# AN ALTERNATIVE EM FORMULATION FOR RANDOMIZED RESPONSE DATA

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

Abstract: To facilitate the computation of maximum likelihood (ML) estimates for data arising from Randomized Response (RR) investigations, one can view the data as mixture data, and apply the EM algorithm. The EM formulation presented differs from the earlier formulation, in that now the proportions to be estimated are regarded as the mixing proportions, leading to a simple implementation of the EM algorithm. A general formulation is presented for both related-question and unrelated-question RR designs, and illustrated with applications.

## I. INTRODUCTION

The Randomized Response (RR) technique was introduced in Warner (1965) to deal with non-response and consequent bias associated with surveys of stigmatizing traits. The development of the RR technique since then has been quite extensive, see reviews by Deffaa (1982), Boruch and Cecil (1979) and 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 could produce estimates outside the range (0, 1), as noted by Singh (1976). For the early RR designs, a minor adjustment to the original estimators was sufficient to make them ML, but for many later designs, computation of ML estimates and their standard errors is difficult, as is evident from Gould, Shah and Abernathy (1969), Greenberg et al. (1971), Liu and Chow (1976), and Bourke (1982).

By viewing observations from RR procedures as mixture data, one can apply the EM algorithm described in Dempster, Laird and Rubin (1977) to find ML estimates. The standard errors of these estimates can moreover be easily obtained using the results of Louis (1982). There are two ways of viewing the data as a mixture of distributions. One can view the randomizing device as the mixing mechanism so that the mixing proportions are the chosen parameters of the randomizing device, as in Bourke and Moran (1984). Alternatively the proportions to be estimated can be taken as the mixing proportions. The latter approach is adopted in this paper.

A general formulation for estimation of a multinomial distribution is given in Section 2. In Section 3, the unrelated question design with two trials for each respondent and two samples, first described in Horvitz, Shah and Simmons (1967), is used to illustrate two applications of the methods developed in Section 2. The second application also provides an example of multivariate estimation and a test of independence between the variates is carried out.

## 2. ESTIMATION OF MULTINOMIALS FROM RANDOMIZED RESPONSE DATA

Consider a sensitive variate S having c categories of which at most (c-1) are stigmatizing. Let the population proportion for category k be $\Pi_k$. Our objective is to estimate the $\Pi_k$'s. We consider also a non-sensitive variate U having f categories (usually f = c) and the population proportions in these categories may or may not be known. For brevity we refer to the true state of a respondent with respect to variate S as the S-level and similarly for variate U.

We consider two broad classes of RR designs: related question designs and unrelated question designs. In a related question design, such as the original Warner (1965) design, the possible questions or statements that may be put to a respondent relate only to the variate S, while in an unrelated question design, see Horvitz et al. (1967), some of the questions relate to the variate U. If $\lambda_r$ denotes the probability of the $r^{th}$ response, then

$$\lambda_r = \sum_{k=1}^{c} p_{rk} \Pi_k, \qquad (r = 1, 2, \ldots, d) \qquad (2.1)$$

where $p_{rk}$ = P [response r|S-level = k].

Writing (2.1) in matrix notation we get

$$\underset{\sim}{\lambda} = \underset{\sim}{P} \underset{\sim}{\Pi} \qquad (2.2)$$

where $\underset{\sim}{P}$ the matrix of conditional probabilities $\{p_{rk}\}$, may be termed the design matrix of the RR procedure. If $\underset{\sim}{P}$ is square, the moment estimator for $\underset{\sim}{\Pi}$ may give the ML estimate, but if $\underset{\sim}{P}$ has more rows than columns a numerical procedure is necessary.

Two cases are considered:

Case 1: <u>All the parameters of the design matrix are known</u>

This class of RR designs is considered in Loynes (1976) and corresponds to related question designs, or unrelated question designs with a known distribution of the unrelated question. For example, in the unrelated question design of Horvitz et al. (1967) for estimating a single proportion $\Pi_1$, (2.2) becomes

$$\begin{bmatrix} \lambda_1 \\ \lambda_2 \end{bmatrix} = \begin{bmatrix} p+(1-p)\mu & (1-p)\mu \\ (1-p)(1-\mu) & p+(1-p)(1-\mu) \end{bmatrix} \begin{bmatrix} \Pi_1 \\ \Pi_2 \end{bmatrix} \qquad (2.3)$$

where $\lambda_1$ is the probability of the response 'YES', $\lambda_2 = (1-\lambda_1)$, $\Pi_2 = (1-\Pi_1)$, p is the probability that the sensitive question is put to a respondent, and $\mu$ is the known proportion with the unrelated attribute (for an example of a sensitive and an unrelated attribute, see Section 3).

Case 2: <u>Some of the parameters of the design matrix are unknown</u>

The most familiar example of this is the Horvitz et al. (1967) unrelated question design with an unknown distribution of the unrelated question, so that in (2.3) above, $\mu$ is unknown. Two samples with different values of p are then needed to estimate $\Pi_1$ and $\mu$. Usually the moment estimators for $\Pi_1, \mu$ give ML estimates, but not always, as shown by this data-set:

|  | p | Sample-Size | Number of 'YES' replies |
|---|---|---|---|
| Sample 1 | 0.3 | 1000 | 15 |
| Sample 2 | 0.7 | 1000 | 75 |

Here the moment estimates are 0.12, -0.03 whereas the ML estimates are 0.09, 0.

Analysis of Case 1

The response made by the $i^{th}$ respondent may be represented by a vector $(\underset{\sim}{y}_i)$ of (d-1) zeros and one unity such that the $r^{th}$ element being unity implies that the $r^{th}$ response was given. The S-level of the $i^{th}$ respondent is represented by a similar vector $\underset{\sim}{z}_i$. In EM terminology $\underset{\sim}{z}_i$ corresponds to the missing data on the S level of the $i^{th}$ respondent. If the $\underset{\sim}{z}_i$ were observed, the log-likelihood for the $\Pi_k$'s based on the observations $(\underset{\sim}{z}_i, \underset{\sim}{y}_i)$, i = 1, 2, ..., n, would be

$$\log L(\underset{\sim}{\Pi}) = \sum_i \Big[ \sum_{r,k} y_{ir}\, z_{ik} \log p_{rk} + z_{ik} \log \Pi_k \Big] \qquad (2.4)$$

where $y_{ir}$ and $z_{ik}$ are elements of $\underset{\sim}{y}_i$ and $\underset{\sim}{z}_i$ respectively. The E step of the algorithm consists of estimating the complete data sufficient statistics $\Sigma \underset{\sim}{z}_i$ by replacing the unobserved $\underset{\sim}{z}_i$ by their expectations $\underset{\sim}{z}^*_i$ conditional on the observed $\underset{\sim}{y}_i$ and the current parameter estimates. Thus

$$\begin{aligned} z^*_{ik} &= E(z_{ik} \mid y_{ir} = 1, \underset{\sim}{\Pi}) \\ &= p_{rk}\, \Pi_k / \lambda_r, \ (k = 1, 2, \ldots, c) \qquad (2.5) \end{aligned}$$

The M step of the algorithm then gives immediately $\hat{\underset{\sim}{\Pi}} = \Sigma\, \underset{\sim}{z}^*_i / n$. The E and M steps generate a sequence of estimates converging to the ML estimate $\hat{\underset{\sim}{\Pi}}$. It may be noted that this $\hat{\underset{\sim}{\Pi}}$ automatically satisfies the restrictions on the $\Pi_k$, viz. $\Pi_k \geq 0$ and $\Sigma\, \Pi_k = 1$.

The asymptotic variance-covariance matrix of $\hat{\underset{\sim}{\Pi}}$ is readily estimated using the results of Orchard and Woodbury (1972) or Louis (1982), and is

$$\underset{\sim}{G} = \sum_{i=1}^{n} \underset{\sim}{S}^*_i\, \underset{\sim}{S}^{*t}_i \qquad (2.6)$$

where $\underset{\sim}{S}^*_i$ is the vector of elements $(z^*_{ik} / \hat{\Pi}_k)$ $-(z^*_{ic} / \hat{\Pi}_c)$ from the final iteration of the algorithm. An example of Case 1 is given in Section 3.

Analysis of Case 2

In an unrelated question design suppose that the distribution $(\underset{\sim}{\mu})$ of the U-variate is unknown. In (2.4) some of the $p_{rk}$ will now depend on the unknown $\underset{\sim}{\mu}$, and this can make the M step awkward. To avoid this, we suggest the following approach. Corresponding to (2.1), we have two possible expressions for $\lambda_r$:

$$\lambda_r = \sum_{k=1}^{c} p_{rk}\, \Pi_k \qquad (2.7)$$

$$\lambda_r = \sum_{k=1}^{f} q_{rk}\, \mu_k \qquad (2.8)$$

where $q_{rk}$ = P[response r|U-level = k].

If $\underset{\sim}{\mu}$ were known, the procedure of Case 1 could be applied. Similarly if $\underset{\sim}{\Pi}$ were known, we could reverse the role of $\underset{\sim}{\Pi}$ and $\underset{\sim}{\mu}$ and again apply the EM procedure of Case 1 to estimate $\underset{\sim}{\mu}$. This symmetry can be exploited to maximize the likelihood for $\underset{\sim}{\Pi}$ and $\underset{\sim}{\mu}$ simultaneously. Depending on whether the missing data is taken to be the levels of the S-variate or the U-variate, we have two possible expressions for the log-likelihood:

$$\log L(\underset{\sim}{\Pi}, \underset{\sim}{\mu}) = \sum_i \Big\{ \sum_{r,k} \big[ y_{ir}\, z_{ik} \log p_{rk} + z_{ik} \log \Pi_k \big] \Big\} \qquad (2.9)$$

$$\log L(\underset{\sim}{\Pi}, \underset{\sim}{\mu}) = \sum_i \Big\{ \sum_{r,k} \big[ y_{ir}\, w_{ik} \log q_{rk} + w_{ik} \log \mu_k \big] \Big\} \qquad (2.10)$$

where $z_{ik}$ has the same meaning as $z_{ik}$ of (2.4), and $w_{ik}$ is an element of the corresponding $\underset{\sim}{w}_i$ vector denoting the U-level of the $i^{th}$ respondent. As in (2.5), we see that the E step is

$$z^*_{ik} = p_{rk}\, \Pi_k / \lambda_r$$

$$w^*_{ik} = q_{rk}\, \mu_k / \lambda_r$$

and the M step is $\hat{\underset{\sim}{\Pi}} = \Sigma\, \underset{\sim}{z}^*_i / n$ and $\hat{\underset{\sim}{\mu}} = \Sigma\, \underset{\sim}{w}^*_i / n$.

The observed information matrix is as (2.6) where $\underset{\sim}{S}^*_i$ is the vector of elements $S^*_{ik}$ and

$$S^*_{ik} = (z^*_{ik} / \hat{\Pi}_k) - (z^*_{ic} / \hat{\Pi}_c), \quad 1 \leq k \leq (c-1)$$

$$S^*_{i,k+(c-1)} = w^*_{ik}/\hat{\mu}_k - w^*_{if}/\hat{\mu}_f \quad 1 \le k \le (f-1)$$

An example of Case 2 is given in Section 3.

Although the presentation here has been in terms of a multinomial variate S, the procedure is also applicable to multivariate cases. In Tamhane (1981) and Bourke (1982) RR designs for multivariate estimation are presented, each design being represented as in (2.2). An example of Case 1 multivariate estimation is given in Section 3.

## 3. APPLICATIONS

The 2-Trial Unrelated Question Design with Two Samples.

An application of this design to the estimation of a proportion of illegitimate births is described in Horvitz et al. (1967). The design involved two samples. The statements used were:

S "In the past 12 months there was a baby born in this household to an unmarried woman who was living here at the time".

U "I was born in North Carolina".

The probability ($p_i$) that the randomizing device would select statement S was kept constant for each trial in sample i, i = 1, 2. The relevant data are reproduced in Table 3.1.

TABLE 3.1

| | Frequency of Responses | | | | |
|---|---|---|---|---|---|
| | YY | YN | NY | NN | $p_i$ |
| Sample 1 | 137 | 271 | 253 | 566 | 0.7 |
| Sample 2 | 512 | 291 | 215 | 322 | 0.3 |

We present two applications of the EM procedure to this data-set; the first to illustrate the case where some parameters in the design matrix are unknown, the second to illustrate multivariate estimation where all parameters in the design matrix are known. The first application requires that the S and U variates be independent, an assumption made in previous analyses of this data-set. The second analysis not only avoids this assumption, but shows that it is inconsistent with the data and has a substantial effect on the estimates.

### 3.1 Application 1

Denote the two possible S-levels by $\underset{\sim}{z}$ vectors (1,0) and (0,1), and similarly for the U-levels. Denote the four possible responses YY, YN, NY, NN (YY = YES, YES) by a vector $\underset{\sim}{y}$ where $\underset{\sim}{y} = (1,0,0,0)$ denotes YY. The matrix of conditional probabilities of (2.7) for the first sample is:

$$\begin{bmatrix} p_1^2+2p_1q_1\mu+q_1^2\mu & q_1^2\mu \\ p_1q_1(1-\mu) & p_1q_1\mu \\ p_1q_1(1-\mu) & p_1q_1\mu \\ q_1^2(1-\mu) & p_1^2+2p_1q_1(1-\mu)+q_1^2(1-\mu) \end{bmatrix} \tag{3.1}$$

where $q_1 = 1 - p_1$, and $\mu$ is the unknown proportion of the population having attribute U. The corresponding matrix for the second sample is similar. The matrix of conditional probabilities of (2.8) is similar to (3.1) with $p_i$, $q_i$ interchanged and $\Pi$ replacing $\mu$, where $\Pi$ is the proportion having attribute S.

In Horvitz et al. (1967) moment estimates for $\Pi$ and $\mu$ were presented. Subsequently in Gould et al. (1969, Table 4, Model 1) ML estimates for $\Pi$ and $\mu$ were reported, but not standard errors. The application of the EM algorithm yields the results shown in Table 3.2.

TABLE 3.2

| | $\hat{\Pi}$ | $\hat{\mu}$ | S.E.($\hat{\Pi}$) | S.E.($\hat{\mu}$) |
|---|---|---|---|---|
| EM Algorithm | 0.02829 | 0.8616 | 0.0095 | 0.0112 |
| Gould et al. (1969) | 0.02824 | 0.8616 | -- | ---. |

### 3.2 Application 2

Here, we illustrate Case 1 for multivariate estimation by computing ML estimates for $\Pi_{ij}$ where i denotes the S-level and j denotes the U-level. We note that $\Pi = \Pi_{11} + \Pi_{12}$ and $\mu = \Pi_{11} + \Pi_{21}$.

The $\underset{\sim}{z}$ vector, which now indexes the four possible cross-categories of S and U, has the same form as the $\underset{\sim}{y}$ vector in the first application. For the first sample the design matrix in (2.2) is

$$\underset{\sim}{P} = \begin{bmatrix} 1 & p_1^2 & q_1^2 & 0 \\ 0 & p_1q_1 & p_1q_1 & 0 \\ 0 & p_1q_1 & p_1q_1 & 0 \\ 0 & q_1^2 & p_1^2 & 1 \end{bmatrix}$$

with a similar form for the second sample.

The application of the EM algorithm yields the following ML estimates for the $\Pi_{ij}$, with standard errors in parentheses:

| $\hat{\Pi}_{11}$ | $\hat{\Pi}_{12}$ | $\hat{\Pi}_{21}$ | $\hat{\Pi}_{22}$ |
|---|---|---|---|
| 0.000<br>(0.012) | 0.124<br>(0.024) | 0.779<br>(0.019) | 0.097<br>(0.014) |

The assumption of independence can easily be incorporated in the EM procedure as follows: After each M step compute the estimated marginal proportions $\hat{\Pi}$ and $\hat{\mu}$, then obtain $\hat{\Pi}_{11} = \hat{\Pi}\,\hat{\mu}$, and return to the E step. Not surprisingly, the estimates for $\Pi$ and $\mu$ are as given in Table 3.2.

The estimate of $\Pi$, without assuming independence of S and U, is 0.124. Calculating the corresponding likelihood ratio statistic Λ we find -2logΛ = 37.55 for 1 DF which clearly contradicts the assumption of independence. On reflection, one might expect a lower proportion of births to unmarried women for those who were born in North Carolina (and continue to live there) than for those who were born elsewhere (but now live in North Carolina). One explanation might be that unmarried women who become pregnant leave home, often moving out-of-state; another is that unmarried women living away from home may be less subject to social norms, and thus more likely to become pregnant, or to admit to it.

## REFERENCES

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

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

DEFFAA, W. (1982), *Anonymisierte Befragungen mit zufallsverschlüsselten Antworten: Die Randomized-Response-Technik (RRT),* Frankfort 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.

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.

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.

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.

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.

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

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.

ORCHARD, T., and WOODBURY, M. A. (1972), "A Missing Information Principle: Theory and Applications," *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability,* I, Berkeley: University of California Press, 697-715.

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

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.