give a `yes' answer to the randomized response} \\
&= (\pi - x)p + y(1-p). \qquad (4.1)
\end{aligned}
$$

Let

$$
\begin{aligned}
n_1 &= \text{the number of `yes' answers to the direct question} \\
n_2 &= \text{the number of respondents who say `no' to the direct question} \\
&\quad \text{and say `yes' to the randomized response} \\
n_3 &= \text{the number of respondents who say `yes' to the direct question} \\
&\quad \text{and to the randomized response.} \qquad (4.2)
\end{aligned}
$$

Clearly, $(n_1, n_2, n-n_1-n_2)$ has a multinomial $M(n, \lambda_1, \lambda_2, 1-\lambda_1-\lambda_2)$ distribution and $(n_2, n_3, n-n_2-n_3)$ is $M(n, \lambda_2, \lambda_3, 1-\lambda_2-\lambda_3)$. Before calculating the variances, we first develop the following proposition.

Proposition 4. Let X_i's and Y_i's be *i.i.d.* Bernoulli random variables with parameters λ_1 and λ_2, $i = 1, 2, \cdots, n$, respectively. Suppose that if $Y_i = 1$ then $X_i = 1$, but the converse does not hold. Then

$$
Cov\Big(\sum_{i=1}^{n} X_i, \sum_{i=1}^{n} Y_i\Big) = n\lambda_2(1-\lambda_1).
$$

Proof. The proof is made in the same way developed in that of the Remark (Mood, Graybill, and Boes(1974), pp.507–508, Section 11.2).

It can be shown by (4.1) and (4.2) that the ML estimators of π, x, and y are

$$
\begin{aligned}
\hat{\pi} &= \frac{n_2/n + n_3/n - (1-p)}{2p-1} \\
\hat{x} &= \frac{n_2/n - (1-p)(1-n_1/n)}{2p-1} \\
\hat{y} &= \frac{n_1 p/n - n_3/n}{2p-1}
\end{aligned}
$$

which are all unbiased and, by Proposition 4, in addition to (4.1) and (4.2) the variances are

$$\begin{aligned} Var(\hat{\pi}) &= \frac{\pi(1-\pi)}{n} + \frac{p(1-p)}{n(2p-1)^2} \\ Var(\hat{x}) &= \frac{x(1-x)}{n} + \frac{p(1-p)(1-\pi+x-y)}{n(2p-1)^2} \\ Var(\hat{y}) &= \frac{y(1-y)}{n} + \frac{p(1-p)(\pi-x+y)}{n(2p-1)^2}, \end{aligned}$$

where $p \neq 1/2$. It is very interesting that the variance of $\hat{\pi}$ is the same as that of Warner's estimator. This fact make us to reinterpret the Warner's estimator. Namely, the Warner's estimator for π can be applied when there is certain belief that interviewees are untruthful whether they have a sensitive attribute or not.

We shall consider some specific cases where (i) $x = c$, (ii) $y = c$, (iii) $x = \pi$, and (iv) $y = 1 - \pi$ (c is a known constant).

Case 1 : $x = c$. If we have a prior information about the proportion of respondents who report untruthfully to a direct question, the generalized procedure can be easily reduced to

$$\begin{aligned} \lambda_1 &= \pi - c + y \\ \lambda_3 &= (\pi - c)p + y(1-p). \end{aligned}$$

The ML estimators of π and y are

$$\begin{aligned} \hat{\pi} &= \frac{n_3/n - (1-p)n_1/n}{2p-1} + c \\ \hat{y} &= \frac{n_1 p/n - n_3/n}{2p-1}, \end{aligned}$$

with variances

$$\begin{aligned} Var(\hat{\pi}) &= \frac{\pi(1-\pi)}{n} + \frac{p(1-p)(\pi+y)}{n(2p-1)^2} - \frac{c}{n}\left\{1 - 2\pi + c + \frac{p(1-p)}{(2p-1)^2}\right\} \\ Var(\hat{y}) &= \frac{y(1-y)}{n} + \frac{p(1-p)(\pi+y-c)}{n(2p-1)^2}, \end{aligned}$$

where $p \neq 1/2$, since n_1 and n_3 are dependently binomially distributed with parameters (n, λ_1) and (n, λ_3), respectively.

Case 2 : $y = c$. Let the proportion y of a population not having the stigmatizing characteristic A and giving a 'yes' answer to a direct question be known, that is, $y = c$. Then

$$\begin{aligned} \lambda_1 &= \pi - x + c \\ \lambda_2 &= xp + (1 - \pi - c)(1 - p). \end{aligned}$$

Since $(n_1, n_2, n - n_1 - n_2)$ follows $M(n, \lambda_1, \lambda_2, 1 - \lambda_1 - \lambda_2)$, the ML estimators

$$\begin{aligned} \hat{\pi} &= \frac{n_1 p/n + n_2/n - (1-p)}{2p-1} - c \\ \hat{x} &= \frac{n_2/n + (1-p)n_1/n - (1-p)}{2p-1} \end{aligned}$$

has variances

$$\begin{aligned} Var(\hat{\pi}) &= \frac{\pi(1-\pi)}{n} + \frac{p(1-p)(1-\pi+x-c)}{n(2p-1)^2} + \frac{c(1-c-2\pi)}{n} \\ Var(\hat{x}) &= \frac{x(1-x)}{n} + \frac{p(1-p)(1-\pi+x-c)}{n(2p-1)^2}, \end{aligned}$$

where $p \neq 1/2$. Note that if $c = 0$, then this corresponds to the procedure of Section 2.

Case 3 : $x = \pi$. Under the situation where every people who has the characteristic A tells a lie, $(n_1, n_2, n - n_1 - n_2)$ has $M(n, \lambda_1, \lambda_2, 1 - \lambda_1 - \lambda_2)$ distribution. The reduced procedure can be represented by

$$\begin{aligned} \lambda_1 &= y \\ \lambda_2 &= \pi p + (1 - \pi - y)(1 - p). \end{aligned}$$

The ML estimators of π and y are

$$\begin{aligned} \hat{\pi} &= \frac{(1-p)n_1/n + n_2/n - (1-p)}{2p-1} \\ \hat{y} &= \frac{n_1}{n} \end{aligned}$$

with variances

$$\begin{aligned} Var(\hat{\pi}) &= \frac{\pi(1-\pi)}{n} + \frac{p(1-p)(1-y)}{n(2p-1)^2} \\ Var(\hat{y}) &= \frac{y(1-y)}{n}, \end{aligned}$$

respectively, where $p \neq 1/2$.

Case 4 : $y = 1 - \pi$. In the case where $y = 1 - \pi$, that is, every respondent not having the attribute A reports 'yes' to a direct question, the method is rewritten as follows:

$$\begin{aligned} \lambda_1 &= 1 - x \\ \lambda_3 &= (\pi - x)p + (1-\pi)(1-p) \\ \hat{\pi} &= \frac{n_3/n - n_1 p/n}{2p-1} + 1 \\ \hat{x} &= 1 - \frac{n_1}{n} \\ Var(\hat{\pi}) &= \frac{\pi(1-\pi)}{n} + \frac{p(1-p)(1-x)}{n(2p-1)^2} \\ Var(\hat{x}) &= \frac{x(1-x)}{n}, \end{aligned}$$

where $p \neq 1/2$. Note that n_1 and n_3 follow dependently $b(n, \lambda_1)$ and $b(n, \lambda_3)$ distributions, respectively.

There is often the case where we might make the assumption that the proportion y of individuals not having the attribute A and reporting 'yes' to a direct question is known, that is, $y = c$(Case 2). Under this circumstance, the variance comparison of $\hat{\pi}$ and $\hat{\pi}_W$ can be carried out to obtain the condition in which the former has smaller variance than the latter.

Theorem 5. Let $y = c$. Then $\hat{\pi}$ has smaller variance than Warner estimator $\hat{\pi}_W$ if $c_1 < c < c_2$, where

$$\begin{aligned} c_1 &= \frac{(1-2\pi)(2p-1)^2 - p(1-p)}{2(2p-1)^2} \\ &- \frac{\sqrt{(2p-1)^2\{(1-2\pi)^2(2p-1)^2 - 2p(1-p)(1-2x)\} + p^2(1-p)^2}}{2(2p-1)^2}, \end{aligned}$$

$$
\begin{aligned}
c_2 &= \frac{(1-2\pi)(2p-1)^2 - p(1-p)}{2(2p-1)^2} \\
&+ \frac{\sqrt{(2p-1)^2\{(1-2\pi)^2(2p-1)^2 - 2p(1-p)(1-2x)\} + p^2(1-p)^2}}{2(2p-1)^2},
\end{aligned}
$$

and $c_2 < 1 - \pi$.

REFERENCES

(1) Cochran, W. G.(1977) *Sampling Techniques*, 3rd edition. New York: John Wiley and Sons, Inc.

(2) Kuk, Anthony Y. C.(1990) Asking sensitive question indirectly. *Biometrika*, **77**, 436–438.

(3) Mangat, N. S.(1994) An improved randomized response strategy. *Journal of the Royal Statistical Society, Series B*, **56**, No.1, 93–95.

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

(5) Mood, A. M., Graybill, F. A., and Boes, D. C.(1974) *Introduction to the Theory of Statistics*, 3rd edition. New York: McGraw-Hill.

(6) Warner, S. L.(1965) Randomized response: a survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, **60**, 63–69.