APPLICATION OF THE EM ALGORITHM TO RANDOMIZED RESPONSE DATA

Patrick D. Bourke and Michael A. Moran
University College, Cork, Ireland

1. INTRODUCTION

The Randomized Response (RR) technique was introduced in Warner (1965) as a means of dealing with the bias and non-response associated with surveys of stigmatizing traits. The development of the RR technique since then has been quite extensive and recent reviews are provided in Deffaa (1982), Boruch and Cecil (1979) or Horvitz, Greenberg and Abernathy (1976). Many of the estimators for proportions presented in the early RR literature were claimed to be Maximum Likelihood (ML), although they were capable of producing estimates outside the range (0,1), as pointed out in Singh (1976). For the early RR designs, concerned primarily with the estimation of simple proportions, a minor adjustment was all that was required to the original estimators to make them ML, but for many subsequent RR designs, derivation of ML estimators and associated standard errors has been difficult, as is evident from Horvitz, Shah and Simmons (1967), Gould, Shah and Abernathy (1969), Greenberg et al. (1971), and Liu and Chow (1976).

In this paper it is shown that by viewing the observations from RR procedures as incomplete data, one can apply the EM algorithm described in Dempster, Laird and Rubin (1977) to find ML estimators. The standard error of these estimators can moreover be easily obtained using the results of Louis (1982). The estimation of a multinomial distribution for categorical-type data is considered in Section 2, and a general formulation, similar to that of Warner (1971), is presented for both related question designs and unrelated question designs. Expressions for ML estimators are developed which are very appealing in form due to their similarity to the case of direct questioning. These expressions are such that the natural restrictions on estimated proportions are automatically satisfied. In Section 3 a number of numerical examples based on well-established RR designs for estimating proportions are used to illustrate the procedure. These include the initial Warner design and the Simmons unrelated question design, as well as a design without explicit expressions for its ML estimators discussed originally by Horvitz, Shah and Simmons (1967).

In Section 4, the case of *quantitative* data arising from RR procedures, as discussed in Greenberg et al (1971) is considered. If one is willing to assume a parametric form for the distribution of the sensitive variate, then it is a straightforward matter to compute ML estimates of the parameters. On the other hand without the assumption of a parametric form, it is shown that one can use the EM procedure to estimate the distribution function of the sensitive variate (and indeed of the unrelated variate).

2. ESTIMATION OF MULTINOMIALS

Consider a multinomial variate with c categories, the probability of the k^{th} category being Π_k where $\Sigma\,\Pi_k = 1$. Let $\underset{\sim}{x}_i$, $1 \leq i \leq n$, represent the true response of the i^{th} individual in the sample where $\underset{\sim}{x}_i$ is a cx1 vector with all elements but one equal to zero, such that the k^{th} element being unity identifies the respondent as belonging to the k^{th} category.

A RR design is one which employs a randomizing device to produce a coded response $\underset{\sim}{y}_i$ rather than the true response $\underset{\sim}{x}_i$, where $\underset{\sim}{y}_i$ is a (dx1) vector($d \geq c$) of similar structure to $\underset{\sim}{x}_i$. The true response $\underset{\sim}{x}_i$ may be determined from the coded (i.e. observed) response $\underset{\sim}{y}_i$ by means of the transformation

$$\underset{\sim}{x}_i = \underset{\sim}{T}_j\, \underset{\sim}{y}_i$$

where $\underset{\sim}{T}_j$, $1 \leq j \leq t$ is one of a set of cxd transformation matrices, the actual transformation used being chosen according to the probability distribution $\{p_j;\ \Sigma\, p_j = 1\}$. The set of transformation matrices and their associated probabilities essentially define the particular RR design. Note that we are considering only RR designs where the response $\underset{\sim}{y}_i$ and the particular transformation used will uniquely determine the true category of the respondent. Such designs may be termed one-to-one.

In the case of unrelated question designs, (Greenberg et al. (1969)) some of the questions do not deal with the sensitive attribute and this is taken up in Section 2.3.

To begin with we consider RR designs in which all questions relate to the sensitive attribute. Such designs may be termed "related question" designs.

2.1 Related Question Designs

If $\underset{\sim}{z}_i$ is a tx1 vector, of similar structure to $\underset{\sim}{x}_i$ and $\underset{\sim}{y}_i$ above indicating which transformation matrix was used by the i^{th} individual then the log likelihood for the observation $(\underset{\sim}{z}_i, \underset{\sim}{y}_i)$, say, may be written as

$$\log L = \log f(\underset{\sim}{z}_i) + \log f(\underset{\sim}{y}_i | \underset{\sim}{z}_i) \qquad (2.1)$$

where $f(\underset{\sim}{z}_i)$ and $f(\underset{\sim}{y}_i|\underset{\sim}{z}_i)$ denote the marginal density of $\underset{\sim}{z}_i$ and the conditional density of $\underset{\sim}{y}_i$ given $\underset{\sim}{z}_i$ respectively.

This in turn may be expressed as

$$\log L = \sum_{j=1}^{t} z_{ij} \log p_j + \sum_{k=1}^{c} x_{ik} \log \Pi_k \qquad (2.2)$$

where

$$\underset{\sim}{x}_i = \left(\sum_{j=1}^{t} z_{ij}\, \underset{\sim}{T}_j\right)\underset{\sim}{y}_i$$

and $\{z_{ij},\ j=1,2,..,t\}$, $\{x_{ik},\ k=1,2,..,c\}$ denote the elements of vectors $\underset{\sim}{z}_i$ and $\underset{\sim}{x}_i$ respectively.

The ML estimator of $\underset{\sim}{\Pi}$ is simply

$$\hat{\underset{\sim}{\Pi}} = \underset{\sim}{x}_i \Big/ \left(\sum_{k=1}^{c} x_{ik}\right) = \underset{\sim}{x}_i\ ,$$

since $\sum_{k=1}^{c} x_{ik} = 1$, because of the structure of $\underset{\sim}{T}_j$ and $\underset{\sim}{y}_i$.

With observations $\{(\underset{\sim}{z}_i, \underset{\sim}{y}_i),\ i=1,2...n\}$ on a sample of n individuals, the ML estimator of $\underset{\sim}{\Pi}$ is

$$\hat{\Pi} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad (2.3)$$

The E step of the EM algorithm in this context consists of replacing the unobserved values z_{ij} by their expectations conditional on the observed values y_i, assuming the values of the parameters Π are known.

As each z_{ij} is a binary variate, the required expectation is simply a conditional probability and

$$E(z_{ij}|y_i,\Pi) = P[y_i|\Pi, z_{ij}=1]p_j/\lambda_i = z^*_{ij} \text{ (say)} \qquad (2.4)$$

where λ_i is the probability of response y_i as follows

$$\lambda_i = \sum_{j=1}^{t} P[y_i|\Pi,\ z_{ij} = 1]p_j \qquad (2.5)$$

The M step of the algorithm then gives

$$\hat{\Pi} = \frac{1}{n}\sum_{i=1}^{n} x^*_i \qquad (2.6)$$

where $x^*_i = \left(\sum_{j=1}^{t} z^*_{ij}\, T_j\right) y_i$ is the expected true response given the observed response y_i.

For the d distinct response-types, we form the d-element vector λ of distinct elements λ_i, and using (2.5) it is possible to write $\lambda = R\,\Pi$. For identifiability, we require that the rank of R should be $\geqslant (c-1)$, which imposes restrictions on the choice of values for the p_j's.

Equations (2.4) and (2.6) provide the basis of an iterative algorithm which will produce from an initial value $\Pi_{(o)}$ a sequence of estimates converging to the ML estimate of Π given observations on the $\{y_i\}$ alone.

It may be noted that the resulting estimator automatically satisfies the natural restrictions on the Π_k, viz. $\Pi_k \geqslant 0$, $\forall k$, and $\sum_{k=1}^{c} \Pi_k = 1$.

2.2 Standard Errors of the Estimates

The asymptotic variance-covariance matrix of $\hat{\Pi}$ is readily estimated using the results of Louis (1982). The derivative or efficient score for Π is a $(c-1)\times 1$ vector with elements

$$\frac{\delta \log L}{\delta \Pi_k} = \sum_{i=1}^{n} S_{ik}\ , \text{ with } S_{ik} = \frac{x_{ik}}{\Pi_k} - \frac{x_{ic}}{\Pi_c} \qquad 1 \leqslant k \leqslant (c-1)$$

Since each x_i is an indicator for a multinomial distribution the observed information matrix is simply

$$I = \sum_{i=1}^{n} S^*_i\, S^{*t}_i$$

where S^*_i is the vector of elements $\left(\frac{x^*_{ik}}{\hat{\Pi}_k} - \frac{x^*_{ic}}{\hat{\Pi}_c}\right)$.

2.3 Unrelated Question Designs for Estimating Proportions

In an unrelated question RR design the vector Π may be replaced by two subvectors $\Pi^{(1)}$ and $\Pi^{(2)}$ of lengths c and f corresponding to the categories of the sensitive and unrelated question respectively, where $\Sigma\, \Pi^{(1)}_k = \Sigma\, \Pi^{(2)}_k = 1$. The elements of $\Pi^{(2)}$ will be assumed known (see Section 3.3 for an example with unknown $\Pi^{(2)}$). The vector x_i representing the true response of the i^{th} individual, may be similarly replaced by two subvectors $x^{(1)}_i$ and $x^{(2)}_i$ each with structure similar to x_i in Section 2.

Corresponding to (2.2) the log-likelihood is again

$$\log L = \sum_{j=1}^{t} z_{ij}\, p_j + \sum_{k=1}^{c+f} x_{ik} \log \Pi_k\ , \qquad (2.7)$$

while (2.3) is replaced by

$$\hat{\Pi}^{(1)} = \frac{\sum_{i=1}^{n} x^{(1)}_i}{\sum_{i=1}^{n}\sum_{k=1}^{c} x_{ik}} \qquad (2.8)$$

where $\sum_{i=1}^{n}\sum_{k=1}^{c} x_{ik}$, the number of respondents to whom a sensitive question is posed, is assumed non-zero. The expressions for z^*_{ij} and x^*_i and the rest of the algorithm proceed as in Section 2.1.

3. EXAMPLES

3.1 The Initial Warner Design

Assume that it is desired to estimate the proportion of people who submitted an incorrect tax return last year. The questionnaire used has two questions

(1) Did you submit an incorrect tax return last year?

(2) Did you submit a correct tax return last year?

each of which has a yes or no answer. Each individual agrees to answer question 1 or 2 depending on the outcome of a randomizing device with probabilities 0.75 and 0.25 respectively for the two questions. The sample size is 1000 and 306 individuals answer yes. In this example we have (see Section 2.1) c = 2, d = 2, t = 2. The transformations in use are :

$$T_1 = \begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix},\ T_2 = \begin{pmatrix}0 & 1\\ 1 & 0\end{pmatrix}, \text{ with probabilities } p_1=0.75,\ p_2=0.25 \text{ respectively.}$$

The vector y_i can only assume two distinct values viz. $(1,0)^t$ and $(0,1)^t$ denoting yes and no respectively. Corresponding to (2.4) we have

$$E(z_{i1}|\text{yes}) = p_1\, \Pi_1/(p_1\, \Pi_1 + p_2\, \Pi_2) = z^*_1$$

$$E(z_{i1}|\text{no}\ ) = p_1\, \Pi_2/(p_1\, \Pi_2 + p_2\, \Pi_1) = z^*_2$$

Starting with $\hat{\Pi}_1 = 0.15$, the first iterations proceed as follows

Iteration	$\hat{\Pi}_1$	z^*_1	z^*_2
1	0.1500	-	-
2	0.1445	0.3462	0.9444
3	0.1399	0.3363	0.9467
4	0.1360	0.3279	0.9486
.	.	.	.
.	.	.	.
12	0.1190	0.2902	0.9565
∞	0.1120	0.2745	0.9597

The estimated variance of $\hat{\Pi}_1$ is 0.000849. Both the estimate of Π_1 and its estimated variance are in agreement with the values obtained by applying the appropriate expressions given in Warner (1965).

3.2 The Simmons Unrelated Question Design

In this design, the second question of the Warner design is replaced by an unrelated question, so that the two questions on the questionnaire might read :

(1) Have you ever been convicted of drunken driving?

(2) Were you born in the first six months of the calendar year?

A randomizing device directs the respondent to answer either question 1 (with probability 0.5 say) or question 2. Suppose the sample size was 250 and 101 respondents answered yes, and let us assume that the proportion of the population born in the first six months of the calendar year is 0.5.

Thus $$\underset{\sim}{\Pi}^{(1)} = \begin{pmatrix}\Pi_1\\ \Pi_2\end{pmatrix}, \; \Pi_1 + \Pi_2 = 1$$

$$\underset{\sim}{\Pi}^{(2)} = \begin{pmatrix}\Pi_3\\ \Pi_4\end{pmatrix} = \begin{pmatrix}0.5\\ 0.5\end{pmatrix}$$

The observed response vector $\underset{\sim}{y}_i$ takes only two distinct values $(1,0)^t$ and $(0,1)^t$, denoting yes and no respectively while the true response vector $\underset{\sim}{x}_i$ takes four values $(1,0,0,0)^t$, $(0,1,0,0)^t$, etc. There are only two possible transformations, viz.

$$\underset{\sim}{T}_1 = \begin{bmatrix}1&0\\0&1\\0&0\\0&0\end{bmatrix} \qquad \underset{\sim}{T}_2 = \begin{bmatrix}0&0\\0&0\\1&0\\0&1\end{bmatrix}$$

with selection probabilities p_1 and p_2 respectively.

The conditional expectations of the z_{ij} are

$$E(z_{i1}|\text{yes}) = \frac{p_1\Pi_1}{p_1\Pi_1 + p_2\Pi_3} = z_1^*$$

$$E(z_{i1}|\text{no}) = \frac{p_1\Pi_2}{p_1\Pi_2 + p_2\Pi_4} = z_2^*$$

Using the initial value $\hat{\Pi}_1 = 0.2$, the first iterations are as follows :

Iteration	$\hat{\Pi}_1$	z_1^*	z_2^*
1	.2000		
2	.2394	.2857	.6154
3	.2667	.3238	.6034
4	.2840	.3479	.5946
.	.	.	.
.	.	.	.
12	.3079	.3810	.5806
∞	.3080	.3812	.5805

The estimate of the asymptotic variance is 0.00385. Again, the estimate of Π_1 and its estimated variance are in agreement with the values obtained by applying the appropriate expressions given in Greenberg et al (1969).

3.3 The Simmons Unrelated Question Design with Two Trials

In the previous examples, closed form expressions for the ML estimator are available. In this Section we look at an example, Simmons Unrelated Question Design with two trials per respondent, where this is not the case. In contrast to the previous example 3.2, the proportion of the population with the non-sensitive attribute is not assumed known, although the sensitive and unrelated characteristics are assumed independently distributed in the population. (The independence assumption ensures that each Π_k in (2.7) can be expressed as a product of two terms so that the form of the log-likelihood in (2.7) is preserved).

Our objective is to estimate $\underset{\sim}{\Pi}^{(1)}$ and $\underset{\sim}{\Pi}^{(2)}$ where

$$\underset{\sim}{\Pi}^{(1)} = \begin{pmatrix}\Pi_1\\ \Pi_2\end{pmatrix}, \; (\Pi_1+\Pi_2=1), \text{ for the sensitive question.}$$

and

$$\underset{\sim}{\Pi}^{(2)} = \begin{pmatrix}\Pi_3\\ \Pi_4\end{pmatrix}, \; (\Pi_3+\Pi_4=1), \text{ for the unrelated question.}$$

The observed response vector $\underset{\sim}{y}_i$ takes four distinct values

1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	1
YY	YN	NY	NN

where YN, for example, indicates a response yes to the first question and no to the second.

There are four possible transformations corresponding to the possible orders SS, SU, US, UU in which the sensitive (S) and unrelated (U) questions may be asked :

$$\begin{array}{cccc} \underset{\sim}{T}_1 & \underset{\sim}{T}_2 & \underset{\sim}{T}_3 & \underset{\sim}{T}_4 \\ \begin{bmatrix}1&0&0&0\\0&0&0&1\\0&0&0&0\\0&0&0&0\end{bmatrix} & \begin{bmatrix}1&1&0&0\\0&0&1&1\\1&0&1&0\\0&1&0&1\end{bmatrix} & \begin{bmatrix}1&0&1&0\\0&1&0&1\\1&1&0&0\\0&0&1&1\end{bmatrix} & \begin{bmatrix}0&0&0&0\\0&0&0&0\\1&0&0&0\\0&0&0&1\end{bmatrix} \\ \text{SS} & \text{SU} & \text{US} & \text{UU} \end{array}$$

The selection probabilities for these four transformations are p_1^2, p_1p_2, p_2p_1, p_2^2 respectively, where p_1 and p_2 are the selection probabilities of the sensitive and unrelated questions in a single trial. The conditional expectations of the z_{ij} are :

	Possible Responses (χ_i)			
	YY	YN	NY	NN
$E(z_{i1}\|\chi_i)$	$\frac{p_1^2\Pi_1}{\lambda_{11}}$	0	0	$\frac{p_1^2\Pi_2}{\lambda_{22}}$
$E(z_{i2}\|\chi_i)$	$\frac{p_1p_2\Pi_1\Pi_3}{\lambda_{11}}$	$\frac{p_1p_2\Pi_1\Pi_4}{\lambda_{12}}$	$\frac{p_1p_2\Pi_2\Pi_3}{\lambda_{21}}$	$\frac{p_1p_2\Pi_2\Pi_4}{\lambda_{22}}$
$E(z_{i3}\|\chi_i)$	$\frac{p_1p_2\Pi_1\Pi_3}{\lambda_{11}}$	$\frac{p_1p_2\Pi_2\Pi_3}{\lambda_{12}}$	$\frac{p_1p_2\Pi_1\Pi_4}{\lambda_{21}}$	$\frac{p_1p_2\Pi_2\Pi_4}{\lambda_{22}}$
$E(z_{i4}\|\chi_i)$	$\frac{p_2^2\Pi_3}{\lambda_{11}}$	0	0	$\frac{p_2^2\Pi_4}{\lambda_{22}}$

where the λ_{ij} denote the probability of responses YY, YN, etc., and are as follows :

$$\lambda_{11} = p_1^2\, \Pi_1 + 2\, p_1p_2\, \Pi_1\, \Pi_3 + p_2^2\, \Pi_3$$

$$\lambda_{12} = p_1p_2\, \Pi_1\, \Pi_4 + p_2\, p_1\, \Pi_2\, \Pi_3$$

$$\lambda_{21} = p_1p_2\, \Pi_2\, \Pi_3 + p_2\, p_1\, \Pi_1\, \Pi_4$$

$$\lambda_{22} = p_1^2\, \Pi_2 + 2\, p_1p_2\, \Pi_2\, \Pi_4 + p_2^2\, \Pi_4$$

In Horvitz, Shah and Simmons (1967), the unrelated question design with two trials was used in a survey to estimate the proportion (Π_1) of births where the mother was unmarried. Two samples, each with different design parameters, were used and moment estimates for Π_1 and Π_3 were presented. Part of the data is as follows :

	Frequency of Responses			
	YY	YN	NY	NN
Sample 1	137	271	253	566
Sample 2	512	291	215	322

Subsequently, in Gould, Shah and Abernathy (1969, Table 4, Model 1) ML estimates for Π_1 and Π_3 were reported for this data, but not standard errors because of computational difficulties. The application of the EM procedure to the above data produces the following estimates :

	$\hat{\Pi}_1$	$\hat{\Pi}_3$	S.E. $(\hat{\Pi}_1)$	S.E. $(\hat{\Pi}_3)$
EM Procedure	0.02829	0.8616	0.0095	0.0112
Gould, Shah & Abernathy (1969)	0.02824	0.8616	-	-

It should be noted that for this design, two separate samples are not necessary for the estimation of Π_1 and Π_3.

In Liu and Chow (1976), the Fisher scoring method is used to derive ML estimates for the data arising from a design in which the Warner related-question procedure is applied three times to each respondent. The EM procedure has been applied to the Liu-Chow data, with complete agreement on the estimates.

4. QUANTITATIVE DATA DESIGNS

In Greenberg et al. (1971) an RR design for a quantitative sensitive variate was presented. This design was used in an abortion study in North Carolina and each respondent was asked one or other of the following two questions :

Question A : How many abortions have you had during your lifetime?

Question Y : If a woman has to work full-time to make a living, how many children do you think she should have?

Since the distribution for the non-sensitive variate was unknown, two samples were required, and let $\lambda_j(x)$ be the probability function (p.f.) of the response x from a randomly selected respondent in sample j, while f(x) and g(x) are the p.f.'s of the sensitive and non-sensitive variates. Then,

$$\text{Sample 1 : } \lambda_1(x) = p_1 f(x) + (1-p_1)g(x), \tag{4.1}$$

$$\text{Sample 2 : } \lambda_2(x) = p_2 f(x) + (1-p_2)g(x), \tag{4.2}$$

where p_j is the selection probability of the sensitive question in sample j.

The response for each question is discrete (with relatively few distinct values) and the estimation of the two distributions is merely a special case of the multinomial estimation considered in Section 2. Means and other parameters of the estimated distributions can then be computed. The following data were reported in Greenberg et al. (1971).

Response	0	1	2	3	4	5
Frequency (Sample 1)	304	14	56	10	7	2
Frequency (Sample 2)	114	10	30	6	1	1

After applying the EM procedure of Section 2, the following estimates of f and g were obtained :

x Value	$\hat{f}(x)$	$\hat{g}(x)$
0	.83 (.046)	.65 (.065)
1	.016 (.022)	.081 (.034)
2	.11 (.038)	.22 (.055)
3	.017 (.018)	.046 (.027)
4	.025 (.013)	0 (.013)
5	.004 (.008)	.007 (.011)

(Standard errors in parentheses)

Estimates for the average values (μ_A, μ_Y) of the sensitive and non-sensitive variates are as follows :

	Greenberg et al. (1971)	Present Approach
$\hat{\mu}_A$	.415 (.107)	.408 (.107)
$\hat{\mu}_Y$	.678 (.145)	.686 (.145)

(Standard errors in parentheses)

It may not be necessary to use the EM procedure to compute ML estimates for f(x) and g(x). One should first look at the simple estimators in terms of linear combinations of $\hat{\lambda}_1(x)$ and $\hat{\lambda}_2(x)$, the proportions responding 'x' in samples 1 and 2 :

$$\hat{f}_s(x) = \frac{(1-p_2)\hat{\lambda}_1(x) - (1-p_1)\hat{\lambda}_2(x)}{p_1(1-p_2) - p_2(1-p_1)} \qquad (4.3)$$

with a similar expression for $\hat{g}_s(x)$. If these estimators produce values all lying in [0,1], then one has found the ML estimates for f and g. In the case of the above data not all values of $\hat{f}_s(x)$ and $\hat{g}_s(x)$ lie in [0,1], so that some numerical procedure, such as EM, was necessary. This accounts for the difference between the present estimates for μ_A, μ_Y and those presented in Greenberg et al. (1971) which were based on linear combinations of the $\hat{\lambda}_j(x)$.

In the case of a continuous variate, or a discrete variate with a large number of distinct values, one can split the range of the variate into a number of disjoint categories, so that one can again apply the multinomial estimation of Section 2. Alternatively, one can derive ML estimates of the distribution functions. Let $\Lambda_j(x)$ denote the distribution function of the response from a respondent in sample j, while F(x) and G(x) are the distribution functions for the sensitive and non-sensitive variates. Then,

$$\text{Sample 1: } \Lambda_1(x) = p_1F(x) + (1-p_1)G(x), \qquad (4.4)$$

$$\text{Sample 2: } \Lambda_2(x) = p_2F(x) + (1-p_2)G(x), \qquad (4.5)$$

By analogy with (4.3) one can suggest simple estimators of F(x) and G(x) such as

$$\hat{F}(x) = \frac{(1-p_2)\hat{\Lambda}_1(x) - (1-p_1)\hat{\Lambda}_2(x)}{p_1(1-p_2) - p_2(1-p_1)}$$

where $\hat{\Lambda}_j(x)$ is the proportion of responses in sample j which are less than or equal to x. If $\hat{F}(x)$ and $\hat{G}(x)$ are monotone and in [0,1] for all x (which is unlikely) then $\hat{F}(x)$ is the ML estimator for F(x).

Notice that in estimating F(x) for a *given* x one is in fact estimating a simple proportion using what is in effect the Simmons unrelated question design. Thus one can build the ML estimate for F(x), and for G(x), by taking each x value in turn and applying the EM procedure to estimate the unknown proportions F(x) and G(x). The latter can then be used to estimate means, variances, percentiles and other summary statistics, as well as to motivate parametric forms for F(x) and G(x).

An alternative approach in both the discrete and continuous cases would be to assume parametric models (e.g. Poisson) for the sensitive and non-sensitive variates. With this approach, only one sample is needed, and the problem of estimation reduces to a special case of the classical mixture problem in which the mixing proportion is known. A straightforward application of the EM procedure will produce the ML estimates for the model parameters, and their standard errors.

REFERENCES

[1] BORUCH, R.F. & CECIL, J.S. (1979), *Assuring the Confidentiality of Social Research Data.* University of Pennsylvania Press, 127-173.

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

[3] 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.

[4] GOULD, A.L., SHAH, B.V. and ABERNATHY, J.R. (1969), "Unrelated Question Randomized Response Techniques with Two Trials per Respondent", *Proceedings of the American Statistical Association: Social Statistics Section*, pp. 351-59.

[5] GREENBERG, B.G., ABUL-ELA, A.A., SIMMONS, W.R. and HORVITZ, D.G. (1969), "The Unrelated Question Randomized Response Model: Theoretical Framework", *Journal of the American Statistical Association* 64: 520-39.

[6] GREENBERG, B.G., KUEBLER, R.R., ABERNATHY, J.R. and HORVITZ, D.G. (1971), "Application of the Randomized Response Technique in Obtaining Quantitative Data", *Journal of the American Statistical Association* 66: 243-50.

[7] HORVITZ, D.G., GREENBERG, B.G. and ABERNATHY, J.R. (1976), "Randomized response: A Data Gathering-Device for Sensitive Questions", *International Statistical Review* 44: 181-96.

[8] HORVITZ, D.G., SHAH, B.V. and SIMMONS, W.R. (1967), "The Unrelated Question Randomized Response Model", *Proceedings of the American Statistical Association: Social Statistics Section*, pp. 65-72.

[9] LIU, P.T. and CHOW, L.P. (1976), "The Efficiency of the Multiple Trial Randomized Response Technique", *Biometrics*, 32: 607-18.

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

[11] SEN, P.K. (1974), "On Unbiased Estimation for Randomized Response Models", *Journal of the American Statistical Association*, 69, 997-1001.

[12] SINGH, J. (1976), "A Note on the Randomized Response Technique", *Proceedings of the American Statistical Association: Social Statistics Section:* 772.

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

[14] WARNER, S.L. (1971), "The Linear Randomized Response Model", *Journal of the American Statistical Association* 66: 884-88.