ESTIMATION FOR A NON-SENSITIVE VARIATE OVER SENSITIVE SUB-POPULATIONS OBSERVED USING RANDOMIZED RESPONSE

P.D. Bourke
University College, Cork, Ireland

KEY WORDS: Survey Technique; Maximum Likelihood; EM Algorithm.

Abstract: Two variates are observed for each member of a sample. Variate S is sensitive and is observed using Randomized Response, while variate X is non-sensitive and is directly observed. It is required to estimate the distribution of X for each category of S. Simple estimators are developed without making any distributional assumptions about X. If distributional assumptions are made it is shown that the EM algorithm may be used to compute Maximum Likelihood estimates. Based on computational comparisons between the estimators it is concluded that the simple estimators perform well, particularly for large sample sizes.

1. INTRODUCTION

In surveys of human populations, one or more of the variates under study may be embarrassing or stigmatizing in some way. The randomized response (RR) technique introduced in Warner (1965) is by now well known as a method of providing each respondent with a degree of protection, and thereby encouraging greater co-operation by respondents. Recent reviews of randomized response are contained in Chaudhuri and Mukerjee (1988), Fox and Tracy (1986), Deffaa (1982) and Boruch and Cecil (1979).

Most of the methodological work on randomized response has been concentrated on the design and analysis of the RR procedures themselves, ranging from extensions to the case of multinomial and multivariate estimation (e.g. Abul-Ela et al. (1967), Bourke and Dalenius (1973), Tamhane (1981), Bourke (1982)) to the development of general RR models (Warner (1971), Loynes (1976), Godambe (1980)). However little attention has been paid to the presence of other non-sensitive variates in a survey which includes an RR procedure. In this paper we consider the problem of estimating the distribution of a non-sensitive variate for each category of a sensitive variate. For example we may be concerned with the variate number of years of full-time education for women who engage in shop-lifting and for women who do not.

In Section 2 estimators are developed without any distributional assumptions about the non-sensitive variate, whereas in Section 3 such assumptions are made, and the EM algorithm (Dempster, Laird and Rubin (1977)) is used to compute ML estimates. Some empirical analyses of the estimators are given in Section 4.

2. ESTIMATION WITHOUT DISTRIBUTIONAL ASSUMPTIONS

In this Section we develop estimators for various features of the distribution of X (denoting the non-sensitive variate) for each level of S (denoting the sensitive variate) without making any distributional assumptions about X. For clarity of presentation, we assume here that S has two levels (one stigmatizing and one non-stigmatizing) but there is little difficulty in dealing with more than two levels for S, using an appropriate multinomial RR design.

Initially, we consider the estimation of the proportions in various categories of variate X for each level of S. Suppose that there are m categories of X and we wish to estimate the proportions $\{Q_{1j}\}$, $j = 1, 2, \ldots, m$, when S is at level 1, and similarly for Q_{2j}. We note that $\Sigma Q_{1j} = \Sigma Q_{2j} = 1$. Let π_k be the proportion of the population at level k of S, and π_k is not assumed known. The data available are the n ordered pairs (x_i, r_i), x_i denoting the value of X for respondent i and r_i denoting the response given in the RR procedure for S by respondent i. Let z_i denote the level of S for respondent i, and z_i takes on values 1 and 0. We assume here that there are just two distinct values for r_i, denoted by 1 and 0. (There are RR designs where the number of response categories exceeds the number of levels of S, and again there is little difficulty in adapting to these designs). Let $t_j|1$ denote the probability that a randomly selected respondent is in category j of X given that the response $r_i = 1$ has been observed in the RR procedure, and similarly for $t_j|0$. Then,

$$t_j|1 = Q_{1j}\, P(z_i=1|r_i=1) + Q_{2j}\, P(z_i=0|r_i=1)$$

$$t_j|0 = Q_{1j}\, P(1|0) + Q_{2j}\, P(0|0)$$

where, with an obvious contraction of notation, $P(1|0)$ denotes the probability that the respondent has S at level 1 given that the response $r_i = 0$ has been observed. Rewriting :

$$\begin{vmatrix} t_j|1 \\ t_j|0 \end{vmatrix} = \begin{vmatrix} P(1|1) & P(0|1) \\ P(1|0) & P(0|0) \end{vmatrix} \begin{vmatrix} Q_{1j} \\ Q_{2j} \end{vmatrix} \qquad (2.1)$$

for $j = 1, 2, \ldots, m$. One can estimate $t_j|1$ and $t_j|0$ from the proportions in category j of X for those with $r_i = 1$ and $r_i = 0$ respectively in the RR procedure. The quantities such as $P(z_i=1|r_i=1)$ depend on π_k and on the parameters of the chosen RR procedure. For example if the RR procedure used is the Simmons unrelated question design (Greenberg et al (1969)), then the matrix in (2.1) is :

$$\begin{matrix} (p+(1-p)\beta)\pi_1/\lambda & (1-p)\beta\,\pi_2/\lambda \\ (1-p)(1-\beta)\pi_1/(1-\lambda) & [p+(1-p)(1-\beta)]\pi_2/(1-\lambda) \end{matrix} \qquad (2.2)$$

where p is the known probability that the randomizing device selects the sensitive question, β is the known proportion having the unrelated characteristic, and

$$\lambda = p\,\pi_1 + (1-p)\beta \ .$$

Since the π_k's are unknown, they must be estimated from the RR data (the r_i values) using whatever estimators are appropriate to the chosen RR procedure. For the Simmons unrelated question design, the usual estimator for π_1 is

$$\hat{\pi}_1 = [(n_1/n) - (1-p)\beta]/p \qquad (2.3)$$

where n_1 is the number replying 'YES' (i.e. r_i = 1) in the RR procedure, and n is the sample size used. Thus inserting the RR estimates for the π_k's in (2.2) one could extract estimates for Q_{1j} and Q_{2j} from (2.1), provided that the matrix is non-singular.

If the variate X is quantitative one may wish to estimate, for each level k of S, its mean and variance which we denote by μ_k, σ_k^2, k = 1, 2. For the estimation of μ_1 and μ_2 we have, corresponding to (2.1)

$$\begin{vmatrix} \mu|1 \\ \mu|0 \end{vmatrix} = \begin{vmatrix} P(1|1) & P(0|1) \\ P(1|0) & P(0|0) \end{vmatrix} \begin{vmatrix} \mu_1 \\ \mu_2 \end{vmatrix} \qquad (2.4)$$

where $\mu|1$ denotes the mean of X given that r_i=1 has been observed for a randomly selected respondent, and similarly for $\mu|0$. The quantities $\mu|1$ and $\mu|0$ may be estimated by $\bar{x}_1$, $\bar{x}_0$ the sample means of X for those with r_i = 1 and 0 respectively, and estimation of μ_1, μ_2 then proceeds as described earlier for Q_{1j} and Q_{2j}.

The procedure suggested for estimating σ_1^2 and σ_2^2 is similar. Let $\sigma^2|1$ denote the variance of X given that r_i = 1 has been observed for a randomly selected respondent, and similarly for $\sigma^2|0$. We have

$$\begin{aligned} \sigma^2|1 &= V_{11}\,P(1|1) + V_{21}\,P(0|1) \\ \sigma^2|0 &= V_{10}\,P(1|0) + V_{20}\,P(0|0) \end{aligned} \qquad (2.5)$$

where

$$V_{11} = \sigma_1^2 + (\mu_1 - \mu|1)^2$$

$$V_{21} = \sigma_2^2 + (\mu_2 - \mu|1)^2$$

$$V_{10} = \sigma_1^2 + (\mu_1 - \mu|0)^2$$

$$V_{20} = \sigma_2^2 + (\mu_2 - \mu|0)^2$$

The quantities $\sigma^2|1$, $\sigma^2|0$ may be estimated by s_1^2, s_0^2 the sample variances of the X values for those respondents with r_i = 1 and 0 respectively. Using the estimators for the other quantities (μ_1, $\mu|1$, $P(1|1)$, etc) described earlier, it is possible to estimate σ_1^2, σ_2^2 .

The estimators presented in this section seem intuitively reasonable. However, formal analysis of these estimators is difficult, and will not be attempted here. In Section 4, some empirical results on the performance of these estimators are presented, and comparisons made with the ML estimates of Section 3.

3. ESTIMATION WITH DISTRIBUTIONAL ASSUMPTIONS FOR THE NON-SENSITIVE VARIATE

In this section we consider the estimation of the distribution of X where we are willing to make assumptions about the distributional form of X at each level of S. We shall use the EM algorithm to compute ML estimates of the parameters of the distributions. (Two distinct EM formulations for RR data are described in Bourke and Moran (1984 , 1986)).

The quantities x_i, r_i, z_i have the same meanings as described in Section 2, and z_i differs from x_i, r_i in that it is not observed. Let g_1, g_2 denote the p.d.f.s of X for the sub-populations corresponding to the two levels of S. The log-likelihood for the data $(x_1,r_1,z_1),\ldots,(x_n,r_n,z_n)$, if the z_i were observed, is

$$\log L(\pi_1,g_1,g_2) \propto \sum_i \{ z_i[\log \pi_1 + \log g_1(x_i)] + (1-z_i)[\log \pi_2 + \log g_2(x_i)] \} \qquad (3.1)$$

To proceed further, it is necessary to assume parametric forms for g_1 and g_2. To illustrate the procedure, we will assume that g_1 is Poisson (m_1) and g_2 is Poisson (m_2). In the E step of the algorithm, each z_i is replaced by its expectation z_i^* conditional on the observed (x_i,r_i) and the current parameter estimates. Thus, at iteration t

$$\begin{aligned} z_i^* &= E[z_i|x_i,\ r_i,\ (\pi_1=\pi_1^{(t)}),\ (m_1=m_1^{(t)}),\ (m_2=m_2^{(t)})] \\ &= P(z_i = 1 \mid x_i,\ r_i, \pi_1^{(t)},\ m_1^{(t)},\ m_2^{(t)}) \\ &= \frac{P(x_i,\ r_i|(z_i=1),\ \pi_1^{(t)},\ m_1^{(t)},\ m_2^{(t)})\ \pi_1^{(t)}}{P(x_i,\ r_i)} \\ &= \frac{g_1(x_i)P(r_i|z_i=1)\ \pi_1^{(t)}}{g_1(x_i)P(r_i|z_i=1)\pi_1^{(t)} + g_2(x_i)P(r_i|z_i=0)\pi_2^{(t)}} \end{aligned} \qquad (3.2)$$

using the fact that x_i and r_i are independent of

TABLE 1 : Comparison of the Estimators:
Estimates of Expected Values and Standard Errors of the Estimators of m_1, m_2 and π_1 for a Range of Sample Sizes.

Sample Size	Estimation Method	$\hat{E}(\hat{m}_1)$	$S\hat{.}E.(\hat{m}_1)$	$\hat{E}(\hat{m}_2)$	$S\hat{.}E.(\hat{m}_2)$	$\hat{E}(\hat{\pi}_1)$	$S\hat{.}E.(\hat{\pi}_1)$
500	Simple	1.96	.55	3.00	.097	.099	.025
	EM	1.97	.47	3.01	.085	.099	.024
1000	Simple	1.96	.38	2.99	.061	.101	.020
	EM	1.97	.39	3.01	.055	.099	.014
1500	Simple	1.95	.28	3.01	.057	.100	.016
	EM	1.96	.30	3.02	.041	.099	.009
2000	Simple	2.00	.25	3.00	.047	.100	.013
	EM	1.95	.24	3.02	.032	.098	.006
3000	Simple	1.99	.22	3.00	.041	.100	.010
	EM	1.99	.19	3.00	.030	.099	.006

each other, conditional on z_i. The M step of the algorithm then gives up-dated estimates as follows :

$$\pi_1^{(t+1)} = \sum^{n} z_i^* / n$$

$$m_1^{(t+1)} = (\sum^{n} x_i z_i^*) / \sum^{n} z_i^* \qquad (3.3)$$

$$m_2^{(t+1)} = (\sum^{n} x_i(1-z_i^*)) / \sum^{n} (1-z_i^*) .$$

The E and M steps generate a sequence of iterates converging to the ML estimates for π_1, m_1, m_2 . The observed information matrix for the parameter estimates may be found using the results of Louis (1982).

4. EMPIRICAL COMPARISON OF THE ESTIMATORS

In this Section we present some results from a comparison of the performance of the estimators using estimates computed from simulated data. The data on the non-sensitive variate were generated by mixing two Poisson distributions with means $m_1 = 2$ and $m_2 = 3$, using the mixing proportion $\pi_1 = 0.10$. This corresponds to 10% of a population having the sensitive variate S at level 1 (the stigmatizing level). The resulting data correspond to the x_i values in the notation of Section 2. The RR procedure was also simulated, assuming the use of the Simmons unrelated question design with parameters $p = 0.7$ and $\beta = 0.5$ (see Section 2), and the resulting data correspond to the r_i values of Section 2.

It was decided to compare the estimators for a range of sample sizes (n), and the following values for n were chosen: 500, 1000, 1500, 2000, 3000. For each sample size, 400 replicated samples of that size were generated, and the procedures of Section 2 and 3 were used to compute estimates of m_1, m_2 and π_1 for each sample. Using the 400 replicated sets of estimates it was then possible to estimate the expected values and the standard errors of the estimators, and the results are presented in Table 1.

The results of the computations indicate that the simple estimators compare quite favourably with the ML estimators, even for smaller sample sizes. The estimated standard errors of the ML estimators are somewhat smaller, and there is some indication that the ML estimators have noticeably smaller standard errors for quite large sample sizes, but further computations are needed. The poorer estimates and larger standard errors associated with the estimation of m_1 are of course due to the much smaller number of sample elements contributing information on m_1. The estimators of Section 2 for m_1 and m_2 are ratio estimators and one thus expects them to be

biased. However the bias appears to be negligible.

Overall one may conclude that the simple estimators of Section 2 perform quite well compared with ML estimators, and especially so for the large sample sizes that are likely to be used in RR surveys.

REFERENCES

ABUL-ELA, A.L., GREENBERG, B.G., HORVITZ, D.G. (1967). A Multi-Proportions Randomized Response Model. Journal of the American Statistical Association, 62, 990-1008.

BORUCH, R.F. and CECIL, J.S. (1979). Assuring the Confidentiality of Social Research Data, University of Pennsylvania Press.

BOURKE, P.D. (1982). Randomized Response Multivariate Designs for Categorical Data. Communications in Statistics: Theory and Methods, 11(25), 2889-2901.

BOURKE, P.D. and DALENIUS, T. (1973). Multi-proportions Randomized Response Using a Single Sample. Report No. 68 of the Errors in Surveys Research Project, Department of Statistics, University of Stockholm.

BOURKE, P.D. and MORAN, M.A. (1984). Application of the EM Algorithm to Randomized Response Data. Proceedings of the American Statistical Association: Section on Survey Research Methods, 788-793.

BOURKE, P.D. and MORAN, M.A. (1986). An Alternative EM Formulation for Randomized Response Data. Proceedings of the American Statistical Association: Section on Survey Research Methods, 444-447.

CHAUDHURI, A. and MUKERJEE, R. (1988). Randomized Response: Theory and Techniques. New York: Marcel Dekker.

DEFFAA, W. (1982). Anonymisierte Befragungen mit zufallsverschlusselten Antworten: Die Randomized-Response-Technik (RRT), Frankfurt am Main: Verlag Peter Lang.

DEMPSTER, A.P., LAIRD, N.M. and RUBIN, D.B. (1977) Maximum Likelihood Estimation from Incomplete Data via the EM Algorithm. J.R. Statist. Soc. B, 39, 1-38.

FOX, J.A. and TRACY, P.E. (1986). Randomized Response: A Method for Sensitive Surveys. Sage University Paper series on Quantitative Applications in the Social Sciences, 32 Beverly Hills: Sage Publications.

GOODAMBE, V.P. (1980). Estimation in Randomized Response Trials. International Statistical Review, 48, 29-32.

GREENBERG, B.G., ABULA-ELA, A.L., SIMMONS, W.R., HORVITZ, D.G. (1969). The Unrelated Question Randomized Response Model: Theoretical Framework. Journal of the American Statistical Association, 64, 520-539.

LOUIS, T.A. (1982). Finding the Observed Information Matrix when Using the EM Algorithm. J. R. Statist. Soc. B, 44, 226-233.

LOYNES, R.M. (1976). Asymptotically Optimal Randomized Response Procedures. Journal of the American Statistical Association, 71, 924-928.

TAMHANE, A.C. (1981). Randomized Response Techniques for Multiple Sensitive Attributes. Journal of the American Statistical Association, 76, 916-923.

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

WARNER, S.L. (1971). The Linear Randomized Response Model. Journal of the American Statistical Association, 66, 884-888.