ON SCRAMBLED RESPONSE OF SENSITIVE QUANTITATIVE DATA

M. Ahsanullah and B. H. Eichhorn, Rider College

1. INTRODUCTION

The randomized response method for estimating the proportion of people with a sensitive characteristic has been extensively studied since its introduction by Warner (1965). The object is to reduce the frequency of false answers by giving the respondent two questions, one of which is the sensitive one. The respondent then selects one of the questions using some specified randomization device. The interviewer does not know which question has been selected, and so does not know whether the answer he receives is for the sensitive or non-sensitive question. This hopefully will make the respondent less likely to give a false answer. A survey of the field may be found in the paper by Horvitz, et al. (1975).

Greenberg, et al. (1971) extended the method to the case where the responses to the sensitive question are quantitative, rather than a simple 'Yes' or 'No.' The respondent selects, by means of a randomization device, one of two questions; the sensitive question, and an unrelated question, the answers to which are of about the same order of magnitude as for the sensitive question. However, there are several difficulties which arise when using this unrelated question method. The main one is in choosing the unrelated question. As Greenberg et al. (1971) notes, it is essential that the mean and variance of the responses to the unrelated question be close to those for the sensitive question; otherwise, it will often be possible to recognize from the response which question was selected. However, the mean and variance of the responses to the sensitive question are unknown, making it difficult to choose a good unrelated question. A second difficulty is that in some cases the answers to the unrelated question may be more rounded or regular, making it possible to recognize which question was answered. For example, Greenberg, et al. (1971) considered the sensitive question: about how much money did the head of this household earn last year This was paired with the question: about how much money do you think the average head of a household of your size earns in a year An answer such as $26,350 is more likely to be in response to the sensitive question, while an answer such as $18,618 is almost certainly in response to the sensitive question. A third difficulty is that some people will be hesitant in disclosing their answer to the sensitive question (even though they know that the interviewer cannot be sure that the sensitive question was selected). For example, some respondents may not want to reveal their income even though they know that the interviewer can only be 3/4 certain, say, that the figure given is the respondent's income.

In the Scrambled Randomized Response method which was introduced by Eichhorn and Hayre (1) these difficulties are no longer present. Each respondent scrambles his response x by multiplying it by a random scrambling variable S and only then reveals the scrambled result y = X x S to the interviewer. The mean of the response, E(X) can be estimated from a sample of y's and a known scrambling distribution of S. Here we explore two questions concerning this scrambled response method.

The first is, can we improve the estimates of the expected value of X, the sensitive variable, if we knew its underlying distribution The second question is, are there other parameters besides the mean which can be obtained from scrambled responses, like the median or percentiles In section 2 we give some improved estimating procedures for a special family of distributions for X. It seems likely that other families exist where similar improvements may be made. In section 3 we introduce another scrambling method which may be used for certain sensitive questions and for which we are able to retrieve information to estimate the cumulative distribution function of X, F(X). This way we may estimate the median or any other parameter of F(X).

2. ESTIMATION OF THE MEAN WHEN x IS UNIFORMLY DISTRIBUTED

Suppose $x \overset{d}{=} U(0,\theta)$, i.e., the pdf f(x) of X is given by

$$f(x) = \frac{1}{\theta} \quad 0<x<\theta \ .$$

We assume $S \overset{d}{=} U(0,2)$ and Y = SX. We are interested in estimating θ when y's are observable.

$$P[Y < y] = \frac{y}{2\theta}\left(n\frac{2\theta}{y} + 1\right).$$

It can be shown that $Y_{n,n} = \max(Y_1, Y_2, \ldots, Y_n)$ is the sufficient statistic for θ. Let $f_n(y)$ be the pdf of $Y_{n,n}$ thus

$$f_n(y) = \frac{ny^{n-1}}{(2\theta)^n}\left(\ln\frac{2\theta}{y}\right)\left(\ln\frac{2\theta}{y} + 1\right)^{n-1}, \quad 0\le y\le 2\theta$$

$$E(Y_{n,n}) = \int_0^{2\theta} \frac{ny^n}{(2\theta)^n}\left(\ln\frac{2\theta}{y}\right)\left(\ln\frac{2\theta}{y} + 1\right)^{n-1} dy$$

$$= 2n\theta \int_0^{\infty} (1+t)^{n-1} e^{-(n+1)t}\, t\, dt$$

$$= 2n\theta \int_0^{\infty} (1+t)^{n} e^{-(n+1)t}\, dt - \int_0^{\infty} (1+t)^{n} e^{-(n+1)t}\, dt$$

$$= 2n\theta\, c_n' \ ,$$

where $c_n' = a_n - b_n$,

$$a_n = \int_0^{\infty} (1+t)^n e^{-(n+1)t} dt = \sum_{r=0}^{n} \frac{n!}{(n-r)!}\frac{1}{(n+1)^{r+1}}$$

$$b_n = \int_0^\infty (1+t)^{n-1} e^{-(n+1)t}\, dt =$$

$$\sum_{r=0}^{n-1} \frac{(n-1)!}{(n-1-r)!} \frac{1}{(n+1)^{r+1}}$$

The minimum variance unbiased estimator of θ is

$$T_n \text{ where } T_n = Y_{n,n}/(2nc_n') \quad .$$

Using the following expression of $E(Y_{n,n}^2)$, the variance of T_n can easily be calculated.

$$E(Y_{n,n}^2) = \int_0^{2\theta} y^2 \frac{ny^{n-1}}{(2\theta)^n} \left(\ln \frac{2\theta}{y} + 1\right)^{n-1} dy$$

$$= 4n\theta^2 \int_0^\infty t(1+t)^{n-1} e^{-(n+2)t}\, dt$$

$$= 4n\theta^2\, d_n' \text{ , where}$$

$$d_n' = a_n' - b_n' \text{ ,}$$

$$a_n' = \int_0^\infty (1+t)^n e^{-(n+2)t}\, dt =$$

$$\sum_{r=0}^{n} \frac{n!}{(n-r)!} \frac{1}{(n+2)^{r+1}}$$

$$b_n' = \int_0^\infty (1+t)^{n-1} e^{-(n+2)t}\, dt =$$

$$\sum_{r=0}^{n-1} \frac{(n-1)!}{(n-1-r)!} \frac{1}{(n+2)^{r+1}}$$

3.

Let X be the sensitive quantity that we are interested in, let us assume that X has a continuous distribution with an unknown C.D.F., $F_X(x)$. Let A be a guess of the median value of X. We let the respondents scramble their value of X by subtracting A from it and then multiplying it by S. As

$$S = \begin{cases} +1 \\ -1 \end{cases} \text{ with prob } \begin{matrix} P \\ 1-P \end{matrix}, \text{ p is a value dif-}$$

ferent from 0, 1 or 0.5 a good value for P is $\frac{2}{3}$. Let us observe a random sample of n respondents with responses x_i, $i = 1,\ldots,n$ and let $Y_i = S_i(x_i - A)$ be their scrambled response that we get to observe. Let us group the y values into positive ones which will be denoted by $y_1^+, \ldots, y_r^+$, their number being r, and the absolute value of the negative y's we'll denote by $\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_{n-r}$. The first stage will be to estimate the percentile rank of A or alternatively the probability of X exceeding A.

Denote by $\alpha = P(X > A)$ an estimate of α given by a method like the one for randomized response; $\hat{\alpha} = \frac{\frac{r}{n} - (1-P)}{2P-1}$ restricting α to be in [0,1]. If $\hat{\alpha}$ is greater than 1 we use $\alpha = 1$ and if α is less than 0 we use $\alpha = 0$. Now we continue to estimate $F_X(x)$ for different x values. Let us define the C.D.F.'s of y^+ and y^- as

$$F_y^+(t) \text{ and } F_y^-(t)$$

$$\text{and let } F_x^+(t) = \begin{cases} \frac{F_X(A+t)-F_X(A)}{1-F_X(A)} & \text{for } t > 0 \\ 0 & \text{for } t < 0 \end{cases}$$

be the conditional distribution of X-A given $X > A$ and

$$F_x^-(t) = \begin{cases} \frac{F_X(A)-F_X(A-t)}{F_X(A)} & \text{for } t > 0 \\ 0 & \text{for } t < 0 \end{cases}$$

be the conditional distribution of X-A given $X < A$.

$$\text{So } F_x^+(t) = \begin{cases} \frac{F_X(A+t)-(1-\alpha)}{\alpha} & \text{for } t > 0 \\ 0 & \text{for } t < 0 \end{cases}$$

$$\text{and} \quad F_x^-(t) = \begin{cases} \frac{1-\alpha-F_X(A-t)}{1-\alpha} & \text{for } t > 0 \\ 0 & \text{for } t < 0 \end{cases}$$

we have

$$F_y^+(t) = P_1 F_x^+(t) + (1-P_1)F_x^-(t) \quad \text{for } t > 0$$

$$F_y^-(t) = (1-P_2)F_x^+(t) + P_2 F_x^-(t)$$

as P_1 is the posterior probability that $X>0$ given $y>0$ and P_2 is the posterior probability that $X<0$ given $y<0$

$$P_1 = \frac{\alpha P}{\alpha P + (1-\alpha)(1-P)} \quad ; \qquad \hat{P}_1 = \frac{\hat{\alpha} P}{\hat{\alpha} P + (1-\hat{\alpha})(1-P)}$$

estimated by

$$P_2 = \frac{(1-\alpha)P}{(1-\alpha)P + \alpha(1-P)} \quad ; \qquad \hat{P}_2 = \frac{(1-\hat{\alpha})P}{(1-\hat{\alpha})P + \hat{\alpha}(1-P)}$$

$F_y^+(t)$ and $F_y^-(t)$ can be estimated from the empirical distributions of y^+ and y^-, from these we can solve for estimates of $F_x^+(t)$ and $F_x^-(t)$ and together with α we estimate $F_X(x)$.

If $F_X(x)$ can be estimated, so can any other parameters of X.

Let us take a specific value $t > 0$.

$$\widehat{F_y^+}(t) = \frac{\text{number of } y^+ \text{ values} < t}{r}$$

$$\widehat{F_y^-}(t) = \frac{\text{number of } y^- \text{ values} < t}{n-r}$$

To solve for $\widehat{F_x^+}(t)$ and $\widehat{F_x^-}(t)$ we solve the equations

$$\widehat{F_y^+}(t) = \hat{P}_1 \widehat{F_x^+}(t) + (1-\hat{P}_1)\widehat{F_x^-}(t)$$

$$\widehat{F_y^-}(t) = (1-\hat{P}_2)\widehat{F_x^+}(t) + \hat{P}_2\widehat{F_x^-}(t)$$

and finally

$$\widehat{F_X(x)} = \begin{cases} 1-\hat{\alpha}-(1-\hat{\alpha})\widehat{F_x^-}(A-x) & \text{for } x < A \\ 1-\hat{\alpha} & \text{for } x = A \\ 1-\hat{\alpha}+\hat{\alpha}\widehat{F_x^+}(x-A) & \text{for } x > A \end{cases}$$

REFERENCES

[1] Eichhorn, B., and Hayre, L. Scrambled Randomized Response Methods for Obtaining Sensitive Quantitative Data. Journal of Statistical Planning and Inference 7 (1983) 307-316.

[2] Greenberg, B.G., Kuebler, R.R., Abernathy, J.R., and Horvitz, D.G. (1971). Application of the Randomized Response Technique in Obtaining Quantitative Data. J. Am. Statist. Assoc., 66, 243-250.

[3] Horvitz, D.G., Greenberg, B.G., and Abernathy, J.R. (1975). Recent Developments in Randomized Response Designs. A Survey of Statistical Designs and Linear Models, J. N. Srivastava, ed. North-Holland Publishing Company, New York.

[4] Pollock, K.H. and Bek, Y. (1976). A Comparison of Three Randomized Response Models for Quantitative Data. J. Am. Statist. Assoc., 71, 884-886.

[5] Warner, S.L. (1965). Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias. J. Am. Statist. Assoc., 60, 63-69.

[6] Warner, S.L. (1971). The Linear Randomized Response Model. J. Am. Statist. Assoc., 66, 884-888.