THE COVARIATE UNRELATED QUESTION RANDOMIZED RESPONSE MODEL

N.J. Scheers and C. Mitchell Dayton, University of Maryland

INTRODUCTION

The randomized response technique (Warner, 1965) is a data collection procedure that allows researchers to collect sensitive information while guaranteeing privacy to respondents. For example, if one question from a pair of questions that have the same response options is selected randomly and in private, the question answered by a given individual will not be known to the researcher. However, there is sufficient information from a sample of respondents (e.g., the total number of "yes" responses and the probability of selecting the sensitive question) to estimate the proportion of the population with the sensitive characteristic.

Additional variance for the estimator of the population with the sensitive characteristic results from sampling items with the randomized response technique. The technique described in this paper allows one to improve accuracy of estimation by using a covariate that correlates with the sensitive characteristic but which is itself not sensitive. Theory has been developed for a covariate randomized response model which is an extension of the Warner procedure (Dayton and Scheers, 1980); the present study involves a similar extension for the unrelated question randomized response technique (see Greenberg, et al., 1969).

This study comprises three stages: (1) theoretical development of the covariate model for both π_Y known and π_Y unknown cases, (2) construction of efficiency tables that compare the covariate model to the unrelated question model and (3) a demonstration study of the covariate model.

2. THEORETICAL DEVELOPMENT

With the unrelated question design, one statement from a pair of statements is randomly selected to answer by each respondent. One of these statements is embarrassing or sensitive (A), while the other is nonsensitive (Y). For example:

A: I have copied answers on an exam.

Y: I was born in May.

For the covariate modification, an additional nonsensitive statement is answered anonymously by each respondent; for example:

X: Estimate your grade-point average.

It should be noted, however, that an assumption of the covariate model is that the covariate is nonsensitive. If the covariate is sensitive, either because respondents feel they may be identified by the anonymous question used to obtain the covariate information, or because they feel the question itself is an invasion of their privacy, cooperation from respondents will be introduced into the estimate.

It is necessary to model the relationship between the covariate X and π_A such that X, which (in theory) has infinite range, can be mapped onto the 0 to 1 interval for π_A. Of course, any cumulative frequency distribution has this characteristic. Also, the function should be nonlinear, since (a) a linear cumulative frequency distribution results from a rectangular distribution which would be an unusually restrictive model and (b) a linear function does not remain positive for all values of X, since X has infinite range. Along with the conditions described above, a function would be preferred that has been widely used in psychological research. Two such functions are the normal ogive and logistic functions. It is known that the logistic function, "...differs by less than 0.01, uniformly in X, from the normal cdf with mean zero and standard deviation 1.7...." (Birnbaum, 1968, p. 399). Since the functions are so similar, the logistic function was chosen because it is mathematically easier to use.

In the present context, the form of the logistic function is:

$$\pi_{A|X_i} = [(1+e^{-\alpha-\beta X_i})^{-1}] \qquad (2.1)$$

where X_i is the i^{th} level of the covariate.

The π_Y Known Covariate Model

For the unrelated question design, π_Y known, the probabilities of a "yes" response and a "no" response are:

$$P(Y = \text{yes}) = P\pi_A + (1-P)\pi_Y$$
$$P(Y = \text{no}) = P(1-\pi_A) + (1-P)(1-\pi_Y) \qquad (2.2)$$

where P = the probability of selecting the sensitive question,
π_A = the proportion of people in the population with the sensitive characteristic,
π_Y = the proportion of people in the population with the nonsensitive characteristic.

With the covariate modification, π_A is rewritten in terms of the logistic function; letting Y_i be the value of Y conditional on X_i:

$$P(Y_i = \text{yes}) = P(1+e^{-\alpha-\beta X_i})^{-1} + (1-P)\pi_Y$$
$$P(Y_i = \text{no}) = P[1-(1+e^{-\alpha-\beta X_i})^{-1}] + (1-P)(1-\pi_Y) \qquad (2.3)$$

Assuming that n_{1i} respondents at the i^{th} level of the covariate report "yes" and n_{oi} respondents at the i^{th} level of the covariate report "no", the likelihood function is:

$$L = \prod_{i=1}^{k} \{[P(1+e^{-\alpha-\beta X_i})^{-1} + (1-P)\pi_Y]^{n_{1i}} \cdot [1-P(1+e^{-\alpha-\beta X_i}) - (1-P)\pi_Y]^{n_{oi}}\} \qquad (2.4)$$

Since the normal equations found by taking derivatives of the log likelihood function are

nonlinear in the parameters, the Fisher Method of Scoring (Rao, 1965) can be used to estimate the parameters iteratively.

The variance of π_A at each level of X can be calculated using the law of propagation of error (Kendall and Stuart, 1961):

$$\text{Var}(\hat{\pi}_{Ai}) = \left(\frac{\partial \hat{\pi}_{Ai}}{\partial \alpha}\right)^2 \text{Var}(\hat{\alpha}) + \left(\frac{\partial \hat{\pi}_{Ai}}{\partial \beta}\right)^2 \text{Var}(\hat{\beta}) + 2\left(\frac{\partial \hat{\pi}_{Ai}}{\partial \alpha}\right)\left(\frac{\partial \hat{\pi}_{Ai}}{\partial \beta}\right)\text{Cov}(\hat{\alpha}, \hat{\beta}) \quad (2.5)$$

where $\hat{\pi}_{Ai}$ = the estimate of π_A at the i^{th} level of the covariate and

$$\frac{\partial \hat{\pi}_{Ai}}{\partial \alpha} = \frac{(e^{-\alpha-\beta X_i})}{(1+e^{-\alpha-\beta X_i})^2}$$

$$\frac{\partial \hat{\pi}_{Ai}}{\partial \beta} = X_i\left(\frac{\partial \hat{\pi}_{Ai}}{\partial \beta}\right)$$

where the $\text{Var}(\hat{\alpha})$, the $\text{Var}(\hat{\beta})$ and the $\text{Cov}(\hat{\alpha}, \hat{\beta})$ are elements in $-D^{-1}$, the negative inverse of the matrix of second order partial derivatives from the Fisher procedure evaluated at the point of the final estimates for α and β. A computer program has been developed to accomodate the π_Y known covariate model.

The π_Y Unknown Covariate Model

When π_Y is unknown and must be estimated, two groups of respondents must be used. The probabilities of a "yes" response and a "no" response are:

$$P(Y = \text{Yes}) = P_j\pi_A + (1-P_j)\pi_Y$$
$$P(Y = \text{No}) = P_j(1-\pi_A) + (1-P_j)(1-\pi_Y) \quad (2.6)$$

where P_j = the probability of selecting the sensitive question in group j (j = 1, 2)

and $P_1 \neq P_2$

For the covariate modification, π_A is rewritten as a logistic function, where X_i is the i^{th} level of the covariate, which is assumed to have known, fixed levels:

$$\pi_{A|X_i} = (1+e^{-\alpha-\beta X_i})^{-1}$$

Substituting this function in equation (2.6) results in:

$$P(Y_i = \text{yes}) = P_j(1+e^{-\alpha-\beta X_i})^{-1} + (1-P_j)\pi_Y \quad (2.7)$$
$$P(Y_i = \text{no}) = 1 - [P_j(1+e^{-\alpha-\beta X_i})^{-1}] - (1-P_j)\pi_Y$$

Assuming that n_{11i} respondents at the i^{th} level of the covariate report "yes" and n_{01i} respondents at the i^{th} level of the covariate report "no" for group 1, and that n_{12i} respondents at the i^{th} level of the covariate report "yes" and n_{02i} respondents at the i^{th} level of the covariate report "no" for group 2, the likelihood function is:

$$L = \prod_{j=1}^{2}\prod_{i=1}^{k}\{[P_j(1+e^{-\alpha-\beta X_i})^{-1} + (1-P_j)\pi_Y]^{n_{1ji}} \cdot [1-(P_j(1+e^{-\alpha-\beta X_i})^{-1} - (1-P_j)\pi_Y]^{n_{0ji}}\} \quad (2.8)$$

The resulting normal equations are difficult to solve directly; thus, the Fisher Method of Scoring is used to estimate the unknown parameters α, β, and π_Y. A program has been developed to accomodate the π_Y unknown covariate model.

Assessing Fit of Models

A Pearson or likelihood ratio (Rao, 1965), chi-square goodness-of-fit test can be used to determine if the logistic function adequately represents a set of data. After parameter estimates for α and β have been found, the expected proportions for "yes" responses and "no" responses can be determined for each level of the covariate using equation (2.3) for the π_Y known model or equation (2.7) for the π_Y unknown model. If the non-covariate model of equation (2.2) were fitted to each of the k levels of the covariate, a total of k independent values of π_A would be estimated. The covariate model fits the same data table with only two estimated parameters, α and β. Thus, the degrees of freedom for a goodness-of-fit test are k-2. In addition, if π_Y must be estimated, as in the case of the π_Y unknown model, the appropriate degrees of freedom become k-3.

3. RELATIVE EFFICIENCY TABLES

There appears to be no completely satisfactory way to determine the efficiency of the covariate unrelated question model compared to the usual unrelated question model. An efficiency comparison may be made between a weighted average variance over covariate levels for the covariate model and a single variance based on the total sample, disregarding covariate levels. For this comparison, the variance of the usual unrelated question model should be augmented by a bias term since the single parameter estimate will differ systematically from the true values at the various levels. If, however, the efficiency comparison is made at each level of the covariate and a weighted average variance over covariate levels determined for each model, relative efficiency is influenced by the reduction in number of respondents per level. Despite this limitation, the second approach to assessing relative efficiency has been used in this paper.

Relative efficiency tables presented in this section compare the variance of the covariate unrelated question model to the variance of the original unrelated question model, assuming truthful responses, such that:

$$E = \frac{\text{Mean Square Error (Covariate)}}{\text{Mean Square Error (Unrelated Question)}}$$

Covariates with 3 and 5 levels are paired with various π_A values. The total sample size (n=200) was assumed to be normally distributed over the covariate levels and initial frequencies were calculated from the sample size and π_A at each covariate level. Since the variances in the numerator and denominator of the

efficiency ratio both involve N, the sample size, this factor cancels out and the comparison is constant across all sample sizes. Other parameter values used to calculate the efficiency tables for both models are: $\pi_Y = 1, .3$; $P = .7, .8$; and for the π_Y unknown case, P_1, $P_2 = .7, .3$ and $.8, .2$. For the $_Y$ unknown case, n_1 and n_2 were not optimally allocated since the purpose was to compare models, not to minimize variance.

Results are presented in Table 1 for a covariate with 3 levels ($\pi_A = .10, .35, .60$) and 5 levels ($\pi_A = .10, .225, .35, .475, .60$) and with $\pi_Y = .1$. For other π_A and π_Y values at 3 and 5 covariate levels, relative efficiency is virtually the same. The following conclusions result from an examination of these tables:

1. The reduction in variance by the covariate model relative to the unrelated question model is much larger for the π_Y unknown model than for the π_Y known model.

2. The efficiency of the covariate model increases relative to the unrelated question model as the number of covariate levels increases from k = 3 to k = 5.

3. There was little change in the efficiency of the covariate model relative to the unrelated question model across P values.

4. DEMONSTRATION STUDY

Estimates of 5 different types of academic cheating behavior were obtained for a group of students questioned anonymously (N = 194) and a group questioned by the unrelated question technique (N = 184) at the University of Maryland in the Spring of 1981. The covariate was grade-point average and respondents were asked to categorize their grade-point average as (1) 3.76 - 4.00, (2) 3.51 - 3.75, (3) 3.26 - 3.50, (4) 3.00 - 3.25, or (5) 2.99 or less.

For the covariate unrelated question group, respondents selected one question from each of 5 pairs of questions to answer "true" or "false" by using a spinner. The spinner was constructed so that the probability of selecting the sensitive question was .70. An example of one pair of questions is:

A. I have lied to a teacher to avoid taking an exam.
B. I was born in August.

For the anonymous questionnaire, respondents simply circled "true" or "false" after each question about academic cheating behavior.

For the randomized response questionnaire, nonsensitive questions were devised from students' social security numbers, birth month, and number of course credits taken in Spring, 1981. An estimate of the proportion of people with each nonsensitive characteristic (π_{Yi}) was obtained before the survey was administered in order to determine the distribution of each nonsensitive characteristic.

When the covariate model is used to allow variation in π_A as a function of GPA, substantially different estimates of π_A occur at the various covariate levels. Thus, ignoring GPA produces biased results and this bias is greater for some cheating behaviors than for others. Table 2 shows the variation in the estimate of π_A over GPA levels for both the covariate unrelated question model and the usual unrelated question model for the 5 questions of the demonstration study. The most dramatic variation in π_A occurred in question 5, "Copied Answers," where the percent of students who admitted to copying answers on an exam ranged from 21% at the highest GPA level to 86% at the lowest GPA level for the covariate model.

The logistic function was found to be a parsimonious representation of the relationship between π_A and the covariate (GPA). A chi-square goodness-of-fit test was calculated for each of the 5 questions, and the chi-square statistics, 1.36, 3.66, 0.14, 1.185, and 3.10, were each nonsignificant with 3 degrees of freedom. This may be interpreted to mean that the logistic model fits the observed data as well as separate estimates at each covariable level from the usual unrelated question model.

All covariate estimates of the sensitive behaviors were larger than the estimates obtained from the anonymous questionnaire. If a weighted average for estimates of the academic cheating behavior is calculated for each item, the differences between the estimates from the covariate model and the anonymous questionnaire range from .13 to .18 (see Table 3). This indicates relatively severe under-reporting since the proportions being estimated were in a range from .15 to .48. Thus the percent of under-reporting is from 43% to 83%, suggesting that the questions about academic cheating behavior were indeed sensitive.

REFERENCES

Birnbaum, A., "The Logistic Test Model," in Lord, F. and Novick, M., Statistical Theories of Mental Test Scores. Reading, Mass: Addison-Wesley Publishing Co., (1968), pp. 397-423.

Dayton, C.M. and Scheers, N.J., "The Covariate Warner Randomized Response Model," (unpublished work, University of Maryland) (1980).

Dixon, W.J. (ed.), Biomedical Computer Programs, P-Series, Los Angeles: University of California Press, (1979).

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

Kendall, M. and Stuart, A., The Advanced Theory of Statistics, Volume 2, London: Griffin (1961).

Rao, C.R., Linear Statistical Inference and Its Applications, New York: John Wiley and Sons, Inc., (1965).

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

Table 1

Efficiency of the Covariate Model
Compared to the Unrelated Question Model

$\pi_Y = .1$

	π_Y Known Model		π_Y Unknown Model	
		Relative Efficiency		Relative Efficiency
X = 3	$P_1 = .7$	.699	$P_1 = .7$, $P_2 = .3$	.250
	$P_1 = .8$	.700	$P_1 = .8$, $P_2 = .2$	.309
X = 5	$P_1 = .7$	.411	$P_1 = .7$, $P_2 = .3$	.148
	$P_1 = .8$	.412	$P_1 = .8$, $P_2 = .2$	.182

Table 2

Results: Survey of Academic Cheating Behaviors
Using the Covariate Unrelated Question Model

P = .70

GPA		Question 1	2	3	4	5
3.76 to 4.00	A_C	.13	.16	.14	.17	.21
	A_{UQ}	.11	.15	.14	.13	.25
3.51 to 3.75	A_C	.18	.22	.15	.20	.37
	A_{UQ}	.20	.14	.15	.30	.24
3.26 to 3.50	A_C	.23	.30	.16	.22	.56
	A_{UQ}	.33	.40	.18	.17	.63
3.00 to 3.25	A_C	.30	.38	.17	.25	.74
	A_{UQ}	.23	.56	.14	.26	.85
2.99 or less	A_C	.37	.48	.18	.28	.86
	A_{UQ}	.38	.36	.19	.27	.79

Table 3

Comparison of the Survey Results from
the Covariate Model and the Anonymous Questionnaire

Question	Covariate Estimate	Anonymous Estimate	Percent of Under-reporting
1	.223863	.128382	42.65
2	.280996	.139168	50.47
3	.154305	.025763	83.30
4	.214832	.087628	59.21
5	.482353	.293810	39.09