THE COMPREHENSION FACTOR IN RANDOMIZED RESPONSE

Dennis M. O'Brien, University of Wisconsin - La Crosse
Robert S. Cochran, University of Wyoming

I. INTRODUCTION

Since the introduction of the randomized response technique for questioning interviewees on sensitive topics by Stanley L. Warner [4] in 1965, several modifications and extensions of the procedure have been presented. For two such extensions, see Greenberg, et al. [1] and Horvitz, et al. [2]. Oftentimes, the primary motivation for refinements has been to encourage further cooperation on the part of the potential respondent and thus provide more accurate information and make more precise estimates possible. In all applications of the technique, close adherence to the instructions and control over the implementation and mechanics is required. These later attempts to further assure anonymity sometimes carry with them a more complex set of instructions which the interviewee is expected to understand and then follow. Some investigations have been performed into the effects of truthfulness of the respondent on some of the questioning models. However, very little has ever been mentioned on the ability or the desire to follow instructions.

Comments contributed by respondents to a 'Consumer Opinion Survey' (see O'Brien, et al. [3]), where variations of the randomized response technique were used, indicate that there is reason to suspect less than complete comprehension and an unwillingness to follow instructions. Hence this paper will introduce a 'comprehension factor', which includes the idea of truthfulness as well as the interest in (and/or ability to) following instructions. Its effect on estimation and variance formulas will be shown for three randomized response models. Also considered will be the action taken by the 'non-comprehenders'.

II. INTRODUCTION OF THE COMPREHENSION FACTOR INTO TWO QUALITATIVE MODELS

The Warner related question procedure (see Warner [4]) requires that the respondent be given two statements of the form:

1) I am a member of Group A
2) I am not a member of Group A (1)

and a randomizing device. The respondent will use the randomizing device to determine to which statement he is to respond. His answer is then 'yes' or 'no'.

The Simmons unrelated question procedure (see Horvitz, et al. [2]) also uses a randomizing device, but the statements are now of the form:

1) I am a member of Group A
2) I am a member of Group B, (2)

In both models Group A is considered to be of a sensitive nature so that an individual when asked directly about his affiliation with that group may refuse to answer or may answer, but will give false information. Group B is of a non-sensitive nature and is assumed to generate no hesitancy in admitting membership. The goal of the Warner and Simmons procedures is to estimate π, the proportion of the population who are members of the sensitive Group A. For this paper it is assumed that π_Y, the proportion in Group B of the Simmons method, is known and hence only a single simple random sample of size n is needed. If π_Y is unknown, two samples are needed. For a discussion of this case, see Horvitz, et al. [2].

The maximum likelihood estimators and their variances for these two procedures are as follow:

Warner-

$$\hat{\pi}_W = \frac{P-1}{2P-1} + \frac{n_1}{(2P-1)n} \qquad (3)$$

$$V(\hat{\pi}_W) = \frac{\pi(1-\pi)}{n} + \frac{P(1-P)}{n(2P-1)^2} \qquad (4)$$

where P = the probability of the random device indicating Group A ($P \neq \frac{1}{2}$) and n_1 = the number of 'yes' responses.

Simmons-

$$\hat{\pi}_S = [\frac{n_1}{n} - (1-P)\pi_Y]/P \qquad (5)$$

$$V(\hat{\pi}_S) = \frac{1}{P^2 n} [\pi P+\pi_Y(1-P)][(1-\pi)P+(1-\pi_Y)(1-P)]. \qquad (6)$$

Under the assumptions of complete comprehension and truthfulness in responses, equal sample size, and equal probabilities of a Group A indication, $V(\hat{\pi}_S)$ is always less than $V(\hat{\pi}_W)$. However if, for any reason, a proportion of the respondents do not answer the question in the proper fashion then there is a possibility of circumstances developing where the Warner method may prove to have a lower mean-square-error. In the following all reasons for not answering in the proper fashion are grouped under the general heading of "comprehension".

To handle this concept the following additional parameters are introduced:

Θ_W, Θ_S for the levels of comprehension of the Warner and Simmons procedures, respectively (proportion of the sample that responds correctly and honestly);

Θ_{YW}, Θ_{YS} for the probability of responding with a 'yes' in the event of miscomprehending in the respective procedures.

For the present time, assume the values of these new parameters are unknown and thus cannot be allowed for in the estimators. Under this assumption the estimator expressions are unchanged but they are now biased and the variance expressions change. The bias and variance expressions are:

Warner-

$$\text{Bias}_W = (1-\Theta_W)\left(\frac{P+\Theta_{YW}-1}{2P-1} - \pi\right) \tag{7}$$

(Note that Bias_W is independent of the sample size n.)

$$V(\hat{\pi}_W) = \frac{1}{(2P-1)^2 n}[\Theta_W\pi P+\Theta_W(1-\pi)(1-P)+(1-\Theta_W)\Theta_{YW}]\cdot[\Theta_W\pi(1-P)+\Theta_W(1-\pi)P+(1-\Theta_W)(1-\Theta_{YW})]. \tag{8}$$

Simmons-

$$\text{Bias}_S = (1-\Theta_S)\left(\frac{P\pi_Y+\Theta_{YS}-\pi_Y}{P} - \pi\right) \tag{9}$$

(Note that Bias_S is independent of the sample size n.)

$$V(\hat{\pi}_S) = \frac{1}{P^2 n}[\Theta_S\pi P+\Theta_S\pi_Y(1-P)+(1-\Theta_S)\Theta_{YS}]\cdot[\Theta_S(1-\pi)P+\Theta_S(1-\pi_Y)(1-P)+(1-\Theta_S)(1-\Theta_{YS})]. \tag{10}$$

Comparing the two procedures on the basis of mean-square-errors, MSE = $V(\hat{\pi})+\text{Bias}^2$, under various parameter conditions it can be shown that situations exist where $\text{MSE}_W < \text{MSE}_S$. As an illustration, consider the situation where n = 500, P = .8, π = .3, π_Y = .1, Θ_{YW} = .5, and Θ_{YS} = .0. The following is observed as the comprehension factor increases:

$\Theta_W = \Theta_S$	$\text{MSE}_W/\text{MSE}_S$
.80	.62
.85	.77
.90	1.07
.95	1.68

Thus the comprehension levels and the action taken by the non-comprehenders should be considered when deciding which of these two models is to be implemented.

At this time, assume that pre-sampling or past experience has provided values for these new parameters. Allowing for the comprehension factor in the estimators makes them unbiased. The new estimators and their variances are:

Warner-

$$\tilde{\pi}_W = \frac{P-1}{2P-1} + \frac{n_1 - n(1-\Theta_W)\Theta_{YW}}{(2P-1)n\Theta_W} \tag{11}$$

$$V(\tilde{\pi}_W) = \frac{1}{(2P-1)^2 n\Theta_W^2}[\Theta_W\pi P+\Theta_W(1-\pi)(1-P)+(1-\Theta_W)\Theta_{YW}]\cdot[\Theta_W\pi(1-P)+\Theta_W(1-\pi)P+(1-\Theta_W)(1-\Theta_{YW})]. \tag{12}$$

Simmons-

$$\tilde{\pi}_S = \left[\frac{n_1-n(1-\Theta_S)\Theta_{YS}}{n\Theta_S} - (1-P)\pi_Y\right]/P \tag{13}$$

$$V(\tilde{\pi}_S) = \frac{1}{P^2 n\Theta_S^2}[\Theta_S\pi P+\Theta_S\pi_Y(1-P)+(1-\Theta_S)\Theta_{YS}]\cdot[\Theta_S(1-\pi)P+\Theta_S(1-\pi_Y)(1-P)+(1-\Theta_S)(1-\Theta_{YS})]. \tag{14}$$

Comparisons between the MSE of the 'standard' Warner and $V(\tilde{\pi}_W)$ reveal situations where the 'standard' is best, that is, where $\text{MSE}_W < V(\tilde{\pi}_W)$, as well as parameter combinations where the latter estimator is best. As an illustration, consider the case where n = 500, P = .8, π = .3, and Θ_{YW} = .3. The following is observed as Θ_W increases:

Θ_W	$\text{MSE}_W/V(\tilde{\pi}_W)$
.7	1.12
.8	1.00
.9	.94

Similar situations occur for the 'standard' Simmons vs. the 'modified' Simmons. For example, consider the situation where n = 500, P = .8, Θ_S = .8, and Θ_{YS} = .5. The following is observed:

$\pi=\pi_Y$	$\text{MSE}_S/V(\tilde{\pi}_S)$
.1	15.29
.5	.67
.9	12.43

III. INTRODUCTION OF THE COMPREHENSION FACTOR INTO A QUANTITATIVE MODEL

The Greenberg quantitative model (see Greenberg, et al. [1]) uses the unrelated question randomized response procedure to obtain quantitative information on sensitive topics. A randomizing device is used to indicate the question to which the interviewee is to respond. The questions are of the form:

1) How many abortions have you had during your lifetime?
2) If a woman had to work full-time to make a living, how many children do you think she should have? (15)

Question 1) is considered the sensitive question, while 2) is considered the non-sensitive question. As for the Simmons model, the investigator may or may not know the parameter values for the responses to the non-sensitive question. In this case they are the mean and variance, denoted by (μ_Y, σ^2_Y). In this paper it is assumed they are known and thus a single sample of size n is needed. The object is to estimate the mean of the sensitive question response distribution, $\mu_{\underline{X}}$.

An unbiased estimator for $\mu_{\underline{X}}$ and its variance are given by:

$$\hat{\mu}_{\underline{X}} = [\bar{Z}-(1-P)\mu_Y]/P, \qquad (16)$$

where $\bar{Z}$ is the sample response mean.

$$V(\hat{\mu}_{\underline{X}}) = \frac{1}{nP^2}[P\sigma^2_{\underline{X}} + (1-P)\sigma^2_Y + P(1-P)(\mu_{\underline{X}}-\mu_Y)^2], \qquad (17)$$

where $\sigma^2_{\underline{X}}$ is the variance of the sensitive question response distribution.

Letting Θ be the unknown proportion comprehending and following all instructions and assuming all non-comprehenders respond as if answering the non-sensitive question, $V(\hat{\mu}_{\underline{X}})$ becomes

$$V(\hat{\mu}_{\underline{X}}) = \frac{1}{nP^2}[P\Theta\sigma^2_{\underline{X}} + (1-P\Theta)\sigma^2_Y + P\Theta(1-P\Theta)(\mu_{\underline{X}}-\mu_Y)^2]. \qquad (18)$$

The estimator now has a bias of

$$\text{Bias}_G = (1-\Theta)(\mu_Y-\mu_{\underline{X}}). \qquad (19)$$

The standard direct question estimator

$$\hat{\mu} = \bar{Z} \qquad (20)$$

has a variance of

$$V(\hat{\mu}) = \sigma^2_{\underline{X}}/n \qquad (21)$$

under complete truthfulness. Letting T be the probability of obtaining a truthful response in a direct question interview and assuming those not responding truthfully respond according to a distribution with mean and variance of μ_T and σ^2_T, the estimator has a bias and variance as follows:

$$\text{Bias}_D = (1-T)(\mu_T-\mu_{\underline{X}}) \qquad (22)$$

$$V(\hat{\mu}) = \frac{1}{n}[T\sigma^2_{\underline{X}} + (1-T)\sigma^2_T + T(1-T)(\mu_{\underline{X}}-\mu_T)^2]. \qquad (23)$$

Comparisons of the Greenberg MSE under varying degrees of comprehension and the direct MSE under varying levels of truthfulness reveal cases where the Greenberg procedure is best as well as cases where the direct question approach is best. As an illustration, consider the case where $n = 500$, $P = .75$, $\sigma^2_Y = \sigma^2_T = .5\sigma^2_{\underline{X}}$, $\mu_Y = \mu_T = .5\mu_{\underline{X}}$, $\sigma_{\underline{X}} = .1\mu_{\underline{X}}$, and $T = .7$. As Θ increases, the following is observed:

Θ	MSE_G/MSE_D
.6	1.78
.7	1.01
.8	.45
.9	.12

References

[1] Greenberg, B. G., Abernathy, J. R. and Horvitz, D. G., "Application of the Randomized Response Technique in Obtaining Quantitative Data," Proceedings of Social Statistics Section, ASA (Aug. 1969), 40-3.

[2] Horvitz, D. G., Shah, B. V. and Simmons, W. R., "The Unrelated Question Randomized Response Model," Proceedings of Social Statistics Section, ASA (1967), 65-72.

[3] O'Brien, Dennis M., Cochran, Robert S., Marquardt, Ray S. and Makens, James C., "Randomized Response vs. Direct Question in a Mail vs. Personal Interview Consumer Opinion Survey," College of Commerce and Industry Research Paper No. 85, University of Wyoming, July 1975, Laramie, Wyoming.

[4] Warner, S. L., "Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias," JASA, 60 (1965), 63-9.