

On Mixing Direct and Randomized Response Procedures

Rajiv Saksena, N. Rastogi and R. Karan Singh

Department of Statistics, Lucknow University, Lucknow- 226 007

(Received July 8, 2002)

Abstract: For the estimation of the proportion of individuals having a sensitive character, a procedure based on direct and randomized responses is proposed and its properties are analyzed. The procedure is general in the sense that the results of procedures based on direct responses and randomized responses may be seen as its special cases.

Key words: Direct responses, randomized responses, mixed responses.

1. Introduction

Consider a population consisting the proportion π_A of individuals having a sensitive character A and the proportion $1 - \pi_A$ of individuals having the character $\bar{A}$. For the estimation of the proportion π_A , Warner¹ developed a procedure known as Randomized Response Technique (RRT) based on selecting a simple random sample with replacement of n individuals from the population, providing a randomized response device for all the individuals in the sample and asking each of the individuals to answer in either 'Yes' or 'No' if the randomized response device dealing with a random experiment results in the outcome according to individual's actual status or contrary to the individual's actual status. For further details, see Sukhatme, Sukhatme, Sukhatme and Asok². Warner's RRT is based on the fact that no respondent is ready to give direct answer but it may be possible that due to changes in socio-economic behaviour and living patterns, modernization, deviations from orthodoxy and other developments, some of the respondents may prefer to give direct answers truthfully. In such situations, it is advisable to deal with a procedure which utilizes both, the direct answers and the answers obtained by using RRT. Here, it is an attempt to solve the problem of estimation of π_A in this spirit and direction along with giving an idea about the changes in the attitude of the individuals towards giving direct answers or not giving direct answers.

2. The Estimation Procedure and Related Results

Let π_D and $\pi_{\bar{D}}$ be the proportions of individuals giving 'direct' and 'not direct' answers in the population respectively. Also, let π_{AD} , $\pi_{\bar{A}D}$, $\pi_{A\bar{D}}$ and $\pi_{\bar{A}\bar{D}}$ be the

population proportions of individuals 'giving direct answers and having the sensitive character A ', 'giving direct answers and not having the sensitive character A ', 'not giving direct answers and having the sensitive character A ' and 'not giving direct answers and not having the sensitive character A ' respectively. It is clear that

$$\begin{aligned}\pi_D + \pi_{\bar{D}} &= 1, & \pi_D &= \pi_{AD} + \pi_{\bar{A}D}, & \pi_{\bar{D}} &= \pi_{A\bar{D}} + \pi_{\bar{A}\bar{D}}, \\ \pi_A &= \pi_{AD} + \pi_{A\bar{D}}, & \pi_{AD} + \pi_{\bar{A}D} + \pi_{A\bar{D}} + \pi_{\bar{A}\bar{D}} &= 1, \\ \pi_A + \pi_{\bar{A}D} + \pi_{\bar{A}\bar{D}} &= 1 \text{ and } 1 - \pi_A &= \pi_{\bar{A}D} + \pi_{\bar{A}\bar{D}}.\end{aligned}$$

Further, let n_D and $n_{\bar{D}}$ be the number of individuals giving 'direct' and 'not direct' answers in the simple random sample with replacement (SRSWR) of size n respectively, and n_{AD} , $n_{\bar{A}D}$ be the number of individuals 'having the sensitive character A and giving direct answers', 'not having A and giving direct answers' in SRSWR of size n respectively. It may be easily seen that n_D , $n_{\bar{D}}$, n_{AD} and $n_{\bar{A}D}$ follow binomial distributions with parameters (n, π_D) , $(n, \pi_{\bar{D}})$, (n, π_{AD}) and $(n, \pi_{\bar{A}D})$ respectively.

Now, in SRSWR of n individuals, n_D individuals give direct answers saying 'Yes' or 'No' according to their actual status having the sensitive character A or not having A , and $n_{\bar{D}}$ individuals (not giving direct answers) are asked to follow Warner's randomized response technique. In this set-up dealing with direct answers and answers under Warner's randomized response device consisting of two statements (i) I have the attribute A and (ii) I have not the attribute A with probabilities p and $1-p$ respectively and individual's reporting 'Yes' or 'No' according to the outcome of the randomized device and the actual status that individual has with respect to A , the probability of a 'Yes' answer denoted by θ is

$$\begin{aligned}(2.1) \quad \theta &= \text{Prob. (a 'Yes')} \\ &= \pi_D \cdot \pi_{A|D} + \pi_{\bar{D}} \{p \cdot \pi_{A|\bar{D}} + (1-p) \pi_{\bar{A}|\bar{D}}\},\end{aligned}$$

where (i) $\pi_{A|D}$, (ii) $\pi_{A|\bar{D}}$ and (iii) $\pi_{\bar{A}|\bar{D}}$ are the conditional probabilities of (i) getting an individual with A given D , (ii) getting an individual with A given $\bar{D}$ and (iii) getting an individual with $\bar{A}$ given $\bar{D}$ respectively.

Simplifying (2.1), we have

$$\begin{aligned}(2.2) \quad \theta &= \pi_{AD} + \{p \pi_{A\bar{D}} + (1-p) \pi_{\bar{A}\bar{D}}\} \\ &= \pi_{AD} + \theta_{\bar{D}} \quad (\text{where } \theta_{\bar{D}} = p \pi_{A\bar{D}} + (1-p) \pi_{\bar{A}\bar{D}}) \\ &= \pi_{AD} + p \pi_{A\bar{D}} + (1-p)(\pi_{\bar{D}} - \pi_{A\bar{D}}) \\ &= \pi_{AD} + (2p-1) \pi_{A\bar{D}} + (1-p) - (1-p) \pi_D.\end{aligned}$$

Let n_{AD} be the number of direct 'Yes' answers (number of individuals having A in n_D individuals giving direct answers) in n_D and $n_{Y\bar{D}}$ be the number of 'Yes' answers under the randomized response device in $n_{\bar{D}}$ so that in the entire sample of size n , the

number (n_1) of 'Yes' answers becomes $n_1 = n_{AD} + n_{Y\bar{D}}$ following the binomial distribution with parameters (n, θ) . Thus the unbiased estimator of θ is given by

$$(2.3) \quad \hat{\theta} = \frac{n_1}{n} = \frac{n_{AD} + n_{Y\bar{D}}}{n}.$$

Let $n_{N\bar{D}}$ be the number of 'No' answers under the randomized response device in $n_{\bar{D}}$ so that

$$(2.4) \quad n_{AD} + n_{\bar{A}D} + n_{Y\bar{D}} + n_{N\bar{D}} = n,$$

where n_{AD} , $n_{\bar{A}D}$, $n_{Y\bar{D}}$, $n_{N\bar{D}}$ jointly follow the multinomial distribution with

$$(2.5) \quad \begin{cases} E\left(\frac{n_{AD}}{n}\right) = \pi_{AD}, \quad E\left(\frac{n_{\bar{A}D}}{n}\right) = \pi_{\bar{A}D}, \quad E\left(\frac{n_{Y\bar{D}}}{n}\right) = \theta_{\bar{D}}, \quad E\left(\frac{n_{N\bar{D}}}{n}\right) = 1 - \theta_{\bar{D}}, \\ V\left(\frac{n_{AD}}{n}\right) = \frac{\pi_{AD}(1 - \pi_{AD})}{n}, \quad V\left(\frac{n_{\bar{A}D}}{n}\right) = \frac{\pi_{\bar{A}D}(1 - \pi_{\bar{A}D})}{n}, \quad V\left(\frac{n_{Y\bar{D}}}{n}\right) = \frac{\theta_{\bar{D}}(1 - \theta_{\bar{D}})}{n}, \\ Cov\left(\frac{n_{AD}}{n}, \frac{n_{\bar{A}D}}{n}\right) = -\frac{\pi_{AD}\pi_{\bar{A}D}}{n}, \quad Cov\left(\frac{n_{AD}}{n}, \frac{n_{Y\bar{D}}}{n}\right) = -\frac{\pi_{AD}\theta_{\bar{D}}}{n} \text{ and} \\ Cov\left(\frac{n_{\bar{A}D}}{n}, \frac{n_{Y\bar{D}}}{n}\right) = -\frac{\pi_{\bar{A}D}\theta_{\bar{D}}}{n}. \end{cases}$$

Substituting the unbiased estimators $\hat{\theta} = \frac{n_{AD} + n_{Y\bar{D}}}{n}$, $\hat{\pi}_{AD} = \frac{n_{AD}}{n}$ and $\hat{\pi}_D = \frac{n_D}{n}$ of θ , π_{AD} and π_D respectively in (2.2) and solving for $\pi_{A\bar{D}}$, we get the unbiased estimator $\hat{\pi}_{A\bar{D}}$ as

$$(2.6) \quad \begin{aligned} \hat{\pi}_{A\bar{D}} &= \frac{\hat{\theta} - \hat{\pi}_{AD} - (1-p) + (1-p)\hat{\pi}_D}{(2p-1)} \\ &= \frac{\left(\frac{n_{AD} + n_{Y\bar{D}}}{n}\right) - \left(\frac{n_{AD}}{n}\right) - (1-p) + (1-p)\left(\frac{n_D}{n}\right)}{(2p-1)} \\ &= \frac{\left(\frac{n_{Y\bar{D}}}{n}\right) - (1-p) + (1-p)\left(\frac{n_{AD} + n_{\bar{A}D}}{n}\right)}{(2p-1)}, \quad p \neq 1/2. \end{aligned}$$

Since $\pi_A = \pi_{AD} + \pi_{A\bar{D}}$, unbiased estimator of π_A is given by

$$(2.7) \quad \hat{\pi}_A = \hat{\pi}_{AD} + \hat{\pi}_{A\bar{D}}$$

$$= \frac{n_{AD}}{n} + \frac{\left(\frac{n_{YD}}{n}\right) - (1-p) + (1-p)\left(\frac{n_{AD} + n_{AD}}{n}\right)}{(2p-1)}.$$

Further, variance of $\hat{\pi}_A$ is

$$(2.8) \quad V(\hat{\pi}_A) = V\left[\frac{n_{AD}}{n} + \frac{\left(\frac{n_{YD}}{n}\right) + (1-p)\left(\frac{n_{AD} + n_{AD}}{n}\right)}{(2p-1)}\right]$$

$$= \frac{\pi_{AD}(1-\pi_{AD})}{n} + \frac{1}{n(2p-1)^2} \left[\theta_D(1-\theta_D) + (1-p)^2 \{ \pi_{AD}(1-\pi_{AD}) \right.$$

$$+ \pi_{AD}(1-\pi_{AD}) - 2\pi_{AD}\pi_{AD} \} - 2(1-p) \{ \pi_{AD}\theta_D + \pi_{AD}\theta_D \}]$$

$$+ \frac{2}{n(2p-1)} \left[-\pi_{AD}\theta_D + (1-p) \{ \pi_{AD}(1-\pi_{AD}) - \pi_{AD}\pi_{AD} \} \right].$$

Some Remarks

(a) As Warner¹ uses randomized response device for all the individuals, it seems that Warner's procedure has the presumption that the population consists of individuals not giving direct answers, that is $\pi_D = 1$, $\pi_D = \pi_{AD} + \pi_{AD} = 0$ implying

$$\pi_{AD} = \pi_{AD} = 0, \quad \pi_A = \pi_{AD} + \pi_{AD} = \pi_{AD}, \quad 1 - \pi_A = \pi_{AD} + \pi_{AD} = \pi_{AD} \text{ and}$$

$$\theta_D = p\pi_{AD} + (1-p)\pi_{AD} = p\pi_A + (1-p)(1-\pi_A) = \theta^* \text{ (say).}$$

Substituting all these values in (2.8), we get

$$(2.9) \quad V(\hat{\pi}_A) = \frac{1}{n(2p-1)^2} \theta^*(1-\theta^*) = \frac{\pi_A(1-\pi_A)}{n} + \frac{p(1-p)}{n(2p-1)^2},$$

which is the same expression of the variance of the estimator for Warner's procedure. Also, for Warner's RRT, $n_D = 0$ implying $n_{AD} = n_{AD} = 0$ which when substituted in (2.7), we see that the estimator $\hat{\pi}_A$ developed here reduces to Warner's estimator. Thus, Warner's study may be seen as the special case of the present study.

(b) For the case of all the individuals giving direct answers, we have $\pi_D = 1$, $\pi_D = \pi_{AD} + \pi_{AD} = 0$ implying that

$$\pi_{AD} = \pi_{AD} = 0, \quad \theta_D = p\pi_{AD} + (1-p)\pi_{AD} = 0, \quad \pi_A = \pi_{AD} + \pi_{AD} = \pi_{AD} \text{ and}$$

$1 - \pi_A = 1 - \pi_{AD} = \pi_{\bar{A}D} + \pi_{\bar{A}\bar{D}} = \pi_{\bar{A}D}$. Substituting all these values in (2.8), we see that $V(\hat{\pi}_A)$ reduces to

$$(2.10) \quad V(\hat{\pi}_A) = \frac{\pi_A(1 - \pi_A)}{n},$$

which is the same variance-expression as obtained in case of direct answers. Further, in case of direct answers, $n_{\bar{D}} = n_{Y\bar{D}} + n_{N\bar{D}} = 0$ implying $n_{Y\bar{D}} = n_{N\bar{D}} = 0$, $n = n_{AD} + n_{\bar{A}D} + n_{Y\bar{D}} + n_{N\bar{D}} = n_{AD} + n_{\bar{A}D}$. Substituting all these values in (2.7), we get

$$\hat{\pi}_A = \frac{n_{AD}}{n},$$

which is the same usual estimator in case of direct answers. Thus the findings based on direct answers may also be seen as the special case of this study.

(c) In case of various mixed population containing individuals with direct answers and randomized responses, the variance-expression given by (2.8) will lie between (2.9) and (2.10). Hence the suggested procedure utilizing both types of responses, direct and randomized, is an improvement over Warner's RRT in such situations.

References

1. S.L. Warner: *J. Amer. Statist. Assoc.* (1965) 60-63.
2. P.V. Sukhatme, B.V. Sukhatme, S. Sukhatme and C. Asok: *Sampling Theory of Surveys with Applications*, Iowa State University Press, Ames (U.S.A.) and Indian Society of Agricultural Statistics, New Delhi, 1984.

