THE OMITTED DIGIT RANDOMIZED RESPONSE MODEL FOR TELEPHONE APPLICATIONS

Stanley L. Warner, York University

1. INTRODUCTION

Recent articles by Godambe (1980) and Adhikari, Chaudhuri and Vijayan (1984) illustrate the continuing progress in the development and understanding of theoretical randomized response model efficiencies. Innovations in practical applications continue as well; experiments with telephone surveys such as that by Stem and Steinhorst (1984) show the interest but also the difficulty in making randomized response practical for the less expensive survey methods. Since current problems such as those related to social disease also suggest that the scope for applications is increasing, a question of importance is how to improve the practical as well as the theoretical efficiency of randomized response procedures.

Efficiency of randomized response in practical applications depends upon the interviewee's understanding that privacy can be protected by randomization. As emphasized by recent experimental telephone applications, how privacy is protected through randomization is a difficult concept to convey to general populations. Explanations are long and complicated; yet many interviewees remain skeptical. Interviewees often act as if they believe the random device simply determines whether they are required to reveal or not to reveal a secret. Thus, for example, they sometimes appear less willing to cooperate after the random outcome calls for reporting the word or symbol that would be stigmatizing if there were no randomization. Protecting secrecy through randomization is not yet a familiar concept.

The concept of secrecy most familiar to general populations is that something is revealed or not revealed with probability one. If the interviewer cannot appeal to this fundamental notion in his explanation of randomized response, explanations may be more difficult than they need to be. For those not familiar with probability, requiring the "actual answer" or "actual value" to be reported with probability P may be misinterpreted as requiring that the secret which the interviewee wishes to hide will be revealed with probability P. A simple procedure in which a hidden symbol is never revealed might be easier to explain.

While the question as to which explanations work best is a matter for experimentation, both explanations and experimentation have largely been restricted to familiar models which are of the form "report the actual answer with probability P and another answer with probability (1-P)." This form was assumed for an example in Warner (1965), "the interviewee makes a statement that is true with probability P as to which of two groups he belongs." It is the natural form of the appealing and widely used unrelated question model developed in Greenberg, Abul-Ela, Simmons and Horvitz (1969), "the respondent is asked to reply 'Yes' or 'No' to one of two statements selected on a probability basis." It is again the form used for most theoretical developments such as those of Godambe (1980) and Adhikari, Chaudhuri and Vijayan (1984) cited in the introduction. It is also the form commonly used for most applications including the telephone interviews of Stem and Steinhorst (1984), "if the last digit of the 'selected phone number' is 3 or above, give your actual answer to the question."

The requirement that interviewees report the actual answer according to the outcome of a random device has consistently been associated with some concerns, however. In the original development of the unrelated question model reported in Greenberg, Abul-Ela Simmons and Horvitz (1969), concern was expressed that

> "Even if membership in Y is not stigmatizing, a person might deny membership therein because he knows that a 'Yes' answer might be embarrassing whereas with a'No' answer there is never any possibility of embarrassment."

This point evidently applies to all respondents whether or not they are in the stigmatizing population, and it has been a continuing concern. This concern is expressed again in the cited telephone application. Here, in commenting on the need to facilitate general understanding, the authors emphasize that

> "...Even more critical is the task of convincing the respondents of the importance of supplying the surrogate answer even though they may be innocent of any sensitive behavior."

In addition to investigator apprehension that interviewees may hesitate to say "Yes" if that may be interpreted as the "actual answer" and so might be embarrassing, there are evidently more general concerns regarding how symmetrical and tamper-proof the procedures appear to the interviewee. While models and explanations in use are accepted improvements over the earlier examples, continuing concerns suggest that the search for improvements should continue. The next section outlines a possibility.

2. THE OMITTED DIGIT MODEL

The first requirement is a random device which is easy to understand, widely available, and adaptable for a broad range of applications. As concluded by Stem and Steinhorst for their telephone experiments, random draws from subsets of the digits 0 through 9 appear to be particularly convenient. Depending on the application, random digits might be taken from random numbers in a textbook (any digit), serial numbers from paper money (last digit), page numbers from a book (next to last digit), numbers on a die (only digit), or telephone numbers (last digit). Subsets of the digits if needed are easily described.

Practical approximations to random draws from

the above sources are straightforward. For example, in the case of telephone numbers, instructions might specify opening the book, looking away while placing a pencil point on the page, and taking the last digit of the nearest number above and to the right of the pencil point. The discrete uniform distribution being approximated will not be perfect, but its error should be unsystematic and small relative to other errors. Complications from non-equal probabilities for the digits are ignored in this paper.

Possible applications and procedures for the model are introduced with two examples. The first considers the estimation of an income distribution; the second considers the estimation of a population proportion. In each case, interviews are presumed to be accomplished by telephone, with random digits identified with the last digit of a telephone number randomly drawn from a telephone directory.

For the income distribution example, suppose that a 10-class distribution is to be estimated. Each interviewee is asked to write down and keep secret the digit corresponding to the income class to which he belongs. The procedure is based on the interviewee's reporting a randomly drawn number from the remaining 9 digits not written down.

As an illustration, if instructions paired "0" with incomes greater than 0 but less than $10,000, "1" with incomes greater than $10,000 but less than $20,000, and "9" with incomes greater than $90,000, an interviewee with $25,000 would be expected to choose and write down the digit "2" and then draw and report a random number from the set {1,3,4,5,6,7,8,9}. The chosen digit "2" would thus be omitted from the set of possible digits to report. In particular, instructions might specify that after writing down the digit to be kept secret, the interviewee draw a random telephone number, examine the last digit of that number, and report that digit provided that it differed from the digit written down. If the last digit of a number were the same as the digit written down, the telephone number next in the listing would be examined.

For the second example, suppose that the problem is to estimate the proportion of the population that has attribute B. In this case, only the digits 1 through 6 might be used, with the interviewee instructed to choose and write down one of the digits 1, 2, or 3 if he has attribute "B" and to choose and write down one of the digits 4, 5, or 6 if he does not have attribute "B". The procedure is the same as in the first example, except that, in addition to the digit chosen by the interviewee, the digits 0, 7, 8 and 9 are also omitted from the set of possible digits to report. Thus, if an individual with attribute "B" chose the digit "3", the digit reported would be randomly drawn from the set {1,2,4,5,6}.

While these examples illustrate applications of the general procedure to a 10-class problem and a 2-class problem, in-between numbers of classes are obviously possible. A 3-class income-distribution problem, for example, could be based on identifying digits 1 or 2 with the first class, digits 3 or 4 with the second class, and digits 5 or 6 with the third class. Some flexibility in design is apparent by deciding on the number of classes and the number of digits per class for a given application. Of some advantage for understanding by diverse users, simple estimates and variances for general applications are easily summarized.

In particular, for r classifications, the general problem is to find

$$\Pi_j = \text{probability of being in group } j$$

or some linear function of the Π_j, $j = 1,2,\ldots,$ r with

$$\sum_{j=1}^{r} \Pi_j = 1. \tag{1}$$

Let G_j be the set of m_j digits assigned to the jth group with $G_j \cap G_k = 0$ for $j \neq k$ and $G = \bigcup_{j=1}^{r} G_j \subset \{0,1,2,3,4,5,6,7,8,9\}$ so that

$$\sum_{j=1}^{r} m_j = m \leq 10. \tag{2}$$

Then, identifying X_i as the digit written down and Y_i as the digit reported by the ith interviewee, $X_i \in G_k$ implies the ith interviewee is a member of the kth group, and that Y_i will be chosen from the (m-1) digits in the set $G-\{X_i\}$ for $k = 1,2,\ldots,r$.

The conditional probability that the digit reported is in G_j given that the digit written is in G_k is thus given by

$$P(Y_i \in G_j | X_i \in G_k) = (m_j - \delta_{jk})/(m - 1) \tag{3}$$

with $\delta_{jk} = 1$ if $j = k$ and $= 0$ otherwise. The probability that the interviewee i reports a digit in G_j is then

$$P_j = P(Y_i \in G_j) = \sum_{k=1}^{r} \Pi_k (m_j - \delta_{jk})/(m-1)$$

$$= (m_j - \Pi_j)/(m-1) \tag{4}$$

with

$$(1 - P_j) = (m - 1 - m_j + \Pi_j)/(m - 1). \tag{5}$$

Supposing n observations and n_j of the reported n digits are in G_j, $j = 1,2,\ldots r$, estimates of the P_j are

$$\hat{P}_j = n_j/n, \tag{6}$$

with variances using (4) and (5) given by

$$VAR(\hat{P}_j) = [1/(m - 1)]^2 (m_j - \Pi_j)$$

$$(m - 1 - m_j + \Pi_j)/n \tag{7}$$

and covariances for $j \neq k$ given by

$$COV(\hat{P}_j, \hat{P}_k) = -[1/(m - 1)]^2 (m_j - \Pi_j)$$

$$(m_k - \Pi_k)/n. \tag{8}$$

Expression (4) shows simple linear estimates of the Π_j to be

$$\hat{\Pi}_j = m_j - (m - 1)\hat{P}_j \qquad (9)$$

for which variances and covariances are given from (7) and (8) as

$$\text{VAR}\ (\hat{\Pi}_j) = (m_j - \Pi_j)\ (m - 1 - m_j + \Pi_j)/n \qquad (10)$$

and

$$\text{COV}\ (\hat{\Pi}_j, \hat{\Pi}_k) = -(m_j - \Pi_j)\ (m_k - \Pi_k)/n \qquad (11)$$

for $j = 1,2,\ldots,r$ and $k \neq j$. Since (9) can be negative, evidently improved estimates are possible as have been suggested in other contexts for randomized response estimates.

As an illustration of the variances implied by (10) for simple problems, suppose that $r = 2$ and that $m_1 = m_2 = m/2$. Then (10) becomes

$$\text{VAR}\ (\hat{\Pi}_1) = [(m/2) - \Pi_1][(m/2) - (1 - \Pi_1)]/n.$$

Thus, if $\Pi_1 = .50$ and $m_1 = m_2 = 2$, sample sizes of 225, 900, and 3600 would be required to respectively provide standard deviations of 0.10, 0.05, and 0.025. If $\Pi_1 = .50$ and $m_1 = m_2 = 3$, then sample sizes of 625, 2500, and 10,000 would be required for standard deviations of 0.10, 0.05, and .025.

As a second illustration, if $r = 10$ and each $m_j = 1$, then

$$\text{VAR}\ (\hat{\Pi}_j) = (1 - \Pi_j)(8 + \Pi_j)/n.$$

Supposing in this case that each $\Pi_j = 1/10$, sample sizes of 729, 2916, and 11,664 are required for standard deviations of 0.10, 0.05, and 0.025.

Not much information is extracted per observation, and for most problems flexibility is limited. For the dichotomous problem with equal digits per class, more than two or three digits per class would evidently result in prohibitive variances. For the ten-class problem, only one digit is possible, and the calculations again show the large samples required for precision. If the procedures save interviewer time and increase interviewee cooperation, however, in many applications more efficient estimates might be achieved for a given expenditure.

3. CONCLUSIONS

The appeal and flexibility of the unrelated question and similar models made randomized response practical, but for some applications, such as telephone interviewing, other models may prove easier to explain. The model of this paper allows r-class problems to be estimated by procedures which address previously expressed concerns. The digit that is kept secret is never reported; interviewees are not asked to respond truthfully or not by random outcome; and the procedure is symmetrical for persons in different groups. In particular, reporting a digit assigned to the stigmatizing population is never embarrassing because reported digits are never secret digits.

Randomized response procedures will not likely be efficient for general applications until there is general familiarity with the concept of protecting privacy through randomization. Procedures such as opinion polling became more efficient as knowledge of the survey concepts behind those procedures became widespread. The importance of more widespread familiarity with randomized response procedures emphasizes the need for experimentation to identify procedures which can be explained in similar terms for a variety of applications. Previous models might still prove best, but the omitted digit model provides an additional possibility to consider.

ACKNOWLEDGMENTS

The author would like to thank G.P. Patil, M. Zelen, and other participants at the 1985 Conference on Weighted Distributions held at Pennsylvania State University for helpful suggestions.

REFERENCES

Adhikari, A.K., Chaudhuri, A. and Vijayan, K. (1984). Optimum sampling strategies for randomized response trials. *Int. Statist. Rev.* 52, 115-125.

Godambe, V.P. (1980). Estimation in randomized response trials. *Int. Statist. Rev.* 48, 29-32.

Greenberg, B.G., Abul-Ela, A.A., Simmons, W.R., and Horvitz, D.G. (1969). The unrelated question randomized response model: theoretical framework. *J. Am. Statist. Assoc.* 64, 520-539.

Stem, D.E. and Steinhorst, R.K. (1984). Telephone interview and mail questionnaire applications of the randomized response model. *J. Am. Statist. Assoc.* 79, 555-564.

Warner, S.L. (1965). Randomized response: a survey technique for eliminating evasive answer bias. *J. Am. Statist. Assoc.* 60, 63-69.