

# Randomized Response, Statistical Disclosure Control and Misclassification: a Review

Ardo van den Hout and Peter G.M. van der Heijden

Utrecht University, Faculty of Social Sciences, Dept. of Methodology and Statistics, Heidelberglaan 2, 3584 CS Utrecht, The Netherlands

## Summary

This paper discusses analysis of categorical data which have been misclassified and where misclassification probabilities are known. Fields where this kind of misclassification occurs are randomized response, statistical disclosure control, and classification with known sensitivity and specificity. Estimates of true frequencies are given, and adjustments to the odds ratio are discussed. Moment estimates and maximum likelihood estimates are compared and it is proved that they are the same in the interior of the parameter space. Since moment estimators are regularly outside the parameter space, special attention is paid to the possibility of boundary solutions. An example is given.

*Key words:* Contingency table; EM algorithm; Misclassification; Odds ratio; Randomized response; Sensitivity; Specificity; Statistical disclosure control.

## 1 Introduction

When scores on categorical variables are observed, there is a possibility of misclassification. By a categorical variable we mean a stochastic variable of which the range consists of a limited number of discrete values called the categories. Misclassification occurs when the observed category is  $i$  while the true category is  $j$ ,  $i \neq j$ . This paper discusses analysis of categorical data subject to misclassification with known misclassification probabilities.

There are four fields in statistics where the misclassification probabilities are known. The first is randomized response (RR). RR is an interview technique which can be used when sensitive questions have to be asked. Warner (1965) introduced this technique and we use a simple form of the method as an introductory example. Let the sensitive question be 'Have you ever used illegal drugs?' The interviewer asks the respondent to roll a dice and to keep the outcome hidden. If the outcome is 1, 2, 3 or 4 the respondent is asked to answer question  $Q$ , if the outcome is 5 or 6 he is asked to answer  $Q^c$ , where

$$\begin{aligned} Q &= \text{'Have you ever used illegal drugs?'} \\ Q^c &= \text{'Have you never used illegal drugs?'} \end{aligned}$$

The interviewer does not know which question is answered and observes only 'yes' or 'no'. The respondent answers  $Q$  with probability  $p = 2/3$  and answers  $Q^c$  with probability  $1 - p$ . Let  $\pi$  be the unknown probability of observing a yes-response to  $Q$ . The probability of a yes-response is  $\lambda = p\pi + (1 - p)(1 - \pi)$ . So, with the observed proportion as an estimate  $\hat{\lambda}$  of  $\lambda$ , we can estimate  $\pi$  by

$$\hat{\pi} = \frac{\hat{\lambda} - (1 - p)}{2p - 1}. \quad (1)$$

The main idea behind RR is that perturbation by the misclassification design (in this case the dice) protects the privacy of the respondent and that insight in the misclassification design (in this case the knowledge of the value of  $p$ ) can be used to analyse the observed data.

It is possible to create RR settings in which questions are asked to get information on a variable with  $K > 2$  categories (Chaudhuri & Mukerjee, 1988, Chapter 3). We restrict ourselves in this paper to those RR designs of the form

$$\lambda = P\pi, \quad (2)$$

where  $\lambda = (\lambda_1, \dots, \lambda_K)^T$  is a vector denoting the probabilities of the observed responses with categories 1, ...,  $K$ ,  $\pi = (\pi_1, \dots, \pi_K)^T$  is the vector of the probabilities of the true responses and  $P$  is the  $K \times K$  transition matrix of conditional misclassification probabilities  $p_{ij}$ , with

$$p_{ij} = \mathbb{P}(\text{category } i \text{ is observed} | \text{true category is } j).$$

Note that this means that the columns of  $P$  add up to 1. In the Warner model above we have  $\lambda = (\lambda_1, 1 - \lambda_1)^T$ ,

$$P = \begin{pmatrix} p & 1-p \\ 1-p & p \end{pmatrix},$$

and  $\pi = (\pi_1, 1 - \pi_1)^T$ . Further background and more complex randomized response schemes can be found in Fox & Tracy (1986) and Chaudhuri & Mukerjee (1988).

The second field where the misclassification probabilities are known is the post randomisation method (PRAM), see Kooiman, Willenborg & Gouwelleuw (1997). The idea of PRAM is to misclassify the values of categorical variables after the data have been collected in order to protect the privacy of the respondents by preventing disclosure of their identities. PRAM can be seen as applying RR after the data have been collected. More information about PRAM and a comparison with RR is given in Section 2.

The third field is statistics in medicine and epidemiology. In these disciplines, the probability to be correctly classified as a case given that one is a case is called the sensitivity, and the probability to be correctly classified as a non-case given that one is a non-case is called the specificity. In medicine, research concerning the situation with known sensitivity and specificity is presented in Chen (1989) and Greenland (1980, 1988). In epidemiology, see Magder & Hughes (1997) and Copeland *et al.* (1977).

The fourth field is the part of statistical astronomy that discusses rectification and deconvolution problems. Lucy (1974), for instance, considers the estimation of a frequency distribution where observations might be misclassified and where the misclassification probabilities are presumed known.

To present RR and PRAM as misclassification seems to be a logical approach, but a note must be made on this usage. Misclassification is a well known concept within the analysis of categorical data and different methods to deal with this kind of perturbation have been proposed, see the review paper by Kuha & Skinner (1997), but the situation in which misclassification probabilities are known does not often occur. In most situations, these probabilities have to be estimated which makes analyses of misclassified data more complex.

The focus of this paper is on RR and PRAM. The discussion is about the analysis of the misclassified data, not about the choice of the misclassification probabilities. The central problem is: given the data subject to misclassification and given the transition matrix, how should we adjust standard analysis of frequency tables in order to get valid results?

Special attention is given to the possibility of boundary solutions. By a boundary solution we mean an estimated value of the parameter which lies on the boundary of the parameter space. For instance, in formula (1) the unbiased moment estimate of  $\pi$  is given. It is possible that this estimate is negative and makes no sense. In this case the moment estimate differs from the maximum likelihood estimate

which is zero and therefore lies on the boundary of the parameter space. (This was already noted by, for instance, Singh, 1976.)

The possibility that the moment estimator yields estimates outside the parameter space is an awkward property, since standard analyses as, e.g., univariate probabilities and the odds ratio, are in that case useless. However, negative estimates are likely to occur when RR is used. Typically, RR is applied when sensitive characteristics are investigated and often sensitivity goes hand in hand with rareness. Therefore, some of the true frequencies in a sample may be low and when these frequencies are unbiasedly estimated, random error can easily cause negative estimates. The example discussed in Section 7 illustrates this situation. Regarding PRAM the same problem can occur, see Section 2.

This analysis of misclassified data has also been discussed by other authors, see the references above. Our present aim is to bring together the different fields of misclassification, compare the different methods, and propose methods to deal with boundary solutions. Noticeably lacking in some literature is a discussion of the properties of proposed estimators such as unbiasedness and maximum likelihood. Where appropriate, we try to fill this gap.

Section 2 provides more information about PRAM. A comparison with RR is made. Section 3 discusses the moment estimator of the true frequency table, i.e., the not-observed frequencies of the correctly classified scores. Point estimates and estimations of covariances are presented. In Section 4 we consider the maximum likelihood estimation of the true frequency table. Again point estimation and variances are discussed, this time using the EM algorithm. Section 5 relates the moment estimator to the maximum likelihood estimator. In Section 6 we consider the estimation of the odds ratio. In Section 7 an example is given with RR data stemming from research into violating regulations of social benefit. Section 8 evaluates the results and concludes.

## 2 Protecting Privacy

The post randomisation method (PRAM) was introduced by Kooiman *et al.* (1997) as a method for statistical disclosure control of microdata files. A microdata file is a data matrix where each row, called a record, corresponds to one respondent and where the columns correspond to the variables. Statistical disclosure control (SDC) aims at safeguarding the identity of respondents. Because of the privacy protection, data producers, such as national statistical institutes, are able to pass on data to a third party.

PRAM can be applied to variables in the microdata file that are categorical and identifying. Identifying variables are variables that can be used to re-identify individuals represented in the data. The perturbation of these identifiers makes re-identification of individuals less likely. The PRAM procedure yields a new microdata file in which the scores on certain categorical variables in the original file may be misclassified into different scores according to a given probability mechanism. In this way PRAM introduces uncertainty in the data: the user of the data cannot be sure that the information in the file is original or perturbed due to PRAM. In other words, the randomness of the procedure implies that matching a record in the perturbed file to a record of a known individual in the population could, with a high probability, be a mismatch.

An important aspect of PRAM is that the recipient of the perturbed data is informed about the misclassification probabilities. Using these probabilities he can adjust his analysis and take into account the extra uncertainty caused by applying PRAM.

As with RR, the misclassification scheme is given by means of a  $K \times K$  transition matrix  $P$  of conditional probabilities  $p_{ij}$ , with

$$p_{ij} = \mathbb{P}(\text{category } i \text{ is released} | \text{true category is } j).$$

Since national statistical institutes, which are the typical users of SDC methods, prefer model free approaches to their data, PRAM is presented in the form

$$E[T^*|T] = PT, \quad (3)$$

where  $T^*$  is the vector of perturbed frequencies and  $T$  is the vector of the true frequencies. So instead of using probabilities as in (2), frequencies are used in (3) to avoid commitment to a specific parametric model.

PRAM is currently under study and is by far not the only way to protect microdata against disclosure (see, e.g., Willenborg & De Waal, 2001). Two common methods used by national statistical institutes are global recoding and local suppression. Global recoding means that the number of categories is reduced by pooling, so that the new categories include more respondents than the original categories. This can be necessary when a category in the original file contains just a few respondents. For example, in microdata file where the variable profession has just one respondent with the value 'mayor', we can make a new variable 'working for the government' and include in this category not only the mayor, but also the people in the original file who have governmental jobs. The identity of the mayor is then protected not only by the number of people in the survey with governmental jobs, but also by the number of people in the population with governmental jobs.

Local suppression means protecting identities by making data missing. In the example above, the identity of the mayor can be protected by making the value 'mayor' of the variable profession missing.

When microdata are processed by using recoding or suppression, there is always loss of information. This is inevitable: losing information is intrinsic to SDC. Likewise, there will be loss of information when data are protected by applying PRAM.

PRAM is not meant to replace existing SDC techniques. Using the transition matrix with the misclassification probabilities to take into account the perturbation due to PRAM, requires extra effort and becomes of course more complex when the research questions become more complex. This may not be acceptable to all researchers. Nevertheless, existing SDC methods are also not without problems. Especially global recoding can destroy detail that is needed in the analysis. For instance, when a researcher has specific questions regarding teenagers becoming 18 years old, it is possible that the data he wants to use is globally recoded before it is released. It is possible that the variable age is recoded from year of birth to age categories going from 0 to 5, 5 to 10, 10 to 15, 15 to 20, etcetera. In that case the researcher has lost his object of research.

PRAM can be seen as a SDC method which can deal with specific requests concerning released data (such as in the foregoing paragraph) or with data which are difficult to protect using current SDC methods (meaning the loss of information is too large). PRAM can of course also be used in combination with other SDC methods. Further information about PRAM can be found in Gouweteeuw, Kooiman, Willenborg & De Wolf (1998) and Van den Hout (1999).

The two basic research questions concerning PRAM are (i) how to choose the misclassification probabilities in order to make the released microdata safe, and (ii) how should statistical analysis be adjusted in order to take into account the misclassification probabilities? As already stated in the introduction, this paper concerns (ii). Our general objective is not only to present user-friendly methods in order to make PRAM more user-friendly, but also to show that results in more than one field in statistics can be used to deal with data perturbed by PRAM. Regarding (i), see Willenborg (2000) and Willenborg & De Waal (2001), Chapter 5.

Comparing (2) with (3), it can be seen that RR and PRAM are mathematically equivalent. Therefore, PRAM is presented in this paper as a special form of RR. In fact, the idea of PRAM dates back from Warner (1971), the originator of RR, who mentions the possibilities of the RR procedure to protect data after they have been collected. PRAM can be seen as applying RR after the data have been collected. Rosenberg (1979, 1980) elaborates the Warner idea and calls it ARRC: additive RR contamination. PRAM turns out to be the same as ARRC. Rosenberg discusses multivariate analysis of data protected by ARRC: multivariate categorical linear models and the chi-square test for contingency tables, in particular.

In the remainder of this section we make some comparisons between PRAM and RR. Since the

methods serve different purposes, important differences may occur in practice. First, PRAM will be typically applied to those variables which may give rise to the disclosure of the identity of a respondent, i.e., covariates as, e.g., gender, age and race. RR, on the other hand, will be typically applied to response variables, since the identifying covariates are obvious from the interview situation. Secondly, the usefulness of the observed response in the RR setting is dependent on the cooperation of the respondent, whereas applying PRAM is completely mechanic. Although RR may be of help in eliciting sensitive information, the method is not a panacea, see Van der Heijden *et al.* (2000). The third important difference concerns the choice of the transition matrix. When using RR the matrix is determined *before* the data are collected, but in the case of PRAM the matrix can be determined conditionally on the original data. This means that the extent of randomness in applying PRAM can be controlled better than in the RR setting, see Willenborg (2000).

PRAM is similar to RR regarding the possibility of boundary solutions, see Section 1. PRAM is typically used when there are respondents in the sample with rare combinations of scores. Therefore, some of the true frequencies in a sample may be low and when PRAM has been applied and these frequencies are unbiasedly estimated, random error can easily cause negative estimates. So, also regarding PRAM, methods to deal with boundary solutions are important.

### 3 Moment Estimator

This section generalizes (1) in order to obtain a moment estimator of the true contingency table. A contingency table is a table with the sample frequencies of categorical variables. For example, the 2-dimensional contingency table of two binary variables has four cells, each of which contains the frequency of a compounded class of the two variables. Section 3.1 presents the moment estimator for a  $m$ -dimensional table ( $m > 1$ ). In Section 3.2 formulas to compute covariances are presented.

#### 3.1 Point Estimation

If  $P$  in (2) is non-singular and we have an unbiased point estimate  $\hat{\lambda}$  of  $\lambda$ , we can estimate  $\pi$  by the unbiased moment estimator

$$\hat{\pi} = P^{-1}\hat{\lambda}. \quad (4)$$

see, Chaudhuri & Mukerjee (1988), and Kuha & Skinner (1997).

In practice, assuming that  $P$  in (2) is non-singular does not impose much restriction on the choice of the misclassification design.  $P^{-1}$  exists when the diagonal of  $P$  dominates, i.e.,  $P_{ii} > 1/2$  for  $i \in \{1, \dots, K\}$ , and this is reasonable since these probabilities are the probabilities that the classification is correct.

In this paper we assume that the true response is multinomially distributed with parameter vector  $\pi$ . The moment estimator (4) is not a maximum likelihood estimator since it is possible that for some  $i \in \{1, \dots, K\}$ ,  $\hat{\pi}_i$  is outside the parameter space  $(0,1)$ .

In Section 1, we have considered the misclassification of one variable. The generalization to a  $m$ -dimensional contingency table with  $m > 1$  is straightforward when we have the following independence property between each possible pair  $(A, B)$  of the  $m$  variables:

$$P(A^* = i, B^* = k | A = j, B = l) = P(A^* = i | A = j)P(B^* = k | B = l). \quad (5)$$

Regarding RR, this property means, that the misclassification design is independently applied to the different respondents and, when more than one question is asked, the design is independently applied to the different questions. So, in other words, answers from other respondents or to other questions do not influence the misclassification design in the RR survey. Regarding PRAM, this property means that the misclassification design is independently applied to the different records and independently to the different variables.

In this situation we structure the  $m$ -dimensional contingency table as an 1-dimensional table of the compounded variable. For instance, when we have three binary variables, we get an 1-dimensional table with rows indexed by 111, 112, 121, 122, 211, 212, 221, 222. (The last index changes first.) Due to property (5) it is easy to create the transition matrix of the compounded variable using the transition matrices of the underlying separate variables. Given the observed compounded variable and its transition matrix we can use the moment estimator as described above.

To give an example, assume we have an observed cross-tabulation of the misclassified variables  $A$ , and  $B$ , where row variable  $A$  has  $K$  categories and transition matrix  $P_A$ , and column variable  $B$  has  $S$  categories and transition matrix  $P_B$ . (When one of the variables is not misclassified, we simply take the identity matrix as the transition matrix.) Together  $A$  and  $B$  can be considered as one compounded variable with  $KS$  categories. When property (5) is satisfied we can use the Kronecker product, denoted by  $\otimes$ , to compute the  $KS \times KS$  transition matrix  $P$  as follows:

$$P = P_B \otimes P_A = \begin{pmatrix} P_{11}^B P_A & P_{12}^B P_A & \dots & P_{1S}^B P_A \\ \vdots & \vdots & \ddots & \vdots \\ P_{S1}^B P_A & P_{S2}^B P_A & \dots & P_{SS}^B P_A \end{pmatrix},$$

where each  $P_{ij}^B P_A$  for  $i, j \in \{1, \dots, S\}$ , is itself a  $K \times K$  matrix.

3.2 Covariances

Since the observed response is multinomially distributed with parameter vector  $\lambda$ , the covariance matrix of (4) is given by

$$V(\hat{\pi}) = P^{-1}V(\lambda)(P^{-1})' \\ = n^{-1}P^{-1}(\text{Diag}(\lambda) - \lambda\lambda') (P^{-1})', \tag{6}$$

where  $\text{Diag}(\lambda)$  denotes the diagonal matrix with the elements of  $\lambda$  on the diagonal. The covariance matrix (6) can be unbiasedly estimated by

$$\hat{V}(\hat{\pi}) = (n-1)^{-1}P^{-1}(\text{Diag}(\hat{\lambda}) - \hat{\lambda}\hat{\lambda}') (P^{-1})', \tag{7}$$

see Chaudhuri & Mukerjee (1988, Section 3.3).

As stated before, national statistical institutes prefer a model free approach. Consequently, Kooiman *et al.* (1997) present only the extra variance due to applying PRAM, and do not assume a multinomial distribution. The variance given by Kooiman *et al.* (1997) can be related to (6) in the following way. Chaudhuri & Mukerjee (1988, Section 3.3) present a partition of (6) in two terms, where the first denotes the variance due to the multinomial scheme and the second represents the variance due to the perturbation:

$$V(\hat{\pi}) = \Sigma_1 + \Sigma_2, \tag{8}$$

where

$$\Sigma_1 = \frac{1}{n}(\text{Diag}(\pi) - \pi\pi')$$

and

$$\Sigma_2 = \frac{1}{n}P^{-1}(\text{Diag}(\lambda) - P \text{Diag}(\pi)P') (P^{-1})'.$$

Analysing  $\Sigma_2$  it turns out that it is the same as the variance due to PRAM given in Kooiman *et al.* (1997), as was to be expected, see Appendix A.

4 Maximum Likelihood Estimator

As already noted in Sections 1 and 2, it is possible that the moment estimator yields estimates outside the parameter space when the estimator is applied to RR data or PRAM data. Negative estimates of frequencies are awkward, since they do not make sense. Furthermore, when there are more than two categories and the frequency of one of them is estimated by a negative number, it is unclear how the moment estimate must be adjusted in order to obtain a solution in the parameter space. This is a reason to look for a maximum likelihood estimate (MLE). Another reason to use MLEs is that in general, unbiasedness is not preserved when functions of unbiased estimates are considered. Maximum likelihood properties on the other hand, are in general preserved (see Mood, Graybill & Boes, 1985).

This section discusses first the estimation of the MLE of the true contingency table using the EM algorithm and, secondly, in 4.2, the covariances of this estimate.

4.1 Point Estimation

The expectation-maximization (EM) algorithm (Dempster, Laird & Rubin, 1977) can be used as an iterative scheme to compute MLEs when data are incomplete, i.e., when some observations are missing. The EM algorithm is in that case an alternative to maximizing the likelihood function using methods as, e.g., the Newton Raphson method. Two appealing properties of the EM algorithm relative to Newton-Raphson are its numerical stability and, given that the complete data problem is a standard one, the use of standard software for complete data analysis within the steps of the algorithm. These properties can make the algorithm quite user-friendly. More background and recent developments can be found in McLachlan & Krishnan (1997).

We will now see how the EM algorithm can be used in a misclassification setting, see also Bourke & Moran (1988), Chen (1989), and Kuba & Skinner (1997). For ease of exposition we consider the  $2 \times 1$  frequency table of a binary variable  $A$ . As stated before, we assume multinomial sampling.

When the variable is subject to misclassification, say with given transition matrix  $P = (p_{ij})$ , we do not observe values of  $A$ , but instead we observe values of a perturbed  $A$ , say  $A^*$ .

Let  $A^*$  be tabulated in

$A^*$	1	$n_1^*$
	2	$n_2^*$
		$n$

where  $n_i^*$  for  $i = \{1, 2\}$  is the observed number of values  $i$  of  $A^*$  and  $n_1^* + n_2^* = n$  is fixed. Let  $\pi = P(A = 1)$  and  $\lambda = P(A^* = 1)$ . When transition probabilities are given, we know  $\lambda = p_{11}\pi + p_{12}(1 - \pi)$ . So, ignoring constants, the observed data log likelihood is given by

$$\log l^*(\pi) = n_1^* \log \lambda + n_2^* \log(1 - \lambda) \\ = n_1^* \log(p_{11}\pi + p_{12}(1 - \pi)) + n_2^* \log(p_{21}\pi + p_{22}(1 - \pi)). \tag{9}$$

The aim is to maximize  $\log l^*(\pi)$  for  $\pi \in (0, 1)$ .

In this simple case of a  $2 \times 1$  frequency table, the maximization of  $\log l^*(\pi)$  is no problem. By computing the analytic solution to the root of the first derivative, we can locate the maximum. Nevertheless, in the case of a  $K \times 1$  frequency table, finding the analytic solution can be quite tiresome and we prefer an iterative method. The  $2 \times 1$  table will serve as an example.

To explain the use of the EM algorithm, we can translate the problem of maximizing (9) into an incomplete-data problem. We associate with each observed value of  $A^*$  its not-observed non-perturbed value of  $A$ . Together these pairs form an incomplete-data file with size  $n$ . (In the framework of Rubin (1976): the missing data are missing at random, since they are missing by design.) When we tabulate this incomplete-data file we get

	A		
	1	2	
A*	$n_{11}$	$n_{12}$	$n_1^*$
	$n_{21}$	$n_{22}$	$n_2^*$
	$n_1$	$n_2$	$n$

(10)

where for  $i, j \in \{1, 2\}$ ,  $n_{ij}$  is the frequency of the combination  $A^* = i$  and  $A = j$ . Only the marginals  $n_1^*$  and  $n_2^*$  are observed.

When we would have observed the complete data, i.e.,  $n_{ij}$  for  $i, j \in \{1, 2\}$ , we would only have to consider the bottom marginal of (10) and the complete-data log likelihood function of  $\pi$  would be given by

$$\log l(\pi) = n_1 \log \pi + n_2 \log(1 - \pi), \tag{11}$$

from which the maximum likelihood estimate  $\hat{\pi} = n_1/n$  follows almost immediately.

The idea of the EM algorithm is to maximize the incomplete-data likelihood by iteratively maximizing the expected value of the complete-data log likelihood (11), where the expectation is taken over the distribution of the complete-data given the observed data and the current fit of  $\pi$  at iteration  $p$ , denoted by  $\pi^{(p)}$ . That is, in each iteration we look for the  $\pi$  which maximizes the function

$$Q(\pi, \pi^{(p)}) = \mathbb{E}[\log l(\pi) | n_1^*, n_2^*, \pi^{(p)}]. \tag{12}$$

In the EM algorithm it is not necessary to specify the corresponding representation of the incomplete-data likelihood in terms of the complete-data likelihood (McLachlan & Krishnan, 1997, Section 1.5.1). In other words, we do not need (9), the function which plays the role of the incomplete-data likelihood, but we can work with (11) instead.

Since (11) is linear with respect to  $n_i$ , we can rewrite (12) by replacing the unknown  $n_i$ 's in (11) by the expected values of  $n_i$ 's given the observed  $n_i^*$ 's and  $\pi^{(p)}$ . Furthermore, since  $n = n_1^* + n_2^*$ , and  $n$  is known,  $n_2^*$  does not contain extra information. Therefore, (12) is equal to:

$$Q(\pi, \pi^{(p)}) = \mathbb{E}[N_1 | n_1^*, \pi^{(p)}] \log \pi + \mathbb{E}[N_2 | n_1^*, \pi^{(p)}] \log(1 - \pi), \tag{13}$$

where  $N_1$  and  $N_2$  are the stochastic variables with values  $n_1$  and  $n_2$ , and, of course,  $N_1 + N_2 = n$ . The EM algorithm consists in each iteration of two steps: the E-step and the M-step. In this situation the E-step consist of estimating  $\mathbb{E}[N_1 | n_1^*, \pi^{(p)}]$ . We assume that  $(n_{11}, n_{12}, n_{21}, n_{22})$  are values of the stochastic variables  $(N_{11}, N_{12}, N_{21}, N_{22})$  which are multinomially distributed with parameters  $(n, \pi_{11}, \pi_{12}, \pi_{21}, \pi_{22})$ . A property of the multinomial distribution is that the conditional distribution of  $(N_{11}, N_{12})$  given  $n_{1+} = n_1^*$  is again multinomial with parameters  $(n_1^*, \pi_{11}/\pi_{1+}, \pi_{12}/\pi_{1+})$ , for  $i \in \{1, 2\}$ . So we have

$$\mathbb{E}[N_{ij} | n_1^*, \pi_{11}, \pi_{12}, \pi_{21}, \pi_{22}] = n_1^* \frac{\pi_{ij}}{\pi_{1+}}.$$

And consequently

$$\mathbb{E}[N_1 | n_1^*, \pi_{11}, \pi_{12}, \pi_{21}, \pi_{22}] = \frac{\pi_{11}}{\pi_{1+}} n_1^* + \frac{\pi_{21}}{\pi_{2+}} n_2^*. \tag{14}$$

See also Schafer (1997, Section 3.2.2.).

In order to use the updates  $\pi^{(p)}$  of  $\pi = \mathbb{P}(A = 1)$  and the fixed misclassification probabilities we note that

$$\pi_{11} = \mathbb{P}(A^* = i, A = 1) = \mathbb{P}(A^* = i | A = 1) \mathbb{P}(A = 1), \tag{15}$$

and

$$\pi_{i+} = \mathbb{P}(A^* = i) = \sum_{k=1}^2 \mathbb{P}(A^* = i | A = k) \mathbb{P}(A = k). \tag{16}$$

Next, we use (14), (15) and (16) to estimate  $\mathbb{E}[N_{ij} | n_1^*, \pi^{(p)}]$  by

$$n_{ij}^{(p)} = \sum_{i=1}^2 \frac{p_{i1} \pi^{(p)}}{p_{i1} \pi^{(p)} + p_{i2} (1 - \pi^{(p)})} n_i^*,$$

which ends the E-step.

The M-step gives an update for  $\pi$ , which is the value of  $\pi$  that maximizes (13), using the current estimate of  $\mathbb{E}[N_{ij} | n_1^*, \pi^{(p)}]$ , which also provides an estimate of  $\mathbb{E}[N_2 | n_1^*, \pi^{(p)}] = n - \mathbb{E}[N_1 | n_1^*, \pi^{(p)}]$ . Maximizing is easy due to the correspondence between the standard form of (11) and the form of (13):  $\pi^{(p+1)} = n_1^{(p)}/n$ .

The EM algorithm is started with an initial value  $\pi^{(0)}$ . The following can be stated regarding the choice of the initial value and convergence of the algorithm. When there is a unique maximum in the interior of the parameter space, the EM algorithm will find it, see the convergence theorems of the algorithm as discussed in McLachlan & Krishnan (1997, Section 3.4). Furthermore, as will be explained in Section 5, in the RR/PRAM setting, the incomplete-data likelihood is from a regular exponential family and is therefore strictly concave, so finding the maximum should not pose any difficulties when the starting point is chosen in the interior of the parameter space and the maximum is also achieved in the interior.

In general, let  $A$  have  $K$  categories and for  $i, j \in \{1, 2, \dots, K\}$ : let  $\pi_j = \mathbb{P}(A = j)$ , let  $n_{ij}$  denote the cell frequencies in the  $K \times K$  table of  $A^*$  and  $A$ , let  $n_j$  denote the frequencies in the  $K \times 1$  table of  $A$ , and let  $n_j^*$  denote the frequencies in the observed  $K \times 1$  table of  $A^*$ . The observed data log likelihood is given by

$$\log l^*(\pi) = \sum_{i=1}^K n_i^* \log \lambda_i + C \tag{17}$$

where  $\lambda_i = \sum_{k=1}^K p_{ik} \pi_k$  and  $C$  is a constant.

The EM algorithm in this situation and presented as such in Kuha & Skinner (1997) is

Initial values:  $\pi_j^{(0)} = \frac{n_j^*}{n}$

E-step:  $n_{ij}^{(p)} = \frac{p_{ij} \pi_j^{(p)}}{\sum_{k=1}^K p_{ik} \pi_k^{(p)}} n_i^*$

$n_j^{(p)} = \sum_{i=1}^K n_{ij}^{(p)}$

M-step:  $\pi_j^{(p+1)} = \frac{n_j^{(p)}}{n}$ .

Note that  $\pi_j^{(p)} < 0$  is not possible for  $j \in \{1, 2, \dots, K\}$ .

This section discussed the misclassification of one variable, but as shown in section 3, the generalization to a  $m$ -dimensional contingency table with  $m > 1$  is straightforward when we have property (5) for each possible pair of the  $m$  variables. In that case we create a compounded variable, put together the transition matrix of this variable and use the EM algorithm as described above.

4.2 Covariances

Consider the general case where  $A$  has  $K$  categories and the observed data log likelihood is given by (17). Assuming that the MLE of  $\pi$  lies in the interior of the parameter space, we can use the information matrix to estimate the asymptotic covariance matrix of the parameters. Using  $\pi_k = 1 - \sum_{l=1}^{k-1} \pi_l$ , we obtain for  $k, l \in \{1, \dots, K-1\}$  the  $kl$ -component of the information matrix:

$$-\frac{\partial}{\partial \pi_k \partial \pi_l} \log l^*(\pi) = \sum_{i=1}^K \frac{n_i^*}{\lambda_i^2} (p_{il} - p_{ik})(p_{ik} - p_{il} \kappa). \tag{18}$$

Incorporating the estimate  $\hat{\lambda}_i = n_i^*/n$  in (18) we get an approximation of the information matrix where for  $k, l \in \{1, \dots, K-1\}$  the  $kl$ -component is given by

$$\sum_{i=1}^K \frac{n}{\lambda_i} (p_{ik} - p_{il} \kappa)(p_{il} - p_{ik} \kappa), \tag{19}$$

see Bourke & Moran (1988). The inverse of this approximation can be used as an estimator of the asymptotic covariance matrix.

When the MLE of  $\pi$  is on the boundary of the parameter space, using the information matrix is not appropriate and we suggest to use the bootstrap percentile method to estimate a 95% confidence interval (regarding the bootstrap, see Efron & Tibshirani, 1993). The bootstrap scheme we propose is the following. We draw  $B$  bootstrap samples from a multinomial distribution with parameter vector  $\hat{\lambda} = (\hat{\lambda}_1, \dots, \hat{\lambda}_K)'$  and estimate for each bootstrap replication the corresponding parameter  $\hat{\pi}^* = (\hat{\pi}_1^*, \dots, \hat{\pi}_K^*)'$  of the multinomial distribution of the true table. The resulting  $B$  bootstrap estimates  $\hat{\pi}_i^*$  are sorted for each  $i \in \{1, \dots, K\}$  from small to large and the confidence interval for each  $\hat{\pi}_i$  is constructed by deleting 5% of the sorted values: 2.5% of the smallest estimates and 2.5% of the largest.

Note that this scheme incorporates the double stochastic scheme of the RR setting: the variance due to the multinomial distribution and the extra variance due to applying RR. A disadvantage of the bootstrap in this setting is that computations can take some time since the bootstrap is combined with the EM algorithm.

5 The MLE Compared to the Moment Estimate

In this section we prove that the observed log likelihood function  $\log l^*(\pi)$  given in (17) is the log likelihood of a distribution from a regular exponential family. Using this property of  $l^*(\pi)$ , the uniqueness of a solution of the likelihood equations is established when this solution is found in the interior of the parameter space. Furthermore, we prove that when the MLE is in the interior of the parameter space, the MLE is equal to the estimate provided by the moment estimator. This equality has been observed by several authors (Schwartz, 1985, app. A, Bourke & Moran, 1988, and Chen, 1989) but theoretic proof is not given. By using the exponential family we prove this equality and thus provide an alternative to results in Lucy (1974) as far as they apply to misclassification of categorical variables.

First, to determine that  $l^*(\pi)$  is from an exponential family, we have to show that this function can be written in the following form

$$l^*(\pi) = a(\pi) b(\mathbf{n}^*) \exp(\theta'(\pi) \mathbf{t}(\mathbf{n}^*)), \tag{20}$$

see Barndorff-Nielsen (1982).

Let

$$a(\pi) = 1,$$

$$b(\mathbf{n}^*) = \frac{n!}{n_1^! \dots n_K^!},$$

the sufficient statistic

$$\mathbf{t}(\mathbf{n}^*) = (t_1(\mathbf{n}^*), \dots, t_K(\mathbf{n}^*))' = (n_1^*, \dots, n_K^*)',$$

and the canonical parameter

$$\begin{aligned} \theta'(\pi) &= (\theta_1(\pi), \dots, \theta_K(\pi)) \\ &= (\log \lambda_1, \dots, \log \lambda_K) \\ &= (\log \sum_{j=1}^K p_{1j} \pi_j, \dots, \log \sum_{j=1}^K p_{Kj} \pi_j). \end{aligned}$$

Due to the affine constraint  $n_1^* + \dots + n_K^* = n$ , the exponential representation in (20) where the functions are defined as above, is not minimal, i.e., it is possible to define  $\mathbf{t}$  and  $\theta$  in such a way that their dimensions are smaller than  $K$ . Since we need a minimal representation in order to establish regularity, we provide alternative definitions of the functions in (20).

A minimal representation is obtained by taking

$$\mathbf{t}(\mathbf{n}^*) = (n_1^*, \dots, n_{K-1}^*)' \tag{21}$$

and

$$\theta'(\pi) = (\theta_1(\pi), \dots, \theta_{K-1}(\pi)) = (\log \frac{\lambda_1}{\lambda_K}, \dots, \log \frac{\lambda_{K-1}}{\lambda_K}), \tag{22}$$

where again  $\lambda_i = \sum_{j=1}^K p_{ij} \pi_j$ . We get as a minimal representation

$$l^*(\pi) = (1 + e^{\theta_1} + \dots + e^{\theta_{K-1}})^{-n} \frac{n!}{n_1^! \dots n_K^!} \exp(\theta_1 n_1^* + \dots + \theta_{K-1} n_{K-1}^*), \tag{23}$$

where  $\theta_i$  stands for  $\theta_i(\pi)$ ,  $i \in \{1, \dots, K-1\}$ .

Having established that  $l^*(\pi)$  is from an exponential family, we now prove, using (23), that the function is from a regular exponential family. We follow the definitions of regularity as given by Barndorff-Nielsen (1982). Let  $\Omega$  be the domain of variation for  $\pi$  and  $\Theta = \theta(\Omega)$  the canonical parameter domain. We must prove two properties:

(i)  $\Theta$  is an open subset of  $\mathbb{R}^{K-1}$ , and

(ii)

$$\Theta = \{\theta \mid \theta \in \theta(\Omega) \mid \int_X \frac{n!}{x_1! \dots x_K!} e^{\theta'(\mathbf{x})} d\mathbf{x} < \infty\}, \tag{24}$$

where  $X = \{\mathbf{x} \mid \mathbf{x} = (x_1, \dots, x_K)' \mid x_1, \dots, x_K > 0, x_1 + \dots + x_K = n\}$  and  $\theta$  and  $\mathbf{t}$  are given in (22) and (21) respectively.

Regarding property (i):  $\Omega = \{\pi \mid \pi \in (0, 1)^K \mid \pi_1 + \dots + \pi_K = 1\}$ . Since  $p_{ij} \geq 0$  and  $\pi_j > 0$  for  $i, j \in \{1, \dots, K\}$ , and no column in the transition matrix  $P = (p_{ij})$  consists only of zeroes, it follows that  $\lambda_i > 0$  for  $i \in \{1, \dots, K\}$ . Furthermore, again using the properties of the transition matrix, from  $\pi_1 + \dots + \pi_K = 1$  it follows that  $\lambda_1 + \dots + \lambda_K = 1$ . So  $\Theta = \{\theta \mid \theta = \log \lambda_i \lambda_K^{-1} \mid \lambda_1 + \dots + \lambda_K = 1, \lambda_i > 0\}$ . For each  $r = (r_1, \dots, r_{K-1}) \in \mathbb{R}^{K-1}$ , there is a choice for  $\lambda_1, \dots, \lambda_K$  such that  $\lambda_1 + \dots + \lambda_K = 1$ ,  $\lambda_i > 0$  for  $i \in \{1, \dots, K\}$ , and  $\log \lambda_i \lambda_K^{-1} = r_i$  for  $i \in \{1, \dots, K-1\}$ . So property (i) is satisfied by the equality  $\Theta = \mathbb{R}^{K-1}$ .

Regarding property (ii):

$$\begin{aligned} \int_X \frac{n!}{x_1! \cdots x_k!} e^{\theta(x)} dx &\leq n! \int_X \left(\frac{\lambda_1}{\lambda_k}\right)^{x_1} \cdots \left(\frac{\lambda_{k-1}}{\lambda_k}\right)^{x_{k-1}} dx \\ &= n! \int_X \lambda_1^{x_1} \cdots \lambda_k^{x_k} \frac{1}{\lambda_k} dx < \infty, \\ &\leq n! \int_X \left(\frac{1}{\lambda_k}\right)^n dx < \infty, \end{aligned}$$

for every  $\lambda_k \in (0, 1)$  and  $n = x_1 + \dots + x_k$ . This means that (24) is satisfied.

Having shown that the observed data log likelihood  $l^*(\pi)$  is from a regular exponential family, we can use the powerful theory that exists for this family. A property that is of practical use is that the maximum of the observed data likelihood is unique when found in the interior of the parameter space, since the likelihood is strictly concave (Barndorff-Nielsen, 1982). This justifies the use of the maximum found by the EM algorithm in Section 4.1.

A second property concerns the comparison of the MLE and the estimate provided by the moment estimator. The two estimates are equal when both are in the interior of the parameter space. The equality can be proved as follows. We continue to use the minimal representation as given in (23) where  $\theta$  is the canonical parameter and where  $\alpha(\pi)$  is given by

$$\alpha(\pi) = (1 + e^{\theta_1} + \dots + e^{\theta_{k-1}})^{-n}.$$

When  $\log l^*(\pi)$  is maximized, we solve the likelihood equations

$$\frac{\partial}{\partial \theta} \log l^*(\pi) = 0.$$

That is

$$\frac{\partial}{\partial \theta} (\theta^t \mathbf{t}(\mathbf{n}^*)) = \frac{\partial}{\partial \theta} (-\log \alpha(\pi)). \tag{25}$$

We have

$$\frac{\partial}{\partial \theta} (\theta^t \mathbf{t}(\mathbf{n}^*)) = (n_1^*, \dots, n_{k-1}^*)^t \tag{26}$$

and according to the theory of the exponential family (Barndorff-Nielsen, 1982)

$$\frac{\partial}{\partial \theta} (-\log \alpha(\pi)) = \mathbb{E} [\mathbf{t}(\mathbf{N}^*)] = n \begin{pmatrix} p_{11} & p_{12} & \dots & p_{1k} \\ \vdots & \vdots & \ddots & \vdots \\ p_{k-1,k} & \dots & \dots & p_{k-1,k} \end{pmatrix} \begin{pmatrix} \pi_1 \\ \pi_2 \\ \vdots \\ \pi_k \end{pmatrix}, \tag{27}$$

where  $\mathbf{N}^*$  is the random variable which has value  $\mathbf{n}^*$ . So, combining (26) and (27) in (25) shows that the likelihood equations (25) are equal to the equations (2) on which the moment estimator is based. Conclusion: in the interior of the parameter space MLE is equal to the ME.

Of course, the above properties of  $l^*(\pi)$  can be derived without references to exponential families. Lucy (1974) discusses the estimation of a frequency distribution where observations are subject to measurement error and the error distribution is presumed known. The difference with our setting is that the observed variable is a continuous one. However, the observations are categorized in intervals and correction of the observations is on the basis of these intervals, so measurement error can be easily translated to misclassification of categorical variables. Lucy (1974) advocates an EM algorithm comparable with the EM algorithm given above. Furthermore, it is proved that in the interior of the parameter space the MLE is equal to the moment estimate and the maximum of the likelihood is unique. In Appendix B we have translated Lucy's proof regarding the equivalence between the MLE

and the moment estimate to our setting.

### 6 Odds Ratio

This section discusses the estimation of the odds ratio when data are perturbed by PRAM or RR. The odds ratio  $\theta$  is a measure of association for contingency tables. We will not go into the rationale of using the odds ratio, information about this measure can be found in most textbooks on categorical data analysis, see, e.g., Agresti (1990, 1996).

Section 6.1 discusses point estimation both in the situation without and with misclassification. Two estimates of the odds ratio given by different authors are the same, but are not always the MLE. Section 6.2 discusses the variance. Again, it is important whether the estimates of the original frequencies are in the interior of the parameter space or not.

#### 6.1 Point Estimation

We start with the situation without misclassification. Let  $\pi_{ij} = \mathbb{P}(A = i, B = j)$  for  $i, j \in \{1, 2\}$  denote the probability that the scores for A and B fall in the cell in row  $i$  and column  $j$ , respectively. The odds ratio is defined as

$$\begin{aligned} \theta &= \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}}. \\ \hat{\theta} &= \frac{n_{11}n_{22}}{n_{12}n_{21}}. \end{aligned} \tag{28}$$

With  $n_{ij}$  the observed frequency in the cell with probability  $\pi_{ij}$ , the sample odds ratio equals

The sample odds ratio equals 0 or  $\infty$  if any  $n_{ij} = 0$ , and it is not defined if both entries in a row or column are zero. The value 1 means independence of A and B. For multinomial sampling, this is the MLE of the odds ratio (Agresti, 1996).

It is possible to use sample proportions to compute the sample odds ratio. With  $p_{A|B}(i,j) = n_{ij}/(n_{1j} + n_{2j})$  we get

$$\hat{\theta} = \frac{p_{A|B}(1|1) - p_{A|B}(1|2)}{1 - p_{A|B}(1|1)} \left( \frac{p_{A|B}(1|2) - p_{A|B}(1|2)}{1 - p_{A|B}(1|2)} \right)^{-1}. \tag{29}$$

Next, we consider the situation with misclassification. Two estimates of the odds ratio are proposed in the literature. Let only variable A be subject to misclassification, and the  $2 \times 2$  transition matrix be given by  $P = (p_{ij})$ . First, Magder & Hughes (1997) suggest to adjust formula (29) as

$$\hat{\theta}_1 = \frac{p_{A^*|B}(1|1) - p_{12}}{p_{11} - p_{A^*|B}(1|1)} \left( \frac{p_{A^*|B}(1|2) - p_{12}}{p_{11} - p_{A^*|B}(1|2)} \right)^{-1}, \tag{30}$$

where  $p_{A^*|B}(i|j) = n_{ij}^*/(n_{1j}^* + n_{2j}^*)$  with  $n_{ij}^*$  the observed cell frequencies. This formula can be used only if all the numerators and denominators in the formula are positive. If one of these is negative, the estimate is 0 or  $\infty$ . According to Magder & Hughes (1997), (30) is the MLE of  $\theta$ . Assuming that  $\theta_1$  is not equal to zero or infinity, it will always be further from 1 than the odds ratio  $\theta$  which is computed in the standard way using the observed table. Incorporating the information of the transition matrix in the estimation process compensates for the bias towards 1 (Magder & Hughes, 1997).

Secondly, Greenland (1988) suggests to estimate the probabilities of the true frequencies using the moment estimator, yielding estimated frequencies  $\hat{n}_{ij} = n\hat{\pi}_{ij}$ , and then estimate the odds ratio

using its standard form:

$$\hat{\theta}_2 = \frac{\hat{n}_{11}\hat{n}_{22}}{\hat{n}_{12}\hat{n}_{21}} \quad (31)$$

This procedure can also be used when  $A$  and  $B$  are both misclassified.

In order to compare (30) and (31), we distinguish two situations concerning the misclassification of only  $A$ . First, the situation where estimated frequencies are in the interior of the parameter space, or, in other words, where the moment estimate of the frequencies is equal to the MLE. In this case, (30) and (31) are identical, which can be easily proved by writing out. Furthermore, (31), and thus (30), is the MLE due to the invariance property of maximum likelihood estimation (Mood *et al.*, 1985).

Secondly, if the moment estimator yields probabilities outside the parameter space, we should compute (31) using the MLE, and consequently (30) and (31) differ. In fact, (30) is not properly defined, since it might be a negative value corresponding to the negative cell frequencies estimated by the moment estimator. Therefore, as noted in Magder & Hughes (1997), the estimate of the odds ratio should be adjusted to be either 0 or  $\infty$ .

The advantage of formula (30) is that we can use the observed table. A disadvantage is that (30) is not naturally extended to the situation where two variables are misclassified.

## 6.2 Variance

We now turn to the variance estimator of the odds ratio.

First we describe the situation without misclassification. Since outcomes  $n_{ij} = 0$  have positive probability, the expected value and variance of  $\hat{\theta}$  do not exist. It has been shown that

$$\hat{\theta} = \frac{(n_{11} + 0.5)(n_{22} + 0.5)}{(n_{12} + 0.5)(n_{21} + 0.5)},$$

has the same asymptotic normal distribution around  $\theta$  as  $\hat{\theta}$  (compare Agresti, 1990). Note that  $\hat{\theta}$  has a variance.

The close relation between  $\hat{\theta}$  and  $\hat{\theta}$  is the reason we will discuss asymptotic standard error (ASE) of  $\log \hat{\theta}$ , although it is not mathematically sound to do so.

There are at least two methods available to estimate the ASE of  $\log \hat{\theta}$ . The first method is using the delta method. The estimated ASE is then given by

$$ASE(\log \hat{\theta}) = \left( \frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}} \right)^{1/2},$$

see Agresti (1990, Sections 3.4 and 12.1).

The second method to estimate the ASE in the situation without misclassification is to use the bootstrap. For instance, we can use the bootstrap percentile method to estimate a 95% confidence interval. When we assume the multinomial distribution, we take the vector of observed cell proportions as MLE of the cell probabilities. With this MLE we simulate a large number of multinomial tables and each time compute the odds ratio. Then we estimate a 95% confidence interval is the same way as described in Section 4.2.

Next, we consider the situation with misclassification. Along the line of the two methods described above, we discuss two methods to estimate the variance of the estimate of the odds ratio. First, when the moment estimator is used, the delta-method can be applied to determine the variance of the log odds ratio. Greenland (1988) shows how this can be done when the transition matrix is estimated with known variances. Our situation is easier, since the transition matrix is given. We use the multivariate delta method (Bishop, Fienberg & Holland, 1975, Section 14.6.3). The random vector is  $\hat{\pi}$

$= (\hat{\pi}_{11}, \hat{\pi}_{12}, \hat{\pi}_{21}, \hat{\pi}_{22})'$  with  $4 \times 4$  asymptotic covariance-variance matrix  $V(\hat{\pi})$ , see Section 3. We take the function  $f$  to be

$$f(\pi) = \log \left( \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}} \right),$$

which has a derivative at  $\pi \in (0, 1)^4$ . The delta method provides the asymptotic variance  $V_f$  for  $f(\hat{\pi})$ :

$$V_f = (Df)'V(\hat{\pi})Df,$$

where  $Df$  is the gradient vector of  $f$  at  $\hat{\pi}$  and  $V(\hat{\pi})$  is given by (6).

The problem with this method is that it makes use of the moment estimator which only makes sense when this estimator yields a solution in the interior of the parameter space. A second way to estimate the variance is using the bootstrap method and can also be applied in combination with the EM algorithm. Since the table provided by the algorithm is an estimate, we should take into account the variance of this estimate when estimating the variance of the estimate of the odds ratio. This motivates the following procedure. (i) We assume that the observed table is multinomially distributed and we draw a large number of samples from this distribution. Let  $B$  be the number of samples. (ii) For each of the  $B$  samples we estimate the true table using the EM algorithm and furthermore we estimate the odds ratio. This results in the numbers  $\theta_1^*, \dots, \theta_B^*$ . We then use the bootstrap percentile method, see Section 4.2, to estimate from  $\theta_1^*, \dots, \theta_B^*$  a 95% confidence interval.

## 7 An Example

This section illustrates the foregoing by estimating tables of true frequencies on the basis of data collected using RR. Also, in Section 7.2, an estimate of the odds ratio will be discussed. The example makes clear that boundary solutions can occur when RR is applied and that we need to apply methods such as the EM algorithm and the bootstrap.

### 7.1 Frequencies

The RR data we want to analyse stem from a research into violating regulations of social benefit (Van Gils, Van der Heijden & Rosebeek, 2001). Sensitive items were binary: respondents were asked whether or not they violated certain regulations.

The research used the RR procedure introduced by Kuk (1990) where the misclassification design is constructed by using stacks of cards. Since the items in the present research were binary, two stacks of cards were used. In the right stack the proportion of red cards was 8/10, and in the left stack it was 2/10. The respondent was asked to draw one card from each stack. Then the sensitive question was asked, and when the answer to it was 'yes', the respondent should name the color of the card of the right stack, and when the answer was 'no', the respondent should name the color of the card of the left stack.

We associate violations with the color red. In this way the probability to be correctly classified is 8/10 both for respondents who violated regulations and for those who did not. The transition matrix is therefore given by

$$P = \begin{pmatrix} 8/10 & 2/10 \\ 2/10 & 8/10 \end{pmatrix}. \quad (32)$$

We discuss observed frequencies of red or black regarding two questions,  $Q_1$  and  $Q_2$ , which were asked using this RR procedure. Both questions concern the period in which the respondent received a benefit. In translation, question  $Q_1$ : Did you turn down a job offer, or did you endanger on

purpose an offer to get a job? And  $Q_2$ : Was the number of job applications less than required? In our discussion, we deal first with the frequencies of the separate questions, and secondly, we take them together, meaning that we tabulate the frequencies of the four possible profiles: red-red, red-black, black-red, black-black, and we want to know the true frequencies of the profiles violation-violation, violation-no violation, no violation-violation and no violation-no violation.

We assume that the data are multinomially distributed, so the correspondence between the probabilities and the frequencies is direct, given  $n = 412$ , the size of the data set. Univariate observed frequencies are

$Q_1$	red	120	$Q_2$	red	171
	black	292		black	241

and

$Q_1$	violation	62.67	$Q_2$	violation	147.67
	no violation	349.33		no violation	264.33

and

$Q_1$	red	68	$Q_2$	red	120
	black	103		black	189
		171			241
					292
					412

Table 1

Response Profiles

		$Q_2$	
		red	black
$Q_1$	red	68	52
	black	103	189
		171	241
			292
			412

The moment estimate of the true frequencies is equal to the MLE since the solution is in the interior of the parameter space:

Since we have for each respondent the answers to both RR questions, we can tabulate the frequencies of the 4 possible response profiles, see table 1. Using the Kronecker product to determine the

$4 \times 4$  transition matrix, see Section 3.1, the moment estimate yields a negative cell entry, see table 2. The MLE can be computed by using the EM algorithm as described in Section 4.1 and is given

Table 2

Moment Estimate

		$Q_2$	
		violation	no violation
$Q_1$	violation	73.00	-10.33
	no violation	74.67	274.66
		147.67	264.33
			412

by table 3.

Table 3

MLE

		$Q_2$	
		violation	no violation
$Q_1$	violation	67.98	0.00
	no violation	78.33	265.69
		146.31	265.69
			412

There is a discrepancy which shows up in this example. From table 3, we can deduce estimated univariate frequencies of answers to  $Q_1$  and  $Q_2$ . These estimates, which are based on the MLE of the true multivariate frequencies are different from the univariate moment estimates which are also

MLE's. Differences however are small.

Next, we turn to the estimation of variance. First, the univariate case, where we only discuss question  $Q_1$ . The estimated probability of violation is  $\hat{\pi} = 62.67/412 = 0.152$ . The estimated standard error of  $\hat{\pi}$  can be computed by using (7) and is estimated to be 0.037. The bootstrap scheme as explained in Section 4.2 can also be used. With  $B = 500$  we compute the bootstrap standard error of  $\hat{\pi}$  from the sample covariance matrix of the bootstrap estimates  $\hat{\pi}_1^*, \dots, \hat{\pi}_B^*$ . This procedure yields the same estimate of the standard error as above.

Secondly, we compute the variance of the four estimated probabilities concerning profiles of violation. From table 3 we obtain  $\hat{\pi} = (\hat{\pi}_1, \dots, \hat{\pi}_4)' = (0.17, 0.00, 0.19, 0.64)'$ . Since the MLE is on the boundary of the parameter space, estimating a 95% confidence interval is more useful than estimating standard errors. We use the bootstrap percentile method as explained in Section 4.2. Using  $B = 500$  we obtain the four intervals [0.10, 0.22], [0.00, 0.04], [0.12, 0.28], and [0.56, 0.72], for  $\hat{\pi} = (\hat{\pi}_1, \dots, \hat{\pi}_4)'$

### 7.2 Odds Ratio

To determine whether the items corresponding to  $Q_1$  and  $Q_2$  are associated, we want to know the odds ratio. The starting point is the  $2 \times 2$  contingency table of observed answers to  $Q_1$  and  $Q_2$ , given by table 1. Without any adjustment, the estimated odds ratio is  $(68 \cdot 189) / (103 \cdot 52) = 2.40$ .

Since we have two misclassified variables, we can not use (30) to estimate the odds ratio. Instead, we estimate the  $2 \times 2$  contingency table of the true frequencies and then compute the odds ratio in the standard way, as in (31). The moment estimate in table 2 of the true frequencies yields a negative frequency, so the MLE in table 3 is used. The estimate of the odds ratio is  $\hat{\theta}_2 = \infty$ . This means, that given that rule 1 is violated, the probability that rule 2 is also violated is estimated to be 1. The bootstrap percentile method is used to construct a 95% confidence interval, see Section 4.2. In this case the interval is infinite and we are interested in the lower bound. We delete the smallest 5% of the 500 bootstrap estimates of the odds ratio and obtain the 95% confidence interval [11.32628,  $\infty$ ). So there is no reason to believe in independence between the answers to the questions. Furthermore, adjusting for the misclassification shows that the estimate of the odds ratio is much further away from 1 than the estimate based on the observed table alone.

### 8 Conclusion

The aim of this paper is to review the different fields of misclassification where misclassification probabilities are known, and to compare estimators of the true contingency table and the odds ratio. Special attention is devoted to the possibility of boundary solutions. The matrix based moment estimator is quite elegant, but there are problems concerning solutions outside the parameter space. We have explained and illustrated with the example that these problems are likely to occur when randomized response or PRAM is applied, since these procedures are often applied to skewed distributions. The maximum likelihood estimator is a good alternative to the moment estimator but demands more work, since the likelihood function is maximized numerically using the EM algorithm. When boundary solutions are obtained, we suggest the bootstrap method to compute confidence intervals.

The proof of the equality of the moment estimate and the maximum likelihood estimate, when these estimates are in the interior of the parameter space, is interesting because it establishes theoretically what was conjectured by others on the basis of numerical output.

Regarding PRAM, the results are useful in the sense that they show that frequency analysis with the released data is possible and that there is ongoing research in the field of RR and misclassification which deals with the problems that are encountered. This is important concerning the acceptance of

PRAM as a SDC method.

Regarding RR, the example illustrates that a boundary solution may be encountered in practice. This possibility was also noted by others but is, as far as we know, not investigated in the multivariate situation with attention to the estimation of standard errors.

## References

- Agresti, A. (1990). *Categorical Data Analysis*. New York: Wiley.
- Agresti, A. (1996). *An Introduction to Categorical Data Analysis*. New York: Wiley.
- Barndorff-Nielsen, O. (1982). Exponential Families. *Encyclopedia of Statistical Sciences*, Eds. S. Kotz and N.L. Norman. New York: Wiley.
- Bishop, Y.M.M., Fienberg, S.E. & Holland, P.W. (1975). *Discrete Multivariate Analysis*. Cambridge: MIT Press.
- Bourke, P.D. & Moran, M.A. (1988). Estimating Proportions From Randomized Response Data Using the EM Algorithm. *J. Am. Stat. Assoc.* **83**, 964–968.
- Chaudhuri, A. & Mukerjee, R. (1988). *Randomized Response: Theory and Techniques*. New York: Marcel Dekker.
- Chen, T.T. (1989). A Review of Methods for Misclassified Categorical Data in Epidemiology. *Stat. Med.*, **8**, 1095–1106.
- Copeland, K.T., Checkoway, H., McMichael, A.J. & Holbrook, R.H. (1977). Bias Due to Misclassification in the Estimation of Relative Risk. *Am. J. Epidemiol.*, **105**, 488–495.
- Dempster, A.P., Laird, N.M. & Rubin, D.B. (1977). Maximum Likelihood from Incomplete Data via the EM algorithm. *J. R. Stat. Soc.*, **39**, 1–38.
- Efron, B. & Tibshirani, R.J. (1993). *An Introduction to the Bootstrap*. New York: Chapman and Hall.
- Fox, J.A. & Tracy, P.E. (1986). *Randomized Response: A method for Sensitive Surveys*. Newbury Park: Sage.
- Gouweleew, J.M., Kooiman, P., Willenborg, L.C.R.J. & De Wolf, P.-P. (1998). Post Randomisation for Statistical Disclosure Control: Theory and Implementation. *J. Off. Stat.*, **14**, 463–478.
- Greenland, S. (1980). The Effect of Misclassification in the Presence of Covariates. *Am. J. Epidemiol.*, **112**, 564–569.
- Greenland, S. (1988). Variance Estimation for Epidemiologic Effect Estimates under Misclassification. *Stat. Med.*, **7**, 745–757.
- Kooiman, P., Willenborg, L.C.R.J. & Gouwewouw, J.M. (1997). PRAM: A Method for Disclosure Limitation of Microdata. Research paper no. 9705. Voorburg/Heerlen: Statistics Netherlands.
- Kuba, J. & Skinner, C. (1997). Categorical Data Analysis and Misclassification. *Survey Measurement and Process Quality*, Eds. L. Lyberg *et al.* New York: Wiley.
- Kuk, A.Y.C. (1990). Asking Sensitive Questions Indirectly. *Biometrika*, **77**, 436–438.
- Lucy, L.B. (1974). An Iterative Technique for the Rectification of Observed Distributions. *Astron. J.*, **79**, 745–754.
- Magder, L.S. & Hughes, J.P. (1997). Logistic Regression When the Outcome Is Measured with Uncertainty. *Am. J. Epidemiol.*, **146**, 195–203.
- McLachlan, G.F. & Krishnan, T. (1997). *The EM Algorithm and Extensions*. New York: Wiley.
- Mood, A.M., Graybill, F.A. & Boes, D.C. (1985). *Introduction to the theory of Statistics*. Auckland: McGraw-Hill.
- Rosenberg, M.J. (1979). Multivariate Analysis by a Randomized Response Technique for Statistical Disclosure Control. Ph.D. Dissertation, University of Michigan.
- Rosenberg, M.J. (1980). Categorical Data Analysis by a Randomized Response Technique for Statistical Disclosure Control. *Proceedings of the Survey Research Section, American Statistical Association*.
- Rubin, D.B. (1976). Inference and Missing Data. *Biometrika*, **63**, 581–592.
- Schafer J.L. (1997). *Analysis of Incomplete Multivariate Data*. London: Chapman and Hall.
- Schwartz, J.E. (1985). The Neglected Problem of Measurement Error in Categorical Data. *Soc. Meth. Research*, **13**, 435–466.
- Singh, J. (1976). A Note on RR Techniques. *Proc. ASA. Soc. Statist. Sec.*, 772.
- Van Gils, G., Van der Heijden, P.G.M. & Rosebeek, A. (2001). Onderzoek naar regelovertrekking. Resultaten ABW, WAO en WW. Amsterdam: NIPO. (In Dutch)
- Van den Hout, A.D.L. (1999). *The Analysis of Data Perturbed by PRAM*. Delft: Delft University Press.
- Van der Heijden, P.G.M., Van Gils, G., Bouts, J. & Hox, J.J. (2000). A Comparison of Randomized Response, Computer-Assisted Self-Interview, and Face-to-Face Direct Questioning. *Soc. Meth. Research*, **28**, 505–537.
- Warner, S.L. (1965). Randomized Response: a Survey Technique for Eliminating Answer Bias. *J. Am. Stat. Assoc.*, **60**, 63–69.
- Warner, S.L. (1971). The Linear Randomized Response Model. *J. Am. Stat. Assoc.*, **66**, 884–888.
- Willenborg, L. (2000). Optimality Models for PRAM. In *Proceedings in Computational Statistics*, Eds. J.G. Bethlehem and P.G.M. van der Heijden, pp. 505–510. Heidelberg: Physica-Verlag.
- Willenborg, L. & De Waal, T. (2001). *Elements of Statistical Disclosure Control*. New York: Springer.

## Résumé

Cet article traite de l'analyse de variables catégorielles, lorsqu'il y a des erreurs dans les affectations entre modalités et que les probabilités d'erreurs d'affectation sont connues. Les réponses randomisées constituent un des domaines où ce type d'erreurs existe. Des estimations des fréquences vraies sont données, et on procède à une discussion sur les ajustements aux rapports de chances (odds ratio). Les estimateurs du moment et du maximum de vraisemblance sont comparés, et on prouve qu'ils sont les mêmes à l'intérieur de l'espace des paramètres. Comme les estimateurs du moment sont régulièrement en dehors de l'espace des paramètres, on accorde une attention particulière à la possibilité de solutions limites. Un exemple est proposé.

## Appendix A

As stated in Section 3.2,  $V(\hat{\pi})$  can be partitioned as

$$V(\hat{\pi}) = \Sigma_1 + \Sigma_2, \quad (33)$$

where

$$\Sigma_1 = \frac{1}{n} (\text{Diag}(\pi) - \pi\pi')$$

and

$$\Sigma_2 = \frac{1}{n} P^{-1} (\text{Diag}(\lambda) - P \text{Diag}(\pi) P') (P^{-1})'. \quad (34)$$

To understand (33):

$$\begin{aligned} \Sigma_1 + \Sigma_2 &= \frac{1}{n} (\text{Diag}(\pi) - \pi\pi') + \frac{1}{n} P^{-1} (\text{Diag}(\lambda) - P \text{Diag}(\pi) P') (P^{-1})' \\ &= \frac{1}{n} (P^{-1} \text{Diag}(\lambda) (P^{-1})' - \pi\pi') \\ &= \frac{1}{n} P^{-1} (\text{Diag}(\lambda) - P \pi \pi' P') (P^{-1})' \\ &= \frac{1}{n} P^{-1} (\text{Diag}(\lambda) - \lambda \lambda') (P^{-1})' \\ &= V(\hat{\pi}) \end{aligned}$$

The variance due to PRAM as given in Kooiman *et al.* (1997) equals

$$\begin{aligned} V(\hat{T}|T) &= P^{-1} V(T^*|T) (P^{-1})' \\ &= P^{-1} \left( \sum_{j=1}^K T(j) V_j \right) (P^{-1})' \end{aligned} \quad (35)$$

where for  $j \in \{1, \dots, K\}$ ,  $T(j)$  is the true frequency of category  $j$ , and  $V_j$  is the  $K \times K$  covariance matrix of two observed categories  $h$  and  $i$  given the true category  $j$ :

$$V_j(h, i) = \begin{cases} p_{ij}(1 - p_{ij}) & \text{if } h = i \\ -p_{hj}p_{ij} & \text{if } h \neq i \end{cases}, \text{ for } h, i \in \{1, \dots, K\},$$

(Kooiman *et al.*, 1997).

In order to compare (34) with (35), we go from probabilities to frequencies in the RR data. This is no problem since we assume the RR data to be distributed multinomially. So we have  $V(\hat{T}|T) = \pi^2 V(\hat{\pi}|\pi)$  where, analogous to the PRAM data,  $T$  denotes the true frequencies.

In order to prove that  $\pi^2 \Sigma_2$  is the same as (35), it is sufficient to prove that

$$\begin{aligned} \sum_{j=1}^K T(j) V_j &= \text{Diag}(T^*) - P \text{Diag}(T) (P)' \\ &= \begin{pmatrix} \sum_j p_{1j} T(j) & 0 & \dots & 0 \\ 0 & \sum_j p_{2j} T(j) & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \dots & \dots & \sum_j p_{Kj} T(j) \end{pmatrix} - P \text{Diag}(T) (P)' \end{aligned}$$

which follows by writing out.

### Appendix B

As stated in Section 5, Lucy (1974) proves that in the interior of the parameter space the MLE of the true frequencies is equal to the moment estimate. In the following we have translated this proof to our setting and put in some explanatory steps.

The function we want to maximize for  $\pi \in (0, 1)^K$  under the constraint  $\sum_{j=1}^K \pi_j = 1$ , is

$$\log l^*(\pi) = \sum_{i=1}^K n_i^* \log \lambda_i + C \quad (36)$$

where  $n_i^*$  is given,  $\lambda_i = \sum_{k=1}^K p_{ik} \pi_k$ , for  $i \in \{1, \dots, K\}$ , and  $C$  is a constant.

We start by maximizing for  $\pi \in \mathbb{R}^K$  and we look for the stationary points of the function

$$G(\pi, \mu) = \sum_{i=1}^K n_i^* \log \lambda_i - \mu \left( \sum_{j=1}^K \pi_j - 1 \right) \quad (37)$$

where  $\mu$  is the Lagrange multiplier. Setting the derivatives of  $G$  with respect to  $\pi_j$  and  $\mu$  equal to zero, we obtain

$$\frac{\partial}{\partial \pi_j} G(\pi, \mu) = \sum_{i=1}^K n_i^* \frac{p_{ij}}{\lambda_i} + \mu = 0 \quad (38)$$

and

$$\frac{\partial}{\partial \mu} G(\pi, \mu) = \sum_{j=1}^K \pi_j - 1 = 0. \quad (39)$$

Multiplying (38) with  $\pi_j$  and summing over  $j$  yields

$$\sum_{j=1}^K \sum_{i=1}^K n_i^* \frac{p_{ij} \pi_j}{\lambda_i} = -\mu \sum_{j=1}^K \pi_j.$$

Using (39) we find that  $\mu = -\sum_{j=1}^K n_i^* \pi_j = -n$ . With this result it follows that the equality in (38) holds if  $\pi_j$  for  $j \in \{1, \dots, K\}$  is such that

$$\sum_{i=1}^K \hat{\lambda}_i \frac{p_{ij}}{\lambda_i} = 1,$$

where  $\hat{\lambda}_i = n_i^*/n$  for  $i \in \{1, \dots, K\}$ . Since the transition matrix  $P$  has the property that  $\sum_{i=1}^K p_{ij} = 1$ , the equality in (38) holds if  $\pi_j$ , for  $j \in \{1, \dots, K\}$ , is such that  $\hat{\lambda}_i = \lambda_i$  for  $i \in \{1, \dots, K\}$ .

In other words, a stationary point of (37) is found for such  $\pi$  that

$$\hat{\lambda} = P\pi.$$

To conclude, when (36) has one maximum under the constraint  $\sum_{j=1}^K \pi_j = 1$ , this maximum is attained at the moment estimator  $\hat{\pi} = P^{-1}\hat{\lambda}$ .

When we include the constraint  $\pi \in (0, 1)^K$  and we maximize under this extra constraint, the MLE is not equal to the moment estimate when the moment estimate is outside the parameter space  $(0, 1)^K$ .

[Received March, 2001, accepted February, 2002]