The Analysis of Data Perturbed by PRAM

A.D.L. van den Hout

November 1999

Contents

Acknowledgements

This report is written under the authority of Statistics Netherlands (CBS). The research took place at the CBS at the Department of Statistical Methods and was part of a two-year post-Master's program at the University of Delft (TU Delft). The report was presented and defended at 30 November 1999 at the TU Delft in front of a committee.

The committee members:

- prof dr R.M. Cooke, TU Delft
- prof dr ir P. Kooiman, CBS, Voorburg
- dr L.C.R.J. Willenborg, CBS, Voorburg
- dr C. Kraaikamp, TU Delft
- dr H.P. Lopuhaä, TU Delft

I would like to thank Cor Kraaikamp, Leon Willenborg, Peter Kooiman and Peter-Paul De Wolf for their comments and suggestions.

Ardo van den Hout
Voorburg, November 1999

Chapter 1

General Introduction

1.1 Introduction

The Post RAndomisation Method (PRAM) was introduced in Kooiman et al. (1997) as a method for disclosure protection of microdata files (see also Gouweleeuw et al., 1998). A microdata file consists of records and each record contains individual data of respondents. The PRAM procedure yields a new microdata file in which the scores on certain variables in the original file may be changed into different scores according to a prescribed probability mechanism. 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. The recipient of the perturbed data is informed about the probability mechanism which is used in order that he can adjust his statistical analysis and take into account the extra uncertainty caused by applying PRAM.

This report explains PRAM and discusses how statistical analysis can be adjusted when variables are perturbed by PRAM. Because PRAM always concerns categorical variables - variables with a finite number of values - we discuss mainly categorical data analysis. We do *not* discuss the extent of randomness which the PRAM procedure needs to protect the data satisfactory. This randomness, that is, the transition probabilities that scores on certain variables change into different scores, should be determined before the microdata file is perturbed, see Willenborg (1999). In this report we assume that the randomness of the PRAM procedure is known. So the main problem in this report is the problem of the recipient of the data: given the perturbed file and the information on the extent of the perturbation, how can standard statistical analysis be adjusted?

PRAM is not yet applied in practice but research in the applicability of the

method is in progress. We hope that this report contributes to the understanding of the possibilities and the problems involved.

Although the PRAM method is fairly new and, as far as we know, there is no literature concerning adjustment of data perturbed by PRAM, the situation bears close resemblance to analysis of data subject to misclassification, analysis of data obtained by using randomized response and analysis of incomplete data. Methods used in these situations provide the main tools of the methods presented in this report.

This report is not only a theoretical research into analyses of data perturbed by PRAM, it is also intended as a guidebook to people who actually have to work with these data. Therefore we have taken some effort to explain statements and provide the reader with examples and references. The much used scheme is to start with an easy case and extend it to the general case. Not every discussed analysis is considered in its most general form though, but we think that concerning the analyses discussed and the techniques used, the first and most important steps are made.

The report is organized as follows. The remaining of this chapter explains the need for statistical disclosure control of microdata files and describes current techniques of disclosure control in particular as used at Statistics Netherlands (CBS). Furthermore, it introduces PRAM and discusses an estimator of the original frequencies in the microdata file perturbed by PRAM. This estimator is called the *moment estimator*.

In Chapter 2 we compare the situation in which data are perturbed by PRAM with data subject to misclassification, data obtained by randomized response and incomplete data. The similarity between these situations: the scores which are missing or the scores which are only known via perturbed scores can be considered as stochastic variables.

Chapter 3 describes a special way to apply PRAM, invariant PRAM, and discusses an alternative to the way the analyst is provided with data perturbed by PRAM. In this discussion we briefly consider disclosure risk when data are perturbed by PRAM.

In the remaining chapters the discussion concerns the adjustment of statistical analysis when variables are perturbed by PRAM. Chapter 4 discusses basic analysis of contingency tables. Quantities as the difference of proportions, relative risk and the odds ratio are considered when data are perturbed by PRAM.

Chapter 5 is introductory to the EM algorithm which is used in the remaining chapters. An alternative to the moment estimator is given, we called it the *EM estimator*. The two estimators are compared and an important conjecture is

formulated.

Chapter 6 and 7 introduce two analyses, loglinear analysis and logistic regression respectively, and discuss how these analyses can be adjusted to take into account the perturbation caused by applying PRAM.

In the conclusion of the report we summarize the problems concerning analysis of data perturbed by PRAM and we give recommendations for future research.

1.2 The Need for Disclosure Control

In recent years there is a growing demand in society to have access to microdata files which are collected at the CBS. The files which are passed on to analysts should be protected against disclosure of identities of respondents.

Imagine you are the mayor of a big city in a small country and you are asked as a citizen to participate in a survey. Would you cooperate? Of course it is guaranteed that your name and address are not mentioned in the file with your answers, but is that enough to protect your identity? If the city is mentioned together with your profession, your identity is clear to whoever sees the file and is familiar with the situation in your country. And even when the city is not mentioned, your identity is liable to disclosure: there are not many mayors in a small country. Your gender, for example, can in that case be a clue to your identity.

What to do? Of course you can decide not to cooperate. An alternative is to take measures to prevent disclosure of your identity. For instance, you can report that you work for the government, instead of reporting that you are the mayor. It is also possible not to mention your profession at all and skip the related question.

In general, these measures are not necessary. The privacy of the respondents is well protected. CBS protects the privacy of respondents, which can be individuals as well as companies, by what is called *statistical disclosure control*. Disclosure control applies not only to microdata, but also applies for instance to tables in which data of respondents are processed. When a company is large, it is for instance possible that information concerning the financial state of affairs of the company can be deduce from unprotected tables with annual turnovers.

The example with the mayor shows that there is more to disclosure control than simply leaving out name and address.

1.3 Current Techniques of Disclosure Control

In order to protect the identity of respondents in a microdata file several measures can be issued. I explain two of them which are currently in use at CBS and postpone the third, PRAM which is not yet applied, to the next section.

When the concerning variable is categorical, the identity of the respondent can be protected by *global recoding*. By this we mean that new categories are made which include more respondents that the categories in the original file. This can be necessary when a category in the original file contains just a few respondents. For example, when we have a microdata file in which 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.

Another way to protect the identity of the mayor is to leave out his profession. This procedure is called *local suppression*. When data are locally suppressed it should always be checked whether there are other combinations which become rare and can therefore lead to other disclosures. Locally suppressing scores in a microdata file yields incomplete data.

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

More information on statistical disclosure control in the Netherlands and abroad can be found in a special issue of Netherlands Official Statistics (1999) which is devoted to current disclosure control practices as well as research into the assessment of the disclosure risk.

1.4 PRAM, Basic Ideas

This section describes PRAM, gives an idea how it can be applied as a method for statistical disclosure control and introduces the notation used throughout the report.

PRAM is a method for disclosure protection of microdata files. A microdata file consists of records at the respondent level. Each record represents one respondent and consists of scores on characteristic variables for that respondent.

The main idea of PRAM as a method for disclosure protection of microdata files is that PRAM produces a microdata file in which the scores on some variables for certain records in the original file are changed into a different score according to a prescribed Markovian probability mechanism. 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.

Note, that as soon as we want to use a file for statistical analysis and we treat the respondents as a sample from a larger population, we have to take into account the sample model. So, in the context of statistical analysis, applying PRAM means introducing a second stochastic element when the original file is a sample from a larger population. (The first stochastic element is the stochastic nature of the sample model). It is important to realize that the perturbation issued by applying PRAM is independent of the chosen sample model.

Let A, B, C, ... be categorical variables in the original file to which PRAM is applied and let A^*, B^*, C^*, ... be the corresponding variables in the perturbed file. Suppose A has K categories with scores $1, ..., K$. PRAM is applied using transition probabilities: let $p_{kl} = I\!P(A^* = l|A = k)$ denote the probability that $A^* = l$ given that the original score A equals k, for all $k, l = 1, ..., K$. The transition probabilities are aggregated in a matrix. P_A denotes the $K \times K$ matrix that has p_{kl} as its (k, l)-th entry. $P_A = (p_{kl})$ is called a *PRAM matrix*. Note that a PRAM matrix is a Markov matrix, i.e., each row sums up to 1.

The PRAM matrix plays a central role when data are perturbed by PRAM. First, the entries of the matrix determine the extent of the perturbation caused by applying PRAM. Secondly, the PRAM matrix is needed when statistical analyses on the perturbed data have to be corrected to take into account the perturbation. As is mentioned before, in this report we do not discuss how the values of p_{kl} should be chosen, instead this report considers the problem of the recipient of the data: given the perturbed file and the PRAM matrix, how can standard statistical analysis be adjusted?

In this report the assumption is made that P_A is invertible, although this is strictly speaking not necessary for applying PRAM. However, it turns out that P_A^{-1} comes in handy when we want to estimate the frequency distribution of A in the original file.

Applying PRAM to a microdata file means that per record the score on A is liable to change in a different score. The scores per record which are the result of this randomness are the scores on variable A^* in the perturbed file. In this report we describe this procedure as though A has been perturbed to A^*.

Let T_A be the $K \times 1$ table (or vector) with the frequencies of the K possible scores

on the variable A in the original file. (We do not make a distinction in notation according to whether the original file is a sample or not.) $T_A(k)$ denotes the number of records in the original file with score $A = k$. Let T_{A^*} be the $K \times 1$ table with the frequencies in the perturbed file.

Given T_A, we can estimate $I\!P(A = k)$ if the sampling design is known. In general this is a complicated formula, unless the sampling design is self-weighting, in which case $I\!P(A = k)$ is estimated by $T_A(k)/n$, where n is the number of respondents. In this report we will assume that $I\!P(A = k)$ can be estimated by $T_A(k)/n$. To estimate $I\!P(A^* = l)$ we use

$$\begin{aligned} I\!P(A^* = l) &= I\!P(A^* = l, A = 1) + ... + I\!P(A^* = l, A = K) \\ &= I\!P(A^* = l|A = 1)I\!P(A = 1) + ... + I\!P(A^* = l|A = K)I\!P(A = K) \\ &= p_{1l}I\!P(A = 1) + ... + p_{Kl}I\!P(A = K). \end{aligned} \tag{1.1}$$

In other words, given T_A, the distribution of $T_{A^*}(l)$ $(= nI\!P(A^* = l))$ is a sum of K independent binomial distributions: $X_1 \sim \text{Bin}(T_A(1), p_{1l})$, $X_2 \sim \text{Bin}(T_A(2), p_{2l})$, ..., $X_K \sim \text{Bin}(T_A(K), p_{Kl})$

We give an example. Suppose that the variable A is gender, with scores 1 = male and 2 = female and suppose PRAM is applied using

$$P_A = \begin{pmatrix} 9/10 & 1/10 \\ 2/10 & 8/10 \end{pmatrix}. \tag{1.2}$$

When the original file contains 100 individuals, 99 male and 1 female, the perturbed file will also contain 100 individuals. However, 10.7 of these individuals are expected to be female. The idea is that when it is known that the original file contains 1 female, the probability that the identity of this female is disclosed in the perturbed file should be small. In other words, given $A_i^* = 2$, the probability that female i in the perturbed file is *the* female of the original file should be small. This is the case. One has

$$I\!P(A_i = 2|A_i^* = 2) = \frac{I\!P(A_i^* = 2|A_i = 2)I\!P(A_i = 2)}{I\!P(A_i^* = 2)}, \tag{1.3}$$

and (1.3) can be estimated using P_A and the assumption that $T_A(k)/n$ is an estimate of $I\!P(A = k)$. This yields

$$\frac{8/10 \cdot 1/100}{1/10 \cdot 99/100 + 8/10 \cdot 1/100} \simeq 0.075.$$

So far, we discussed applying PRAM to the values of one variable. Of course it is possible to apply PRAM independently to different variables by using the method sequentially.

Three types of independence can be distinguished regarding PRAM. Applying PRAM to A is called *nondifferential with respect to* B if for all l:

$$I\!P(A^* = l|\ A = k,\ B = s) = I\!P(A^* = l|\ A = k).$$

Inconsistencies can occur in a file when PRAM is applied non-differentially. When, for instance, being in the possession of a driving licence or not being in the possession of a driving licence is perturbed by PRAM non-differentially with respect to age, it is possible that in the perturbed file a ten years old has a driving licence. Inconsistency in a record can be a clue to disclosure, because in that case it is obvious that the record has been affected by PRAM.

The term independent is used when PRAM is applied to more than one variable and is as usual. Applying PRAM to A is *independent* of applying PRAM to B when

$$I\!P(A^* = l,\ B^* = t|\ A = k,\ B_i = s) =$$
$$I\!P(A^* = l|\ A = k,\ B_i = s)I\!P(B^* = t|\ A = k,\ B = s)$$

The third type of independence is *independence* of the perturbation *with respect to different records*: applying PRAM to several records is independent when the perturbation of the values in record j is not influenced by the values in record i $(i \neq j)$. For instance, when there is an upper limit to the amount of records that can be perturbed, the perturbation is not independent with respect to the records. In this report this type of independence will not play a role.

Besides applying PRAM to more than one variable sequentially, it is possible to apply PRAM to more than one variable simultaneously. When we want to apply PRAM to A and B with categories $1, ..., K$ and $1, ..., S$ respectively, we define

$$p_{kl,st} = I\!P(A^* = l, B^* = t|A = k, B = s).$$

Applying PRAM now means that given record i with $A_i = k$ and $B_i = s$, the values $A_i^* = l$ and $B_i^* = s$, are determined using the transition probabilities $p_{kl,st}$.

When PRAM is applied to A and B independently and non-differentially with PRAM matrices $P_A = \left(p_{kl}^A\right)$ and $P_B = \left(p_{st}^B\right)$, the transition probabilities can be written as

$$\begin{aligned} p_{kl,st} &= I\!P(A^* = l, B^* = t|A = k, B = s) \\ &= I\!P(A^* = l|A = k)I\!P(B^* = t|B = s) \\ &= p_{kl}^A p_{st}^B \end{aligned}$$

Let T_{AB} be the cross-tabulation of variables A and B and let $vec(T_{AB})$ denote the $KS \times 1$ table of stacked columns of T_{AB}. Together A and B can be considered as one compounded variable with KS categories and frequencies given by $vec(T_{AB})$. We can define the PRAM matrix $P_{AB} \in \mathbb{R}^{KS}$ for applying PRAM to the compounded variable by:

$$P_{AB} = \begin{pmatrix} p_{11}^B P_A & p_{12}^B P_A & \cdots & p_{1L}^B P_A \\ \vdots & \ddots & \ddots & \vdots \\ p_{L1}^B P_A & \cdots & \cdots & p_{LL}^B P_A \end{pmatrix}, \tag{1.4}$$

where each $p_{st}^B P_A$ is itself a $K \times K$ matrix.

More theory and ideas regarding PRAM can be found in Kooiman et al. (1997), Gouweleeuw et al. (1998) and De Wolf et al. (1997).

1.5 The Moment Estimator

What is the effect on statistical analyses when PRAM is applied to microdata? Can standard analyses be adjusted to take into account the perturbation effected by applying PRAM? These questions are essential when PRAM is discussed as a possible method for disclosure control - indeed, they form the main subject of this report. Different analyses demand different corrections. For some analyses the correction is relatively easy. This section discusses the estimation of the original frequencies. In the remainder of this report this estimation will be useful when other statistical analyses have to be adjusted.

Let the notation be as in the previous section and let PRAM be applied to variable A with categories $1, ..., K$. Using the fact that $T_{A^*}(l)$ is a sum of K independent binomial distributions, it follows from (1.1) that

$$\mathbb{E}\left[T_{A^*}|T_A\right] = P_A^t T_A. \tag{1.5}$$

Thus, T_A can be unbiasedly estimated by

$$\widehat{T}_A = \left(P_A^{-1}\right)^t T_{A^*}. \tag{1.6}$$

We call the estimator (1.6) of the original frequencies the *moment estimator*. Note that P_A has to be invertible in order for (1.6) to be well defined. Note also that it is possible that the estimation yields negative frequencies.

The conditional covariance matrix of $\widehat{T}_A$ given T_A is

$$V\left(\widehat{T}_A|T_A\right) = \left(P_A^{-1}\right)^t V\left(T_{A^*}|T_A\right) P_A^{-1}.$$

(Kooiman et al., 1997). To deduce $V\left(T_{A^*}|T_A\right)$, we consider first the case of a 3×1-table, afterwards we give the (co)variances in the general case.

When A has three categories, $T_{A^*}(l)$ is a sum of three independent binomial variables: $X_l \sim \text{Bin}(T_A(1), p_{1l})$, $Y_l \sim \text{Bin}(T_A(2), p_{2l})$ and $Z_l \sim \text{Bin}(T_A(3), p_{3l})$, for $l = 1, 2, 3$. Because of the independence we can sum the variances:

$$V\left(T_{A^*}(l)|T_A\right) = T_A(1)p_{1l}(1-p_{1l}) + T_A(2)p_{2l}(1-p_{2l}) + T_A(3)p_{3l}(1-p_{3l}).$$

Regarding the covariances we note that for $l \neq j$

$$\begin{aligned} C\left(T_{A^*}(l), T_{A^*}(j)|T_A\right) &= C\left(X_l + Y_l + Z_l, X_j + Y_j + Z_j|T_A\right) \\ &= C\left(X_l, X_j|T_A\right) + C\left(X_l, Y_j|T_A\right) + ... + C\left(Z_l, Z_j|T_A\right) \\ &= C\left(X_l, X_j|T_A\right) + C\left(Y_l, Y_j|T_A\right) + C\left(Z_l, Z_j|T_A\right) \\ &= -T_A(1)p_{1l}p_{1j} - T_A(2)p_{2l}p_{2j} - T_A(3)p_{3l}p_{3j}. \end{aligned}$$

Since, because of the independence,

$$\begin{aligned} C\left(X_l, Y_j|T_A\right) = C\left(X_l, Z_j|T_A\right) &= C\left(Y_l, X_j|T_A\right) = C\left(Y_l, Z_j|T_A\right) \\ = C\left(Z_l, X_j|T_A\right) &= C\left(Z_l, Y_j|T_A\right) = 0. \end{aligned}$$

An equivalency as $C\left(X_l, X_j|T_A\right) = -T_A(1)p_{1l}p_{1j}$ is a property of the multinomial distribution and can be found for instance in Johnson and Kotz (1969), Chapter 11, or Agresti (1990), Chapter 12.

In the same way, the general case is given by

$$V\left(T_{A^*}|T_A\right) = \sum_{k=1}^{K} T_A(k)V_k,$$

where, for $k = 1, ..., K$, V_k is the $K \times K$ covariance matrix of the outcomes $l, j = 1, ..., K$ of the transition process of an element with true score k:

$$V_k(l,j) = \begin{cases} p_{kl}(1-p_{kl}) & \text{if} \quad l = j \\ \\ -p_{kl}p_{kj} & \text{if} \quad l \neq j \end{cases}, \text{ for } l, j = 1, ..., K$$

To conclude, the derivations in this section show that when PRAM is applied, perturbed frequencies can be corrected when the PRAM matrix is known. Using the PRAM matrix, it is also possible to estimate variances and covariances of the estimated true frequencies. The moment estimator to estimate the original frequencies is easy to understand and easy to implement. The possibility that this estimator can yield negative estimated cell frequencies is a disadvantage because it conflicts with standard analysis of a frequency table.

Chapter 2

Perturbations Similar to PRAM

2.1 Introduction

Although PRAM is a new technique for statistical disclosure control, the problems the analyst encounters when he wants to apply statistical analysis to data perturbed by PRAM are not completely new. We can understand applying PRAM as misclassification on purpose and investigate statistical methods which work with misclassified data. Also, data perturbed by PRAM bears close resemblance to data provided by randomized response and results of research in this field should also be applicable. Furthermore, there is extensive literature on incomplete data and although data perturbed by PRAM is not incomplete in the sense that some records do not have scores on a variable, the similarity is that we can consider scores which are missing and scores which are only known via perturbed scores as stochastic variables. This chapter goes to some length into the use of literature on misclassification, randomized response and incomplete data. We make the effort because we think that it can be useful to further research on the analysis of data perturbed by PRAM. Fact is that it was the basis of the results in this report. We did not make an extensive and systematic research of literature and just want to mention what came along.

2.2 Misclassification

Misclassification is an easy concept. When a respondent is classified, there is a certain probability that his score is classified as l, while it is in fact k. It has long been recognized that categorical variables are often subject to misclassification and that this can distort the results of statistical analysis. In Kuha and Skinner (1997)

an example is given: misclassification in the 1991 population census of England and Wales. Following this census, a census validation survey was carried out. In contrast to the much larger census, in which data were collected by self-completion forms, the validation survey employed experienced interviewers who used more detailed questionnaires which were designed to probe for the most accurate answer. On the basis of the comparison between the census and the more reliable validation survey, the following misclassification matrix was estimated regarding ethnic group.

		Census	
		Caucasian	Other
Validation Survey	Caucasian	0.9986	0.0014
	Other	0.0566	0.9434

Statistical analysis of misclassified data which explicitly works with a misclassification matrix, will work fine when the data are perturbed by PRAM and the PRAM matrix is taken to be the misclassification matrix. An important advantage in the case of PRAM is that the PRAM matrix is known while in the case of misclassification the matrix always has to be estimated. Another difference is that misclassification is often differential with respect to subgroups, i.e., transition probabilities can differ per subgroup, while applying PRAM is, in general, non-differential (see section 1.4).

Often it is the case that analysis of misclassified data does not use a misclassification matrix explicitly, but incorporates the validation survey in the analysis of the misclassified data. In that case there is no parallel to the situation where data are perturbed by PRAM. A validation survey as a supplement to data perturbed by PRAM is not possible.

Work in which the analysis of misclassified data makes explicit use of the misclassification matrix are the following. Kuha and Skinner (1997) discuss bivariate analysis such as difference of proportion, relative risk and tests of association. As to multivariate analysis they consider the moment estimator and the EM algorithm as a tool for loglinear analysis. Chen (1979, 1989) uses a EM algorithm to apply loglinear analysis. His form of the EM algorithm is quite obscure and this report favours the EM algorithm as described in Kuha and Skinner (1997) because it is easier to combine with standard loglinear analysis. Assakul and Proctor (1967) discuss testing independence in two-way tables. Copeland et al. (1977) present 'an empirical description' of the extent and direction of the bias in situations with misclassification and give a explicit adjustment of the relative risk. Goldberg (1975) discusses the effects of misclassification on difference of proportions and on the relative odds. Greenland (1980,1988) considers effects in the odds ratio and gives

estimates of the variance of the odds ratio. Magder and Hughes (1997) introduce the EM algorithm for the logistic regression model when the dependent variable is subject to misclassification.

2.3 Randomized Response

The similarity between PRAM and randomized response is that in both procedures data are perturbed on purpose. Randomized response is an interview technique which can be used when sensitive questions have to be asked. We explain a simple form of the method (Warner, 1965) with an example.

Let the sensitive question be 'Have you ever used illegal drugs?' When interviewing, have the respondent roll a dice. If the result is 1,2,3 or 4 ask him to answer question Q_S, if the result is 5 or 6 ask to answer Q_S^c, where

$$Q_S = \text{'Have you ever used illegal drugs?'}$$
$$Q_S^c = \text{'Have you never used illegal drugs?'}$$

The interviewer does not know the outcome of the dice, he only receives a 'yes' or a 'no'. The respondent answers Q_S with probability $\theta = 2/3$ and answers Q_S^c with probability $1-\theta$. Let p_y be the proportion in the population for which the true response to Q_S is yes. Then the proportion of yes-responses should be given by $\theta p_y + (1-\theta)(1-p_y)$. So, when the number of all yes-responses in a sample of size n_0 is Y, we estimate p_y by

$$\widehat{p}_y = \frac{Y/n_0 - (1-\theta)}{2\theta - 1}.$$

The analogy with PRAM is as follows. Let A be a binary variable which is perturbed using PRAM matrix

$$P_A = \begin{pmatrix} \theta & 1-\theta \\ 1-\theta & \theta \end{pmatrix}.$$

So, $I\!P(A^* = 0) = \theta I\!P(A = 0) + (1-\theta) I\!P(A = 1)$. Let $n_1 = T_{A^*}(0) + T_{A^*}(1)$. Given T_{A^*} we estimate $p_A = I\!P(A = 0)$ with the moment estimator as

$$\widehat{p}_A = \frac{T_{A^*}(0)/n_1 - (1-\theta)}{2\theta - 1}.$$

This example illustrates the resemblance between PRAM and randomized response. There are all sorts of randomized response techniques, but the principle is

always the same: perturbation with a known probability mechanism. There is also a difference between the two procedures. When using randomized response the probability mechanism is determined *before* the data are collected, but the PRAM 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 randomized response setting.

Literature on randomized response is used in this report. Maddala (1983) discusses a method to adjust the likelihood function in the logistic regression model with a dependent variable which is perturbed. Van der Heijden et al. (1998) uses this method. Chaudhuri and Mukerjee (1988) have written a monograph about randomized response and consider in Appendix A1.3 the maximum likelihood in the Warner model. Bourke and Moran (1984, 1988) describe how the EM algorithm can be applied in several randomized response procedures. In Bourke and Moran (1988) randomized response is discussed in a form that is close to the PRAM procedure; in this article the authors work with a matrix that contains the information concerning the perturbation. In Bourke and Moran (1984) observations from randomized response procedures are viewed explicitly as incomplete data and the EM algorithm is used.

2.4 Incomplete Data

The link between PRAM and incomplete data problems is less obvious. In a data file perturbed by PRAM no data are missing. But when incomplete data are analysed, this can be done by considering missing values as random variables and using the known values to estimate these missing values. In the same way: when data perturbed by PRAM are analysed it is possible to estimate the original values using the perturbed values. So, when we are confronted with data perturbed by PRAM and we want to use incomplete data methods, we can treat each value in the perturbed file as a random variable value which has to be estimated. Also, we can consider a data file twice the size of the original file where half of the data is missing (the original data) and the other half is observed (the perturbed data).

The literature on incomplete data that we use concerns the EM algorithm. The main article is Dempster et al. (1977), but the monograph by McLachlan and Krishnan (1997) on this subject is easier to read and more up to date.

The EM algorithm plays an important role in this report. An alternative to the moment estimator is presented in which the algorithm is used. Also, the algorithm is applied in the logistic regression model, which is nonlinear. We conjecture that the

EM algorithm can be used as a general tool for problems regarding statistical analyses of data perturbed by PRAM. This is the reason we discuss the EM algorithm at length in Chapter 5.

2.5 Conclusion

To summarize, although perturbation of categorical variables by applying PRAM is new and produces new problems for standard statistical analysis, it is possible that solutions to these problems can be found in existing methods which deal with similar perturbation problems. The similarity between these situations: the scores which are missing or the scores which are only known via perturbed scores can be considered as values of stochastic variables which have to be estimated. An important advantage in the case of PRAM is that the probability mechanism used is known, which simplifies these methods.

Of course, it can also be considered the other way round: results regarding the analysis of data perturbed by PRAM can be of use in situations with incomplete data, misclassified data or data provided by randomized response.

Chapter 3

Invariant PRAM

3.1 Introduction

This chapter discusses a suggestion which was made at Statistic Netherlands and which is partly stated in De Wolf et al. (1997, section 6). The suggestion concerns the availability of the PRAM matrix to the recipient of the perturbed data and the idea is to apply PRAM twice: first, the original scores on variable A are perturbed using PRAM matrix P_A and, secondly, PRAM is applied several times 'backwards' to the perturbed scores using a matrix $\overleftarrow{P}_A$. This matrix $\overleftarrow{P}_A$ should be chosen in such a way that the twice perturbed scores are 'close' to the original scores in order that these perturbed scores can be used directly as an estimate of the original scores. The idea for applying PRAM *several times* backwards is to be able to withhold the PRAM matrices because the analyst can estimate variances using several realizations of the perturbed scores. The main motivation behind this procedure is that not providing the PRAM matrix is beneficial to statistical disclosure control. Furthermore, the recipient can use standard analyses on the twice perturbed data.

To investigate the suggestion of applying PRAM twice, we consider in the next two sections two PRAM procedures introduced in Kooiman et al. (1997) and De Wolf et al. (1997) respectively: *invariant PRAM* and *two-stage PRAM*. The applicability of these procedures is not restricted to the investigated suggestion though; viewed apart they are interesting variations on standard PRAM. The fourth section recommends to use invariant PRAM to protect tabular data and can be viewed as an excursion outside the subject of this report. In the last section of this chapter we pay some attention to disclosure control in the PRAM situation and discuss the usefulness of applying PRAM twice.

3.2 Invariant PRAM

So far, non-singularity is the only restriction regarding the Markov matrix P_A which is used to apply PRAM to variable A. This section demonstrates that the analysis of the perturbed data can be simplified if a special choice is made for P_A.

When P_A can be chosen in such a way that

$$P_A^t T_A = T_A, \tag{3.1}$$

we have (see also section 1.4):

$$I\!E\left[T_{A^*}|T_A\right] = P_A^t T_A = T_A,$$

which means that given T_{A^*}, the original table T_A can be estimated unbiasedly by

$$\widehat{T}_A = T_{A^*}.$$

Of course this simplifies the analysis, since no pre-multiplication by a matrix is needed to estimate the original table.

A non-trivial solution for a Markov matrix P_A satisfying (3.1) is given in Kooiman et al. (1997) and in Gouweleeuw et al. (1998): assume without loss of generality that $T_A(k) \geq T_A(K) > 0$, for categories $k = 1, ..., K$, and let, for some $0 < \theta < 1$

$$p_{kl} = \begin{cases} 1 - (\theta T_A(K)/T_A(k)) & \text{if} \quad l = k \\ \\ \theta T_A(K)/\left((K-1)T_A(k)\right) & \text{if} \quad l \neq k \end{cases}. \tag{3.2}$$

When $P_A = (p_{kl})$, P_A is a Markov matrix satisfying (3.1).

To give an example: let T_A be $(75, 25, 50)^t$. When P_A is computed using (3.2), it is given by

$$P_A = \begin{pmatrix} 1-\theta/3 & \theta/6 & \theta/6 \\ \theta/2 & 1-\theta & \theta/2 \\ \theta/4 & \theta/4 & 1-\theta/2 \end{pmatrix}.$$

With a choice for $\theta \in (0, 1)$ the PRAM matrix can be fixed, i.e., the parameter θ can be use to fix the transition probabilities and fine-tune the extent of randomness.

Note that a transformation satisfying (3.1) is invariant with respect to the original frequencies of A. The method which perturbs variable A using this special choice for P_A, is therefore called *invariant PRAM*. Note also that P_A does not need to have a dominant diagonal, which means that the matrix can be singular. This is not an immediate problem since T_A is estimated by T_{A^*} when invariant PRAM is applied.

The invariance of invariant PRAM does not entail that the transformation is invariant with respect to table-crossings of A with other variables in the microdata file. Consider the case with two variables A and B which are cross-tabulated in table T_{AB}. Variable A has K categories and variable B has J categories. Let $vec(T_{AB})$ denote the $KJ \times 1$ table of stacked columns of T_{AB}.

When PRAM must be applied to both A and B, then together A and B can be considered as one compounded variable with KJ categories. This is essentially the same situation as in the case where only one variable has to be perturbed: invariant PRAM can be applied as explained above.

The situation is different when B has to remain unperturbed. Kooiman et al. (1997) present a method to apply invariant PRAM in this case. The two restrictions regarding the invariant matrix P_{AB} are

$$P_{AB}^t vec(T_{AB}) = vec(T_{AB}) \tag{3.3}$$

and

$$\sum_{k=1}^{K} T_{A^*B}(k,j) = \sum_{k=1}^{K} T_{AB}(k,j), \text{ for } j = 1, ..., J. \tag{3.4}$$

The first condition implies that $I\!E\,[T_{A^*B}|T_{AB}] = T_{AB}$ holds, the second implies that the variable B is not perturbed.

Now, let the $KJ \times KJ$ matrix P_{AB} be given by the following block matrix

$$P_{AB} = \begin{pmatrix} P_1 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & P_J \end{pmatrix},$$

where P_j is a $K \times K$ block, for $j = 1, \ldots, J$. Since $vec(T_{AB})$ is the table of stacked columns of T_{AB}, condition (3.3) can be written as

$$P_j^t T_{AB,j} = T_{AB,j} \text{ , for } j = 1, ..., J \tag{3.5}$$

where $T_{AB,j}$ is the jth column of T_{AB}. For each j, equation (3.5) is an univariate invariant PRAM condition. So, solutions P_j for $j = 1, ..., J$ to equation (3.5) can be found by using (3.2). Also note that by choosing this structure for P_{AB}, condition (3.4) is satisfied since transition from category (k, s) to (l, t) can only occur if $s = t$, hence B is not perturbed.

Note that for example in the case that A and B are binary variables, it is very likely that

$$I\!P\Big((A^*, B) = (l, 0)|(A, B) = (k, 0)\Big) \neq I\!P\Big((A^*, B) = (l, 1)|(A, B) = (k, 1)\Big).$$

This is because the two probabilities are determined on the basis of the first column and the second column of T_{AB}, respectively. This is interesting because it means that perturbing A by applying PRAM in this way is not nondifferential with respect to B (see section 1.4).

The advantage of invariant PRAM is obvious: the estimation of the original T_A is easy. Furthermore, applying invariant PRAM means that there will never be negative estimated frequencies.

Drawbacks of invariant PRAM are less freedom in the choice of P_A and the impracticable size of P_A when A is a compounded variable and there are a lot of cross-tabulations to be preserved. Finally it should be noted that an analyst always needs extra information to determine the extra variance due to applying PRAM.

3.3 Two-Stage PRAM

As stated in the introduction of this chapter, we are looking for a way to apply PRAM twice. When the original scores on variable A are perturbed using PRAM matrix P_A, we want to apply PRAM several times 'backwards' to the perturbed scores using matrix $\overleftarrow{P}_A$. Matrix $\overleftarrow{P}_A$ should be chosen in such a way that the twice perturbed scores are 'close' to the original scores.

In De Wolf et al. (1997) a procedure called *two-stage PRAM* is presented. The idea of applying two-stage PRAM to a variable A is to apply PRAM twice. The second PRAM matrix, $\overleftarrow{P}_A$, is determined using the probabilities that the original value of A is k, given that A^* is l. $\overleftarrow{P}_A$ is used to apply PRAM backwards and perturb variable A^*.

$\overleftarrow{P}_A$ is constructed as follows. Note that

$$I\!P(A = k|A^* = l) = \frac{I\!P(A^* = l|A = k)I\!P(A = k)}{I\!P(A^* = l)}.$$

These probabilities can be estimated by

$$\overleftarrow{p}_{lk} = \frac{p_{kl}T_A(k)}{\sum_j p_{jl}T_A(j)}.$$

Let $\overleftarrow{P}_A$ be the matrix that has $\overleftarrow{p}_{lk}$ as its (l,k)-th entry. Matrix $\overleftarrow{P}_A$ is a Markov matrix and can therefore be used to apply PRAM to the perturbed file. When A^* is perturbed to A^{**}, we have

$$I\!E\left[T_{A^{**}}|T_A\right] = \overleftarrow{P}_A^t \left(P_A^t T_A\right) = \left(P_A \overleftarrow{P}_A\right)^t T_A = T_A. \tag{3.6}$$

That $\left(P_A \overleftarrow{P}_A\right)^t T_A = T_A$ is not trivial, but follows from direct calculations.

So far, we did not define what we meant by 'twice perturbed scores 'close' to the original scores', but now, inspired by equation (3.6) we consider twice perturbed scores on A close to the original scores when $I\!E\left[T_{A^{**}}|T_A\right] = T_A$ holds. When this is the case, $T_{A^{**}}$ can be used as an unbiased estimate of T_A.

Let $R = P_A \overleftarrow{P}_A$, then R is an invariant matrix due to (3.6). If R was used to perturb the original file, we would have obtained the same result as when we applied PRAM twice using P_A and $\overleftarrow{P}_A$ respectively. So, starting with an arbitrary matrix P_A, if we apply $\overleftarrow{P}_A$ to the perturbed file, we have in fact applied invariant PRAM.

Considering the application of PRAM in a software package for statistical disclosure control, two-stage PRAM is interesting. The data protector can choose any invertible Markov matrix to start with, and the package will compute the corresponding invariant matrix.

There are still issues to be investigated when two-stage PRAM is discussed as a possible way to apply PRAM. For instance, when P_A is provided, the effect of the resulting invariant matrix R on the disclosure risk is not clear.

3.4 Using PRAM to Protect Tabular Data

So far, and so it will be in the rest of the report, we focussed on statistical disclosure control of microdata. However, is also possible to use invariant PRAM to protect tables. There exist several techniques to avoid disclosure in tabular data: cell suppression, rounding of values and data perturbation, see for example Willenborg and De Waal (1996). When tables are protected by processing the tables themselves, there can be problems regarding disclosure control when more than one table is constructed using a single microdata file. Two tables of a single microdata file which are safe when viewed apart, can still disclose information when they are combined. Of course, at the CBS this problem is known and measures are taken.

Another technique concerning this problem is to process the microdata which provide the material for the tables. The advantage of this idea is that as soon as the microdata file is safe, all the tables constructed using that file are safe, see for

reference Zayatz e.a. in NOS (1999). This is the approach we have in mind when using PRAM as a technique to protect tables against disclosure.

When invariant PRAM can be implemented easily using the idea of two-stage PRAM, it is possible that a specific and protected cross-tabulation is provided by the institute. By which we mean that, when an analyst asks for a cross-tabulation, the institute provides a tabulation which is protected by applying invariant PRAM. Instead of giving the used PRAM matrix to the analyst, the institute can limit the information on the perturbation by giving only the extra variances per cell due to applying PRAM. Note that the institute protects in this way the table by perturbing the microdata file and that the choice of the PRAM matrices depends on the specific cross-tabulation.

The advantage of this procedure is that, given a microdata file with certain variables, the analyst can ask for *any* cross-tabulation he wants and, furthermore, receives information on the extra variance due to the statistical disclosure control.

Note, that when statistical disclosure control on microdata level is used to protected tabular data, the extent of the protection - in the case of PRAM: the extent of the randomness - should be determined with an eye to the tabular data level. In other words, there has to be a translation of disclosure control on the level of tabular data to disclosure control on the level of microdata.

3.5 Applying PRAM Several Times

Invariant PRAM and two-stage PRAM are introduced in this chapter to investigate a suggestion to apply PRAM twice in such a way that the twice perturbed scores are close to the original scores and can be used as a direct estimate of the original scores. The idea is that the first execution of PRAM perturbs the scores and that the second perturbation, which can be seen as perturbation 'backwards', is executed several times in order that the analyst has information on variances and can do without the two PRAM matrices.

The concept of two-stage PRAM provides us with a idea of closeness: when A^{**} is the twice perturbed variable A, then

$$I\!E\left[T_{A^{**}}|T_A\right] = T_A$$

is as close as we can get (apart from variances).

At first sight, with two-stage PRAM at hand, the suggestion seems promising because the analyst can use standard software to analyse the twice perturbed scores

and he can apply his analysis several times to the several realizations of the perturbation in order to gain insight in variances. But on second thought, it does not seem worthwhile to investigate this idea any further before the following problem regarding disclosure control is solved.

Two-stage PRAM is a special form of invariant PRAM and invariant PRAM is a special form of PRAM. When records are perturbed by *any* form of PRAM some scores are changed into other scores. When PRAM is applied once in the way it is assumed in this report, i.e., with a dominant diagonal in the PRAM matrix, most scores will not be changed. So, when PRAM is applied several times to the same scores, it will be easy to detect which scores are changed in the different realizations, since for each record there are several realizations and most of them will be the same. In other words, as De Wolf et al. (1997, section 6) put it: comparing the same record in all the provided data files and picking the score that is present in the majority of them, will - with a high probability - result in the true score. Hence, the disclosure limitation accomplished by PRAM is essentially eliminated in this way.

We will make this more exact. Let m be the number of times PRAM is applied to a variable A with categories 1 and 2 and let $A_1^*, ..., A_m^*$ be the variables with the perturbed scores. Let the PRAM matrix $P_A = (p_{kl})$be given by (1.2).

Assume that the score is picked that is present in the majority of the perturbed scores $A_1^*, ..., A_m^*$ and let the selected score be denoted by A^*.Within this procedure we consider the transition probabilities that $A_i^* = 2$ given that $A_i = 2$ in order to compare them with the transition probabilities in P_A. Note that when the disclosure control is approached in this way, the distribution of the variable in the original file is *not* important; the disclosure control is looked at per record.

Without any further information, two stochastic variables $A_{m_1}^*$ and $A_{m_2}^*$ ($m_1 \neq m_2$) are not independent, since $A_{m_1}^*$ provides a clue to the original score on A, which in turn gives a clue to the score on $A_{m_2}^*$. But conditioned on the score of A the two variables are independent.

When $m = 1$, we get $I\!P_1(A_i^* = 2|A_i = 2) = p_{22} = 0.8$. When $m = 3$, one has

$$\begin{aligned} I\!P_3(A_i^* = 2|A_i = 2) &= I\!P(A_{1i}^* = 2, A_{2i}^* = 2, A_{3i}^* = 2|A_i = 2) \\ &\quad + \binom{3}{1} I\!P(A_{1i}^* = 2, A_{2i}^* = 2, A_{3i}^* = 1|A_i = 2) \\ &= p_{22}^3 + 3 \cdot p_{22}^2 p_{21} \\ &= 0.896 \end{aligned}$$

The general case is given by

$$I\!P_m\ (\ A_i^* = 2|A_i = 2) = I\!P(A_{1i}^* = 2, ..., A_{mi}^* = 2|A_i = 2) +$$

$$+ \binom{m}{1} I\!P(A_{1i}^* = 1, A_{2i}^* = 2, ..., A_{mi}^* = 2|A_i = 2) + ... +$$

$$+ \binom{m}{\lfloor \frac{m}{2} \rfloor} I\!P(A_{1i}^* = 1, .., A_{\lfloor \frac{m}{2} \rfloor i}^* = 1, A_{\lceil \frac{m}{2} \rceil i}^* = 2, ..., A_{mi}^* = 2|A_i = 2) \quad (3.7)$$

Instead of proving mathematically that (3.7) converges to 1 for $m \to \infty$, we tested this conjecture numerically (with $p_{22} = 0.8$):

m	$I\!P_m(A_i^* = 2\|A_i = 2)$
2	0.64
3	0.896
4	0.8192
5	0.94208
6	0.90112
10	0.967206
25	0.999631
50	0.999997
100	1.000000

Because of the procedure to pick the score that is present in the majority of the perturbed scores, there is discontinuity between the cases that m is odd and m is even. When $m = 2$, $I\!P_2(A_i^* = 2|A_i = 2) = 0.64$ which is even smaller than p_{22}. The trend is not really interrupted by this; in general, when m becomes large $I\!P_m(A_i^* = 2|A_i = 2)$ tends to go towards 1. The disclosure limitation per record accomplished by using the PRAM matrix P_A is essentially eliminated in this way: the transition probabilities in P_A are overruled by the transition probabilities $I\!P_m(A_i^* = l|A_i = k)$.

When A has more than two categories, the notation becomes elaborate since the possible number of combinations of values increases. However, the concept does not change and we think that also in this situation the disclosure limitation per record accomplished by using the PRAM matrix P_A in the standard way is eliminated when PRAM is executed several times to the same scores.

Another approach to the disclosure limitation is possible. Consider the probability that given the m perturbed scores of one record, the original score of the record is disclosed, i.e., we are interested in $I\!P(A_i = k|A_{1i}^* = l_1, ..., A_{mi}^* = l_n)$ where

$k, l_1, ..., l_m \in \{1, ..., K\}$. It will turn out that when the disclosure control is approached in this way, the distribution of the variable in the original file is important. Of course, $I\!P(A_i = k|A^*_{1i} = l_1, ..., A^*_{mi} = l_n)$ is not the exact probability that the original score is disclosed; disclosure is an act of the recipient of the perturbed data and is an act far to complex to describe with this probability alone. Nevertheless, it is obvious that $I\!P(A_i = k|A^*_{1i} = l_1, ..., A^*_{mi} = l_n)$ is important regarding disclosure limitation.

One has

$$\begin{aligned} &I\!P(A_i = k|A^*_{1i} = l_1, ..., A^*_{mi} = l_m) = \\ &= \frac{I\!P(A_i = k, A^*_{1i} = l_1, ..., A^*_{mi} = l_m)}{I\!P(A^*_{1i} = l_1, ..., A^*_{mi} = l_m)} \\ &= \frac{I\!P(A^*_{1i} = l_1|A_i = k)...I\!P(A^*_{mi} = l_m|A_i = k)I\!P(A_i = k)}{\sum_{k=1}^{K} I\!P(A^*_{1i} = l_1|A_i = k)...I\!P(A^*_{mi} = l_m|A_i = k)I\!P(A_i = k)} \\ &= \frac{p_{kl_1}...p_{kl_m}I\!P(A_i = k)}{\sum_{k=1}^{K} p_{kl_1}...p_{kl_m}I\!P(A_i = k)}. \end{aligned} \tag{3.8}$$

When we go back to the example in Chapter 1 with the file containing 100 individuals, 99 male and 1 female, we can compare probabilities of disclosure. Assume that the recipient of the perturbed data knows that the original file contains 1 female. When PRAM is applied once with P_A as given by (1.2), the probability that the female in the perturbed file is indeed the female in the original file, $I\!P(A_i = 2|A^*_i = 2)$, is estimated to be 0.075.

When PRAM is applied twice we can use (3.8) and estimate $I\!P(A_i = 2|A^*_{1i} = 2, A^*_{2i} = 2)$ to be 0.393, which is regarding disclosure control too large. When $m = 3$ we get $I\!P(A_i = 2|A^*_{1i} = 2, A^*_{2i} = 2, A^*_{3i} = 2) = 0.838$.

Note that

$$I\!P(A_i = 2|A^*_{1i} = 2, ..., A^*_{mi} = 2) = \frac{I\!P(A_i = 2)}{(p_{12}/p_{22})^m I\!P(A_i = 1) + I\!P(A_i = 2)}.$$

So $I\!P(A_i = 2|A^*_{1i} = 2, ..., A^*_{mi} = 2) \to 1$, when $m \to \infty$.

But there is more to consider. Although $I\!P(A_i = 2|A^*_{1i} = 2, ..., A^*_{mi} = 2)$ converges to 1, when m converges to infinity, $I\!P(A^*_{1i} = 2, ..., A^*_{mi} = 2)$ converges to zero when m converges to infinity. For example, $I\!P(A^*_{1i} = 2, A^*_{2i} = 2, A^*_{3i} = 2) = 0.006$. In other words, a particular realization of the perturbation by PRAM can disclose the identity a the female with probability close to 1, but at the other hand, this

realization has a small probability. Therefore, it is more realistic to consider the probability that $A = k$ *given* that the majority of the perturbed scores has the value k.

Again, let the score that is present in the majority of the perturbed scores $A_1^*, ..., A_m^*$ be denoted by A^*. We are interested in $I\!P_m(A_i = k|A_i^* = k) = I\!P(A_i = k|\text{majority of } A_1^*, ..., A_m^* \text{ has value } k)$ where $k \in \{1, ..., K\}$. The general case is given by

$$I\!P_m(A_i = k|A_i^* = l) = \frac{I\!P_m(A_i^* = l|A_i = k)I\!P(A_i = k)}{\sum_{p=1}^{K} I\!P_m(A_i^* = l|A_i = p)I\!P(A_i = p)}, \tag{3.9}$$

where $I\!P_m(A_i^* = l|A_i = p)$ can be computed for $l, p \in \{1, ..., K\}$, using the same idea as in (3.7). Since $I\!P_m(A_i^* = l|A_i = k)$ for $k = l$ converges to 1 when $m \to \infty$, it follows that $I\!P_m(A_i^* = l|A_i = k)$ for $k \neq l$ converges to 0 when $m \to \infty$. So, we conjecture that $I\!P_m(A_i = k|A_i^* = k)$ converges to 1 when $m \to \infty$. This is not proven wrong by the little test we did with the example in Chapter 1, where $K = 2$, $p_{11} = 0.9$ and $p_{22} = 0.8$:

m	$I\!P_m(A_i = 2\|A_i^* = 2)$
2	0.392638
3	0.2442748
4	0.6910164
5	0.5264428
6	0.8775576
10	0.9851863
25	0.9999839
50	1.0000000
100	1.0000000

Notation becomes elaborate when A has more than two categories, but the concept does not change.

The above discussion of $I\!P_m(A_i = 2|A_i^* = 2)$ also supports our idea that the disclosure limitation accomplished by using the PRAM matrix P_A in the standard way is eliminated when PRAM is executed several times to the same scores.

For the moment we will leave aside whether disclosure control should be approached by considering the transition probabilities as in (3.7) or by considering also the distribution in the original file as we did in (3.9).

The effect of PRAM on disclosure limitation is also discussed in Kooiman et al. (1997) and Gouweleeuw et al. (1998). In these papers the effect is considered in the case PRAM is applied once and a measure is presented for the amount of confusion

introduced by the perturbation. It may be worthwhile to investigate the possibility that this measure is adapted in order that it can be used in the case PRAM is applied more than once.

Out of limitation regarding the subject of this report, we did not go into the matter deeply. Nevertheless, we conjecture that because of the disclosure control problems mentioned in this section the suggestion to release several realizations of applying PRAM to the original scores is not a good idea.

Chapter 4

Two-Way Contingency Tables

4.1 Introduction

This chapter discusses analysis of two-way contingency tables which are constructed using a microdata file to which PRAM has been applied to one or more variables. It considers basic analysis as difference of proportions, relative risk, odds ratio and the Pearson chi-squared test.

When an analyst has at his disposal a microdata file perturbed by PRAM, he can use the moment estimator to estimate the original table and perform analyses on the basis of this estimate. It is also possible that the analyst wishes to consider a particular analysis and wishes to adjust the analysis in order to work directly with the table of the perturbed variables. Both these situations will be considered.

4.2 Difference of Proportions

The comparison of proportions between two subgroups is a simple form of analysis and is a good starting point to explore the influence of PRAM on a 2×2 contingency table.

Let A and B be two binary variables, taking only the value 1 or 2, and let T_{AB} be their contingency table with A the row variable and B the column variable. Furthermore, let $\pi_{A|B}(i|j)$ denote $I\!P(A = i|B = j)$ and let $P_{AB}(i|j)$ be the estimator of $\pi_{A|B}(i|j)$ with realization:

$$p_{A|B}(i|j) = \frac{T_{AB}(i,j)}{T_{AB}(2,j) + T_{AB}(1,j)}, \tag{4.1}$$

with the convention that $p_{A|B}(i|j) = 0$, when the denominator in (4.1) equals zero. So, given T_{AB}, the proportion difference $\pi_{A|B}(2|2) - \pi_{A|B}(2|1)$ can be estimated by $p_{A|B}(2|2) - p_{A|B}(2|1)$.

For example, assume that we want to know whether a certain disease is associated with gender and we have a contingency table in which A denotes the presence or absence of the disease and B denotes gender (presence of the disease: $A = 1$ and males: $B = 1$):

T_{AB}		B		
		1	2	
A	1	35	30	65
	2	45	70	115
		80	100	180

Using (4.1), $\pi_{A|B}(2|2) - \pi_{A|B}(2|1)$ is estimated to be $0.70 - 0.56 = 0.14$. The difference of 0.14 is an indication that gender and the disease are associated: $\pi_{A|B}(2|2) > \pi_{A|B}(2|1)$. So being a female diminishes the risk of having the disease.

Note that $\pi_{A|B}(1|2) + \pi_{A|B}(2|2) = 1 = \pi_{A|B}(1|1) + \pi_{A|B}(2|1)$. So $\pi_{A|B}(2|2) - \pi_{A|B}(2|1) = \pi_{A|B}(1|1) - \pi_{A|B}(1|2)$. Comparison on $A = 2$ is equivalent to comparison on $A = 1$. See Agresti (1990,1996) for more information concerning the difference of proportions.

Consider the situation in which PRAM is applied to variable A where the perturbed variable is denoted by A^*, and the analyst has a microdata file which contains, amongst others, A^* and a not-perturbed variable B.

Let PRAM be applied to A non-differentially with respect to B according to the Markov matrix

$$P_A = \begin{pmatrix} \alpha & 1-\alpha \\ 1-\beta & \beta \end{pmatrix}.$$

The analyst can estimate the original table using the unbiased moment estimator of T_{AB} given by

$$\widehat{T}_{AB} = (P_A^{-1})^t T_{A^*B},$$

where T_{A^*B} is the table of A^* and B (Kooiman et al., 1997).

With $\widehat{T}_{AB}$ at hand, we can define an estimator of $\pi_{A|B}(2|j)$, denoted by $\widehat{P}_{AB}(2|j)$ with realization $\widehat{p}_{AB}(2|j)$, as in the situation with T_{AB}, and estimate $\pi_{A|B}(2|2) - \pi_{A|B}(2|1)$ by

$$\widehat{p}_{A|B}(2|2) - \widehat{p}_{A|B}(2|1) = \frac{\widehat{T}_{AB}(2,2)}{\widehat{T}_{AB}(2,2) + \widehat{T}_{AB}(1,2)} - \frac{\widehat{T}_{AB}(2,1)}{\widehat{T}_{AB}(2,1) + \widehat{T}_{AB}(1,1)} \qquad (4.2)$$

A second possibility is to estimate $\pi_{A|B}(2|2) - \pi_{A|B}(2|1)$ directly by means of T_{A^*B}. When PRAM is applied, the realization $p_{A|B}(i|j)$ is perturbed. Let $P_{A^*|B}(i|j)$ denote the stochastic function of this process and let $p_{A^*|B}(i|j)$ be its realization.

When A is perturbed, the number of records with $B = j$ $(j = 1, 2)$ does not change. So $T_{AB}(2,2) + T_{AB}(1,2) = T_{A^*B}(2,2) + T_{A^*B}(1,2)$. Therefore

$$\begin{aligned} I\!E\left[P_{A^*|B}(2|2)\right] &= I\!E\left[\frac{T_{A^*B}(2,2)}{T_{A^*B}(2,2) + T_{A^*B}(1,2)}\right] \\ &= \frac{1}{T_{AB}(2,2) + T_{AB}(1,2)} I\!E\left[T_{A^*B}(2,2)\right]. \end{aligned} \tag{4.3}$$

Because of this property and from (1.1) it follows that for $j = 1, 2$

$$I\!E\left[P_{A^*|B}(2|j)\right] = \beta p_{A|B}(2|j) + (1-\alpha) p_{A|B}(1|j). \tag{4.4}$$

Using (4.4) yields

$$I\!E\left[P_{A^*|B}(2|2) - P_{A^*|B}(2|1)\right] = (\alpha + \beta - 1)\left(p_{A|B}(2|2) - p_{A|B}(2|1)\right) \tag{4.5}$$

(Kuha and Skinner, 1997). Analogous to (4.1), $p_{A^*|B}(i|j)$ can be determined using T_{A^*B}. Therefore, $\pi_{A|B}(2|2) - \pi_{A|B}(2|1)$ can be estimated by

$$\frac{p_{A^*|B}(2|2) - p_{A^*|B}(2|1)}{\alpha + \beta - 1}. \tag{4.6}$$

The estimates (4.2) and (4.6) of $\pi_{A|B}(2|2) - \pi_{A|B}(2|1)$ are identical, which can be proved by analysing the moment estimator. Furthermore, this estimator is unbiased: the moment estimator yields an unbiased estimate of the cell frequency $T_{AB}(i,j)$ and (4.3) shows that this is enough to obtain an unbiased estimate of $\pi_{A|B}(i|j)$.

Two remarks as a result of (4.5). First, it turns out that applying PRAM diminishes the difference of sample proportions: because α en β are transition probabilities close to 1, it follows that $0 < \alpha + \beta - 1 < 1$. Secondly, (4.5) makes clear that the use of invariant PRAM is limited in the situation where a perturbed microdata file is given to the analyst: in the univariate case we can build the PRAM matrix in such a way that T_{A^*} is a unbiased estimator of T_A, but then T_{A^*B} is not an unbiased estimator of T_{AB}, which means that $p_{A^*|B}(2|2) - p_{A^*|B}(2|1)$ is not an unbiased estimator of $\pi_{A|B}(2|2) - \pi_{A|B}(2|1)$.

The difference between the two methods to estimate $\pi_{A|B}(2|2) - \pi_{A|B}(2|1)$ becomes interesting when the original table is not estimated with the moment estimator. For instance, when the original table is estimated by a maximum likelihood

estimator to prevent negative cell frequencies from occurring, see Chapter 5, it is possible that the two methods to estimate the difference of proportion yield different values.

Up till now, we have considered a point estimate of the original difference of proportions in case one of the variables is perturbed by PRAM. Next we discuss how the extra variance due to applying PRAM should be incorporated in this situation.

Regarding the difference of proportions, the columns in the table are treated as independent binomial samples: in column j of the 2×2 table T_{AB}, $T_{AB}(i,j)$ is considered to be a realization of a binomial distribution with sample size $T_{AB}(+,j)$ and probability $\pi_{A|B}(i|j)$, where $T_{AB}(+,j) = T_{AB}(1,j) + T_{AB}(2,j)$. Note that the sum $T_{AB}(+,j)$ is not considered to be stochastic.

This means that

$$V\left[P_{A|B}(i|j)\right] = \frac{\pi_{A|B}(i|j)\left(1-\pi_{A|B}(i|j)\right)}{T_{AB}(+,j)}.$$

Since estimators $P_{A|B}(2|2)$ and $P_{A|B}(2|1)$ are independent, their difference has expectation

$$I\!E\left[P_{A|B}(2|2) - P_{A|B}(2|1)\right] = \pi_{A|B}(2|2) - \pi_{A|B}(2|1)$$

and variance

$$V\left[P_{A|B}(2|2) - P_{A|B}(2|1)\right] = \sum_{j=1}^{2} \frac{\pi_{A|B}(2|j)\left(1-\pi_{A|B}(2|j)\right)}{T_{AB}(+,j)}. \tag{4.7}$$

To estimate $V\left[P_{A|B}(2|2) - P_{A|B}(2|1)\right]$ the term $\pi_{A|B}(2|j)$ in formula (4.7) is replaced by $p_{A|B}(2|j)$, for $j = 1, 2$.

So far for the standard model, source: Agresti (1990).

When PRAM is applied, it is applied independently of the binomial model assumed. We would like to use the estimated original table, denoted by $\widehat{T}$, as though it is the original table in order to calculate difference op proportions in a standard way and take into account extra variance because of PRAM. Strictly speaking, this is not possible since the columns in $\widehat{T}$ are not independent binomial samples due to the perturbation because of PRAM.

For the time being we suggest the following method, which should be investigated in future research. Regarding the variance, we propose to sum the variance due to PRAM and the variance due to the independent binomial model. Let V^* denote the extra variance per cell due to PRAM, then for $j = 1, 2$

$$V\left[\widehat{T}_{AB}(2,j)\right] = T_{AB}(+,j)\pi_{A|B}(2|j)\left(1-\pi_{A|B}(2|j)\right) + V^*\left[\widehat{T}_{AB}(2,j)\right] \tag{4.8}$$

and therefore

$$V\left[\widehat{P}_{A|B}(2|2) - \widehat{P}_{A|B}(2|1)\right] = \frac{V\left[\widehat{T}_{AB}(2,2)\right]}{T_{AB}(+,2)^2} + \frac{V\left[\widehat{T}_{AB}(2,1)\right]}{T_{AB}(+,1)^2} + \frac{2C\left[\widehat{T}_{AB}(2,2), \widehat{T}_{AB}(2,1)\right]}{\widehat{T}_{AB}(+,2)\widehat{T}_{AB}(+,1)} \tag{4.9}$$

The covariance in (4.9) is the covariance due to PRAM, since the covariance due to the binomial model is zero because of the independence between the columns.

To estimate $V\left[\widehat{P}_{A|B}(2|2) - \widehat{P}_{A|B}(2|1)\right]$, the term $\pi_{A|B}(2|j)$ in formula (4.8) is replaced by $\widehat{p}_{A|B}(2|j)$ for $j = 1, 2$ and (4.8) is used in (4.9).

4.3 Relative Risk

It is not always the case that analyses on the basis of the table of perturbed variables can be adjusted as simply as in (4.6).

Let the notation be as in the previous section. The relative risk is defined as the ratio of two proportions, e.g. $\pi_{A|B}(2|2)/\pi_{A|B}(2|1)$. The distance to 1 of this ratio is a measure of the difference between the subgroups, see also Agresti (1990,1996).

Let PRAM be applied to A. Because of the equations

$$I\!E\left[P_{A^*|B}(2|j)\right] = \beta p_{A|B}(2|j) + (1-\alpha)(1 - p_{A|B}(2|j)) \tag{4.10}$$

for $j = 1, 2$, the sample proportion $p_{A|B}(2|j)$ can be unbiasedly estimated for $j = 1, 2$ by using T_{A^*B} to get

$$\frac{p_{A^*|B}(2|j) - 1 + \alpha}{\alpha + \beta - 1}. \tag{4.11}$$

But (4.11) does not provide a unbiased estimate of the ratio $p_{A|B}(2|2)/p_{A|B}(2|1)$. In general: $I\!E[X^{-1}] \neq I\!E[X]^{-1}$.

A possible way out is to determine the maximum likelihood estimate (MLE) of $p_{A|B}(i|j)$, say $\widetilde{p}_{A|B}(i|j)$, and estimate $1/p_{A|B}(i|j)$ by $1/\widetilde{p}_{A|B}(i|j)$. This is possible because when $\widetilde{\psi}$ maximizes the likelihood function $l(\psi)$, $g(\widetilde{\psi})$ is the MLE of $g(\psi)$, on condition that g^{-1} exists. This follows from noting that the likelihood function of $\phi = g(\psi)$ is $l(g^{-1}(\phi))$, which is maximized when $\phi = g(\widetilde{\psi})$ (Little and Rubin (1987), section 5.1, see also: Kotz and Johnson (1982)). Note that in general $\widetilde{p}_{A|B}(i|j)$ need not to be an unbiased estimator of $p_{A|B}(i|j)$.

When we use the moment estimator to estimate the original table and then estimate the sample proportion $p_{A|B}(2|j)$, we get the same estimate as in (4.11).

As will be explained in Chapter 5, we conjecture that the moment estimator yields the MLE of the original table on the condition that the estimated cell frequencies are positive. So, on this condition, we can use $\widehat{T}_{AB}$ to compute the MLE of $p_{A|B}(2|j)$, that is $\widehat{p}_{A|B}(i|j) = \widetilde{p}_{A|B}(i|j)$, and estimate $p_{A|B}(2|2)/p_{A|B}(2|1)$ by

$$\frac{\widehat{p}_{A|B}(2|2)}{\widehat{p}_{A|B}(2|1)}.$$

Or equivalently, we can work with T_{A^*B} and estimate $p_{A|B}(2|2)/p_{A|B}(2|1)$ by using (4.11) to get

$$\frac{p_{A^*B}(2|2) - 1 + \alpha}{p_{A^*B}(2|1) - 1 + \alpha}.$$

We use this estimate of $p_{A|B}(2|2)/p_{A|B}(2|1)$ as an estimate of the true ratio $\pi_{A|B}(2|2)/\pi_{A|B}(2|1)$. With this procedure we drop the requirement of unbiasedness and settle for maximum likelihood properties. Copeland et al. (1977) provide the same estimator, but do not mention its properties.

Since

$$\frac{p_{A^*|B}(2|2)}{p_{A^*|B}(2|1)} = \frac{\widehat{p}_{A|B}(2|2) + \frac{1-\alpha}{\beta-1+a}}{\widehat{p}_{A|B}(2|1) + \frac{1-\alpha}{\beta-1+a}},$$

it turns out that applying PRAM diminishes the difference between the subgroups: the 'effect' of B on A is made to seem less than it is. Applying PRAM attenuates the ratio away from $p_{A|B}(2|2)/p_{A|B}(2|1)$ towards the null value of one (Kuha and Skinner, 1997).

We did not consider thoroughly the extra variance due to applying PRAM, but we think that the discussion in Greenland (1988) concerning the extra variance in the estimation of the odds ratio due to misclassification is useful.

4.4 The Odds Ratio

The odds ratio θ is a measure of association for 2×2 contingency tables. Let $\pi_{ij} = I\!P(A = i,\ B = j)$ denote the probability that (A, B) falls in the cell in row i and column j. The odds ratio is defined as

$$\theta = \frac{\pi_{11}/\pi_{12}}{\pi_{21}/\pi_{22}} = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}}.$$

(Agresti, 1990,1996, and Fienberg, 1980). The odds ratio is also called the cross-product ratio. With T_{AB} the observed table, the *sample* odds ratio equals

$$\widehat{\theta} = \frac{T_{AB}(1,1)T_{AB}(2,2)}{T_{AB}(1,2)T_{AB}(2,1)}.$$

For multinomial and Poisson sampling, this is the ML estimator of the true odds ratio (Agresti, 1996).

If we think of row totals as fixed, then π_{11}/π_{12} is the odds of being in the first column given that one is in the first row, and π_{21}/π_{22} is the corresponding odds for the second row. When the two cross classified variables A and B are independent, $\pi_{11}/\pi_{12} = \pi_{21}/\pi_{22}$, so $\theta = 1$. This value of θ serves as a standard measure for comparison. When $1 < \theta < \infty$, the odds of $B = 1$ are higher given $A = 1$ than given $A = 2$. For instance, when $\theta = 4$, the odds of $B = 1$ given $A = 1$ are four times the odds of $B = 1$ given $A = 2$. When $0 < \theta < 1$, $B = 1$ given $A = 1$ is less likely than $B = 1$ given $A = 2$. Values of θ farther from 1 in a given direction represent stronger levels of association.

It is possible to use proportions to define the odds ratio:

$$\theta = \frac{\pi_{A|B}(1|1)}{\pi_{A|B}(2|1)}\left(\frac{\pi_{A|B}(1|2)}{\pi_{A|B}(2|2)}\right)^{-1} = \frac{\pi_{A|B}(1|1)}{1-\pi_{A|B}(1|1)}\left(\frac{\pi_{A|B}(1|2)}{1-\pi_{A|B}(1|2)}\right)^{-1}. \tag{4.12}$$

Let PRAM be applied to A non-differentially with respect to B and according to the PRAM matrix P_A. Using the information of the PRAM matrix, the MLE of $p_{A|B}(1|j)$ for $j = 1, 2$ can be determined analogously to (4.11).

Using these MLE's in (4.12) yields the MLE of the odds ratio:

$$\widehat{\theta} = \frac{p_{A^*|B}(1|1) - (1-\beta)}{\alpha - p_{A^*|B}(1|1)}\left(\frac{p_{A^*|B}(1|2) - (1-\beta)}{\alpha - p_{A^*|B}(1|2)}\right)^{-1}.$$

This formula can be used only if all the numerators and denominators in the formula are positive.

Assuming that $\widehat{\theta}$ is not equal to zero or infinity, it will always be farther from 1 than the odds ratio which is computed in the standard way using T_{A^*B}. Incorporating the information of the PRAM matrix in the estimation process compensates for the bias towards 1 (Magder and Hughes, 1997).

We will not consider the extra variance due to applying PRAM. Instead, we refer to Greenland (1988) who discusses the extra variance in the estimation of the odds ratio due to misclassification.

4.5 Pearson Chi-Square Test

Regarding the often used Pearson chi-squared test to detect association in a two-way table: PRAM also influences the size of this test.

The null hypothesis of independence is $H_0 : \pi_{ij} = \pi_{i+}\pi_{+j}$. To test H_0 it is possible to estimate the expected frequencies under H_0 by $\widehat{m} = n_{i+}n_{+j}/N$ and calculate

$$X^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(n_{ij} - \widehat{m}_{ij})^2}{\widehat{m}_{ij}}.$$

X^2 has an asymptotic chi-squared distribution with degrees of freedom equal to $(I-1)(J-1)$.

In papers concerning misclassification and contingency tables, it is stated that misclassification influences the size of the chi-squared test (Kuha and Skinner, 1997, Assakul and Proctor, 1967). Obviously, this will also be the case when PRAM is applied to variables in the table. Association in a table is undermined by applying PRAM non-differentially to one variable or non-differentially and independently to two variables. When there is no association between A and B, there certainly will not be association between A^* and B^*. But if there is association between A^* and B^*, then there is even stronger association between A and B (Kuha and Skinner, 1997, section 3.2.3).

4.6 Conclusion

It is interesting to see that because of the attenuation produced by applying PRAM regarding the analyses discussed, a conclusion concerning the *presence* of association is justified on the basis of the observed table alone. Indeed, when the observed table contains a significant difference in proportion, the original table will contain an even larger difference. And the same applies to the relative risk, the odds ratio, and the Pearson chi-squared test: when the risk, the ratio, and chi-squared test is computed without adjustment for PRAM and the outcome suggests association, then there is certainly association in the original table.

In this chapter we showed that adjustment is possible in the case of simple analysis of a 2×2-table. Still, the estimation of extra variances due to PRAM should be worked out in future research.

Chapter 5

EM Algorithm

5.1 Introduction

The expectation-maximization (EM) algorithm is used as an iterative scheme to compute maximum likelihood estimates (MLEs). The algorithm is broadly applied in statistical problems concerning incomplete or erroneous data and it is an alternative to maximizing the likelihood function using methods as e.g. the Newton Raphson method. The algorithm can also be applied in situations bearing a close resemblance to situations with incomplete data: e.g. censored data (Tanner, 1993, Ch. 4), blurred images (McLachlan and Krishnan, 1997, section 2.5), statistical models such as random effects (Magder and Hughes, 1997) or latent class structures (Hagenaars, 1993, app. B). In the following it will be shown that the EM algorithm can be used in situations where PRAM has been applied.

There are many variants of the EM algorithm and the first EM-type of algorithm was proposed by Newcomb in 1886. The much used reference is the seminal paper of Dempster et al. (1977), in which the authors describe the most general form of the algorithm, prove some convergence properties, and give several examples of the use of the algorithm.

This chapter introduces the EM algorithm in a general form and in the form in which it can be used when data have been perturbed by applying PRAM. An alternative to the moment estimator of the original table is presented and analysed.

5.2 General Form of the Algorithm

The standard incomplete data scheme considers the observable incomplete data $y \sim g(y|\phi)$ as resulting from partial observation of complete data $x \sim f(x|\phi)$, where g and f are densities and ϕ a parameter vector in a parameter space Ω. When the vector x can be written as a compounded vector (y, z) where z denotes the missing data, we have

$$g(y|\phi) = \int_Z f((y, z)|\phi)dz.$$

When x can not be written in this way, only the notation becomes more complex (Dempster et al., 1977).

The EM algorithm can be used when the observed data likelihood function is complicated. The algorithm makes use of the often less complicated complete data likelihood function, which sometimes means that standard software for complete data can be used within the iteration. Because of this last mentioned feature, the algorithm can be quite user-friendly.

The general form of the algorithm is as follows. Let $\log f(x|\phi)$ be the complete data log likelihood function and, with $\phi^{(p)}$ the current fit of ϕ, let $Q(\phi|\phi^{(p)})$ be the function

$$\phi \to \int_Z \log f((y, z)|\phi)h(z|y, \phi^{(p)})dz \tag{5.1}$$

where h is the conditional density $h(z|y, \phi) = f((y, z)|\phi)/g(y|\phi)$. Notice that (5.1) is equivalent to

$$\phi \to I\!E_z \left[\log f((y, z)|\phi)|y, \phi^{(p)}\right]. \tag{5.2}$$

The EM algorithm $\phi^{(p)} \to \phi^{(p+1)}$ is defined by:

Choose a starting point $\phi^{(0)}$.

E-step: Calculate $Q(\phi|\phi^{(p)})$.

M-step: Choose $\phi^{(p+1)}$ to be the value of ϕ which maximizes $Q(\phi|\phi^{(p)})$.

The E- and M-step are alternated repeatedly.

As Dempster et al. (1977) put it: 'The heuristic idea here is that we would like to choose ϕ^* to maximize $\log f(x|\phi)$. Since we do not know $\log f(x|\phi)$, we maximize instead its current expectation given the data y and the current fit $\phi^{(p)}$.'

The general property of the algorithm is that the observed data likelihood function, denoted by $g(y|\phi)$, is non-decreasing along any EM sequence $\phi^{(p)}$. Thus for a bounded sequence of likelihood values $\{g(y|\phi^{(p)})\}$, $g(y|\phi^{(p)})$ converges monotically to some value g^*. As McLachlan and Krishnan (1997, section 3.4) state: in almost all applications, g^* is a stationary point. That is $g^* = g(y|\phi^*)$ for some ϕ^* at which $\partial g(y|\phi)/\partial\phi = 0$. In general, if $g(y|\phi)$ has several stationary points, convergence of

the EM sequence to either type (local or global maximizers, saddle points) depends on the choice of the starting point $\phi^{(0)}$.

In the following we mention some results concerning the EM algorithm.

Regarding convergence to a stationary point: there is a convergence theorem for an EM sequence. It is presented in McLachlan and Krishnan (1997, section 3.4). First, we give the following regularity conditions.

$$\Omega \text{ is a subset in } d\text{-dimensional Euclidian space } \mathbb{R}^d.$$

$$\{\phi \in \Omega : g(y|\phi) \geq g(y|\phi_0)\} \text{ is compact for any } g(y|\phi_0) > -\infty.$$

$$g(y|\phi) \text{ is continuous in } \Omega \text{ and differentiable in the interior of } \Omega.$$

(The condition $g(y|\phi_0) > -\infty$ is only significant when we work with the log likelihood. In that case $g(y|\phi)$ is replaced in the regularity conditions by $\log g(y|\phi)$.)

Theorem 1 *Suppose that the regularity conditions hold and that $Q(\phi|\psi)$ is continuous in ϕ and ψ. Then all the limits points of any instance $\{\phi^{(p)}\}$ of the EM algorithm are stationary points of $g(y|\phi)$, and $g(y|\phi^{(p)})$ converges monotically to some value $g^* = g(y|\phi^*)$ for some stationary point ϕ^*.*

A second result is about convergence to an unique maximum and can also be found in McLachlan and Krishnan (1997, section 3.5).

Theorem 2 *Suppose that the regularity conditions hold. When $g(y|\phi)$ has an unique maximum in Ω with ϕ^* being the only stationary point and when $\partial Q(\phi|\psi)/\partial\phi$ is continuous in ϕ and ψ. Then any EM sequence $\{\phi^{(p)}\}$ converges to the unique maximizer ϕ^* of $g(y|\phi)$.*

Of course, the EM algorithm is not a panacea concerning optimization problems. In situations where there are more than one maximizers, there is often no guarantee that the algorithm converges to the global maximum. (This it shares, of course, with other optimization algorithms.) But, in the situation where the data have been perturbed by PRAM there seems to be a good starting point. Because the perturbation by PRAM is small, we can find a starting point by taking the perturbed data as though it were the original data and using standard methods to determine the parameter which will play the role of $\phi^{(0)}$.

5.3 The EM Estimator

This section provides an alternative to the moment estimator. In Chapter 1 we noted that when data are perturbed by applying PRAM, the unbiased moment estimator of the original table can yield negative cell frequencies. When negative estimated frequencies occur, we know that they are not maximum likelihood estimates of the original frequencies. In this section we use the EM algorithm to obtain maximum likelihood estimates of the original frequencies when PRAM is applied to perturb the data.

For ease of exposition we consider the simple case of a 2×1 frequency table of variable A:

T_A	0	$T_A(0)$
	1	$T_A(1)$
		n

When we consider A as a stochastic variable with $\mathbb{P}(A = 0) = \phi$ where $\phi \in \Omega = [0, 1]$, then $A \sim \text{Bin}(1, \phi)$ and $V(A) = \phi(1 - \phi)$. By taking n fixed we assume multinomial sampling. The likelihood function is given by

$$\left(\begin{array}{c} n \\ T_{A(0)} \end{array} \right) \prod_{i=1}^{n} \mathbb{P}(A_i = 0)^{(1-a_i)} \mathbb{P}(A_i = 1)^{a_i}$$

where the a_i's are the realizations of the stochastic functions A_i. Since the aim is to maximize the likelihood function with the a_i fixed, we ignore the constant and apply the natural logarithm. We call it the original data log likelihood:

$$\begin{aligned} \log l(a_i, \phi) &= \sum_{i=1}^{n} \Big((1 - a_i) \log \mathbb{P}(A_i = 0) + a_i \log \mathbb{P}(A_i = 1) \Big) \\ &= \sum_{i=1}^{n} \Big((1 - a_i) \log \phi + a_i \log(1 - \phi) \Big). \end{aligned}$$

The MLE of ϕ is $\widehat{\phi} = \frac{1}{n} \sum_{i=1}^{n} (1 - a_i)$, which is $T_A(0)$ divided by n. The variance of this estimator is $V(\widehat{\phi}) = V(A)/n$.

Next, consider the situation when PRAM has been applied to the a_i, so the values a_i are perturbed to a_i^* according to the Markov matrix

$$P_A = \left(\begin{array}{cc} p_{00} & p_{01} \\ p_{10} & p_{11} \end{array} \right). \tag{5.3}$$

With A perturbed by PRAM, the objective is not changed: the MLE of ϕ. We will use the EM algorithm to achieve this objective.

In the PRAM situation we know $I\!P(A^* = k) = p_{0k}I\!P(A = 0) + p_{1k}I\!P(A = 1)$ for $k = 0, 1$ and $n^* = n$. So, the observed data log likelihood function is given by

$$\begin{aligned}\log l^*(a_i^*, \phi) &= \sum_{i=1}^{n^*} \Big((1 - a_i^*) \log I\!P(A_i^* = 0) + a_i^* \log I\!P(A_i^* = 1) \Big) \\ &= \sum_{i=1}^{n} \Big((1 - a_i^*) \log \left(p_{00}\phi + p_{10}(1 - \phi)\right) + a_i^* \log \left(p_{01}\phi + p_{11}(1 - \phi)\right) \Big)\end{aligned} \tag{5.4}$$

In the following, $l^*(a_i^*, \phi)$ is maximized for $\phi \in [0, 1]$ and a_i^* fixed, by using the EM algorithm.

The EM algorithm maximizes the observed data likelihood by iteratively maximizing the expected value of the original data log likelihood, where the expectation is taken over the distribution of the original data given the perturbed data and the current value of ϕ, denoted by $\phi^{(p)}$. That is, we take function $Q(\phi|\phi^{(p)})$ in (5.2) to be

$$\phi \to I\!E_{A_i} \left[\log l(A_i, \phi) | a^*, \phi^{(p)} \right].$$

Because in this case $l(a_i, \phi)$ is linear with respect to the a_i's, this can be done by maximizing $l(a_i, \phi)$ after each unknown a_i in $l(a_i, \phi)$ is replaced with the expected value of a_i given the observed a_i^* and $\phi^{(p)}$:

$$Q(\phi|\phi^{(p)}) = \sum_{i=1}^{n} \Big((1 - I\!E_{A_i} \left[A_i | a_i^*, \phi^{(p)} \right]) \log \phi + I\!E_{A_i} \left[A_i | a_i^*, \phi^{(p)} \right] \log(1 - \phi) \Big).$$

So in this case the EM algorithm is equivalent to a procedure which first estimates the original data and them determines $\widehat{\phi}$ (see also Dempster et al., 1977, and Magder and Hughes, 1997).

Since the objective is the MLE of ϕ, it is not necessary to estimate $I\!E_{A_i} \left[A_i | a_i^*, \phi^{(p)} \right]$ for each i; it is enough to estimate $I\!E_A \left[T_A | a_i^*, \phi^{(p)} \right]$. These expected frequencies of a_i can be determined by first considering the distribution of A and A^* together and subsequently considering the distribution of A.

We start the EM algorithm with an initial value of $\phi^{(0)}$, which we determine by using the perturbed data: we take $\phi^{(0)}$ equal to the number of observations $a_i^* = 0$ divided by n.

In the first E-step we estimate $I\!E_A \left[T_A | a_i^*, \phi^{(p)} \right]$ using the observed a_i^* and $\phi^{(p)} = \phi^{(0)}$. This goes as follows.

Let T_{AA^*} be the cell frequencies in the 2×2 table of A and A^*, and T_{A^*} the observed cell frequencies in the 2×1 table of A^*. Note that the analyst of the

perturbed data does not have T_{AA^*}, since he has A^* and the PRAM matrix only. One has for $i, j \in \{0, 1\}$

$$\begin{aligned}
\mathbb{E}\left[T_{AA^*}(i,j)\right] &= \mathbb{P}(A=i, A^*=j)n \\
&= \frac{\mathbb{P}(A=i, A^*=j)}{\mathbb{P}(A^*=j)}\mathbb{P}(A^*=j)n \\
&= \frac{\mathbb{P}(A=i, A^*=j)}{\mathbb{P}(A^*=j)}\mathbb{E}\left[T_{A^*}(j)\right] \\
&= \frac{\mathbb{P}(A^*=j|A=i)\mathbb{P}(A=i)}{p_{0j}\mathbb{P}(A=0)+p_{1j}\mathbb{P}(A=1)}\mathbb{E}\left[T_{A^*}(j)\right]
\end{aligned}$$

So, when we estimate $\mathbb{P}(A=0)$ by $\phi^{(0)}$ and $\mathbb{E}\left[T_{A^*}(j)\right]$ by the realization $T_{A^*}(j)$, we get for instance

$$\widehat{T}^{(1)}_{AA^*}(0,1) = \frac{p_{01}\phi^{(0)}}{p_{01}\phi^{(0)} + p_{11}(1-\phi^{(0)})} T_{A^*}(1).$$

We obtain the expected frequencies of A by $\widehat{T}^{(1)}_A(j) = \widehat{T}^{(1)}_{AA^*}(j,0) + \widehat{T}^{(1)}_{AA^*}(j,1)$ for $j = 0, 1$. This estimation of $T^{(1)}_A$ ends the E-step.

The first M-step is the next estimation of ϕ using $\widehat{T}^{(1)}_A$ which is easy: $\phi^{(1)} = \widehat{T}^{(1)}_A(0)/n$.

This EM procedure is iterated: with $\phi^{(1)}$ we compute $\widehat{T}^{(2)}_{AA^*}$ in the second E-step and determine $\phi^{(2)}$ in the second M-step, etcetera (based on Kuha and Skinner, 1997).

Since for each $\phi^{(p)}$

$$Q(\phi|\phi^{(p)}) = \sum_{i=1}^{n}(1 - \mathbb{E}\left[A_i|a_i^*, \phi^{(p)}\right])\log\phi + \mathbb{E}\left[A_i|a_i^*, \phi^{(p)}\right]\log(1-\phi)$$

is continuous in both ϕ and $\phi^{(p)}$, and the regularity conditions are fulfilled, the sequence $\log l^*(a_i^*, \phi^{(p)})$ converges to $\log l^*(a_i^*, \phi^*)$ for some stationary point ϕ^* (see theorem 1).

In general, let A have K categories and let $\phi(i) = \mathbb{P}(A=i)$, the EM algorithm in this situation:

$$\text{Initial values:} \quad \phi^{(0)}(i) = \frac{T_{A^*}(i)}{n}$$

$$\text{E-step:} \qquad T^{(v)}_{AA^*}(i,j) = \frac{p_{ij}\phi(i)^{(v)}}{\sum_{i=1}^{K} p_{ij}\phi(i)^{(v)}} T_{A^*}(j)$$

$$T^{(v)}_{A}(i) = \sum_{j=1}^{K} T^{(v)}_{AA^*}(i,j)$$

$$\text{M-step:} \qquad \phi^{(v+1)}(i) = \frac{T^{(v)}_{A}(i)}{n}$$

As an example of the method we consider the situation where

T_A	65
	87
	152

and $P_A = \begin{pmatrix} 9/10 & 1/10 \\ 2/10 & 8/10 \end{pmatrix}$.

So, in the original situation $\widehat{\phi} = 65/152 = 0.428$ and because $V(\widehat{\phi}) = \widehat{\phi}(1-\widehat{\phi})/152$, the standard error is estimated to be 0.040.

Let T_{A*} be a possible realization when A is perturbed A^* by applying PRAM and be given by

T_{A^*}	75
	77
	152

Firstly, we use the moment estimator to analyse the perturbed data. Let $\widehat{T}_A$ be the estimation of the original table as a result of the moment estimator. $\widehat{T}_A$ is given by

$\widehat{T}_A$	63.714
	88.286
	152

Where the standard error of the frequencies in $\widehat{T}_A$ is 6.366.

In this situation, $\widehat{\phi}$ is estimated by $\widehat{\phi} = 63.714/152 = 0.419$ and its standard error is a sum because to the variance $\widehat{\phi}(1-\widehat{\phi})/152$ should be added the extra variance due to PRAM which is $N^{-2}V(\widehat{T}_A) = 152^{-2} \cdot 6.366^2$. Therefore $SE(\widehat{\phi}) = 0.058$

Secondly, we use the EM algorithm. Note that the EM algorithm as explained in this section has as its objective the MLE of ϕ and not an estimate of T_A. Nevertheless, we can speak of an MLE estimate of T_A because of the statistical method (in the normal situation) to relate an estimate of ϕ directly to a realization of a table. Therefore, the MLE estimate of ϕ provides also an estimate of the original table.

The EM algorithm provides the same estimate as the moment estimator:

iterations	A=0	A=1
1	69.919	82.081
2	66.892	85.108
5	64.159	87.841
10	63.732	88.268
15	63.715	88.285
20	63.714	88.286
25	63.714	88.286
50	63.714	88.286
100	63.714	88.286
500	63.714	88.286
1000	63.714	88.286

To estimate the standard error of this EM result, we use a bootstrap and Monte Carlo simulation (McLachlan and Krishnan, 1997):

Step 1. A new set of 'perturbed' data $T^*_{A^*}$ is created using $\widehat{\phi} = 0.419$.

Step 2. The EM algorithm is applied to the bootstrap observed data $T^*_{A^*}$ and its output is denoted by $\widehat{\phi}^*$.

Steps (1) and (2) are repeated independently a number of times (say B) to give estimates $\widehat{\phi}^*_1$, $\widehat{\phi}^*_2$, ..., $\widehat{\phi}^*_B$. Then the bootstrap covariance matrix of $\widehat{\phi}^*$ can be approximated by the sample covariance matrix of these B bootstrap replications to give

$$cov^*(\widehat{\phi}^*) = \sum_{b=1}^{B}(\widehat{\phi}^*_b - \overline{\widehat{\phi}^*})(\widehat{\phi}^*_b - \overline{\widehat{\phi}^*})^t/(B-1)$$

where

$$\overline{\widehat{\phi}^*} = \sum_{b=1}^{B}\widehat{\phi}^*_b/B.$$

In this example $(\widehat{\phi}^*_b - \overline{\widehat{\phi}^*})(\widehat{\phi}^*_b - \overline{\widehat{\phi}^*})^t = (\widehat{\phi}^*_b - \overline{\widehat{\phi}^*})^2$ because $\widehat{\phi}^*$ is a scalar. Furthermore, estimates $\widehat{\phi}^*_1$, $\widehat{\phi}^*_2$, ..., $\widehat{\phi}^*_B$ can be used directly to construct a 95%-confidence interval. When $B = 1000$ for instance and the estimates are ordered in order of magnitude, number 26 and number 974 form the 95%-confidence interval.

Note that in step 1 we need to create a perturbed data set. So we have to use $\widehat{\phi}$ in the distribution of A^*: $I\!P(A^*_i = 0) = p_{00}I\!P(A_i = 0) + p_{10}I\!P(A_i = 1)$ which is therefore estimated by $p_{00}\widehat{\phi} + p_{10}(1-\widehat{\phi})$. Note also that in using the EM algorithm and the bootstrap, we have incorporated not only the perturbation due to PRAM,

but also the multinomial sampling scheme. This means that the variance which is estimated using the bootstrap, is the total variance of $\widehat{\phi}$.

We applied the bootstrap method four times taking each time $B = 500$. This yielded three times a rounded standard error of 0.058 and once a standard error of 0.059. So we estimate $SE(\widehat{\phi}) = 0.058$, which is the same as the estimate obtained using the moment estimator.

We illustrated the EM algorithm for the 2×1 table, but as we noted, the algorithm can be used also when A has K categories. Therefore, we can stack the columns of a $I \times J$-table and build the accompanying PRAM matrix using as in (1.4), the original table can always be estimated by using the EM algorithm.

An advantage of this estimator, which we will call the *EM estimator*, compared to the moment estimator is that there will never be estimated cell frequencies with negative values. For example, consider a new situation where A and B are cross-classified and A is perturbed by using PRAM matrix P_A as specified in (1.2). Let the original table be given by

T_{AB}	B		
A	214	12	226
	14	0	14
	228	12	240

Two possible realizations of a perturbed table after applying PRAM to A are given by

T_{A^*B}	B		
A^*	189	11	200
	39	1	40
	228	12	240

T_{A^*B}	B		
A^*	196	12	208
	32	0	32
	228	12	240

(5.5)

Correcting the perturbed table on the left in (5.5), the moment estimator and the EM estimator yield

$\widehat{T}_{AB}$	B		
A	204.86	12.29	217.15
	23.14	-0.29	22.85
	228	12	240

$\widehat{T}_{AB}$	B		
A	204.86	12	216.86
	23.14	0	23.14
	228	12	240

respectively.

And regarding the perturbed table on the right in (5.5), the moment estimator

and the EM estimator yield

$\widehat{T}_{AB}$	B		
A	214.86	13.71	228.57
	13.14	-1.71	11.43
	228	12	240

$\widehat{T}_{AB}$	B		
A	214.86	12	226.86
	13.14	0	13.14
	228	12	240

respectively.

5.4 Two Estimators Compared

In the case of a $I \times J$-table with data perturbed by PRAM, we have two estimators to estimate the original table: the EM estimator and the moment estimator. How are these estimators related and what are their properties?

The moment estimator is unbiased: $I\!E[\widehat{T}_{AB}] = T_{AB}$.

The EM estimator yields the maximum likelihood estimate (MLE). The MLE can be useful because with $\widehat{\theta}$ the MLE of θ, $g(\widehat{\theta})$ is the MLE of $g(\theta)$, on condition that g^{-1} exists. In Chapter 4 this property turned out to be useful in estimating the relative risk and the odds ratio.

In the case of a 2×1-table the two estimators can be compared easily: they provide the same value as long as the moment estimator does not yield negative cell frequencies. In the following this will be explained.

The EM algorithm searches the maximum of (5.4) which can also be formulated as

$$\begin{aligned}\log l^*(a_i^*, \phi) &= \log\left(p_{00}\phi + p_{10}(1-\phi)\right)\sum_{i=1}^{N}(1-a_i^*) + \log\left(p_{01}\phi + p_{11}(1-\phi)\right)\sum_{i=1}^{N} a_i^* \\ &= T_{A^*}(0)\log\left(p_{00}\phi + p_{10}(1-\phi)\right) + T_{A^*}(1)\log\left(p_{01}\phi + p_{11}(1-\phi)\right).\end{aligned}$$

Putting

$$\frac{\partial}{\partial\phi}\log l^*(a_i^*, \phi) = 0,$$

results in the root

$$\phi_0 = \frac{(p_{10}-p_{00})p_{11}T_{A^*}(0) + (p_{11}-p_{01})p_{10}T_{A^*}(1)}{(T_{A^*}(0)+T_{A^*}(1))(p_{01}p_{00} - p_{10}p_{01} - p_{11}p_{00} + p_{11}p_{10})}. \tag{5.6}$$

When we write out the moment estimator

$$\widehat{T}_A = (P_A^{-1})^t T_{A^*},$$

we get the following estimate of ϕ

$$\frac{T_{A^*}(0)p_{11} - p_{10}T_{A^*}(1)}{(p_{00}p_{11} - p_{01}p_{10})(T_{A^*}(0) + T_{A^*}(1))}. \tag{5.7}$$

It turns out that (5.6) is equal to (5.7).

It should be decided whether $\log l^*(a_i^*, \phi)$ has a maximum or a minimum in ϕ_0, or if ϕ_0 is a point of inflection. In this case $\log l^*(a_i^*, \phi)$ is a sum of two logarithms which depend on the values of p_{ij}. Because the differential of $\log l^*(a_i^*, \phi)$ is given by

$$\frac{\partial}{\partial \phi} \log l^*(a_i^*, \phi) = T_A(0)\frac{p_{00} - p_{10}}{p_{00}\phi + p_{10}(1-\phi)} + T_A(1)\frac{-p_{11} + p_{10}}{p_{11}(1-\phi) + p_{10}\phi},$$

we see that when $p_{00} - p_{10} > 0$ and $-p_{11} + p_{10} < 0$, say condition (i), $\log(a_i^*, \phi)$ is for $\phi \in [0, 1]$ a sum of an increasing logarithm and a decreasing logarithm. When the diagonal is dominant in matrix P_A, condition (ii), the asymptotes of these logarithms lie outside [0,1]. Because P_A is a PRAM matrix, conditions (i) and (ii) are fulfilled and $\log l^*(a_i^*, \phi_0)$ is a maximum.

When $\phi_0 \in [0, 1]$, the moment estimator and the EM algorithm provided the same estimate (see also Ghaudhuri and Mukerjee (1988), app. A1.3). But when $\phi_0 < 0$, the moment estimator yields ϕ_0 while the EM algorithm, which searches for a maximum within the parameter space, yields $\widehat{\phi} = 0$.

The general case of a $I \times J$-table in which the variables are subject to misclassification or PRAM is not clear. Bourke and Moran (1988), Chen (1989) and Schwartz (1985, app. A) maintain that the moment estimator yields the MLE as long as negative cell frequencies do not occur, but proof of this statement is not provided. Out of caution we formulate the statement as a conjecture. We hope that proof will be provided by future research.

Conjecture *Let T_A be the frequency table of the categorical variable A and let PRAM be applied to the scores on A using matrix P_A. The moment estimator of the original table of frequencies given by $\widehat{T}_A = (P_A^{-1})^t T_{A^*}$, where T_{A^*} is the frequency table of the perturbed scores, is the maximum likelihood estimator of T_A as long as $\widehat{T}_A$ does not contains negative frequencies.*

5.5 Conclusion

When the original contingency table has to be estimated, using the EM algorithm is laborious and the algorithm cannot compete with the user-friendliness of the

moment estimator, despite the advantage concerning negative estimated cell frequencies. Furthermore, we did not give the EM algorithm in this chapter a firm mathematical basis. The choice of the starting point is rather intuitive and the faith we put in convergence to the global maximum is partly based on tests in which the EM algorithm yielded the same solution as the moment estimator.

The conjecture regarding the two estimators is important, since, if it is right, it provides a important property of the moment estimator: when the estimated cell frequencies are nonnegative, the moment estimator yields the maximum likelihood estimate.

This chapter can serve as an introduction to other applications of the EM algorithm. In the following chapters we will see that the EM algorithm can be useful in situations where data have been perturbed by PRAM.

Chapter 6

Loglinear Analysis

6.1 Introduction

Loglinear analysis is a method to explore relations in contingency tables. With the knowledge of those relations the most simple model is determined which describes the data satisfactory. The principles of loglinear analysis can be found for instance in Fienberg (1980) and in Agresti (1990, 1996).

Section 2 introduces the basic ideas of loglinear analysis using these references. Sections 2 and 3 consider the situation when one or two variable are perturbed by PRAM. In section 2 the moment estimator is used to adjust the analysis, in section 3 the EM algorithm is used. Section 4 is the conclusion of this chapter.

6.2 Standard Model

Although the merits of loglinear analysis become clear with tables of three or more dimensions, the 2×2 table can serve as an introduction. Consider the contingency table

		B		
		1	2	
A	1	π_{11}	π_{12}	π_{1+}
	2	π_{21}	π_{22}	π_{2+}
		π_{+1}	π_{+2}	π_{++}

where π_{ij} is the cell probability of cell (i, j), $\pi_{i+} = \pi_{i1} + \pi_{i2}$, $\pi_{+j} = \pi_{1j} + \pi_{2j}$ and $\pi_{++} = \sum_{i,j} \pi_{ij} = 1$. The analysis of the observed cell frequencies requires assumptions about the random mechanism that generated the data. It is common

to assume multinomial sampling: a fixed sample of size n and cross-classification of each member of the sample according to its values for the underlying variable A and B.

The basic model is the model of independence: $\pi_{ij} = \pi_{i+}\pi_{+j}$. This model states that the values of A are independent of the values of B. The expected cell frequencies under this model are given by

$$m_{ij} = n\pi_{i+}\pi_{+j}. \tag{6.1}$$

By taking the logarithm we get an additive structure:

$$\log m_{ij} = \log n + \log \pi_{i+} + \log \pi_{+j},$$

which is equivalent to

$$\log m_{ij} = \mu + \mu_{1(i)} + \mu_{2(j)}. \tag{6.2}$$

With $I = 2$ and $J = 2$, the parameters in (6.2) are given by:

$$\mu = \frac{1}{IJ} \sum_i \sum_j \log m_{ij} \tag{6.3}$$

$$\mu + \mu_{1(i)} = \frac{1}{J} \sum_j \log m_{ij} \tag{6.4}$$

$$\mu + \mu_{2(j)} = \frac{1}{I} \sum_i \log m_{ij} \tag{6.5}$$

(Note that $\mu \neq \log n$.) Because $\mu_{1(i)}$ and $\mu_{2(i)}$ represent deviations from the mean μ,

$$\sum_i \mu_{1(i)} = \sum_j \mu_{2(j)} = 0. \tag{6.6}$$

This is not a deduction from (6.3) to (6.5) but follows from the parameterization in (6.2): 5 parameters is to much for a 2×2 table, but when (6.6) holds, the model has in fact 3 parameters and is identifiable. (Other choices of parameters are possible. For instance, in (6.2) we can put $\mu_{1(2)} = 0$ and $\mu_{2(1)} = 0$.)

When independence is not the case, the model can be extended by interaction terms:

$$\log m_{ij} = \mu + \mu_{1(i)} + \mu_{2(j)} + \mu_{12(ij)}.$$

Where

$$\mu_{12(ij)} = \log m_{ij} - \frac{1}{J} \sum_j \log m_{ij} - \frac{1}{I} \sum_i \log m_{ij} + \frac{1}{IJ} \sum_i \sum_j \log m_{ij},$$

and the $\mu_{12(ij)}$ satisfy

$$\sum_i \mu_{12(ij)} = \sum_j \mu_{12(ij)} = 0.$$

This is the *saturated model* when the table is 2×2. It fits the data precisely.

We give an example. Assume $n = 164$ and we observe

	B		
A	32	11	43
	86	35	121
	118	46	164

Let $T_{AB}(i,j)$ denote the observed cell frequencies in cell (i,j) and $T_{AB}(i,+) = T_{AB}(i,1) + T_{AB}(i,2)$. Because of (6.1), we calculate the expected cell frequencies under the model of independence by

$$\widehat{m}_{ij} = \frac{T_{AB}(i,+)T_{AB}(+,j)}{n}.$$

This results in the following table:

	B	
A	30.94	12.06
	87.06	33.94

With $\widehat{m}_{ij}$ we can estimate the mean $\widehat{\mu} = 3.4783$ and the deviations from the mean $\widehat{\mu}_{1(1)} = -0.5173 = -\widehat{\mu}_{1(2)}$ and $\widehat{\mu}_{2(1)} = 0.4710 = -\widehat{\mu}_{2(2)}$.

Once we have estimated the expected values under a certain loglinear model, we can check the goodness-of-fit of the model using either of the following statistics:

$$X^2 = \sum \frac{(\text{Observed-Expected})^2}{\text{Expected}} \tag{6.7}$$

$$G^2 = \sum (\text{Observed}) \log\left(\frac{\text{Observed}}{\text{Expected}}\right),$$

where the summation in both cases is over all cells in the table. If the model fitted is correct and the total sample size is large, both X^2 and G^2 have approximate χ^2 distributions with degrees of freedom

$$df = \#\text{cells - }\#\text{parameters fitted.}$$

X^2 and G^2 are asymptotically equivalent.

In our example: $X^2 = 0.1755$ and $G^2 = 0.1777$, where $df = 4 - 3 = 1$. So the model of independence fits the data.

In the following we use vector notation. Let π be the vector of the cell probabilities and $\mathbf{p}$ the vector of the sample proportions. The saturated model has $\widehat{\pi} = \mathbf{p}$. Let $\mathrm{Cov}(\mathbf{p})$ denote the covariance matrix. Under multinomial sampling one has

$$\mathrm{Cov}(\mathbf{p}) = \left[\mathrm{Diag}(\pi) - \pi\pi^t\right]/n.$$

Where $\mathrm{Diag}(\pi)$ has the elements of π on the main diagonal. In our example, using $\mathrm{Cov}(n\mathbf{p}) = n^2\mathrm{Cov}(\mathbf{p})$, the standard errors of the observed cell frequencies are given by

SEs, saturated model	B	
A	5.08	3.20
	6.40	5.25

(6.8)

When we consider covariances of an unsaturated model we use the following form of loglinear models:

$$\log \mathbf{m} = X\mu.$$

Where X is a model matrix containing known constants and μ is a column vector of parameters. In our example, the independence model, $\log m_{ij} = \mu + \mu_{1(i)} + \mu_{2(j)}$:

$$\begin{pmatrix} \log m_{11} \\ \log m_{12} \\ \log m_{21} \\ \log m_{22} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & -1 \\ 1 & -1 & 1 \\ 1 & -1 & -1 \end{pmatrix} \begin{pmatrix} \mu \\ \mu_1 \\ \mu_2 \end{pmatrix}.$$

For instance, $\log m_{12} = \mu + \mu_1 - \mu_2$. Where $\mu_1 = \mu_{1(1)} = -\mu_{1(2)}$ and $\mu_2 = \mu_{2(1)} = -\mu_{2(2)}$. (When the model describes a $I \times J$ table, with $I > 2$ or $J > 2$, an other choice of the parameters is necessary in order that X is non-singular. This can be done by setting parameters equal to zero.)

For the unsaturated model, the estimated covariance matrix of the estimated cell probabilities equals

$$\widehat{\mathrm{C}}\mathrm{ov}(\widehat{\pi}) = \widehat{\mathrm{C}}\mathrm{ov}(\mathbf{p})X\left(X^t\widehat{\mathrm{C}}\mathrm{ov}(\mathbf{p})X\right)^{-1}X^t\widehat{\mathrm{C}}\mathrm{ov}(\mathbf{p}). \tag{6.9}$$

In our example:

SEs, indep. model	B	
A	4.43	2.03
	5.89	4.62

(6.10)

So far for the standard model.

6.3 Loglinear Analysis and the Moment Estimator

How should loglinear analysis be adjusted when the data are perturbed by PRAM?

When multinomial sampling is assumed, we can work as follows. Firstly, the frequencies of the original table are estimated by either the moment estimator or the EM estimator, and the covariance matrix of this PRAM correction is determined, say Cov_1.

Secondly, we choose a loglinear model in the standard way which fits the estimated original cell frequencies.

Thirdly, we estimate the covariance matrix of the frequencies in the saturated model using the estimated original frequencies, say Cov_2.

Lastly, because PRAM is independent of the sample model, we can simply sum Cov_1 and Cov_2 and take the sum as $\widehat{C}ov(\mathbf{p})$. The covariances of reduced models can be computed by using $\widehat{C}ov(\mathbf{p})$ as in (6.9).

As an example we consider the situation discussed in the previous section:

T_{AB}	B	
A	32	11
	86	35

Assume that PRAM is applied non-differentially and independently to A and B using

$$P_A = \begin{pmatrix} 9/10 & 1/10 \\ 2/10 & 8/10 \end{pmatrix} \qquad \text{and} \qquad P_B = \begin{pmatrix} 9/10 & 1/10 \\ 1/10 & 9/10 \end{pmatrix},$$

respectively.

A possible observed table and its corresponding estimate of the original table:

$T_{A^*B^*}$	B^*	
A^*	47	17
	71	29

$\widehat{T}_{AB}$	B	
A	36.22	8.36
	90.78	28.64

(6.11)

The diagonal of the covariance matrix of this estimate, Cov_1, is (49.20, 23.79, 59.73, 34.32). The diagonal of the covariance matrix of the saturated loglinear model, Cov_2, is (28.22, 7.93, 40.53, 23.64). We take $\widehat{C}ov(\mathbf{p})$ as $Cov_1 + Cov_2$, which has the diagonal (77.42, 31.72, 100.26, 57.96) and yields the following standard errors:

SEs, satured model	B	
A	8.80	5.64
	10.01	7.61

We can compare these standard errors to the situation without PRAM, see table (6.8).

The expected frequencies under the model of independence and its standard errors:

$\widehat{T}_{AB}$, ind. model	B	
A	34.52	10.06
	92.48	26.94

SEs, ind. model	B	
A	7.25	2.64
	8.69	5.76

(6.12)

Again, we can compare the errors to the situation without PRAM, see (6.10). Using the X^2 statistic (6.7), we can show that the model of independence fits the estimated data.

6.4 Loglinear Analysis and the EM Algorithm

The second method to adjust loglinear analysis when data are perturbed by PRAM makes use of the EM algorithm. We illustrate this method with two variables A and B and corresponding perturbed variables A^* and B^*. We assume multinomial sampling.

As before, let π_{ij} denote the cell probabilities and $T_{A^*B^*}$ the observed cell frequencies. Suppose we want to fit model L between A and B.

We start with computing the expected cell frequencies under L, denoted by $T^L_{A^*B^*}$, using $T_{A^*B^*}$, then we compute the cell probabilities of $T^L_{A^*B^*}$ and use them as the initial values in the EM algorithm: $\pi^{(0)}_{ij}$.

At the E-step of the EM algorithm the expected value of the complete table between A, B and A^*, B^* is computed using

$$I\!E\left[T^{(v)}_{ABA^*B^*}(i,j,k,,l)\right] = \frac{\pi^{(v)}_{ij} I\!P(A^*=k,B^*=l|A=i,B=j)}{\sum_{i,j}\pi^{(v)}_{ij} I\!P(A^*=k,B^*=l|A=i,B=j)} I\!E\left[T^L_{A^*B^*}(k,l)\right]. \tag{6.13}$$

Where $\pi^{(v)}_{ij}$ is the current estimate of π_{ij}. An estimate $\widehat{T}^{(v)}_{ABA^*B^*}$ is determined by using in (6.13) realization $T^L_{A^*B^*}$ as an estimate of $I\!E\left[T^L_{A^*B^*}\right]$.The expected table of the original variables is obtained by

$$I\!E\left[T^{(v)}_{AB}(i,j)\right] = \sum_{l,k} I\!E\left[T^{(v)}_{ABA^*B^*}(i,j,k,l)\right].$$

At the M-step a new estimate $\pi^{(v)}_{ij}$ is obtained by fitting model L to $T^{(v)}_{AB}$ in a standard way.

The process is iterated until convergence.

Kuha and Skinner (1997, section 28.5.2) describe this method but do not use $T^L_{A^*B^*}$ in (6.13), they use $T_{A^*B^*}$ instead, which seems to be incorrect, because:

$$\begin{aligned}
&\mathbb{E}\left[T_{ABA^*B^*}(i,j,k,,l)\right] \\
&= \mathbb{P}(A=i,B=j,A^*=k,B^*=l)n \\
&= \frac{P(A=i,B=j,A^*=k,B^*=l)}{\mathbb{P}(A^*=k,B^*=l)}\mathbb{P}(A^*=k,B^*=l)n \\
&= \frac{\mathbb{P}(A^*=k,B^*=l|A=i,B=j)\mathbb{P}(A=i,B=j)}{\mathbb{P}(A^*=k,B^*=l)}\mathbb{P}(A^*=k,B^*=l)n
\end{aligned}$$

Assuming model L: $\mathbb{P}(A^*=k,B^*=l)n = \mathbb{E}\left[T^L_{A^*B^*}\right]$.

In the following example we work with the same data as in the previous section. We want to fit the independence model and due to the assumption that PRAM is applied non-differentially and independently, it follows that

$$\mathbb{P}(A^*=k,B^*=l|A=i,B=j) = \mathbb{P}(A^*=k|A=i)\mathbb{P}(B^*=l|B=j).$$

We have

$T_{A^*B^*}$	B^*	
A^*	47	17
	71	29

$T^L_{A^*B^*}$	B	
A	46.05	17.95
	71.95	28.05

Using the table on the right we determine initial values $\pi^{(0)}_{ij}$ by computing the corresponding cell probabilities. Within 25 iterations the EM algorithm yields the following table

$\widehat{T}^L_{AB}$	B	
A	34.52	10.06
	92.48	26.94

which is the same as (6.12).

To estimate to standard errors of this estimation of the frequencies, we use the bootstrap.

Step 1. A new set of 'perturbed' data is created, say $T^{L*}_{A^*B^*}$, using $\widehat{\pi}(i,j) = \widehat{T}^L_{AB}(i,j)/n$.

Step 2. The EM algorithm is applied to the bootstrap observed data $T^{L*}_{A^*B^*}$ and its output is denoted by $\widehat{\pi}^*$.

Step 1 and step 2 are repeated independently a number of times (say B) to give estimates $\widehat{\pi}_1^*$, $\widehat{\pi}_2^*$, ..., $\widehat{\pi}_B^*$. The bootstrap covariance matrix of $\widehat{\pi}^*$ can be approximated by the sample covariance matrix of these B bootstrap replications. To investigate the efficiency of the bootstrap, we applied the bootstrap method four times with $B = 500$. We give the estimated standard errors of the estimated frequencies which can be compared with the standard errors in (6.12).

Run	Standard errors per cell			
	(1,1)	(1,2)	(2,1)	(2,2)
1	7.172	2.878	8,745	5.654
2	7.636	2.946	9.096	6.167
3	7.040	2.920	8.777	5.391
4	7.161	2.939	8.885	5.447

When we take L the saturated model, we get the same estimate of the original table as in (6.11).

6.5 Conclusion

In Chapter 2 we already observed that statistical analysis assumes a certain distribution of the data and that estimates of the original data (needed because of the perturbations by PRAM) have different distributions than the original data.

In the case of loglinear analyses the data are assumed to be distributed according to the multinomial distribution. (Other models are also possible, we did not go into this.) When we estimate the original table and apply loglinear analyses as in section 3, we did not incorporate that the table which estimates the original table is not distributed multinomially. Strictly speaking, this is not correct, but for the time being we assumed that this causes no real problems in practice.

The same problem occurs in section 4 where we work with the EM algorithm. In the procedure suggested, loglinear analyses is applied several times to a table which is not distributed multinomially. The first time loglinear analysis is applied in the algorithm it concerns the observed table which is assumed to be a product of a multinomial distribution process and of a second distribution process because of PRAM.

In this chapter we did not prove mathematically that the suggested procedures are sound. Also, we only tested the procedures on small tables. We acknowledge therefore that further research is necessary, but nevertheless like to state our trust

in the procedures. We conjecture that regarding loglinear analysis the perturbation by PRAM can be corrected.

Chapter 7

Logistic Regression

7.1 Introduction

In many fields logistic regression has become the standard method of analysis in the case that the dependent variable is discrete, taking on two or more possible values, and the independent variables are either discrete or continuous. After an introduction to the logistic regression model, this chapter considers the situation in which either the dependent variable or an independent categorical variable is perturbed by PRAM and discusses methods to adjust the regression. The EM algorithm will be used in sections 4 and 5.

7.2 Standard Model

To discuss the standard logistic regression model, we use introductory literature as for example Hosmer and Lemeshow (1989) and Agresti (1996).

Concerning logistic regression it is common to call the dependent variable the *outcome variable*. Logistic regression is most frequently employed to model the relationship between a binary outcome variable and a set of independent variables. For ease of exposition we only consider this situation. (Hosmer and Lemeshow (1989, section 8.1) discuss the standard logistic regression model in the case the outcome variable is polytomous and show that this model is based on the same concepts which are used when the outcome variable is binary.)

Let Y be the outcome variable, y its realization, and let $\mathbf{x} = (x_0, x_1, .., x_p)^t$ denote the vector with p independent variables (continuous or interval scaled). In practice, Y can denote the presence or absence of a disease and $\mathbf{x}$ can be a vector containing

variables such as age group (interval scaled) or weight (continuous).

The form of the logistic regression model is given by

$$I\!E\,[Y|\mathbf{x}] = \frac{e^{\beta^t \mathbf{x}}}{1 + e^{\beta^t \mathbf{x}}},$$

where $\beta = (\beta_0, \beta_1, ..., \beta_p)^t$ is the parameter vector. We take $x_0 = 1$ which makes β_0 the intercept coefficient.

In the *linear* regression model, $y = I\!E\,[Y|\mathbf{x}] + \varepsilon$, it is common to assume that the error ε follows a normal distribution with mean zero and some variance that is constant across levels of the independent variable. Therefore the conditional distribution of the outcome variable given $\mathbf{x}$ will be normal with mean $I\!E\,[Y|\mathbf{x}]$, and a variance that is constant.

In the *logistic* regression model the outcome variable is binary. Without loss of generality we assume $y \in \{0, 1\}$ and find

$$I\!E\,[Y|\mathbf{x}] = 0 \cdot I\!P(Y = 0|\mathbf{x}) + 1 \cdot I\!P(Y = 1|\mathbf{x}) = I\!P(Y = 1|\mathbf{x}).$$

We introduce the notation $\pi(\mathbf{x}) = I\!P(Y = 1|\mathbf{x})$. So, the logistic regression model is given by $y = \pi(\mathbf{x}) + \varepsilon$. Note that given $\mathbf{x}$, the error ε has two possible values: since $y \in \{0, 1\}$ it must the case that $\varepsilon \in \{1 - \pi(\mathbf{x}), -\pi(\mathbf{x})\}$, and $I\!P(\varepsilon = 1 - \pi(\mathbf{x})|\mathbf{x}) = I\!P(Y = 1|\mathbf{x}) = \pi(\mathbf{x})$ and $I\!P(\varepsilon = -\pi(\mathbf{x})|\mathbf{x}) = I\!P(Y = 0|\mathbf{x}) = 1 - \pi(\mathbf{x})$. Thus, the error ε follows a distribution with mean $(1-\pi(\mathbf{x}))\pi(\mathbf{x}) + (-\pi(\mathbf{x}))(1-\pi(\mathbf{x})) = 0$ and variance equal to $\pi(\mathbf{x})(1-\pi(\mathbf{x}))$. That is, the conditional distribution of the outcome variable follows a Bernoulli distribution with probability given by the conditional mean, $\pi(\mathbf{x})$.

Therefore, given observations $(y_1, \mathbf{x}_1), (y_2, \mathbf{x}_2), ..., (y_n, \mathbf{x}_n)$ the likelihood function is given by

$$\begin{aligned} l(\beta) &= \prod_{y_i=0} I\!P(Y = 0|\mathbf{x}_i) \prod_{y_i=1} I\!P(Y = 1|\mathbf{x}_i) \\ &= \prod_{y_i=0} (1 - \pi(\mathbf{x}_i)) \prod_{y_i=1} \pi(\mathbf{x}_i). \end{aligned}$$

The principle of maximum likelihood states that we use as our estimate of β the value which maximizes the likelihood function. To bring that about, it is common to maximize the log likelihood function:

$$L(\beta) = \log\, l(\beta) = \sum_{i=1}^{n} \Big(y_i \log[\pi(\mathbf{x}_i)] + (1 - y_i) \log[1 - \pi(\mathbf{x}_i)] \Big).$$

Let x_{ik} indicate the k^{th} element of $\mathbf{x}_i = (x_{i0}, x_{i1}, .., x_{ip})^t$. To determine the maximum of $L(\beta)$, we put $\frac{\delta}{\delta\beta_k}L(\beta) = 0$, which yields

$$\sum_{i=1}^{n} x_{ik}(y_i - \pi(\mathbf{x}_i)) = 0.$$

For $k = 0, 1, ..., p$ this equation is nonlinear in β, the maximum of the log likelihood cannot be derived analytically. Numerical (iterative) procedures have to be employed to obtain the implied maximum likelihood estimate of β. Hence, the Newton-Raphson method is used to solve the equations. Let

$$S(\beta) = \left(\frac{\delta}{\delta\beta_0}L(\beta), \frac{\delta}{\delta\beta_1}L(\beta), ..., \frac{\delta}{\delta\beta_p}L(\beta)\right)^t.$$

The information matrix $I(\beta)$ is a $(p+1) \times (p+1)$ matrix and is made up of

$$I(\beta)_{kl} = -\frac{\delta^2}{\delta\beta_k\delta\beta_l}L(\beta) = \sum_{i=1}^{n} x_{ik}x_{il}\pi(\mathbf{x}_i)(1 - \pi(\mathbf{x}_i).$$

Starting with an initial value $\beta^{(0)}$ we compute the values $S(\beta^{(0)})$ and $I(\beta^{(0)})$ and the new estimate of β is

$$\beta^{(1)} = \beta^{(0)} + \left[I(\beta^{(0)})\right]^{-1} S(\beta^{(0)}).$$

This iterative procedure is repeated until it is clear that $\beta^{(0)}, \beta^{(1)}, \beta^{(2)}, ...$ converges. When this is the case and the difference between $\beta^{(n)}$ and $\beta^{(n+1)}$ for a certain n is small or imperceptible we take $\beta^{(n)}$ as the maximum likelihood estimate (MLE) of β. (Note that there is uncertainty whether this is indeed the MLE; Newton-Raphson does not always yield the global maximum.)

An important feature of the information matrix is that it provides us with estimates of the covariances. If the MLE of β is denoted by $\widehat{\beta}$, then the asymptotic covariance matrix is estimated by $\left[I(\widehat{\beta})\right]^{-1}$. These estimated covariances will enable us to test hypotheses about the different elements of $\widehat{\beta}$.

Several tests were performed to gain insight in the behaviour and the speed of the method. We also present the results because it makes a comparison possible with methods in later sections.

We estimated parameters for different data sets with observations of a binary outcome variable y and a continuous independent variable $\mathbf{x} = (x_0, x_1)$ with $x_0 = 1$ always and x_1 a continuous variable. It is not a surprise that the speed of the

method is dependent on the size of the data set, but it turned out that the number of iterations is not. In each test we started with an initial value $\beta^{(0)} = (0, 0)$. Three tests with $n = 500$ as the number of records and with different choices of variables showed that the iteration converges fast: two or three iterations were enough to establish a difference between two following estimates of less than 1/1000, that is $|\beta_0^k - \beta_0^{k+1}| + |\beta_1^k - \beta_1^{k+1}| < 1/1000$. Test with $n = 20$ and $n = 1365$, again with different variables also showed convergence after three iterations. De time needed between the iterations increases with the size of the data set. The test with $n = 1365$ for instance, took four times longer to finish then the tests with $n = 500$.

Once there is a fit of a particular logistic regression model, the process of assessment of the model begins. In the framework of this report it is sufficient to state that the significance of variables and the comparison between models are based on the likelihood functions of the different models. Hence, when the likelihood function is adjusted correctly to account for the perturbation, logistic regression can proceed in the usual way, as though no perturbation has occurred.

So far for the standard model.

7.3 A Perturbed Outcome Variable: Newton Raphson

Consider the situation in which realizations of a binary outcome variable Y are perturbed using the PRAM matrix

$$P_Y = \begin{pmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{pmatrix}.$$

In the following we apply a method which is suggested for data obtained by randomized response, see Maddala (1983). Let $(y_1^*, \mathbf{x}_1), (y_2^*, \mathbf{x}_2), ..., (y_n^*, \mathbf{x}_n)$ be the perturbed observations. In the logistic model with $\pi(\mathbf{x})$ as in the previous section it follows that

$$I\!P(Y_i^* = 1|\mathbf{x}) = p_{11} I\!P(Y_i = 1|\mathbf{x}) + p_{01} I\!P(Y_i = 0|\mathbf{x}) = p_{01} + (p_{11} - p_{01})\pi(\mathbf{x}_i)$$

and

$$I\!P(Y_i^* = 0|\mathbf{x}) = 1 - I\!P(Y_i^* = 1|\mathbf{x}) = p_{10} + (p_{00} - p_{10})(1 - \pi(\mathbf{x}_i)).$$

So that the observed likelihood function becomes

$$l^*(\beta) = \prod_{y_i^*=0} \Big(p_{10} + (p_{00} - p_{10})(1 - \pi(\mathbf{x}_i)) \Big) \prod_{y_i^*=1} \Big(p_{01} + (p_{11} - p_{01})\pi(\mathbf{x}_i) \Big). \quad (7.1)$$

With $L^*(\beta) = \log l^*(\beta)$ we obtain

$$\frac{\delta}{\delta\beta_k} L^*(\beta) = \sum_{i=1}^{n} \left((1 - y_i^*) \frac{(p_{10} - p_{00}) e^{\beta^t \mathbf{x}_i}}{(p_{10} e^{\beta^t \mathbf{x}_i} + p_{00})(1 + e^{\beta^t \mathbf{x}_i})} x_{ik} + y_i^* \frac{(p_{11} - p_{01}) e^{\beta^t \mathbf{x}_i}}{(p_{01} + p_{11} e^{\beta^t \mathbf{x}_i})(1 + e^{\beta^t \mathbf{x}_i})} x_{ik} \right) \tag{7.2}$$

and

$$\frac{\delta^2}{\delta\beta_k \delta\beta_l} L^*(\beta) =$$

$$\sum_{i=1}^{n} \left((1 - y_i^*) \frac{p_{10}p_{00} e^{\beta^t \mathbf{x}_i} - p_{00}^2 e^{\beta^t \mathbf{x}_i} - p_{10}^2 e^{3\beta^t \mathbf{x}_i} + p_{10}p_{00} e^{3\beta^t \mathbf{x}_i}}{(p_{10} e^{\beta^t \mathbf{x}_i} + p_{00})^2 (1 + e^{\beta^t \mathbf{x}_i})^2} x_{ik} x_{il} + y_i^* \frac{p_{01}p_{11} e^{\beta^t \mathbf{x}_i} - p_{01}^2 e^{\beta^t \mathbf{x}_i} - p_{11}^2 e^{3\beta^t \mathbf{x}_i} + p_{11}p_{10} e^{3\beta^t \mathbf{x}_i}}{(p_{01} + p_{11} e^{\beta^t \mathbf{x}_i})^2 (1 + e^{\beta^t \mathbf{x}_i})^2} x_{ik} x_{il} \right). \tag{7.3}$$

Note that these computations can be quite laborious when there are more than one independent variable, i.e., more than two parameters.

Using the derivatives (7.2) and (7.3), the estimate of the parameter vector β and its covariance matrix can be computed with the Newton-Raphson method as described in the previous section. The initial values for the iterative procedure can be determined by applying standard logistic regression using the perturbed data. The idea is that because the perturbation is small, these initial values are more likely to be close to the 'true' β's than random initial values.

Again, several tests were performed in which we used the same data sets as in the previous section, be it with perturbed outcome variables. In each of the data sets PRAM is applied to the outcome variable using PRAM matrix given by

$$P_Y = \begin{pmatrix} 9/10 & 1/10 \\ 2/10 & 8/10 \end{pmatrix}.$$

We encountered no problems regarding convergence. Again, the iterations stopped when the difference between two following estimates was less than 1/1000 (see the previous section for this boundary). The test with $n = 500$ took considerable more time compared to the situation where the outcome is not perturbed. Roughly speaking, the number of iterations doubled and the iterations lasted four to five times longer.

As expected, the variances were larger compared to the situation where the outcome variable is not perturbed.

7.4 A Perturbed Outcome Variable: EM Algorithm

There is an another method to adjust logistic regression when PRAM has been applied to the outcome variable. Magder and Hughes (1997) show that it is possible to incorporate the information of a misclassification matrix concerning the outcome variable into the estimate of the parameters by using an EM algorithm. By exchanging the PRAM matrix for the misclassification matrix we tackle the situation where the outcome variable is perturbed by PRAM.

Let the notation be as in the previous sections and as in Chapter 5.

In the logistic regression model where the outcome variable is perturbed, the EM algorithm has a simple form. In the E-step we calculate $Q(\beta, \beta^{(k)})$:

$$\begin{aligned} Q(\beta, \beta^{(k)}) &= I\!E_Y \left[\log \left(l(\beta) | Y, \mathbf{x} \right) | y^*, \beta^{(k)} \right] \\ &= I\!E_Y \left[\sum_{i=1}^{n} Y_i \log[\pi(\mathbf{x}_i)] + (1 - Y_i) \log[1 - \pi(\mathbf{x}_i)] | y^*, \beta^{(k)} \right]. \end{aligned}$$

Since the original data log likelihood $\log \left(l(\beta) | y, \mathbf{x} \right)$ is linear in y_i, $Q(\beta, \beta^{(k)})$ is given by

$$Q(\beta, \beta^{(k)}) = \sum_{i=1}^{n} I\!E_{Y_i} \left[Y_i | y^*, \beta^{(k)} \right] \log[\pi(\mathbf{x}_i)] + 1 - \left(I\!E_{Y_i} \left[Y_i | y^*, \beta^{(k)} \right] \right) \log[1 - \pi(\mathbf{x}_i)].$$

In the logistic regression model y_i is discrete and since $I\!E_{Y_i} \left[Y_i | y_i^*, \beta^{(k)} \right]$ is not discrete, it is not possible to use $I\!E_{Y_i} \left[Y_i | y_i^*, \beta^{(k)} \right]$ as a substitution of y_i and apply standard software. A solution to this problem is to work with weights: the idea is to compensate for PRAM by performing standard logistic regression considering each individual as both having the property $(Y = 1)$ and not having the property $(Y = 0)$ with weights determined by the probability that the individual originally has the property given the perturbed data.

Of course, the probability of having the property depends in part on the value of the logistic regression parameters, therefore these probabilities have to be recalculated after the parameters are estimated. This is precisely the essence of the EM algorithm: estimating the probabilities and creating the new data set with weights is the E-step, and maximizing $Q(\beta, \beta^{(k)})$ is the M-step.

One has

$$I\!P(Y = 1 | Y^* = 1, \mathbf{x}) = \frac{I\!P(Y = 1, Y^* = 1 | \mathbf{x})}{I\!P(Y^* = 1 | \mathbf{x})} = \frac{I\!P(Y = 1 | \mathbf{x}) p_{11}}{I\!P(Y = 1 | \mathbf{x}) p_{11} + I\!P(Y = 0 | \mathbf{x}) p_{01}}$$

and similarly

$$I\!P(Y = 1|Y^* = 0, \mathbf{x}) = \frac{I\!P(Y = 1|\mathbf{x})p_{10}}{I\!P(Y = 1|\mathbf{x})p_{10} + I\!P(Y = 0|\mathbf{x})p_{00}}.$$

To find the maximum likelihood estimate of β, proceed as follows: with an initial estimate of β, $I\!P(Y_i = 1|Y_i^* = 1, \mathbf{x}_i)$ and $I\!P(Y_i = 1|Y_i^* = 0, \mathbf{x}_i)$ are calculated. Standard logistic regression is then performed with each individual included as both having the property and not having the property with corresponding weights: when $y_i^* = 1$ create a record of individual i with $y = 1$ with weight $I\!P(Y_i = 1|Y_i^* = 1, \mathbf{x}_i)$ and create a second record of individual i with $y = 0$ with weight$1 - I\!P(Y_i = 1|Y_i^* = 1, \mathbf{x}_i)$, when $y_i^* = 0$ work with $I\!P(Y_i = 1|Y_i^* = 0, \mathbf{x}_i)$ and $1 - I\!P(Y_i = 1|Y_i^* = 0, \mathbf{x}_i)$ respectively. Standard logistic regression on the new records yields an updated estimate of β.

As in the previous section: a good choice for the initial estimate of β is the $\widehat{\beta}$ which is the result of standard logistic regression on the perturbed data. Because the perturbation by PRAM is not large, this value should provided a good starting point for the EM algorithm.

Contrary to the Newton-Raphson method, the EM algorithm does not provide a estimation of the covariance matrix as a by-product of the parameters estimates. In the situation of logistic regression it is possible to compute the information matrix, but the idea of using the EM algorithm is to avoid this computation.

We therefore present two ways to estimate the covariance matrix. The first is general and is likely to be of use in other applications of the EM algorithm, the second is purpose-made for logistic regression with a perturbed outcome variable.

Firstly, McLachlan and Krishnan (1997, section 4.3) describe how in the independent and identically distributed case the empirical information matrix, $I_e(\beta)$, can be used as an approximation to the observed information matrix evaluated at the maximum likelihood estimate of the parameters.

In the logistic regression model, the realizations of the outcome variable are independent and identically distributed and the observed log likelihood can be expressed in the form

$$L(\beta) = \log\, l(\beta) = \sum_{j=1}^{n} \log l_j(\beta).$$

We can write the score vector $S(\beta)$ as

$$S(\beta) = \sum_{j=1}^{n} s_j(\beta) = \sum_{j=1}^{n} \frac{\delta}{\delta\beta} \log l_j(\beta).$$

On evaluation at $\beta = \widehat{\beta}$, the empirical information matrix is given by

$$I_e(\widehat{\beta}) = \sum_{j=1}^{n} s_j(\widehat{\beta}) s_j^t(\widehat{\beta}).$$

So when PRAM is applied to the outcome variable and the observed log likelihood is taken to be the logarithm of (7.1), we can use $I_e(\widehat{\beta})$ as an approximation of $I(\widehat{\beta})$.

Secondly, Magder and Hughes (1997) provide a way to estimate the covariance matrix when the outcome variable of the logistic regression model is misclassified.

Let

$$I(\beta) = \sum_{i=1}^{n} I(\beta)_i,$$

where $I(\beta)_i$ is the contribution of the ith subject to the information matrix. For standard logistic regression,

$$I(\beta)_i = \mathbf{x}_i \mathbf{x}_i^t (\widehat{\pi}(\mathbf{x}_i)(1 - \widehat{\pi}(\mathbf{x}_i))$$

When the outcome variable is perturbed the information matrix is corrected as follows

$$I(\beta)_i = \mathbf{x}_i \mathbf{x}_i^t \left(\widehat{\pi}(\mathbf{x}_i)(1 - \widehat{\pi}(\mathbf{x}_i) - \widehat{Y}_i(1 - \widehat{Y}_i) \right),$$

where $\widehat{Y}_i = I\!P(Y = 1 | Y^* = 1, \mathbf{x})$ when $y_i^* = 1$ and $\widehat{Y}_i = I\!P(Y = 1 | Y^* = 0, \mathbf{x})$ when $y_i^* = 0$.

We tested the EM algorithm and used SPSS to execute standard logistic regression with weighted records. The same perturbed data sets as in the previous section were used. The EM algorithm works fine, although the convergence is slower: more iterations are needed, roughly two times more, compared to the Newton-Raphson method. The point estimates of the parameters as provided by the EM algorithm did not differ significantly from the results in the previous section. There are however small differences in the estimation of the variance.

Three methods to estimate the variance were compared: using the information matrix as in the previous section (i), using the empirical information matrix (ii) and using the estimate given by Magder and Hughes (1997) (iii).

Although the differences were small between the variances estimated by the three methods, they were distinctive: method (i) yields larger variance than method (iii), which again yields larger variance than method (ii).

The method which uses the EM algorithm is user-friendly. The recipient of the perturbed data has to implement the EM algorithm, but can use the standard software for logistic regression within each iteration. It is not necessary to construct and maximize a likelihood function as in the method described in the previous section, which can be hard work when there are several independent variables.

7.5 A Perturbed Independent Variable

Next we consider the situation in which PRAM has been applied to a categorical independent variable. In the logistic regression model we have to distinguish between interval scale variables (e.g. rank of income, appraisal) and nominal scaled variables (e.g. race, religious affiliation). When the categorical variable is interval scaled the variable can be included in the model directly, when the categorical variable is nominal scaled we should use dummy variables.

We consider the model with a binary outcome variable Y and a binary independent variable X with categories zero and 1. In this simple model the distinction between interval or nominal scaled makes no difference. For sake of generality we write $\mathbf{x}_i = (x_{i0}, x_{i1})^t = (1, x_i)^t$ and $\beta = (\beta_0, \beta_1)^t$.

Let X be perturbed to X^* using the PRAM matrix

$$P_X = \begin{pmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{pmatrix}.$$

Because the complete-data log likelihood is not linear in $\mathbf{x}_i$ the simple form of the EM algorithm as used in the previous section can not be applied. A possible alternative is to estimate the conditional expectation of the complete-data log likelihood, that is $Q(\beta, \beta^{(k)})$, in the E-step by Monte Carlo simulation and maximize this estimation in the M-step. An EM algorithm where the E-step is executed by Monte Carlo is known as a *Monte Carlo EM* (MCEM) (McLachlan and Krishnan, 1997). In general this algorithm can be use when the E-step is complex and does not admit a closed-form solution to $Q(\beta, \beta^{(k)})$.

In the case of logistic regression where X is perturbed to X^* we have the following:

$$\begin{aligned} Q(\beta, \beta^{(k)}) &= I\!E_X \left[\log \left(l(\beta) | y, \mathbf{X} \right) | \mathbf{x}^*, \beta^{(k)} \right] \\ &= I\!E_X \left[\sum_{i=1}^{n} y_i \log \left(\pi(\mathbf{X}_i) \right) + (1 - y_i) \log \left(1 - \pi(\mathbf{X}_i) \right) | \mathbf{x}^*, \beta^{(k)} \right]. \end{aligned}$$

To execute Monte Carlo we need the distribution of X_i given X_i^*, Y_i and $\beta^{(k)}$:

$$I\!P(X_i = h | X_i^* = j, Y = k, \beta^{(k)})$$

$$= \frac{I\!P(Y_i = k, X_i = h, X_i^* = j | \beta^{(k)})}{I\!P(Y_i = k, X_i^* = j | \beta^{(k)})}$$

$$
\begin{aligned}
&= \frac{I\!P(Y_i = k|X_i = h, X_i^* = j, \beta^{(k)}) I\!P(X_i = h, X_i^* = j)}{I\!P(Y_i = k, X_i^* = j, X_i = 1|\beta^{(k)}) + I\!P(Y_i = k, X_i^* = j, X_i = 0|\beta^{(k)})} \\
&= \frac{I\!P(X_i^* = j|X_i = h) I\!P(Y_i = k|X_i = h, \beta^{(k)}) I\!P(X_i = h)}{\sum_{h=0}^{1} I\!P(Y_i = k|X_i^* = j, X_i = h, \beta^{(k)}) I\!P(X_i = h, X_i^* = j)} \qquad (7.4) \\
&= \frac{p_{hj} I\!P(Y_i = k|X_i = h, \beta^{(k)}) I\!P(X_i = h)}{\sum_{h=0}^{1} p_{hj} I\!P(Y_i = k|X_i = h, \beta^{(k)}) I\!P(X_i = h)}.
\end{aligned}
$$

Note that in the step to (7.4) we use that $I\!P(Y_i = k|X_i = h, X_i^* = j, \beta^{(k)}) = I\!P(Y_i = k|X_i = h, \beta^{(k)})$, since when X_i is known, X_i^* is not important regarding the probability of $Y_i = k$.

We do not have $I\!P(X_i = h)$, but we can estimate this probability by using the moment estimator. Let $T_{x^*} = (T_{x^*}(0), T_{x^*}(1))^t$ be the vector where $T_{x^*}(h)$ is the number of x_i^* with value h. If T_x is the vector with the original frequencies, then according to the moment method we estimate the original frequencies by

$$\widehat{T}_x = (P_X)^{-1} T_{x^*}.$$

Therefore we estimate $I\!P(X_i = h)$ by $\widehat{T}_x(h)/n$.

Let M be a positive integer. The MCEM runs as follows:

Monte Carlo E-step. On the kth iteration, draw $\mathbf{x}_i^1, ..., \mathbf{x}_i^M$ from the distribution of $\mathbf{X}_i$ given $\mathbf{x}_i^*$, y_i and $\beta^{(k)}$. Then approximate the Q-function by

$$\widehat{Q}(\beta, \beta^{(k)}) : \beta \mapsto \frac{1}{M} \sum_{m=1}^{M} \log\left(l(\beta^{(k)}|y, \mathbf{x}^m\right).$$

M-step. Maximize $\widehat{Q}(\beta, \beta^{(k)})$ over β to obtain $\beta^{(k+1)}$.

Because $\log\left(l(\beta^{(k)}|y, \mathbf{x}^m\right)$is itself a summation we can use standard logistic regression software to maximize $\widehat{Q}(\beta, \beta^{(k)})$: we need to maximize a complete-data log likelihood which is made up of a data set which is M times larger than the original data set:

$$\max_{\beta} \sum_{m=1}^{M} \sum_{i=1}^{n} y_i \log[\pi(\mathbf{x}_i^m)] + (1 - y_i) \log[1 - \pi(\mathbf{x}_i^m)].$$

In the MCEM algorithm the choice of M and the monitoring of convergence of the algorithm are somewhat difficult. It is recommended (McLachlan and Krishnan, 1997) to use small values of M in the initial stages of the algorithm and to increase M as the algorithm moves closer to convergence. As to monitoring convergence, it is recommended that the values of $\beta^{(k)}$ be tabulated against k and when the

convergence is indicated by the stabilization of the process with random fluctuations about a value $\widehat{\beta}$, the process may be terminated or continued with a larger value of M.

To give an indication of the choice of M: we tested the method on a data set with a binary outcome variable Y, a binary independent variable X with categories zero and 1 and $n = 500$. In the model we include an intercept coefficient. In the situation without perturbation: $\widehat{\beta}_0 = 1.29$ with $SE(\widehat{\beta}_0) = 0.154$ and $\widehat{\beta}_1 = -1.952$ with $SE(\widehat{\beta}_0) = 0.204$. After X was perturbed by applying PRAM with $p_{00} = 1 - p_{01} = 9/10$ and $p_{11} = 1 - p_{10} = 8/10$, we used MCEM to estimate the parameters. We started with initial values calculated by estimating the parameters as if no perturbation had been taken place.

With $M = 5$ and 25 EM iterations we notice after 5 iterations fluctuations about central values for $\widehat{\beta}_0$ and $\widehat{\beta}_1$. The rounded variances of these fluctuations are calculated using the last 15 iterations: 0.004 and 0.08 respectively. We continued with $M = 10$ and the variance became a lot less: 0.001 and 0.004.

7.6 Conclusion

The fact that the EM algorithm can be slow is a drawback. Although we can always use the bootstrap method to estimate the standard errors, this can be tedious because of the slowness of the algorithm. Another limitation is the choice of M in the MCEM algorithm. Most standard software will be able to handle a large M, but a large M will probably slow down the analyses.

The MCEM algorithm as discussed in the situation where PRAM is applied to an independent variable can also be used in the situation where PRAM is applied to the outcome variable. The advantage in that case is that logistic regression can be applied without weights.

Because of the general use of the MCEM algorithm, we can also tackle the situation when both the outcome and the independent variable are perturbed. In the E-step of the MCEM we can simple draw from the conditional distribution of the outcome variable and from the conditional distribution of the independent variable.

The main advantage of the EM algorithm in the context of logistic regression is that the recipient of the data perturbed by PRAM can use standard software to execute the regression. It this way the analyst can get round the required computations and implementation of the Newton Raphson method, which quickly becomes laborious when the number of parameters goes up.

As in the previous chapter regarding loglinear analysis, we did not prove math-

ematically that the suggested procedures are sound. The uncertainty lies in the maximization of the adjusted log likelihood function and the choice of the starting point of the maximization. (Alternatives to the adjustments of the likelihood function itself do not seem at hand.) These problems are also encountered in the standard logistic regression model; numerical maximization often does not provide certainty about global maximums versus local maximums. So, to conclude, in the situation with a variable perturbed by PRAM adjustment seems possible with the procedures suggested in this chapter. Although the maximization in this situation can be harder because of a more complex likelihood.

Conclusion

This report explains PRAM and discusses how statistical analysis can be adjusted when data are perturbed by PRAM. Adjustment of basic analysis of a 2×2-table (difference of proportions, relative risk and the odds ratio) is considered in Chapter 4 and loglinear analyses and logistic regression are considered in Chapters 6 and 7. The report shows that regarding these analyses, adjustments can be made in order to work with the perturbed data. Of course, this report has its limitations: we did not always consider extra variances (relative risk, odds ratio) and we only discussed simple models with few variables. Nevertheless, we hope to have showed that adjustment is often possible and, furthermore, we hope to have provided tools (in particular the EM algorithm) for further research concerning adjustment of statistical analyses.

As stated in the introduction of Chapter 1: we did *not* discuss the extent of randomness which the PRAM procedure needs to protect the data satisfactory. This randomness, that is, the transition probabilities that scores on certain variables change into different scores, should be determined before the microdata file is perturbed. Clearly, the benefit of adjustment of statistical analyses is related to the choice of the transition probabilities. When probabilities that original scores change into different scores are high, adjustment of analyses can not prevent that the extra variance due to PRAM makes the analyses unworkable. So, in the end, research into the choice of the PRAM matrix and research into the possibilities to adjust analyses should be combined and tests on real data sets should be performed. This is the first recommendation for future research.

Other subjects of future research are provided by some mathematical questions which are not yet solved. Most important is the conjecture in Chapter 4 regarding the differences between the moment estimator and the EM estimator. This is interesting because of the different properties of the estimators which we would like to combine, i.e., we would like to know when the unbiased moment estimator yields the same output as the maximum likelihood estimator. Another subject for research is the use of the bootstrap to estimate variances after the EM algorithm is applied.

In Chapter 4 and 5 the bootstrap is used and we would like to investigate if this use of the bootstrap is mathematically sound. Using the bootstrap in combination with the EM algorithm can provide a important tool when data perturbed by PRAM have to be analysed. A problem which is encountered in all analysis of data perturbed by PRAM is the fact that standard analyses assume certain models for the original data and that these models are strictly speaking not appropriate after the perturbations and estimation of the original data. For instance, the estimated 2×2-table of the original data does not have the same multinomial distribution as the original 2×2-table. It may be that this problem can be ignored in practice, but in view of mathematics, it is awkward.

There seems to be no end when it comes to different analyses which may be adjusted in order to be able to work with data perturbed by PRAM; each analysis which deals with categorical variables can be discussed. In the future it will be worthwhile to get in touch with disciplines as for instance psychometrics, sociometrics and econometrics. Contacts with these disciplines can outline research and may also provide new insights. As discussed in Chapter 3, problems regarding perturbed data are not new and in some situations (misclassification, randomized response and incomplete data) problems are close to the problems encountered when working with data perturbed by PRAM.

The final acceptance of PRAM as a method of statistical disclosure control depends not only on the possibility to produce safe microdata and the possibility to adjust analyses, but also on the user-friendliness of the adjustments.

References

Agresti, A. (1990), *Categorical Data Analysis*, New York: Wiley.

Agresti, A. (1996), *An Introduction to Categorical Data Analysis,* New York: Wiley.

Assakul, K., and Proctor, C.H. (1967), Testing Independence in Two-Way Contingency Tables With Data Subject to Misclassification, *Psychometrika*, **32**, pp 67-76.

Bourke, P.D., and Moran, M.A. (1988) Estimating Proportions from Randomized Response Data Using the EM Algorithm, *Journal of the American Statistical Association*, **83**, pp. 964-968.

Chaudhuri, A., and Mukerjee, R. (1988) *Randomized Response: Theory and Techniques*, New York: Marcel Dekker.

Chen, T.T. (1979), Analysis of Randomized Response as Purposively Misclassified Data, in *Proceedings of the Section on Survey Research Methods*, American Statistical Association, pp 158-163.

Chen, T.T. (1989), A Review of Methods for Misclassified Categorical Data in Epidemiology, *Statistics in Medicine*, **8**, pp 1095-1106.

Copeland, K.T., Checkoway, H., McMichael, A. J., and Holbrook, R. H. (1977) Bias Due to Misclassification in the Estimation of Relative Risk, *American Journal of Epidemiology*, **105**, pp 488-495.

Dempster, A. P., Laird, N.M., and Rubin, D.B. (1977) Maximum Likelihood from Incomplete Data via the EM algorithm, *Journal of the Royal Statistical Society*, **39**, pp 1-38.

De Wolf, P.-P., Gouweleeuw, J.M., Kooiman, P., and Willenborg L.C.R.J (1997) Reflections on PRAM, Research paper no. 9742, Voorburg/Heerlen: Statistics Netherlands.

Diebolt, J. (1999), Book review of *The EM Algorithm and Extensions* by G.J. McLachlan and T. Krishnan, in Mathematical Reviews on the Web, American Mathematical Society.

Fienberg, S.E. (1980) *The Analysis of Cross-Classified Data*, Cambridge: MIT Press.

Goldberg, J.D. (1975) The Effects of Misclassification on the Bias in the Difference Between Two Proportions and the Relative Odds in the Fourfold Table, *Journal of the American Statistical Association*, **70**, pp 561-567.

Gouweleeuw, J.M., Kooiman, P., Willenborg, L.C.R.J., and De Wolf, P.-P. (1998) Post Randomisation for Statistical Disclosure Control: Theory and Implementation, *Journal of Official Statistics*, **14**, pp 463-478.

Greenland, S. (1988) Variance Estimation for Epidemiologic Effect Estimates under Misclassification, *Statistical in Medicine*, **7**, pp 745-757.

Hagenaars, J. A. (1993) Loglinear Models With Latent Variables, Sage University Paper series on Quantitative Applications in the Social Sciences, 07-094, Newbury Park, CA: Sage.

Hosmer, D. W., and Lemeshow, S. (1989) *Applied Logistic Regression*, New York: Wiley.

Johnson, N. L., and Kotz, S. (1969) *Discrete Distributions*, New York: Wiley.

Kooiman, P., Willenborg, L.C.R.J., and Gouweleeuw, J.M. (1997), PRAM: a Method for Disclosure Limitation of Microdata, Research paper no. 9705, Voorburg/Heerlen: Statistics Netherlands.

Kotz, S., and Johnson, N.L. (1982) *Encyclopedia of Statistical Sciences*, New York: Wiley.

Kuha, J., and Skinner, C. (1997), Categorical Data Analysis and Misclassification,

in *Survey Measurement and Process Quality*, by L. Lyberg et al. (eds), New York: Wiley.

Little, R.J.A., and Rubin, D.B. (1987) *Statistical Analysis with Missing Data*, New York: Wiley.

Maddala, G.S. (1983), *Limited Dependent and Qualitative Variables in Econometrics*, Cambridge: Cambridge University Press.

Magder, L. S., and Hughes, J. P. (1997) Logistic Regression When the Outcome Is Measured with Uncertainty, *American Journal of Epidemiology*, **146**, pp. 195-203.

McLachlan, G. J., and Krishnan, T. (1997) *The EM Algorithm and Extensions*, New York: Wiley.

Netherlands Official Statistics (NOS) (1999) **14**, Voorburg /Heerlen: Statistics Netherlands.

Schwartz, J.E. (1985) The Neglected Problem of Measurement Error in Categorical Data, *Sociological Methods and Research*, **13**, pp 435-466.

Van der Heijden, P.G.M., Van Gils, G., Bouts, J., and Hox, J. (1998), A Comparison of RR, CASAQ, and Direct Questioning; Eliciting Sensitive Information in the Context of Social Security Fraud, *Kwantitatieve methoden*, **59**, pp 15-34.

Warner, S. (1965) Randomized Response: a Survey Technique for Eliminating Answer Bias, *Journal of the American Statistical Association*, **60**, pp 63-69.

Willenborg, L.C.R.J, and De Waal, A.G. (1996) *Statistical Disclosure Control in Practice,* Lecture Notes in Statistics **111**, New-York: Springer-Verlag.

Willenborg, L.C.R.J (1999) Optimization Models for PRAM Matrices, Research paper no. 9927, Voorburg/Heerlen: Statistics Netherlands.

Summary

This reports discusses the Post Randomisation Method (PRAM). PRAM was introduced in 1997 as a method for disclosure protection of microdata files. A microdata file consists of records and each record contains individual data of respondents. The PRAM procedure yields a new microdata file in which the scores on certain variables in the original file may be changed into different scores according to a prescribed probability mechanism. 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. The recipient of the perturbed data is informed about the probability mechanism which is used in order that he can adjust his statistical analysis and take into account the extra uncertainty caused by applying PRAM.

This report explains PRAM and discusses how statistical analysis can be adjusted when variables are perturbed by PRAM. Because PRAM always concerns categorical variables - variables with a finite number of values - we discuss mainly categorical data analysis. We do *not* discuss the extent of randomness which the PRAM procedure needs to protect the data satisfactory, instead it is assumed that the randomness of the PRAM procedure is known and provided in the form of a Markov matrix.

Although perturbation of categorical variables by applying PRAM is new and produces new problems for standard statistical analysis, it is possible that solutions to these problems can be found in existing methods which deal with similar perturbation problems such as data subject to misclassification, incomplete data, and data obtained by randomized response. The similarity between these situations: the scores which are missing or the scores which are only known via perturbed scores can be considered as values of stochastic variables which have to be estimated. An important advantage in the case of PRAM is that the probability mechanism used is known, which simplifies these methods.

In Chapter 4 we show that adjustment regarding the perturbation is possible in the case of simple analysis of a 2×2-table. Because of the attenuation produced

by applying PRAM regarding the analyses discussed, a conclusion concerning the *presence* of association is justified on the basis of the observed table alone. When the observed table contains a significant difference in proportion, the original table will contain an even larger difference. And the same applies to the relative risk, the odds ratio, and the Pearson chi-squared test: when the risk, the ratio, and chi-squared test is computed without adjustment for PRAM and the outcome suggests association, then there is certainly association in the original table.

In Chapter 5 we consider the EM algorithm which is currently used in different fields of applied mathematics and which is constructed to find maximum likelihood estimates. For instance, the algorithm can be used when data are incomplete or censored. When data are perturbed by PRAM the recipient of the data can consider a file which is twice the size of the original file and which consist both of the perturbed scores and of the original scores. This new file is incomplete since the original scores are not provided. Using the Markov matrix and the EM algorithm the analyst can determine maximum likelihood estimates for analyses regarding the original data. We used the EM algorithm for frequency estimation (Chapter 5), loglinear analysis (Chapter 6) and logistic regression (Chapter 7). Of course, techniques other than the EM algorithm are possible and some of them are also mentioned in this report. The disadvantage of the EM algorithm is its slowness, but advantages are numerical stability and user-friendliness due to the possibility to use standard software within the iterations of the algorithm.

Research concerning PRAM is still necessary. Some of the techniques discussed in this report demand a firmer mathematical basis. The use of the bootstrap in connection with the EM algorithm, for instance, should be considered more carefully. Furthermore, there is no end when it comes to different analyses which may be adjusted in order to be able to work with data perturbed by PRAM; each analysis which deals with categorical variables can be discussed. In the future it will be worthwhile to get in touch with disciplines as for instance psychometrics, sociometrics and econometrics. Contacts with these disciplines can outline research and may also provide new insights.

The final acceptance of PRAM as a method of statistical disclosure control depends not only on the possibility to produce safe microdata and the possibility to adjust analyses, but also on the user-friendliness of the adjustments.