] Applications of randomized response methodology in e-commerce.

STATISTICAL METHODS IN ECOMMERCE RESEARCH

Chapter: Applications of randomized response methodology in e-commerce

Peter G.M. van der Heijden
Department of Methodology and Statistics , Utrecht University, PO Box 80.140, 3508 TC Utrecht, the Netherlands

Ulf Böckenholt
Faculty of Management, McGill University, 1001 Sherbrooke Street West, Montreal, QC H3A 1G5, Canada

A JOHN WILEY & SONS, INC., PUBLICATION

CHAPTER 1

APPLICATIONS OF RANDOMIZED RESPONSE METHODOLOGY IN E-COMMERCE

Abstract: Randomized response is a method for intentionally misclassifying or perturbing part of a data set. This misclassification can be done either in data collection by a respondent following instructions for misclassification laid out by a researcher, or in a statistical disclosure control setting by a researcher after the data are collected. Both applications have the aim to protect the privacy of individuals. In this chapter we review the literature on randomized response, focussing on similarities and differences between the two applications. In particular we summarize a new approach to take into account the problem that in the data collection context some respondents do not follow the randomized response instructions. We end with a section a short discussion of building a classification tree when the dependent variable is collected with randomized response. Privacy protection is a crucial objective in e-commerce work and we argue that randomized response methodology can make major contributions to this objective.

1.1 INTRODUCTION

Randomized response is a method for intentionally misclassifying or perturbing part of a data set. This misclassification can either be done by a researcher or by a respondent in a survey or other research study. The motivation is the same in both cases: Misclassification or data perturbation provide a convenient mechanism for controlling the amount of information that can be extracted from a data set. The web age has brought us an avalanche of information about individual behavior related to e-commerce, ranging from search for information about commodities to shopping and post-purchase activities. At a result, there has never been a stronger need to apply and develop statistical methods that enforce individual rights for controlling the dissemination of personal information. This chapter reviews and discusses the use of randomized response methodology as a means to provide this privacy protection in the context of two e-commerce problems.

The first problem is collecting e-commerce self-report data on sensitive topics while protecting the privacy of individuals. For example, in online-surveys individuals may be asked for self-reports on such sensitive topics as their financial situation (e.g., loans, or income), their personal uses of male cosmetics or drugs to overcome impotence, their health problems such as urinary incontinence, or medical prescriptions, but also on such potentially less sensitive issues as shopping expenditures, and default rates in credit card usage. It is well–known that surveys into topics like these lead to refusals to answer or untruthful answers. Randomized response has been shown to be the currently best method of dealing with such sensitive issues (see Lensvelt et al., 2005), and to also work work well in web-based applications (Lensvelt et al., 2006). Therefore we see much potential for RR in e-commerce data collection applications.

The second problem that randomized response can address in the context of e-commerce is statistical disclosure control (SDC). The need for SDC in the area of e-commerce has been reviewed recently by Fienberg (2006). Randomized response can be used to misclassify part of a data source so that the data source can be analyzed by third parties who are not the original owners of the data source. A special application in this area is in the context of data integration of multiple databases. Here multiple data bases are linked with the aim to mine the data while preserving data confidentiality. Privacy-preserving data-mining randomized-response methods misclassify the parts of the data that are *not* used for linking the databases (see, for a review, $www.csee.umbc.edu/\ kunliu1/research/privacy_review.html$; see for example Ambianis, Jakobsson and Lipmaa, 2003; Du and Zhan, 2003; Evfimievski, 2002). Related work is reported by Karr and Fienberg (2005), Fienberg (2005) and Trottini et al. (2004). Although statistical disclosure control applications of randomized response are well known and recently further developed under the name of PRAM (Gouweleeuw, Kooiman, Willenborg and de Wolf, 1998; Kooiman, Wil-

lenborg and Gouweleeuw, 1997), an abbreviation of post- randomization, these methods have not yet become mainstream.

Originally, randomized response was proposed by Warner (1965) as a tool for collecting information on sensitive topics. He offered a respondent two complementary questions such as A. I did use hard drugs last year and B. I did not use hard drugs last year. A randomizing device such as a dice, with an outcome unknown by the interviewer, indicates to the respondent which question had to be answered. If, for example for one dice, an outcome of 1, 2, 3 or 4 is linked to answering A and an outcome of 5 or 6 to answering B, one could view the answer of A as a correct answer and the answer to B as a misclassified answer, with a misclassification probability of 2/6. Due to the potential misclassification that a respondent may or may not encounter, respondents may feel safe to answer the sensitive topic, yet it is possible to estimate the prevalence of the sensitive topic defined in A.

Characteristic for this application of randomized response is that respondents are offered a sensitive question and the misclassification is carried out by the respondent. However, very early in the development of randomized response it was noticed that the same methodology could be applied as a tool for statistical disclosure control (Warner, 1971). I.e., after data are collected the partial misclassification of the data is carried out by the owner of the data. Thus the owner of the data can hand over the data to others without risking privacy violations of individuals. The receiving agency only needs to know about the misclassification key for the sample, so that statistical analyses can be carried out that take into account the misclassification (compare Chen, 1989).

Randomized response as a tool for data collection and as a tool for statistical disclosure control are mathematically identical. For example, the data collection application as well as the statistical disclosure control application use the same equations to obtain univariate estimates and variances. Yet, as was pointed out by van den Hout and van der Heijden (2002), in practice the procedures differ. Differences are

A. Randomized response as a data collection tool presupposes the cooperation of respondents in the sense that they are allowed to follow the rules prescribed by the randomized response design. There is evidence that some of the respondents do not do this. This problem has to be handled in the analysis phase. In contrast, randomized response as a SDC tool does not have this problem.

B. Whereas randomized response as a data collection tool is used mainly to randomize response variables (the sensitive topics), randomized response as a SDC tool can be used both to randomize explanatory variables (background characteristics of respondents) as well as to randomize the response variables of a study. When a single source of data is to be protected by SDC, it may be most efficient to use SDC to hide the identity of the respondent by misclassifying

his background characteristics, so that answers to sensitive topics in a questionnaire cannot be linked directly to individuals. When a data source is released with the aim to integrate it with other data sources, background characteristics should not be misclassified as these are necessary for linking but randomized response can be used for misclassifying the response variable. In applications where distinctions between response and explanatory variables are important, this may lead to different statistical analyses for both applications.

C. Related to B., randomized response as a data collection tool should use a randomization scheme that is perceived as protective by respondents, and, at the same time, lead to as much information as possible. The protective aspect calls for large misclassification probabilities, whereas the information demand calls for small misclassification probabilities. In randomized response as a tool for SDC the misclassification probabilities can be chosen in such a way that the outcome is optimal, without the necessity to take the perceptions of respondents into account.

In this chapter we give an overview on statistical methods for the analysis of randomized responses. We discuss bivariate and multivariate methods and discuss recent developments in this field. In section 1.2 and 1.3 of this chapter we give an overview of the literature on randomized response, with an emphasis on the similarities and differences of the two applications of randomized response. In regard to the differences, we focus on the statistical analysis issues A and B, introduced above. For issue C, we refer to van den Hout and Elamir (2006). In section 1.4 we summarize a new development that takes into account that respondents may not follow the regulations laid out by the randomized response design. As e-commerce data are used often for data mining purposes and classification trees are a popular mining tool, in section 1.5 we provide some new results regarding the use of classification trees in the context of randomized response data.

We note that in this chapter we focus on the methodology for misclassifying categorical data. Similar ideas are available for perturbing continuous data or mixtures of continuous and discrete data (see Fox & Tracy, 1986; Chaudhuri & Mukerjee, 1988). Since the same principles apply in these more general settings, we do not consider different data types explicitly.

1.2 UNIVARIATE ANALYSIS

This section introduces our notation and gives a short overview of available tools of analysis for applications of randomized response that are shared by the data collection and SDC approaches.

As an example, consider the following form of the so-called forced classification design (Boruch, 1971). Let the outcome of two dice determine whether the answers will be misclassified. For the

data collection approach this would mean that if the sum of the outcomes of the two dice is 2, 3 or 4 (probability 1/6), then the respondent is asked to give the answer "yes"; if the sum equals 11 or 12 (probability if 1/12), then the respondent is asked to give the answer "no"; if the sum is in the range 5-10 (probability is 3/4), the respondent is asked to reveal his true status. Let θ_1 be the probability of "yes" for the true status ("no" has probability $\theta_2 = 1 - \theta_1$) and let θ_1^* be the probability of "yes" for the observed (and partly misclassified) status ("no" has probability $\theta_2^* = 1 - \theta_1^*$). It follows that

$$\begin{aligned} \theta_1^* &= 1/6 + (3/4)\theta_1 \text{ and} \\ \theta_2^* &= 1/12 + (3/4)\theta_2. \end{aligned} \tag{1.1}$$

These equations can be rewritten in terms of conditional misclassification probabilities. Let the conditional misclassification probabilities be

$$p_{ij} = I\!P(\text{category } i \text{ is observed}| \text{ true category is } j), \tag{1.2}$$

then these probabilities can be collected in a matrix given by

$$\boldsymbol{P} = \begin{pmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{pmatrix} = \begin{pmatrix} 11/12 & 2/12 \\ 1/12 & 10/12 \end{pmatrix}. \tag{1.3}$$

If we collect the probabilities θ_1 and θ_2 in a vector $\boldsymbol{\theta}$, and θ_1^* and θ_2^* in a vector $\boldsymbol{\theta^*}$ we can write

$$\boldsymbol{\theta^* = P\theta} \tag{1.4}$$

(compare Chaudhuri and Mukkerjee, 1988; van den Hout and van der Heijden, 2002). Many randomized response designs can be written in this way, and this formulation in (1.4) can also be extended to situations where the number of categories is larger than two. Further background and more complex randomized response designs can be found in Chaudhuri and Mukerjee (1988) and Fox and Tracy (1986).

1.2.1 Estimation

Estimation of $\boldsymbol{\theta^*}$ and $\boldsymbol{\theta}$ is discussed in detail in van den Hout and van der Heijden (2002) and we summarize their results here. If we want to estimate the probabilities for the true status $\boldsymbol{\theta}$ in (1.4), we plug in the observed proportions $\boldsymbol{p^*}$ for $\boldsymbol{\theta^*}$, and an estimate for $\boldsymbol{\theta}$ is obtained by

$$\boldsymbol{\hat{\theta} = P^{-1}p^*}. \tag{1.5}$$

Thus $\boldsymbol{\hat{\theta}}$ is a moment estimator. If the elements of $\boldsymbol{\hat{\theta}}$ are between zero and 1, then $\boldsymbol{\hat{\theta}}$ is also a maximum likelihood estimator. Note that (1.1) shows that if p_1^* is below chance level, the moment estimator $\hat{\theta}_1$ is negative. In SDC applications this can happen only because of sampling fluctuation. In data

collection applications this can also happen when respondents do not follow the instructions for the randomized response design, and we come back to this issue in section 1.4. If a moment estimate is negative, the maximum likelihood estimate will be on the boundary of the parameter space (i.e., the estimate is zero). For binary variables, maximum likelihood estimation is trivial as the estimate on the boundary of zero will place the other estimate on the boundary of one. For variables with more than two levels, an EM algorithm could be used to find the maximum likelihood estimates (compare van den Hout and van der Heijden, 2002). In later sections we discuss this boundary problem and solutions in more detail.

1.3 BIVARIATE ANALYSIS AND EXTENSIONS

Bivariate analyses can be carried out by generalizing equation (1.4). Two types of extensions have to be taken care of: first, the matrix $\boldsymbol{P}$ in (1.3) has to be adjusted so that it can handle bivariate problems; second, bivariate problems allow for the definition of restrictive models for $\boldsymbol{\theta}$.

First we consider three types of extensions of equation (1.3). These are

1. the situation that both variables are randomized response variables and can be considered as misclassified;
2. the situation that the dependent variable is misclassified, a condition we encounter in the data collection approach of randomized response as well as in applications where multiple data sources are integrated, and
3. the situation that the explanatory variable is misclassified, a result we encounter in the SDC approach of a single data source.

We consider the situation that both variables are binary (but this can be easily generalized to a situation with two polytomous variables). Thus we can collect the four probabilities referring to the true status into a vector $\boldsymbol{\theta}$ and the four probabilities referring to the observed status into a vector $\boldsymbol{\theta}^*$. In all three situations we generalize (1.3) and (1.4).

We first generalize (1.3). Let variable A and variable B be two variables that are possibly misclassified. For variables A and B misclassification probabilities are p_{ij}^A and p_{ij}^B respectively, defined as in (1.2), and these misclassification probabilities are collected in misclassification matrices $\boldsymbol{P_A}$ for variable A and $\boldsymbol{P_B}$ for variable B. Thus variables A and B are misclassified independently.

The joint misclassification probabilities can now be collected in a 4×4 transition matrix $\boldsymbol{P_{AB}}$ that can be constructed using a Kronecker product:

$$\boldsymbol{P_{AB}} = \boldsymbol{P_A} \otimes \boldsymbol{P_B} = \begin{pmatrix} p_{11}^A \boldsymbol{P_B} & p_{12}^A \boldsymbol{P_B} \\ p_{21}^A \boldsymbol{P_B} & p_{22}^A \boldsymbol{P_B} \end{pmatrix}. \tag{1.6}$$

We consider again the three situations 1, 2 and 3. Situation 1, where both variables are misclassified, yields simply the general form of (1.6). For situations 2 and 3, let variable A be the explanatory variable and variable B be the response variable. In situation 2, the response variable A will be misclassified by randomized response. In (1.6) $\boldsymbol{P_{AB}}$ simplifies since $\boldsymbol{P_B} = \boldsymbol{I}$. In situation 3, the explanatory variable B will be misclassified by randomized response, so $\boldsymbol{P_{AB}}$ simplifies since $\boldsymbol{P_A} = \boldsymbol{I}$.

If we collect the bivariate probabilities for the true status in a 4-vector θ and the bivariate probabilities for the observed status in a 4-vector $\boldsymbol{\theta^*}$, then $\boldsymbol{\theta}$ and $\boldsymbol{\theta^*}$ are related by

$$\boldsymbol{\theta^*} = \boldsymbol{P_{AB}\theta}. \tag{1.7}$$

Two interesting statistical models for the bivariate probabilities for the true status $\boldsymbol{\theta}$ are, first, the model where variables A and B are dependent and the elements in $\boldsymbol{\theta}$ are free, and the model where variables A and B are independent

$$\theta_{ij}^{AB} = \theta_i^A \theta_j^B. \tag{1.8}$$

We note that in both situations when the explanatory variable or the response variable is misclassified by randomized response, (1.8) refers to the situation of no subgroup differences in the probabilities of the response variable (for example, the prevalence of the sensitive behavior).

1.3.1 Estimation

This section considers the dependence and the independence models. We will discuss two methods for estimating these models, one by setting up the likelihood and maximizing it, and a second one that is based on analyzing the observed data directly.

Summarizing the literature, Van den Hout and van der Heijden (2004) describe two ways to find maximum likelihood estimates for the dependence and independence models, one is through the EM algorithm and one is by maximizing the likelihood directly. The idea of maximizing the likelihood directly is that (i) the likelihood is set up in terms of the observed responses and the probabilities for the observed status $\boldsymbol{\theta^*}$; (ii) for the dependence model we replace $\boldsymbol{\theta^*}$ by $\boldsymbol{P_{AB}\theta}$ and maximize the likelihood over $\boldsymbol{\theta}$; and (iii) for the independence model we replace $\boldsymbol{\theta^*}$ by $\boldsymbol{P_{AB}\theta}$ with elements of $\boldsymbol{\theta}$ equal $\theta_{ij}^{AB} = \theta_i^A \theta_j^B$ and maximize the likelihood over θ_i^A and θ_j^B; (iv) by comparing the likelihoods of the two models a likelihood ratio chi-square test can be carried out testing the null-hypothesis of independence against the alternative hypothesis of dependence (see also Maddala, 1983).

More conventional methods analyze the counts for the observed status directly - ignoring the fact that one or more variables are misclassified by randomized response. It may perhaps come as a surprise that a conventional analysis of the observed counts cross-classifying variables A and B gives

a correct answer about the cross-classified status of true variables. Thus, if a chi-square test is carried out testing for independence in the 2x2 table of observed responses, this test provides the correct answer about independence in the 2x2 table of the (unobserved) true responses. If only one of the two variables is misclassified, then parameter estimates for θ_i^A and θ_j^B can be found using univariate methods. If both variables are misclassified, the joint probabilities can be found in a slightly more complicated but similar way (see Fox and Tracy, 1986, p. 52).

We illustrate with an example taken from van den Hout and van der Heijden (2004), where a randomized response design is used that employs red and black cards. In Table 1.1. (a) two variables are cross-classified that were collected in as study on compliance with social benefit regulations (Van Gils, et al. 2001). The variable G denotes gender. The observed red/black answers to the RR

Table 1.1. (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

question are denoted by F^*. The question is whether or not 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.

First we use the conventional method. Applying the χ^2-test to the observed values in Table 1.1. (a) yields $X^2 = 3.377$ with 1 degree of freedom and p-value 0.066. When we choose significance level $\alpha = 0.05$, the data do not support rejecting the null hypothesis.

We now show that maximizing the likelihood 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 1.1. (a). We first construct $\boldsymbol{P_{GF}}$. Since gender (G) is not misclassified, we obtain

$$\boldsymbol{P_{GF}} = \begin{pmatrix} 8/10 & 2/10 & 0 & 0 \\ 2/10 & 8/10 & 0 & 0 \\ 0 & 0 & 8/10 & 2/10 \\ 0 & 0 & 2/10 & 8/10 \end{pmatrix}. \tag{1.9}$$

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

$$\boldsymbol{n} = \boldsymbol{P_{GF}^{-1}} \boldsymbol{n}^*, \tag{1.10}$$

see Table 1.1. (b). (We note that all counts are positive, so the moment estimate is equal to the maximum likelihood estimate in this case. If one or more of the counts would have been negative, we would have to maximize the likelihood equation through the EM algorithm or directly.) Next, we estimate the expected frequencies in this table, denoted by $\hat{\boldsymbol{m}} = \hat{m}_{11}, \hat{m}_{12}, \hat{m}_{21}, \hat{m}_{22}$, under the model of independence by $\hat{m}_{ij} = \hat{n}_{i+}\hat{n}_{+j}/N$. Since we want to fit the model of independence, we compute the fitted frequencies under this model, denoted by $\hat{\boldsymbol{m}}^*$, by

$$\hat{\boldsymbol{m}}^* = \boldsymbol{P_{GF}} \hat{\boldsymbol{m}} \tag{1.11}$$

and compare them with the observed $\boldsymbol{n}^*$. Again we get $X^2 = 3.377$. Thus, this example illustrates that prevalence estimates can be obtained either from Table 1.1. (a) by applying univariate estimation methods for males and females separately, or directly from Table 1.1. (b).

1.3.2 Extensions to multivariate analyses

The approach to maximize the likelihood directly, presented in the former paragraph, is also useful in other situations. Basically the framework can be summarized by the following steps:

1. setting up the likelihood in terms of the observed responses and the probabilities for the observed status $\boldsymbol{\theta}^*$;
2. replacing the probabilities for observed status $\boldsymbol{\theta}^*$ by the product $\boldsymbol{P\theta}$, and
3. defining a (restrictive) model for $\boldsymbol{\theta}$.

For step 3, many models can be chosen and thus the framework is very powerful. We give a few examples from the literature. Van den Hout and van der Heijden (2004) generalize the results for the dependence and the independence model to loglinear models in general. In this situation, the P-matrix accommodates possibly more than two variables. Böckenholt and van der Heijden (2004,

2007) and Fox (2005) use this framework to estimate the Rasch model (Rasch, 1980) designed for measuring individual differences with items that form a psychometric test.

An early application of this approach can be found in Maddala (1983) for logistic regression in the data collection approach of randomized response. Here the response variable is misclassified and the framework can be applied in a straightforward way (compare also Scheers and Dayton, 1988; Lensvelt et al., 2006). Van den Hout, van der Heijden and Gilchrist (2007) also apply the model to multivariate logistic regression models where the response variables are all misclassified by using the data collection approach of randomized response. On the other hand, Van den Hout and Kooiman (2005) generalize the linear regression model to the situation where one of the explanatory variables is misclassified using randomized response. This latter class of models may be more useful in the SDC approach of a single data source.

1.4 NONCOMPLIANCE IN DATA COLLECTION APPROACH

Since in the SDC approach of randomized response the misclassification process is computerized, we know the misclassification probabilities collected in $\boldsymbol{P}$. This is not necessarily true in the data collection approach of randomized response, where people may not follow the rules indicated by the randomized response design. If this problem is present in the data but ignored in the analysis, prevalence estimates are biased downwards with the possible consequence that one or more elements of $\hat{\boldsymbol{\theta}}$ fall on the boundary of the parameter space.

Recently, Böckenholt and van der Heijden (2007) and Cruyff et al. (2007) tackled this problem by defining models for multivariate randomized response data that allow for a specific form of noncompliance of respondents. This specific form of noncompliance is that a respondent either follows the rules indicated by the randomized response design or answers negatively to every sensitive question. Let λ be the probability to follow the rules of the randomized response design. Then

$$\boldsymbol{\theta}^* \quad = \quad \lambda \boldsymbol{P}\boldsymbol{\theta} + (1-\lambda)\boldsymbol{v}, \tag{1.12}$$

where $\boldsymbol{v}$ is the D-dimensional vector $(0,\ldots,0,1)^t$. Clearly, if the model specified for $\boldsymbol{\theta}$ has as many parameters as $\boldsymbol{\theta}$ has elements, then (1.12) is not identified (see Cruyff et al (2007), for an illustration). Therefore restrictive models are needed for $\boldsymbol{\theta}$ in order to arrive at an identified model. Böckenholt and van der Heijden (2007) use the Rasch model, a restrictive latent variable model well-known in psychometrics. Cruyff at al. (2007) use constrained loglinear models.

Interestingly, as was noted by Cruyff et al. (2007), the model can also be rewritten as

$$\boldsymbol{\theta}^* = \boldsymbol{Q}\boldsymbol{\theta} \tag{1.13}$$

where the transition matrix $\boldsymbol{Q}$ has elements

$$q_{ij} = \begin{cases} (\lambda)p_{ij} & \text{for } i \neq D, j \in \{1, \ldots, D\} \\ \\ (\lambda)p_{ij} + (1-\lambda) & \text{for } i = D, j \in \{1, \ldots, D\} \end{cases} \tag{1.14}$$

This shows that the framework that we presented in section 3, using $\boldsymbol{\theta}^*$, $\boldsymbol{P}$ and $\boldsymbol{\theta}$, can now be replaced by the following steps using $\boldsymbol{\theta}^*$, $\boldsymbol{Q}$ and $\boldsymbol{\theta}$:

1. setting up the likelihood in terms of the observed responses and the probabilities for the observed status $\boldsymbol{\theta}^*$;
2. replacing the probabilities for observed status $\boldsymbol{\theta}^*$ by the product $\boldsymbol{Q\theta}$, and
3. defining a (restrictive) model for $\boldsymbol{\theta}$.

Even though the form of noncompliance that a respondent is allowed to exhibit is limited in comparison to the types of noncompliance one can theoretically envisage, our experience with these models is that in practical situations, models based on (1.12) fit the data rather well. More sophisticated models for noncompliance (i.e., involving the matrix $\boldsymbol{Q}$) can be found in Böckenholt, Barlas and van der Heijden (2007), where also more sophisticated psychometric models for $\boldsymbol{\theta}$ can be found. The models defined in Bockenholt and van der Heijden (2007) also allow for covariates, both for the parameter λ as well as for the latent variable.

In the framework just presented, estimates of $\boldsymbol{\theta}$ yield unbiased estimates of prevalence of the sensitive behavior for those respondents that answer the sensitive questions following the rules laid out by the randomized response design. Univariate estimates for a variable are obtained by adding up elements in $\boldsymbol{\theta}$ over the other variables. An example from Böckenholt and van der Heijden (2007) is shown in Table 1.2.. In a survey on social benefit fraud six randomized response questions were posed to the respondents concerning violation of health regulation and work regulations. The table illustrates the importance of taking into account that some respondents do not follow the rules dictated by the randomized response design. The estimated probability of not following the design is 17 % for the work items and 13 % for the health items. When these percentages are taken into account the prevalence for the work items, for example, increases from .030 and .110 to .074 and .159, respectively.

Estimating $\boldsymbol{\theta}$ also allows for studying the probabilities that relate to the number of sensitive characteristics that are displayed (i.e., for three items this can be zero, one, two or three). This is illustrated in Table 1.3.. Here we find that - corrected for the respondents that do not follow the randomized response design - 71 percent of the respondents follows both the work as well as health regulations.

Table 1.2. Non–Compliance Estimates and 95% Bootstrap Confidence Intervals

Domain	Items	No Bias Correction	Bias Correction
Health	1	.002 (.000, .015)	.033 (.010, .050)
	2	.014 (.000, .034)	.053 (.033, .075)
	3	.048 (.027, .070)	.083 (.055, .112)
	4	.085 (.063, .107)	.130 (.087, .159)
Work	1	.030 (.009, .052)	.074 (.050, .096)
	2	.110 (.086, .133)	.159 (.104, .190)

Table 1.3. Estimated Compliance Percentages for Health and Work Domains

Counts	Work			Total
Health	0	1	2	
0	71	7	3	81
1	7	2	1	11
2	3	1	1	5
3	1	1	1	2
4	0	0	0	1
Total	82	11	6	100

1.5 CLASSIFICATION TREES

Frequently, e-commerce data are analyzed with exploratory data mining tools such as classification trees. We now discuss classification trees in the context of randomized response data. To our knowledge, this methodology has not yet been considered in this context.

Classification trees provide a breakdown of a population in sub-populations. In our situation, where the response variable is a sensitive characteristic, the aim is to find sub-populations that differ as much as possible in terms of their prevalence in the sensitive behavior. A typical result could be this: let there be an overall prevalence of .20; if the explanatory variables (in this context also called

splitting variables) are age and gender, the first split is between males and females (with a respective prevalence of .25 and .11, say); and a second split is only found for the males, namely males younger than 22 years with a prevalence of .18 and males older then 22 years with a prevalence of .28.

We discuss two applications of randomized response when the response variable is randomized, and when the explanatory variables are randomized. We first note that the estimation problem of a classification tree differs from the estimation problems discussed thus far because for classification trees there is no overall likelihood that is to be maximized, but instead a subproblem has to be solved for each separate split in the classification tree. That is, if we consider the example above, the first sub-problem is determining the splitting variable used for the first split. Specifically, we need to determine whether there is a split on gender or a split on age. Then, once the first split is determined, for every node separately (here: males separately and females separately) again a split has to be determined.

One possible criterion to determine the splitting variable is the chi-square test. If the p-value associated with the chi-square for gender is smaller than the p-value of the chi-square for age, then the variable gender will be used for the first split. Thus, the correct tree can be found as follows when the dependent variable is a randomized response variable by:

1. determining the classification tree on the observed (i.e. randomized) responses;

2. transforming the prevalence estimates found for each subpopulation using equation (1.5).

We now present an informal proof that this approach leads to the correct classification tree when the dependent variable is randomized. For this result to hold, it is only necessary to show that the order of the splitting variables for the observed data is identical to the order of the splitting variables for the true data. In order to show this, we use the results from estimation section 1.3.1 for bivariate problems. There we illustrated that there is a one-to-one correspondence between, on the one hand, a chi-square test on the frequencies in an observed 2x2 table where one variable is measured with randomized response and the other variable has two subgroups, and, on the other hand a chi-square test on the frequencies in the corresponding true 2x2 table. It follows immediately that, if the choice between splitting variables in a classification tree is made using the highest possible value of the chi-square test, then the order of chi-squares for the different splitting variables derived for the observed table is identical to the order of chi-squares for the different splitting variables derived for the observed table, which concludes the proof.

Table (1.4.) provides an illustration. On the left of this table we find the true answers and on the right the observed (i.e. randomized) answers. At the top we find frequencies for the $2 \times 2 \times 2$ table of Gender by Age by Answer on the sensitive question. In this $2 \times 2 \times 2$ table the question is whether the first splitting variable will be Gender or Age. For the true answers the chi-square for gender is

5.694 and the chi-square for age is 1.007, showing that the gender chi-square is 5.694/1.007 = 5.656 times as large. For the observed answers the chi-square for gender is 1.137 and the chi-square for age is .201, showing that the gender chi-square is also (1.137/.201 =) 5.656 times as large. This illustrates that the relative order of p-values for different splitting variables in the true table will be identical to the relative order of p-values for different splitting variables in the observed table.

Table 1.4. (a) Gender by Age by True Answer and (b) Gender by Age by Randomized Answer. . Pearson chi-square for marginal gender table in (a) is 5.694 and in (b) 1.137; Pearson chi-square for marginal age table in (a) is 1.007 and in (b) .201

(a)

2x2x2	answer		
	yes	no	total
male,young	2	100	102
male, old	5	60	65
female, young	5	30	35
female, old	2	15	17
total	14	205	219

gender	answer		
	yes	no	total
male	7	160	167
female	7	45	52
total	14	205	219

age	answer		
	yes	no	total
young	7	130	137
old	7	75	82
total	14	205	219

(b)

2x2x2	answer		
	yes	no	total
male, young	18.5	83.5	102
male, old	14.6	50.4	65
female, young	9.6	25.4	35
female, old	4.3	12.7	17
total	47.0	172.0	219

gender	answer		
	yes	no	total
male	33.1	133.9	167
female	13.9	38.1	52
total	47.0	172.0	219

age	answer		
	yes	no	total
young	28.1	108.9	137
old	18.9	63.1	82
total	14	205	219

This result holds when the response variable is a randomized variable. However, when the explanatory variables are randomized response, the results are more complicated. One of the reasons is that some of the explanatory variables are not necessarily randomized, and as a result the chi-square tests using these explanatory variables will have more power than the explanatory variables that are randomized. Therefore, in this situation it seems preferable to transform first the observed proportions $\boldsymbol{p}^*$ into $\boldsymbol{p}$ using $\boldsymbol{p} = \boldsymbol{P}^{-1}\boldsymbol{p}^*$, and then to construct a classification tree on the elements of $\boldsymbol{p}$.

1.6 DISCUSSION

Privacy protection is a crucial objective for both data collection and statistical analyses in e-commerce work (Fienberg, 2006). The randomized response methodology can make major contributions to this objective. As we have shown in this chapter, much statistical machinery is already available to address directly problems of privacy protection in analyses of publicly available databases, and "privacy-preserving statistical databases", where the data are altered prior to delivery for data mining. Moreover, randomized response methods can also be beneficially employed for collecting information about sensitive topics in e-commerce. Although some arguments could be raised against the use of randomized response for data collection purposes, there are counterarguments that support the use of this methodology as we show below:

1. It could be questioned whether the randomized response method is toot difficult to be understood by respondents , especially when it is used in an internet survey. Yet, we have good experience with this data collection method, see Lensvelt et al. (2006) for the instruction that we offer respondents to explain randomized response to them, and also see our website www.randomizedresponse.nl. It took considerable fine-tuning to arrive at these instructions and we believe that they can be used beneficially as a starting point for anyone who want to apply randomized response. In general, we emphasize that any randomized response instructions should be tested carefully using cognitive survey lab methods (see Boeije and Lensvelt, 2002, for an example).

2. Related to this, it could be questioned whether respondents are willing to follow randomized response instructions. However, recent developments described in this chapter allow taking care of respondents who do not follow the instructions, making the randomized response methodology more suited to real life applications.

3. A drawback of randomized response is that it reduces the power in statistical analyses. Clearly, misclassification has the effect that associations between variables in the populations are harder

to detect. Yet, we believe that, because web–based data collection is becoming increasingly cheaper, this is not a serious drawback for the effective application of randomized response.

In conclusion, the randomized response framework has much to offer to protect the privacy rights of individuals both during and after the collection of data. Although the framework scores high on usability and transparency criteria, much work remains to be done to make this methodology a mainstream and routine component of statistical work. The practical importance and urgency of this work cannot be underestimated, especially in view of the mounting tensions between confidentiality and the ever increasing e-commerce data availability.

REFERENCES

1. Ambainis, A., M. Jakobsson and H Lipmaa (2003), Cryptographic Randomized Response Techniques. *Cryptology ePrint Archive, Report 2003/027*, http://eprint.iacr.org/.

2. Böckenholt, U. , S. Barlas & P.G.M. van der Heijden (2007) Do randomized-response designs eliminate response biases? An empirical study of non-compliance. *Journal of Applied Econometrics*.

3. Böckenholt, U. & van der Heijden, P. G. M. (2004). Measuring noncompliance in insurance benefit regulations with randomized response methods for multiple items. In: A. Biggeri, E. Dreassi, C. Lagazio and M. Marchi, In (Eds.), *Proceedings of 19th International Workshop on Statistical Modelling*. Florence, Italy, pp. 106-110.

4. Böckenholt, U. & 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.

5. Boeije, Henny and Gerty Lensvelt-Mulders. 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.

6. Boruch, Robert F. 1971. "Assuring Confidentiality of Responses in Social Research: a Note on Strategies", *The American Sociologist* 6:308-311.

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

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

9. Cruyff, M., van den Hout, A., van der Heijden, P.G.M. , & Böckenholt, U. (2007). Loglinear randomized response models taking cheating into account. *Sociological Methods and Research*.

10. Du, W. and Z. Zhan (2003). Using randomized response techniques for privacy-preserving data mining. *SIGKKDD*.

11. Evfimievski, A (2002). Randomization in Privacy-Preserving Data Mining. *SIGKDD Explorations: Newsletter of the ACM Special Interest Group on Knowledge Discovery and Data Mining, 4(2)*, 43-48.

12. Fienberg, S.E. (2005). Confidentiality and disclosure limitation. *Encyclopedia of Social Measurement*, 463-469. Amsterdam: North Holland.

13. Fienberg, S.E. (2006). Privacy and confidentiality in an e-commerce world: data mining, data warehousing, matching and disclosure limitation. *Statistical Science, 21*, 143-154.

14. Fox, J.-P. (2005). Randomized item response theory models. *Journal of Educational and Behavioral Statistics*, 30, 1-24.

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

16. 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, *J. Off. Stat.*, **14**, pp 463-478.

17. Karr, A.F. and S.E. Fienberg (2005). *Data confidentiality, data quality and data integration for federal databases: foundations to software prototypes.* Digital Government II Home Page: $http://www.niss.org/dgii/index.html$.

18. 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.

19. Kuha, J., and Skinner, C. (1997) Categorical Data Analysis and Misclassification, *Survey Measurement and Process Quality*, Ed. L. Lyberg et al., New York: Wiley.

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

21. Lensvelt, G.J.L.M., Hox, J.J., Van der Heijden, P.G.M., and Maas, C. (2005). Meta-analysis of Randomized Response Research: 35 Years of Validation. *Sociological Methods and Research, 33*, , 319–348.

22. Lensvelt-Mulders, G.J.L.M., van der Heijden, P.G.M., Laudy, O., & van Gils, G. (2006). A validation of a computer assisted randomized response survey for measuring fraud in social security. *Journal of the Royal Statistical Society, Series A, vol. 169*, 305-318.

23. Maddala, G.S. (1983), Limited dependent and qualitative variables in econometrics. Cambridge University Press, New York.

24. Rasch, G. (1980). *Probabilistic Models for Some Intelligence and Attainment Tests*. Chicago: The University of Chicago Press. (Original published 1960, Copenhagen: The Danish Institute of Educational Research)

25. Rosenberg, M.J. (1979). *Multivariate Analysis by a Randomized Response Technique for Statistical Disclosure Control*, Ph.D. Dissertation, University of Michigan.

26. Scheers, N.J. and Dayton, C.M. (1988). Covariate randomized response models. *Journal of the American Statistical Association, 83*, 969-974.

27. Trottini, M, S.E. Fienberg, U.E. Makov and M.M. Meyer (2004). Additive noise and multiplicative bias as disclosure limitation techniques for continuous microdata: a simulation study. *Journal of Computational Methods in Sciences and Engeneering, 4*, 5-16.

28. Van den Hout, A., & E.A.H. Elsayed (2006). Statistical Disclosure Control Using Post Randomisation: Variants and Measures for Disclosure Risk. textitJournal of official statistics, 22, pp. 711–731

29. Van den Hout, A., and P. Kooiman (2005). Estimating the linear regression model with categorical covariates subject to randomized response. *Computational Statistics and Data Analysis, 50*, 3311-3323.

30. Van den Hout, A., van der Heijden, P.G.M., & Gilchrist, R. (2007). The logistic regression model with response variables subject to randomized response. *Computational Statistics and Data Analysis*.

31. Van den Hout, A., & van der Heijden, P. G. M. (2002). Randomized response, statistical disclosure control and misclassification: A review. *International Statistical Review*, 70, 269-288.

32. Van den Hout, A., & 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.

33. Van Gils, Ger, Peter G.M van der Heijden, and Annemarie Rosebeek. 2001. *Onderzoek naar regelovertreding, Resultaten ABW, WAO en WW*. Amsterdam: NIPO. (In Dutch)

34. Warner, S.L. (1965). Randomized response: A survey technique for eliminating answer bias, *Journal of the American Statistical Association, 60*, 63-69.

35. Warner, S.L. (1971). The linear randomized response model. *Journal of the American Statistical Association, 66*, 884-888.