STATISTICA, anno LXIII, n. 4, 2003

PROTECTION OF PRIVACY WITH OBJECTIVE PRIOR DISTRIBUTION IN RANDOMIZED RESPONSE

M. Ruiz Espejo, H. P. Singh

1. INTRODUCTION

In surveys of human population, it may happen that one of the characteristics under study is very perplexing or stigmatizing. It is natural that a respondent may be hesitant or prevaricate in providing information on such characteristic(s) which may show deviation from a social or legal norm and which he feels may be used against him after sometime.

In such situation, to improve respondent cooperation and to encourage truthful answers, Warner (1965) suggested a procedure called randomized response technique (RRT). The RRT due to Warner (1965), for asking a question resulting from the randomization between two equally probable questions, is analysed.

The question of interest (or of study) is the 'intimate' and the question alternative to the 'intimate' is another whose probability (or proportion) of afirmative (or yes) answer is known and objective a priori. The proportion of afirmative answer to the intimate question is the study parameter and the aim is to estimate it.

The RR interviewing technique is used as a means of reducing the bias of deliberate or false responses due to traditional and direct ask approach to an intimate question and the parameter can be estimated with great protection of interviewee's privacy. The privacy is ensured since the interviewer does not know which question any respondent answered and thus the veracity of the randomized responses is assumed complete for the sample selected from the population.

Both questions (intimate and alternative) are randomly selected, and in our case such mechanism consists in the result of a throw of a perfect coin with equiprobable sides.

Objective Bayesian methodology with objective prior distribution and Hartley entropy are used to compare the uncertainty decrease due to RR with respect to the one existing if the intimate question is formulated directly.

2. OBJECTIVE BAYESIAN MODEL

Consider a simple random sample of size n drawn from the population with replacement sampling design. Let X, Y and Z be the random variables defined by $X = 1$ (if the respondent reports 'yes' to the 'intimate' question) and $X = 0$ (if the respondent reports 'no' to the 'intimate' question), $Y = 1$ (if the respondent reports 'yes' to the 'non intimate' question) and $Y = 0$ (if the respondent reports 'no' to the 'non intimate' question), and $Z = 1$ (if the respondent reports 'yes' to the randomized question) and $Z = 0$ (if the respondent reports 'no' to the randomized question). We designate the probabilities $A = p(X = 0)$ and $1 - A = p(X = 1)$ which are the basis for our study and inference. The probability $B = p(Y = 0)$ for a collection of available questions is known, for example, $B = 0.1, 0.2, ..., 0.9$. It is observed that the sample proportion c of negative answers is an unbiased estimator of the probability $C = p(Z = 0)$.

We assume that the probability of selecting the 'intimate' or 'non intimate' question are a priori objective and equal to $P = ½$. Then, we have

$$C = PA + (1 - P)B = (A + B)/2$$

or

$$A = 2C - B.$$

Thus the estimate of A is defined by

$$a = 2c - B$$

with variance

$$V(a) = V(2c - B) = 4V(c) = 4C(1 - C)/n = (A + B)(2 - A - B)/n.$$

From a traditional inferential viewpoint, this variance is maximized for $B = 1 - A$. If $A = ½$, $B = ½$; if A increases (resp. decreases), B decreases (resp. increases). But these properties which are unrecommendable from an efficiency viewpoint, are recommendable for protecting the interviewee's privacy as we can see via maximization of the conditional entropies.

For this end, the following probabilities are obtained by Bayes theorem:

$$p(X = 0 | Z = 0) = \frac{A(1+B)/2}{A(1+B)/2+(1-A)B/2} = A(1 + B)/(A + B),$$

since

$$p(Z = 0 | X = 0) = p(Z = 0 | X = 0 \text{ and 'intimate'})p(\text{'intimate'})$$
$$+ p(Z = 0 | X = 0 \text{ and 'non intimate'})p(\text{'non intimate'}) = (1 + B)/2,$$

and similarly it can be seen that

$$p(Z = 0 \,|\, X = 1) = B/2.$$

Further, since

$$p(Z = 1 \,|\, X = 0) = (1 - B)/2 \quad \text{and} \quad p(Z = 1 \,|\, X = 1) = (2 - B)/2,$$

it can be obtained that

$$p(X = 1 \,|\, Z = 0) = B(1 - A)/(A + B),$$
$$p(X = 0 \,|\, Z = 1) = A(1 - B)/(2 - A - B)$$

and

$$p(X = 1 \,|\, Z = 1) = (2 - B)(1 - A)/(2 - A - B).$$

3. ENTROPIES AND COMPARISONS

In this section we shall compare the Hartley entropies. ‘A priori’ Hartley entropy is

$$H(X) = -p(X = 0) \log p(X = 0) - p(X = 1) \log p(X = 1)$$
$$= -A \log A - (1 - A) \log (1 - A) = \log \{A^{-A}(1 - A)^{A-1}\},$$

and ‘a posteriori’

$$H(X|Z) = p(Z = 1)\, H(X|Z = 1) + p(Z = 0)\, H(X|Z = 0)$$
$$= C\,\{-p(X = 1 \,|\, Z = 1) \log p(X = 1 \,|\, Z = 1)$$
$$-p(X = 0 \,|\, Z = 1) \log p(X = 0 \,|\, Z = 1)\}$$
$$+ (1 - C)\{-p(X = 1 \,|\, Z = 0) \log p(X = 1 \,|\, Z = 0)$$
$$-p(X = 0 \,|\, Z =0) \log p(X = 0 \,|\, Z = 0)\}$$
$$= C\,[-\{A(1 - B)/(A + B)\} \log \{A(1 + B)/(A + B)\}$$
$$-\{B(1 - A)/(A + B)\} \log \{B(1 - A)/(A + B)\}]$$
$$+ (1 - C)[-\{A(1 - B)/(2 - A - B)\} \log \{A(1 - B)/(2 - A - B)\}$$
$$-\{(2 - B)(1 - A)/(2 - A - B)\} \log \{(2 - B)(1 - A)/(2 - A - B)\}].$$

From the information theory we have $H(X|Z) \le H(X)$ with equality if and only if X and Z are independent random variables. But our objective is to maximize $H(X|Z)$ for several values of A and B. We have computed entropies $H(X|Z)$ for (fixed) A = 0.2, 0.5 and 0.8, and (variable for each fixed A) B = 0.1(0.1)0.9; and tabled in Table 1.

TABLE 1

Entropies $H(X|Z)$ for variable A and B

B	$A = 0.2$	$A = 0.5$	$A = 0.8$
0.1	0.162	0.228	0.175
0.2	0.171	0.237	0.178
0.3	0.176	0.241	0.180
0.4	0.179	0.243	0.181
0.5	0.180	0.244	0.180
0.6	0.181	0.243	0.179
0.7	0.180	0.241	0.176
0.8	0.178	0.237	0.171
0.9	0.175	0.228	0.162

Table 1 exhibits that for $A = 0.2$ (0.5 or 0.8) we have the maximum values of $H(X|Z)$ whose are 0.181, 0.244 and 0.181 respectively, and they correspond to $B = 0.6$ (0.5 or 0.4) respectively. The maximum values of $H(X|Z)$ for fixed A, are surligned in the interior of the table, which corresponds to different values of B. These values indicate the maximum protection of interviewee's privacy interviewed by RRT with equiprobable questions. Concretely for $A = 0.1(0.1)0.9$, it has been researched by Ruiz Espejo (1988). In our example, for $A = 0.1$, 0.2 and 0.3, the optimal protection is given for $B = 0.6$; for $A = 0.4$, 0.5 and 0.6, the value $B = 0.5$ is optimal; for $A = 0.7$, 0.8 and 0.9, the value $B = 0.4$ is optimal. Since $B = p(Y = 0)$, when the value A is low ($A \leq 3.5$) the value of B must be approximately 0.6; when $3.5 < A < 6.5$ the value of B must be approximately 0.5; and when $6.5 \leq A$, B must be approximately 0.4. These privacity approximately optimal values of B can be used from three selected 'non intimate' and available questions with $B \cong 0.4$, 0.5 and 0.6, for the RRT described in this article.

Departamento de Matemáticas Fundamentales
Universidad Nacional de Educación a Distancia, Madrid
MARIANO RUIZ ESPEJO

School of Studies in Statistics
Vikram University, Ujjain
HOUSILA P. SINGH

REFERENCES

M. RUIZ ESPEJO, (1988), *Study of models of applied sampling*, Doctoral Thesis, Universidad Complutense, Madrid.

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

RIASSUNTO

Protezione della privacy con distribuzioni a priori oggettive in risposte randomizzate

Usando la metodologia Baysiana oggettiva e l'entropia di Hartley, si analizza il metodo di risposta randomizzata con due questioni, "personale" e "alternativa non personale" le cui probabilità di risposta (sì o no) sono rispettivamente “non note” e “note”. La massimizzazione dell’entropia condizionata è il criterio per la scelta di una domanda alternativa per proteggere la privacy di un soggetto intervistato, per domande “non personale”.

SUMMARY

Protection of privacy with objective prior distribution in randomized response

Using objective Bayesian methodology and Hartley entropy, we analyse the method of randomized response with two questions, “intimate” and “alternative non intimate”, whose probabilities of response (Yes or No) are “unknown” and “known” respectively. A maximization of conditioned entropy is the criterion for selecting the alternative question to protect the privacy of the interviewed people, for several given “non intimate” questions.