VARYING LEVELS OF PROBABILITY FOR SELECTING SENSITIVE QUESTIONS USING A RANDOMIZED RESPONSE TECHNIQUE/1

Robin Laurie Pearl and Walter T. Federer, Cornell University

1. Introduction

Obtaining information pertaining to sensitive or stigmatizing characteristics has been a vexing problem that is encountered in sample surveys. The questions that make the respondent suffer embarrassment if he answers the question affirmatively prompt him to select the path that is least likely to jeopardize his reputation. This would then entail data that are mostly unreliable. Research in statistical methodology to devise schemes to elicit answers in the above context has been in the direction of finding methods that ensure anonymity to the respondent in as far as his answer is concerned. It is believed that if the interviewer does not know what the answer from the respondent to the sensitive question is, then the respondent feels safe in responding truthfully to the sensitive question. In this direction, an attempt has been made by Warner [5], who devised a scheme which is currently known as randomized response technique. It is assumed that the population consists of two groups A and B; and the membership in A bears a stigma. It is desired to estimate the proportion of individuals belonging to the group A. The respondent is asked to use a randomized device such as a six-sided die, for example. If a one or two, say, turns up, he is asked to answer the sensitive question, without revealing the result of the toss,

1) I am a member of group A,

truthfully by marking "yes" or "no". If a 3, 4, 5 or 6 turns up, he is asked to choose the nonsensitive question,

2) I am a member of group B,

truthfully by marking "yes" or "no". Thus, the interviewer does not know which statement is answered by the respondent and the answer to the statement. Therefore it can be reasonably assumed that the respondent can be persuaded indirectly to answer truthfully without experiencing any embarrassment. This assumption remains to be tested in field trials because its validity is related to the degree of sensitivity of the question or the nature of the stigmatizing characteristic.

Later Simmons modified [cf. [2]] Warner's procedure by introducing an unrelated question along with the sensitive question. Greenberg <u>et al</u>. [1] further explored Simmon's method and studied its statistical properties. Briefly, the method is as follows.

An experiment is conducted with known probability P of occurrence of an event (say, a one or a two turning up in the die tossing experiment). If the event is realized in the experiment, without revealing the outcome to the interviewer, the respondent is asked to choose to answer truthfully the sensitive question.

A) I had an abortion in the last few months by marking "yes" or "no". If the event is not realized, then the respondent is asked to choose the nonsensitive question.

B) I was born in New York and answer truthfully by marking "yes" or "no". Again, apparently the anonymity of the respondent is preserved. Symbolically writing, if P denotes the probability of occurrence of an event in the experiment conducted, Π_A the true proportion in population of the membership of A, Π_B the proportion of membership of B, then

$$\begin{aligned}&\text{Prob(yes answer)}\\&=\text{Prob(A is chosen)}\cdot\text{P(yes answer}\mid\text{A is chosen)}\\&\quad+\text{Prob(B is chosen)}\cdot\text{P(yes answer}\mid\text{B is chosen)}\end{aligned}$$

$$\lambda=\Pi_A\cdot P+\Pi_B(1-P)\quad .$$

In Warner's method Π_B is selected to be $1-\Pi_A$. Greenberg <u>et al</u>. [1] compared the efficiency of this unrelated question method with that of Warner and found when Π_A is small that it has better efficiency. Even though the anonymity is theoretically ensured, it is always desirable to find if the respondent in an actual survey shares the same view, and gives truthful answers to sensitive questions. In the present paper, the results of a sample survey pertaining to three sensitive questions are reported. Two randomization techniques are employed--one using the respondent's Social Security Number, and the other using a box containing a fixed proportion of marbles of two colors. The two randomization methods are used to determine if one method performs "better" in ensuring the respondents that the confidentiality is preserved or not. It is assumed in these randomized response methods that if the value of P is high, it becomes difficult for the respondent to give truthful answers if he is prone to feel embarrassment in answering the sensitive questions. To ascertain the truth of such a belief, the following experiment was conducted at three levels of P (.50, .70, and .90). The results are mixed and are reported in the following section.

2. Description of Randomization Methods, the Questionnaire, and Results

The two randomization methods are described in (i) and (ii). The procedure of conducting the survey is described in (iii), and the results are described briefly in (iv) and (v).

(i) <u>Use of social security number (SSN)</u>: The last digit of SSN of the persons in a population is assumed to be distributed uniformly with a probability of 0.1 on each number $0,1,\cdots,9$. The level of P could be controlled through this last digit. For example, if P is to be chosen to have a value 0.5, the respondent is asked to think of the last digit of his SSN and if it is less than or equal to 4 he is asked to select to answer the sensitive question. Similarly, if the respondent is asked to answer the sensitive question if his last digit of SSN is less than or equal to 6, then the probability, P, of choosing to answer a sensitive question is 0.7, and so on.

(ii) <u>Use of marbles</u>: For a given value of P, 10P blue marbles and 10(1-P) yellow marbles were placed in a decorated box. The respondent was free to look inside the box since we attempted to convince him that no tricks were involved. The respondent is asked to shake the box and pick a marble without looking at the box and notice its color. He is then given a questionnaire with instructions to answer the A (sensitive) questions if the marble selected is blue and the B (nonsensitive) questions if the marble selected is yellow.

Before the start of the survey it was felt that as P increases towards unity, there would be a greater proportion of individuals who would give untruthful answers; and also that the randomization procedure using marbles is better to employ as it is easier to convince a respondent that the interviewer does not know which colored marble was selected than to convince him that his SSN is not or will not be known. In addition, the marble procedure could serve as a distracting device whereby the respondent might be unaware of the proportion of blue and yellow marbles in the box and thus the probability of answering the sensitive question.

(iii) Procedure: It was decided to use 120 Cornell University students in the study with 20 different individuals for each of the three P levels (.50, .70, and .90) and the two randomization methods (SSN and box of marbles) for obtaining P. Π_B is assumed to be known and is equal to .30 in the population for this study. Since a comparison of methods and P levels was being made, the question of representativeness of the sample is not involved; however a combination of a P level and method (a treatment) was randomly assigned to an individual subject to the proviso that 20 students were allotted to each treatment. The students were briefed in groups on Simmons' model in an attempt to convince them their anonymity in answering either the sensitive or nonsensitive questions was assured, and of the procedure.

A copy of a questionnaire presented to a student where P=.50 and the randomization method was the use of the Social Security number is as follows:

You are to answer one of two questions in each pair according to the following procedure: If the last digit of your Social Security number is 0, 1, 2, 3, or 4, answer questions A;if the last digit of your Social Security number is 5, 6, 7, 8, or 9, answer questions B. You give only one answer in each pair of questions, and because of the way this method is designed the interviewer cannot tell which questions you have answered. The interviewer only knows the proportion of times the question is answered.

question A: While at Cornell, have you ever stolen money or any article worth over $5.00 from a friend, roommate, employer, or anyone else?

question B: Were you born on the 11th through the 19th of the month?

Yes______ No______

question A: Have you ever either given or received anything tangible, monetary or otherwise, for sexual activity?

question B: Were you born on the 15th through the 23rd of the month?

Yes______ No______

question A: Have you masturbated at least once within the last two weeks?

question B: Were you born between Sept. 13th and Dec. 31st?

Yes______ No______

Your Name______________________________

The A questions are highly sensitive ones to many students, whereas each of the B questions should not be sensitive to any student. Responses to the questions should be independent of each other. The questions may have been so sensitive that the respondent decided not to answer regardless of the value of P.

If this were an actual survey, the name of the respondent would not be needed since information is confidential and it would help to insure his anonymity even further. However, in this study the purpose for requesting the student's name is two-fold. The interviewing was done in groups for ease and convenience, and in many cases the names of respondents were unknown. One intention of the present study was to simulate a personal interview since the randomized response technique is conducted in this manner. Since a respondent's name, or at least place of residence, is usually known if he is an element of a sample, it was desired to have names associated with answers. In addition, if this technique alone insures that the chance of embarassment is removed, the respondents should not have any qualms about signing their names.

(iv) Results: The following tables summarize the findings of the survey pertaining to three pairs of questions. The values indicated represent the number of "yes" responses for 20 respondents.

1st pair of questions

# of "yes" responses

		METHOD	
		SS No.	Marble
P Level	.50	3	2
	.70	4	4
	.90	3	3

2nd pair of questions

# of "yes" responses

		METHOD	
		SS No.	Marble
P Level	.50	6	3
	.70	2	1
	.90	1	0

3rd pair of questions

# of "yes" responses

		METHOD	
		SS No.	Marble
P Level	.50	5	7
	.70	8	10
	.90	8	6

A detailed statistical analysis of these data is given in the Bachelor of Science thesis by Pearl [3]. It is not reported here due to space considerations. An analysis of variance of the above data transformed to arcsines was performed and comparisons were made with the theoretical variance 821/20. The results are more homogeneous than might be expected. Also, a contingency chi-square analysis was performed with comparable results for the two procedures.

(v) Negative Estimates: For each of the three sensitive questions and six treatments, an estimate of Π_A is given below.

1st question

Estimate of Π_A

		METHOD	
		SS No.	Marble
P Level	.50	.000	-.100
	.70	.157	.157
	.90	.133	.133

2nd question

Estimate of Π_A

	METHOD	
	SS No.	Marble
.50	.300	.000
.70	.014	-.057
.90	.411	-.033

3rd question

Estimate of Π_A

		METHOD	
		SS No.	Marble
P Level	.50	.200	.400
	.70	.443	.586
	.90	.411	.300

An undesirable feature of the randomized response methods is that negative estimates may be obtained. The estimating equation is

$$\hat{\lambda} = P\hat{\Pi}_A + (1-P)\Pi_B \ .$$

In order to obtain an estimate $\hat{\Pi}_A$ of Π_A we require

$$(1-P)\Pi_B < \hat{\lambda} < P + (1-P)\Pi_B$$

and since $\hat{\lambda}$ is a random quantity we cannot ensure all the time that it lies between these two prescribed values. Several ad hoc methods can be suggested which are in a sense arbitrary and do not exploit the properties of the randomized procedure. One method of obtaining an estimate of $\hat{\Pi}_A$ between 0 and 1 is to adopt a method of truncation. Define

$$\hat{\Pi}_A = \begin{cases} 0 & \text{if } \hat{\lambda} < (1-P)\Pi_B \\ \hat{\Pi}_A & \text{if } (1-P)\Pi_B < \hat{\lambda} < P + (1-P)\Pi_B \\ 1 & \text{if } \hat{\lambda} < P + (1-P)\Pi_B \ . \end{cases}$$

Another method is as follows. $\hat{\lambda}$ is a maximum likelihood estimate of λ. One can solve the maximum likelihood equation subject to the constraint that λ lie in the interval $((1-P)\Pi_B, P+(1-P)\Pi_B)$.

In the original work (Pearl [3]), a detailed table is presented illustrating values of P, Π_B, and n necessary to be at least 95% certain that non-negative estimates of $\hat{\Pi}_A$ are obtained for a minimum value of Π_A or of λ. These results are not reproduced here because of space considerations.

3. Discussion

Seven out of nine students interviewed in one group refused to sign their names on their completed questionnaires. However, since their responses did not alter the results more than a negligible amount, their answers were included. Only one student, who was in that same group, outwardly refused to respond after reading the questionnaire. It should be noted that his P level was 0.9 and the randomization method was the use of his Social Security number.

For the first and third pair of questions, there was no significant difference in the number of "yeses" at the 5% level among the three P levels and between the two randomization methods used for all three analyses done. These results are contrary to the authors' hypotheses. For the second pair of questions, however, there was a significant difference at the 5% level among P levels for each of the three analyses. For the first analysis of variance, there also was a significant difference at the 5% level between the randomization methods used. However, only at the 25% level was there a significant difference among methods for the other two analyses. Perhaps for some sensitive questions, the number of untruthful answers will increase as P approaches unity; for other questions, the value of P will make no difference. Unfortunately, there is no way of knowing whether it will make a difference prior to the survey. Overall, the analyses show no significant difference at the 5% level in the number of "yeses", with one exception, between the two methods.

The absence of any trend in the number of "yeses" in the first and third pair of questions might also be explained by the following reasons:

1) Repondents decide to answer truthfully or untruthfully independent of their P level.
2) Students influence the responses of one another since they were interviewed as a group.
3) Even a P level of .50 is too high in that respondents will answer untruthfully even at this level.
4) Although students were instructed to answer the questionnaire as if an interviewer from Gallup or Harris polls was there instead of a student interviewer, it is the authors' feeling that the students answered honestly at all P levels and treatments since they wanted to help a peer with her study.

Perhaps for these reasons and/or others not mentioned, the randomized response technique gave similar responses with different P levels and randomization methods. One should use whatever method produces the most truthful information. Perhaps the block total response or a randomized form of it [4] are procedures which have advantages over the randomized response technique with regard to eliciting reliable information in certain situations.

[1] This work is based on the first author's thesis for a B.S. degree at Cornell University. The authors appreciate the efforts of Dr. K. C. Rao for editorial help and suggestions.

References

[1] Greenberg, B.G., Abdul-Ela, Abdel-Latif, A., Simmons, W.R. and Horvitz, D.G., "The Unrelated Question Randomized Response Model, Theoretical Framework", Journal of the American Statistical Association, 64, (1969), 520-39.

[2] Horvitz, D.G., Shah, B.V. and Simmons, W.R., "The Unrelated Question Randomized Response Model", Proceedings of the Social Statistics Section, American Statistical Association, (1969), 65-72.

[3] Pearl, R.L., "Varying Levels of Probability for Selecting Sensitive Questions Using the Randomized Response Technique". B.S. Thesis in Statistics and Biometry, Biometrics Unit, Cornell University (June, 1975).

[4] Smith, L.L., Federer, W.T. and Raghavarao, D., "A Comparison of Three Techniques for Eliciting Answers to Sensitive Questions", Proceedings of the Social Statistics Section, American Statistical Association, (1974), 447-52.

[5] Warner, S.L., "Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias", Journal of the American Statistical Association, 60, (March 1965), 63-9.