Metrika (2003) 57: 195–200

Metrika

A generalized randomized response technique

Tasos C. Christofides[1]

[1] Department of Mathematics and Statistics, University of Cyprus, P.O. Box 20537,
1678 Nicosia, CYPRUS

Abstract. To eliminate a major source of bias in surveys of human populations resulting from respondents refusal to cooparate in cases where a question of sensitive nature is involved, the idea of "randomized response" was introduced by Warner (1965). In this paper, an alternative randomized response technique is presented which improves upon the pioneering work of Warner (1965). The procedure includes Warner's method as a special case for a specific choice of the parameters. In addition, a generalization of the proposed method is presented.

AMS classification: 62D05

Key words: randomized response, estimation of a proportion, respondent protection, sampling

1 Introduction

A major source of bias in surveys of human populations results from the refusal of individuals to respond to certain questions of sensitive nature. In addition, some of those responding provide untruthful answers. This is something which is to be expected because it is very difficult to persuade interviewees that their responses to questions, for example about sexual behaviour, drug abuse or tax evasion will only be used for statistical purposes. To extract indirectly the desired information, Warner (1965) proposed a procedure termed "randomized response". Very briefly the technique is as follows: Suppose we want to estimate the proportion θ of people belonging to a certain group A. Each individual is required to respond "yes" or "no" to one of two statements:

(a) I am a member of group A
(b) I am not a member of group A.

The interviewee responds to statement (a) with probability p and to statement (b) with probability $1-p$ using a random device, e.g., by the toss of a (biased) coin and in the absence of the interviewer. Therefore, the individual does not reveal whether he/she is a member of group A or not. The parameter θ is estimated based on the indirect responses of all individuals via the estimator

$$\hat{\theta} = \frac{p-1}{2p-1} + \frac{n_1}{(2p-1)n} \tag{1.1}$$

where n_1 is the number of individuals responding "yes". The estimator is unbiased with variance given by the expression

$$Var(\hat{\theta}) = \frac{1}{n}\theta(1-\theta) + \frac{0.25}{n}[(2p-1)^{-2} - 1]. \tag{1.2}$$

A similar procedure is the unrelated question model proposed by Greenberg et al. (1969). According to this model one of the two statements is chosen to have no relation to the stigmatizing characteristic. Many other authors such as Horvitz et al. (1967), Mangat and Singh (1990), Kuk (1990), and Chaudhuri (2001) proposed similar models in order to improve the efficiency of the estimators and the level of respondent cooperation or to allow unequal probabilities of selection.

2 The proposed procedure

Suppose we want to estimate the population proportion θ of individuals possessing a certain sensitive characteristic. We select from the population a simple random sample with replacement of size n. Each person sampled is provided with a random device which produces the integers $1, \ldots, L$ with frequencies $p_1, \ldots, p_L$ respectively. For example the device might be a deck of M cards with exactly Mp_j of those cards showing the integer j, $j = 1, \ldots, L$, or perhaps the device is a die with L faces, each one showing one of the integers $1, \ldots, L$ with probability $p_1, \ldots, p_L$ respectively. Using the random device in the absence of the interviewer, the individual produces one of these numbers and he/she reports how far away this number is from $L+1$ if he/she has the characteristic or from 0 if he/she does not have it. The information provided to the interviewer is not sufficient to determine whether the individual possesses the characteristic or not. For instance, suppose that L is 8 and the individual reports the number 5. This means that either the individual has the characteristic and the number produced is 4 or the individual does not have the characteristic and the number produced is 5.

The procedure is very easy to apply. The interviewee is only required to report the difference of two nonnegative integers, something that every individual participating in a survey is expected to be capable of.

To formulate the procedure mathematically, let x_i take on the value $L+1$ if individual i has the characteristic and the value 0 if not. Clearly $P(x_i = L+1) = \theta$ and $P(x_i = 0) = 1 - \theta$. Let y_i be the integer produced by individual i using the random device. Then the number reported is d_i where $d_i = |x_i - y_i|$. Then

$$P(d_i = k) = (1 - \theta)p_k + \theta p_{L+1-k}, \quad k = 1, \dots, L.$$

Direct calculation shows that

$$E(d_i) = \sum_{k=1}^{L} kp_k + \theta\left(L + 1 - 2\sum_{k=1}^{L} kp_k\right).$$

Observe that $\sum_{k=1}^{L} kp_k = E(y)$ where y is a random variable identically distributed with the random observations $y_1, \dots, y_n$. Thus

$$E(d_i) = E(y) + \theta(L + 1 - 2E(y)). \tag{2.1}$$

Similarly, one can verify that

$$Var(d_i) = Var(y) + \theta(1 - \theta)(L + 1 - 2E(y))^2. \tag{2.2}$$

Let $\bar{d}$ denote the sample average of $d_1, \dots, d_n$. Define the estimator

$$\hat{\theta} = (\bar{d} - E(y))(L + 1 - 2E(y))^{-1} \tag{2.3}$$

provided that $L + 1 - 2E(y) \neq 0$. Then it is easily verified that $\hat{\theta}$ is an unbiased estimator of θ with variance

$$\begin{aligned} Var(\hat{\theta}) &= Var(\bar{d})(L + 1 - 2E(y))^{-2} \\ &= \frac{1}{n} Var(d_1)(L + 1 - 2E(y))^{-2} \\ &= \frac{1}{n}\theta(1 - \theta) + \frac{1}{n} Var(y)(L + 1 - 2E(y))^{-2}. \end{aligned} \tag{2.4}$$

Observe that the first term of the right hand side of (2.4) is the variance due to random sampling and the second term is the variance due to the randomized procedure.

Remark. Being a sample average by construction, for large n the estimator $\hat{\theta}$ has approximately a normal distribution and estimation of θ by means of a confidence interval follows in the usual way.

3 Choice of the random device

Considering Warner's estimator, from (1.2) one can see that the second term gets smaller as p moves away from 0.5. However, values of p close to 0 or 1 are not desirable as the cooperation of the interviewee would be in jeopardy since questions would be raised whether the procedure really protects him/her from disclosing the confidential information.

Using our procedure, it follows from (2.4) that the variance of the estimator depends on the choice of the random device via the parameters L, $p_1, \ldots, p_L$. One can verify that for $L = 2$ and $p_1 = p$ (and trivially $p_2 = 1 - p$), the right hand side of (2.4) is the same as the right hand side of (1.2), something which is expected since the mechanism of our procedure in this special case is the same as that of Warner's. This implies that by suitably choosing the parameters $L, p_1, \ldots, p_L$ with $L \geq 3$, we can construct an estimator given by (2.3) which will have smaller MSE than the estimator in (1.1). For instance, for $p = 0.6$ the value of the quantity $0.25[(2p - 1)^{-2} - 1]$ appearing in the right hand side of (1.2) is 6 whereas the value for the quantity $Var(y)(L + 1 - 2E(y))^{-2}$ appearing in the right hand side of (2.4) for $L = 6$ and $(p_1, \ldots, p_6) = (0.26, 0.05, 0.1, 0.19, 0.02, 0.38)$ is 3.76. Since all other terms in (1.2) and (2.4) are the same, it follows that the estimator resulting from our procedure has smaller variance than the estimator in (1.1).

Remark. The quantity in the right hand side of (2.4) representing the variance due to the randomized procedure can be written as

$$\begin{aligned} Var(y)(L + 1 - 2E(y))^{-2} &= \frac{1}{4}\, Var(y)\left(\frac{L+1}{2} - E(y)\right)^{-2} \\ &= \frac{1}{4}\, Var\left(y - \frac{L+1}{2}\right)\left(E\left(y - \frac{L+1}{2}\right)\right)^{-2} \\ &= \frac{1}{4}(CV(z))^2, \end{aligned}$$

where $CV(z)$ is the coefficient of variation of the random variable $z \equiv y - (1/2)(L + 1)$. For theoretical interest, one could consider the problem of choosing the distribution of z in such a way that the coefficient of variation is minimum.

4 A modified procedure

Suppose now that individual i is asked to use the random device m_i times and each time to report the distance of the number produced from $L + 1$ or 0 depending on whether he/she has the characteristic or not. The m_i repetitions of the procedure must be independent from each other. Let y_{ij} be the number produced by individual i at the jth time he/she uses the device and let d_{ij} be the number reported. Let

$$\hat{\theta}_{m.} = (\bar{d}_{m.} - E(y))(L + 1 - 2E(y))^{-1} \tag{4.1}$$

where

$$\bar{d}_{m.} = \left(\sum_{i=1}^{n} m_i\right)^{-1} \sum_{i=1}^{n} \sum_{j=1}^{m_i} d_{ij},$$

and y has the same distribution as the y_{ij}'s, $j = 1, \ldots, m_i$, $i = 1, \ldots, n$. Then $\hat{\theta}_m$ is unbiased for θ. The variance of the estimator follows easily from the variance of $\overline{d}_m$. To that end,

$$\begin{aligned}
Var(\overline{d}_{m.}) &= Var\left(\left(\sum_{i=1}^{n} m_i\right)^{-1} \sum_{i=1}^{n}\sum_{j=1}^{m_i} d_{ij}\right)\\
&= \left(\sum_{i=1}^{n} m_i\right)^{-2} \sum_{i=1}^{n} Var\left(\sum_{j=1}^{m_i} d_{ij}\right)\\
&= \left(\sum_{i=1}^{n} m_i\right)^{-2} \sum_{i=1}^{n}\left(\sum_{j=1}^{m_i} Var(d_{ij}) + 2\sum_{1\le s<t\le m_i} Cov(d_{is}, d_{it})\right). \qquad (4.2)
\end{aligned}$$

Calculating the covariances appearing in the right hand side of (4.2) we have that

$$\begin{aligned}
Cov(d_{is}, d_{it}) &= E(|x_i - y_{is}|\,|x_i - y_{it}|) - E(|x_i - y_{is}|)E(|x_i - y_{it}|)\\
&= E((x_i - y_{is})(x_i - y_{it})) - (E(y) + \theta(L + 1 - 2E(y)))^2 \qquad (4.3)\\
&= E(x_i^2) - E(x_i(y_{is} + y_{it})) + E(y_{is}y_{it})\\
&\quad - (E(y) + \theta(L + 1 - 2E(y)))^2\\
&= \theta(L+1)^2 - 2\theta(L+1)E(y) + (E(y))^2\\
&\quad - (E(y) + \theta(L + 1 - 2E(y)))^2 \qquad (4.4)\\
&= \theta(1-\theta)(L + 1 - 2E(y))^2, \qquad (4.5)
\end{aligned}$$

where the first term of (4.3) follows from the fact that $x_i - y_{is}$ and $x_i - y_{it}$ have the same sign, the second term of (4.3) from (2.1), and (4.4) from the independence of x_i, y_{is}, y_{it}. From (2.2) and (4.5) we have that

$$\begin{aligned}
&\sum_{j=1}^{m_i} Var(d_{ij}) + 2\sum_{1\le s<t\le m_i} Cov(d_{is}, d_{it})\\
&\quad = m_i\, Var(y) + m_i\theta(1-\theta)(L + 1 - 2E(y))^2\\
&\qquad + m_i(m_i - 1)\theta(1-\theta)(L + 1 - 2E(y))^2\\
&\quad = m_i\, Var(y) + m_i^2\theta(1-\theta)(L + 1 - 2E(y))^2.
\end{aligned}$$

Substituting the previous expression in (4.2) we get

$$Var(\bar{d}_{m.}) = \left(\sum_{i=1}^{n} m_i\right)^{-2} \sum_{i=1}^{n} (m_i\, Var(y) + m_i^2 \theta(1-\theta)(L+1-2E(y))^2)$$

$$= \left(\sum_{i=1}^{n} m_i\right)^{-1} Var(y) + \left(\sum_{i=1}^{n} m_i\right)^{-2} \left(\sum_{i=1}^{n} m_i^2\right) \theta(1-\theta)(L+1-2E(y))^2,$$

and finally from (4.1) it follows that

$$Var(\hat{\theta}_{m.}) = \left(\sum_{i=1}^{n} m_i\right)^{-2} \left(\sum_{i=1}^{n} m_i^2\right) \theta(1-\theta) + \left(\sum_{i=1}^{n} m_i\right)^{-1} Var(y)(L+1-2E(y))^{-2}. \tag{4.6}$$

It is evident from (4.6) that by allowing each individual to use the random device more than once we can have more control over the precision of the estimator. Although the coefficient of $\theta(1-\theta)$ in the right hand side of (4.6) is greater than or equal to $1/n$, the coefficient of $Var(y)(L+1-2E(y))^{-2}$ can be considerably less than $1/n$. However, in the special case where $m_i = m$, $i = 1, \ldots, n$, i.e., each respondent is asked to use the random device m times, (4.6) reduces to

$$Var(\hat{\theta}_m) = \frac{1}{n}\theta(1-\theta) + \frac{1}{mn}\, Var(y)(L+1-2E(y))^{-2}. \tag{4.7}$$

Comparing (4.7) and (2.4) we can see that the variance of the estimator can be improved by allowing multiple use of the random device. It has to be pointed out however, that for practical purposes this option cannot be abused as the respondents might not like the idea to use the random device over and over again.

References

[1] Chaudhuri A (2001) Using randomized response from a complex survey to estimate a sensitive proportion in a dichotomous finite population. J. Statist. Plann. Inference 94:37–42

[2] Greenberg BG, Abul-Ela ELA, Simmons WR, Horvitz DG (1969) The unrelated question randomized response model: theoretical framework. J. Amer. Stat. Assoc. 64:520–539

[3] Horvitz DG, Shah BV, Simmons WR (1967) The unrelated question randomized response model. Proc. Social Statist. Sect., Am. Statist. Assoc. 65–72

[4] Kuk AYC (1990) Asking sensitive questions inderictly. Biometrika 77:436–438

[5] Mangat NS, Singh R (1990) An alternative randomized response procedure. Biometrika 77:439–442

[6] Warner SL (1965) Randomized response: a survey technique for eliminating evasive answer bias. J. Amer. Stat. Assoc. 60:63–69