Untruthful answering in repeated randomized response procedures

K. KRISHNAMCOURTH and D. RAGHAVARAO

University of Southwestern Louisiana and Temple University

Key words and phrases: Binomial distribution, discordant answers, normal distribution, randomized response, sample size.

AMSA 1985 subject classifications: 62D05.

ABSTRACT

A simple and clear procedure is given that allows to some degree the estimation and testing of the possibility of an untruthful answer in addition to estimating the population proportion by using Warner's (1965) randomized response technique twice.

RÉSUMÉ

Cet article présente une procédure simple et claire qui permet, dans une certaine mesure, l'estimation et la vérification de la probabilité d'une réponse erronée. De plus, cette méthode permet également d'estimer la proportion de la population en utilisant deux fois la méthode de réponse aléatoire de Warner (1965).

1. INTRODUCTION

The randomized response procedure introduced by Warner (1965) is a device for eliciting information on sensitive questions without exposing the true identity of the respondents. The monograph of Chaudhuri and Mukerjee (1988) gives a good exposition of this and related work.

Essentially applying Warner's technique twice, a procedure is given here to test the hypothesis of truthful answers. It is reasonably assumed that the respondents answer independently with the same degree of truthfulness on the two occasions on which different randomization devices are used. Although the test procedure is applicable more generally, a simplistic assumption is made that the probabilities of untruthful answers are the same for people in the sensitive and nonsensitive categories for estimation purposes.

This common probability is estimated by the method of moments. Finally, the proportion of persons belonging to the sensitive category is estimated by the maximum-likelihood method and also from the (yes, yes) answers.

2. TEST FOR TRUTHFUL RESPONSES

Let each individual in the population choose either from the sensitive category (C) or from the nonsensitive category (C'). Let p be the probability that an individual belongs to C. A simple random sample with replacement of n respondents will be taken, and each respondent is given two random devices R1 and R2. Each random device involves two statements equivalent to (1) "I belong to C" and (2) "I do not belong to C". The random device R1 selects the statement (i) with probability p, i = 1, 2. Without revealing
the statement responded, the respondent will answer a yes or so to $R_1$ and another yes or no to $R_2$.

It is possible that people may not truthfully respond to the selected statement. In this section, two cases of untruthful answering by the respondents are considered: (1) untruthful answering depends on the population to which the interviewee belongs, and (2) untruthful answering depends on whether a stigmatizing or non-stigmatizing answer is required. Interestingly, it turns out that (2) is a particular case of (1). One may question whether the probability of untruthful answering remains the same on the two occasions of this procedure, but as it is felt that interviewers who lie first will also lie the second time, the proportion of untruthful answers remains the same on both occasions. Thus the degree of truthfulness is assumed the same in both instances.

In case (1), let $L_1$ and $L_2$ be the probability that a person belonging to $C (\bar{C})$ lies, that is,

$$L_i = P(\text{untruthful answer}|C), \quad L_2 = P(\text{untruthful answer}|\bar{C}).$$

Define $Y_i = 1$ (0) according as the ith respondent responds yes (no) to $R_1$, and $Y_i = 1$ (0) according as the ith respondent responds yes (no) to $R_2$. Then, assuming that $X_i$ and $Y_i$ are independently distributed,

$$P(X_i = 1, Y_i = 0) = \pi P(X_i = 1, Y_i = 0|C) + \pi P(X_i = 1, Y_i = 0|\bar{C})$$

$$= \pi (p_1 L_1 + p_1 L_2)(p_2 L_1 + p_2 L_2)$$

$$+ \pi (p_1 L_2 + p_1 L_1)(p_2 L_2 + p_2 L_1)$$

where $\pi = 1 - \pi$, $L_1 = 1 - L_2$, and $p_i = 1 - p_i$, $i = 1, 2$. Similarly,

$$P(X_i = 0, Y_i = 1) = \pi (p_1 L_1 + p_1 L_2)(p_2 L_1 + p_2 L_2)$$

$$+ \pi (p_1 L_2 + p_1 L_1)(p_2 L_2 + p_2 L_1)$$

A respondent is considered to be giving discordant responses if $X_i = 1, Y_i = \bar{y}$ or $X_i = 0, Y_i = 1$. Then

$$P(\text{discordant response}) = \theta_1 + 2(1 - 2\pi)(1 - 2\pi) (\pi_1 L_1 + \pi_2 L_2)$$

$$= \theta_1 L_1 L_2, \quad \text{say.} \quad (2.1)$$

where $\theta_1 = p_1 + p_2 - 2p_1 p_2$. Note that $\theta(L_1, L_2) = \theta(L_1, L_2) = \theta(L_1, L_2)$, and $\theta(0, 0) = \theta_0$.

In case (2), let $L = P(\text{untruthful answer}|\text{interviewee chooses statement (1) and has to answer yes}) = P(\text{untruthful answer}|\text{interviewee chooses statement (2) and has to answer no})$. It can be shown in this case also that the probability of discordant response is (2.1) with $L_1 = L$ and $L_2 = \bar{L}$. Here people belonging to $C \bar{C}$ never have to give untruthful answers.

Noting that for $p_1, p_2 < 0.5$, testing

$$H_0 : L_1 = L_2 = 0 \quad \text{vs.} \quad H_a : \bar{y} < L_1 < 1, 0 < L_2 < 1$$

is equivalent to testing

$$H_0 : \theta(L_1, L_2) = \theta_0 \quad \text{vs.} \quad H_a : \theta(L_1, L_2) \leq 0.5, \quad (2.2)$$
a test for (2.2) will be presented.
Let \( N_1, N_0, n_1, \) and \( n_0 \) be the numbers of respondents who answered (yes, yes), (yes, no), (no, yes), and (no, no) respectively for \( R_1 \) and \( R_2 \). When \( L_1 = 0 \) and \( L_2 = 0 \), \( n_0 + n_0 \) follows a binomial distribution with parameters \( n \) and \( \theta_0 \). Thus, the critical region for testing (2.2) is
\[
N_0 > n_0, \quad n_0 > n_0, \quad \text{where} \quad P(N_0 > n_0) = \alpha.
\]
Asymptotically for large \( n \), the critical region for testing (2.2) is
\[
\left( \frac{n_0 + n_0}{n} - \theta_0 \right) \sqrt{\frac{\theta_0(1 - \theta_0)}{n}} > z_\alpha,
\]
where \( z_\alpha \) is the \( 100(1 - \alpha) \) percentile point of standard normal distribution. For an alternative test procedure see Lakshmi and Raychaudhuri (1992).

3. ESTIMATION OF PARAMETERS
One can explicitly write the likelihood function for the multinomial case and estimate the parameters of interest by iterative procedures. However, closed analytical expressions can be given for certain special cases and will be discussed in this section.

When \( 2L_0 \) is rejected, the probability for untruthful answering can be obtained assuming \( L_1 = L_2 = L \). Since
\[
L = 0.5 \pm \sqrt{0.25 - \frac{(n_0 + n_0)(n - n_0)}{2(1 - 2p)(1 - 2p)'}}.
\]

one may estimate \( L \) by
\[
L = 0.5 \pm \sqrt{0.25 - \frac{(n_0 + n_0)(n - n_0)}{2(1 - 2p)(1 - 2p)'}}.
\]

Noting that \( (n_0 + n_0)/n \) \( \rightarrow \theta_0 \) and
\[
0.25 - \frac{(n_0 + n_0)(n - n_0)}{2(1 - 2p)(1 - 2p)'} \leq 0.25 - L(1 - L) \geq 0,
\]

\( L \) always exists for large \( n \). The approximate variance of \( L \) is
\[
Var L = \frac{\theta_0(1 - \theta_0)}{4n[(1 - 2p)(1 - 2p)']^2} \left( 1 - 4L(1 - L) \right).
\]

When \( M \) is retained and \( p_1 = p_2 = p \), the maximum-likelihood estimator of \( \pi \) can easily be derived as
\[
\hat{\pi} = \frac{[1 - p^2 + p^2'n_0]}{(1 - 2p)[n_0 + n_1]} = \frac{p^2}{1 - 2p'},
\]

Putting
\[
\theta_1 = P(X = 1, Y = 1) = (1 - p)^2 + \theta_1(1 - 2p),
\]
\[
\theta_0 = P(X = 0, Y = 0) = p^2 + \pi(1 - 2p')
\]

Putting
\[
\theta_1 = P(X = 1, Y = 1) = (1 - p)^2 + \theta_1(1 - 2p),
\]
\[
\theta_0 = P(X = 0, Y = 0) = p^2 + \theta_0(1 - 2p')
\]
clearly,

$$E[R_{ii}] = \frac{(1 - p_i^2 + p_i^2)}{1 - 2p} \theta_i \left( \frac{n_0 + n_i}{n_0 + n_{i1}} \right) \left( \frac{n_0 + n_{i1} + N - n}{1 - 2p} \right) - \frac{p_i^2}{1 - 2p}$$

and

$$Var[R_{ii}] = \frac{1}{1 - 2p^2} \left( \frac{\theta_{ii}}{(n + 1)(\theta_0 + \theta_{i1}) - 1} \right)^2 \theta_{ii}$$

using an approximation for the inverse moment of a binomial distribution given by Grab and Savage (1954).

When $H_0$ is rejected, an estimator of $\pi$ can be obtained by considering the expected value of $n_{i1}/n$ and is

$$\hat{\pi} = \frac{n_{i1}/n}{1 + \hat{L} \hat{L}} = \frac{n_{i1}/n - (\hat{p}_1 \hat{L} + \hat{p}_2 \hat{L} + \hat{p}_1 \hat{L} + \hat{p}_2 \hat{L})}{1 - 2\hat{L} \hat{L}}$$

where $\hat{L} = 1 - \hat{L}$. The exact variance of (3.1) is difficult to compute. However, an approximate variance can be computed using $\hat{L}$ as fixed and is

$$Var[\hat{\pi}] = \frac{\hat{\pi}(1 - \hat{\pi})}{n(1 - 2\hat{L} \hat{L} + \hat{p}_1 \hat{p}_2 - 1)}$$

where $\hat{\pi} = \frac{n_{i1}/n - (\hat{p}_1 \hat{L} + \hat{p}_2 \hat{L} + \hat{p}_1 \hat{L} + \hat{p}_2 \hat{L})}{1 - 2\hat{L} \hat{L} + \hat{p}_1 \hat{p}_2 - 1}$.

ACKNOWLEDGMENTS

The authors wish to thank the referees and an associate editor for many helpful comments in improving the original submission of this paper. The work of the first author was partially supported financially by a Laura H. Cornell professorship fund for the second author.

REFERENCES


Received 30 January 1981
Revised 3 November 1982
Accepted 12 November 1982