Survey Methodology, June 1995
Vol. 21, No. 1, pp. 21–23
Statistics Canada

On Efficiency of Using Distinct Respondents in a Randomized Response Survey

N.S. MANGAT, R. SINGH, S. SINGH, D.R. BELLHOUSE and H.B. KASHANI[1]

ABSTRACT

It is well known that the sample mean based on the distinct sample units in simple random sampling with replacement is more efficient than the sample mean based on all units selected including repetitions (Murthy 1967, pp. 65-66). Seth and Rao (1964) showed that the mean of the distinct units is less efficient than the sample mean in sampling without replacement under the same average sampling cost. Under Warner's (1965) method of randomized response we compare simple random sampling without replacement and sampling with replacement when only the distinct number of units in the sample are considered.

KEY WORDS: Simple random sampling with and without replacement; Inferences with distinct units; Warner's technique.

1. INTRODUCTION

The randomized response (RR) technique to procure trustworthy data for estimating the proportion of the population belonging to a sensitive group was first introduced by Warner (1965). Since then many developments have taken place in this area. Recently, among others, Franklin (1989), Kuk (1990), Mangat and Singh (1990, 1991), Mangat, Singh and Singh (1992) and Mangat (1994) have suggested alternative RR procedures/estimators.

In the usual simple random sampling (SRS) with replacement (WR) surveys, it is well known that the estimator of population mean based on the distinct units is always more efficient than the mean based on all selections (Murthy 1967, pp. 65-66). Also, Seth and Rao (1964) showed that, under the same average cost to sample, sampling without replacement was more efficient than with replacement sampling using the mean of the distinct sample units. This motivated the authors to investigate whether the above observation also holds in the case of Warner's pioneer RR model which is widely used in practice for selecting the respondents in the case of a survey dealing with sensitive characters. To investigate the problem we shall consider the use of four sampling strategies.

1.1 Strategy I

According to this (Warner's) procedure, each respondent included in the sample using the SRSWR method is provided with a suitable randomization device consisting of two statements of the form: (i) "I belong to sensitive group" and (ii) "I do not belong to sensitive group", represented with probabilities p and $(1 - p)$, respectively. The respondent answers "yes" or "no" according to the randomly selected statement and to his actual status with respect to the attribute, without revealing the statement chosen. If n' persons in the sample (including repetitions) answered "yes", Warner's estimator

$$\hat{\pi} = \frac{n'/n - 1 + p}{2p - 1}, \quad p \neq .5, \tag{1}$$

is unbiased for π and its variance is given by

$$V_1(\hat{\pi}) = \frac{\pi(1 - \pi)}{n} + \frac{p(1 - p)}{n(2p - 1)^2}. \tag{2}$$

The value of p should be chosen as close to 1 or 0 as possible without threatening the degree of co-operation by respondents.

1.2 Strategy II

A sample of n respondents is drawn from a finite population of N units using SRSWR but the information from the d distinct units in the sample, $1 \leq d \leq n$, is used in the construction of the estimator. Let d' denote the respondents reporting a "yes" answer in the interview conducted with the RR device. We then consider the following estimator for π:

$$\hat{\pi}_d = \frac{d'/d - 1 + p}{2p - 1}, \quad p \neq .5. \tag{3}$$

Conditional on d distinct units, the resulting sample is a simple random sample without replacement of size d from N units. The estimator $\hat{\pi}_d$ is, therefore, unbiased for the population π.

[1] N.S. Mangat, R. Singh and S. Singh, Punjab Agricultural University, Ludhiana-141004 (India); D.R. Bellhouse, University of Western Ontario, London, Ontario, Canada, N6A 5B7; H.B. Kashani, West Oregon State College, Monmouth, OR 97361, U.S.A.

In order to study the performance of the proposed estimator $\hat{\pi}_d$, we need its variance. We give here the expression for the conditional variance $V_2(\hat{\pi}_d)$ for a given value of d. Thus

$$V_2(\hat{\pi}_d) = \frac{N-d}{N-1}\,\frac{\pi(1-\pi)}{d} + \frac{p(1-p)}{d(2p-1)^2}. \quad (4)$$

If E_1 and V_1 are the expectation and variance over all values of d, then we have $V_{\text{II}}(\hat{\pi}_d) = E_1V_2(\hat{\pi}_d) + V_1E_2(\hat{\pi}_d)$. On using (4) one gets

$$V_{\text{II}}(\hat{\pi}_d) = \left[NE_1\left(\frac{1}{d}\right) - 1\right]\frac{\pi(1-\pi)}{N-1} + \frac{p(1-p)}{(2p-1)^2}E_1\left(\frac{1}{d}\right) \quad (5)$$

since the second term in $V_{\text{II}}(\hat{\pi}_d)$ is zero as $E_2(\hat{\pi}_d) = \pi$.

1.3 Strategy III

The sample of n respondents is selected using SRSWOR (Kim and Flueck 1978). In this case the variance of the estimator $\hat{\pi}$ in (1) can be written by replacing d in (4) by n. Thus we have

$$V_{\text{III}}(\hat{\pi}) = \frac{N-n}{N-1}\,\frac{\pi(1-\pi)}{n} + \frac{p(1-p)}{n(2p-1)^2}. \quad (6)$$

1.4 Strategy IV

Here the estimator is based on a WOR simple random sample of size $E(d)$. This yields the same expected cost for both in SRSWR and SRSWOR. For this scheme the estimator will be

$$\hat{\pi}_E = \frac{d'/E(d) - 1 + p}{2p-1}, \quad p \neq .5$$

with variance

$$V_{\text{IV}}(\hat{\pi}_E) = \frac{N/E(d) - 1}{N-1}\pi(1-\pi) + \frac{p(1-p)}{E(d)(2p-1)^2}. \quad (7)$$

2. EFFICIENCY COMPARISONS

It has been shown by Korwar and Serfling (1970) that, for $n \geq 3$,

$$Q - \frac{1}{720N} < E\left(\frac{1}{d}\right) \leq Q$$

where

$$Q = \frac{1}{n} + \frac{1}{2N} + \frac{n-1}{12N^2}.$$

Let us now examine the variance expression in (5). Using Q, it is easily verified that

$$\frac{NE_1(1/d) - 1}{N-1} \leq \frac{1}{n}, \quad (8)$$

in the first term on the right of (5) but that $E_1(1/d) \geq 1/n$ in the second term on the right of (5). Thus the relative efficiency of the SRSWR estimator in (1) using repeated units with respect to the SRSWR estimator in (3) using the distinct number of units will depend on the relative sizes of π and p. This is due to the fact that the repeated units can give rise to different responses because of the randomizing device and hence can provide some additional information. A sufficient condition for the inequality $V_{\text{II}}(\hat{\pi}_d) - V_{\text{I}}(\hat{\pi}) < 0$ to hold is obtained by using $E_1(d) = Q$. Thus we get the condition as

$$\pi(1-\pi) > \frac{n(N-1)(6N+n-1)}{N\{6Nn - 12N - n(n-1)\}}\,\frac{p(1-p)}{(2p-1)^2}. \quad (9)$$

The above inequality is likely to hold for values of p closer to 0 or 1, the situations in which respondent jeopardy would be of concern. For example, if $N = 100$, $n = 10$ and $p = 0.9$, the inequality (9) will hold for $0.236 \leq \pi \leq 0.764$.

Similarly, Strategy II will be inferior to Strategy I if $V_{\text{II}}(\hat{\pi}_d) - V_{\text{I}}(\hat{\pi}) > 0$. Using $E_1(1/d) = Q - 1/720N$ this inequality reduces to

$$\pi(1-\pi) < \frac{n(N-1)\{359N + 60(n-1)\}}{N\{361Nn - 720N - 60n(n-1)\}}\,\frac{p(1-p)}{(2p-1)^2}.$$

This inequality will hold for the example considered for inequality (9) whenever either $\pi \leq 0.234$ or $\pi \geq 0.764$.

On using the Cauchy-Schwarz inequality, $E(1/d) > 1/E(d)$, as in Seth and Rao (1964) we find that $V_{\text{II}}(\hat{\pi}_d) > V_{\text{IV}}(\hat{\pi}_E)$. This implies that Strategy IV is more efficient than Strategy II.

It is trivial to note that Strategy III is more efficient than Strategy I.

We know that $E(1/d) \geq 1/n$. This means $V_{\text{II}}(\hat{\pi}_d) > V_{\text{III}}(\hat{\pi})$, implying that Strategy III is more efficient than Strategy II.

Since $1/E(d) \geq 1/n$, Strategy III is more efficient than Strategy IV.

The last pair to consider consists of Strategies I and IV. Since $E(1/d) > 1/E(d)$ for $n > 1$, on using (8) we have

$$\frac{N/E(d) - 1}{N - 1} \leq \frac{1}{n}$$

implying that in (7) and (2)

$$\frac{N - E(d)}{N - 1} \frac{\pi(1 - \pi)}{E(d)} \leq \frac{\pi(1 - \pi)}{n}$$

for $n > 1$. Also $1/E(d) \geq 1/n$. This shows that the second term of (7) on the right hand side will be more than the corresponding term of (2). Thus the relative efficiencies of Strategies I and IV depend on relative values of π and p. As a numerical illustration, if $N = 100$, $n = 10$ and $p = 0.9$ then Strategy IV will be more efficient than Strategy I for $0.18 \leq \pi \leq 0.82$.

REFERENCES

FRANKLIN, L.A. (1989). Randomized response sampling from dichotomous populations with continuous randomization. *Survey Methodology*, 15, 225-235.

KIM, J.-I., and FLUECK, J.A. (1978). Modifications of the randomized response technique for sampling without replacement. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 346-350.

KORWAR, R.M., and SERFLING, R.J. (1970). On averaging over distinct units in sampling with replacement. *Annals of Mathematical Statistics*, 41, 2132-2134.

KUK, A.Y.C. (1990). Asking sensitive questions indirectly. *Biometrika*, 77, 436-438.

MANGAT, N.S. (1994). An improved randomized response strategy. *Journal of the Royal Statistical Society*, Series B, 56, 93-95.

MANGAT, N.S., and SINGH, R. (1990). An alternative randomized response procedure. *Biometrika*, 77, 439-442.

MANGAT, N.S., and SINGH, R. (1991). An alternative approach to randomized response survey. *Statistica*, anno LI, 327-332.

MANGAT, N.S., SINGH, R., and SINGH, S. (1992). An improved unrelated question randomized response strategy. *Calcutta Statistical Association Bulletin*, 42, 277-281.

MURTHY, M.N. (1967). *Sampling Theory and Methods*, Calcutta, India: Statistical Publishing Society.

SETH, G.R., and RAO, J.N.K. (1964). On the comparison between simple random sampling with and without replacement. *Sankhyā* (A), 26, 85-86.

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