# Analyzing Misclassified Data: Randomized Response and Post Randomization

ISBN 90-393-3788-8

# Analyzing Misclassified Data: Randomized Response and Post Randomization

Over de analyse van misgesclassificeerde data: randomized response
en post randomization

(met een samenvatting in het Nederlands)

**Proefschrift**

ter verkrijging van de graad van doctor
aan de Universiteit Utrecht
op het gezag van de Rector Magnificus, Prof. dr. W. H. Gispen,
ingevolge het besluit van het College voor Promoties
in het openbaar te verdedigen
op vrijdag 24 september 2004, des middags te 16.15 uur

door

**Arie Daniël Leo van den Hout**

geboren op 8 maart 1968, te Dirksland

Promotoren: Prof. dr. P.G.M. van der Heijden
Faculteit der Sociale Wetenschappen
Universiteit Utrecht

Prof. dr. ir. P. Kooiman
Faculteit der Economische Wetenschappen
Erasmus Universiteit Rotterdam

Dit proefschrift werd mede mogelijk gemaakt met financiële steun van de Nederlandse Organisatie voor Wetenschappelijk Onderzoek.

# Acknowledgements

I would like to thank my supervisors Peter van der Heijden and Peter Kooiman for discussion, encouragement, and the joint work that is presented in this Ph.D. thesis. Chapter 2 and 3 is joint work with Peter van der Heijden, Chapter 5 is joint work with Peter Kooiman.

Several persons helped me with their comments and suggestions. Some of their remarks will be used in future research, some are used in this book. I would like to thank Wicher Bergsma, Antonio Forcina, Jacques Hagenaars, Cor Kraaikamp, Gerty Lensvelt-Mulders, Don Rubin, Jeroen Vermunt, Leon Willenborg, and Peter-Paul de Wolf. I would especially like to thank Elsayed Elamir for the joint work that is presented in Chapter 6.

Most of the Ph.D. research was done at the Department of Methodology and Statistics, Faculty of Social Sciences, Utrecht University. I would like to thank my colleagues for a pleasant and inviting place to work. Special thanks are due to Olav Laudy for discussion and computer help.

I would like to thank Marije Altorf for assistance regarding the English language and for being around.

# Contents

# Chapter 1

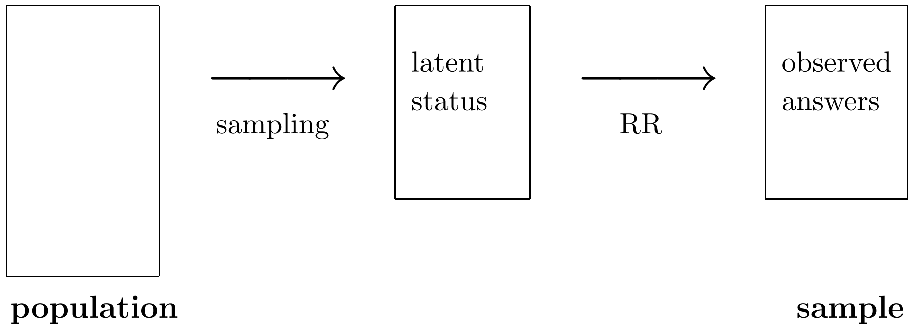
# Introduction

This book is about the analysis of randomized response data and the analysis of data that are subject to the post randomization method (PRAM). The following introduces randomized response and PRAM, provides an outline of the subsequent chapters, and describes important assumptions and choices that are made throughout the book.

## 1.1 Randomized Response

Randomized response (RR) is an interview technique that can be used when sensitive questions have to be asked and respondents are reluctant to answer directly (Warner 1965). Examples of sensitive questions are questions about alcohol consumption, sexual behavior or fraud. RR variables can be seen as misclassified categorical variables where conditional misclassification probabilities are known. The misclassification protects the privacy of the individual respondent.

The general scheme of RR is given by



Assume that a researcher wants to assess a sensitive item and that he uses a question to which the answer is either *yes* or *no*. After the sample is drawn from the population, the RR design is applied to the selected respondents. The latent status of the respondents with respect to the sensitive item are unknown. Observed data consist of observed answers after RR is applied.

A possible choice of a RR design is the forced response design (Boruch 1971). In this design, the respondent throws two dice after the sensitive question is asked. The outcome of the dice is hidden from the interviewer. If the outcome is 2, 3 or 4, the respondent answers *yes*. If the outcome 5, 6, 7, 8, 9 or 10, he answers according to the truth. If the outcome 11 or 12, he answers *no*.

The RR design can be seen as a misclassification design. Assume that the sensitive question concerns fraud and that respondent Z indeed committed fraud. When the question "Did you commit fraud?" is asked in a direct response situation, the truthful answer of Z is *yes*. Assume next that the forced response design is applied. When Z throws the dice and the outcome is 2, 3, 4, 5, 6, 7, 8, 9 or 10, he answers *yes* and he is correctly classified as a person that committed fraud. When the outcome is 11 or 12, the answer of Z is *no* and Z is misclassified as a person that did not commit fraud. The misclassification probability conditional on the fact that Z committed fraud can be computed using the distribution of the outcome of the dice and is given by

$$\mathbb{P}(\text{ Z is misclassified}|\text{ Z committed fraud}) = 1/12. \tag{1.1}$$

When Z did not commit fraud, the misclassification probability is given by

$$\mathbb{P}(\text{ Z is misclassified}|\text{ Z did not commit fraud}) = 1/6.$$

Since the interviewer does not know the outcome of the dice, the interviewer does not know whether the observed answer corresponds with the latent status of Z. In other words, an observed *yes* does not necessarily mean that Z committed fraud. Hence the privacy of Z is protected.

Probability (1.1) is rather small and one might wonder whether the respondent is satisfied with the privacy protection that is offered. However, Moriarty and Wiseman (1976) showed that respondents tend to overestimate (1.1) due to an inaccurate idea of the distribution of the outcome of the two dice.

More formally, let $X$ be the binary RR variable that models the latent status, $X^*$ the binary variable that models the observed answer, and *yes* $\equiv 1$ and *no* $\equiv 2$. Given the forced response design, the distribution of $X^*$ is the 2-component finite mixture given by

$$\mathbb{P}(X^* = x^*) = \sum_{k=1}^{2} \mathbb{P}(X^* = x^* | X = k)\mathbb{P}(X = k), \tag{1.2}$$

where $x^* \in \{1, 2\}$. The conditional probabilities $p_{jk} = \mathbb{P}(X^* = j|X = k)$ for $j, k \in \{1, 2\}$ are fixed by the forced response design and the known distribution of the sum of the two dice. Formulation (1.2) shows that RR variables can be seen as misclassified variables. The transition matrix of $X$ that contains the conditional misclassification probabilities $p_{jk}$ for $j, k \in \{1, 2\}$ is given by

$$\boldsymbol{P}_X = \begin{pmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{pmatrix} = \begin{pmatrix} 11/12 & 2/12 \\ 1/12 & 10/12 \end{pmatrix}.$$

Similar expressions hold for more that two RR variables or RR variables with more than two categories.

Other RR designs are possible. A second example is the design where the misclassification is based on drawing playing cards from two stacks (Kuk 1990). In this design, the misclassification is based on chosen distributions of the colors in the stacks. By choosing different distributions different misclassification probabilities can be determined.

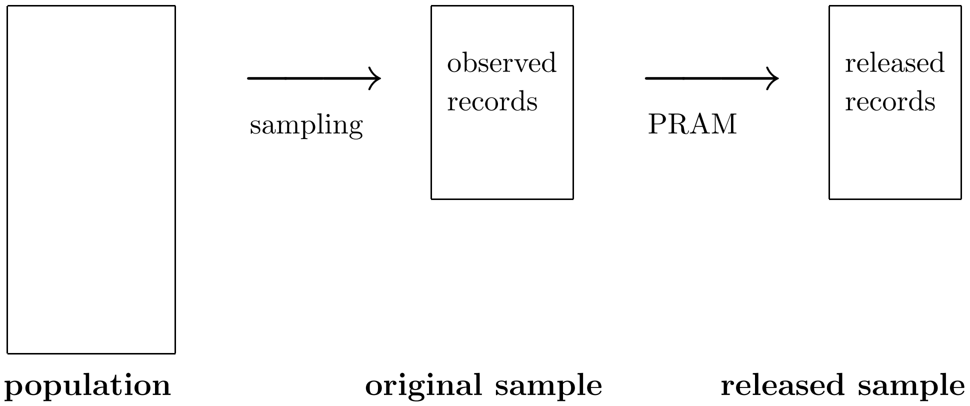
In recent years, RR techniques have been investigated and applied in the Netherlands. Van der Heijden, Van Gils, Bouts, and Hox (2000) compare two RR designs with face-to-face direct questioning. Boeije and Lensvelt-Mulders (2002) investigate compliance and non-compliance in RR surveys. Van Gils, Van der Heijden, Laudy, and Ross (2003) report about rule transgression with respect to social benefits. The transgression was investigated using the RR design by Boruch (1971). Elffers, Van der Heijden, and Hezemans (2003) use RR to study rule transgression for two Dutch instrumental laws. RR has also been studied outside the Netherlands. The monograph on RR by Chaudhuri and Mukerjee (1988) gives an overview of existing theory and techniques.

The basic idea of RR is that the perturbation induced by the misclassification design (in the first example, using the dice) protects the privacy of the respondent and that insight into the misclassification design (in the first example, the known distribution of the dice) can be used to analyze the data. A researcher who wants to apply RR should reflect on two issues. The first is about the choice of the RR design and efficiency. How much protection does the design offer? Will respondents understand the design? How expensive is the application? And, closely connected with cost, how many respondents are necessary? The second issue is connected with the first and is about the analysis of the RR data. It is obvious that one should take care of the misclassification due to the use of RR. How should that be done?

## 1.2 Post Randomization Method

The founder of RR suggested that the idea of RR can also be used to protect data that have already been collected (Warner 1971). The post randomization method (PRAM) was introduced by Kooiman, Willenborg and Gouweleeuw (1997) and can be seen as an application of RR where the misclassification is applied using a computer. PRAM is designed for the situation in which a statistical institute wants to release data to researchers outside the institute. When data are released, the privacy of the respondents should be protected. The field in statistics that studies the problems in this situation is called *statistical disclosure control.*

PRAM is a method for statistical disclosure control of microdata files. A microdata file is a data matrix where each row, called a record, corresponds to one respondent and where the columns correspond to the variables. PRAM can be applied to variables in the microdata file that are categorical and identifying. Identifying variables are variables that can be used to re-identify individuals represented in the data, e.g., Age, Gender or Ethnic Background. The misclassification of these kind of identifiers makes re-identification of individuals less likely. An essential aspect of PRAM is that the recipient of the misclassified data is informed about the misclassification probabilities. Using the probabilities he can adjust his analysis and take into account the extra uncertainty caused by applying PRAM. The general scheme of PRAM is given by



The need for statistical disclosure control of microdata files is illustrated by the following example. Assume that a general practitioner in the Netherlands is a respondent in a Dutch survey and that she was born in Bolivia. It is possible that this doctor is the only respondent in the sample that has the values (GP, Bolivia) of the combination of variables (Profession, Native Country). This means that her record might attract attention when the data are released without protection. It might be that a fellow doctor who happens to browse the sample file recognizes a former fellow student. Another more aggressive scenario is that someone tries to match records from the current survey to records from another survey in order to look for discriminating information. The intentions of such an intruder may be obscure, yet a statistical institute that releases data should take the possibility of such an attack into account. Hence the need for statistical disclosure control.

PRAM is not the only way to protect microdata against disclosure. Willenborg and De Waal (2001) discuss alternative methods such as global recoding and local suppression. At Statistic Netherlands, PRAM was introduced by Kooiman, Willenborg and Gouweleeuw (1997). Subsequent research is presented by De Wolf, Gouweleeuw, Kooiman, and Willenborg (1997), Gouweleeuw, Kooiman, Willenborg, and De Wolf (1998), and Van den Hout (1999). PRAM is also one of the methods discussed by Domingo-Ferrer and Torra (2001) who make a quantitative comparison of disclosure control methods for microdata.

The two main issues concerning PRAM are comparable to the issues discussed in the preceding section on RR: First, how to choose the misclassification probabilities in order to make the released microdata safe? Second, how should statistical analysis be adjusted in order to take into account the misclassification? It is not difficult to protect the privacy of respondents by perturbing data, the problem is to perturb the data in such a way that the privacy is protected *and* the released data are useful for research.

## 1.3 Outline of the Subsequent Chapters

The basis of the subsequent chapters consists of five papers that are written for individual publication. This structure has the advantage that the chapters are self-contained, a disadvantage is that there is some overlap in the discussion, especially in the introduction of the chapters. The outline is as follows.

Given the chosen conditional misclassification probabilities in the RR design or the PRAM design, Chapter 2 discusses the estimation of proportions and the estimation of the odds ratio. The methods in this chapter can be used for questions like: What is the percentage of persons that committed fraud? Or: Is there an association between committing fraud and gender? Moment estimates and maximum likelihood estimates of the proportions are compared and it is proven that they are the same in the interior of the parameter space. Special attention is paid to the possibility of boundary solutions.

Chapter 3 can be seen as a generalization of the discussion in Chapter 2. The method in Chapter 3 can be used to investigate more dimensional association patterns. For example, a study is possible regarding the association between committing fraud, gender and population size of the place of residence. The chapter describes the fitting of loglinear models to RR data and PRAM data. The misclassification is described by a latent class model. Since a latent class model is a loglinear model with one or more categorical latent variables, it is possible to investigate relations between misclassified variables. Methods to fit loglinear models for the latent table are discussed, including an EM algorithm. Again, attention is paid to problems with boundary solutions. In an example, RR data are analyzed which were collected using the RR design by Kuk (1990).

Chapter 4 also discusses the fitting of loglinear models to RR data. There is some overlap with Chapter 3, but the situation is slightly different since a different RR design is used and the use of RR is a factor in a 2 × 2 factorial design. Some of the respondents used the forced response design (Boruch 1971), others did not. The likelihood for this estimation problem is formulated and it is shown that also in this situation latent class software can be used to analyze the data. An example including a power analysis is discussed. This chapter shows the versatility of the modeling that is presented in Chapter 3.

Chapter 5 is about maximum likelihood estimation of the iid normal linear regression model when some of the independent variables are subject to RR or PRAM. An example of an application is the investigation of the relation between misclassified independent variables Age, Gender, and Ethnic Background and non-perturbed dependent variable Income. The likelihood of the linear regression model with misclassified independent variables is derived and a fast and straightforward EM algorithm is developed to obtain maximum likelihood estimates. The basis of the algorithm consists of elementary weighted least squares steps.

The discussion in Chapter 6 concerns the application of PRAM. The chapter discusses two variants of the initial idea of PRAM regarding the information about the misclassification that is given along with the released data. The first variant concerns calibration probabilities and the second variant concerns misclassification proportions. It is shown that the distinction between the univariate case and the multivariate case is important. In addition, the chapter discusses two measures for disclosure risk when PRAM is applied.

## 1.4 Assumptions and Choices

The following describes the main assumptions and choices that are made throughout the book.

- The emphasis of the book is on analysis of misclassified data. How should we adjust standard statistical models in order to take into account the misclassification induced by either RR or PRAM? Chapter 6, which explicitly discusses PRAM, is an exception, since one of its topics is the relation between the choice of the misclassification probabilities and the protection that is offered.

- Throughout the book the assumption is that respondents follow the RR design. This is a rather strong assumption. It is easy to imagine scenarios where respondents do not follow the design, either because they do not understand it or because they do not trust the protection offered. At the end of Chapter 3 this topic is briefly discussed.

- This book contains some real RR data examples. It is assumed that the research methods underlying the RR data are proper. Issues as sampling, questionnaires, interviewing, and data editing are not discussed.

- Since this book concerns applied statistics an effort is taken to make the discussion accessible. Especially Chapters 2 and 3 go at some length to introduce concepts, methods and solutions. Chapters 3 and 4 contain computer programs that can be used to analyze the misclassified data. Another way in which the discussion is made more accessible is the linking of RR and PRAM to some well-known issues in social statistics such as the analysis of incomplete data and latent class analysis.

# Chapter 2

# Proportions and the Odds Ratio

## 2.1 Introduction

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

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

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

The interviewer does not know which question is answered and observes only *yes* or *no*. The respondent answers $Q$ with probability $p = 2/3$ and answers $Q^c$ with probability $1 - p$. Let $\pi$ be the unknown probability of observing a yes-response to

[1] Published as Van den Hout and Van der Heijden (2002). Randomized response, statistical disclosure control and misclassification: a review, *International Statistical Review* **70**, 269-288.

$Q$. The probability of a yes-response is $\lambda = p\pi + (1-p)(1-\pi)$. So with the observed proportion as an estimate $\widehat{\lambda}$ of $\lambda$, we can estimate $\pi$ by

$$\widehat{\pi} = \frac{\widehat{\lambda} - (1-p)}{2p-1}. \tag{2.1}$$

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

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

$$\boldsymbol{\lambda} = \boldsymbol{P}\boldsymbol{\pi}, \tag{2.2}$$

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

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

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

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

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

The second field where the misclassification probabilities are known is the post randomization method (PRAM), see Kooiman, Willenborg and Gouweleeuw (1997). The idea of PRAM is to misclassify the values of categorical variables after the data have been collected in order to protect the privacy of the respondents by preventing disclosure of their identities. PRAM can be seen as applying RR after the data have been collected. More information about PRAM and a comparison with RR is given in Section 2.2.

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

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

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

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

Special attention is given to the possibility of boundary solutions. By a boundary solution we mean an estimated value of the parameter which lies on the boundary of the parameter space. For instance, in formula (2.1) the unbiased moment estimate of $\pi$ is given. It is possible that this estimate is negative and makes no sense. In this case the moment estimate differs from the maximum likelihood estimate which is zero and therefore lies on the boundary of the parameter space. This was already noted by Singh (1976).

The possibility that the moment estimator yields estimates outside the parameter space is an awkward property, since standard analyses as, e.g., univariate probabilities and the odds ratio, are in that case useless. However, negative estimates are likely to occur when RR used. Typically, RR is applied when sensitive characteristics are investigated and often sensitivity goes hand in hand with rareness. Therefore, some of the true frequencies in a sample may be low and when these frequencies are unbiasedly estimated, random error can easily cause negative estimates. The example discussed in Section 2.7 illustrates this situation. Regarding PRAM the same problem can occur, see Section 2.2.

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

Section 2.2 provides more information about PRAM. A comparison with RR is made. Section 2.3 discusses the moment estimator of the true frequency table, i.e., the not-observed frequencies of the correctly classified scores. Point estimates and estimation of covariances are presented. In Section 2.4, we consider the maximum likelihood estimation of the true frequency table. Again point estimation and variances are discussed, this time using the EM algorithm. Section 2.5 relates the moment estimator to the maximum likelihood estimator. In Section 2.6, we consider the estimation of the odds ratio. In Section 2.7, an example is given with RR data stemming from research into violating regulations of social benefit. Section 2.8 evaluates the results and concludes.

## 2.2 Protecting Privacy Using PRAM

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

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

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

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

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

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

$$\mathbb{E}[\boldsymbol{T}^*|\boldsymbol{T}] = \boldsymbol{P}\boldsymbol{T}, \tag{2.3}$$

where $\boldsymbol{T}^*$ is the stochastic vector of perturbed frequencies and $\boldsymbol{T}$ is the vector of the true frequencies. So instead of using probabilities as in (2.2), frequencies are used in (2.3) to avoid commitment to a specific parametric model.

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

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

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

PRAM is not meant to replace existing SDC techniques. Using the transition matrix with the misclassification probabilities to take into account the perturbation due to PRAM, requires extra effort and becomes of course more complex when the research questions become more complex. This may not be acceptable to all researchers. Nevertheless, existing SDC methods are also not without problems.

Especially global recoding can destroy detail that is needed in the analysis. For instance, when a researcher has specific questions regarding teenagers becoming 18 years old, it is possible that the data he wants to use is globally recoded before it is released. It is possible that the variable Age is recoded from year of birth to age categories going from 0 to 5, 5 to 10, 10 to 15 , 15 to 20, etcetera. In that case, the researcher has lost his object of research.

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

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

Comparing (2.2) with (2.3), it can be seen that RR and PRAM are mathematically equivalent. Therefore, PRAM is presented in this paper as a special form of RR. In fact, the idea of PRAM dates back from Warner (1971), the originator of RR, who mentions the possibilities of the RR procedure to protect data after they have been collected. PRAM can be seen as applying RR after the data have been collected. Rosenberg (1979, 1980) elaborates the Warner idea and calls it additive RR contamination (ARRC). PRAM turns out to be the same as ARRC. Rosenberg discusses multivariate analysis of data protected by ARRC, he discusses multivariate categorical linear models and the chi-square test for contingency tables, in particular.

In the remainder of this section we make some comparisons between PRAM and RR. Since the methods serve different purposes, important differences may occur in practice. First, PRAM will be typically applied to those variables which may give rise to the disclosure of the identity of a respondent, i.e., covariates as, e.g., Gender, Age and Race. RR, on the other hand, will be typically applied to response variables, since the identifying covariates are obvious from the interview situation. Secondly, the usefulness of the observed response in the RR setting is dependent on the cooperation of the respondent, whereas applying PRAM is completely mechanic.

Although RR may be of help in eliciting sensitive information, the method is not a panacea (Van der Heijden, Van Gils, Bouts, and Hox 2000). The third important difference concerns the choice of the transition matrix. When using RR the matrix is determined *before* the data are collected, but in the case of PRAM the matrix can be determined conditionally on the original data. This means that the extent of randomness in applying PRAM can be controlled better than in the RR setting (Willenborg 2000).

PRAM is similar to RR regarding the possibility of boundary solutions, see Section 2.1. PRAM is typically used when there are respondents in the sample with rare combinations of scores. Therefore, some of the true frequencies in a sample may be low and when PRAM has been applied and these frequencies are unbiasedly estimated, random error can easily cause negative estimates. So also regarding PRAM, methods to deal with boundary solutions are important.

## 2.3 Moment Estimator

This section generalizes (2.1) in order to obtain a moment estimator of the true contingency table. A contingency table is a table with the sample frequencies of categorical variables. For example, the 2-dimensional contingency table of two binary variables has four cells, each of which contains the frequency of a compounded class of the two variables. Section 2.3.1 presents the moment estimator for a m-dimensional table ($m > 1$). In Section 2.3.2 formulas to compute covariances are presented.

### 2.3.1 Point Estimation

If $\boldsymbol{P}$ in (2.2) is non-singular and we have an unbiased point estimate $\widehat{\boldsymbol{\lambda}}$ of $\boldsymbol{\lambda}$, we can estimate $\boldsymbol{\pi}$ by the unbiased moment estimator

$$\widehat{\boldsymbol{\pi}} = \boldsymbol{P}^{-1}\widehat{\boldsymbol{\lambda}}, \tag{2.4}$$

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

In practice, assuming that $\boldsymbol{P}$ in (2.2) is non-singular does not impose much restriction on the choice of the misclassification design. Matrix $\boldsymbol{P}^{-1}$ exists when the diagonal of $\boldsymbol{P}$ dominates, i.e., $p_{ii} > 1/2$ for $i \in \{1, ..., K\}$, and this is reasonable since these probabilities are the probabilities that the classification is correct.

In this paper, we assume that the true response is multinomially distributed with parameter vector $\boldsymbol{\pi}$. The moment estimator (2.4) is not a maximum likelihood estimator since it is possible that for some $i \in \{1, ..., K\}$, $\widehat{\pi}_i$ is outside the parameter space (0,1).

In Section 2.1, we have considered the misclassification of one variable. The generalization to a $m$-dimensional contingency table with $m > 1$ is straightforward when we have the following independence property between each possible pair $(A, B)$ of the $m$ variables:

$$I\!\!P(A^* = i, B^* = k | A = j, B = l) = I\!\!P(A^* = i | A = j) I\!\!P(B^* = k | B = l). \qquad (2.5)$$

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

In this situation we structure the $m$-dimensional contingency table as an 1-dimensional table of the compounded variable. For instance, when we have three binary variables, we get an 1-dimensional table with rows indexed by 111, 112, 121, 122, 211, 212, 221, 222. (The last index changes first.) Due to property (2.5) it is easy to create the transition matrix of the compounded variable using the transition matrices of the underlying separate variables. Given the observed compounded variable and its transition matrix we can use the moment estimator as described above.

To give an example, assume we have an observed cross-tabulation of the misclassified variables $A$, and $B$, where row variable $A$ has $K$ categories and transition matrix $\boldsymbol{P}_A$, and column variable $B$ has $S$ categories and transition matrix $\boldsymbol{P}_B$. (When one of the variables is not misclassified, we simply take the identity matrix as the transition matrix.) Together $A$ and $B$ can be considered as one compounded variable with $KS$ categories. When property (2.5) is satisfied we can use the Kronecker product, denoted by $\otimes$, to compute the $KS \times KS$ transition matrix $\boldsymbol{P}$ as follows:

$$\boldsymbol{P} = \boldsymbol{P}_A \otimes \boldsymbol{P}_B = \begin{pmatrix} p^A_{11}\boldsymbol{P}_B & p^A_{12}\boldsymbol{P}_B & \cdots & p^A_{1K}\boldsymbol{P}_B \\ \vdots & \ddots & \ddots & \vdots \\ p^A_{K1}\boldsymbol{P}_B & \cdots & \cdots & p^A_{KK}\boldsymbol{P}_B \end{pmatrix},$$

where each $p^A_{ij}\boldsymbol{P}_B$, for $i, j \in \{1, ..., K\}$, is a $S \times S$ matrix.

### 2.3.2 Covariances

Since the observed response is multinomially distributed with parameter vector $\boldsymbol{\lambda}$, the covariance matrix of (6.3) is given by

$$\begin{aligned} V(\widehat{\boldsymbol{\pi}}) &= \boldsymbol{P}^{-1} V(\boldsymbol{\lambda}) \left(\boldsymbol{P}^{-1}\right)^t \\ &= n^{-1} \boldsymbol{P}^{-1} \left(\mathrm{Diag}(\boldsymbol{\lambda}) - \boldsymbol{\lambda}\boldsymbol{\lambda}^t\right) \left(\boldsymbol{P}^{-1}\right)^t, \end{aligned} \tag{2.6}$$

where Diag($\boldsymbol{\lambda}$) denotes the diagonal matrix with the elements of $\boldsymbol{\lambda}$ on the diagonal. The covariance matrix (2.6) can be unbiasedly estimated by

$$\widehat{V}(\widehat{\boldsymbol{\pi}}) = (n-1)^{-1} \boldsymbol{P}^{-1} \left(\mathrm{Diag}(\widehat{\boldsymbol{\lambda}}) - \widehat{\boldsymbol{\lambda}}\widehat{\boldsymbol{\lambda}}^t\right) \left(\boldsymbol{P}^{-1}\right)^t, \tag{2.7}$$

see Chaudhuri and Mukerjee (1988, Section 3.3).

As stated before, national statistical institutes prefer a model free approach. Consequently, Kooiman et al. (1997) present only the extra variance due to applying PRAM, and do not assume a multinomial distribution. The variance given by Kooiman et al. (1997) can be related to (2.6) in the following way. Chaudhuri and Mukerjee (1988, Section 3.3) present a partition of (2.6) in two terms, where the first denotes the variance due to the multinomial scheme and the second represents the variance due to the perturbation:

$$V(\widehat{\boldsymbol{\pi}}) = \boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2, \tag{2.8}$$

where

$$\boldsymbol{\Sigma}_1 = \frac{1}{n} \left(\mathrm{Diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}^t\right)$$

and

$$\boldsymbol{\Sigma}_2 = \frac{1}{n} \boldsymbol{P}^{-1} \left(\mathrm{Diag}(\boldsymbol{\lambda}) - \boldsymbol{P}\mathrm{Diag}(\boldsymbol{\pi})\boldsymbol{P}^t\right) \left(\boldsymbol{P}^{-1}\right)^t.$$

Analyzing $\boldsymbol{\Sigma}_2$ it turns out that it is the same as the variance due to PRAM given in Kooiman et al. (1997), as was to be expected, see Appendix 2.A.

## 2.4 Maximum Likelihood Estimator

As already noted in Sections 2.1 and 2.2, it is possible that the moment estimator yields estimates outside the parameter space when the estimator is applied to RR data or PRAM data. Negative estimates of frequencies are awkward, since they do not make sense. Furthermore, when there are more than two categories and the frequency of one of them is estimated by a negative number, it is unclear how the moment estimate must be adjusted in order to obtain a solution in the parameter space. This is a reason to look for a maximum likelihood estimate (MLE). Another reason to use MLEs is that in general, unbiasedness is not preserved when functions of unbiased estimates are considered. Maximum likelihood properties on the other hand, are in general preserved, see Mood, Graybill and Boes (1985).

This section discusses first the estimation of the MLE of the true contingency table using the EM algorithm and, secondly, in 2.4.2, the covariances of this estimate.

### 2.4.1 Point Estimation

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

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

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

| $A^*$ | |
|---|---|
| 1 | $n_1^*$ |
| 2 | $n_2^*$ |
| Total | $n$ |

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

$$\begin{aligned}\log l^*(\pi) &\propto n_1^* \log \lambda + n_2^* \log(1-\lambda)\\ &\propto n_1^* \log\left(p_{11}\pi + p_{12}(1-\pi)\right) + n_2^* \log\left(p_{21}\pi + p_{22}(1-\pi)\right). \qquad (2.9)\end{aligned}$$

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

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

To explain the use of the EM algorithm, we can translate the problem of maximizing (2.9) into an incomplete-data problem. We associate with each observed value of $A^*$ its not-observed non-perturbed value of $A$. Together these pairs form an incomplete-data file with size $n$. In the framework of Rubin (1976): The missing data are missing at random, since they are missing by design. When we tabulate this incomplete-data file we obtain the following table.

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

In this table, number $n_{ij}$, for $i,j \in \{1,2\}$, is the frequency of the combination $A^* = i$ and $A = j$ . Only the marginals $n_1^*$ and $n_2^*$ are observed. When we would have observed the complete data, i.e., $n_{ij}$ for $i,j \in \{1,2\}$, we would only have to consider the bottom marginal and the complete-data loglikelihood function of $\pi$ would be given by

$$\log l(\pi) \propto n_1 \log \pi + n_2 \log(1-\pi), \qquad (2.10)$$

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

The idea of the EM algorithm is to maximize the incomplete-data likelihood by iteratively maximizing the expected value of the complete-data loglikelihood (2.10), where the expectation is taken over the distribution of the complete-data given the observed data and the current fit of $\pi$ at iteration $p$, denoted by $\pi^{(p)}$. That is, in each iteration we look for the $\pi$ which maximizes the function

$$Q\left(\pi, \pi^{(p)}\right) = I\!E\left[\log l(\pi) | n_1^*, n_2^*, \pi^{(p)}\right]. \qquad (2.11)$$

In the EM algorithm it is not necessary to specify the corresponding representation of the incomplete-data likelihood in terms of the complete-data likelihood (McLachlan and Krishnan 1997, Section 1.5.1). In other words, we do not need (2.9), the function which plays the role of the incomplete-data likelihood, but we can work with (2.10) instead.

Since (2.10) is linear with respect to $n_i$, we can rewrite (2.11) by replacing the unknown $n_i$'s in (2.10) by the expected values of $n_i$'s given the observed $n_i^*$'s and $\pi^{(p)}$. Furthermore, since $n = n_1^* + n_2^*$, and $n$ is known, $n_2^*$ does not contain extra information. Therefore, (2.11) is equal to:

$$Q\left(\pi, \pi^{(p)}\right) = I\!E\left[N_1 | n_1^*, \pi^{(p)}\right] \log \pi + I\!E\left[N_2 | n_1^*, \pi^{(p)}\right] \log(1-\pi), \tag{2.12}$$

where $N_1$ and $N_2$ are the stochastic variables with values $n_1$ and $n_2$, and, of course, $N_1 + N_2 = n$.

The EM algorithm consists in each iteration of two steps: the E-step and the M-step. In this situation, the E-step consists of estimating $I\!E\left[N_1 | n_1^*, \pi^{(p)}\right]$. We assume that $(n_{11}, n_{12}, n_{21}, n_{22})$ are values of the stochastic variables $(N_{11},\ N_{12}, N_{21}, N_{22})$ which are multinomially distributed with parameters $(n, \pi_{11},\ \pi_{12}, \pi_{21}, \pi_{22})$. A property of the multinomial distribution is that the conditional distribution of $(N_{i1}, N_{i2})$ given $n_{i+} = n_i^*$ is again multinomial with parameters $(n_i^*, \pi_{i1}/\pi_{i+}, \pi_{i2}/\pi_{i+})$, for $i \in \{1,2\}$. So we have

$$I\!E\left[N_{ij} | n_1^*, \pi_{11}, \pi_{12}, \pi_{21}, \pi_{22}\right] = n_i^* \frac{\pi_{ij}}{\pi_{i+}}.$$

And consequently

$$I\!E\left[N_1 | n_1^*, \pi_{11}, \pi_{12}, \pi_{21}, \pi_{22}\right] = \frac{\pi_{11}}{\pi_{1+}} n_1^* + \frac{\pi_{21}}{\pi_{2+}} n_2^*. \tag{2.13}$$

See also Schafer (1997, Section 3.2.2.).

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

$$\pi_{i1} = I\!P(A^* = i, A = 1) = I\!P(A^* = i | A = 1) I\!P(A = 1). \tag{2.14}$$

and

$$\pi_{i+} = I\!P(A^* = i) = \sum_{k=1}^{2} I\!P(A^* = i | A = k) I\!P(A = k). \tag{2.15}$$

Next, we use (2.13), (2.14) and (2.15) to estimate $I\!\!E\left[N_1|n_1^*, \pi^{(p)}\right]$ by

$$n_1^{(p)} = \sum_{i=1}^{2} \frac{p_{i1}\pi^{(p)}}{p_{i1}\pi^{(p)} + p_{i2}\left(1 - \pi^{(p)}\right)} n_i^*,$$

which ends the E-step.

The M-step gives an update for $\pi$, which is the value of $\pi$ that maximizes (2.12), using the current estimate of $I\!\!E\left[N_1|n_1^*, \pi^{(p)}\right]$, which also provides an estimate of $I\!\!E\left[N_2|n_1^*, \pi^{(p)}\right] = n - I\!\!E\left[N_1|n_1^*, \pi^{(p)}\right]$. Maximizing is easy due to the correspondence between the standard form of (2.10) and the form of (2.12): $\pi^{(p+1)} = n_1^{(p)}/n$.

The EM algorithm is started with an initial value $\pi^{(0)}$. The following can be stated regarding the choice of the initial value and convergence of the algorithm. When there is a unique maximum in the interior of the parameter space, the EM algorithm will find it, see the convergence theorems of the algorithm as discussed in McLachlan and Krishnan (1997, Section 3.4). Furthermore, as will be explained in Section 2.5, in the RR/PRAM setting, the incomplete-data likelihood is from a regular exponential family and is therefore strictly concave, so finding the maximum should not pose any difficulties when the starting point is chosen in the interior of the parameter space and the maximum is also achieved in the interior.

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

$$\log l^*(\boldsymbol{\pi}) = \sum_{i=1}^{K} n_i^* \log \lambda_i + C \tag{2.16}$$

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

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

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

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

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

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

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

This section discussed the misclassification of one variable, but as shown in section 2.3, the generalization to a $m$-dimensional contingency table with $m > 1$ is straightforward when we have property (2.5) for each possible pair of the $m$ variables. In that case, we create a compounded variable, put together the transition matrix of this variable and use the EM algorithm as described above.

### 2.4.2 Covariances

Consider the general case where $A$ has $K$ categories and the observed data log-likelihood is given by (2.16). Assuming that the MLE of $\boldsymbol{\pi}$ lies in the interior of the parameter space, we can use the information matrix to estimate the asymptotic covariance matrix of the parameters. Using $\pi_K = 1 - \sum_{i=1}^{K-1} \pi_i$, we obtain for $k, l \in \{1, ..., K-1\}$ the $kl$-component of the information matrix:

$$-\frac{\partial}{\partial \pi_k \partial \pi_l} \log l^*(\boldsymbol{\pi}) = \sum_{i=1}^{K} \frac{n_i^*}{\lambda_i^2} (p_{il} - p_{iK})(p_{ik} - p_{iK}). \tag{2.17}$$

Incorporating the estimate $\widehat{\lambda}_i = n_i^*/n$ in (2.17) we get an approximation of the information matrix where for $k, l \in \{1, ..., K-1\}$ the $kl$-component is given by

$$\sum_{i=1}^{K} \frac{n}{\widehat{\lambda}_i} (p_{ik} - p_{iK})(p_{il} - p_{iK}), \tag{2.18}$$

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

When the MLE of $\boldsymbol{\pi}$ is on the boundary of the parameter space, using the information matrix is not appropriate and we suggest to use the bootstrap percentile method to estimate a 95% confidence interval. Regarding the bootstrap, see, e.g., Efron and Tibshirani (1993). The bootstrap scheme we propose is the following. Draw $B$ bootstrap samples from a multinomial distribution with parameter vector $\widehat{\boldsymbol{\pi}} = (\widehat{\pi}_1, ..., \widehat{\pi}_K)^t$. For each bootstrap sample the RR design is applied using a computer. Using the $B$ simulated observed tables, parameters $\widehat{\boldsymbol{\pi}}_b^{\text{boot}} = \left(\widehat{\pi}_{b1}^{\text{boot}}, ..., \widehat{\pi}_{bK}^{\text{boot}}\right)^t$, $b = 1, ..., B$, are estimated. Next, the bootstrap estimates are sorted, i.e., for each

$i \in \{1,..,K\}$, estimates $\widehat{\pi}_{1i}^{\text{boot}}, ..., \widehat{\pi}_{Bi}^{\text{boot}}$ are sorted from small to large. A confidence interval for $\widehat{\pi}_i$ is constructed by deleting 5% of the sorted values: 2.5% of the smallest estimates and 2.5% of the largest.

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

## 2.5 The MLE Compared to the Moment Estimate

In this section, we prove that the observed loglikelihood function $\log l^*(\boldsymbol{\pi})$ given in (2.16) is the loglikelihood of a distribution from a regular exponential family. Using this property of $l^*(\boldsymbol{\pi})$, the uniqueness of a solution of the likelihood equations is established when this solution is found in the interior of the parameter space. Furthermore, we prove that when the MLE is in the interior of the parameter space, the MLE is equal to the estimate provided by the moment estimator. This equality has been observed by several authors (Schwartz 1985, Appendix A, Bourke and Moran 1988, and Chen 1989) but theoretic proof is not given. By using the exponential family we prove this equality and thus provide an alternative to results in Lucy (1974) as far as they apply to misclassification of categorical variables.

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

$$l^*(\boldsymbol{\pi}) = a(\boldsymbol{\pi})b(\mathbf{n}^*)\exp\{\boldsymbol{\theta}^t(\boldsymbol{\pi})\mathbf{t}(\mathbf{n}^*)\}, \tag{2.19}$$

see Barndorff-Nielsen (1982).

Let

$$a(\boldsymbol{\pi}) = 1,$$
$$b(\mathbf{n}^*) = \frac{n!}{n_1^*! \cdots n_K^*!},$$

the sufficient statistic

$$\mathbf{t}(\mathbf{n}^*) = (t_1(\mathbf{n}^*), ..., t_K(\mathbf{n}^*))^t = (n_1^*, ..., n_K^*)^t,$$

and the canonical parameter

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

Due to the affine constraint $n_1^* + ... + n_K^* = n$, the exponential representation in (2.19) where the functions are defined as above, is not minimal, i.e., it is possible to define $\mathbf{t}$ and $\boldsymbol{\theta}$ in such a way that their dimensions are smaller than $K$. Since we need a minimal representation in order to establish regularity, we provide alternative definitions of the functions in (2.19).

A minimal representation is obtained by taking

$$\mathbf{t}(\mathbf{n}^*) = (n_1^*, ..., n_{K-1}^*)^t \tag{2.20}$$

and

$$\boldsymbol{\theta}^t(\boldsymbol{\pi}) = (\theta_1(\boldsymbol{\pi}), .., \theta_{K-1}(\boldsymbol{\pi})) = (\log \frac{\lambda_1}{\lambda_K}, ..., \log \frac{\lambda_{K-1}}{\lambda_K}), \tag{2.21}$$

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

$$l^*(\boldsymbol{\pi}) = \left(1 + e^{\theta_1} + ... + e^{\theta_{K-1}}\right)^{-n} \frac{n!}{n_1^*! \cdots n_K^*!} \exp\{\theta_1 n_1^* + ... + \theta_{K-1} n_{K-1}^*\}, \tag{2.22}$$

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

Having established that $l^*(\boldsymbol{\pi})$ is from a exponential family, we now prove, using (2.22), that the function is from a *regular* exponential family. We follow the definitions of regularity as given by Barndorff-Nielsen (1982). Let $\Omega$ be the domain of variation for $\boldsymbol{\pi}$ and $\Theta = \boldsymbol{\theta}\,(\Omega)$ the canonical parameter domain. We must prove two properties:

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

*(ii)*

$$\Theta = \{\boldsymbol{\theta} | \boldsymbol{\theta} \in \boldsymbol{\theta}\,(\Omega) \mid \int_X \frac{n!}{x_1! \cdots x_K!} e^{\boldsymbol{\theta}^t \mathbf{t}(\mathbf{x})} dx < \infty\}, \tag{2.23}$$

where $X = \{\mathbf{x} | \mathbf{x} = (x_1, ..., x_K)^t | x_1, ..., x_K > 0, x_1 + ... + x_K = n\}$ and $\boldsymbol{\theta}$ and $\mathbf{t}$ are given in (6.11) and (2.20) respectively .

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

Regarding property *(ii)*:

$$\begin{aligned}\int_X \frac{n!}{x_1!\cdots x_K!}e^{\boldsymbol{\theta}^t\mathbf{t}(\mathbf{x})}dx &\leq n!\int_X \left(\frac{\lambda_1}{\lambda_K}\right)^{x_1}\cdots\left(\frac{\lambda_{K-1}}{\lambda_K}\right)^{x_{K-1}}dx\\ &= n!\int_X \lambda_1^{x_1}\cdots\lambda_K^{x_K}\frac{1}{\lambda_K^n}dx\\ &\leq n!\int_X \left(\frac{1}{\lambda_K}\right)^n dx < \infty,\end{aligned}$$

for every $\lambda_K \in (0,1)$ and $n = x_1 + ... + x_K$. This means that (2.23) is satisfied.

Having shown that the observed data loglikelihood $l^*(\boldsymbol{\pi})$ is from a regular exponential family, we can use the powerful theory that exists for this family. A property that is of practical use is that the maximum of the observed data likelihood is unique when found in the interior of the parameter space, since the likelihood is strictly concave (Barndorff-Nielsen 1982). This justifies the use of the maximum found by the EM algorithm in Section 2.4.1.

A second property concerns the comparison of the MLE and the estimate provided by the moment estimator. The two estimates are equal when both are in the interior of the parameter space. The equality can be proved as follows. We continue to use the minimal representation as given in (2.22) where $\boldsymbol{\theta}$ is the canonical parameter and where $a(\boldsymbol{\pi})$ is given by

$$a(\boldsymbol{\pi}) = \left(1 + e^{\theta_1} + ... + e^{\theta_{K-1}}\right)^{-n},$$

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

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

That is

$$\frac{\partial}{\partial\boldsymbol{\theta}}(\boldsymbol{\theta}^t\mathbf{t}(\mathbf{n}^*)) = \frac{\partial}{\partial\boldsymbol{\theta}}(-\log a(\boldsymbol{\pi})). \tag{2.24}$$

We have

$$\frac{\partial}{\partial\boldsymbol{\theta}}(\boldsymbol{\theta}^t\mathbf{t}(\mathbf{n}^*)) = (n_1^*, ..., n_{K-1}^*)^t \tag{2.25}$$

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

$$\frac{\partial}{\partial\theta}\left(-\log a(\boldsymbol{\pi})\right) = I\!E\left[\mathbf{t}(\mathbf{N}^*)\right] = n\begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1K}\\ \vdots & \ddots & \ddots & \vdots\\ p_{K-1,K} & \cdots & \cdots & p_{K-1,K}\end{pmatrix}\begin{pmatrix}\pi_1\\ \pi_2\\ \vdots\\ \pi_K\end{pmatrix}, \tag{2.26}$$

where $\mathbf{N}^*$ is the random variable which has value $\mathbf{n}^*$. Combining (2.25) and (2.26) in (2.24) shows that the likelihood equations (2.24) are equal to the equations (2.2) on which the moment estimator is based. So in the interior of the parameter space, the MLE is equal to the ME.

Of course, the above properties of $l^*(\boldsymbol{\pi})$ can be derived without references to exponential families. Lucy (1974) discusses the estimation of a frequency distribution where observations are subject to measurement error and the error distribution is presumed known. The difference with our setting is that the observed variable is a continuous one. However, the observations are categorized in intervals and correction of the observations is on the basis of these intervals, so measurement error can be easily translated to misclassification of categorical variables. Lucy (1974) advocates an EM algorithm comparable with the EM algorithm given above. Furthermore, it is proven that in the interior of the parameter space the MLE is equal to the moment estimate and the maximum of the likelihood is unique. In Appendix 2.B we have translated Lucy's proof regarding the equivalence between the MLE and the moment estimate to our setting.

## 2.6 Odds Ratio

This section discusses the estimation of the odds ratio when data are perturbed by PRAM or RR. The odds ratio $\theta$ is a measure of association for contingency tables. We will not go into the rationale of using the odds ratio, information about this measure can be found in most textbooks on categorical data analysis, see, e.g., Agresti (2002).

Section 2.6.1 discusses point estimation both in the situation without and with misclassification. Two estimates of the odds ratio given by different authors are the same, but are not always the MLE. Section 2.6.2 discusses the variance. Again, it is important whether the estimates of the original frequencies are in the interior of the parameter space or not.

### 2.6.1 Point Estimation

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

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

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

$$\widehat{\theta} = \frac{n_{11}n_{22}}{n_{12}n_{21}}. \tag{2.27}$$

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

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

$$\widehat{\theta} = \frac{p_{A|B}(1|1)}{1 - p_{A|B}(1|1)} \left( \frac{p_{A|B}(1|2)}{1 - p_{A|B}(1|2)} \right)^{-1}. \tag{2.28}$$

Next, we consider the situation with misclassification. Two estimates of the odds ratio are proposed in the literature. Let only variable $A$ be subject to misclassification, and the 2×2 transition matrix be given by $P = (p_{ij})$. First, Magder and Hughes (1997) suggest to adjust formula (2.28) as

$$\widehat{\theta}_1 = \frac{p_{A^*|B}(1|1) - p_{12}}{p_{11} - p_{A^*|B}(1|1)} \left( \frac{p_{A^*|B}(1|2) - p_{12}}{p_{11} - p_{A^*|B}(1|2)} \right)^{-1}, \tag{2.29}$$

where $p_{A^*|B}(i|j) = n^*_{ij}/(n^*_{1j} + n^*_{2j})$ with $n^*_{ij}$ the observed cell frequencies. This formula can be used only if all the numerators and denominators in the formula are positive. If one of these is negative, the estimate is 0 or $\infty$. According to Magder and Hughes (1997), (2.29) is the MLE of $\theta$. Assuming that $\widehat{\theta}_1$ is not equal to zero or infinity, it will always be further from 1 than the odds ratio $\widehat{\theta}$ which is computed in the standard way using the observed table. Incorporating the information of the transition matrix in the estimation process compensates for the bias towards 1 (Magder and Hughes 1997).

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

$$\widehat{\theta}_2 = \frac{\widehat{n}_{11}\widehat{n}_{22}}{\widehat{n}_{12}\widehat{n}_{21}}. \tag{2.30}$$

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

In order to compare (2.29) and (2.30), we distinguish two situations concerning the misclassification of only $A$. First, the situation where estimated frequencies are in the interior of the parameter space, or, in other words, where the moment estimate of the frequencies is equal to the MLE. In this case, (2.29) and (2.30) are identical, which can be easily proved by writing out. Furthermore, (2.30), and thus (2.29), is the MLE due to the invariance property of maximum likelihood estimation (Mood et al. 1985).

Secondly, if the moment estimator yields probabilities outside the parameter space, we should compute (2.30) using the MLE, and consequently (2.29) and (2.30) differ. In fact, (2.29) is not properly defined, since it might be a negative value corresponding to the negative cell frequencies estimated by the moment estimator. Therefore, as noted in Magder and Hughes (1997), the estimate of the odds ratio should be adjusted to be either 0 or $\infty$.

The advantage of formula (2.29) is that we can use the observed table. A disadvantage is that (2.29) is not naturally extended to the situation where two variables are misclassified.

### 2.6.2 Variance

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

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

has the same asymptotic normal distribution around $\theta$ as $\widehat{\theta}$ (Agresti 2002). Note that $\tilde{\theta}$ has a variance. The close relation between $\tilde{\theta}$ and $\widehat{\theta}$ is the reason we will discuss asymptotic standard error (ASE) of $\log \widehat{\theta}$, although it is not mathematically sound to do so.

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

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

see Agresti (2002, Sections 3.1.5 and 14.1).

The second method to estimate the ASE in the situation without misclassification is to use the bootstrap. For instance, we can use the bootstrap percentile method to estimate a 95% confidence interval. When we assume the multinomial distribution, we take the vector of observed cell proportions as MLE of the cell probabilities.

With this MLE we simulate a large number of multinomial tables and each time compute the odds ratio. Then we estimate a 95% confidence interval in the same way as described in Section 2.4.2.

Next, we consider the situation with misclassification. Along the line of the two methods described above, we discuss two methods to estimate the variance of the estimate of the odds ratio. First, when the moment estimator is used, the delta-method can be applied to determine the variance of the log odds ratio. Greenland (1988) shows how this can be done when the transition matrix is estimated with known variances. Our situation is easier, since the transition matrix is given. We use the multivariate delta method (Bishop, Fienberg and Holland 1975, Section 14.6.3). The random vector is $\widehat{\boldsymbol{\pi}} = (\widehat{\pi}_{11}, \widehat{\pi}_{12}, \widehat{\pi}_{21}, \widehat{\pi}_{22})^t$ with $4 \times 4$ asymptotic covariance-variance matrix $V(\widehat{\pi})$, see Section 2.3. We take the function $f$ to be

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

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

$$V_f = (\boldsymbol{Df})^t V(\widehat{\boldsymbol{\pi}}) \boldsymbol{Df},$$

where $\boldsymbol{Df}$ is the gradient vector of $f$ at $\widehat{\boldsymbol{\pi}}$ and $V(\widehat{\boldsymbol{\pi}})$ is given by (2.6).

The problem with this method is that it makes use of the moment estimator which only makes sense when this estimator yields a solution in the interior of the parameter space. A second way to estimate the variance is using the bootstrap method as explained in Section 2.4.2. in combination with the EM algorithm. Given that $B$ is the number of bootstraps, the bootstrap will yield $\widehat{\theta}_1^{\text{boot}}, .., \widehat{\theta}_B^{\text{boot}}$ and the bootstrap percentile method can be used to estimate a 95% confidence interval.

## 2.7 Example

This section illustrates the foregoing by estimating tables of true frequencies on the basis of data collected using RR. Also, in Section 2.7.2, an estimate of the odds ratio will be discussed. The example makes clear that boundary solutions can occur when RR is applied and that we need to apply methods such as the EM algorithm and the bootstrap.

### 2.7.1 Frequencies

The RR data we want to analyze stem from a research into violating regulations of social benefit (Van Gils, Van der Heijden, and Rosebeek 2001). Sensitive items were

Table 2.1: Frequencies of observed answers to questions $Q_1$ and $Q_2$.

|  | $Q_2$ |  |  |
|---|---|---|---|
| $Q_1$ | Red | Black | Total |
| Red | 68 | 52 | 120 |
| Black | 103 | 189 | 292 |
| Total | 171 | 241 | 412 |

binary: Respondents were asked whether or not they violated certain regulations.

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

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

$$\boldsymbol{P} = \begin{pmatrix} 8/10 & 2/10 \\ 2/10 & 8/10 \end{pmatrix}, \tag{2.31}$$

We discuss observed frequencies of the colors red or black regarding two questions, $Q_1$ and $Q_2$, which were asked using this RR procedure. Both questions concern the period in which the respondent received a benefit. In translation, question $Q_1$: Did you turn down a job offer, or did you endanger on purpose an offer to get a job? And $Q_2$: Was the number of job applications less than required? In our discussion, we deal first with the frequencies of the separate questions, and secondly, we take them together, meaning that we tabulate the frequencies of the four possible profiles: red-red, red-black, black-red, black-black, and we want to know the true frequencies of the profiles violation-violation, violation-no violation, no violation-violation and no violation-no violation.

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

Table 2.2: Estimated frequencies of true answers to questions $Q_1$ and $Q_2$ using the moment estimator.

| | $Q_2$ | | |
|---|---|---|---|
| $Q_1$ | Violation | No violation | Total |
| Violation | 73.00 | -10.33 | 62.67 |
| No violation | 74.67 | 274.66 | 349.33 |
| Total | 147.67 | 264.33 | 412 |

Table 2.3: Estimated frequencies of true answers to questions $Q_1$ and $Q_2$ using the MLE.

| | $Q_2$ | | |
|---|---|---|---|
| $Q_1$ | Violation | No violation | Total |
| Violation | 67.98 | 0.00 | 67.98 |
| No violation | 78.33 | 265.69 | 344.02 |
| Total | 146.31 | 265.69 | 412 |

the data set. Univariate observed frequencies are

| $Q_1$ | |
|---|---|
| Red | 120 |
| Black | 292 |

and

| $Q_2$ | |
|---|---|
| Red | 171 |
| Black | 241 |

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

| $Q_1$ | |
|---|---|
| Violation | 62.67 |
| No violation | 349.33 |

and

| $Q_2$ | |
|---|---|
| Violation | 147.67 |
| No violation | 264.33 |

Since we have for each respondent the answers to both RR questions, we can tabulate the frequencies of the 4 possible response profiles, see Table 2.1. Using the Kronecker product to determine the 4 × 4 transition matrix, see Section 2.3.1, the moment estimate yields a negative cell entry, see Table 2.2. The MLE can be computed using the EM algorithm as described in Section 2.4.1 and is given by Table 2.3.

There is a discrepancy which shows up in this example. From Table 2.3, we can deduce estimated univariate frequencies of answers to $Q_1$ and $Q_2$. These estimates, which are based on the MLE of the true multivariate frequencies are different from the univariate moment estimates which are also MLEs. Differences however are small.

Next, we turn to the estimation of variance. First, the univariate case, where we only discuss question $Q_1$. The estimated probability of violation is $\widehat{\pi} = 62.67/412 = 0.152$. The estimated standard error of $\widehat{\pi}$ can be computed using (2.7) and is estimated to be 0.037. Secondly, we compute the variance of the four estimated probabilities concerning profiles of violation. From Table 2.3 we obtain $\widehat{\boldsymbol{\pi}} = (\widehat{\pi}_1, ..., \widehat{\pi}_4)^t = (0.17, 0.00,\ 0.19, 0.64)^t$. Since the MLE is on the boundary of the parameter space, estimating a 95% confidence interval is more useful than estimating standard errors. We use the bootstrap percentile method as explained in Section 2.4.2. and with $B = 500$ we obtain the four intervals [0.09, 0.23], [0.00, 0.08], [0.11, 0.28], and [0.53, 0.72], for $\widehat{\boldsymbol{\pi}} = (\widehat{\pi}_1, ..., \widehat{\pi}_4)^t$.

### 2.7.2 Odds Ratio

To determine whether the items corresponding to $Q_1$ and $Q_2$ are associated, we want to estimate the odds ratio. The starting point is the $2 \times 2$ contingency table of observed answers to $Q_1$ and $Q_2$, given by Table 2.1. Without any adjustment, the estimated odds ratio is $(68 \cdot 189)/(103 \cdot 52) = 2.40$.

Since we have two misclassified variables, we cannot use (2.29) to estimate the odds ratio. Instead, we estimate the $2 \times 2$ contingency table of the true frequencies and then compute the odds ratio in the standard way, as in (2.30). The moment estimate in Table 2.2 of the true frequencies yields a negative frequency, so the MLE in Table 2.3 is used. The estimate of the odds ratio is $\widehat{\theta}_2 = \infty$. This means, that given that rule 1 is violated, the probability that rule 2 is also violated is estimated to be 1. The bootstrap percentile method is used to construct a 95% confidence interval, see Section 2.4.2. In this case the interval is infinite and we are interested in the lower bound. We delete the smallest 5% of the 500 bootstrap estimates of the odds ratio and obtain the 95% confidence interval $[5.78, \infty\rangle$. So there is no reason to believe in independence between the answers to the questions. Furthermore, adjusting for the misclassification shows that the estimate of the odds ratio is much further away from 1 than the estimate based on the observed table alone.

## 2.8 Conclusion

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

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

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

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

# Appendix 2.A

As stated in Section 2.3.2, $V\left(\widehat{\pi}\right)$ can be partitioned as

$$V\left(\widehat{\boldsymbol{\pi}}\right) = \boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2, \tag{2.32}$$

where

$$\boldsymbol{\Sigma}_1 = \frac{1}{n}\left(\mathrm{Diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}^t\right)$$

and

$$\Sigma_2 = \frac{1}{n}\boldsymbol{P}^{-1}\left(\mathrm{Diag}(\boldsymbol{\lambda}) - \boldsymbol{P}\mathrm{Diag}(\boldsymbol{\pi})\boldsymbol{P}^t\right)\left(\boldsymbol{P}^{-1}\right)^t. \tag{2.33}$$

To understand (2.32):

$$\begin{aligned}
\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2 &= \frac{1}{n}\left(\mathrm{Diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}^t + \boldsymbol{P}^{-1}\mathrm{Diag}(\boldsymbol{\lambda})\left(\boldsymbol{P}^{-1}\right)^t - \mathrm{Diag}(\boldsymbol{\pi})\right) \\
&= \frac{1}{n}\left(\boldsymbol{P}^{-1}\mathrm{Diag}(\boldsymbol{\lambda})\left(\boldsymbol{P}^{-1}\right)^t - \boldsymbol{\pi}\boldsymbol{\pi}^t\right) \\
&= \frac{1}{n}\boldsymbol{P}^{-1}\left(\mathrm{Diag}(\boldsymbol{\lambda}) - \boldsymbol{P}\boldsymbol{\pi}\boldsymbol{\pi}^t\boldsymbol{P}^t\right)\left(\boldsymbol{P}^{-1}\right)^t \\
&= \frac{1}{n}\boldsymbol{P}^{-1}\left(\mathrm{Diag}(\boldsymbol{\lambda}) - \boldsymbol{\lambda}\boldsymbol{\lambda}^t\right)\left(\boldsymbol{P}^{-1}\right)^t \\
&= V\left(\widehat{\boldsymbol{\pi}}\right)
\end{aligned}$$

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

$$\begin{aligned}
V\left(\widehat{\boldsymbol{T}}|\boldsymbol{T}\right) &= \boldsymbol{P}^{-1}V\left(\boldsymbol{T}^*|\boldsymbol{T}\right)\left(\boldsymbol{P}^{-1}\right)^t \\
&= \boldsymbol{P}^{-1}\left(\sum_{j=1}^{K} T(j)\boldsymbol{V}_j\right)\left(\boldsymbol{P}^{-1}\right)^t \qquad (2.34)
\end{aligned}$$

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

$$\boldsymbol{V}_j(h,i) = \begin{cases} p_{ij}(1-p_{ij}) & \text{if} \quad h = i \\ \\ -p_{hj}p_{ij} & \text{if} \quad h \neq i \end{cases}, \text{ for } h, i \in \{1, ..., K\}, \qquad (2.35)$$

(Kooiman et al. 1997).

In order to compare (2.33) with (6.5), we go from probabilities to frequencies in the RR data. This is no problem since we assume the RR data to be distributed multinomially. So we have $V\left(\widehat{\boldsymbol{T}}|\boldsymbol{T}\right) = n^2 V\left(\widehat{\boldsymbol{\pi}}|\boldsymbol{\pi}\right)$ where, analogous to the PRAM data, $\boldsymbol{T}$ denotes the vector with the true frequencies.

In order to prove that $n^2\boldsymbol{\Sigma}_2$ is the same as (6.5), it is sufficient to prove that

$$\begin{aligned}
\sum_{j=1}^{K} T(j)\boldsymbol{V}_j &= \mathrm{Diag}(\boldsymbol{T}^*) - \boldsymbol{P}\mathrm{Diag}(\boldsymbol{T})\boldsymbol{P}^t \\
&= \begin{pmatrix} \sum_j p_{1j}T(j) & 0 & \cdots & 0 \\ 0 & \sum_j p_{2j}T(j) & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & \sum_j p_{Kj}T(j) \end{pmatrix} - \boldsymbol{P}\mathrm{Diag}(\boldsymbol{T})\left(\boldsymbol{P}\right)^t,
\end{aligned}$$

which follows by writing out.

# Appendix 2.B

As stated in Section 2.5, Lucy (1974) proves that in the interior of the parameter space the MLE of the true frequencies is equal to the moment estimate. In the following we have translated this proof to our setting and put in some explanatory steps.

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

$$\log l^*(\boldsymbol{\pi}) = \sum_{i=1}^{K} n_i^* \log \lambda_i + C \tag{2.36}$$

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

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

$$G(\boldsymbol{\pi},\mu) = \sum_{i=1}^{K} n_i^* \log \lambda_i - \mu \left( \sum_{j=1}^{K} \pi_j - 1 \right) \tag{2.37}$$

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

$$\frac{\partial}{\partial \pi_j} G(\boldsymbol{\pi},\mu) = \sum_{i=1}^{K} n_i^* \frac{p_{ij}}{\lambda_i} + \mu = 0 \tag{2.38}$$

and

$$\frac{\partial}{\partial \mu} G(\boldsymbol{\pi},\mu) = \sum_{j=1}^{K} \pi_j - 1 = 0. \tag{2.39}$$

Multiplying (2.38) with $\pi_j$ and summing over $j$ yields

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

Using (2.39) we find that $\mu = -\sum_{i=1}^{K} n_i^* = -n$. With this result it follows that the equality in (2.38) holds if $\pi_j$ for $j \in \{1,...,K\}$ is such that

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

where $\widehat{\lambda}_i = n_i^*/n$ for $i \in \{1,...,K\}$. Since the transition matrix $\boldsymbol{P}$ has the property that $\sum_{i=1}^{K} p_{ij} = 1$, the equality in (2.38) holds if $\pi_j$, for $j \in \{1,...,K\}$, is such that $\widehat{\lambda}_i = \lambda_i$ for $i \in \{1,...,K\}$.

In other words, a stationary point of (2.37) is found for such $\boldsymbol{\pi}$ that

$$\widehat{\boldsymbol{\lambda}} = \boldsymbol{P}\boldsymbol{\pi}.$$

To conclude, when (2.36) has one maximum under the constraint $\sum_{j=1}^{K} \pi_j = 1$, this maximum is attained at the moment estimator $\widehat{\boldsymbol{\pi}} = \boldsymbol{P}^{-1}\widehat{\boldsymbol{\lambda}}$.

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

# Chapter 3

# Loglinear Analysis

## 3.1 Introduction

Randomized response (RR) is an interview technique that can be used when sensitive questions have to be asked (Warner 1965; Chaudhuri and Mukerjee 1988). RR data can be seen as misclassified data where conditional misclassification probabilities are known. The main purpose of this paper is to show how research questions concerning association patterns in multivariate RR data can be assessed using latent class models (Haberman 1979; Hagenaars 1993). Describing the RR design by a latent class model (LCM) is an advantage in practice since software to assess LCMs is widely available, e.g, the program $\ell EM$ (Vermunt 1997).

In addition, this paper considers problems with respect to boundary solutions in the loglinear models that we want to fit. As far as we know, these problems are not discussed in the literature. As will be shown by examples, boundary solutions can occur when analyzing RR data and in that situation one should take care regarding the formulation of the EM algorithm. We review the discussion in Kuha and Skinner (1997) and Chen (1989), in which the EM algorithm is used for loglinear analysis of misclassified data.

As an example, RR data concerning violations of regulations for social benefit are analyzed. Sensitive items were binary: Respondents were asked whether they had violated certain regulations (Van Gils, Van der Heijden and Rosebeek 2001).

The outline of the paper is as follows. Section 3.2 introduces the RR design and misclassification designs that are closely related. Section 3.3 discusses the chi-

[1] Published as Van den Hout and Van der Heijden (2004). The analysis of multivariate misclassified data with special attention to randomized response data, *Sociological Methods and Research* **32**, 310-336.

square test of independence and introduces some of the techniques that are used in the following sections. In Section 3.4, the RR design is described by a latent class model and consequently loglinear models for RR data are presented and an example is given. Section 3.5 presents techniques for fitting the loglinear models to RR data. The likelihood is given, and the EM algorithm in the literature is discussed. In Section 3.6, attention is paid to boundary solutions and bias in RR data. Section 3.7 concludes.

## 3.2 The Randomized Response Design

The research by Van Gils et al. (2001) used the RR design introduced by Kuk (1990). In this design, there are two stacks of cards each containing black and red cards. In the right stack the proportion of red cards is 8/10, and in the left stack 2/10. The respondent is asked to draw one card from each stack and to keep the color of the cards hidden from the interviewer. Next, the sensitive question is asked. Instead of answering the question directly with *yes* or *no*, the respondent names the color of the card he took from the related stack, i.e., when the answer is *yes*, the respondent names the color of the card he or she took from the right stack, and when the answer is *no*, he or she names the color of the card from the left stack.

RR data can be described as misclassified data. We associate violations with the color red. In this way, the probability to be correctly classified is 8/10 both for respondents who violated regulations and for those who did not. The RR matrix that contains the conditional misclassification probabilities

$$p_{ij} = \mathbb{P}(\text{category } i \text{ is observed}| \text{ true category is } j) \tag{3.1}$$

is therefore given by

$$\boldsymbol{P} = \begin{pmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{pmatrix} = \begin{pmatrix} 8/10 & 2/10 \\ 2/10 & 8/10 \end{pmatrix}. \tag{3.2}$$

The main idea behind RR is that the perturbation induced by the misclassification design (in this case, the red and black cards) protects the privacy of the respondent and that insight in the misclassification design (in this case, the knowledge of the proportions red/black) can be used to analyze the observed data.

It is possible to create RR designs in which questions are asked to get information about a variable with $K > 2$ categories, see, e.g., Chaudhuri and Mukerjee (1988, Chapter 3). The general form of the RR designs we discuss is

$$\boldsymbol{\theta}^* = \boldsymbol{P}\boldsymbol{\theta}, \tag{3.3}$$

where $\boldsymbol{\theta^*} = (\theta_1^*, ..., \theta_K^*)^t$ is a vector denoting the probabilities of the observed answers with categories $1, ..., K$; $\boldsymbol{\theta} = (\theta_1, ..., \theta_K)^t$ is the vector of the probabilities of the true answers; and $\boldsymbol{P}$ is the $K \times K$ transition matrix of conditional misclassification probabilities $p_{ij}$, as given in (3.1). Note that this means that the columns of $\boldsymbol{P}$ add up to 1. Further background and more complex randomized response designs can be found in Chaudhuri and Mukerjee (1988) and Fox and Tracy (1986).

Since we are dealing with the general form of misclassification as given in (3.3), the methods discussed in this paper can also be used in categorical data analysis where misclassification occurs and the probabilities given by (3.1) are known. An example is known sensitivity and specificity in epidemiologic research, see, e.g., Magder and Hughes (1997).

There is also a similarity between RR designs and the post randomization method (PRAM), introduced by Kooiman, Willenborg, and Gouweleeuw (1997) as a method for statistical disclosure control of data matrices. Statistical disclosure control aims at safeguarding the identity of respondents. Because of the privacy protection, data producers, such as national statistical institutes, are able to pass on data to a third party. The PRAM procedure yields a new data matrix in which the values of certain categorical variables in the original matrix may be misclassified into different values according to a given probability mechanism. In this way, PRAM introduces uncertainty in the data: The user of the data cannot be sure whether the information in the matrix is original or perturbed due to PRAM. As with RR, the misclassification scheme is given by means of a $K \times K$ transition matrix $\boldsymbol{P}$ of conditional probabilities $p_{ij}$, where

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

The role of the transition matrix in the analysis of PRAM data is the same as the role of the transition matrix in the analysis of RR data. More about PRAM and the similarity with RR can be found in Van den Hout and Van der Heijden (2002).

A third field that may benefit from results regarding misclassification with known misclassification probabilities is data mining. In this field, huge amounts of data are collected from surfers on the web, and privacy concerns initiated research into ways to protect the privacy of surfers by intentional statistical perturbation (Evfimievski, Srikant, Agrawal, and Gehrke 2002).

Specific to the misclassification induced by RR is that it is nondifferential and independent. Let $A$ and $B$ denote two categorical variables, where $A$ has $I$ categories and $B$ has $J$ categories. Let $A^*$ and $B^*$ be the misclassified versions of $A$ and $B$. Misclassification of $A$ is called nondifferential with respect to $B$ if

$$I\!P(A^* = k|\ A = i,\ B = j) = I\!P(A^* = k|\ A = i), \tag{3.4}$$

where $k, i \in \{1, 2, ..., I\}$ and $j \in \{1, 2, ..., J\}$, see Kuha and Skinner (1997). The notion of independence is used when there are more than two misclassified variables. The misclassification is independent if

$$\begin{aligned} I\!P(A^* = k,\ B^* = l|\ A = i,\ B = j) = I\!P(A^* = k|\ A = i,\ B = j) \\ \times\ I\!P(B^* = l|\ A = i,\ B = j), \end{aligned} \qquad (3.5)$$

where $k, i \in \{1, 2, ..., I\}$ and $l, j \in \{1, 2, ..., J\}$.

If $\boldsymbol{P}$ in (3.3) is non-singular and we have an unbiased estimate $\widehat{\boldsymbol{\theta}}^*$ of $\boldsymbol{\theta}^*$, we can estimate $\boldsymbol{\theta}$ by the unbiased moment estimator

$$\widehat{\boldsymbol{\theta}} = \boldsymbol{P}^{-1}\widehat{\boldsymbol{\theta}}^*, \qquad (3.6)$$

(Chaudhuri and Mukerjee 1988; Kuha and Skinner 1997). In practice, assuming that $\boldsymbol{P}$ in (3.3) is non-singular does not impose much restriction on the choice of the misclassification design. Matrix $\boldsymbol{P}^{-1}$ exists when the diagonal of $\boldsymbol{P}$ dominates — that is, $p_{ii} > 1/2$ for $i \in \{1, ..., K\}$ — and this is reasonable since these probabilities are the probabilities that the classification is correct.

Due to the fact that the misclassification in a RR design is independent and nondifferential, the generalization to an $m$-dimensional contingency table with $m > 1$ is straightforward. First, the $m$-dimensional contingency table is structured as an 1-dimensional table of a compounded variable. For instance, when we have three binary variables, we obtain an one-dimensional table with rows indexed by 111, 112, 121, 122, 211, 212, 221, 222 (the last index changes first). Second, due to the properties (3.4) and (3.5), it is possible to create the transition matrix of the compounded variable using the transition matrices of the underlying variables. Given the observed compounded variable and its transition matrix, we can use the moment estimator as described above.

## 3.3 Chi-Square Test of Independence

This section discusses testing independence between two categorical variables where one variable or both variables are subject to misclassification due to a RR design such as (3.3).

Consider the cross-tabulation of the variables $A$ and $B$, which are defined in the previous section. Let $\pi_{ij} = I\!P(A = i, B = j)$ for each $i \in \{1, 2, .., I\}$ and $j \in \{1, 2, .., J\}$. The data are assumed to be distributed multinomially. The null hypothesis of independence is $H_0 : \pi_{ij} = \pi_{i+}\pi_{+j}$, where the plus sign denotes summation over the related index, e.g., $\pi_{i+} = \pi_{i1} + ... + \pi_{iJ}$. In the standard situation

Table 3.1: *(a)* Classification by Gender ($G$) and RR Answer ($F^*$) and *(b)* estimated classification by Gender ($G$) and True Answer ($F$).

*(a)*

| | $F^*$ | | |
|---|---|---|---|
| G | Red | Black | Total |
| Male | 218 | 500 | 718 |
| Female | 152 | 438 | 590 |
| Total | 370 | 938 | 1308 |

*(b)*

| | $F$ | | |
|---|---|---|---|
| G | Yes | No | Total |
| Male | 124.00 | 594.00 | 718 |
| Female | 56.67 | 533.33 | 590 |
| Total | 180.67 | 1127.33 | 1308 |

without misclassification, the expected frequencies in the (i,j) cell under $H_0$ are estimated by $\widehat{m}_{ij} = n_{i+}n_{+j}/N$, where $n_{ij}$ denotes the observed frequencies in the (i,j) cell of the cross-tabulation of $A$ and $B$, and $N$ is the sample size. The test statistic is the standard chi-square test of independence.

When one or two variables are misclassified and the misclassification is nondifferential and independent, the Collapsibility Theorem (Bishop, Fienberg and Holland 1975) can be used to show that the standard chi-square test of independence can be applied to the observed table (Korn 1981). As a result, when the misclassification is due to RR, and $A^*$ and $B^*$ denote the misclassified versions of $A$ and $B$, it is possible to make inference about the independence between $A$ and $B$ by applying the chi-square test to the observed cross-classification of $A^*$ and $B^*$. The test has the correct significance level, but power is reduced compared to the situation without misclassification. Several other authors discussed the chi-square test when one or more variables are misclassified — see, for example, Mote and Anderson (1965), Assakul and Proctor (1967), who give attention to the reduction of power, and Rosenberg (1979).

An example shows how this works with RR data. In Table 3.1 (a), two variables are cross-classified that come from research into violating regulations of social benefit (Van Gils et al. 2001). The variable $G$ denotes gender. The observed *red/black* answers to the RR question are denoted by $F^*$. The question is whether the respondents earned money by doing some odd jobs without informing the office that provides their social benefit. This is a sensitive question since not informing the office is against regulations. Let the binary variable $F$ denote the not-observed *yes/no* answers that we will call the true answers.

Applying the chi-square test to the observed values in Table 3.1 (a) yields $X^2 = 3.377$ with 1 degree of freedom and $p$ value of 0.066. When we choose a significance level of $\alpha = 0.05$, the data do not give a reason to reject the null hypothesis.

We now show that ignoring the results of Korn (1981) and taking the misclassification into account leads to the same value of $X^2$. Let $\boldsymbol{n}^* = (n^*_{11}, n^*_{12}, n^*_{21}, n^*_{22})^t$ denote the observed frequencies in Table 3.1 (a). To use the moment estimator (6.3), we first define the transition matrix $\boldsymbol{P}_{GF}$ of the compounded variable. Since the RR design for $F$ is applied with matrix (3.2) and gender ($G$) is not perturbed, we obtain

$$\boldsymbol{P}_{GF} = \begin{pmatrix} 8/10 & 2/10 & 0 & 0 \\ 2/10 & 8/10 & 0 & 0 \\ 0 & 0 & 8/10 & 2/10 \\ 0 & 0 & 2/10 & 8/10 \end{pmatrix}. \qquad (3.7)$$

This matrix is used to estimate frequencies $\widehat{\boldsymbol{n}} = (\widehat{n}_{11}, \widehat{n}_{12}, \widehat{n}_{21}, \widehat{n}_{22})^t$ in the classification by $G$ and $F$ by

$$\widehat{\boldsymbol{n}} = \boldsymbol{P}^{-1}_{GF}\boldsymbol{n}^*,$$

see Table 3.1 (b). Next, we estimate the expected frequencies in this table, denoted by $\widehat{\boldsymbol{m}}= (\widehat{m}_{11}, \widehat{m}_{12}, \widehat{m}_{21}, \widehat{m}_{22})^t$, under the model of independence by $\widehat{m}_{ij} = \widehat{n}_{i+}\widehat{n}_{+j}/N$. Since we want to fit the model of independence, we compute the fitted frequencies under this model, denoted by $\widehat{\boldsymbol{m}}^*$, by

$$\widehat{\boldsymbol{m}}^* = \boldsymbol{P}_{GF}\widehat{\boldsymbol{m}}$$

and compare them with the observed $\boldsymbol{n}^*$. Again we get $X^2 = 3.377$.

When measuring the association between $G$ and $F$ by estimating the odds ratio, the misclassification should be taken into account explicitly. To show this, we will first ignore the misclassification and, second, give the adjusted estimate of the odds ratio.

Using only Table 3.1 (a) to compute an estimate of the odds ratio $\eta$ in the standard way, yields $\widehat{\eta}^* = (218 \times 438)/(500 \times 152) = 1.26$. The large-sample standard error of $\log \widehat{\eta}^*$ is $(1/218 + 1/500 + 1/152 + 1/438)^{1/2} = 0.12$, see Agresti (2002), so that $\widehat{\eta}^*$ has the 95% confidence interval [0.99; 1.61]. This interval includes 1 and does therefore not justify rejecting the null hypothesis of independence. However, this estimate of $\eta$ is biased towards 1 (Magder and Hughes 1997) and is therefore not trustworthy.

An adjusted estimate can be deduced from Table 3.1 (b): $\widehat{\eta} = (124.00 \times 533.33)/(594.00 \times 56.67) = 1.96$. The logarithm of this estimate has a large-sample standard error of 0.40, so that $\widehat{\eta}$ has 95% confidence interval [0.90; 4.29], see Greenland (1988) and Van den Hout and Van der Heijden (2002). As expected, the interval is larger than the interval of $\widehat{\eta}^*$ since the extra variance due to the RR design is taken into account. We see that, in accordance with the chi-square test to the observed values in Table 3.1 (a), the adjusted estimation of the odds ratio does not justify rejecting $H_0$.

## 3.4 The Loglinear Model

This section discusses loglinear analysis where one or more categorical variables are observed using a RR design such as (3.3). This section can be seen as an extension to Section 3.3 since testing the loglinear model of independence for two variables is equal to the chi-square test of independence. First, we use the loglinear parameterization of the LCM to show that the RR design can be described by a LCM. Second, we give an example of loglinear analysis where one of the variables is an RR variable. The link between RR and LCMs is useful since it turns out that widely available latent class software can be used to fit loglinear models that contain RR variables.

When one or more variables in the standard loglinear model concern observed values in a RR design, loglinear analysis using only the observed table may lead to wrong inference about the parameters. Consider, for instance, the variables $G$ and $F^*$ that are cross-classified in Table 3.1 (a). The standard saturated model $(GF^*)$ to describe this table is given by

$$\log m_{gf^*} = \lambda_0 + \lambda_g^G + \lambda_{f^*}^{F^*} + \lambda_{gf^*}^{GF^*},$$

where $m_{gf^*}$ denotes the expected frequency in the $(g, f^*)$ cell, and $g, f^* \in \{1, 2\}$. The $\lambda$ terms are restricted by

$$\sum_{g=1}^{2} \lambda_g^G = \sum_{f^*=1}^{2} \lambda_{f^*}^{F^*} = \sum_{g=1}^{2} \lambda_{gf^*}^{GF^*} = \sum_{f^*=1}^{2} \lambda_{gf^*}^{GF^*} = 0.$$

The estimate $\exp(4\widehat{\lambda}_{11}^{GF^*})$ is equal to the estimate of the odds ratio $\widehat{\eta}^* = 1.26$ as given in Section 3.3 It was already noted that this estimate is biased.

In order to apply loglinear models correctly, we should take into account the misclassification due to the RR design. In the standard application of LCMs there are two kinds of variables: directly observed manifest variables and indirectly observed latent variables. The general idea is that the latent variables explain relationships among the manifest variables. Say we have one latent variable, $X$, and three manifest variables, $S$, $T$, and $U$. An important assumption in LCMs is local independence: Given the latent variable, manifest variables are independent. The loglinear parameterization of the LCM is therefore

$$\log m_{stux} = \lambda_0 + \lambda_s^S + \lambda_t^T + \lambda_u^U + \lambda_x^X + \lambda_{sx}^{SX} + \lambda_{tx}^{TX} + \lambda_{ux}^{UX}, \tag{3.8}$$

where the possible values $x$ of the latent variable $X$ and the number of categories of $X$ is not known beforehand. An example of latent class analysis is the situation where the manifest variables concern attitudes towards political issues and the latent variable is binary and indicates political orientation, for instance, left wing vs. right wing. The idea here is that the latent variable explains dependencies between the attitudes. More about this example and the general LCM can be found Hagenaars (1993).

The RR situation is rather different from the standard latent class situation. Say we have an observed *red/black* variable $A^*$ that is the misclassified version of the *yes/no* variable $A$. The relation between the variables is one to one: Manifest variable $A^*$ corresponds to latent variable $A$, and the assumption of local independence does not apply since there are no other manifest variables besides $A^*$. Furthermore, we do not have to investigate how many categories $A$ has, since the number is equal to the number of categories of $A^*$. The loglinear parameterization of this LCM is

$$\log m_{a^*a} = \lambda_0 + \lambda_{a^*}^{A^*} + \lambda_a^A + \lambda_{a^*a}^{A^*A}, \tag{3.9}$$

where $a^*, a \in \{1, 2\}$. An important property of (3.9) is that $\lambda_{a^*}^{A^*}$ and $\lambda_{a^*a}^{A^*A}$ are fixed since the conditional probabilities $I\!P(A^* = a^*|\ A = a)$ are fixed by the RR design. The relations between these terms and conditional probabilities are, given in, for example, Heinen (1996, Chapter 2); see also Section 3.5.

Once we have a loglinear parameterization of the RR design, we can add manifest variables that are not RR variables and investigate different loglinear models. We elaborate the social benefit example by considering, besides variables $F$ and $G$, the categorical variable $P$, which denotes the population size of the place of residence and has five levels. Consider Table 3.2, which cross-classifies $F^*$ with $G$ and $P$, and the loglinear model $(FGP, FF^*)$, given by

$$\begin{aligned}\log m_{f^*gpf} = \lambda_0 &+ \lambda_{f^*}^{F^*} + \lambda_g^G + \lambda_p^P + \lambda_f^F \\ &+ \lambda_{fg}^{FG} + \lambda_{fp}^{FP} + \lambda_{gp}^{GP} + \lambda_{f^*f}^{F^*F} + \lambda_{fgp}^{FGP},\end{aligned} \tag{3.10}$$

where $\lambda_{f^*}^{F^*}$ and $\lambda_{f^*f}^{F^*F}$ are fixed by the RR design, and $f^*, g, f \in \{1, 2\}$, $p \in \{1, .., 5\}$. We call (3.10) the saturated model for the latent table $FGP$. In what follows, we will assess different loglinear models for the latent table $FGP$. All the models

Table 3.2: Classification by RR Answer ($F^*$), Gender (G), and Population Size of the Place of Residence ($P$).

| | | $P$ ($\times$1000) | | | | |
|---|---|---|---|---|---|---|
| $F^*$ | $G$ | $\geq$ 400 | 100-400 | 50-100 | 20-50 | $\leq$ 20 |
| Red | Male | 12 | 34 | 51 | 79 | 42 |
| | Female | 19 | 30 | 33 | 47 | 23 |
| Black | Male | 32 | 89 | 79 | 198 | 102 |
| | Female | 35 | 101 | 105 | 150 | 47 |

Table 3.3: Goodness-of-fit statistics for loglinear models for table $FGPF^*$.

| Model | df | $X^2$ | $p$ value | $L^2$ | $p$ value |
|---|---|---|---|---|---|
| 1. $(FG, FP, GP, FF^*)$ | 4 | 6.78 | 0.15 | 6.70 | 0.15 |
| 2. $(FG, GP, FF^*)$ | 8 | 11.54 | 0.17 | 11.10 | 0.20 |
| 3. $(FP, GP, FF^*)$ | 5 | 10.34 | 0.07 | 10.39 | 0.06 |
| 4. $(GP, FF^*)$ | 9 | 14.85 | 0.10 | 14.49 | 0.11 |

include the fixed terms $\lambda_{f^*}^{F^*}$ and $\lambda_{f^*f}^{F^*F}$ since the RR design should always be taken into account.

The preceding discussion shows that we can use latent class software when this software allows for restrictions on conditional probabilities. The program *ℓEM* (Vermunt 1997) is an example of this kind of software. Since the LCMs that correspond to the RR design are very restricted, estimation of the models become less complex when they describe RR data. Using *ℓEM* to estimate loglinear models for RR data is easy and fast. The code that we used for the example with variables $F^*$, $G$ and $P$ is given in Appendix 3.A. Apart from the fixed interaction terms, the models for the RR data also differ from standard loglinear models because they concern an incomplete contingency table, i.e., in table $FGPF^*$ variable $F$ is not observed. Because of this incompleteness, the EM algorithm (Dempster, Laird, and Rubin 1977) is applicable. The estimation of the models and the formulation of the EM algorithm are discussed in Section 3.5.

Since we are not interested in the relation between $G$ and $P$, we only consider models that contain $G$ and $P$ jointly. In Table 3.3, the values of the test statistics of several models are given. Estimating the frequencies under the saturated loglinear model for the latent table $FGP$ yields the estimates in Table 3.4. This table can

Table 3.4: Estimated classification by True Answer ($F$), Gender (G), and Population Size of the Place of Residence ($P$).

| | | $P$ ($\times$1000) | | | | |
|---|---|---|---|---|---|---|
| $F$ | $G$ | $\geq$ 400 | 100-400 | 50-100 | 20-50 | $\leq$ 20 |
| Yes | Male | 5.3 | 15.7 | 41.7 | 39.3 | 22.0 |
| | Female | 13.7 | 6.3 | 9.0 | 12.7 | 15.0 |
| No | Male | 38.7 | 107.3 | 88.3 | 237.7 | 122.0 |
| | Female | 40.3 | 124.7 | 129.0 | 184.3 | 55.0 |

Table 3.5: Hypothesis test for various pairs of nested models in Table 3.3.

| Comparison models | $\Delta$df | $\Delta L^2$ | $p$ value |
|---|---|---|---|
| 2 versus 1 | 4 | 4.40 | 0.36 |
| 4 versus 2 | 1 | 3.39 | 0.07 |
| 3 versus 1 | 1 | 3.69 | 0.05 |
| 4 versus 3 | 4 | 4.10 | 0.39 |

also be estimated by using the unbiased moment estimator (6.3).

We partitioned the likelihood ratio goodness-of-fit statistic to find the best model, see Table 3.5, and this leads to model $(GP, FF^*)$. This means that there is no convincing evidence for the dependence between $F$, on one hand, and $G$ and $P$, taken jointly, on the other hand. Estimated $\lambda$ terms and their estimated standard errors for model $(GP, FF^*)$ are given in Table 3.6. The data indicate that not informing the social benefit office about money earned, is independent of gender and population size of the place of residence taken together.

In the remainder of this section, we make some general remarks with respect to the loglinear models for RR data. Certain loglinear models for RR data can be tested when standard loglinear analysis is applied to observed variables, even when some of the variables are misclassified. Korn (1981) showed that a hierarchical model is preserved by misclassification if the misclassified variable appears only once in the specification of the model. *Preserved* means that the misclassification will not change the fact that the observed table satisfies the model. When testing the model to the observed table, the same significance level is achieved, but power is reduced. An example is applying the chi-square test to a two-dimensional table, see Section 3.2. Another example is the model $(AB, BC)$ that is preserved under misclassification in $A$ and in $C$ but not under misclassification in $B$.

In practice, it often will be the case that we want to investigate loglinear mod-

Table 3.6: Estimates of $\lambda$ parameters in model $(GP, FF^*)$ and their estimated standard errors.

| Parameter | Estimate | Standard errors | Parameter | Estimate | Standard errors |
|---|---|---|---|---|---|
| $\lambda_1^F$ | -0.92 | 0.09 | $\lambda_4^P$ | 0.72 | 0.05 |
| $\lambda_1^G$ | 0.07 | 0.03 | $\lambda_{11}^{GP}$ | -0.18 | 0.08 |
| $\lambda_1^P$ | -0.85 | 0.08 | $\lambda_{12}^{GP}$ | -0.10 | 0.06 |
| $\lambda_2^P$ | 0.11 | 0.06 | $\lambda_{13}^{GP}$ | -0.10 | 0.06 |
| $\lambda_3^P$ | 0.16 | 0.06 | $\lambda_{14}^{GP}$ | 0.10 | 0.05 |

els that do not meet the criterion formulated by Korn (1981), and we still need the adjustments described by the methods above. Also, even when a model is preserved, the estimation of the $\lambda$ terms in the model should take the misclassification into account. In the example, only model $(FG, FP, GP, FF^*)$ does not satisfy the assumptions of Korn.

Another important point is whether local maxima of the likelihood at hand are possible. In the standard hierarchical loglinear model, the likelihood function has a unique maximum when the solution is in the interior of the parameter space, see Birch (1963). Regarding the general LCM, it is known that it is possible that the likelihood function has local maxima (Haberman 1979). However, the restricted LCM in this paper that describes the RR variables seems to have different properties than the general LCM. In the example above and in other not reported loglinear analyses, we did not encounter local maxima of the likelihood functions. As an illustration, Figure 3.1 depicts the likelihood given by (3.12) of the independence model for latent table $FG$ denoted by $(G, FF^*)$ and applied to Table 3.1 (a). As can be seen in Figure 3.1, the parameter space of the likelihood seems to have one maximum. We obtain $(\widehat{\lambda}_1^G, \widehat{\lambda}_1^F) = (0.098, -0.92)$ as the point where the maximum is attained.

We conjecture that the loglinear models for RR data have a unique maximum if there is a solution in the interior of the parameter space. With respect to the saturated model for the latent table, this is true since the saturated model is just a reparametrization of the multinomial distribution and Van den Hout and Van der Heijden (2002) prove that in that case there is a unique maximum. The conjecture for more parsimonious models might be investigated using research concerning product models (Haberman 1977) or research into marginal models, see, e.g., Bergsma and Rudas (2002). Both fields seem to address related problems. We hope to provide

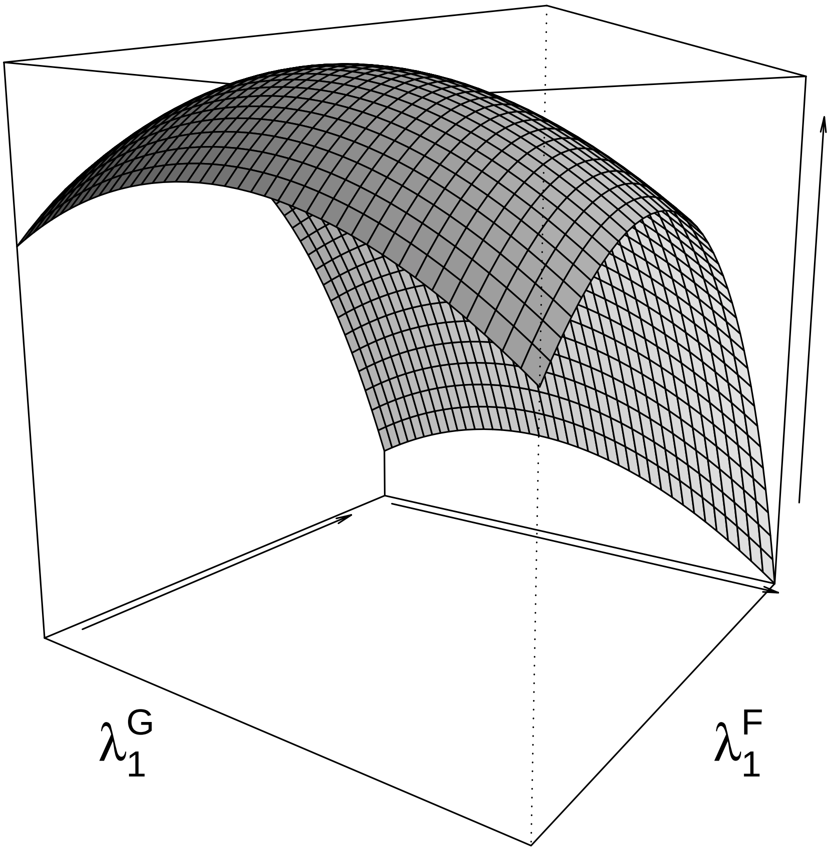


Figure 3.1: Likelihood of model $(G, FF^*)$

a decisive answer in future research.

## 3.5 Estimating The Loglinear Model

This section presents techniques for estimating the loglinear models for RR data discussed in Section 3.4. First, we specify the likelihood. Second, we discuss the EM algorithm that can be used to maximize the likelihood. For loglinear models with latent variables, the algorithm was formulated by Haberman (1979). Both Chen (1989) and Kuha and Skinner (1997) use the algorithm in the situation of misclassification when misclassification probabilities are known, although the formulations of the algorithm differ. Chen (1989) explicitly discusses RR data. We will review the two formulations since the difference is important when a boundary solution is encountered. By a *boundary solution*, we mean an estimated expected cell frequency in the latent table that equals zero. Section 3.6 will give examples of RR data where boundary solutions occur.

To give the general formula of the likelihood, let the latent frequencies $\boldsymbol{n} = (n_1, ..., n_D)^t$ be multinomially distributed with parameters $N$ and $\boldsymbol{\theta} = (\theta_1, ..., \theta_D)^t$. We specify loglinear models by $\eta_d = \log \theta_d$, $d \in \{1, ..., D\}$, and $\boldsymbol{\eta} = \boldsymbol{M}\boldsymbol{\lambda}$, where $\boldsymbol{\eta} = (\eta_1, ..., \eta_D)^t$, $\boldsymbol{M}$ is the $D \times r$ design matrix that defines the loglinear model, and $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_r)^t$ is the parameter vector of the model. Ignoring constants, the likelihood is given by

$$L(\boldsymbol{\lambda}|\boldsymbol{n}^*) = \prod_{i=1}^{D} (\theta_i^*)^{n_i^*} = \exp\{\sum_{i=1}^{D} n_i^* \log(p_{i1}e^{\eta_1} + .. + p_{iD}e^{\eta_D})\}, \qquad (3.11)$$

where $\boldsymbol{n}^*$ is the vector with observed frequencies, and $p_{ij}$, $i, j \in \{1, ..., D\}$, are the entries of the transition matrix that describe the misclassification with respect to $\boldsymbol{\theta}$. Since $\boldsymbol{n}$ is assumed to be multinomially distributed, $\boldsymbol{n}^*$ is also multinomially distributed due to the specific form of the transition matrix (Van den Hout and Van der Heijden 2002, Section 5).

As an example, consider the likelihood of the independence model $(G, F)$ applied to Table 3.1 (a), given by

$$L(\boldsymbol{\lambda}|\boldsymbol{n}^*) = \prod_{i=1}^{4} (\theta_i^*)^{n_i^*} = \exp\{\sum_{i=1}^{4} n_i^* \log(p_{i1}e^{\eta_1} + p_{i2}e^{\eta_2} + p_{i3}e^{\eta_3} + p_{i4}e^{\eta_4})\}, \qquad (3.12)$$

where $\boldsymbol{n}^*$ are the frequencies in Table 3.1 (a); $p_{ij}$, $i, j \in \{1, ..., 4\}$, are the entries in $\boldsymbol{P}_{GF}$ given by (3.7); and the non-redundant parameters are $\lambda_1^G$ and $\lambda_1^F$. Note that parameter $\lambda_0$ is not a free parameter in the loglinear model — in this case

$$\lambda_0 = -\log \sum_{gf} \exp(\lambda_g^G + \lambda_f^F),$$

where $g, f \in \{1, 2\}$.

We can maximize (4.6) directly using the Newton-Raphson method, but we can also maximize it using an EM algorithm. The program $\ell EM$ (Vermunt 1997) uses both procedures. The program starts with an EM algorithm and uses Newton-Raphson when close to the maximum. The applicability of the EM algorithm to RR data becomes clear when RR data are viewed as incomplete data. For each respondent, we can associate with the observed value of $F^*$ the not-observed nonperturbed value of $F$. Together, these pairs form an incomplete data matrix. In the framework of Rubin (1976): The missing data are missing at random since they are missing by design.

Next we review the formulations of the EM algorithm given by Chen (1989) and Kuha and Skinner (1997). We use the example in the preceding section and start with the formulation by Chen (1989). Say we want to fit model $(FGP, FF^*)$. With $v = 0, 1, 2...$ denoting the cycles, the algorithm is given by

*Initial estimate*: $m^{(0)}_{fgp}$

*E-step*: $n^{(v)}_{f^*gpf} = n_{f^*gp}\Big(m^{(v)}_{fgp}\pi_{f^*|f}\Big)\Big/\Big(\sum_{t=1}^{2} m^{(v)}_{tgp}\pi_{f^*|t}\Big)$

*M-step*: Fit $(FGP, FF^*)$ to $n^{(v)}_{f^*gpf}$ and use estimated expected frequencies to compute $m^{(v+1)}_{fgp}$,

where $m^{(0)}_{fgp}$ is the initial estimate of the frequencies in latent table $FGP$, $n_{f^*gp}$ are the observed frequencies in the $F^*GP$ table, and in each step $f^*, g, f \in \{1, 2\}$, $p \in \{1, .., 5\}$. The conditional probabilities $\pi_{f^*|f}$ are fixed and provided for by the transition matrix for $F$ given by (3.2).

To test the model after convergence, compare $m^{(\infty)}_{f^*gp+}$ with $n_{f^*gp}$ using, for instance, the chi-square test or the likelihood ratio test. The degrees of freedom of the chi-square distributions of these test statistics are the number of cells that are compared minus the number of parameters fitted. For model $(FGP, FF^*)$, we have 20 - 20 = 0 degrees of freedom since the $\lambda^{F^*}_{f^*}$ and $\lambda^{F^*F}_{f^*f}$ are fixed due to the RR design.

Chen (1989) is not explicit with respect to the fixed $\lambda$ -terms in models such as $(FGP, FF^*)$. We think that it is important to stress that when fitting a loglinear model in the M-step, one should check whether the restrictions due to the RR design are maintained. The relations between $\lambda$ terms and conditional probabilities in our example are

$$\pi_{f^*|f} = \frac{\exp(\lambda^{F^*}_{f^*} + \lambda^{F^*F}_{f^*f})}{\sum_{f^*}\exp(\lambda^{F^*}_{f^*} + \lambda^{F^*F}_{f^*f})}, \tag{3.13}$$

where $f^*, f \in \{0, 1\}$, and the summation is over values $f^* \in \{0, 1\}$, see, e.g., Heinen (1996, Chapter 2). The relation (3.13) is the same for more parsimonious models for the latent table $FGP$. Using (3.13), we obtain

$$\lambda_1^{F^*} = -1/4\Big(\log \pi_{2|1} - \log \pi_{1|1} - \log \pi_{1|2} + \log \pi_{2|2}\Big) = 0$$

$$\lambda_{11}^{F^*F} = -1/4\Big(\log \pi_{2|1} - \log \pi_{1|1} + \log \pi_{1|2} - \log \pi_{2|2}\Big) = 0.69. \qquad (3.14)$$

When fitting $(FGP, FF^*)$ in the M-step in a standard way without maintaining the restrictions, $\widehat{\lambda}_1^{F^*}$ and $\widehat{\lambda}_{11}^{F^*F}$ converge to the fixed values of $\lambda_1^{F^*}$ and $\lambda_{11}^{F^*F}$. However, when there is a boundary solution, this may not be the case. In Section 3.6, we will give examples of RR data with boundary solutions. When the restrictions are not maintained, estimated expected frequencies $m_{f^*gp+}^{(\infty)}$ are wrong; consequently, the value of the test statistic is wrong. This means that the formulation of the EM algorithm in Chen (1989) does not always yield the EM algorithm that we want.

To apply the EM algorithm correctly — that is, in such a way that it also yields the right estimates in the case of boundary solutions — there are two possible adjustments. We will explain these two adjustments and show that they are one and the same due to the Collapsibility Theorem. First, we can fit $(FGP, FF^*)$ in the M-step using the fixed values of $\lambda_{f^*}^{F^*}$ and $\lambda_{f^*f}^{F^*F}$, given by (3.14). The disadvantage is that this is not completely standard loglinear analysis since we should take care of these restrictions in estimating expected frequencies.

Second, we can use the E-step and M-step as given in Kuha and Skinner (1997), who refer to Chen (1989) but nevertheless give a different formulation of the algorithm — namely

*E-step*:
$$n_{f^*gpf}^{(v)} = n_{f^*gp}\Big(m_{fgp}^{(v)}\pi_{f^*|f}\Big)\Big/\Big(\textstyle\sum_{t=1}^{2} m_{tgp}^{(v)}\pi_{f^*|t}\Big)$$
$$n_{fgp}^{(v)} = n_{+gpf}^{(v)}$$

*M-step*: Fit $(FGP)$ to $n_{fgp}^{(v)}$ to obtain estimated expected frequencies $m_{fgp}^{(v+1)}$.

The advantage is that the M-step is standard. With respect to the testing of the model after convergence, Kuha and Skinner (1997) do not explicitly give a procedure, but we suggest doing the following. Structure the estimated frequencies $\widehat{m}_{fgp}$ in the three-dimensional latent table $FGP$ as a one-dimensional table of a compounded variable — say $\widehat{\boldsymbol{m}}$ — and compute the fitted frequencies under this model, denoted

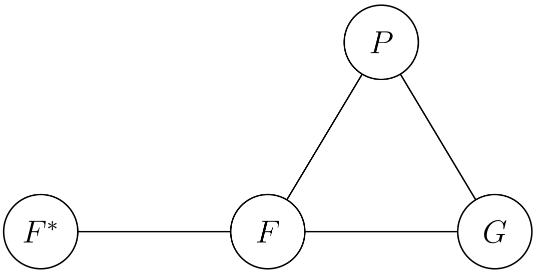


Figure 3.2: Saturated model for the latent table $FGP$

by $\widehat{\boldsymbol{m}}^*$, by

$$\widehat{\boldsymbol{m}}^* = \boldsymbol{P}_{FGP}\widehat{\boldsymbol{m}}, \tag{3.15}$$

where $\boldsymbol{P}_{FGP}$ is the transition matrix of the compounded variable. Next, $\widehat{\boldsymbol{m}}^*$ can be compared with the observed $\boldsymbol{n}^*$.

The above adjustments yield one and the same EM algorithm. This follows from the applicability of the Collapsibility Theorem (Bishop et al. 1975). The theorem states that the interaction between $F$, $G$ and $P$ in model $(FGP, FF^*)$ can be measured from the table of sums obtained by collapsing table $F^*GPF$ over $F^*$, see Figure 3.2.

This is why, in the EM algorithm, we can collapse the estimated complete table in the E-step *before* we apply loglinear analysis in the M-step.

To test the fitted model after convergence of the EM algorithm formulated in Kuha and Skinner (1997), we can combine the estimated $\lambda$ terms of the latent table $FGP$, i.e., the main effects and the interactions, with the fixed $\lambda$ terms given by (3.14), compute $\lambda_0$, and estimate the complete table $F^*GPF$ to which the model $(FGP, F^*F)$ exactly fits. However, since the information of the fixed $\lambda$ terms is completely given by the transition matrix of $F$, we can also proceed after the EM algorithm as described by (3.15). In this way, we can stay away from the estimation of $\lambda$ terms and work with cell frequencies instead.

When the reader wants to implement the EM algorithm, we advocate using the EM algorithm in Kuha and Skinner (1997) and testing the models using (3.15). Note, however, that the EM algorithm does not yield estimated standard errors for

Table 3.7: *(a)* Classification by RR Answers $F_1^*$ and $F_2^*$, and *(b)* estimated classification by True Answers $F_1$ and $F_2$.

*(a)*

| | $F_2^*$ | | |
|---|---|---|---|
| $F_1^*$ | Red | Black | Total |
| Red | 133 | 237 | 370 |
| Black | 147 | 791 | 938 |
| Total | 280 | 1028 | 1308 |

*(b)*

| | $F_2$ | | |
|---|---|---|---|
| $F_1$ | Yes | No | Total |
| Yes | 107.21 | 66.22 | 173.43 |
| No | 0.00 | 1134.57 | 1134.57 |
| Total | 107.21 | 1200.79 | 1308 |

the estimated $\lambda$ terms. For estimated standard errors, one could use a method like Newton-Raphson, as is done in $\ell EM$.

## 3.6 Boundary Solutions

This section discusses boundary solutions that we encountered in the RR data concerning violations of regulations for social benefit (Van Gils et al. 2001). On the basis of these examples, a more general discussion is given with respect to boundary solutions in RR data and in PRAM data. This section generalizes the discussion of boundary solutions in Van den Hout and Van der Heijden (2002) to the situation with more than one variable.

A boundary solution is encountered when an estimated cell frequency in the latent table equals zero. This situation might occur when we combine several RR variables. From the research concerning violations of regulations for social benefit, we consider three binary RR variables — $F_1^*$, $F_2^*$, and $F_3^*$ — with latent counterparts that are denoted by $F_1$, $F_2$ and $F_3$. Variable $F_1^*$ is the same as $F^*$ in the Sections 3.3 and 3.4. Variable $F_2^*$ denotes observed answers concerning the question of whether the respondents had a (temporary) legal job without informing the office that provides their social benefit. Variable $F_3^*$ concerns the question of whether the respondent had an illegal job without informing the office. One transition matrix is used for each of the three variables and is given by (3.2).

As an example, consider Table 3.7 (a), which contains observed frequencies of RR variables $F_1^*$ and $F_2^*$, and the estimated latent Table 3.7 (b), which contains estimated expected frequencies under model $(F_1F_2, F_1F_1^*, F_2F_2^*)$. Testing the saturated model for the latent table $F_1F_2$ yields $X^2 = 18.67$ and $L^2 = 20.12$. If there would not have been estimated zeroes in Table 3.7 (b), $X^2$ and $L^2$ would have been

Table 3.8: *(a)* Classification by RR Answers $F_1^*$, $F_2^*$ and $F_3^*$, and *(b)* estimated classification by True Answers $F_1$, $F_2$ and $F_3$.

*(a)*

| | | $F_3^*$ | |
|---|---|---|---|
| $F_1^*$ | $F_2^*$ | Red | Black |
| Red | Red | 66 | 67 |
| | Black | 68 | 169 |
| Black | Red | 52 | 95 |
| | Black | 123 | 668 |

*(b)*

| | | $F_3$ | |
|---|---|---|---|
| $F_1$ | $F_2$ | Yes | No |
| Yes | Yes | 101.92 | 11.07 |
| | No | 18.56 | 45.38 |
| No | Yes | 0.00 | 0.00 |
| | No | 0.00 | 1131.06 |

zero.

A second example is given by Table 3.8 (a), which contains observed frequencies, and Table 3.8 (b), which contains estimated expected frequencies under model $(F_1F_2F_3, F_1F_1^*, F_2F_2^*, F_3F_3^*)$. The sample size is again 1308. Testing the saturated model for the latent table $F_1F_2F_3$ yields $X^2 = 38.53$ and $L^2 = 41.61$. It is clear from Table 3.8 (b) that the questions are strongly related.

Boundary solutions might occur due to random error. When some of the latent frequencies are close to zero, an estimate of these frequencies after RR has been executed might result in a boundary solution. As an example, consider the binary latent variable $H$ with latent frequencies $(98, 2)^t$ and assume that the transition matrix is given by (3.2). Possible frequencies of $H^*$ due to the misclassification by the RR design are $(85, 15)^t$, and on the basis of these frequencies, the latent frequencies are estimated as $(100, 0)^t$ — a boundary solution. When fitted frequencies are estimated by left-multiplying the transition matrix with $(100, 0)^t$, we get fitted frequencies $(80, 20)^t$ and $X^2 > 0$.

So, the fact that $X^2 > 0$ for the saturated model for the latent table is in itself not an indication that something is amiss. However, if the difference between zero and the $X^2$ of model $(F_1F_2F_3, F_1F_1^*, F_2F_2^*, F_3F_3^*)$ is large, it might be an indication that the perturbation of the latent frequencies is not due to misclassification alone. This can be shown by two reasonings. First, a parametric bootstrap in which data are sampled from the estimated expected frequencies under the model can show that the large value of $X^2$ is unlikely when only misclassification is taken into account. We carried out such a bootstrap to investigate $X^2 = 38.53$ for the model $(F_1F_2F_3, F_1F_1^*, F_2F_2^*, F_3F_3^*)$ given the RR design. To describe the bootstrap, we switch to binary RR variables $H_1^*$, $H_2^*$, and $H_3^*$. From the estimated expected frequencies in the latent table $F_1F_2F_3$ under model $(F_1F_2F_3, F_1F_1^*, F_2F_2^*, F_3F_3^*)$, we sampled 100 tables $H_1H_2H_3$ and next simulated 100 tables $H_1^*H_2^*H_3^*$ using the RR design. For each of these tables $H_1^*H_2^*H_3^*$ the test statistic $X^2$ is computed. The mean value of the 100 simulated $X^2$'s is 2.00, and the maximum is 9.25. This maximum is not even close to the $X^2 = 38.53$ of model $(F_1F_2F_3, F_1F_1^*, F_2F_2^*, F_3F_3^*)$. Results for the likelihood ratio test $L^2$ are very similar.

A second way to investigate whether the RR data can be described by misclassification alone is the following. Continuing with the example: Assume that none of the respondents committed fraud, or in other words, that the latent score is 2 for each variable $F_1$, $F_2$, and $F_3$. Under this assumption, the expected number of respondents with observed score *black* for each variable $F_1^*$, $F_2^*$, and $F_3^*$ is $(8/10)^3 \times 1308 = 669.70$. The observed number in the survey is 668. This might suggest that the assumption is correct and that there are no frauds at all in the survey. However, this is contradicted by the 66 respondents who have the score *red* for each variable $F_1^*$, $F_2^*$, and $F_3^*$ — a frequency that is much higher than the expected $(2/10)^3 \times 1308 = 10.46$ under the assumption of no fraud at all.

The two reasonings show that, given the RR design in the research and the estimated expected frequencies in Table 3.8 (b), it is rather unlikely that Table 3.8 (a) is the observed table. However, since the observed data are the starting point of statistical inference, we should state the conclusion the other way around: Given Table 3.8 (a) and the RR design, Table 3.8 (b) is probably not a good estimate of the latent frequencies. The cause for this estimation problem is probably that some respondents do not always follow the RR design and answer *black* too often, irrespective of the question asked. A reason for this might be that some respondents do not trust the privacy protection offered by the RR design and answer *black* since *black* is associated with *no*. These respondents bring about a second perturbation of the latent frequencies besides the misclassification due to the RR design. In the conclusion, we will return to this problem.

To make the discussion more general, note that when the statistical disclosure method PRAM is applied, $X^2 > 0$ for the saturated model for the latent table can only occur due to random error since the misclassification is executed by the computer. Also, in the case where RR data are not unlikely in the sense as discussed above, a method is needed to deal with the fact that $X^2$ might be unequal to zero. In the related field of incomplete data it also may occur that $X_0^2 > 0$ (Schafer 1997). We suggest following Schafer (1997), who proposes taking the deviation from the null as a baseline for assessing nonsaturated models and defines an adjusted test statistic

$$X_{adj}^2 = X^2 - X_0^2,$$

where $X_0^2$ is the $X^2$ of the saturated model. The likelihood ratio test is adjusted in the same way. The behavior of this adjusted test might be studied using Gelman, Meng and Stern (1996), who use Bayesian analysis to assess model fit in situations when $X^2 > 0$ but when $X^2$ is expected to be zero if the model is true.

## 3.7 Conclusion

This paper discusses loglinear analysis of randomized response data. It is shown that this kind of analysis can be executed using existing latent class software.

The RR data from the example in this paper are difficult data. The problem is not the theoretic misclassification due to the RR design since this paper shows that we can handle this misclassification. The problem is that some respondents do not follow the RR design and — of course — that we cannot identify these "cheaters". To some extent we can use $X_0^2$, i.e., the test statistic for the saturated model for the latent table, as a measure of the bias of the RR data but it provides not a decisive answer.

Asking sensitive questions will always produce incomplete or biased data. So one must make do with what one has got, and our idea is that RR performs relatively well (Van der Heijden, Van Gils, Bouts, and Hox 2000). Analysis of RR data in the future might profit from research into more methodological aspects of RR designs, see Boeije and Lensvelt-Mulders (2002), who discuss cheating in RR designs. A possible form of cheating is when a respondent answers *black*, irrespective of the question asked since he or she does not trust the privacy protection. When respondents understand the privacy protection offered by the RR design better, data might be less biased. Another approach might be to add extra parameters to the model in order to describe cheating behavior. This is not straightforward since it is difficult to model cheating behavior, and also we might run into identifiability problems, see, e.g., Goodman (1974).

The question of how the bias in the data influences the analysis is difficult to answer. Obvious is that results should be interpret with care and that cross-classifying several RR questions might increase the bias of the results. We suggest the following: When the saturated model fits perfectly, i.e., $X_0^2 = 0$, we advocate the loglinear modeling as described in Section 3.4. When $X_0^2 > 0$, one should be more careful, and the reasonings used in Section 3.6 can be used to assess bias in the RR data due to the "cheaters". When the parametric bootstrap in Section 3.6 makes the value of $X_0^2$ unlikely, it is unclear how to interpret the results of the loglinear analysis. When $X_0^2$ is not too large, i.e., the deviation from the null can be explained by random error, an adjusted test statistic can be used to test the fitting of loglinear models.

# Appendix 3.A

We present two input files that can be use in $\ell EM$ to fit the models in Section 3.4. The program $\ell EM$ and the manual can be downloaded for free from www.kub.nl/faculteiten/fsw/ organisatie/departementen/mto/software2.html. The symbol * denotes comment. By removing and adding this symbol in the text, different models can be fitted. The first input fits the saturated model for the latent table $FGP$ and follows the loglinear parameterization of the LCM.

```
lat 1            * 1 latent variable
man 3            * 1 manifest variable
dim 2 2 2 5      * dimensions of variables
lab F R G P      * labels:
                 * F = fraud, R = observed RR answer
                 * G = gender, R = pop. size of place of residence

mod {FGP,wei(FR)}        * sat. mod. with weighted interaction FR
*mod {FG,FP,PG,wei(FR)}  * no 3-way interaction model
*mod {FG,PG,wei(FR)}     * conditional independence
*mod {FP,GP,wei(FR)}     * conditional independence
*mod {F, GP,wei(FR)}     * model of joint independence
sta wei(FR) [.8 .2 .2 .8]    * misclassification prob. determine weights

dat [12 34 51 79 42 19 30 33 47 23       * observed data
32 89 79 198 102 35 101 105 150 47]
```

The second input also fits the saturated model for the latent table $FGP$ but follows the loglinear modified path model parameterization (Goodman 1973; Hagenaars 1993, p.15). We only give the input for the the saturated model for the latent table $FGP$, but restrictive models can easily be formulated.

```
lat 1            * 1 latent variable
man 3            * 1 manifest variable
dim 2 2 2 5      * dimensions of variables
lab F R G P      * labels
```

```
mod FGP {FGP}                     * saturated model
R|F {wei(RF)}                     * specifying weights
sta wei(RF) [.8 .2 .2 .8]         * using misclassification probabilities

dat [12 34 51 79 42 19 30 33 47 23     * observed data
32 89 79 198 102 35 101 105 150 47]
```

# Chapter 4

# Randomized Response in a $2 \times 2$ Factorial Design

## 4.1 Introduction

Randomized response (RR) is an interview technique that can be used when sensitive questions have to be asked (Warner 1965; Chaudhuri and Mukerjee 1988). Examples of sensitive questions are questions about alcohol consumption, sexual behavior or fraud. Respondents might be reluctant to answer sensitive questions directly. RR techniques have in common that the true status of the individual respondent is not revealed since his observed answer depends not only on his status but also on a specified probability mechanism.

As an example, assume that the sensitive question is whether the respondent has committed fraud. The RR technique introduced by Boruch (1971), the *force response method*, goes as follows. After the sensitive question is asked, the respondent throws two dice and keeps the outcome hidden from the interviewer. If the outcome of the dice is 2, 3 or 4, the respondent answers *yes*. If the outcome 5, 6, 7, 8, 9 or 10, he answers according to the truth. If the outcome 11 or 12, he answers *no*. The observed answer depends both on the true status of the respondent and on the dice as the probability mechanism. Due to the use of the probability mechanism, the privacy of the individual respondent is guaranteed. RR techniques have been applied in the Netherlands (Van Gils, Van der Heijden, Laudy, and Ross 2003; Elffers, Van der Heijden, and Hezemans 2003).

Chen (1989) used the concept of misclassification to describe RR data that are collected using the RR design by Warner (1965). With each RR variable a transition matrix is associated that contains conditional misclassification probabilities. When the true status regarding the sensitive question is modeled by the discrete stochastic variable $X$ with sample space $\{1, 2, ..., J\}$, only the misclassified values of $X$ are observed. The misclassified version of $X$ is denoted $X^*$ and it has the same sample space as $X$. The distribution of $X^*$ is the $J$-component finite mixture given by

$$I\!P(X^* = i) = \sum_{j=1}^{J} I\!P(X^* = i|X = j) I\!P(A = j), \tag{4.1}$$

where $i = 1, ..., J$, and $I\!P(X^* = i|X = j)$ for all $i, j \in \{1, 2, ..., J\}$ are fixed and given by the transition matrix of $X$.

The basis for this paper is a survey where some respondents were asked sensitive questions about fraudulent behavior directly and others were asked the same questions using RR. The RR was performed using the forced response method. Besides the use of RR as a factor, the use of a PC was a second factor, i.e., some respondents were asked sensitive questions using a PC and others were asked the same questions without the use of a PC. Research questions concern the association between fraud and the PC/RR-classification. For instance, is fraud more easily admitted when RR is used?

Due to the fact that the use of RR induces a misclassification, standard statistical models cannot be used since they do not take into account the misclassification. This paper discusses the adjustment of the standard loglinear model so that the model can deal with the particulars of the current survey. The approach consists of two steps. First, the paper shows that the forced response method can be described using a transition matrix with conditional misclassification probabilities. Understanding the forced response method as a misclassification design generalizes for instance the two formulations of the likelihood of the logistic regression model for RR data in Van der Heijden and Van Gils (1996). Second, the paper shows that, with an adaptation regarding the misclassification, the current survey can be analyzed using the framework in Van den Hout and Van der Heijden (2004), in which the general RR design is described by a restrictive latent class model. To illustrate the theory, the paper discusses an example. The example includes a power analysis for testing loglinear models for RR data. The power analysis is important since using RR causes extra variability in the data and extra variability decreases power.

The outline of the paper is as follows. Section 4.2 introduces the survey. Section 4.3 describes the forced response method as a misclassification design. Section 4.4 discusses the estimation of loglinear models and shows how the design of the current survey fits into the latent class framework. Section 4.5 illustrates the theory by analyzing the RR data of the survey, and investigates the power of testing loglinear

Table 4.1: Observed classification by $F^*$, $P$ and $R$.

| | | $R$ | |
|---|---|---|---|
| $F^*$ | $P$ | 1 | 2 |
| Yes | 1 | 246 | 24 |
| | 2 | 246 | 24 |
| No | 1 | 628 | 226 |
| | 2 | 604 | 193 |

models. Section 4.6 concludes. Appendix 4.B contains code for the latent class software $\ell EM$ (Vermunt 1997).

## 4.2 The Survey

The sample size of the survey is $n = 2191$. All respondents receive a social benefit due to inability to work. Several sensitive questions were asked about whether the respondent committed fraud with respect to his or her social benefit. An example of fraud is doing odd jobs without reporting the extra income. The survey is a $2 \times 2$ factorial design where the 4 groups were formed according to two binary variables $P$ and $R$. Variable $P$ denotes the use of a PC, i.e., $P = 1$ means a PC was used, both to ask the questions and to answer them, and $P = 2$ means no PC was used. Variable $R$ denotes the use of RR by way of the forced response method, i.e., $R = 1$ means RR was used, and $R = 2$ means no RR was used.

The objective of the survey was to study whether the use of a PC and/or the use of RR is associated with reporting fraud. Note that in general RR surveys are expensive. There is extra variance due to the misclassification and this is reflected in the need for a relatively large sample size. When the use of a PC does not influence reporting fraud, cost of future RR surveys can be reduced by using PCs.

One of the RR question in the survey was ”Have you ever done some odd job for family or acquaintances and received money for this job without reporting the extra income?” This is a sensitive question since the reporting is obligatory and can influence the benefit. Let binary $F^*$ denote the observed answer with respect to this question. Table 4.1 presents observed frequencies in the table $F^*PR$. (The equality between the first two rows is not a typo.) Since all variables in the survey are discrete, loglinear analysis is the appropriate tool, see, e.g., Fienberg (1980). The standard loglinear model however does not take the misclassification due to the

RR design into account. The following sections discuss how to adjust the loglinear model when the use of RR is a factor. Note that adjustment does not concern $R$ itself, but the observed answers concerning fraud when $R = 1$.

## 4.3 Misclassification

Consider the forced response method (Boruch 1971) as described in the introduction. Let $X$ be the binary RR variable, and *yes* $\equiv 1$ and *no* $\equiv 2$. Given the forced response method where $D$ models the outcome of the sum of the two dice, it follows that

$$I\!\!P(X^* = 1) = I\!\!P(D = 2, 3 \text{ of } 4) + I\!\!P(X = 1)I\!\!P(D = 5, 6, 7, 8, 9 \text{ of } 10)$$

and

$$I\!\!P(X^* = 2) = I\!\!P(D = 11 \text{ of } 12) + I\!\!P(X = 2)I\!\!P(D = 5, 6, 7, 8, 9 \text{ of } 10).$$

An alternative formulation is

$$I\!\!P(X^* = i) = I\!\!P(X^* = i|X = 2)I\!\!P(X = 2) + I\!\!P(X^* = i|X = 1)I\!\!P(X = 1) \tag{4.2}$$

where $i = 1, 2$, and the conditional misclassification probabilities are given by the forced response method and the known distribution of the sum of the two dice. Note that (4.2) has the same structure as (5.2). It follows that the transition matrix of $X$ that contains the conditional misclassification probabilities $p_{ij} = I\!\!P(X^* = i|X = j)$ is given by

$$\boldsymbol{P}_X = \begin{pmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{pmatrix} = \begin{pmatrix} 11/12 & 2/12 \\ 1/12 & 10/12 \end{pmatrix}.$$

In $\boldsymbol{P}_X$, columns sum up to one.

From the fact that this forced response method can be described by misclassification and a transition matrix, it follows that the results in Van den Hout and Van der Heijden (2002, 2004) can be used to analyze data that have been collected using this method.

## 4.4 Loglinear Analysis

Van den Hout and Van der Heijden (2004) describe the misclassification due to RR using a restrictive latent class model and show how the software $\ell EM$ (Vermunt

1997) can be used to fit loglinear models for RR data. The use of $\ell EM$ follows from the fact that this environment is very suitable for the analysis of latent class models. In the current survey, the RR design is not always applied, i.e., some respondents answer the sensitive questions directly and others answer via RR. This section starts with the formulation of the misclassification using a transition matrix. This matrix is used in the formulation of the likelihood of the loglinear model. Second, the misclassification is formulated in such a way that it links up with latent class analysis and $\ell EM$. The discussion in this section handles the 2 × 2 factorial design in the survey, but from the presentation it will be clear how to generalize.

### 4.4.1 The Likelihood

Let $F$ denote the binary variable in the survey that models a sensitive question about fraud where *yes* $\equiv 1$ and *no* $\equiv 2$. The objective is to estimate loglinear models for the table $FPR$ in order to investigate how the use of a PC and/or the use of RR influences reporting fraud. Loglinear models describe cell probabilities in the $FPR$ table using main effects and interaction terms. The standard saturated model for the $FPR$ table denoted $(FPR)$ is given by

$$\log \pi_{fpr} = \lambda_0 + \lambda_f^F + \lambda_p^P + \lambda_r^R + \lambda_{fp}^{FP} + \lambda_{fr}^{FR} + \lambda_{pr}^{PR} + \lambda_{fpr}^{FPR}, \tag{4.3}$$

where $\pi_{fpr} = I\!P(F =, P = p, R = r)$, $f, p, r \in \{1, 2\}$. For identifiability $\lambda$ terms are constrained to sum to zero over any subscript,

$$\sum_{f=1}^{2} \lambda_f^F = 0, \qquad \sum_{f=1}^{2} \lambda_{fp}^{FP} = \sum_{p=1}^{2} \lambda_{fp}^{FP} = 0, \tag{4.4}$$

and so on. The term $\lambda_0$ is not a free parameter in the loglinear model but a normalizing constant chosen to make the cell probabilities sum to one,

$$\lambda_0 = -\log\left\{\sum_{fpr} \exp\left(\lambda_f^F + \lambda_p^P + .... + \lambda_{fpr}^{FPR}\right)\right\}.$$

Restricted, not saturated, loglinear models are defined by leaving out main effects or interaction effects. For instance, the model $(FR, PR)$ is defined by leaving out $\lambda_{fp}^{FP}$ and $\lambda_{fpr}^{FPR}$ for all $f, p, r \in \{1, 2\}$.

Due to the misclassification induced by the forced response method, table $FPR$ is latent, only frequencies in table $F^*PR$ are observed, where $F^*$ is the misclassified version of $F$. The misclassification design is given

$$\begin{aligned} &I\!\!P(F^* = f^*, F = f, P = p, R = r) = \\ &\qquad I\!\!P(F^* = f^* | F = f, R = r) I\!\!P(F = f, P = p, R = r), \end{aligned} \tag{4.5}$$

where $f^*, f, p, r \in \{1, 2\}$. The misclassification of $F$ is dependent on $R$ and this is different from the standard RR design, but easy to handle. Let $\boldsymbol{\pi} = (\pi_1, ..., \pi_8)^t$ denote the vector with cell probabilities for $FPR$ with the convention that the last index runs fastest: $\pi_1 = \pi_{111}^{FPR}$, $\pi_2 = \pi_{112}^{FPR}$, $\pi_3 = \pi_{121}^{FPR}$ and so on. It follows that the mixture due to the misclassification is given by

$$\boldsymbol{\pi}^* = \boldsymbol{P}_{FPR}\, \boldsymbol{\pi},$$

where $\boldsymbol{\pi}^* = (\pi_1^*, ..., \pi_8^*)^t$ is the vector with cell probabilities for $F^*PR$, and

$$\boldsymbol{P}_{FPR} = \begin{pmatrix} 11/12 & 0 & 0 & 0 & 2/12 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 11/12 & 0 & 0 & 0 & 2/12 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1/12 & 0 & 0 & 0 & 10/12 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1/12 & 0 & 0 & 0 & 10/12 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

Matrix $\boldsymbol{P}_{FPR}$ is *not* a Kronecker product of three transition matrices and is in that sense different from the misclassification discussed in Van den Hout and Van der Heijden (2004), in which a transition matrix per variable (possible the identity matrix) was specified and individual matrices where combined using a Kronecker product to describe the misclassification regarding cell probabilities.

In order to give the general formula of the likelihood, let $\boldsymbol{n} = (n_1, ..., n_8)^t$ denote the frequencies in the latent table $FPR$ and let $\boldsymbol{n}^* = (n_1^*, ..., n_8^*)^t$ denote the frequencies in table $F^*PR$. Frequencies $\boldsymbol{n}$ are assumed to be multinomially distributed with parameters $n$ and $\boldsymbol{\pi}$. It follows that $\boldsymbol{n}^*$ is also multinomially distributed with parameters $n$ and $\boldsymbol{\pi}^*$ due to the specific form of the transition matrix, see Van den Hout and Van der Heijden (2002, Section 5). We specify loglinear models by $\eta_d = \log \pi_d$, $d \in \{1, ..., D\}$, and $\boldsymbol{\eta} = (\eta_1, ..., \eta_D)^t = \boldsymbol{M\lambda}$, where $\boldsymbol{M}$ is the $D \times r$ design matrix that defines the loglinear model and $\boldsymbol{\lambda}$ is the $r \times 1$ parameter vector of the model. In the current survey $D = 8$. The kernel of the loglikelihood is given by

$$\begin{aligned} l(\boldsymbol{\lambda}|\boldsymbol{n}^*) &= \sum_{d=1}^{D} n_d^* \log(\pi_d^*) = \sum_{d=1}^{D} n_d^* \log(p_{d1}\pi_1 + ... + p_{dD}\pi_D) \\ &= \sum_{d=1}^{D} n_d^* \log(p_{d1}e^{\eta_1} + .. + p_{dD}e^{\eta_D}), \end{aligned} \tag{4.6}$$

where $p_{ij}$, $i, j \in \{1, ..., D\}$, are the entries of $\boldsymbol{P}_{FPR}$. Some additional remarks with respect to the maximizing of (4.6) and the testing of a model are given in Appendix 4.A.

### 4.4.2 The Latent Class Analysis Link

Following Goodman's (1974) notation, the latent class model of the misclassification in the current survey is given by

$$\pi^{F^*PR}_{f^*pr} = \sum_{f=1}^{2} \pi^{F^*FPR}_{f^*fpr} = \sum_{f=1}^{2} \pi^{FPR}_{fpr}\pi^{\overline{F^*}FR}_{f^*fr},$$

where $f^*, p, r \in \{1, 2\}$. This formulation follows directly from (4.5). The parameter $\pi^{\overline{F^*}FR}_{f^*fr}$ is the conditional response probability that an individual obtains score $F^* = f^*$, given that this person belongs to the latent class determined by $F = f$ and $R = r$. In standard latent class, analysis conditional response probabilities have to be estimated. When RR is used, these probabilities are fixed. Consequently, software for latent class analysis that allows for fixed conditional response probabilities can be used to analyze RR data.

In order to link up with the code for $\ell EM$, the conditional response probabilities can be described with the $2 \times 2 \times 2$ array

$$\begin{array}{|ll|}\hline p_{1|11} & p_{1|21} \\ p_{2|11} & p_{2|21} \\ \hline\end{array} \qquad \begin{array}{|ll|}\hline p_{1|12} & p_{1|22} \\ p_{2|12} & p_{2|22} \\ \hline\end{array}$$

equal to

$$\begin{array}{|ll|}\hline 11/12 & 2/12 \\ 1/12 & 10/12 \\ \hline\end{array} \qquad \begin{array}{|ll|}\hline 1 & 0 \\ 0 & 1 \\ \hline\end{array}$$

where $p_{f^*|fr} = \mathbb{P}(F^* = f^*|F = f, R = r)$ for $f^*, f, r \in \{1, 2\}$. The $\ell EM$ code for the example in the next section is given in Appendix 4.B.

## 4.5 Example

This section applies the foregoing in the loglinear analysis of the RR data in Table 4.1. Let binary $F$ denote the latent status with respect to the sensitive question, where $F = 1$ means that there are fraudulent activities and $F = 2$ means that there are no fraudulent activities. Section 4.5.1 discusses hypothesis testing and estimation. Section 4.5.2 investigates power.

### 4.5.1 Hypothesis Testing and Estimation

Goodness-of-fit statistics for different loglinear models for latent table $FPR$ are given in Table 4.2, where $X^2$ denotes the Pearson chi-square statistic and $L^2$ denotes the likelihood ratio statistic. Since the $2 \times 2$ factorial design is an experimental design where the 4 groups are formed using $P$ and $R$, we must include the $\lambda_{pr}^{PR}$ terms in all the models to be considered (Fienberg 1980, Chapter 6). In other words, we consider only the loglinear models where the estimated expected marginal totals for $P$ and $R$ are equal to the observed totals fixed by design.

A goodness-of-fit statistic is used to test the current model against the alternative of a saturated model. The $L^2$ statistic can also be used to compare nested models by partitioning the chi-square (Fienberg 1980, Section 4.3). When we choose significance level $\alpha$ equal to 0.10 and use the partitioning, we obtain $(FR, PR)$ as the best model for latent table $FPR$ ($\Delta$df=1, $\Delta L^2 = 0.36$, $p = 0.55$). In the next section the choice of $\alpha$ will be motivated.

Interpretation of $(FR, PR)$: There is an association between reporting fraud and the use of RR, but there is no association between reporting fraud and the use of a computer. As before, let *yes* $\equiv 1$ and *no* $\equiv 2$. Probabilities $\pi_{1|pr}^{F|PR} = I\!\!P(F = 1|P = p, R = r)$ for the 4 groups can be estimated using the fit of model $(FR, PR)$. We obtain

$$\begin{aligned} \hat{\pi}_{1|11} = \hat{\pi}_{1|21} &= \frac{\hat{\pi}_{121}}{\sum_{f=1}^{2} \hat{\pi}_{f21}} = 0.158 \\ \hat{\pi}_{1|12} = \hat{\pi}_{1|22} &= 0.103. \end{aligned}$$

The estimates show that when RR is used, more fraud is reported.

With respect to the interpretation, it is also possible to use a logit model instead of a loglinear model (Fienberg 1980, Chapter 6). The logit model takes explicitly into account that one of the variables is a dependent variable. The estimated logit model presents the results of the estimated loglinear model in a different way. The logit model that corresponds with the loglinear model $(FR, PR)$ is given by

$$\log\left(\pi_{1pr}^{FPR} \Big/ \pi_{2pr}^{FPR}\right) = w + w_r^R, \tag{4.7}$$

where $p, r \in \{1, 2\}$. The interpretation of the logit model is straightforward when the underlying loglinear model is made identifiable using dummy coding, i.e., the redundant $\lambda$ terms are set to zero. In the loglinear model this means

$$\lambda_2^F = \lambda_2^P = \lambda_2^R = \lambda_{12}^{FR} = \lambda_{21}^{FR} = \lambda_{22}^{FR} = \lambda_{12}^{PR} = \lambda_{21}^{PR} = \lambda_{22}^{PR} = 0.$$

Table 4.2: Goodness-of-fit statistics for loglinear models for latent table $FPR$.

| Model | df | $X^2$ | $p$ value | $L^2$ | $p$ value |
|---|---|---|---|---|---|
| 1. $(FPR)$ | 0 | 0.00 | - | 0.00 | - |
| 2. $(FP, FR, PR)$ | 1 | 0.04 | 0.83 | 0.04 | 0.83 |
| 3. $(FR, PR)$ | 2 | 0.40 | 0.82 | 0.40 | 0.82 |
| 4. $(FP, PR)$ | 2 | 6.81 | 0.03 | 7.04 | 0.03 |
| 5. $(F, PR)$ | 3 | 7.23 | 0.07 | 7.51 | 0.06 |
| 6. $(PR)$ | 4 | 750.96 | 0.00 | 802.24 | 0.00 |

The parameter estimates for (4.7) are $\widehat{w} = -2.167$ (0.152) and $\widehat{w}_1^R = 0.496$ (0.187), where the estimated standard errors are between the brackets. It follows that the odds $\pi_{1p1}^{FPR}/\pi_{2p1}^{FPR}$ is estimated to be 0.188, while the odds $\pi_{1p2}^{FPR}/\pi_{2p2}^{FPR}$ is estimated to be 0.115, for $p \in \{1, 2\}$. Estimated model $(FR, PR)$ implies that using RR has a positive effect on the odds of reporting fraud. When RR is used, the estimated odds of reporting fraud is $\exp(\widehat{w}_1^R) = 1.642$ times larger.

### 4.5.2 Power Analysis

The power of a test is defined as the probability that the null hypothesis $H_0$ is rejected given that the alternative hypothesis $H_A$ is true. By partitioning the chi-square in the previous section, we tested model $(FR, PR)$ against $(FP, FR, PR)$. In the present survey there is extra variability due to the RR perturbation. This extra variability causes a loss of power compared to a survey without RR. This is the reason to investigate in this section the power of testing $(FR, PR)$ against $(FP, FR, PR)$.

In general, power depends on the sample size, chosen significance level $\alpha$, and on the difference between $H_0$ and $H_A$. Given a fixed sample size and a chosen $\alpha$, power increases when the difference between $H_0$ and $H_A$ increases. In the following the model under $H_0$ will be $(FR, PR)$. In this model there is no interaction between $F$ and $P$, i.e., $\lambda_{11}^{FP} = 0$. The models under $H_A$ are defined by $(FP, FR, PR)$ where $\lambda_{11}^{FP}$ is fixed. By choosing different fixed values for $\lambda_{11}^{FP}$, the difference between $H_0$ and $H_A$ varies.

When model $(FP, FR, PR)$ is estimated, we obtain $\widehat{\lambda}_{11}^{FP} = -0.027$ (0.026), when the restrictions (4.4) are used. The Wald test can be used to test whether $\lambda_{11}^{FP}$ is equal to zero: $W = 0.027/0.026 = 1.04$, $p = 0.30$. So the difference between estimated models $(FR, PR)$ and $(FP, FR, PR)$ is small. Given this small difference and the fact that power is increased when significance level $\alpha$ is increased, we chose significance level $\alpha = 0.10$.

Simulation will be used to compute the power. Note that RR data are subject to two stochastic processes: the survey sampling and the misclassification design. Both processes are taken into account in the simulation. Let the model under $H_0$ be denoted $M_0$ and the model under $H_A$ be denoted $M_A$. One simulation goes as follows.

1. Estimate expected frequencies $\widehat{\boldsymbol{m}} = (\widehat{m}_1, ..., \widehat{m}_8)^t$ for latent table $FPR$ under $M_A$
2. Draw a sample $s$ from a multinomial distribution with parameters $\widehat{\boldsymbol{\pi}} = \widehat{\boldsymbol{m}}/n$ and $n$.
3. Simulate the RR design for $s$ to obtain simulated frequencies of table $F^*PR$.
4. Fit $M_0$ to simulated table $F^*PR$ and accept or reject $M_0$ using significance level $\alpha$.

The proportion of times $M_0$ is rejected by the goodness-of-fit test is the power of testing $M_0$ against $M_A$, given a choice of $\alpha$. As an illustration of the simulation, Figure 4.1 shows the distribution of the 1000 times simulated goodness-of-fit statistic $L^2$ when $M_0 = (FR, PR)$ and $M_A = (FP, FR, PR)$. The plotted line is the noncentral chi-square distribution with df $= 2$ and non-centrality parameter $\gamma = 0.36$. From Figure 4.1 can be concluded that the simulated $L^2$ follows the noncentral chi-square distribution, which is according to theory (compare Agresti 2002, Section 6.5.4.).

Next, the power of testing model $(FR, PR)$ against alternative models $(FP, FR, PR)$ with fixed $\lambda_{11}^{FP}$ is assessed. Figure 4.2 presents the simulated power curve for varying choices of $\lambda_{11}^{FP}$, where the goodness of fit is tested using $L^2$. Note that when $\lambda_{11}^{FP} = 0$, $(FR, PR)$ is tested against itself and the power equals $\alpha = 0.10$. The black circle in Figure 4.2 indicates the power of testing $(FR, PR)$ against $(FP, FR, PR)$ when fixed $\lambda_{11}^{FP}$ is equal to $\widehat{\lambda}_{11}^{FP} = -0.027$.

From the simulation results it follows that there is adequate power ($\geq 0.80$) in testing $(FR, PR)$ against $(FP, FR, PR)$ when $\lambda_{11}^{FP} \leq -0.13$ or $0.13 \leq \lambda_{11}^{FP}$. Interpretation of $\lambda_{11}^{FP}$ is via $\exp(4\lambda_{11}^{FP})$ which is the odds ratio between $F$ and $P$ (Fienberg 1980, Section 2.5). In model $(FR, PR)$, $\lambda_{11}^{FP} = 0$ and the odds ratio is 1, i.e., there is no association between $F$ and $P$. From the power analysis above, it follows that there is adequate power ($\geq 0.80$) in testing $(FR, PR)$ against $(FP, FR, PR)$ when the odds ratio between $F$ and $P$ lies outside the interval (0.60,1.68).

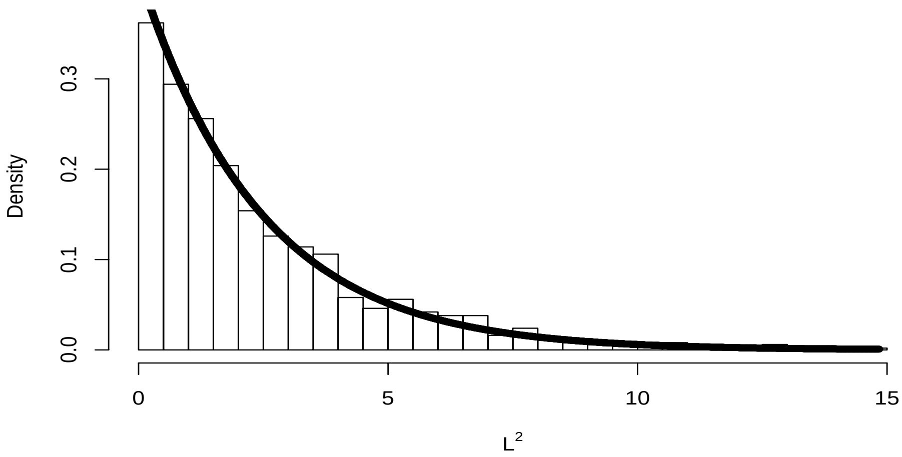


Figure 4.1: Distribution of $L^2$ when $M_0 = (FR, PR)$ and $M_A = (FP, FR, PR)$. Histogram of 1000 simulations and plotted chi-square distribution with df=2 and $\gamma = 0.36$.

The conclusion from the power analysis is that there is not enough power to test $(FR, PR)$ against $(FP, FR, PR)$. This is not a surprise given the small difference between estimated models $(FR, PR)$ and $(FP, FR, PR)$. Only with an odds ratio outside the interval (0.60,1.68) there is adequate power to test $(FR, PR)$ against $(FP, FR, PR)$.

## 4.6 Conclusion

From the formulation of the loglikelihood (4.6) it is clear that all kinds of models for the latent table can be estimated for RR data. The loglikelihood is constructed using the misclassification design with respect to all the variables. This design differs for differ RR designs, but as soon as the misclassification is specified, the method described in this paper can be used.

An important assumption in this discussion is that respondents follow the RR design. This assumption will not always be right. For instance, it might be that some respondents do not trust the privacy protection offered by the RR design and answer *no* irrespective of the outcome of the dice. These respondents bring about

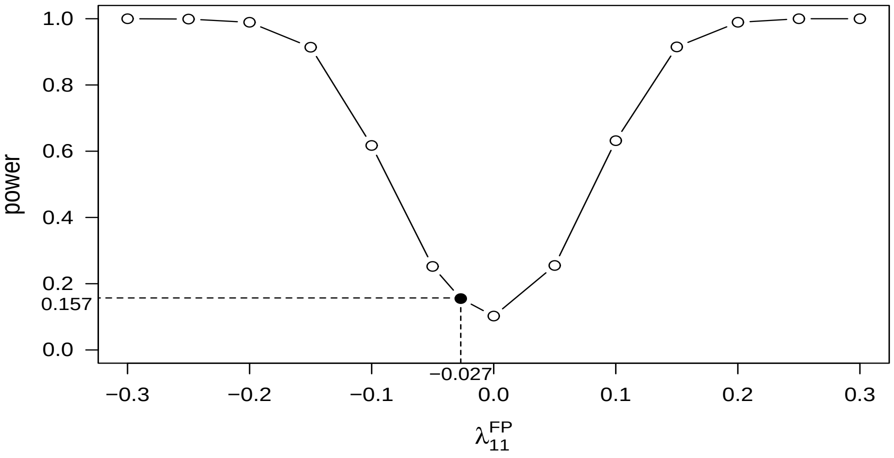


Figure 4.2: Power in testing $(FR, PR)$ against $(FP, FR, PR)$ given fixed values of $\lambda_{11}^{FP}$ and $\alpha = 0.10$. For each fixed value 1000 simulations were used.

a second perturbation besides the misclassification due to the RR design. Modeling this behavior is difficult. A idea for future work is to assign a prior distribution to the conditional misclassification probabilities to take into account the uncertainty of the misclassification process. Using Bayesian inference, the sensitivity of results for different prior distributions can then be investigated.

Asking sensitive questions will always produce incomplete or biased data. So one must make do with what one has got and the idea is that RR performs relatively well, see Van der Heijden, Van Gils, Bouts and Hox (2000). Analysis of RR data in the future might profit from research into more methodological aspects of RR designs, see Boeije and Lensvelt-Mulders (2002), who discuss cheating in RR designs. When respondents understand the privacy protection offered by the RR design better, the measurement will be more valid.

# Appendix 4.A

With respect to the maximization of the likelihood (4.6) and the usual constrains for cell probabilities, note that due to the loglinear transformation estimated probabilities are always positive and due to $\lambda_0$ estimated probabilities sum up to 1. When (4.6) is maximized to estimate a loglinear model, one has to estimate the fitted frequencies in table $F^*PR$ in order to test the model. In the example in Section 4.5 model $(FR, PR)$ is fitted. In this model, we have $D = 8$, $r = 6$, $\boldsymbol{\lambda} = (\lambda_0, \lambda_1^F, \lambda_1^P, \lambda_1^R, \lambda_{11}^{FR}, \lambda_{11}^{PR})^t$, and the design matrix is given by

$$\boldsymbol{M} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & -1 & -1 & -1 \\ 1 & 1 & -1 & 1 & 1 & -1 \\ 1 & 1 & -1 & -1 & -1 & 1 \\ 1 & -1 & 1 & 1 & -1 & 1 \\ 1 & -1 & 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 & -1 & -1 \\ 1 & -1 & -1 & -1 & 1 & 1 \end{pmatrix}.$$

Maximizing (4.6) yields $\widehat{\boldsymbol{\lambda}}$ and the fitted frequencies are given by

$$\widehat{\boldsymbol{m}}^* = n\boldsymbol{P}_{FPR} \ \exp(\boldsymbol{M}\widehat{\boldsymbol{\lambda}}).$$

# Appendix 4.B

The following presents code for $\ell EM$ that was used for the example in Section 4.5. Text after `*` is ignored by $\ell EM$. Different models can be fitted by choosing different `mod`-s. This input fits the logit model $(FR, PR)$ with dummy coding.

```
lat 1              * labels:
man 3              * F = fraud variable question, i.e., the latent variable
dim 2 2 2 2         * P = is PC used?
lab F P R G         * R = is RR used?, G = misclassified version of F

* Loglinear models:
* mod FPR {FPR}         * saturated model
* mod FPR {FR,FR}       * conditional independence

* Logit models:
* mod F|PR {FPR}        * saturated model
mod F|PR {FR,PR}        * conditional independence

G|FR {wei(GFR)}         * the misclassification

* Misclassification prob.  p111,p112,p211,p212,p121,p122,p221,p222, resp.:
sta wei(GFR) [.91667 0.08333 1 0 0.1666 0.8333 0 1]
```

```
dum 2 2 2 2                             * the dummy coding

dat [246 628 24 226 246 604 24 193]     * the data
```

# Chapter 5

# The Linear Regression Model

## 5.1 Introduction

Randomized response (RR) is an interview technique that can be used when sensitive questions have to be asked and respondents are reluctant to answer directly (Warner 1965; Chaudhuri and Mukerjee 1988). Examples of sensitive questions are questions about alcohol consumption, sexual behavior or fraud. RR variables can be seen as misclassified categorical variables where conditional misclassification probabilities are known. The misclassification protects the privacy of the individual respondent.

This paper applies the ideas in Spiegelman, Rosner, and Logan (2000) to iid normal linear regression models where some of the independent variables are subject to RR. Spiegelman et al. (2000) discuss the logistic regression model with misclassified independent variables. Their misclassification model, however, is different from the misclassification model induced by RR. This paper specifies the misclassification model of RR and shows how the misclassification can be taken into account in the maximum likelihood estimation of the linear regression model. Furthermore, as an alternative to Newton-Raphson maximization of the likelihood function an EM algorithm (Dempster, Laird, and Rubin 1977) is presented.

There is quite some literature about RR and the adjustment for the misclassification in the analysis, see, e.g., probability estimation in Chaudhuri and Mukerjee (1988), Bourke and Moran (1988), and Moors (1981), the logistic regression model with a RR dependent variable in Maddala (1983), and loglinear models in Chen (1989) and Van den Hout and Van der Heijden (2004). RR variables as independent

[1] This chapter is joint work with Peter Kooiman, CPB Netherlands Bureau for Economic Policy Analysis, the Netherlands.

variables, however, have not been dealt with. The possibility to include RR variables in regression models enlarges the possible application of RR. As an example, consider the situation where one variable depends on a second variable that models sexual behavior. When respondents are reluctant to answer about their behavior directly, RR can be used. In that case, a standard regression model is incorrect since it does not take into account the misclassification due to the use of RR.

A second field that may benefit from the discussion in this paper is statistical disclosure control. There is a similarity between RR designs and the post randomization method (PRAM) as a method for disclosure control of data matrices, see Gouweleeuw, Kooiman, Willenborg, and De Wolf (1998). Disclosure control aims at safeguarding the identity of respondents, see, e.g., Bethlehem, Keller, and Pannekoek (1990). When privacy is sufficiently protected, data producers, such as national statistical institutes, can safely pass on data to a third party. The idea of PRAM is to misclassify some of the categorical variables in the original data matrix and to release the perturbed data together with information about the misclassification mechanism. In this way PRAM introduces uncertainty in the data, i.e., the user of the data cannot be sure whether the individual information in the matrix is original or perturbed due to PRAM. Since the variables that are perturbed are typically independent variables such as, e.g., Gender, Ethnic Group, Region, it is important to know how to adjust regression models in order to take into account the misclassification. PRAM can be seen as a specific form of RR and the idea to use RR in this way goes back to the founder of RR, see Warner (1971). Similarities and differences between PRAM and RR are discussed in Van den Hout and Van der Heijden (2002).

The outline of the paper is as follows. Section 5.2 introduces the RR model. Section 5.3 discusses the linear regression model with RR independent variables. In Section 5.4 an EM algorithm is presented that maximizes the likelihood formulated in Section 5.3. Section 5.5 discusses the necessity of adjustment for misclassification and presents some simulation results. Section 5.6 concludes.

## 5.2 The Randomized Response Model

This section starts with the forced response design (Boruch 1971) as an example of a RR design and shows how the design can be seen as a misclassification design. Next, the section presents a general model for RR variables. The forced response design has recently been used in a Dutch survey into rule transgression, see Elffers, Van der Heijden, and Hezemans (2003).

Assume that the sensitive question asks for a *yes* or a *no*. The forced response design is as follows. After the sensitive question is asked, the respondent throws two dice and keeps the outcome hidden from the interviewer. If the outcome is 2, 3 or 4, the respondent answers *yes*. If the outcome 5, 6, 7, 8, 9 or 10, he answers according to the truth. If the outcome 11 or 12, he answers *no*.

Let $W$ be the binary RR variable that models the sensitive item, $W^*$ the binary variable that models the observed answer, and *yes* $\equiv 1$ and *no* $\equiv 2$. Given the forced response design where $D$ models the outcome of the sum of the two dice, it follows that

$$I\!\!P(W^* = 1) = I\!\!P(D = 2, 3 \text{ or } 4) + I\!\!P(W = 1)I\!\!P(D = 5, 6, 7, 8, 9 \text{ or } 10)$$

and

$$I\!\!P(W^* = 2) = I\!\!P(D = 11 \text{ or } 12) + I\!\!P(W = 2)I\!\!P(D = 5, 6, 7, 8, 9 \text{ or } 10).$$

An alternative formulation is $I\!\!P(W^* = w^*) = \sum_{j=1}^{2} I\!\!P(W^* = w^*|W = j)I\!\!P(W = j)$, where $w^* \in \{1, 2\}$, and the conditional probabilities are given by the forced response design and the known distribution of the sum of the two dice. This formulation shows that RR variables can be seen as misclassified variables. The transition matrix of $W$ that contains the conditional misclassification probabilities $p_{jk} = I\!\!P(W^* = j|W = k)$ for $j, k \in \{1, 2\}$ is given by

$$\boldsymbol{P}_W = \begin{pmatrix} p_{11} \; p_{12} \\ p_{21} \; p_{22} \end{pmatrix} = \begin{pmatrix} 11/12 & 2/12 \\ 1/12 & 10/12 \end{pmatrix}. \qquad (5.1)$$

For a general model for RR variables let $W^*$ denote the misclassified observed version of latent $W$. We assume that both $W^*$ and $W$ have the same set of categories, say $\{w_1, ..., w_J\}$. Let $\boldsymbol{P}_W$ denote the $J \times J$ nonsingular transition matrix that contains the conditional misclassification probabilities $p_{jk} = I\!\!P(W^* = w_j|W = w_k)$, for all $j, k \in \{1, ..., J\}$. Given the RR design, $\boldsymbol{P}_W$ is known. The distribution of $W^*$ is the $J$-component finite mixture given by

$$I\!\!P(W^* = w^*) = \sum_{j=1}^{J} I\!\!P(W^* = w^*|W = w_j)I\!\!P(W = w_j), \qquad (5.2)$$

for $w^* \in \{w_1, ..., w_J\}$. Since $\boldsymbol{P}_W$ is known, mixture (5.2) is a *known component density model*, see, e.g., Lindsay (1995). Let $\pi_j^* = I\!\!P(W^* = w_j)$ and $\pi_j = I\!\!P(W = w_j)$ for $j \in \{1, ..., J\}$. In matrix notation, we have $\boldsymbol{\pi^*} = \boldsymbol{P}_W\boldsymbol{\pi}$, where $\boldsymbol{\pi^*} = (\pi_1^*, ..., \pi_J^*)^t$ and $\boldsymbol{\pi} = (\pi_1, ..., \pi_J)^t$.

Let $W_{(i)}$ denote the variable $W$ for unit $i$, where $i = 1, ..., n$. Consider independent drawings $W_{(1)}$, $W_{(2)}$, ...,$W_{(n)}$, and assume $\sum_{j=1}^{J} \pi_j = 1$, where $\pi_j = I\!P(W_{(i)} = w_j)$, for $j \in \{1, ..., J\}$. Let $\boldsymbol{N} = (N_1, N_2, ..., N_J)^t$ be the frequency vector, i.e., $N_j = \sum_{i=1}^{n} \delta_{w_j}(W_{(i)})$, where $\delta_{w_j}(W_{(i)}) = 1$ if $W_{(i)} = w_j$ and $\delta_{w_j}(W_{(i)}) = 0$ if $W_{(i)} \neq w_j$. We assume that $\boldsymbol{N}$ is multinomially distributed with parameters $n$ and $\boldsymbol{\pi}$. Next, let $\boldsymbol{N}^* = (N_1^*, N_2^*, ..., N_J^*)^t$ be the frequency vector of the misclassified variables $W_{(1)}^*$, $W_{(2)}^*$, ..., $W_{(n)}^*$, and $\pi_j^* = I\!P(W_{(i)}^* = w_j)$. Due to the properties of $\boldsymbol{P}_W$, it follows that $\boldsymbol{N}^*$ is multinomially distributed with parameters $n$ and $\boldsymbol{\pi}^* = (\pi_1^*, ..., \pi_J^*)^t$. The loglikelihood for $\boldsymbol{\pi}$ is given by

$$l(\boldsymbol{\pi}|w_{(1)}^*, w_{(2)}^*, ..., w_{(n)}^*) \propto \sum_{j=1}^{J} n_j^* \log \sum_{k=1}^{J} p_{jk}\pi_k, \tag{5.3}$$

where $\sum_{k=1}^{J} \pi_k = 1$. When a maximum of (5.3) is in the interior of the parameter space, i.e., $\hat{\boldsymbol{\pi}} \in (0, 1)^J$, it is a global maximum and it is equal to the moment estimate $\boldsymbol{P}_W^{-1}\hat{\boldsymbol{\pi}}^*$, where $\hat{\boldsymbol{\pi}}^* = n^{-1}\boldsymbol{n}^*$, see Van den Hout and Van der Heijden (2002).

Extensions to more than one RR variable are direct. Say we have two binary variables $W_1$ and $W_2$ with values in $\{w_1^1, w_2^1\}$ and $\{w_1^2, w_2^2\}$, respectively. Assume that RR is independently applied to both variables. The above holds for the Cartesian product $\boldsymbol{W} = (W_1, W_2)$ where the transition matrix is given by

$$\boldsymbol{P}_{\boldsymbol{W}} = \boldsymbol{P}_{W_1} \otimes \boldsymbol{P}_{W_2} \tag{5.4}$$

where $\otimes$ denotes the Kronecker product. This follows from the assumption that the randomization is independent between variables, i.e.,

$$I\!P\Big(\boldsymbol{W}^* = (w_1^*, w_2^*)|\boldsymbol{W} = (w_1, w_2)\Big) = I\!P(W_1^* = w_1^*|W_1 = w_1) \times$$
$$I\!P(W_2^* = w_2^*|W_1 = w_2).$$

for $w_1^*, w_1, \in \{w_1^1, w_2^1\}$ and $w_2^*, w_2, \in \{w_1^2, w_2^2\}$. Similar expressions hold for more that two RR variables or RR variables with more than two categories.

## 5.3 Linear Regression

Without misclassification, the density of the scalar dependent variable $Y$ in the normal linear regression model is given by the density of the normal distribution with mean $\boldsymbol{x\beta}$ and variance $\sigma^2$, i.e., by

$$f(Y|\boldsymbol{x}) = \frac{1}{\sigma\sqrt{2\pi}} \exp\bigg( -\frac{1}{2}\Big(\frac{Y - \boldsymbol{x\beta}}{\sigma}\Big)^2\bigg),$$

where $\boldsymbol{x} = (x_1, ..., x_p)$ and $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^t$.

Next, let $\boldsymbol{X} = (\boldsymbol{W}, \boldsymbol{u})$, where $\boldsymbol{W} = (W_1, ..., W_q)$ and $\boldsymbol{u} = (u_{q+1}, ..., u_p)$ denote $q$ RR independent variables and $p - q$ non-RR independent variables, respectively, with $q < p$. Values of $\boldsymbol{W}$ are not observed, instead perturbed values of $\boldsymbol{W}^* = (W_1^*, ..., W_q^*)$ are given. Assume that both $\boldsymbol{W}$ and $\boldsymbol{W}^*$ have the same set of categories, say $\{\boldsymbol{w}_1, ..., \boldsymbol{w}_J\}$. Note that $J = J_1 \times ... \times J_q$, where $J_1, ..., J_q$ denote the number of categories of $W_1, ..., W_q$, respectively. Using the law of total probability it follows that

$$f(Y|\boldsymbol{w}^*, \boldsymbol{u}) = \sum_{j=1}^{J} f(Y|\boldsymbol{w}_j, \boldsymbol{u}) I\!\!P(\boldsymbol{W} = \boldsymbol{w}_j | \boldsymbol{W}^* = \boldsymbol{w}^*), \tag{5.5}$$

for $\boldsymbol{w}^* \in \{\boldsymbol{w}_1, ..., \boldsymbol{w}_J\}$. Note that we use $f(Y|\boldsymbol{w}, \boldsymbol{w}^*, \boldsymbol{u}) = f(Y|\boldsymbol{w}, \boldsymbol{u})$. The tacit assumption in (5.5) is that $I\!\!P(\boldsymbol{W} = \boldsymbol{w}|\boldsymbol{W}^* = \boldsymbol{w}^*, \boldsymbol{u}) = I\!\!P(\boldsymbol{W} = \boldsymbol{w}|\boldsymbol{W}^* = \boldsymbol{w}^*)$, i.e., $\boldsymbol{u}$ does not bear any information about the misclassification process. The formulation of this model is close to the model described in Spiegelman et al. (2000) for logistic regression with misclassified covariates, and following their terminology we call $I\!\!P(\boldsymbol{W} = \boldsymbol{w}|\boldsymbol{W}^* = \boldsymbol{w}^*)$ the *reclassification model*. Using Bayes' rule the reclassification model can be written as

$$I\!\!P(\boldsymbol{W} = \boldsymbol{w}|\boldsymbol{W}^* = \boldsymbol{w}^*) = \frac{I\!\!P(\boldsymbol{W}^* = \boldsymbol{w}^*|\boldsymbol{W} = \boldsymbol{w}) I\!\!P(\boldsymbol{W} = \boldsymbol{w})}{\sum_{j=1}^{J} I\!\!P(\boldsymbol{W}^* = \boldsymbol{w}^*|\boldsymbol{W} = \boldsymbol{w}_j) I\!\!P(\boldsymbol{W} = \boldsymbol{w}_j)}, \tag{5.6}$$

for $\boldsymbol{w}^*, \boldsymbol{w} \in \{\boldsymbol{w}_1, ..., \boldsymbol{w}_J\}$. The conditional probabilities $I\!\!P(\boldsymbol{W}^* = \boldsymbol{w}_j|\boldsymbol{W} = \boldsymbol{w}_k)$ for $j, k \in \{1, ..., J\}$ are the entries $p_{jk}$ in the transition matrix $\boldsymbol{P}_{\boldsymbol{W}}$.

In the standard iid normal linear regression model, independent variables are not stochastic variables. When RR is used, values of the independent variables $W_1, ..., W_q$ are latent and consequently $\boldsymbol{W}$ is a stochastic variable. Let $\pi_j = I\!\!P(\boldsymbol{W} = \boldsymbol{w}_j)$ for $j \in \{1, ..., J\}$. In this paper, $\boldsymbol{\pi} = (\pi_1, ..., \pi_J)^t$ is considered as a parameter that describes the sample distribution of $\boldsymbol{W}$, it does not describe the distribution of $\boldsymbol{W}$ in the population. Conditional on a sample $s$, we assume that the frequency vector $\boldsymbol{N}$ of $\boldsymbol{W}_{(1)}, \boldsymbol{W}_{(2)}, ..., \boldsymbol{W}_{(n)}$ is multinomially distributed in $s$ with parameters $n$ and $\boldsymbol{\pi}$, where $\pi_j = I\!\!P(\boldsymbol{W}_{(i)} = \boldsymbol{w}_j | i \in s)$ for $j \in \{1, ..., J\}$. In the following, the conditioning on the sample will be ignored in the notation.

Let the observed value of $\boldsymbol{W}^*$ of unit $i$ in the sample be denoted $\boldsymbol{w}^*_{(i)}$. The loglikelihood of the regression model with RR independent variables follows from (5.5), (5.6) and (5.3), and is for $n$ iid observations given by

$$l(\boldsymbol{\beta},\sigma,\boldsymbol{\pi}) \propto \sum_{i=1}^{n} \log \left\{ \sum_{j=1}^{J} \frac{1}{\sigma\sqrt{2\pi}} \exp\left( -\frac{1}{2}\Big(\frac{y_i - \boldsymbol{x}_i^j\boldsymbol{\beta}}{\sigma}\Big)^2 \right) \frac{p_{(i)j}\pi_j}{\sum_{k=1}^{J} p_{(i)k}\pi_k} \right\} + \sum_{j=1}^{J} n_j^* \log \sum_{k=1}^{J} p_{jk}\pi_k, \qquad (5.7)$$

where $\boldsymbol{x}_i^j = (\boldsymbol{w}_j, \boldsymbol{u}_i)$, $p_{(i)j} = I\!P(\boldsymbol{W}^* = \boldsymbol{w}_{(i)}^* | \boldsymbol{W} = \boldsymbol{w}_j)$, $n_j^*$ denotes the observed frequency of value $\boldsymbol{w}_j^*$ in the sample, and $\sum_{k=1}^{J} \pi_k = 1$.

From loglikelihood (5.7) it is clear that estimating the linear regression model with misclassified independent variables becomes rapidly complex when $J$ increases. Especially in the case of PRAM, where often several independent variables are perturbed, $J$ might be large and, consequently, implementing a straightforward Newton-Raphson algorithm to obtain the maximum of (5.7) will be quite a burden. The next section presents an alternative.

## 5.4 An EM Algorithm

This section presents an EM algorithm (Dempster, Laird, and Rubin 1977) for the linear regression model with RR independent variables. Loglikelihood (5.7) can be maximized using standard optimization software based on Newton-Raphson type algorithms. An alternative is to use an EM algorithm which is a stable algorithm and relatively easy to implement. An EM can also be used to find good starting values for a Newton-Raphson type algorithm.

The EM algorithm originates from the analysis of incomplete data. The problem of working with RR variables can be translated into an incomplete-data problem. Each observed value of $W^*$ is associated with its not observed not perturbed value of $W$. Together these pairs form an incomplete-data file with size $n$. In the framework of Rubin (1976), the missing data are missing at random, since they are missing by design. The E-step of EM finds the conditional expectation of the complete-data loglikelihood, denoted $Q(\boldsymbol{\phi}|\boldsymbol{\phi}^{(v)})$, where $\boldsymbol{\phi}^{(v)}$ is the current estimate of the parameter of interest. The M-step of EM determines $\boldsymbol{\phi}^{(v+1)}$ by maximizing $Q(\boldsymbol{\phi}|\boldsymbol{\phi}^{(v)})$ over $\boldsymbol{\phi}$.

The following presents an EM algorithm that maximizes the loglikelihood (5.7) over $\boldsymbol{\phi} = (\boldsymbol{\beta}, \sigma, \boldsymbol{\pi})$. The conditional expectation of the complete-data loglikelihood consists of two parts and is given by

$$Q(\boldsymbol{\phi}|\boldsymbol{\phi}^{(v)}) \propto Q_1(\boldsymbol{\beta},\sigma|\boldsymbol{\phi}^{(v)}) + Q_2(\boldsymbol{\pi}|\boldsymbol{\phi}^{(v)}), \qquad (5.8)$$

where

$$Q_1(\boldsymbol{\beta},\sigma|\boldsymbol{\phi}^{(v)}) = I\!\!E_{\boldsymbol{W}}\left[-n\log\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \boldsymbol{X}_i\boldsymbol{\beta})^2\middle| y_i, \boldsymbol{x}_i^*, i\in\{1,..,n\}, \boldsymbol{\phi}^{(v)}\right]$$

$$Q_2(\boldsymbol{\pi}|\boldsymbol{\phi}^{(v)}) = I\!\!E_{\boldsymbol{W}}\left[\sum_{j=1}^{J}\sum_{i=1}^{n}\delta_{\boldsymbol{w}_j}(\boldsymbol{W}_{(i)})\log\pi_j\middle| y_i, \boldsymbol{x}_i^*, i\in\{1,..,n\}, \boldsymbol{\phi}^{(v)}\right],$$

$\boldsymbol{X}_i = (\boldsymbol{W}_{(i)}, \boldsymbol{u}_i)$, and $\boldsymbol{x}_i^* = (\boldsymbol{w}_{(i)}^*, \boldsymbol{u}_i)$. Note that the first part is the expectation of the loglikelihood of a standard linear regression model, and that the second part is the expectation of the loglikelihood of the multinomially distributed frequency vector of $\boldsymbol{W}$. It follows that

$$Q_1(\boldsymbol{\beta},\sigma|\boldsymbol{\phi}^{(v)}) = -n\log\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\sum_{j=1}^{J} q_{ij}^{(v)}(y_i - \boldsymbol{x}_i^j\boldsymbol{\beta})^2, \tag{5.9}$$

where $q_{ij}^{(v)} = I\!\!P\left(\boldsymbol{W}_{(i)} = \boldsymbol{w}_j | y_i, \boldsymbol{x}_i^*, \boldsymbol{\phi}^{(v)}\right)$ and $\boldsymbol{x}_i^j = (\boldsymbol{w}_j, \boldsymbol{u}_i)$, for $i\in\{1,...,n\}$ and $j\in\{1,...,J\}$. The conditional distribution of $\boldsymbol{W}$ is given by

$$I\!\!P(\boldsymbol{W} = \boldsymbol{w}_k | \boldsymbol{w}^* = \boldsymbol{w}_j, y, \boldsymbol{u}, \boldsymbol{\phi}^{(v)}) = \frac{p_{jk} f(y|\boldsymbol{w}_k, \boldsymbol{u}, \boldsymbol{\beta}^{(v)}, \sigma^{(v)})\pi_k^{(v)}}{\sum_{l=1}^{J} p_{jl} f(y|\boldsymbol{w}_l, \boldsymbol{u}, \boldsymbol{\beta}^{(v)}, \sigma^{(v)})\pi_l^{(v)}}, \tag{5.10}$$

for $k, j\in\{1,...,J\}$.

The second part of (5.8) is given by

$$Q_2(\boldsymbol{\pi}|\boldsymbol{\phi}^{(v)}) = \sum_{j=1}^{J}\sum_{i=1}^{n}\sum_{k=1}^{J} q_{ik}^{(v)}\delta_{\boldsymbol{w}_j}(\boldsymbol{w}_k)\log\pi_j = \sum_{j=1}^{J}\sum_{i=1}^{n} q_{ij}^{(v)}\log\pi_j. \tag{5.11}$$

In the M-step, $Q(\boldsymbol{\phi}|\boldsymbol{\phi}^{(v)})$ is maximized over $\boldsymbol{\phi}$. This is done by maximizing the product $JQ(\boldsymbol{\phi}|\boldsymbol{\phi}^{(v)}) = JQ_1(\boldsymbol{\beta},\sigma|\boldsymbol{\phi}^{(v)}) + JQ_2(\boldsymbol{\pi}|\boldsymbol{\phi}^{(v)})$, where $J$ is the number of categories of $\boldsymbol{W}$. The maximization of the two parts can be done separately. Formula $JQ_1(\boldsymbol{\beta},\sigma|\boldsymbol{\phi}^{(v)})$ is the loglikelihood of a weighted linear regression model with sample size $n_0 = Jn$ and weights $r_{ij}^{(v)} = Jq_{ij}^{(v)}$. Maximizing $JQ_1(\boldsymbol{\beta},\sigma|\boldsymbol{\phi}^{(v)})$ over $(\boldsymbol{\beta},\sigma)$ is possible by weighted least squares where the weights $r_{ij}^{(v)}$ are computed using (5.10). Formula $JQ_2(\boldsymbol{\pi}|\boldsymbol{\phi}^{(v)})$ represents the likelihood of a pseudo multinomially distributed variable with $J$ categories and observed frequencies $r_{+j}^{(v)} = \sum_{i=1}^{n} r_{ij}^{(v)}$, where $\sum_{j=1}^{J} r_{+j}^{(v)} = nJ$. Note that in general $r_{+j}^{(v)}$ will not be an integer, hence the adjective pseudo.

The EM algorithm for the linear regression model with RR independent variables runs as follows:

*Initial estimate*: $\boldsymbol{\phi}^{(0)} = (\boldsymbol{\beta}^{(0)}, \sigma^{(0)}, \boldsymbol{\pi}^{(0)})$.

*E-step*: Compute $r_{ij}^{(v)} = J q_{ij}^{(v)}$ for $i = 1, ..., n$ and $j = 1, ..., J$, and create a weighted sample of size $n_0 = Jn$ where each unit $(y_i, \boldsymbol{w}_j, \boldsymbol{u}_i)$ has weight $r_{ij}^{(v)}$.

*M-step*: Construct the $n_0 \times 1$ matrix

$$\boldsymbol{Y} = (y_1, ..., y_1, y_2, ..., y_2, ......, y_n, ..., y_n)^t,$$

construct the $n_0 \times p$ matrix

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{w}_1 & \boldsymbol{u}_1 \\ \boldsymbol{w}_2 & \boldsymbol{u}_1 \\ \vdots & \vdots \\ \boldsymbol{w}_J & \boldsymbol{u}_1 \\ \boldsymbol{w}_1 & \boldsymbol{u}_2 \\ \vdots & \vdots \\ \boldsymbol{w}_J & \boldsymbol{u}_2 \\ \vdots & \vdots \\ \boldsymbol{w}_1 & \boldsymbol{u}_n \\ \vdots & \vdots \\ \boldsymbol{w}_J & \boldsymbol{u}_n \end{pmatrix},$$

and let $\boldsymbol{R}$ be the $n_0 \times n_0$ diagonal matrix

$$\boldsymbol{R} = \mathrm{Diag}\left(r_{11}^{(v)}, ..., r_{1J}^{(v)}, r_{21}^{(v)}, ..., r_{2J}^{(v)}, ......, r_{n1}^{(v)}, ..., r_{nJ}^{(v)}\right).$$

Compute

$$\begin{aligned} \boldsymbol{\beta}^{(v+1)} &= \left(\boldsymbol{X}^t \boldsymbol{R} \boldsymbol{X}\right)^{-1} \left(\boldsymbol{X}^t \boldsymbol{R} \boldsymbol{Y}\right) \\ \sigma^{(v+1)} &= \sqrt{\ } \left(\left(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}^{(v+1)}\right)^t \boldsymbol{R} \left(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}^{(v+1)}\right) / n_0\right) \\ \boldsymbol{\pi}^{(v+1)} &= \left(r_{+1}^{(v)}, ..., r_{+J}^{(v)}\right)^t / n_0. \end{aligned}$$

When perturbation due to the RR design is not too drastic, a reasonable choice for the initial estimates of $\boldsymbol{\beta}$ and $\sigma$ are the estimation results from a standard linear regression on the observed data. The moment estimate $\boldsymbol{P}_W^{-1} \hat{\boldsymbol{\pi}}^*$ can be used as an initial estimate of $\boldsymbol{\pi}$.

## 5.5 Simulation Example

The first part of this section gives an idea of the necessity of the adjustment to iid normal linear regression models when RR is applied. Data are simulated and the estimation of a linear regression model is discussed. The second part uses the same simulated data to illustrate the estimation of confidence intervals when the linear regression model is adjusted with respect to the misclassification.

### 5.5.1 Necessity of Adjustment

One may wonder whether it is necessary to adjust standard regression models as described in the preceding sections. When the probability of misclassification is small it might be that the influence of the misclassification is negligible. Especially in small samples, it might be that the extra variance due to RR is small in comparison with the variance of the regression. The following investigates this idea.

The plan of the simulation is to assess the regression model

$$\mathbb{E}(Y|\boldsymbol{x}) = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4, \tag{5.12}$$

where values of $x_2$ and $x_3$ are 0 or 1, and $x_4$ takes values on a continuous scale. The values of the independent variables are chosen as follows. Given the sample size $n$, there are $n2/10$ units $(x_2, x_3) = (0, 0)$, $n3/10$ units $(x_2, x_3) = (0, 1)$, $n3/10$ units $(x_2, x_3) = (1, 0)$, and $n2/10$ units $(x_2, x_3) = (1, 1)$. The values of $x_4$ are sampled from a normal distribution with mean 20 and $\sigma^2 = 4$. The correlation matrix of the independent variables is

$$\boldsymbol{CR} = \begin{pmatrix} 1 & -0.20 & -0.11 \\ -0.20 & 1 & -0.06 \\ -0.11 & -0.06 & 1 \end{pmatrix}.$$

One simulation of a sample with sample size $n$ goes as follows. For $i = 1, ..., n$ and given the chosen values of $\boldsymbol{x}_i = (1, x_{2i}, x_{3i}, x_{4i})$, value $y_i$ is sampled from a normal distribution with mean $\boldsymbol{x}_i\boldsymbol{\beta}$, where $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3, \beta_4)^t = (8, 4, 15, 8)^t$, and variance $\sigma^2 = 9$. To give an idea, for $n = 1000$ and one sample $y_1, ..., y_n$, the estimation of model (6.8) yields

$$\begin{array}{llllll} \widehat{\beta}_1 = 7.63 & (0.96) & \widehat{\beta}_3 = 14.91 & (0.19) & \widehat{\sigma} = 2.98 \\ \widehat{\beta}_2 = 4.21 & (0.19) & \widehat{\beta}_4 = 8.02 & (0.05) & \end{array} \tag{5.13}$$

where estimated standard errors are in parentheses.

Table 5.1: Actual coverage percentage (ACP) based on 500 simulations when regression model (6.8) is *not* adjusted for the misclassification due to RR.

| | | ACP given misclassification parameter $p_d$ | | |
|---|---|---|---|---|
| Sample size | Regression parameter | $p_d = 1$ | $p_d = \frac{9}{10}$ | $p_d = \frac{8}{10}$ |
| $n = 60$ | $\beta_2$ | 94.1 | 82.0 | 67.7 |
| | $\beta_3$ | 95.1 | 38.4 | 3.1 |
| | $\beta_4$ | 94.6 | 96.1 | 98.1 |
| $n = 100$ | $\beta_2$ | 95.0 | 71.0 | 49.8 |
| | $\beta_3$ | 95.5 | 10.6 | 0.0 |
| | $\beta_4$ | 94.8 | 96.2 | 97.4 |
| $n = 1000$ | $\beta_2$ | 95.0 | 0.0 | 0.0 |
| | $\beta_3$ | 95.3 | 0.0 | 0.0 |
| | $\beta_4$ | 94.6 | 90.4 | 96.0 |

The necessity of adjustment for the misclassification due to RR is investigated using the actual coverage percentages (ACPs) of the regression coefficients $\beta_2$, $\beta_3$, and $\beta_4$. After a number of simulations, the ACP of a parameter is the percentage of estimated confidence intervals that include the true value of the parameter. Three samples sizes are considered, namely $n = 60$, 100, and 1000. Table 5.1 shows the ACPs for different choices of misclassification probabilities and different sample sizes. Table 5.1 is obtained as follows. Given the choice of the sample size, one simulation consists of creating a sample $y_1, ..., y_n$ as describe above and simulating the misclassification due to RR for the values of $x_2$ and $x_3$. The transition matrix of both $x_2$ and $x_3$ is given by

$$\boldsymbol{P} = \begin{pmatrix} p_d & 1 - p_d \\ 1 - p_d & p_d \end{pmatrix}.$$

Next, a standard linear regression is performed using the sample $y_1, ..., y_n$ and the values of $x_2^*$, $x_3^*$, and $x_4$. For each of the regression coefficients a 95% confidence interval is estimated. The number of simulations is 500.

The ACP is a bit of a crude measure to assess the quality of the statistical

Table 5.2: Actual coverage percentage (ACP) based on 500 simulations when regression model (6.8) is adjusted for the misclassification due to RR.

| | | ACP given misclassification parameter $p_d$ | | |
|---|---|---|---|---|
| Sample size | Regression parameter | $p_d = 1$ | $p_d = \frac{9}{10}$ | $p_d = \frac{8}{10}$ |
| $n = 100$ | $\beta_2$ | 95.0 | 91.6 | 88.4 |
| | $\beta_3$ | 95.5 | 95.4 | 93.2 |
| | $\beta_4$ | 94.8 | 94.8 | 94.8 |
| $n = 1000$ | $\beta_2$ | 95.0 | 95.0 | 94.6 |
| | $\beta_3$ | 95.3 | 96.2 | 95.8 |
| | $\beta_4$ | 94.6 | 94.0 | 94.8 |

inference. However, it is clear from Table 5.1 that inference using the standard linear regression model is unreliable when there are RR independent variables. In the large sample $s_3$, the confidence intervals for $\beta_2$ and $\beta_3$ do not cover the true values when RR is applied with either $p_d = 9/10$ or $p_d = 8/10$. In the smaller samples, the results when RR used are a bit better but still not satisfactory. In all three samples $s_1$, $s_2$, and $s_3$, the perturbation due to the use of RR cannot be ignored.

It is hard to obtain an intuition about the specific unreliability caused by RR when the analysis is not adjusted for the misclassification. Some more research is needed to investigate for instance the increase of the ACP of $\beta_4$ in the samples $s_1$ and $s_2$. Is this increase caused by the fact that $x_4$ is not affected by the RR design? And: Is correlation with RR independent variables important? At the other hand, such an investigation is of limited use. It is more interesting to assess the estimation procedure which takes the misclassification into account. The simulation of this procedure is discussed in the next section.

### 5.5.2 Adjusting for Randomized Response

The following discusses the ACP for the parameters in the example in the previous section when the misclassification is taken into account.

The data are the simulated data in the previous section, but now the estimation of the regression model takes into account the misclassification due to RR by maximizing (5.7). The maximization routine is a combination of two routines. First, 20 EM steps are performed. Starting values are obtained by applying the standard linear regression model to the misclassified data. Next, the result of the EM is used as the starting point of the second routine which is the standard maximization routine `nlm()` in the programming environment R. This second routine is a Newton-Raphson type of maximization and yields estimated standard errors.

To give an idea of the increase of the standard errors due to RR, consider the following estimates for a simulated sample. The estimates were obtained by maximizing (5.7).

$$\begin{array}{llllll} \widehat{\beta}_1 = 8.03 & (1.25) & \widehat{\beta}_3 = 14.47 & (0.26) & \widehat{\sigma} = 3.09 \\ \widehat{\beta}_2 = 3.90 & (0.32) & \widehat{\beta}_4 = 8.02 & (0.06) & \end{array} \tag{5.14}$$

These estimates concern the same sample as in (5.13) albeit that this time RR was simulated with $p_d = 8/10$ for the independent variables $x_2$ and $x_3$. Comparing (5.14) with (5.13) shows that standard errors are increased, but that they stay within a reasonable range.

Table 5.2 presents results after 500 simulated samples for different sample sizes and different choices of $p_d$. The overall impression is that the adjustment works well when the sample size is large, see the results for $n = 1000$. For smaller samples, the simulations show that when $p_d$ decreases, the analysis becomes less reliable, see the results for $n = 100$. Apparently, the sample size has to be large so that the asymptotic properties of the maximum likelihood estimator dominate. Perturbation caused by the use of RR or PRAM implies that samples sizes have to larger than in standard situations without misclassification.

## 5.6 Conclusion

This paper presents a method to estimate the iid normal linear regression model with RR independent variables. An EM algorithm is presented as an alternative to Newton-Raphson maximization of the loglikelihood. In general, an EM algorithm is considered a stable but a somewhat slow maximization routine and when a Newton-Raphson type of algorithm is possible, it is preferred since it is faster and the estimation of standard errors is almost automatic. However, the present loglikelihood can be quite complex numerically. Especially in the case of PRAM, there might be a large number of perturbed independent variables some of which may have a large number of categories. Consider a model that includes the variables Gender (2 categories), Ethnic Background (4 categories), and Region (20 categories). Assume that PRAM is applied. Besides the regression parameters, there are $2 \times 4 \times 20 - 1 = 159$ nuisance parameters in the maximization of (5.7).

Using ready-made maximization routines to maximize the loglikelihood is an option, but these routines will also be sensitive to ill-chosen starting values. Furthermore, ready-made routines are blind with respect to the structure of the maximization problem. The EM presented in this paper explicitly uses the structure of the linear regression model. Since the maximization in the M-step is of closed form, the EM algorithm is actually quite fast. It is possible to use EM and extend the maximization with a general method that yields estimated standard errors in EM maximization, see, e.g., the discussion of the supplemented EM in Little and Rubin (2002, Section 9.2).

The simulations in Section 5.5 show that the perturbation caused by using RR or PRAM cannot be ignored. Furthermore, the simulations demonstrate that the method presented in this paper is feasible and that adjustment for the perturbation is possible. However, sample sizes have to be larger than in the standard situation without perturbation. Protecting privacy is not for free.

Estimation of the parameters in the linear regression model is only the first step in fitting a linear regression model to data. Future research should address possibilities to check some of the assumptions of the model. Can outliers for instance be detected when some of the independent variables are subject to misclassification?

Although this paper only discusses the iid normal linear regression model, it might be interesting to investigate the approach for more sophisticated regression models. The reclassification model will stay the same, but it might be that a straightforward EM is not possible anymore.

# Chapter 6

# Variants of PRAM and Measures for Disclosure Risk

## 6.1 Introduction

The post randomization method (PRAM) is discussed in Gouweleeuw, Kooiman, Willenborg, and De Wolf (1998) as a method for statistical disclosure control (SDC). When survey data are released by statistical agencies, SDC protects the identities of respondents. SDC tries to prevent that a user of released data can link data of a respondent in the survey to a specific person in the population. See Willenborg and De Waal (2001) for an introduction into SDC and SDC methods other than PRAM.

There is a close link between PRAM and randomized response, a method to ask sensitive questions in a survey, see Warner (1965) and Rosenberg (1979). Van den Hout and Van der Heijden (2002) sum up some differences and similarities between randomized response and PRAM.

When SDC is used, there will always be a loss of information. This is inevitable since SDC tries to determine the information in the data that can lead to the disclosure of an identity of a respondent, and eliminates this information before data are released. It is not difficult to prevent disclosure, but it is difficult to prevent disclosure *and* release data that is still useful for statistical analysis. Applying SDC means searching for a balance between disclosure risk and information loss.

The idea of PRAM is to misclassify some of the categorical variables in the survey using fixed misclassification probabilities and to release the partly misclassified data together with those probabilities. Say variable $X$, with categories $\{1, ..., J\}$,

[1] This chapter is joint work with Elsayed Elamir, University of Southampton, United Kingdom.

is misclassified into variable $X^*$. The survey containing $X^*$ but not $X$ is released together with conditional probabilities $I\!\!P(X^* = k|X = j)$, for $k, j \in \{1, ..., J\}$. In this way, PRAM introduces uncertainty in the data: The user of the released data cannot be sure that the information is original or perturbed due to PRAM and it becomes harder to establish a correct link between a respondent in the survey and a specific person in the population. Since the user has the misclassification probabilities, he can adjust his analysis by taking into account the perturbation due to PRAM.

This paper discusses two ideas to make PRAM more efficient with respect to the balance between disclosure risk and information loss. First, the paper discusses the use of *calibration probabilities*

$$I\!\!P(\text{true category is } j|\text{category } i \text{ is released}). \tag{6.1}$$

in the analysis of released data and compares this with using *misclassification probabilities*

$$I\!\!P(\text{category } i \text{ is released}|\text{true category is } j). \tag{6.2}$$

The idea of using calibration probabilities is discussed by De Wolf, Gouweleeuw, Kooiman, and Willenborg (1997), who refer to the discussion of calibration probabilities in misclassification literature, see, e.g., Kuha and Skinner (1997). We will elaborate the discussion and show that the advantage of calibration probabilities is limited to the univariate case. Secondly, the paper shows that information loss can be reduced by providing *misclassification proportions* along with the released data. These proportions inform about the actual change in the survey data due to the application of PRAM. (Probabilities (6.1) and (6.2) inform about the expected change.) In addition, the paper discusses two measures for disclosure risk when PRAM is applied. The first is an extension of the measure introduced by Skinner and Elliot (2002). The second measure links up with the SDC practice at Statistic Netherlands. Simulation results are given to illustrate the theory.

The outline of the paper is as follows. Section 6.2 provides the framework and the notation. Section 6.3 describes frequency estimation for PRAM data. Section 6.4 discusses the use of calibration probabilities. In Section 6.5, we introduce the use of misclassification proportions. Section 6.6 discusses measures for disclosure risk, whereas information loss is briefly considered in Section 6.7. Section 6.8 presents some simulations, and Section 6.9 concludes.

## 6.2 Framework and Notation

In survey data, we distinguish between *identifying variables* and *non-identifying variables*. Identifying variables are variables that can be used to re-identify individuals represented in the data. These variables are assumed to be categorical, e.g., Gender, Race, Place of Residence. We assume that the sensitive information of respondents is contained in the non-identifying variables, see Bethlehem, Keller, and Pannekoek (1990), and that we want to protect this information by applying PRAM to (a subset of) the identifying variables.

The notation in this paper is the same as in Skinner and Elliot (2002). Units are selected from a finite population $U$ and each selected unit has one record in the *microdata sample* $s \subset U$. Let $n$ denote the number of units in $s$. Let the categorical variable formed by cross-classifying (a subset of) the identifying variables be denoted $X$ with values in $\{1, ..., J\}$. Let $X_i$ denote the value of $X$ for unit $i \in U$. The *population frequencies* are denoted

$$F_j = \sum_{i \in U} I(X_i = j), \qquad j \in \{1, ..., J\},$$

where $I(\cdot)$ is the indicator function: $I(A) = 1$ if $A$ is true and $I(A) = 0$ otherwise. The *sample frequencies* are denoted

$$f_j = \sum_{i \in s} I(X_i = j), \qquad j \in \{1, ..., J\}.$$

In the framework of PRAM, we call the sample that is released by the statistical agency the *released microdata sample* $s^*$. Note that unit $i \in s^*$ if and only if $i \in s$. Let $X^*$ denote the released version of $X$ in $s^*$. By *misclassification* of unit $i$ we mean $X_i \neq X_i^*$. The *released sample frequencies* are denoted

$$f_k^* = \sum_{i \in s^*} I(X_i^* = k), \qquad k \in \{1, ..., J\}.$$

Let $\boldsymbol{P}_X$ denote the $J \times J$ transition matrix that contains the conditional misclassification probabilities $p_{kj} = I\!P(X^* = k | X = j)$, for $k, j \in \{1, ..., J\}$. Note that the columns of $\boldsymbol{P}_X$ sum up to one. The distribution of $X^*$ conditional on $s$ is the $J$-component finite mixture given by

$$I\!P(X_i^* = k | i \in s) = \sum_{j=1}^{J} I\!P(X_i^* = k | X_i = j) I\!P(X_i = j | i \in s), \qquad k \in \{1, ..., J\},$$

where the component distributions are given by $\boldsymbol{P}_X$ and the component weights are given by the conditional distribution of $X$. The conditional distribution of $X$ in sample $s$ is given by

$$I\!\!P(X_i = j | i \in s) = \frac{1}{n} f_j, \qquad j \in \{1, ..., J\}.$$

## 6.3 Frequency Estimation for PRAM Data

When PRAM is applied and some of the identifying variables are misclassified, standard statistical models do not apply to the released data since these models do not take into account the perturbation. This section shows how the misclassification can be taken into account in frequency estimation.

We have $I\!\!E[\boldsymbol{F}^* | \boldsymbol{f}] = \boldsymbol{P}_X \boldsymbol{f}$, where $\boldsymbol{f} = (f_1, ..., f_J)^t$ and $\boldsymbol{F}^* = (F_1^*, ..., F_J^*)^t$ is the stochastic vector of the released sample frequencies. An unbiased moment estimator of $\boldsymbol{f}$ is given by

$$\hat{\boldsymbol{f}} = \boldsymbol{P}_X^{-1} \boldsymbol{f}^*, \tag{6.3}$$

see Kooiman, Willenborg, and Gouweleeuw (1997). In practice, assuming that $\boldsymbol{P}_X$ is non-singular does not impose much restriction on the choice of the misclassification probabilities. Matrix $\boldsymbol{P}_X^{-1}$ exists when the diagonal of $\boldsymbol{P}_X$ dominates, i.e., $p_{ii} > 1/2$ for $i \in \{1, ..., J\}$. An additional assumption in (6.3) is that the dimensions of $\boldsymbol{f}$ and $\boldsymbol{f}^*$ are the same.

PRAM is applied to each variable independently and a transition matrix is released per variable. When the user of the released sample assesses a compounded variable, he can construct its transition matrix using the transition matrices of the individual variables. For instance, consider identifying variables $X_1$, with categories $\{1, .., J_1\}$ and $X_2$, with categories $\{1, .., J_2\}$, and the cross-classification $X = (X_1, X_2)$, i.e., the Cartesian product of $X_1$ and $X_2$. Since PRAM is applied independently, we have

$$\begin{aligned} I\!\!P\Big(X^* = (k_1, k_2) | X = (j_1, j_2)\Big) &= I\!\!P(X_1^* = k_1 | X_1 = j_1) \\ &\times \; I\!\!P(X_2^* = k_2 | X_1 = j_2), \end{aligned} \tag{6.4}$$

for $k_1, j_1 \in \{1, .., J_1\}$ and $k_2, j_2 \in \{1, .., J_2\}$. In matrix notation, we have $\boldsymbol{P}_X = \boldsymbol{P}_{X_1} \otimes \boldsymbol{P}_{X_2}$, where $\otimes$ is the Kronecker product. Note that when one of two variables is not perturbed by PRAM, the transition matrix of that variable is the identity matrix.

The variance of (6.3) equals

$$V[\hat{\boldsymbol{f}}|\boldsymbol{f}] = \boldsymbol{P}_X^{-1} V[\boldsymbol{F}^*|\boldsymbol{f}](\boldsymbol{P}_X^{-1})^t = \boldsymbol{P}_X^{-1}\Big(\sum_{j=1}^{J} f_j \boldsymbol{V}_j\Big)(\boldsymbol{P}_X^{-1})^t \tag{6.5}$$

where $\boldsymbol{V}_j$ is the $J \times J$ covariance matrix of two released values given the original value $j$, i.e.,

$$\boldsymbol{V}_j(k_1, k_2) = \begin{cases} p_{k_2 j}(1 - p_{k_2 j}) \text{ if } k_1 = k_2 \\ \\ \qquad -p_{k_1 j} p_{k_2 j} \text{ if } k_1 \neq k_2 \end{cases} \qquad \text{for } k_1, k_2 \in \{1, ..., J\},$$

see Kooiman et al. (1997). The variance can be estimated by substituting $\hat{f}_j$ for $f_j$ in (6.5), for $j \in \{1, .., J\}$.

The variance given by (6.5) is the extra variance due to PRAM and does not take into account the sampling distribution. The formulas for the latter are given in Chaudhuri and Mukerjee (1988) for multinomial sampling and compared to (6.5) in Van den Hout and Van der Heijden (2002), see also Appendix 6.B.

## 6.4 Calibration Probabilities

Literature concerning misclassification shows that calibration probabilities (6.1) are more efficient in the analysis of misclassified data than misclassification probabilities (6.2), see the review paper by Kuha and Skinner (1997). Often, calibration probabilities have to be estimated. However, when PRAM is applied, the statistical agency can compute the calibration probabilities using the sample frequencies. The idea of using calibration probabilities for PRAM is mentioned in De Wolf et al. (1997). The following elaborates this idea and makes a comparison with PRAM as explained in the previous section.

The $J \times J$ matrix with calibration probabilities of univariate variable $X$ is denoted by $\overleftarrow{\boldsymbol{P}}_X$ and has entries $\overleftarrow{p}_{jk}$ defined by

$$\mathbb{P}(X_i = j | X_i^* = k, i \in s) = \frac{p_{kj} f_j}{\sum_{j_0=1}^{J} p_{kj_0} f_{j_0}}, \qquad j, k \in \{1, ..., J\}, \tag{6.6}$$

where $p_{kj}$ are the entries of $\boldsymbol{P}_X$. Matrix $\overleftarrow{\boldsymbol{P}}_X$ is again a transition matrix; each column sums up to one. We have

$$\boldsymbol{f} = \overleftarrow{\boldsymbol{P}}_X \mathbb{E}[\boldsymbol{F}^*|\boldsymbol{f}], \tag{6.7}$$

see Appendix 6.A. An unbiased moment estimator of $\boldsymbol{f}$ is therefore given by

$$\tilde{\boldsymbol{f}} = \overleftarrow{\boldsymbol{P}}_X \boldsymbol{f}^*. \tag{6.8}$$

In general, $\overleftarrow{\boldsymbol{P}}_X \neq \boldsymbol{P}_X^{-1}$, see Appendix 6.A. The variance of (6.8) is given by (6.5) where $\boldsymbol{P}_X^{-1}$ is replaced by $\overleftarrow{\boldsymbol{P}}_X$ and $f_j$ is estimated by $\tilde{f}_j$, for $j \in \{1, ..., J\}$.

In the remainder of this section we compare estimators (6.3) and (6.8). The first difference is that (6.3) might yield an estimate where some of the entries are negative, whereas (6.8) will never yield negative estimates, see, e.g., De Wolf et al. (1997).

Secondly, estimator (6.8) is more efficient than (6.3) in the univariate case. This is already discussed in Kuha and Skinner (1997). Consider the case where $X$ has two categories. Say we want to know $\pi = I\!P(X = 1)$. Let $\hat{\pi}$ be the estimate using $\boldsymbol{P}_X$ and $\tilde{\pi}$ the estimate using $\overleftarrow{\boldsymbol{P}}_X$. The efficiency of $\hat{\pi}$ relative to $\tilde{\pi}$ is given by

$$\text{eff}(\hat{\pi}, \tilde{\pi}) = \frac{V[\tilde{p}]}{V[\hat{p}]} = (p_{11} + p_{22} - 1)^2 (\overleftarrow{p}_{22} - \overleftarrow{p}_{21})^2 < 1. \tag{6.9}$$

So $\tilde{\pi}$ is always more efficient than $\hat{\pi}$. An important difference with the general situation of misclassification is that in the situation of PRAM, matrices $\boldsymbol{P}_X$ and $\overleftarrow{\boldsymbol{P}}_X$ are given and do not have to be estimated. Comparison (6.9) is therefore a simple form of the comparison in Kuha and Skinner (1997, Section 28.5.1.3.).

The third comparison is between the maximum likelihood properties of (6.3) and (6.8). Assume that the frequency vector $\boldsymbol{f}$ of $X_1, ...X_n$ is multinomially distributed with parameters $n$ and $\boldsymbol{\pi} = (\pi_1, .., \pi_J)^t$. In the framework of misclassification, Hochberg (1977) proves that estimator (6.8) yields an MLE. When (6.3) yields an estimate in the interior of the parameter space, the estimate is also an MLE. See Appendix 6.B for the maximum likelihood properties of (6.8) and (6.3). The likelihood function corresponding to (6.8) is different from the likelihood function corresponding to (6.3), since the information used is different. This explains why both can be an MLE despite being different estimators of $\boldsymbol{f}$.

The fourth comparison is with respect to transition matrices of Cartesian products and is less favorable for (6.8). It has already been noted that $\boldsymbol{P}_{X_1} \otimes \boldsymbol{P}_{X_2}$ is the matrix with misclassification probabilities for the Cartesian product $X = (X_1, X_2)$, see (6.4). Analogously, given $\overleftarrow{\boldsymbol{P}}_{X_1}$ and $\overleftarrow{\boldsymbol{P}}_{X_2}$ the user can construct matrix $\overleftarrow{\boldsymbol{P}}_{X_1} \otimes \overleftarrow{\boldsymbol{P}}_{X_2}$. However, this matrix does *not* necessarily contain calibration probabilities for $X$. Note that

$$\begin{aligned}\mathbb{P}\Big(X_i = (j_1, j_2)|X_i^* = (k_1, k_2), i \in s\Big) \\ = \frac{p_{k_1 j_1} p_{k_2 j_2} \mathbb{P}\Big(X_i = (j_1, j_2)|i \in s\Big)}{\sum_v^{J_1} \sum_w^{J_2} p_{k_1 v} p_{k_2 w} \mathbb{P}\Big(X_1 = (v, w)|i \in s\Big)}.\end{aligned} \tag{6.10}$$

It follows that $\overleftarrow{\boldsymbol{P}}_X = \overleftarrow{\boldsymbol{P}}_{X_1} \otimes \overleftarrow{\boldsymbol{P}}_{X_2}$ when $X_1$ and $X_2$ are independent. In general, this independence is not guaranteed and since the user of the released data does not have the frequencies of $X$, he cannot construct $\overleftarrow{\boldsymbol{P}}_X$.

The fifth and last comparison is with respect to the creation of subgroups. Consider the situation where a user of the released data creates a subgroup by using a grouping variable that is not part of $X$. When the number of categories in the subgroup is smaller than $J$, estimate (6.8) is not well-defined. When the number of categories is equal to $J$, estimate (6.8) is biased due to the fact that (6.7) does not hold. Note with respect to (6.7) that the frequencies that are used to construct $\overleftarrow{\boldsymbol{P}}_X$ are the frequencies in the whole sample which will differ from the frequencies in the subgroup, see also Appendix 6.A. Estimator (6.3) is still valid for the subgroup.

Since calibration probabilities contain information about the distribution of the sample $s$, they perform better than misclassification probabilities regarding the univariate case. However, in a multivariate setting this advantage may disappear. Section 6.8 presents some simulation results.

## 6.5 Misclassification Proportions

Matrices $\boldsymbol{P}_X$ and $\overleftarrow{\boldsymbol{P}}_X$ inform about the expected change due to PRAM. As an alternative, we can create transition matrices that inform about the actual change due an application of PRAM. These matrices contain proportions and will be denoted $\boldsymbol{P}_X^\circ$ and $\overleftarrow{\boldsymbol{P}}_X^\circ$. Matrix $\boldsymbol{P}_X^\circ$ contains *misclassification proportions* and $\overleftarrow{\boldsymbol{P}}_X^\circ$ contains *calibration proportions*. This section shows how $\boldsymbol{P}_X^\circ$ and $\overleftarrow{\boldsymbol{P}}_X^\circ$ are computed and discusses properties of these matrices.

We start with an example. Say that $X$ has categories {1,2}. Assume that applying PRAM yields the cross-classification in Table 6.1. From this table it follows that the proportion of records with $X = 1$ that have $X^* = 1$ in the released sample is 300/400=3/4 and that the proportion of records with $X^* = 1$ that have $X = 1$ in the original sample is 300/500=3/5. Analogously we get the other entries of

$$\boldsymbol{P}_X^\circ = \begin{pmatrix} 3/4 & 1/3 \\ 1/4 & 2/3 \end{pmatrix} \qquad \text{and} \qquad \overleftarrow{\boldsymbol{P}}_X^\circ = \begin{pmatrix} 3/5 & 1/5 \\ 2/5 & 4/5 \end{pmatrix}.$$

Table 6.1: Classification by $X^*$ and $X$

|  | $X$ |  |  |
|---|---|---|---|
| $X^*$ | 1 | 2 | Total |
| 1 | 300 | 200 | 500 |
| 2 | 100 | 400 | 500 |
| Total | 400 | 600 | 1000 |

For the general construction of $\boldsymbol{P}^{\circ}_X$ and $\overleftarrow{\boldsymbol{P}}^{\circ}_X$, let the cell frequencies in the cross-classification $X^*$ by $X$ be denoted $c_{kj}$, for $k, j \in \{1, .., J\}$. The entries of the $J \times J$ transition matrices with the proportions are given by

$$p^{\circ}_{kj} = \frac{c_{kj}}{f_j} \qquad \text{and} \qquad \overleftarrow{p}^{\circ}_{jk} = \frac{c_{kj}}{f^*_k},$$

where $k, j \in \{1, .., J\}$.

It follows that $\boldsymbol{f}^* = \boldsymbol{P}^{\circ}_X \boldsymbol{f}$ and $\boldsymbol{f} = \overleftarrow{\boldsymbol{P}}^{\circ}_X \boldsymbol{f}^*$. This is the reason to consider the matrices with the proportions more closely, since it is a great improvement compared to (6.3) and (6.8). Note that when the user of the released sample has $\boldsymbol{P}^{\circ}_X$ or $\overleftarrow{\boldsymbol{P}}^{\circ}_X$, he can reconstruct Table 6.1.

Conditional on $\boldsymbol{f}$, $\boldsymbol{P}^{\circ}_X$ and $\overleftarrow{\boldsymbol{P}}^{\circ}_X$ are stochastic, whereas $\boldsymbol{P}_X$ and $\overleftarrow{\boldsymbol{P}}_X$ are not. In expectation $\boldsymbol{P}^{\circ}_X$ equals $\boldsymbol{P}_X$, and $\boldsymbol{P}^{\circ}_{X_1} \otimes \boldsymbol{P}^{\circ}_{X_2}$ equals $\boldsymbol{P}_{X_1} \otimes \boldsymbol{P}_{X_2}$, see Appendix 6.C. However, since $\boldsymbol{f}^*$ is a value of the stochastic vector $\boldsymbol{F}^*$, and $\mathbb{P}(F^*_k = 0) \neq 0$, the expectation of $\overleftarrow{\boldsymbol{P}}^{\circ}_X$ does not exists. Nevertheless, an approximation shows that $\overleftarrow{\boldsymbol{P}}^{\circ}_X$ will be close to $\overleftarrow{\boldsymbol{P}}_X$, see Appendix 6.C.

There is a set back with respect to the use of proportions for Cartesian products and this is comparable to the problem mentioned in the previous section. Given $\boldsymbol{P}^{\circ}_{X_1}$ and $\boldsymbol{P}^{\circ}_{X_2}$, the user can construct $\boldsymbol{P}^{\circ}_{X_1} \otimes \boldsymbol{P}^{\circ}_{X_2}$ for $X = (X_1, X_2)$. However, $\boldsymbol{P}^{\circ}_{X_1} \otimes \boldsymbol{P}^{\circ}_{X_2}$ does *not* contain proportions as defined above. Note that the user does not have the cross-classification of $X$ and $X^*$, so he cannot derive the proportions in $\boldsymbol{P}^{\circ}_X$. The same holds for $\overleftarrow{\boldsymbol{P}}^{\circ}_X$. The optimal use of misclassification proportions is thereby limited to the univariate case.

Since misclassification proportions contain information about the actual perturbation due to PRAM, we expect them to perform well also in the multivariate case. Section 6.8 discusses a multivariate example.

## 6.6 Disclosure Risk

There are several ways to measure disclosure risk (Skinner and Elliot 2002; Domingo-Ferrer and Torra 2001). This section discusses two measures for disclosure risk with respect to PRAM. Section 6.6.1 discusses an extension of the general measure of disclosure risk introduced by Skinner and Elliot (2002). Section 6.6.2 introduces a measure that links up with the way disclosure risk is assessed at Statistics Netherlands.

### 6.6.1 The Measure Theta

The following describes how the general measure for disclosure risk introduced in Skinner and Elliot (2002) can be extended to the situation where PRAM is applied before data are released by the statistical agency. When a disclosure control method such as PRAM has been applied, a measure for disclosure risk is needed to quantify the protection that is offered by the control method. Scenarios that may lead to a disclosure of the identity of a respondent are about persons that aim at disclosure and that may have data that overlap the released data. A common scenario is that a person has a sample from another source and tries to identify respondents in the released sample by matching records. Using an extension of the measure in Skinner and Elliot (2002) we can investigate how applying PRAM reduces the disclosure risk.

Under simple random sampling, Skinner and Elliot (2002) introduced the measure of disclosure risk $\theta = I\!P(\text{correct match}|\text{unique match})$ as

$$\theta = \frac{\sum_{j=1}^{J} I(f_j = 1)}{\sum_{j=1}^{J} F_j \; I(f_j = 1)}.$$

The measure $\theta$ is the proportion of correct matches among those population units which match a sample unique. The measure is sample dependent and a distribution-free prediction is given by

$$\widehat{\theta} = \frac{\pi n_1}{\pi n_1 + 2(1-\pi) n_2},$$

where $\pi$ is the sampling fraction, $n_1 = \sum_{j=1}^{J} I(f_j = 1)$ is the number of uniques and $n_2 = \sum_{j=1}^{J} I(f_j = 2)$ is the number of twins in the sample (Skinner and Elliot 2002). Elamir and Skinner (2003) extended $\theta$ for the situation where misclassification occurs. The extension is given by

$$\theta_{mm} = \frac{\sum_{i \in s} I(f_{X_i} = 1, X_i^* = X_i)}{\sum_{j=1}^{J} F_j \; I(f_j = 1)}$$

and its distribution-free prediction is given by

$$\widehat{\theta}_{mm} = \frac{\pi \sum_{j=1}^{J} \mathrm{I}(f_j = 1) p_{jj}}{\pi n_1 + 2\,(1 - \pi)\, n_2},$$

where $p_{jj}$ is the diagonal entry $(j, j)$ of the transition matrix $\boldsymbol{P}_X$ which describes the misclassification.

Section 6.8 presents some simulation results with respect to the measure $\theta$ before applying PRAM, and $\theta_{mm}$ after applying PRAM.

### 6.6.2 Spontaneous Recognition

Statistics Netherlands releases data in several ways one of which is releasing detailed survey data under contract. Data are released under contract to bona fide research institutes that sign an agreement in which they promise not to look for disclosure explicitly, e.g, by matching the data to other data files. In this situation, SDC concerns the protection against what is called *spontaneous recognition*. This section introduces a measure for disclosure risk for PRAM data that is specific to the control for spontaneous recognition.

Controlling for spontaneous recognition means that one should prevent that certain records attract attention. A record may attract attention when a low dimensional combination of its values has a low frequency. Also, without cross-classifying, a record may attract attention when one of its values is recognized as being very rare in the population. Combinations of values with low frequencies in the sample are called *unsafe combinations*.

Statistics Netherlands uses the rule of thumb that a recognition of a combination of values of more than three variables is not spontaneous anymore. For this reason, only combinations of three variables are assessed with respect to disclosure control for spontaneous recognition.

Note that applying PRAM causes two kinds of modifications in the sample that make disclosure more difficult. First, it is possible that unsafe combinations in the sample change into apparently safe ones in the released sample, and, secondly, it is possible that safe combinations in the sample change into apparently unsafe ones. Since misclassification probabilities are not that large (in order to keep analysis of the released sample possible) and the frequency of unsafe combinations is typically low, the effect of the first modification is negligible in expectation. The second modification is more likely to protect an unsafe combination $j$ when there are a lot of combinations $k$, $k \neq j$, which are misclassified into $j$. This is the reason to focus, for a given record $i$ with the unsafe combination of scores $j$, on the calibration probability

$$\mu = I\!P(X_i = j | X_i^* = j, i \in s).$$

When there are hardly any $k$, $k \neq j$, misclassified into $j$, this probability will be large, and, as a consequence, the released record is unsafe. Note that combinations with frequency equal to zero are never unsafe.

Measure $\mu$ is a simplification since it ignores possible correlation between $X$ and other variables in the sample. Note that $X$ will be a Cartesian product and that the statistical agency can compute the calibration probability $\mu$ using (6.6) since the agency has the frequencies of $X$.

## 6.7 Information Loss

Since we stressed in the introduction that SDC means searching for a balance between disclosure risk and information loss, this section indicates ways to investigate information loss due to PRAM.

First, the transition matrix $\boldsymbol{P}_X$ gives an idea of the loss of information. The more this matrix resembles the identity matrix, the less information gets lost. In general, this requires a definition of a distance between two matrices. However, we can apply PRAM using matrices that are parameterized by one parameter, denoted $p_d$. The idea is as follows. Each time PRAM is applied, the diagonal probabilities in a transition matrix are fixed and equal to $p_d$. In the columns, the probability mass $1 - p_d$ is equally divided over the entries that are not diagonal entries. In this situation $1 - p_d$ is a measure for the deviation from the identity matrix.

Although transition matrices give an idea of the information loss, it is hard to have an intuition about how a certain deviation from the identity matrix affects analysis of the released data. A second way to investigate information loss is the comparison of extra variances due to PRAM with respect to frequency estimation. The idea here is that when this extra variance is already substantial, more complex analyses of the released sample is probably not possible. The variance with respect to frequency estimation can be estimated using (6.5).

The next section will assess information loss due to PRAM using these two approaches.

## 6.8 Simulation Examples

The objective of this section is to illustrate the theory in the foregoing sections and to investigate disclosure risk and information loss for different choices of misclassification parameters. The population is chosen to consist of units with complete records in the British Household Survey 1996-1999. We have $n = 16710$ and we distinguish 5 identifying variables with respect to the household owner: Sex ($S$), Marital Status ($M$), Economic Status ($D$), Socio-Economic Group ($E$), and Age ($A$), with number of categories 2, 7, 10, 7, and 8, respectively. In the following, we consider simple random sampling without replacement from the population where the sample fraction $\pi$ is equal to 0.05, 0.10 or 0.15. The three samples are denoted $s_1$, $s_2$ and $s_3$ and have sample sizes 836, 1671 and 2506 respectively.

The transition matrices used to apply PRAM to the selected variables are mostly of a simple form and determined by one parameter $p_d$, as described in Section 6.7. A more sophisticated construction of the transition matrices can reduce the disclosure risk further. An example of this fine-tuning will be given.

### 6.8.1 Disclosure Risk and the Measure Theta

The following discusses disclosure risk by comparing the measure $\theta$ before PRAM is applied with the measure $\theta_{mm}$ after PRAM has been applied, see Section 6.6.1. The identifying variables are described by $X = (S, M, D, E, A)$ with $J = 7840$ possible categories.

Since the population is known, we can compute the measures and using the samples we can compute their predictions. Table 6.2 presents the simulation results using simple random sampling without replacement, different sampling fractions $\pi$, and different choices of $p_d$. Given a choice of $\pi$ and $p_d$, drawing the sample and applying PRAM is 100 times simulated. The means of the computed and predicted measures are reported in Table 6.2. Note that $\theta$ and $\widehat{\theta}$ reflect the risk before applying PRAM and $\theta_{mm}$ and $\widehat{\theta}_{mm}$ reflect the risk after applying PRAM.

It is clear from Table 6.2 that applying PRAM will reduce the risk. For example, when $p_d = 0.80$ and $\pi = 0.10$, applying PRAM reduces the risk from $\theta = 0.166$ to $\theta_{mm} = 0.055$. When $p_d$ decreases, disclosure risk decreases too, as one might expect. Note that disclosure risk increases when sample size increases. In a larger sample, a record with a unique combination of scores is more likely to be a population unique and therefore the danger of a correct match is higher.

Table 6.2: Simulation results of disclosure risk measures for $X = (S, M, D, E, A)$ before and after applying PRAM with $p_d$.

| $p_d$ | Sample fraction | $\theta$ | $\widehat{\theta}$ | $\theta_{mm}$ | $\widehat{\theta}_{mm}$ |
|---|---|---|---|---|---|
| | 0.05 | 0.084 | 0.087 | 0.061 | 0.065 |
| 0.95 | 0.10 | 0.151 | 0.147 | 0.112 | 0.110 |
| | 0.15 | 0.217 | 0.215 | 0.165 | 0.166 |
| | 0.05 | 0.087 | 0.086 | 0.047 | 0.051 |
| 0.90 | 0.10 | 0.148 | 0.157 | 0.083 | 0.087 |
| | 0.15 | 0.213 | 0.216 | 0.123 | 0.127 |
| | 0.05 | 0.087 | 0.090 | 0.028 | 0.031 |
| 0.80 | 0.10 | 0.166 | 0.151 | 0.055 | 0.054 |
| | 0.15 | 0.213 | 0.221 | 0.065 | 0.064 |

### 6.8.2 Disclosure risk and Spontaneous Recognition

This following illustrates the measure $\mu$ for disclosure risk for spontaneous recognition that is discussed in Section 6.6.2.

Spontaneous recognition is defined for combinations up to three identifying variables, see Section 6.6.1. So there are 10 groups to consider. We will discuss only one of them, namely the group defined by $X = (M, D, E)$. The number of categories of $X$ is 490. The measure for disclosure risk is given by

$$\mu = I\!P\Big((M, D, E) = (m, d, e)\Big|(M^*, D^*, E^*) = (m, d, e)\Big),$$

for those combinations of values $(m, d, e)$ that have frequency 1 in sample $s_1$, $s_2$ or $s_3$. Note that when PRAM is not applied, $\mu = 1$. Table 6.3 shows results with respect to the maximum of $\mu$ when PRAM is applied to $M$, $D$ and $E$. With respect to $X = (M, D, E)$ the number of unique combinations in $s_1$, $s_2$ or $s_3$ are 48, 44, and 53, respectively.

We draw two conclusions from the results. First, the results illustrate that the probability $p_d$ matters, as one might expect. Second, the results show that the size of the sample is important. In order to protect an unsafe combination $j$, it is necessary that there are a lot of combinations that can change into $j$ due to PRAM. Note that this is the other way around compared to the measure $\theta$ where a larger sample size causes a higher disclosure risk. This difference shows that different concepts of disclosure induce different methods for disclosure control.

Table 6.3: Maximum of $\mu$ for values of $(M, D, E)$ with frequency 1 when applying PRAM to $M$, $D$ and $E$.

| | $p_d$ | | | | | |
|---|---|---|---|---|---|---|
| Sample | 0.95 | 0.90 | 0.85 | 0.80 | 0.70 | 0.60 |
| $s_1$ with $n = 836$ | 0.94 | 0.86 | 0.76 | 0.65 | 0.43 | 0.24 |
| $s_2$ with $n = 1671$ | 0.94 | 0.85 | 0.74 | 0.61 | 0.36 | 0.18 |
| $s_3$ with $n = 2506$ | 0.93 | 0.83 | 0.69 | 0.55 | 0.31 | 0.15 |

The following introduces a method to fine-tune a transition matrix and shows that this can help to diminish the disclosure risk. The idea is to adjust one or more columns in the transition matrix of each variable that is part of an unsafe combination. Consider $\boldsymbol{P}_{X_1}$ where variable $X_1$ has $J_1$ categories. The column that is chosen first corresponds to the category of $X_1$ with the highest frequency in sample $s$, say column $j$. Let furthermore $k$ be the number that corresponds to the category of $X_1$ with the lowest frequency in $s$. The columns of $\boldsymbol{P}_{X_1}$ that are not column $j$ are constructed as explained in Section 6.7: $p_d$ on the diagonal and $(1 - p_d)$ equally divided over the other entries. Column $j$ is fine-tuned by

$$p_{lj} = \begin{cases} p_d & \text{if} \quad l = j \\ (1-p_d)/\eta & \text{if} \quad l = k \\ (1-p_d)/(\eta(J_1 - 2)) & \text{if} \quad l \neq j, k \end{cases} , \qquad (6.11)$$

for $l \in \{1, .., J_1\}$ and $\eta > 1$. The idea here is that when we choose $\eta$ close to 1, the category with the highest frequency has a relatively high probability to change into the category with the lowest frequency. Assuming a link between an unsafe combination and a low frequency in the original sample, this idea explicitly supports the concept of PRAM: An unsafe combination $c$ is protected by creating new combinations $c$ from combinations that have high frequencies in the original sample.

In the same way, other columns in $\boldsymbol{P}_{X_1}$ can be fine-tuned. For example, the second column chosen is the column that corresponds to the category of $X_1$ with the second highest frequency in sample $s$, and the chosen row is now the row that corresponds to the category of $X_1$ with the second lowest frequency in sample $s$.

Table 6.4 presents results for sample $s_3$ when the transition matrices of $M$, $D$, and $E$ are fine-tuned. The advantage of fine-tuning the transition matrices is dependent of the data and on the size of the sample. One can see that the idea works, e.g., if

Table 6.4: Maximum of $\mu$ for values of $(M, D, E)$ with frequency 1 in sample $s_3$ when applying PRAM to $M$, $D$ and $E$, and using fine-tuning for all three variables.

| | $p_d$ | | | | |
|---|---|---|---|---|---|
| Construction of transition matrix | 0.95 | 0.90 | 0.85 | 0.80 | 0.70 |
| No fine-tuning | 0.93 | 0.83 | 0.69 | 0.55 | 0.31 |
| Fine-tuning 1 column where $\eta = 1.001$ | 0.92 | 0.80 | 0.66 | 0.51 | 0.28 |
| Fine-tuning 2 columns where $\eta = 1.001$ | 0.91 | 0.74 | 0.57 | 0.42 | 0.20 |
| Fine-tuning 3 columns where $\eta = 1.001$ | 0.90 | 0.72 | 0.52 | 0.34 | 0.15 |

Table 6.5: Maximum of $\mu$ for values of $(M, D, E)$ with frequency 1 in the population when applying PRAM to $M$, $D$ and $E$, and using fine-tuning for all three variables.

| | $p_d$ | | |
|---|---|---|---|
| Construction of transition matrix | 0.95 | 0.90 | 0.85 |
| No fine-tuning | 0.75 | 0.46 | 0.28 |
| Fine-tuning three columns where $\eta = 1.001$ | 0.55 | 0.29 | 0.16 |

$p_d = 0.80$, fine-tuning can decrease the maximum of $\mu$ from 0.55 to 0.34. To show that the size of the sample makes an important difference, we present an additional example where we assess the population with $n = 16710$. In the population there are 53 unique combinations of values of $X = (M, D, E)$. Table 6.5 shows that also in this case, fine-tuning works well and also that the maximum of $\mu$ is notable lower than in the samples.

An additional way to reduce disclosure risk is to consider the histogram of $\mu$ and to combine PRAM with local suppression. By deleting or suppressing some records that have a high $\mu$, disclosure risk goes down. Figure 6.1 is the histogram of $\mu$ for the 53 values of $(M, D, E)$ with frequency 1 in sample $s_3$, where $p_d = 0.80$ and three columns are fine-tuned with $\eta = 1.001$. When we delete the two records with highest $\mu$, the maximum of $\mu$ becomes 0.30 instead of 0.34.

Conclusion and advice: Determine a largest tolerated $\mu$ and check all combinations of three identifying variables and use fine-tuning. The protection offered by PRAM dependents on $p_d$, but also very much on the sample size.

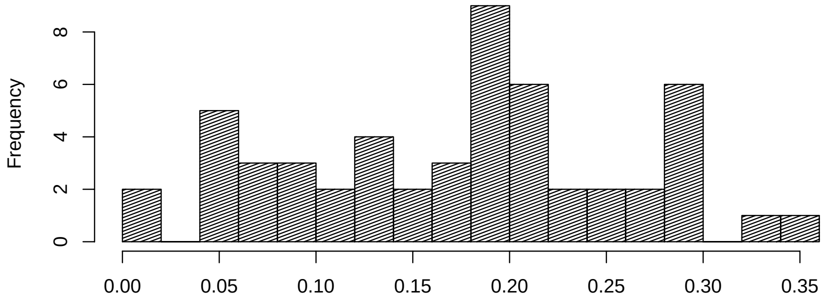


Figure 6.1: Histogram of $\mu$ for the 53 values of $(M, D, E)$ with frequency 1 in sample $s_3$ when applying PRAM to $M$, $D$ and $E$ with $p_d$=0.80.

### 6.8.3 Information Loss in Frequency Estimation

To investigate information loss due to PRAM, this section discusses an example with univariate frequency estimation with respect to the variable $M$ and bivariate frequency estimation with respect to the variables $S$ and $E$ in sample $s_3$. We illustrate the difference between using $\boldsymbol{P}_X$ and $\overleftarrow{\boldsymbol{P}}_X$ by comparing standard errors in estimating the univariate frequencies of variable $M$. In the following, we assume released sample frequencies of $M$ to be equal to the expected released sample frequencies. That is, released sample frequencies $\boldsymbol{f}^*$ are given by $\mathbb{E}(\boldsymbol{F}^*|\boldsymbol{f}) = \boldsymbol{P}_M\boldsymbol{f}$, where $\boldsymbol{F}^*$ and $\boldsymbol{f}$ are defined with respect to $M$. It follows that in this situation $\hat{\boldsymbol{f}} = \boldsymbol{f}$, so that (6.5) can be used to compare variances. To estimate the standard errors when using calibration, we use $\overleftarrow{\boldsymbol{P}}_M$ in (6.5) instead of $\boldsymbol{P}_M^{-1}$. Table 6.6 presents standard errors of estimated frequencies for different choices of $p_d$. The example shows that $\overleftarrow{\boldsymbol{P}}_M$ is more efficient than $\boldsymbol{P}_M$, a difference that becomes more striking when $p_d$ is smaller.

In the bivariate situation, calibration probabilities do not always work well. To illustrate this, the following simulation example is about frequency estimation of variable $X = (S, E)$ that has 14 categories. The chi-square test of independence between $S$ and $E$ yields 529.55, where df $= 6$ and the $p$ value $< 0.00$. It is this lack of independence between the variables that causes calibration probabilities to perform badly. PRAM was applied 10 times to both $S$ and $E$ with $p_d = 0.85$. Figure 6.2 shows the estimation of the frequencies of $X$ using $X^*$ and $\boldsymbol{P}_S \otimes \boldsymbol{P}_E$ versus using

Table 6.6: Frequencies before PRAM and standard errors of estimated frequencies after PRAM of the variable $M$ in sample $s_3$.

| | Standard error of $\hat{\boldsymbol{f}}$ given $p_d$ and the transition matrix | | | | | |
|---|---|---|---|---|---|---|
| | $p_d = 0.95$ | | $p_d = 0.90$ | | $p_d = 0.85$ | |
| $\boldsymbol{f}$ | $\boldsymbol{P}_M$ | $\overleftarrow{\boldsymbol{P}}_M$ | $\boldsymbol{P}_M$ | $\overleftarrow{\boldsymbol{P}}_M$ | $\boldsymbol{P}_M$ | $\overleftarrow{\boldsymbol{P}}_M$ |
| 99 | 5.27 | 4.08 | 7.87 | 4.77 | 10.23 | 4.87 |
| 378 | 6.33 | 5.62 | 9.40 | 7.35 | 12.13 | 8.30 |
| 353 | 6.24 | 5.52 | 9.27 | 7.21 | 11.97 | 8.11 |
| 471 | 6.65 | 5.95 | 9.85 | 7.83 | 12.69 | 8.90 |
| 525 | 6.83 | 6.12 | 10.11 | 8.08 | 13.01 | 9.21 |
| 551 | 6.78 | 6.08 | 10.04 | 8.02 | 12.93 | 9.13 |
| 169 | 5.55 | 4.64 | 8.28 | 5.77 | 10.74 | 6.20 |

$X^*$ and $\overleftarrow{\boldsymbol{P}}_S \otimes \overleftarrow{\boldsymbol{P}}_E$. From the figure it is clear that the misclassification probabilities perform better, i.e., the points $(f_j, \hat{f}_j)$ are closer to the identity line than $(f_j, \tilde{f}_j)$, $j \in \{1, ..., 14\}$. The variance is less when $\overleftarrow{\boldsymbol{P}}_S \otimes \overleftarrow{\boldsymbol{P}}_E$ is used, but the figure shows that in that case estimates are biased. This can be made more precise by estimating the mean squared error (MSE). We define

$$\widehat{MSE}_j = \frac{1}{B}\sum_{b=1}^{B}(\hat{f}_{jb} - f_j)^2 \quad \text{and} \quad \widetilde{MSE}_j = \frac{1}{B}\sum_{b=1}^{B}(\tilde{f}_{jb} - f_j)^2 \quad \text{for } j \in \{1, ..., 14\},$$

where $B$ is the number of simulations, and $\hat{f}_{jb}$ and $\tilde{f}_{jb}$ denote estimates of $f_j$ in the $b$th simulation. In the example, $B = 10$, and $[\min_j\{\widehat{MSE}_j\}, \max_j\{\widehat{MSE}_j\}] = [71.9, 354.9]$ while $[\min_j\{\widetilde{MSE}_j\}, \max_j\{\widetilde{MSE}_j\}] = [256.2, 13079.8]$. Violating the independence assumption regarding the use of $\overleftarrow{\boldsymbol{P}}_S \otimes \overleftarrow{\boldsymbol{P}}_E$ has sever consequences.

Misclassification proportions are close to misclassification probabilities in the above example. Compare for instance

$$\boldsymbol{P}_S = \begin{pmatrix} 0.85 & 0.15 \\ 0.15 & 0.15 \end{pmatrix} \quad \text{and} \quad \boldsymbol{P}_S^\circ = \begin{pmatrix} 0.854 & 0.154 \\ 0.156 & 0.156 \end{pmatrix}.$$

A simulation study can be used to investigate the performance of misclassification probabilities versus misclassification proportions. The study compares using $\boldsymbol{P}_S \otimes \boldsymbol{P}_E$ versus using $\boldsymbol{P}_S^\circ \otimes \boldsymbol{P}_E^\circ$ by looking at the actual coverage percentage (ACP),

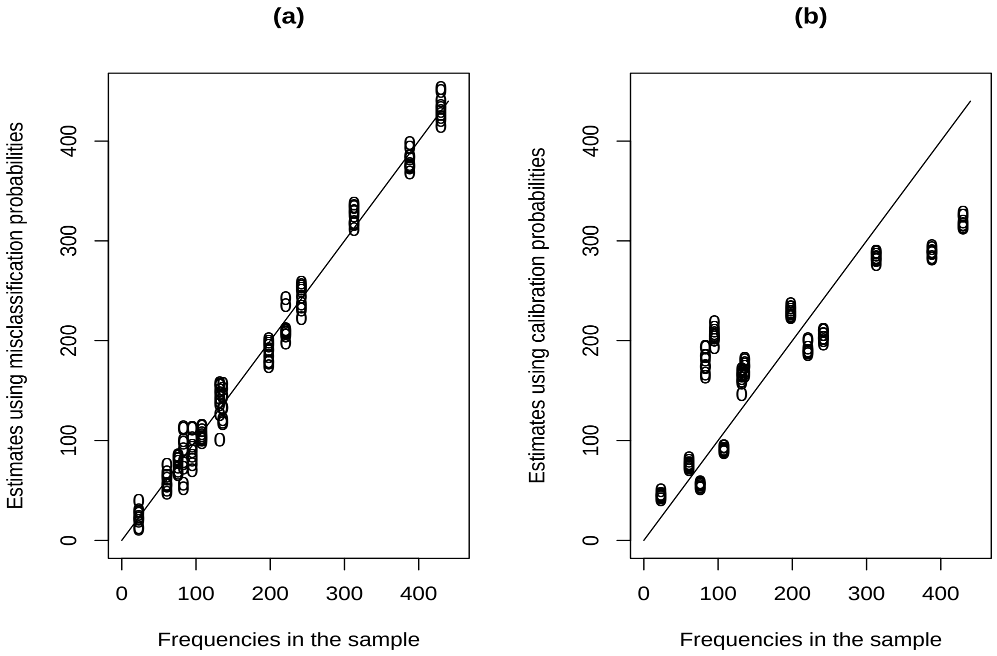


Figure 6.2: Estimating frequencies of $X = (S, E)$ after applying PRAM to $S$ and $E$ in sample $s_3$ with $p_d = 0.85$ in 10 simulations. *(a)* Using misclassification probabilities. *(b)* Using calibration probabilities.

which is the percentage of the replicated perturbed samples for which the confidence interval of the estimated frequency covers the actual frequency in the original sample. We used sample $s_3$, $p_d = 0.85$ and 1000 simulated perturbed samples. Table 6.7 shows that misclassification proportions perform better than then misclassification probabilities. The mean value of ACP when using $\boldsymbol{P}_S \otimes \boldsymbol{P}_E$ equals 95.14, and the mean value of ACP when using $\boldsymbol{P}_S^\circ \otimes \boldsymbol{P}_E^\circ$ equals 98.04. (A paired t-test yields a $p$ value $< 0.00$.) An increase in ACP is only advantageous when it is not caused by an increase in variance. We define

$$\widehat{MSE}_j^\circ = \frac{1}{B}\sum_{b=1}^{B}(\hat{f}_{jb}^\circ - f_j)^2 \quad \text{for } j \in \{1, ..., 14\},$$

where $\hat{f}_{jb}^\circ$ denotes the estimate of $f_j$ using $\boldsymbol{P}_S^\circ \otimes \boldsymbol{P}_E^\circ$ in the $b$th simulation. In the example, $B = 1000$, and $\widehat{MSE}_j^\circ < \widehat{MSE}_j$, for all $j \in \{1, ..., 14\}$. Using $\boldsymbol{P}_S^\circ \otimes \boldsymbol{P}_E^\circ$

Table 6.7: Actual coverage percentage w.r.t. $X = (S, E)$ for sample $s_3$ and 1000 simulated perturbed samples, where $p_d = 0.85$.

| | ACP given the transition matrix | | | ACP given the transition matrix | |
|---|---|---|---|---|---|
| Category | $\boldsymbol{P}_S \otimes \boldsymbol{P}_E$ | $\boldsymbol{P}_S^\circ \otimes \boldsymbol{P}_E^\circ$ | Category | $\boldsymbol{P}_S \otimes \boldsymbol{P}_E$ | $\boldsymbol{P}_S^\circ \otimes \boldsymbol{P}_E^\circ$ |
| (1,1) | 94.8 | 98.2 | (2,1) | 94.5 | 97.8 |
| (1,2) | 95.3 | 98.1 | (2,2) | 95.6 | 97.7 |
| (1,3) | 95.7 | 97.7 | (2,3) | 95.0 | 97.9 |
| (1,4) | 95.4 | 98.0 | (2,4) | 95.7 | 98.8 |
| (1,5) | 95.5 | 99.1 | (2,5) | 93.5 | 97.3 |
| (1,6) | 95.1 | 97.4 | (2,6) | 93.8 | 97.5 |
| (1,7) | 96.0 | 98.3 | (2,7) | 96.0 | 98.7 |

gives the best result. Although the transition matrices are quite alike at first sight, misclassification proportions perform best.

## 6.9 Conclusion

The paper shows that the analysis of PRAM data is more efficient when misclassification proportions are used instead of misclassification probabilities. Calibration probabilities and calibration proportions work fine in the univariate case, but cause serious bias in the multivariate case. Since in most situations the user of PRAM data will be interested in multivariate analysis, it seems wise not to release calibration probabilities or calibration proportions along with the PRAM data. The two measures for disclosure risk that are used in this paper show that PRAM helps in protecting the identity of respondents.

Given that releasing misclassification proportions makes PRAM more efficient with respect to information loss, it is still an open question how this works out when PRAM is compared to other SDC methods, see Domingo-Ferrer and Torra (2001). It might be worthwhile to state that PRAM was never meant to replace existing SDC methods. Working with PRAM data and taking into account the information about the misclassification in the analysis might be quite a burden for some researchers. However, when researchers are interested in specific details in data, details that might disappear, e.g., when global recoding is used, PRAM can be a solution. Note that PRAM is statistically sound. Data are perturbed, but information about the perturbation can be used. Although estimates will have extra variance due to the perturbation, they will be unbiased.

Since the misclassification proportions provide more information about the original sample than the misclassification probabilities, one should consider the question whether providing these proportions increases the disclosure risk. Since the privacy protection that is offered by PRAM is at the record level, we do not think that disclosure risk increases when misclassification proportions are released. With these proportions, sample frequencies of the identifying variables can be deduced, but these frequencies are not sensitive information. Note also that when one works with the measures for disclosure risk discussed in Section 6.6, the risk does not change when misclassification proportions are released.

# Appendix 6.A

The following shows that $\boldsymbol{f} = \overleftarrow{\boldsymbol{P}}_X \mathbb{E}[\boldsymbol{F}^*|\boldsymbol{f}]$. First note that $\mathbb{E}[\boldsymbol{F}^*|\boldsymbol{f}] = \boldsymbol{P}_X \boldsymbol{f}$ and that entries $\overleftarrow{p}_{jk}$ of $\overleftarrow{\boldsymbol{P}}_X$ are defined as $\overleftarrow{p}_{jk} = (p_{kj} f_j)(\sum_{j_0=1}^{J} p_{kj_0} f_{j_0})^{-1}$ for $k, j \in \{1, ..., J\}$. For each $j \in \{1, ..., J\}$ we have

$$\begin{aligned}\Big(\overleftarrow{\boldsymbol{P}}_X \mathbb{E}[\boldsymbol{F}^*|\boldsymbol{f}]\Big)(j) &= \sum_{k=1}^{J} \overleftarrow{p}_{jk} \Big(\mathbb{E}[\boldsymbol{F}^*|\boldsymbol{f}]\Big)(k) \\ &= \sum_{k=1}^{J} \overleftarrow{p}_{jk} \Big(\sum_{j_0=1}^{J} p_{kj_0} f_{j_0}\Big) = \sum_{k=1}^{J} p_{kj} f_j = f_j,\end{aligned}$$

since the columns of $\boldsymbol{P}_X$ sum up to one. So $\boldsymbol{f} = \overleftarrow{\boldsymbol{P}}_X \boldsymbol{P}_X \boldsymbol{f}$ and $\boldsymbol{f}$ is an eigenvector of $\overleftarrow{\boldsymbol{P}}_X \boldsymbol{P}_X$ with eigenvalue 1.

In general, $\overleftarrow{\boldsymbol{P}}_X \neq \boldsymbol{P}_X^{-1}$. To illustrate this, let $\boldsymbol{R} = \overleftarrow{\boldsymbol{P}}_X \boldsymbol{P}_X$. The entries of $\boldsymbol{R}$ are $r_{ij} = \sum_{k=1}^{J} \overleftarrow{p}_{ik} p_{kj}$, for $i, j \in \{1, ..., J\}$. Assume that the entries of $\boldsymbol{P}_X$ are all $> 0$ and that $\boldsymbol{f}_j > 0$, for $j \in \{1, ..., J\}$. Then $\overleftarrow{p}_{jk} > 0$, for $j, k \in \{1, ..., J\}$ and $r_{ij} > 0$, for $i, j \in \{1, ..., J\}$. In this case, $\boldsymbol{R}$ is not the identity matrix and consequently $\overleftarrow{\boldsymbol{P}}_X \neq \boldsymbol{P}_X^{-1}$. A more intuitive explanation is that $\overleftarrow{\boldsymbol{P}}_X$ changes when the survey data change, whereas $\boldsymbol{P}_X$ can be determined independently from the data and does not necessarily change when the data change. Therefore, it is always possible to cause $\overleftarrow{\boldsymbol{P}}_X \neq \boldsymbol{P}_X^{-1}$ by changing the data.

# Appendix 6.B

The following derives the maximum likelihood properties of (6.3) and (6.8). The reasoning is the same as in Hochberg (1977), but simpler, since in the PRAM situation calibration probabilities do not have to be estimated. Also, we show that the reasoning applies both to (6.3) and to (6.8).

Assume that the frequency vector of independent $X_1, ...X_n$ is multinomially distributed with parameters $n$ and $\boldsymbol{\pi} = (\pi_1, .., \pi_J)^t$, where $\pi_j > 0$ for $j \in \{1, .., J\}$, and $\sum_{j=1}^{J} \pi_j = 1$. Consider the transformation $\boldsymbol{\pi}^* = \boldsymbol{P}\boldsymbol{\pi}$, where $\boldsymbol{P}$ is a $J \times J$ transition matrix, i.e., columns sum up to one and $p_{kj} \geq 0$ for $k, j \in \{1, .., J\}$. Assume that $\boldsymbol{P}$ is nonsingular. Let the distribution of $X^*$ be given by $\mathbb{P}(X^* = k) = \pi_k^*$, for $k \in \{1, .., J\}$. It follows that the frequency vector of $X_1^*, X_2^*, ..., X_n^*$ is multinomially distributed with parameters $n$ and $\boldsymbol{\pi}^*$. Indeed, $\pi_k^* = p_{k1}\pi_1 + ... + p_{kJ}\pi_J > 0$ for $k \in \{1, .., J\}$ and

$$\sum_{k=1}^{J} \pi_k^* = \Big(\sum_{l=1}^{J} p_{l1}\Big)\pi_1 + ... + \Big(\sum_{l=1}^{J} p_{lJ}\Big)\pi_J = 1.$$

The likelihood $L^*$ for $\boldsymbol{\pi}^*$ and observed $\boldsymbol{x}^* = (x_1^*, x_2^*, ..., x_n^*)^t$ is well known. Let $\boldsymbol{f}^* = (f_1^*, f_2^*, ..., f_J^*)^t$ denote the observed cell frequencies. The MLE is given by $\hat{\boldsymbol{\pi}}^* = \boldsymbol{f}^*/n$ and has covariance matrix $\boldsymbol{\Omega} = [\text{Diag}(\boldsymbol{\pi}^*) - \boldsymbol{\pi}^*(\boldsymbol{\pi}^*)^t]/n$, where $\text{Diag}(\boldsymbol{\pi}^*)$ is the diagonal matrix with the diagonal entries given by the elements of $\boldsymbol{\pi}^*$.

Next we can use the invariance property of maximum likelihood. Define the transformation $g(\boldsymbol{\pi}^*) = \boldsymbol{P}^{-1}\boldsymbol{\pi}^*$. Since $g$ is one-to-one, it follows from $L^*(\boldsymbol{\pi}^*|\boldsymbol{x}^*)$ and $\boldsymbol{\pi} = g(\boldsymbol{\pi}^*)$ that the likelihood for $\boldsymbol{\pi}$ is given by $L^*(g^{-1}(\boldsymbol{\pi})|\boldsymbol{x}^*)$ which is maximized for $\hat{\boldsymbol{\pi}} = g(\hat{\boldsymbol{\pi}}^*) = \boldsymbol{P}^{-1}\hat{\boldsymbol{\pi}}^*$. Consequently, when $\hat{\boldsymbol{\pi}} \in (0, 1)^J$, it is the MLE. Since $g$ has a first order derivative, the covariance matrix of $\hat{\boldsymbol{\pi}}$ can be obtained using the delta-method, see, e.g., Agresti (2002, Chapter 14). We have $\partial g(\boldsymbol{\pi}^*)/\partial\boldsymbol{\pi}^* = (\boldsymbol{P}^{-1})^t$ and the covariance matrix of $\hat{\boldsymbol{\pi}}$ is given by $n^{-1}\boldsymbol{P}^{-1}\boldsymbol{\Omega}(\boldsymbol{P}^{-1})^t$.

So with respect to (6.3), maximum likelihood properties are proven by taking $\boldsymbol{P} = \boldsymbol{P}_X$ and obtaining $\hat{\boldsymbol{\pi}} = g(\hat{\boldsymbol{\pi}}^*) = \boldsymbol{P}_X^{-1}\hat{\boldsymbol{\pi}}^*$. With respect to (6.8), the misclassification design is described by $\boldsymbol{\pi} = \overleftarrow{\boldsymbol{P}}_X\boldsymbol{\pi}^*$, so $\boldsymbol{P} = \overleftarrow{\boldsymbol{P}}_X^{-1}$ and the MLE is given by $\tilde{\boldsymbol{\pi}} = g(\hat{\boldsymbol{\pi}}^*) = \overleftarrow{\boldsymbol{P}}_X\hat{\boldsymbol{\pi}}^*$.

# Appendix 6.C

Let $P_{kj}^{\circ}$ denote the stochastic variable of the $kj$-th entry of $\boldsymbol{P}_X^{\circ}$ and $C_{kj}$ the stochastic variable of the $kj$-th cell in the cross-classification $X^*$ by $X$. It follows that $C_{kj}$ has a binomial distribution with parameters $f_j$ and $p_{kj}$. Consequently, $\mathbb{E}[P^{\circ}_{kj}|\boldsymbol{f}] = \mathbb{E}[C_{kj}/f_j|\boldsymbol{f}] = f_j p_{kj}/f_j = p_{kj}$ and in expectation $\boldsymbol{P}^{\circ}_X$ equals $\boldsymbol{P}_X$. Since $C_{k_1 j_1}$ and $C_{k_2 j_2}$ are independent given $\boldsymbol{f}$, it follows that $\mathbb{E}[P^{\circ}_{k_1 j_1} P^{\circ}_{k_2 j_2}|\boldsymbol{f}] = p_{k_1 j_1} p_{k_2 j_2}$. So in expectation, $\boldsymbol{P}^{\circ}_{X_1} \otimes \boldsymbol{P}^{\circ}_X$ equals $\boldsymbol{P}_{X_1} \otimes \boldsymbol{P}_{X_2}$.

We define $\overleftarrow{P}^{\circ}_{jk} = C_{kj}/(F^*_k + \varepsilon)$ where $\varepsilon$ is a small positive value. Using the delta method, see, e.g., Rice (1995, Section 4.6), we obtain

$$\mathbb{E}[\overleftarrow{P}^{\circ}_{jk}|\boldsymbol{f}] \approx \frac{\mathbb{E}[C_{kj}|\boldsymbol{f}]}{\mathbb{E}[F^*_k|\boldsymbol{f}]} + \frac{1}{\mathbb{E}[F^*_k|\boldsymbol{f}]^2}\left(V[F^*_k|\boldsymbol{f}]\frac{\mathbb{E}[C_{kj}|\boldsymbol{f}]}{\mathbb{E}[F^*_k|\boldsymbol{f}]} - \rho\sqrt{V[C_{kj}|\boldsymbol{f}]V[F^*_k|\boldsymbol{f}]}\right),$$

where $\mathbb{E}[C_{kj}|\boldsymbol{f}] = f_j p_{kj}$ and $\rho$ is the correlation between $C_{kj}$ and $F^*_k$. From this we see that the difference between $\mathbb{E}[\overleftarrow{P}^{\circ}_{jk}|\boldsymbol{f}]$ and $\overleftarrow{p}_{jk}$ will be small when $V[F^*_k|\boldsymbol{f}]$ is small and $\mathbb{E}[F^*_k|\boldsymbol{f}]$ is large.

# References

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

Agresti, A. (2002). *Categorical Data Analysis, Second Edition*, New York: Wiley.

Assakul, K., and Proctor, C.H. (1967). Testing independence in two-way contingency tables with data subject to misclassification, *Psychometrika* **32**, 67-76.

Barndorff-Nielsen, O. (1982). Exponential families, *Encyclopedia of Statistical Sciences* (S. Kotz and N.L. Norman, eds), New York: Wiley.

Bethlehem, J.G., Keller, W.J., and Pannekoek, J. (1990). Disclosure control of microdata, *Journal of the American Statistical Association* **85**, 38-45.

Bishop, Y.M.M., Fienberg, S.E., and Holland, P.W. (1975). *Discrete Multivariate Analysis*, Cambridge: MIT Press.

Bergsma, W.P., and Rudas, T. (2002). Marginal models for categorical data, *The Annals of Statistics* **30**, 140-159.

Boeije, H., and Lensvelt-Mulders, G. (2002). Honest by chance: a qualitative interview study to clarify respondents' (non-)compliance with computer-assisted randomized response, *Bulletin de Methodologie Sociologique* **75**, 24-39.

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**, 964-968.

Boruch, R.F. (1971). Assuring confidentiality of responses in social research: a note on strategies, *The American Sociologist* **6**, 308-311.

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

Chen, T.T. (1989). A review of methods for misclassified categorical data in epidemiology, *Statistics in Medicine* **8**, 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**, 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 B* **39**, 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.

Domingo-Ferrer, J., and Torra, V. (2001). A quantitative comparison of disclosure control methods for microdata, In *Confidentiality, Disclosure, and Data Access, Theory and Practical Application for Statistical Agencies*, (P. Doyle, J. Lane, L. Zayatz, and J. Theeuwes, eds.) Amsterdam: North-Holland.

Efron, B, and Tibshirani, R.J (1993). *An Introduction to the Bootstrap*, New York: Chapman and Hall.

Elamir, E.A.H., and Skinner, C.J. (2003). Disclosure risk assessment for microdata, the treatment of measurement Error, manuscript.

Elffers, E., Van der Heijden, P.G.M., and Hezemans, M. (2003). Explaining regulatory non-compliance: a survey study of rule transgression for two Dutch instrumental laws, applying the randomized response method, *Journal of Quantitative Criminology* **19**, 409-439.

Evfimievski, A., Srikant, R., Agrawal, R., and Gehrke (2002) Privacy preserving mining of association rules, in *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery in Databases and Data Mining*, Edmonton, Alberta, Canada, 2002, 217-228.

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

Fox, J.A., and Tracy, P.E. (1986). *Randomized Response: A Method for Sensitive Surveys*, Newbury Park: Sage.

Gelman, A., Meng, X.L., and Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies, *Statistica Sinica* **6**, 733-807.

Goodman, L.A. (1973). The analysis of multidimensional contingency tables when some variables are posterior to others: a modified path analysis approach, *Biometrika* **60**, 179-192.

Goodman, L.A. (1974). Explanatory latent-structure analysis using both identifiable and unidentifiable models, *Biometrika* **61**, 251-231.

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**, 463-478.

Greenland, S. (1980). The effect of misclassification in the presence of covariates, *American Journal of Epidemiology* **112**, 564-569.

Greenland, S. (1988). Variance estimation for epidemiologic effect estimates under misclassification, *Statistics in Medicine* **7**, 745-757.

Haberman, S.J. (1977). Product models for frequency tables involving indirect observation, *The Annals of Statistics* **5**, 1124-1147.

Haberman, S.J. (1979). *Analysis of Qualitative Data: New Developments (Vol.2)*, New York: Academic Press.

Hagenaars, J.A. (1993). *Loglinear Models With Latent Variables*, Newbury Park: Sage.

Heinen, T. (1996). *Latent Class and Discrete Latent Trait Models: Similarities and Differences*, Thousand Oaks: Sage Publications.

Hochberg, Y. (1977). On the use of double sampling schemes in analyzing categorical data with misclassification errors, *Journal of the American Statistical Association* **72**, 914-921.

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.

Korn, E.L. (1981). Hierarchical log-linear models not preserved by classification error, *Journal of the American Statistical Association* **76**, 110-113.

Kuha, J., and Skinner, C. (1997). Categorical data analysis and misclassification, in *Survey Measurement and Process Quality*, (L. Lyberg, P. Biemer, M. Collins, E. De Leeuw, C. Dippo, N. Schwarz, and D. Trewin, eds.), New York: Wiley.

Kuk, A.Y.C. (1990). Asking sensitive questions indirectly. *Biometrika* **77**, 436-438.

Lindsay, B.G. (1995). *Mixture Models: Theory, Geometry and Applications,* NSF-CBMS Regional Conference Series in Probability and Statistics Vol. 5, Hayward, California: Institute of Mathematical Statistics.

Little, J.A., and Rubin D.B. (2002). *Statistical Analysis With Missing Data, Second Edition*, New York: Wiley.

Lucy, L.B. (1974). An iterative technique for the rectification of observed distributions, *The Astronomical Journal* **79**, 745-754.

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**, 195-203.

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

Mood, A.M., Graybill, F.A., and Boes, D.C. (1985). *Introduction to the theory of Statistics*, Auckland: McGraw-Hill.

Moors, J.J.A. (1981). Inadmissibility of linearly invariant estimators in truncated parameter spaces, *Journal of the American Statistical Association* **76**, 910-915.

Moriarty, M., and Wiseman, F. (1976). On the choice of a randomization technique with the randomized response model, *Proceedings of the social statistics section of the American Statistical Association*, 624-626.

Mote, V.L., and Anderson, R.L. (1965). An investigation of the effect of misclassification on the properties of $\chi^2$-tests in the analysis of categorical data, *Biometrika* **52**, 95-109.

Rice, J.A. (1995). *Mathematical Statistics and Data Analysis*, Belmont: Duxbury Press.

Rosenberg, M.J. (1979). Multivariate analysis by a randomized response technique for statistical disclosure control, Ph.D. Dissertation, University of Michigan.

Rosenberg, M.J. (1980). Categorical data analysis by a randomized response technique for statistical disclosure control, *Proceedings of the Survey Research Methods Section, American Statistical Association*, 311-318.

Rubin, D.B. (1976). Inference and missing data, *Biometrika* **63**, 581-592.

Schafer J.L.(1997). *Analysis of Incomplete Multivariate Data*, London: Chapman and Hall.

Schwartz, J.E. (1985). The neglected problem of measurement error in categorical data, *Sociological Methods and Research* **13**, 435-466.

Singh, J. (1976). A note on randomized response techniques, *Proceedings of the social statistics section of the American Statistical Association*, 772.

Skinner, C.J., and Elliot, M.J. (2002). A measure of disclosure risk for microdata, *Journal of the Royal Statistical Society B* **64**, 855-867.

Spiegelman, D., Rosner B., and Logan, R. (2000). Estimation and inference for logistic regression with covariate misclassification and measurement error in main study/validation study designs, *Journal of the American Statistical Association* **95**, 51-61.

Van den Hout, A. (1999) *The Analysis of Data Perturbed by PRAM*, WBBM Report Series 45, Delft: Delft University Press.

Van den Hout, A., and Van der Heijden, P.G.M (2002). Randomized response, statistical disclosure control and misclassification: a review, *International Statistical Review* **70**, 269-288.

Van den Hout, A., and Van der Heijden, P.G.M (2004). The analysis of multivariate misclassified data with special attention to randomized response data, *Sociological Methods and Research* **32**, 310-336.

Van der Heijden, P.G.M, and Van Gils, G. (1996). Some logistic regression models for randomized response data, Proceedings of the eleventh international workshop on statistical modelling, (A. Forcina, G. M. Marchetti, R. Hatzinger and G. Falmacci eds.).

Van der Heijden, P.G.M., Van Gils, G., Bouts, J., and Hox, J.J. (2000). A comparison of randomized response, computer-assisted self-interview, and face-to-face direct questioning, *Sociological Methods and Research* **28**, 505-537.

Van Gils, G., Van der Heijden, P.G.M., Laudy, O., Ross, R (2003). Regelovertreding in de sociale zekerheid, Den Haag: Ministerie van Social Zaken en Werkgelegenheid. (In Dutch)

Van Gils, G., Van der Heijden, P.G.M., and Rosebeek, A. (2001). Onderzoek naar regelovertreding. Resultaten ABW, WAO en WW. Amsterdam: NIPO. (In Dutch)

Vermunt, J.K. (1997). LEM: a general program for the analysis of categorical data, user's manual, Tilburg: Tilburg University.

Warner, S.L. (1965). Randomized response: a survey technique for eliminating answer bias, *Journal of the American Statistical Association* **60**, 63-69.

Warner, S.L. (1971). The linear randomized response model, *Journal of the American Statistical Association* **66**, 884-888.

Willenborg, L.C.R.J. (2000). Optimality models for PRAM, *Proceedings in Computational Statistics* (J.G. Bethlehem, J.G., and Van der Heijden, P.G.M., eds.), Heidelberg: Physica-Verlag.

Willenborg, L.C.R.J., and De Waal, T. (2001). *Elements of Statistical Disclosure Control*, New York: Springer.

# Samenvatting

Dit boek gaat over de analyse van *randomized response*-data en over de analyse van *post randomization*-data. Randomized response (RR) is een methode die wordt gebruikt in onderzoek naar sensitieve items (Warner 1965). Voorbeelden van sensitieve items zijn items over fraude, seksueel gedrag of alcoholgebruik. In een situatie waarin vragen over sensitieve items onmiddellijk worden gesteld, kan het zijn dat respondenten niet naar waarheid antwoorden vanwege privacy-overwegingen. De RR-methode zorgt er voor dat de privacy wordt gewaarborgd, omdat het antwoord van de respondent op een sensitieve vraag deels wordt bepaald door een kansmechanisme. De RR-methode is in Nederland toegepast, zie bijvoorbeeld Van Gils, Van der Heijden, Laudy en Ross (2003), en Elffers, Van der Heijden en Hezemans (2003).

Een voorbeeld van een RR-methode is de methode van Boruch (1971). Stel dat het antwoord op de sensitieve vraag *ja* of *nee* is. In Boruch's methode krijgt de respondent twee dobbelstenen. De sensitieve vraag wordt gesteld en voordat de respondent antwoordt, gooit hij met de dobbelstenen. Afhankelijk van het aantal ogen dat wordt gegooid, bepaalt de respondent zijn antwoord. Als het aantal ogen 2, 3 of 4 is, antwoordt de respondent *ja*. Als het aantal ogen 5, 6, 7, 8, 9, of 10 is, antwoordt de respondent naar waarheid. Als het aantal ogen 11 of 12 is, antwoordt de respondent *nee*. In deze methode is het kansmechanisme het gebruik van de dobbelstenen. Het mechanisme beschermt de privacy, want het aantal ogen dat wordt gegooid, wordt voor de interviewer verborgen gehouden. Omdat we het kansmechanisme goed kennen, kunnen we er rekening mee houden in de analyse van de geobserveerde antwoorden.

De post randomization methode (PRAM) kan worden gezien als een speciale toepassing van de RR-methode (Kooiman, Willenborg en Gouweleeuw 1997). PRAM beschermt data die al is verzameld en die wordt vrijgegeven voor onderzoekers buiten het instituut dat de data verzamelde. Ook in deze situatie is de bescherming van de privacy nodig. Het kansmechanisme wordt bij PRAM uitgevoerd door een computer. De verstoorde data worden samen met informatie over het kansmechanisme, en onder bepaalde voorwaarden, vrijgegeven.

Dit boek beschrijft de verstoring door RR of PRAM met behulp van een misclassificatie model. Het eerste hoofdstuk introduceert RR en PRAM.

Hoofdstuk 2 gaat over het schatten van proporties en over het meten van afhankelijkheid. Met de methoden in dit hoofdstuk kunnen vragen worden beantwoord als bijvoorbeeld: Hoeveel mensen hebben er gefraudeerd? en Is fraude plegen afhankelijk van de indeling man/vrouw? Omdat er in het schatten rekening moet worden gehouden met het kansmechanisme, is de analyse van de gegevens anders dan in een standaardsituatie zonder RR of PRAM.

Hoofdstuk 3 behandelt het fitten van loglineare modellen voor RR-data of PRAM-data. Loglineare modellen beschrijven complexe, meer-dimensionale structuren voor afhankelijkheid. Met de methode in dit hoofdstuk kan bijvoorbeeld de volgende vraag worden beantwoord: Is fraude plegen afhankelijk van zowel de indeling man/ vrouw alsook van de grootte van de woonplaats?

In hoofdstuk 4 wordt een specifiek onderzoek beschreven waarin sommige respondenten via RR vragen beantwoordden en andere respondenten direct antwoordden. Opnieuw worden loglineare modellen gefit, maar nu wordt er ook rekening gehouden met het feit dat niet alle respondenten de RR-methode gebruiken.

Regressiemodellen worden geschat in hoofdstuk 5. In dit hoofdstuk wordt de situatie beschreven waarin RR-variabelen onafhankelijke variabelen in een standaard lineair regressiemodel zijn. Een situatie als deze kan ontstaan wanneer seksueel gedrag met behulp van RR wordt gemeten en het verband met een afhankelijke variabele wordt onderzocht door een lineair regressiemodel.

Hoofdstuk 6 gaat over PRAM en over de informatie over het kansmechanisme die wordt meegegeven met de verstoorde data. Omdat PRAM wordt toegepast met behulp van een computer op bestaande data, zijn er meer mogelijkheden voor het verstrekken van informatie over het kansmechanisme dan bij RR. Ook worden er in dit hoofdstuk maten gegeven voor de privacybescherming wanneer PRAM is toegepast.

Het boek laat zien dat, gegeven een aantal assumpties, de analyse van RR-data zeer goed mogelijk is. Het boek laat ook zien, vooral in hoofdstuk 3, dat die assumpties niet altijd houdbaar zijn en dat in sommige gevallen verder onderzoek nodig is voor het beter modelleren van RR-data.

## Curriculum Vitae

Ardo van den Hout was born in Dirksland, the Netherlands, on March 8, 1968. After completing secondary school (VWO) in 1989, he started studying philosophy and mathematics at the University of Nijmegen. In 1996, he graduated in philosophy (cum laude), and in 1997 he graduated in mathematics. From 1997 to 1999, he followed a two-year post-Master's program at the Department of Mathematics and Computer Science at Delft University, which included a one-year work placement at Statistic Netherlands in Voorburg. In 2000, he started his Ph.D. research which was a joint project of Statistic Netherlands and the Department of Methodology and Statistics of the Faculty of Social Sciences at Utrecht University. Currently, he is employed as a postdoc researcher at the Department of Methodology and Statistics at Utrecht University.