Journal of Classification 29:76-90 (2012)
DOI: 10.1007/s00357-011-9094-8

A Property of the CHAID Partitioning Method for Dichotomous Randomized Response Data and Categorical Predictors

Pier Francesco Perri

University of Calabria, Italy

Peter G.M. van der Heijden

Utrecht University, The Netherlands

Abstract: In this paper, we present empirical and theoretical results on classification trees for randomized response data. We considered a dichotomous sensitive response variable with the true status intentionally misclassified by the respondents using rules prescribed by a randomized response method. We assumed that classification trees are grown using the Pearson chi-square test as a splitting criterion, and that the randomized response data are analyzed using classification trees as if they were not perturbed. We proved that classification trees analyzing observed randomized response data and estimated true data have a one-to-one correspondence in terms of ranking the splitting variables. This is illustrated using two real data sets.

Keywords: Pearson chi-square; Forced response method; Prevalence estimation; CHAID method.

Most of the research of Pier Francesco Perri was done during his stay at the Department of Methodology and Statistics, University of Utrecht (The Netherlands). His work was partly supported by the *research voucher* awarded by Regione Calabria, Italy.

Authors' Addresses: P.F. Perri, Department of Economics and Statistics, University of Calabria, Via P. Bucci, Cubo 0C, 87036 Arcavacata di Rende, Italy, email: pierfrancesco.perri@unical.it; P.G.M. van der Heijden, Department of Methodology and Statistics, Utrecht University, PO Box 80.140, 3508 TC Utrecht, The Netherlands, email: p.g.m.vanderheijden@uu.nl

Published online 29 October 2011

1. Introduction

The problem of parsimoniously describing large data sets and/or predicting the level of a dichotomous response variable from a set of measurements is a persistent issue in many fields including the medical and social sciences, environmental research, market segmentation, e-commerce, Internet user profiling and so on. Classification trees are a nonparametric data mining tool for *pattern recognition* formulated as tree structures showing how the level of a response variable can be determined or predicted by a set of explanatory variables. Roughly speaking, classification trees identify a number of exhaustive and mutually exclusive subgroups of a population whose members share common characteristics that influence the dependent variable. Since the work by Breiman, Friedman, Olshen, and Stone (1984) on *classification and regression trees* (CART), a great deal of research has been done in this area. However, as far as we know, the CART of *randomized response* (RR) data has not yet been studied from a methodological point of view.

In this paper we considered a number of classification questions of practical concern that arise in the context of randomized response data. In particular, since the status of surveyed people is represented by a dichotomous response in many surveys, analyzing a dichotomous sensitive variable for classification purposes is the objective of this work.

Randomized response is an ingenious tool introduced by Warner (1965) for collecting data on sensitive topics while ensuring respondent anonymity. The technique is based on a randomization device that respondents use to perturb their true answers. Thus, their true sensitive status remains undisclosed and privacy is protected. Assuming that the observed randomized responses are analyzed by using classification trees as if they were not perturbed, we showed that classification trees analyzing, on the one hand, observed RR data and, on the other hand, estimated true data, have a one-to-one correspondence in terms of the ranking of the splitting variables when the Pearson chi-square test is used as splitting criterion.

Indeed, a similar correspondence between true data and RR data has been illustrated by a numerical example in van der Heijden and Böckenholt (2008). Therefore, the aim of this paper is twofold: (i) to illustrate relations between classification trees for observed RR data and estimated true data and (ii) to formally explain the empirical evidence with a rigorous proof of the intuitive idea in van der Heijden and Böckenholt (2008).

2. Randomized Response Designs

Randomized response is an interview technique that can be used if sensitive, highly personal and stigmatizing questions are asked and respon-

dents are reluctant to answer directly. Posing direct questions in this context can lead to refusal to cooperate or untruthful, non-incriminating or socially acceptable answers (Fox and Tracy 1986; Chaudhuri and Mukerjee 1988). Sensitive questions can pertain, for example, to gambling, alcoholism, sexual violence or abuse, drug addiction, xenophobia or non-compliance with rules and regulations. Many studies have assessed the validity of RR methods and shown that they produce more reliable and honest answers than other conventional data collection methods; see Lensvelt-Mulders, Hox, van der Heijden, and Maas (2005) for a meta analysis.

In the original approach proposed by Warner (1965), each respondent is presented with two statements, A and B, with B the complement of A. For example, statement A is *I used hard drugs last year* and statement B is *I did not use hard drugs last year*. A randomizing device determines which statement is to be selected. Without revealing the outcome of the randomizer, the interviewee gives a *yes* or *no* response. Thus, the true status of the respondent remains uncertain and privacy is protected. Yet, the prevalence of the sensitive characteristic can be estimated from a sample of *yes* or *no* answers.

When data are collected using an RR method, the responses received by the interviewer on sensitive topics can differ from the responses that interviewees would give if they were completely honest in reporting their real status directly. These truthful responses denote what we call "true data" and are generally unknown. The collected *yes* and *no* responses are partly misclassified data, and we refer to these data as "observed RR data". Finally, without loss of generality, we reserve the term "RR data" for misclassified data which are not necessarily observed albeit theoretically definable.

In the Netherlands, RR surveys have mainly been carried out to measure the prevalence of regulatory non-compliance in the area of social welfare, unemployment and disability benefits (see Lensvelt-Mulders, van der Heijden, Laudy, and van Gils 2006). In these surveys, Boruch's (1971) *forced response design* is often used because of its ascertained statistical efficiency. The method works as follows. Consider a dichotomous sensitive question with responses *yes* and *no*. Response *yes* denotes that the true status is stigmatizing. Respondents are instructed to throw two dice and keep the outcome hidden from the interviewer. If the sum of the outcomes is 2, 3 or 4 (probability of $1/6$) the respondent is asked to answer *yes*, irrespective of his or her true status; if the sum is 11 or 12 (probability of $1/12$) the respondent is asked to answer *no*; if the sum is 5, 6, 7, 8, 9 or 10 the respondent is asked to reveal his or her true status.

Let θ_1 and $\theta_2 = 1 - \theta_1$ be the probabilities of a *yes* or *no* response for the true status. These are unknown probabilities which are to be estimated. Analogously, let θ_1^* and $\theta_2^* = 1 - \theta_1^*$ be the probabilities of observing a *yes*

or *no* randomized response. Then, in Boruch's design

$$\theta_1^* = \frac{1}{6} + \frac{3}{4}\theta_1, \quad \theta_2^* = \frac{1}{12} + \frac{3}{4}\theta_2. \tag{1}$$

To estimate the prevalence of *yes* responses in the population, i.e. the proportion of people with social stigma, it is useful to rewrite θ_1^* and θ_2^* in terms of conditional misclassification probabilities.

Consider the sensitive variable Y with the two categories, $yes \equiv 1$ and $no \equiv 2$, and let

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

be the *conditional misclassification probabilities*. These probabilities can be arranged in the *transition matrix*

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

If $\boldsymbol{\theta}^* = (\theta_1^*, \theta_2^*)'$ and $\boldsymbol{\theta} = (\theta_1, \theta_2)'$ the expressions in (1) can be given in the matrix form

$$\boldsymbol{\theta}^* = \mathbf{P}\boldsymbol{\theta}. \tag{2}$$

The *prevalence estimate* of a sensitive behavior can be obtained from model (2). If $\mathbf{P}$ is non-singular and $\widehat{\boldsymbol{\theta}}^*$ denotes the vector of the observed proportions of *yes* and *no* responses in a sample, $\boldsymbol{\theta}$ can be unbiasedly estimated by the moment estimator

$$\widehat{\boldsymbol{\theta}} = \mathbf{P}^{-1}\widehat{\boldsymbol{\theta}}^* \tag{3}$$

and "estimated true data" can be obtained accordingly. We observe that in practice it is possible to obtain non-admissible estimates $\widehat{\boldsymbol{\theta}}$ if the observed proportion of *yes* is lower than the probability of a forced *yes*.

We also note that it is sometimes felt to be a drawback of RR that only univariate prevalence estimates can be obtained. This is incorrect, as there is a whole range of tools that can be used for the analysis of RR data. An early overview can be found in Fox and Tracy (1986). Basically, a likelihood is set up in terms of the observed randomized data using θ_1^* and θ_2^*. In the likelihood, these proportions are replaced by a function of the unknown probabilities θ_1 and θ_2 using the expressions in (1). For example, by defining a logistic regression model for θ_1 and θ_2, Maddala (1983) and Scheers and Dayton (1988) used this approach to develop logistic regression for RR data. Similarly, models for multivariate RR data have been developed, among others, by Fox (2005) and Böckenholt and van der Heijden (2007).

3. Classification Trees and RR

Classification trees are usually built on data collected using direct questioning techniques. Here we study the properties of classification trees if RR data are considered. The main issue is whether splitting the population based on estimated true data differs from splitting the population on the basis of observed RR data if the misclassification of these data is ignored. As a matter of fact, the problem has been partially investigated from a different perspective by van der Heijden and Böckenholt (2008) who used an example to demonstrate that splitting the population based on RR data does not lead to a different splitting if the true data are considered. In this section, we provide rigorous proof of their finding and generalize it to other classification tree problems. We start from motivating examples and then provide a general result that formally establishes some relations between true data, RR data and estimated true data.

3.1 Motivating Examples

Example 1. Multiple Dichotomous Explanatory Variables

A main concern in classification is the selection of the splitting variables. Given a set of explanatory variables, which ones are the most important? The critical issue is how to rank the variables that, while not giving the first best split of a node, may give the second and third node and so on. One possible way to order the splitting variables is by considering the p-value of the Pearson chi-square (χ^2) test for a contingency table. The variable with the smallest significance is used for the first split and after the first split for each of the two nodes separately, the second smallest p-value leads to the second split and so on. Let us assume that the explanatory variables have been ranked on the basis of observed RR data where the misclassification in the data has been ignored. Then the question would be: Is this ranking also valid for the estimated true data?

Let us consider the following example with the two dichotomous explanatory variables Gender (*female*, *male*) and Former Job (*temporary*, *permanent*) and a $2 \times 2 \times 2$ contingency table. Data are taken from a survey on regulatory compliance regarding unemployment benefits (see for details Lensvelt-Mulders et al. 2006) where interviewees are asked to respond with a forced *yes* or *no* to the question: *Have you done an odd job for friends or relatives in the past 12 months and received money for this job?* The observed forced responses are reported in Table 1 together with the estimated true frequencies obtained from (3).

In a classification tree framework, the question is whether the first splitting variable is Gender or Former Job and whether there is a one-to-one

Table 1. (a) Gender by Former Job by Randomized Answers and (b) Gender by Former Job by Estimated True Answers.

(a) Observed RR data			
Gender by Former Job	*yes*	*no*	total
female, temporary	103	277	380
female, permanent	104	336	440
male, temporary	112	199	311
male, permanent	222	472	694
total	541	1284	1825

(b) Estimated true data			
Gender by Former Job	*yes*	*no*	total
female, temporary	52.89	327.11	380
female, permanent	40.89	399.11	440
male, temporary	80.22	230.78	311
male, permanent	141.78	552.22	694
total	315.78	1509.22	1825

Table 2. 2×2 Marginal distributions: (a) Gender [Former Job] by Randomized Answers and (b) Gender [Former Job] by Estimated True Answers. The Pearson chi-square for marginal Gender [Former Job] table is $\chi^2 = 13.822$ [1.153] in (a) and $\chi^2 = 35.816$ [2.988] in (b).

(a) Observed RR data			
Gender	*yes*	*no*	total
female	207	613	820
male	334	671	1005
total	541	1284	1825

(b) Estimated true data			
Gender	*yes*	*no*	total
female	93.78	726.22	820
male	222.00	783.00	1005
total	315.78	1509.22	1825

Former Job	*yes*	*no*	total
temporary	215	476	691
permanent	326	808	1134
total	541	1284	1825

Former Job	*yes*	*no*	total
temporary	133.11	557.89	691
permanent	182.67	951.33	1134
total	315.78	1509.22	1825

correspondence between the order of the splitting variable on the true and observed RR data. First, the marginal 2×2 distributions for Gender and Former Job need to be considered and the Pearson χ^2 calculated. Table 2 provides an illustration.

For the observed RR answers, the χ^2 for Gender and Former Job is, respectively, $\chi^2_{Gender,a} = 13.822$ and $\chi^2_{Job,a} = 1.153$, so that $\chi^2_{Gender,a}$ is 11.99 times larger than $\chi^2_{Job,a}$. Thus, the first splitting variable is Gender. Analogously with the estimated true data, we get $\chi^2_{Gender,b} = 35.816$ and $\chi^2_{Job,b} = 2.988$ so that $\chi^2_{Gender,b}$ is 11.99 time larger than $\chi^2_{Job,b}$.

This shows that the two data sets have an identical ratio of the χ^2's, and the order of the χ^2's for the two explanatory variables derived for the estimated true data is identical to the order of the χ^2's derived for observed RR data.

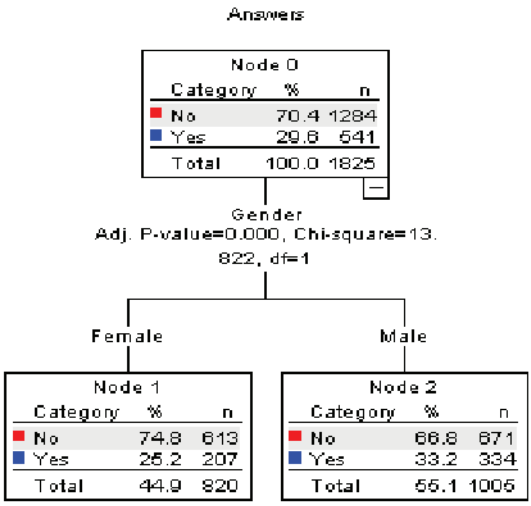

Figure 1. Classification tree built by CHAID method for observed RR data

Figure 1 shows the split of interviewed people by the *chi-square automatic interaction detector* (CHAID), an efficient technique for tree growing developed by Kass (1980) and investigated, among others, by Biggs, de Ville, and Suen (1991). The sample of respondents is first divided into two sub-groups according to the significant splitting variable Gender. In a second step, the prevalence of *yes* in each node of the RR tree can be transformed by using relation in (3) to find estimates of the prevalence in the true data. For example, from the observed proportion of *yes* responses in the parent node, we can estimate the prevalence in the true data as $\widehat{\theta}_1 = \frac{541/1825-1/6}{3/4} = 0.173$, i.e. $315.78(= 0.173 \times 1825)$ stigmatized individuals. Analogously, the prevalence estimation in the female population is $\widehat{\theta}_1 = \frac{207/820-1/6}{3/4} = 0.114$, and $\widehat{\theta}_1 = \frac{334/1005-1/6}{3/4} = 0.221$ in the male population. So, if we had a new individual whose gender only is known, we could use these estimates to classify the individual on the sensitive variable and, thus, employ the classification tree for prediction purposes.

We note that we shall later prove that these results not only hold true for two dichotomous explanatory variables, but can be extended to any number of dichotomous explanatory variables. The reason is that extending the number of variables does not change the marginal bivariate tables such as those reported in Table 2, it only extends the number of marginal bivariate tables. For this larger number of tables, the relative sizes of the χ^2's in the observed RR data will be identical to the relative sizes of the χ^2's in the estimated true data.

Example 2. An Ordinal Explanatory Variable

Let us consider now tree growing with an explanatory variable that is ordinal and has more than two levels. Data are taken again from a survey on regulatory compliance in unemployment benefits (Lensvelt-Mulders et al. 2006). We focus on the ordinal explanatory variable Perceived Detection with five levels. The interviewees are confronted with the statement *The odds that I will be detected when I earn some extra illegal income are not very large* and are asked to respond using a five-point rating scale with labels *1. yes, completely agree; 2. yes, agree; 3. no, opinion; 4. no, disagree* and *5. no, completely disagree*. The observed RR data and the estimated true data are reported in Table 3.

The sample is split into sub-groups by merging two or more consecutive levels so that more homogeneous levels are obtained. A classification tree built on observed RR data splits the sample in two groups. The first group has level 1 and 2 of the explanatory variable, and the second group has levels 3, 4 and 5. If we consider an alternative split, for example $(1,2),(3),(4,5)$ and compare it to the split $(1,2)$, $(3,4,5)$ as in the example for two dichotomous explanatory variables, we find that the ratio between the χ^2's for the RR data is identical to the ratio between the χ^2's on the estimated true data. Table 4 illustrates this one-to-one correspondence.

Moreover, similarly to Example 1, the examples show that the estimated probabilities of finding regulatory non-compliance behavior in the population under study are: 0.202 in the group $(1,2)$, 0.125 in $(3,4,5)$, 0.147 in (3) and 0.0628 in $(4,5)$. The prevalence estimation on the entire population is 0.178.

Similar results can be obtained by a more refined classification, for example by comparing the χ^2's for classifications such as (1), (2), (3), (4) and (5) with $(1,2)$, (3), (4) and (5). The reason is that extending the number of levels does not change the classifications reported in Table 4, it only extends the number of classifications. For this larger number of classifications, the relative sizes of the χ^2's in the observed RR data is identical to the relative sizes of the χ^2's in the estimated true data. By stretching this analogy to the limit, similar results also hold true if the explanatory variable is continuous and is classified into a number of ordered levels. This will make the number of possible classifications very large, but will not change the relative magnitude of the χ^2's.

3.2 Some Theoretical Results

We now prove that the results in the illustrations are theoretically valid. We first give a general theorem that formalizes and extends the one-to-one correspondence between the true data and the RR data suggested in

Table 3. (a) Perceived Detection by Randomized Answers and (b) Perceived Detection by Estimated True Answers.

(a) Observed RR data			
Detection (level)	*yes*	*no*	total
1	90	189	279
2	286	615	901
3	111	290	401
4	24	98	122
5	7	16	23
total	518	1208	1726

(b) Estimated true data			
Detection (level)	*yes*	*no*	total
1	58.00	221.00	279
2	181.11	719.89	901
3	58.89	342.11	401
4	4.89	117.11	122
5	4.22	18.78	23
total	367.33	1418.89	1726

Table 4. Pearson chi-square on reduced contingency tables: (a) $\chi^2 = 6.096$ and $\chi^2 = 8.110$ respectively for 2×2 and 3×2 contingency tables; (b) $\chi^2 = 15.564$ and $\chi^2 = 20.703$, respectively, for 2×2 and 3×2 contingency tables. For contingency tables in (a) the ratio between χ^2's is 0.752. Identical ratio in (b).

(a) Observed RR data			
Detection	*yes*	*no*	total
(1,2)	376	804	1180
(3,4,5)	142	404	546
total	518	1208	1726

(b) Estimated true data			
Detection	*yes*	*no*	total
(1,2)	239.11	940.89	1180
(3,4,5)	68.00	478.00	546
total	307.11	1418.89	1726

Detection	*yes*	*no*	total
(1,2)	376	804	1180
(3)	111	290	401
(4,5)	31	114	145
total	518	1208	1726

Detection	*yes*	*no*	total
(1,2)	239.11	940.89	1180
(3)	58.89	342.11	401
(4,5)	9.11	135.89	145
total	307.11	1418.89	1726

van der Heijden and Böckenholt (2008). Then, we prove the correspondence between the observed RR data and the estimated true data as a consequence of the main result.

Let us consider a population of size n and let Y be a sensitive variable taking values $yes \equiv 1$ and $no \equiv 2$. Let $\theta_1 = P(Y = 1)$ and $\theta_2 = 1 - P(Y = 1)$ be the probabilities of *yes* or *no* answers for the true status. Accordingly, let Y^* be the dichotomous variable that misclassifies the true status using a generic RR device and

$$\theta_1^* = P(Y^* = 1) = a_1 + b_1\theta_1, \quad \theta_2^* = 1 - P(Y^* = 1) = a_2 + b_2\theta_2, \quad (4)$$

with the constants a_j and b_j, $j = 1, 2$, as *design parameters* induced by

the RR method. For instance, under the forced response design, $a_1 = 1/6$, $a_2 = 1/12$ and $b_1 = b_2 = 3/4$.

Let $Z = \{z_1, \ldots, z_k\}$ be an explanatory variable and consider the $k \times 2$ contingency table of Y by Z, with absolute joint and marginal frequencies expressed as $n_{ij} = \text{Freq}(Z = z_i, Y = y_j)$, $n_{i.} = \text{Freq}(Z = z_i) = \sum_{j=1}^{2} n_{ij}$, $n_{.j} = \text{Freq}(Y = y_j) = \sum_{i=1}^{k} n_{ij}$, $i = 1, 2, \ldots, k$ and $j = 1, 2$.

The corresponding frequencies for the RR data contingency table of Y^* by Z follow from (4) and are given by

$$n^*_{ij} = n_{i.}\left(a_j + b_j \frac{n_{ij}}{n_{i.}}\right), \quad n^*_{.j} = n\left(a_j + b_j \frac{n_{.j}}{n}\right), \quad n_{i.} = n^*_{i.}.$$

Finally, let

$$\chi^2 = \sum_{i,j} \frac{(n_{ij} - \widehat{n}_{ij})^2}{\widehat{n}_{ij}}, \quad \widehat{n}_{ij} = \frac{n_{i.} \times n_{.j}}{n}$$

and

$$\chi^{2*} = \sum_{i,j} \frac{(n^*_{ij} - \widehat{n}^*_{ij})^2}{\widehat{n}^*_{ij}}, \quad \widehat{n}^*_{ij} = \frac{n^*_{i.} \times n^*_{.j}}{n}$$

be the Pearson chi-square for the true and the RR contingency tables, respectively. Now, the following result is established.

Theorem 1. *Given an RR device with a probabilistic structure as in (4), there is the following relation between χ^{2*} and χ^2*

$$\frac{\chi^2}{\chi^{2*}} = \frac{\alpha + \beta\delta}{1 + \delta}, \tag{5}$$

with

$$\alpha = \frac{n}{b_1^2 n_{.1}}\left(a_1 + b_1 \frac{n_{.1}}{n}\right), \ \beta = \frac{n}{b_2^2 n_{.2}}\left(a_2 + b_2 \frac{n_{.2}}{n}\right), \ \delta = \frac{b_2^2}{b_1^2} \frac{a_1 + b_1 \dfrac{n_{.1}}{n}}{a_2 + b_2 \dfrac{n_{.2}}{n}}.$$

The proof of the theorem is reported in the Appendix.

Consequences of Theorem 1:

1. The ratio in (5) is only a function of the constants α, β and δ that depend on the marginal distribution of the true answers. Explanatory variables do not influence the marginal frequencies $n_{.1}$ and $n_{.2}$, implying that the ratio χ^2/χ^{2*} always takes the same value $\frac{\alpha+\beta\delta}{1+\delta}$ whatever the explanatory variable. From this, the next result follows.

2. Consider the dichotomous variables Y and Y^* which model the true and the observed randomized responses, respectively. Let $\{Z_i\}_{i=1}^q$ be a set of explanatory variables, each with m_i categories. If χ_i^2 $[\chi_i^{2*}]$ denotes the Pearson chi-square of the marginal contingency table of Y $[Y^*]$ by Z_i, then from (5) we easily obtain

$$\frac{\chi_i^2}{\chi_j^2} = \frac{\chi_i^{2*}}{\chi_j^{2*}}. \tag{6}$$

   This result proves the idea stated in van der Heijden and Böckenholt (2008) according to which the ranking of the splitting variables for the true data is identical to the ranking of the splitting variables for the RR data.

3. The same one-to-one correspondence given at point 2 holds true if a set of ordinal explanatory variables is considered.

4. Equality in (6) can be used to derive another result of practical concern. Consider the sensitive dichotomous variable Y and the explanatory continuous variable Z, whose values are supposed to be grouped in m classes. Given the $m \times 2$ contingency table of Y by Z, let us reduce the dimensionality of the table by aggregating two or more consecutive classes of Z. For example, if $Z = \{c_1, \ldots, c_5\}$, a few possibilities are $i_1 \equiv \{c_1, c_2, c_3, (c_4, c_5)\}$ and $i_2 \equiv \{c_1, c_2, (c_3, c_4, c_5)\}$. In the first case, classes c_4 and c_5 merge, and in the second case classes c_3, c_4 and c_5 collapse into a unique class. Applying the same merging to the RR data, the question is: What happens to the ratio between the χ^2's computed on reduced contingency tables for the true and the RR data? Is a similar conclusion to the one in (6) still valid? The answer is that the ratio of the χ^2's on the true data is still identical to the ratio on the RR data. The reason is immediately clear since, by changing the notation, each possible reduced table can be ascribed to a different explanatory variable. In this way, the reasoning can be redirected to the equivalence in (6).

5. Previous results have been discussed in terms of probabilities for the true status and the probability for the observed (and misclassified) status. In practice, if RR techniques are used to collect data, only the observed RR data are available and the observed RR proportions can nonetheless be used to obtain an estimate of the probabilities for the true status by means of relation in (3). This is the situation that we have actually considered in the examples of Section 3.1.

   The re-parameterization of the expressions in (4), according to (3), allows us to extend the results to the relation between Y^* and the es-

timated true data, say $\widehat{Y}$. Therefore, the same one-to-one correspondence holds true if the observed RR and the estimated true data are considered. This formally justifies the empirical evidence in Examples 1-2.

4. Conclusions

There are many tools for the analysis of RR data in the literature (see Section 2). In this paper we have focused on classification trees built on data intentionally misclassified by respondents following the rules of the forced response design. The usual approach in the context of RR data is to construct adjusted tools for their analysis. However, we have shown that this is not necessary for classification trees. Our basic finding shows that starting from the observed RR data, if one is interested in a classification tree of the estimated true data, it is not necessary to derive the estimated true data. Instead one can employ a two-step procedure: in stage 1 we ignore the fact that the data have been subjected to randomized response and are analyzed directly using classification trees; this yields the correct tree for the estimated true data. In stage 2, the prevalence estimate of the subpopulations found by the classification tree analysis of the observed RR data is rescaled to yield the correct prevalence estimates for the true data. Thus, this two-stage procedure leads to the correct classification tree as well as to unbiased prevalence estimates. Since this rescaling is a very simple operation (see relation in (3)), this result is of great practical importance for researchers who have collected RR data.

Classification trees obtained for observed RR data and estimated true data coincide with those that could be built if the true, not-perturbed, data were considered. In fact, we have confirmed an informal result given in van der Heijden and Böckenholt (2008) that establishes a one-to-one correspondence in terms of splitting variables between the true data and the RR data. In other words, the classification of the population on the basis of RR data, and thus on estimated true data, produces the same classification results as if the true data were really available.

An anonymous reviewer has pointed out that the findings of our paper have broader implications than indicated in an earlier version of this manuscript. First, the invariance property that we have shown does not only hold for the Pearson chi-square statistic but also for the Gini index and the Goodman and Kruskal τ index, due to the close mathematical relation between the Pearson chi-square statistic and these indices (see Bishop, Fienberg, and Holland 1975). Second, if all explanatory variables are categorical and the response variable is dichotomous, then the CHAID method discussed herein and the CART method (Breiman et al. 1984) yield the same tree in the partitioning phase (see Mola and Siciliano 1997); hence, it turns

out that our findings also hold for the CART method. Finally, Siciliano and Mola (2000) extended CART splitting to multivariate trees for multivariate independent response variables where our theoretical results remain still valid. We gratefully acknowledge these remarks.

Appendix

Theorem 1. *For a given RR device with probabilities like those in (4), there is the following relation between χ^{2*} and χ^2*

$$\frac{\chi^2}{\chi^{2*}} = \frac{\alpha + \beta\delta}{1+\delta},$$

with

$$\alpha = \frac{n}{b_1^2 n_{.1}}\left(a_1 + b_1\frac{n_{.1}}{n}\right),\ \beta = \frac{n}{b_2^2 n_{.2}}\left(a_2 + b_2\frac{n_{.2}}{n}\right),\ \delta = \frac{b_2^2}{b_1^2}\frac{a_1 + b_1\dfrac{n_{.1}}{n}}{a_2 + b_2\dfrac{n_{.2}}{n}}.$$

Proof: Let $\xi_{ij} = \frac{(n_{ij}-\widehat{n}_{ij})^2}{\widehat{n}_{ij}}$ and $\xi_{i.} = \sum_j \xi_{ij}$. Then, for the true data, express the Pearson chi-square as $\chi^2 = \sum_{i,j}\xi_{ij} = \sum_i \xi_{i.}$. Analogously, the notation ξ^*_{ij} and $\xi^*_{i.}$ is adopted for the RR data. Consider the ratio ξ_{ij}/ξ^*_{ij}. We have

$$\begin{aligned}
\frac{\xi_{ij}}{\xi^*_{ij}} &= \frac{\left(n_{ij} - \dfrac{n_{i.}n_{.j}}{n}\right)^2}{\dfrac{n_{i.}n_{.j}}{n}} \frac{n_{i.}\left(a_j + b_j\dfrac{n_{ij}}{n_{i.}}\right)}{\left[n_{i.}\left(a_j + b_j\dfrac{n_{ij}}{n_{i.}}\right) - n_{i.}\left(a_j + b_j\dfrac{n_{.j}}{n}\right)\right]^2} \\
&= \frac{n}{n_{.j}} n_{i.}^2 \left(\frac{n_{ij}}{n_{i.}} - \frac{n_{.j}}{n}\right)^2 \frac{a_j + b_j\dfrac{n_{ij}}{n_{i.}}}{b_j^2 n_{i.}^2\left(\dfrac{n_{ij}}{n_{i.}} - \dfrac{n_{.j}}{n}\right)^2} \\
&= \frac{n}{b_j^2 n_{.j}}\left(a_j + b_j\frac{n_{ij}}{n_{i.}}\right). \qquad (7)
\end{aligned}$$

From (7) we obtain

$$\frac{\xi_{i1}}{\xi^*_{i1}} = \frac{n}{b_1^2 n_{.1}}\left(a_1 + b_1\frac{n_{i1}}{n_{i.}}\right) = \alpha\,, \qquad (8)$$

$$\frac{\xi_{i2}}{\xi^*_{i2}} = \frac{n}{b_2^2 n_{.2}}\left(a_2 + b_2\frac{n_{i2}}{n_{i.}}\right) = \beta. \qquad (9)$$

Consider the ratio ξ^*_{i2}/ξ^*_{i1}. From (8) and (9) we get

$$\frac{\xi^*_{i2}}{\xi^*_{i1}} = \frac{\alpha}{\beta}\frac{\xi_{i2}}{\xi_{i1}}$$
$$= \frac{n\left(a_1 + b_1\dfrac{n_{i1}}{n_{i.}}\right)}{b_1^2 n_{.1}} \frac{b_2^2 n_{.2}}{n\left(a_2 + b_2\dfrac{n_{i2}}{n_{i.}}\right)} \frac{\left(n_{12} - \dfrac{n_{1.}n_{.2}}{n}\right)^2}{\dfrac{n_{1.}n_{.2}}{n}} \frac{\dfrac{n_{1.}n_{.1}}{n}}{\left(n_{11} - \dfrac{n_{1.}n_{.1}}{n}\right)^2}$$
$$= \frac{b_2^2}{b_1^2}\frac{a_1 + b_1\dfrac{n_{i1}}{n_{i.}}}{a_2 + b_2\dfrac{n_{i2}}{n_{i.}}}\left(\frac{nn_{12} - n_{1.}n_{.2}}{nn_{11} - n_{1.}n_{.1}}\right)^2,$$

where

$$\frac{nn_{12} - n_{1.}n_{.2}}{nn_{11} - n_{1.}n_{.1}} = -1.$$

In fact

$$(nn_{12} - n_{1.}n_{.2}) + (nn_{11} - n_{1.}n_{.1}) =$$
$$n(n_{12} + n_{11}) - n_{1.}(n_{.2} + n_{.1}) = nn_{1.} - n_{1.}n = 0.$$

Therefore, we obtain

$$\frac{\xi^*_{i2}}{\xi^*_{i1}} = \frac{b_2^2}{b_1^2}\frac{a_1 + b_1\dfrac{n_{i1}}{n_{i.}}}{a_2 + b_2\dfrac{n_{i2}}{n_{i.}}} = \delta. \tag{10}$$

Let us focus now on $\xi_{i.}$ and consider (8) and (9). We have $\xi_{i.} = \xi_{i1} + \xi_{i2} = \alpha\xi^*_{i1} + \beta\xi^*_{i2}$. But, from (10), $\xi^*_{i2} = \delta\xi^*_{i1}$ and then we get $\xi_{i.} = \xi^*_{i1}(\alpha + \beta\delta)$. Finally, consider $\xi^*_{i.} = \xi^*_{i1} + \xi^*_{i2}$. From (10) we have $\xi^*_{i.} = \xi^*_{i1}(1 + \delta)$. Therefore

$$\frac{\xi_{i.}}{\xi^*_{i.}} = \frac{\alpha + \beta\delta}{1 + \delta},$$

and

$$\frac{\chi^2}{\chi^{2*}} = \frac{\sum_i \xi_{i.}}{\sum_i \xi^*_{i.}} = \frac{\alpha + \beta\delta}{1 + \delta}.$$

This concludes the proof.

■

References

BIGGS, D., DE VILLE, B., and SUEN, E. (1991), "A Method of Choosing Multiway Partitions for Classification and Decision Trees", *Journal of Applied Statistics, 18*, 49–62.

BISHOP, Y.M.M., FIENBERG, S.E., and HOLLAND P.W. (1975), *Discrete Multivariate Analysis. Theory and Practice*, Massachusetts: The MIT Press.

BÖCKENHOLT, U., and VAN DER HEIJDEN, P.G.M. (2007), "Item Randomized-Response Models for Measuring Noncompliance: Risk-return Perceptions, Social Influences, and Self-Protective Responses", *Psychometrika, 72*, 245–262.

BORUCH, R.F. (1971), "Assuring Confidentiality of Response in Social Research: A Note on Strategies", *The American Sociologist, 6*, 308–311.

BREIMAN, L., FRIEDMAN, J.H., OLSHEN, R.A., and STONE, C.J. (1984), *Classification and Regression Trees*, Boca Raton: Chapmann & Hall/CRC.

CHAUDHURI, A., and MUKERJEE, R. (1988), *Randomized Response: Theory and Techniques*, New York: Marcel Dekker, Inc.

FOX, J.A., and TRACY, P.E. (1986), *Randomized Response: A Method for Sensitive Survey*, Newbury Park: Sage Publication, Inc.

FOX, J.P. (2005), "Randomized Item Response Theory Models", *Journal of Educational and Behavioral Statistics, 30*, 1–24.

KASS, G.V. (1980), "An Exploratory Technique for Investigating Large Quantities of Categorical Data", *Applied Statistics, 29*, 119–127.

LENSVELT-MULDERS, G.J.L.M., HOX, J.J., VAN DER HEIJDEN, P.G.M., and MAAS, C.J.M. (2005), "Meta-Analysis of Randomized Response Research: Thirty-five Years of Validation", *Sociological Methods and Research, 33*, 319-3-48.

LENSVELT-MULDERS, G.J.L.M., VAN DER HEIJDEN, P.G.M., LAUDY, O., and VAN GILS, G. (2006), "A Validation of a Computer-assisted Randomized Response Survey to Estimate the Prevalence of Fraud in Social Security", *Journal of the Royal Statistical Society, Series A, 169*, 305–318.

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

MOLA, F., and SICILIANO, R. (1997), "A Fast Splitting Procedure for Classification and Regression Trees", *Statistics and Computing, 7*, 209–216.

SCHEERS, N.J., and DAYTON, M.C. (1988), "Covariate Randomized Response Models", *Journal of the American Statistical Association, 83*, 969-974.

SICILIANO, R., and MOLA, F. (2000), "Multivariate Data Analysis through Classification and Regression Trees", *Computational Statistics & Data Analysis, 32*, 285–301.

VAN DER HEIJDEN, P.G.M., and BÖCKENHOLT, U. (2008), "Applications of Randomized Response Methodology in e-Commerce", in *Statistical Methods in e-Commerce Research*, eds. W. Jank and G. Shmueli, New York: Wiley, pp. 401–416.

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