NEW RANDOMIZED RESPONSE TECHNIQUES FOR CONTROL OF NONSAMPLING ERRORS IN SURVEYS

by

W. B. Barksdale

McDonnell Douglas Automation Company
McDonnell Douglas Corporation
Long Beach, CA 90846

ABSTRACT

Randomized response survey techniques were previously developed to estimate frequencies of single stigmatizing characteristics while enabling the respondent to protect his self-image without reporting untruthfully. This paper extends the technique to permit estimation of two attributes one or both of which may be stigmatizing. The correlation coefficient and its variance is defined. The study includes the definition of a generalized mean square error measure and consideration of bias, variance, choice of sample size and randomizing proportions. The consequence of this study demonstrates that appreciable gains in mean square error efficiency are achieved when the characteristics in question are positively correlated.

INTRODUCTION

Randomized response survey techniques have been in existence for some 11 years commencing with the work by S. L. Warner.[4] Abul-Ela,[1,3] Simmons,[3] Greenberg,[1,2,3] Horvitz,[1,2,3] Abernathy,[2] and others have added to these techniques with theoretical and field application experience. This paper continues the development of Warner's idea along lines first apparent to Greenberg. The basic developments that preceded this work are briefly reviewed. The extended alternate question model is presented and is compared to the direct question approach after a basis for comparison is established.

METHODS FOR ESTIMATING THE PROPORTION OF A POPULATION WITH A SINGLE SENSITIVE CHARACTERISTIC

DIRECT QUESTION

One can immediately recognize the difficulties with this approach. Namely, one can expect that some portion of those in the survey will refuse to answer the sensitive question and another portion will answer untruthfully. Interviewers may also be reticent to question in the sensitive areas. In this model the probability of a "yes" response to a direct question about the presence of trait A is given by

$$\lambda_D = [(1 - R_A)(1 - T_{A1})\pi_A + (1 - R_a)T_{A2}(1 - \pi_A)] \cdot n'/n$$

where

R_A, R_a are the proportions with and without characteristic A, respectively, who refuse to answer, T_{A1}, T_{A2} are the proportions with and without characteristic A, respectively, who answer untruthfully, π_A is the true proportion of the population with characteristic A, and n, n′ are the sizes of the questioned and answering samples. In the event of complete and true reporting λ_D is distributed as $b(n, \pi_A)$.

WARNER MODEL

Assuming that persons in a population belong to one of two mutually exclusive groups, a randomizing device is used to permit a respondent to select, unseen by the interviewer, whether he replies to a question about membership in the group characterized by attribute A or the complementary group a. The probability described above is given here for Warner's model by

$$\lambda_W = [P(1 - R_A)(1 - T_{A1})\pi_A + (1 - P)(1 - R_a)(1 - T_{A2})(1 - \pi_A) + P(1 - R_a)T_{A2}(1 - \pi_A) + (1 - P)(1 - R_A)T_{A1}\pi_A] \cdot n'/n$$

where, in addition to the definitions stated above, P is the proportion of the time one can expect the question about characteristic A to be selected. Warner's results indicate that gains in efficiency (mean square error) are obtained with the randomized response technique over the direct question when P and π_A differ from 0.5 and untruthfulness is present in the direct question approach.

ALTERNATE QUESTION MODEL

In this model, initially suggested by W. R. Simmons, the question about the mutually exclusive trait is replaced by a question about a second trait, denoted by π_B, presumably innocuous. Two samples are required in order to estimate the two unknown proportions. The probabilities in this model are given by

$$\lambda_1 = P_1(1 - T_{A1})\pi_A + (1 - P_1)(1 - T_{B1})\pi_B + P_1T_{A2}(1 - \pi_A) + (1 - P_1)T_{B2}(1 - \pi_B)$$

$$\lambda_2 = P_2(1 - T_{A1})\pi_A + (1 - P_2)(1 - T_{B1})\pi_B + P_2T_{A2}(1 - \pi_A) + (1 - P_2)T_{B2}(1 - \pi_B)$$

Theoretical results using this approach indicate significant gains over the Warner model. The variables in this technique

enable a wide selection of parameters for achieving the gains, e.g., P_1, P_2 π_B, and an optimal splitting of the total sample into the two required groups.

ASPECTS OF USING TWO RELATED CHARACTERISTICS

VISUALIZING THE PROBLEM

A rational model can be formed for the direct question case (and similarly for the randomized response cases). We shall assume that trait A is always sensitive. Trait B is related to A and may or may not be stigmatizing. As illustrated in Figure 1, membership in the A group does not influence the subject's answer to the question about B because he is unaware that he will be quizzed on A. Therefore, his response is based solely on characteristic B. Similarly, his response to the question on A will most likely depend on his sensitivity to A and not on the relationship of A to B. Admittedly, if B is stigmatizing, the relationship with A could affect the response to A, but the effect can largely be overcome. In the direct question model with complete and true reporting the response probabilities are

$$\delta_3 = \Pr\{\text{yes-yes}\} = \pi_{AB}$$

$$\delta_2 = \Pr\{\text{yes-no}\} = \pi_B - \pi_{AB}$$

$$\delta_1 = \Pr\{\text{no-yes}\} = \pi_A - \pi_{AB}$$

$$\delta_0 = 1 - \pi_{AB} - \pi_B + \pi_{AB}$$

such that

$$\sum_{i=0}^{3} \delta_i = 1$$

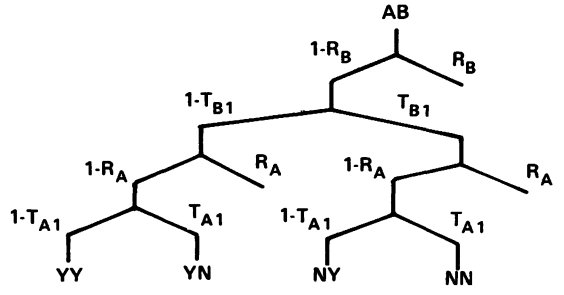

FIGURE 1. DIRECT QUESTION MODEL WHEN A RESPONDENT HAS BOTH CHARACTERISTICS A AND B

MEASURING THE EFFECTS OF BIAS AND NON-RESPONSE

(A Generalized Mean Square Error Definition)

In the univariate studies, mean square error efficiency was employed as the measure of performance for the various schemes; recall that mean square error is the sum of the variance of the estimate and square of the bias in the estimate. In the bivariate case several possible candidates were considered:

- MSE $(\hat{\pi}_A)$
- MSE $(\hat{\pi}_A + \hat{\pi}_B - \hat{\pi}_{AB})$ = MSE $(\hat{\pi}_{ab})$
- MSE $(\hat{\pi}_A + \hat{\pi}_B + \hat{\pi}_{AB}) = \sum_i \sum_j (v_{ij} + b_{ij})$

where v_{ij} is the ij^{th} element of the variance-covariance matrix v and b_{ij} is the ij^{th} element of BB′, the squared biases and "co-biases."

- trace $[V + BB'] = \sum_i (v_{ii} + b_{ii})$
- $|V| + |BB'|$
- $|V + BB'|$

Table 1 below summarizes the results of the analysis to select the most suitable standard for comparison. Such a standard should increase as the variances increase and bias increases. Measures 3 and 4 react in this way; measure 3 was chosen because it is more general than measure 4.

TABLE 1
CANDIDATES FOR EFFICIENCY CRITERIA EVALUATED FOR DIRECT QUESTION METHOD AT $\pi_A = 0.2$, $\pi_B = 0.4$ AND VARYING π_{AB} AT TWO LEVELS OF UNTRUTHFULNESS

CASE	π_{AB}	T_{A1}	CANDIDATE					
			(1)·n	(2)·n	(3)·n	(4)·n	(5)·n^3	(6)·n^3
1	0	0	0.16	0.24	0.24	0.4	0	0
	0.1		0.16	0.25	0.81	0.49	0.0015	0.0015
	0.2		0.16	0.24	1.36	0.56	0	0
2	0	0.1	0.55	0.64	0.64	0.79	0	0
	0.1		0.55	0.35	1.66	0.97	0.0013	0.0048
	0.2		0.55	0.24	2.86	1.34	0	0

THE GENERAL EXTENDED ALTERNATE QUESTION MODEL (GEAQ)

This model is depicted in Figure 2 for an individual with both stigmatizing characteristic A and related characteristic B which may or may not be stigmatizing. B is chosen so that it is no more sensitive than A, i.e., $\pi_A \leq \pi_B$.

Hence

$$\max(0, \pi_A + \pi_B - 1) \leq \pi_{AB} \leq \pi_A$$

In the presence of a 2 by 2 table as described here the correlation coefficient is

$$\rho_{AB} = (\pi_{AB} - \pi_A \pi_B) \div \sqrt{\pi_A (1 - \pi_A) \pi_B (1 - \pi_B)}$$

To define the model, we will for simplicity here assume complete and true reporting. A sample of size n is drawn with replacement from the population. In order to estimate π_A, π_B, and π_{AB}, two responses are required for each individual in the sample. A separate randomizing device will be used for each response.

With δ_3, δ_2, and δ_1 as the probabilities for a "yes-yes," "yes-no," and "no-yes" response respectively we have

$$\begin{bmatrix} \delta_3 \\ \delta_2 \\ \delta_1 \end{bmatrix} = \begin{bmatrix} P_1(1-P_2)+P_2(1-P_1) & (1-P_1)(1-P_2) & P_1P_2 \\ -[P_1(1-P_2)+P_2(1-P_1)] & (1-P_1)P_2 & P_1(1-P_2) \\ -[P_1(1-P_2)+P_2(1-P_1)] & P_1(1-P_2) & P_2(1-P_1) \end{bmatrix} \begin{bmatrix} \pi_{AB} \\ \pi_B \\ \pi_A \end{bmatrix}$$

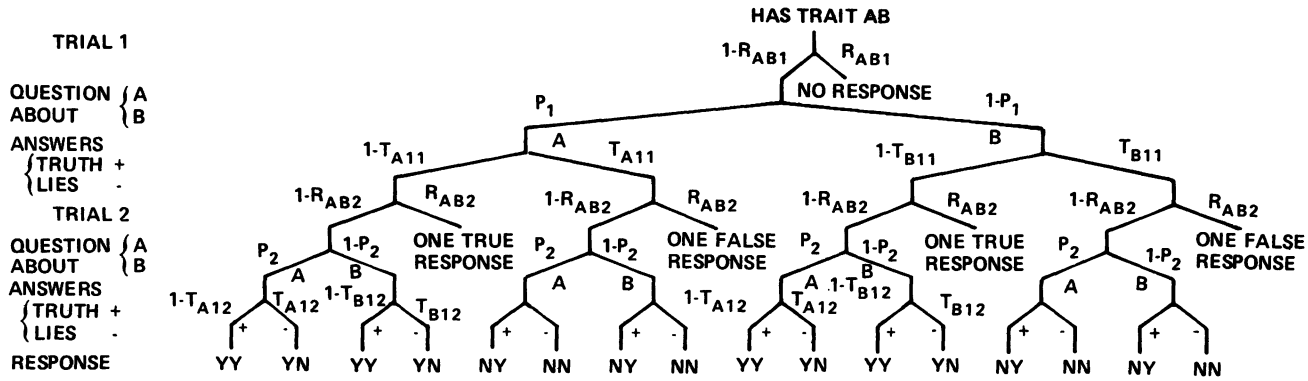

FIGURE 2. THE GENERAL EXTENDED ALTERNATE QUESTION MODEL, REFLECTING INCOMPLETE AND DISHONEST REPORTING FOR AN INDIVIDUAL WITH BOTH TRAITS A AND B

where P_1 and P_2 represent the proportion of time the question about characteristic A is answered with each device.

Maximum likelihood estimates of π may be obtained. Estimates for the variances, and covariances may also be obtained since the observed proportions of responses are multinomially distributed with parameters $(n, \delta_0, \delta_1, \delta_2, \delta_3)$.

Estimates of π may occur outside the range (0, 1) when (1) departures from randomness occurs, (2) complete and honest reporting do not occur, (3) P_1 is too close in value to P_2, or (4) characteristics A and B are negatively correlated.

P_1 should be chosen smaller than P_2 in order to help convince the respondent that no tricks are involved in the technique. A smaller P_1 will enable a more reliable estimate of π_B, also. P_2 can then be made as close to unity as the experimenter believes he can get away with. A good general rule, however, is to make $P_2 = 1 - P_1$.

Efficiency with this technique will be higher when π_B is close to π_A and the two characteristics are positively correlated.

Whereas Figure 2 displays the model applied to a person with both characteristics, the other possibilities can be described similarly and equations produced for this model. A comparison with the direct question model is exhibited in Table 2 when incomplete and/or untruthful reporting exist in the direct question technique. Note that the GEAQ method deserves consideration as an alternative to the DQ approach even at low levels of nonresponse and untruthfulness and especially when attributes A and B are as positively correlated as possible. Also observe that untruthfulness among those who don't have the two characteristics seriously degrades the efficiency.

SUMMARY

An additional way to employ the randomized response techniques to achieve greater efficiency in surveys has been discussed. Moreover, an extension of mean square error has been presented for use in multivariate situations. And still, the iceberg of randomized response is not totally explored.

TABLE 2
EFFECT OF NONRESPONSE AND UNTRUTHFULNESS IN THE DIRECT QUESTION BIVARIATE METHOD ON SAMPLE SIZE AND ON THE EFFICIENCY OF THE ALTERNATE QUESTION METHOD ASSUMING COMPLETE AND TRUE RESPONSES

$n = 1000, P_2 = 1 - P_1 = 0.7$

$$\pi' = \begin{pmatrix} 0 \\ 0.1, \ 0.4, 0.2 \\ 0.2 \end{pmatrix}$$

EFFECTIVE SAMPLE SIZE IN DQ METHOD	PROBABILITIES OF UNTRUTHFULNESS				PROBABILITIES OF NON-RESPONSE				MSEE = MSE DQ / MSE AQ		
n'	T_{A1}	T_{B1}	T_{A2}	T_{B2}	R_A	R_B	R_a	R_b	$\pi_{AB}=0$	$\pi_{AB}=.1$	$\pi_{AB}=.2$
1000	0	0	0	0	0	0	0	0	0.26	0.64	0.86
1000	0.1	0	0	0	0	0	0	0	0.70	1.32	1.81
1000	0.3	0	0	0	0	0	0	0	4.20	6.95	9.75
1000	0.3	0.3	0	0	0	0	0	0	35.65	42.77	50.70
1000	0.3	0.3	0.1	0	0	0	0	0	6.12	13.94	22.87
1000	0.3	0.3	0	0.1	0	0	0	0	12.64	21.8	31.62
1000	0.3	0.3	0.1	0.1	0	0	0	0	0.6	3.24	9.97
980	0	0	0	0	0.1	0	0	0	0.34	1.18	2.10
940	0	0	0	0	0.1	0.1	0	0	1.00	3.41	6.29
980	0.1	0	0	0	0.1	0	0	0	1.04	2.91	4.97

REFERENCES

1. Abul-Ela, A. A., Greenberg, B. G., and Horvitz, D. G., "A Multiproportions Randomized Response Model," *Journal of the American Statistical Association*, 62 (1967) 990-1008.

2. Greenberg, B. G., Abernathy, J. R., and Horvitz, D. G., "A New Survey Technique and Its Application in the Field of Public Health," *Milbank Memorial Fund Quarterly* (1970) 39-55.

3. Greenberg, B. G., Abul-Ela, A. A., Simmons, W. R., and Horvitz, D. G., "The Unrelated Question Randomized Response Model: Theoretical Framework," *Journal of the American Statistical Association*, 64 (1969) 520-539.

4. Warner, S. L., "Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias," *Journal of the American Statistical Association*, 60 (1965) 63-69.