한국통계학회논문집 제3권 2호
1996년 8월 pp. 169-174

The improvement of Mangat Strategy in view of the protection of privacy

Ki-Hak Hong[1], Gi-Sung Lee[2]

Abstract

In the present paper an attempt has been made to improve the Mangat Strategy (1994) in view of the protection of privacy , variance and the range of $\hat{\pi}$. Conditions are obtained under which the proposed model is more efficient than those of Warner (1965) and Mangat and Singh (1990).

1. Introduction

The randomizing response(RR) procedure to procure trustworthy data for estimating the proportion π of the population belonging to a sensitive group was first introduced by Warner (1965). He originally considered the case of estimating proportion of individuals in a population who have some stigmatizing characteristic, e.g. that of being a tax evader. He considered the following maximum likelihood(ML) estimator of π :

$$\hat{\pi}_w = \frac{\frac{n'}{n} - (1-p)}{2p-1} , \qquad p \neq 1/2, \tag{1.1}$$

where n' is the number of 'yes' answers obtained from the n respondents selected by simple random sampling with replacement and p is the proportion of the sensitive character represented in the randomizing device.

The estimator $\hat{\pi}_w$ has been shown to be unbiased and has the variance as follows

$$V(\hat{\pi}_w) = \frac{\pi(1-\pi)}{n} + \frac{p(1-p)}{n(2p-1)^2} . \tag{1.2}$$

1) Assistant Professor, Department of Computer Science and Statistics, Dongshin University, Naju, Chonnam, 520-714, Korea.
2) Assistant Professor, Department of Computer Science and Statistics, Woosuk University, Wanju-gun, Chonbuk, 565-800, Korea.

Since Warner, several other workers have suggested various alternative RR-models. Especially, Mangat and Singh (1990) proposed a two-stage RR model. Mangat (1994) suggested an alternative RR strategy that improved the above Mangat and Singh model under the some conditions. The core of RR is a randomizing device, which gives answers partly random and partly dependent on the value of the respondent's sensitive characteristic.

The idea is that, from the answers, one should be able to obtain estimates of the population parameters without direct knowledge of the sensitive characteristic of the sampled individuals.

Different randomizing device give various degrees of protection to the interviewee's privacy and also allow estimates of different efficiencies.

Leysieffer and Warner (1976) have introduced a quantification based on $p(R \mid A)$ and $p(R \mid A^c)$ to measure jeopardizing response which indicated an invasion of privacy as follows.

$$g(R, A) = \frac{p(R \mid A)}{p(R \mid A^c)} \tag{1.3}$$

for the jeopardy of response (R) with respect to sensitive attribute A and

$$g(R, A^c) = \frac{p(R \mid A^c)}{p(R \mid A)} \tag{1.4}$$

for the jeopardy of R with respect to A^c, with only values g greater than unity indicating R that increase respondent jeopardy from the point of view given by the argument of g.

The function g is called the jeopardy function for the given procedure and its value at the pair (R, A) the level of jeopardy of R with respect to A for the given procedure.

The aim of RR plan is therefore twofold : one is to afford good estimates of the population parameters and the other is to attain a sufficient degree of protection to the interviewee. In this paper, an attempt has been made to improve the Mangat model (1994) in view of the protection of privacy and variance.

2. The Mangat model

Mangat (1994) has proposed a RR procedure. In his method, each interviewee in the simple random sampling with replacement sample of n respondents is instructed to say 'yes' if he or she has the attribute A. If he or she does not have attribute A, the respondent is required to use the Warner randomization device. The probability of a yes answer for this procedure is given by

$$\alpha = \pi + (1-\pi)(1-p). \tag{2.1}$$

The ML unbiased estimator of π and it's variance in this case are respectively given by

$$\hat{\pi}_m = \frac{\frac{n'}{n} - (1-p)}{p}, \tag{2.2}$$

$$V(\hat{\pi}_m) = \frac{\pi(1-\pi)}{n} + \frac{(1-\pi)(1-p)}{np}. \tag{2.3}$$

He compared the efficiency of his estimator with those of Mangat and Singh (1990) and Warner. He showed that his estimator was more efficient than the Mangat and Singh estimator if

$$\pi > 1 - \frac{p(1-T)(1-(1-T)(1-p))}{(2p-1+2T(1-p))^2}, \quad p \neq 1/2$$

and more efficient than the Warner estimator if

$$\pi > 1 - \left(\frac{p}{2p-1}\right)^2, \quad p > 1/3,$$

where T is the probability representing the sensitive attribute in the 1st stage of Mangat and Singh's two stage RR procedure.

Although Mangat has suggested a model that improved Mangat and Singh model and Warner model in some conditions, he did not consider the effect of the protection of privacy and could not overcome the practical defect that the estimator might be taken negative values by p and resulting n'. We consider an alternative model that takes into account both of the protection of privacy and the boundary of the estimator.

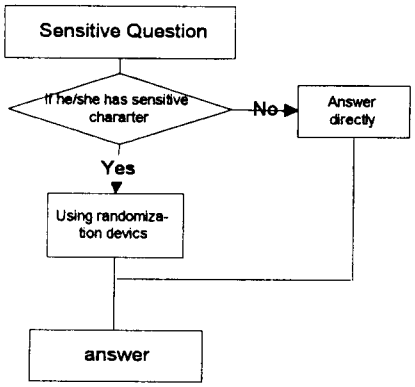
3. The proposed procedure

In this procedure, each of n respondents is instructed to say 'yes' if he or she does not have the sensitive attribute A. If he or she has the attribute A, the respondent is required to use the Warner randomization device consisting of two statements :

(i) ' I belong to attribute A'

(ii) ' I belong to attribute A^c',

represented with probabilities p and $1-p$ respectively. Then he or she is to report 'yes' or 'no' according to the outcome of this randomization device and the actual status that he or she has with respect to attribute A. The Whole procedure is completed by the respondent unobserved by the interviewer.



The probability of a yes answer for this procedure is given by

$$\lambda = (1-\pi) + \pi p. \tag{3.1}$$

Theorem 3.1. The ML unbiased estimator of π and it's variance in this case are respectively given by

$$\hat{\pi}_h = \frac{1-\hat{\lambda}}{1-p}, \tag{3.2}$$

$$V(\hat{\pi}_h) = \frac{\pi(1-\pi)}{n} + \frac{p\pi}{n(1-p)}. \tag{3.3}$$

where $\hat{\lambda} = n'/n$ is the observed proportion of yes answer obtained from the n sampled respondents.

Theorem 3.2. An unbiased estimator of the variance $V(\hat{\pi}_h)$ is

$$v(\hat{\pi}_h) = \frac{\hat{\lambda}(1-\hat{\lambda})}{(n-1)(1-p)^2}. \tag{3.4}$$

Proof. Since $\hat{\lambda} = n'/n$ and n' has the binomial distribution with the sample size n and parameter λ, the expected value of $v(\hat{\lambda})$ on straight forward algebraic simplications is seen to reduce to (3.3). ■

4. Efficiency comparison

In this chapter, we are to compare the efficiency of our proposed model in three ways.

Firstly, if we look at the efficiency in view of the variance, the proposed estimator $\hat{\pi}_h$ will be more efficient than the Mangat estimator $\hat{\pi}_m$ if $V(\hat{\pi}_h) < V(\hat{\pi}_m)$. On using (2.3) and (3.3), the above inequality reduces to

$$p^2(1-2\pi) > 0. \tag{4.1}$$

The above inequality shows that the proposed strategy can always be more efficient than the Mangat strategy by choosing $\pi < 1/2$ for any practical value of p. Since π is the proportion of population members that have the sensitive attribute A, it is reasonable to assume that π is less than $1/2$. In the case of the Warner estimator, our estimator is efficient under the following condition

$$\pi < \left(\frac{1-p}{2p-1}\right)^2, \tag{4.2}$$

which always hold for $p > 2/3$.

Secondly, we appreciate the proposed strategy by using the jeopardy functions described in (1.3) and (1.4). In Mangat model, the jeopardy function of 'yes' with respect to A and the jeopardy function of 'no' with respect to A^c are

$$g_m(y) = g(Y, A) = \frac{1}{1-p}, \tag{4.3}$$

$$g_m(n) = g(N, A^c) \approx \infty. \tag{4.4}$$

The jeopardy functions for 'yes' and 'no' of Warner model are as follows

$$g_w(y) = g(Y, A) = \frac{p}{1-p}, \tag{4.5}$$

$$g_w(n) = g(N, A^c) = \frac{p}{1-p}. \tag{4.6}$$

The jeopardy functions of the proposed model in this paper are as follows

$$g_h(y) = g(Y, A) = p, \tag{4.7}$$

$$g_h(n) = g(N, A^c) = 0. \tag{4.8}$$

In the case of 'yes' answer for attribute A, we can show that from the equations (4.3), (4.5) and (4.7), our proposed model is always more protective than the other two models,

Mangat model and Warner model, irrespective of the value of p while Warner model is more protective than Mangat model if p is less than $1/2$. In the case of 'no' answer for attribute A^c, it is clear that from the equations (4.4), (4.6) and (4.8), our proposed model is always more protective than the other two models, Mangat model and Warner model, while Warner model is more protective than Mangat model irrespective of the value of p.

Finally, we appreciate our estimator $\hat{\pi}_h$ in terms of the range of it. From the equations (1.1), (2.2) and (3.2), we can show that both of the Mangat estimator and Warner estimator may have negative values or have above one values depending on the values of p and $\hat{\lambda}$ i.e. n', on the other hand our estimator has never such values irrespective of the values of p and $\hat{\lambda}$.

References

[1] Chaudhuri, A. and Mukerjee, R. (1988). *Randomized response : theory and techniques*, Marcel Dekker, Inc., New York.

[2] Leysieffer, F.W. and Warner, S.L. (1976). Respondent jeopardy and optimal designs in RR models, *Journal of the American Statistical Association*, Vol. 72, 649-656.

[3] Mangat, N.S. (1994). An improved randomized response strategy, *Journal of the Royal Statistical Society : series B*, 56, 93-95.

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

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