Department for Applied Statistics
Johannes Kepler University Linz

IFAS Research Paper Series
2009-44

A Masking Scheme for Statistical Disclosure Control

Andreas Quatember

August 2009

Abstract
In official statistics statistical diclosure control balances between the laws of data protection and the wish of analysts to have access to the original survey data. Applying Quatember's (2009) standardized randomized response questioning design in this context provides a masking scheme, which can be used for different levels of data protection. In this paper the questioning design is translated into a masking scheme. The instruction for the adequate calculation of the masking parameters in accordance to the wanted level of data protection is given. Furthermore the statistical properties of the estimation of one-dimensional proportions and a manual for the reconstrustion of (2×2)-tables, when one or both variables are masked, are presented.
KEY WORDS: Statistical disclosure control; estimation of proportions; standardized randomized response technique

1 Introduction

There is a continuously growing demand from empirical researchers for access to original survey data. Due to the aspects of data confidentiality and privacy protection of the respondents the release of such microdata files is constrained by law. Therefore statistical disclosure control has to balance between the rights of the survey units and the reasonable preservation of information (cf. Fuller 1993). According for instance to Domingo-Ferrer and Mateo-Sanz (2002) the methods of masking to reduce the risk of reidentification of survey units can be classified into three three categories: global recoding, local suppression and substitution of data (cf. for instance: Willenborg and de Waal 1996). The ideas of the randomized response questioning designs – originally proposed by Warner (1965) as instrument to reduce item-nonresponse and untruthful answering for variables of a highly personal matter – are also applicable in this context (cf. Warner 1971 and for instance: Gouweleeuw et al. 1998 or Katzoff and Kim 2006)).

Quatember (2009) presented a standardization of randomized response techniques for the estimation of the relative size of a certain subpopulation. Let y be a dichotomous (1/0)-variable under study and U be a population of size N. Let furthermore be $U_{y=1}$ be the subset of size $N_{y=1}$, for which $y = 1$ applies. The parameter of interest be the relative size $\pi_{y=1} = \frac{N_{y=1}}{N}$ of $U_{y=1}$. Moreover let $U_{x=1}$ be another subgroup of U of relative size $\pi_{x=1} = \frac{N_{x=1}}{N}$, for which $x = 1$ of another dichotomous variable x applies, which is not related to y (see: Greenberg et al. 1969). Moreover let $U_{y=0} = U - U_{y=1}$ be the subgroup of size $N_{y=0} = N - N_{y=1}$).

Applying the standardized randomized questioning design a respondent has to answer randomly

- with probability p_1 the question "Are you a member of group $U_{y=1}$?",
- the question "Are you a member of group $U_{y=0}$?" with probability p_2 or
- with probability p_3 the question "Are you a member of group $U_{x=1}$?"

or is instructed just to say

- "yes" with probability p_4 or
- "no" with probability p_5

($\sum_{i=1}^{5} p_i = 1$, $0 \leq p_i \leq 1$ for $i = 1, 2, ..., 5$). $\pi_{x=1}$ and the probabilities p_1 to p_5 are the freely chooseable design parameters of the standardized randomized response questioning design. In a probability sample s of size $n \leq N$ the auxiliary variable z with

$$z_k = \begin{cases} 1 & \text{if unit } k \text{ answers "yes"}, \\ 0 & \text{otherwise} \end{cases}$$

is observed. The probability of a "yes"-answer given y is:

$$P(z_k = 1) = p_1 \cdot y_k + p_2 \cdot (1 - y_k) + p_3 \cdot \pi_{x=1} + p_4. \tag{1}$$

This yields the following unbiased "randomized response estimator" $\widehat{\pi}_{y=1}^{RR}$ of parameter $\pi_{y=1}$ with the given design weights d_k of a probability sampling design P, which are the reciproc values of the sample inclusion probabilities π_k ($k = 1, 2, ..., n$):

$$\widehat{\pi}_{y=1}^{RR} = \frac{1}{N} \cdot \sum\nolimits_s \frac{z_k - (p_2 + p_3 \cdot \pi_{x=1} + p_4)}{p_1 - p_2} \cdot d_k \tag{2}$$

($\sum_s$ is the abbreviation for $\sum_{k=1}^{n}$). The variance of this standardized estimator $\widehat{\pi}_{y=1}$ (2) is given by

$$\begin{aligned} V_P(\widehat{\pi}_{y=1}^{RR}) \;=\; & \frac{1}{N^2} \cdot \left(V_P\left(\sum\nolimits_s y_k \cdot d_k \right) + \right. \\ & + \frac{(p_2 + p_3 \cdot \pi_{x=1} + p_4) \cdot (1 - (p_2 + p_3 \cdot \pi_{x=1} + p_4))}{(p_1 - p_2)^2} \cdot \sum\nolimits_U d_k + \\ & \left. + \frac{1 - 2 \cdot (p_2 + p_3 \cdot \pi_{x=1} + p_4) - (p_1 - p_2)}{p_1 - p_2} \cdot \sum\nolimits_U y_k \cdot d_k \right). \end{aligned} \tag{3}$$

(Quatember 2009). $V_P(\sum_s y_k \cdot d_k)$ refers to the variance of the Horvitz-Thompson estimator $\sum_s y_k \cdot d_k)$ for the total $\sum_U y_k$ for a given probability sampling design P (see for instance: Srndal et al. 1992). (3) can be estimated unbiasedly by inserting an unbiased estimator $\widehat{V}_P(\sum_s y_k \cdot d_k)$ for $V_P(\sum_s y_k \cdot d_k)$ and $\sum_s (\frac{z_k - (p_2 + p_3 \cdot \pi_{x=1} + p_4)}{p_1 - p_2} \cdot d_k^2)$ for $\sum_U y_k \cdot d_k$.

2 The Masking Scheme

A translation of the simple probability mechanism behind the standardized questioning design sot hat it can be used as a masking scheme for dichotomous variables, which can be applied on the data after the data collection and before their release, can be done in the following way: Let y be the variable under study and z now be the publishable masked variable. Let

$$z_k | (y_k = 1) = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } 1 - p \end{cases}$$

and

$$z_k|(y_k = 0) = \begin{cases} 1 & \text{with probability } 1 - q \\ 0 & \text{with probability } q \end{cases}$$

The probabilities p and q are the *masking parameters* of this masking scheme ($0 \leq p \leq 1$, $0 \leq q \leq 1$). So as to be able to reasonably fix the values of these parameters we have to include the wanted level of data privacy protection in our considerations. The following ratios λ_1 and λ_0 of conditional probabilities based on the "Leysieffer-Warner-measures of jeopardy" (Leysieffer and Warner 1976) can be objective measures of the loss of data privacy of survey units induced by the masking parameters:

$$\lambda_{i,k} = \frac{max[P(z_k = i|y_k = 1), P(z_k = i|y_k = 0)]}{min[P(z_k = i|y_k = 1), P(z_k = i|y_k = 0)]} \tag{4}$$

($1 \leq \lambda_{i,k} \leq \infty$; $i = 1, 0$, $k \in U$). For $i = 1$ (4) refers to the data protection with respect to $z_k = 1$, for $i = 0$ with respect to $z_k = 0$. For our masking scheme these loss-measures are given by

$$\lambda_{1,k} = \lambda_1 = \frac{max[p; 1 - q]}{min[p; 1 - q]} \tag{5}$$

and

$$\lambda_{0,k} = \lambda_0 = \frac{max[1 - p; q]}{min[1 - p; q]}. \tag{6}$$

($\forall\ k \in U$). $\lambda_1 = \lambda_0 = 1$ indicates a totally protected data privacy. This means that the published variable z contains absolutely no information on y. But the more the λ-measures differ from unity the more information about y is contained in z and the less the individual's data are protected against the data collector. When $z = y$ these measures are given by $\lambda_1 = \lambda_0 = \infty$.

A statistical agency might fix those values of λ_1 and λ_0, which allow enough disclosure control. Without loss of generality let us assume subsequently, that we will choose the two categories of the variable under study in such way, that $y_k = 1$ is at least as worthy of protection as $y_k = 0$ ($p \geq 1 - q$, $1 \leq \lambda_1 \leq \lambda_0 \leq \infty$). From (5) and (6) the masking parameters p and q can be expressed by λ_1 and λ_0:

$$p = \frac{\lambda_1 \cdot \lambda_0 - \lambda_1}{\lambda_1 \cdot \lambda_0 - 1} \tag{7}$$

and

$$q = \frac{\lambda_1 \cdot \lambda_0 - \lambda_0}{\lambda_1 \cdot \lambda_0 - 1}. \tag{8}$$

We have to distinguish between different types of sensitivity of the variable with respect to data protection: For a nonsensitive variable, where $\lambda_1 = \lambda_0 = \infty$ applies, it is easy to see that p and q are equal to 1. For a variable, of which only $y_k = 1$, but not $y_k = 0$ is sensitive $\lambda_1 < \lambda_0 = \infty$ applies. This yields $p = 1$ and $q = \frac{\lambda_1 - 1}{\lambda_1}$. If both $y_k = 1$ and $y_k = 0$ are sensitive, $\lambda_1 \leq \lambda_0 < \infty$ applies and the masking parameters can be calculated directly from (7) and (8).

3 The Statistical Properties

For the publishable masked variable z the probability of $z_k = 1$ is given by

$$P(z_k = 1) = p \cdot y_k + (1-q) \cdot (1-y_k) = (p-(1-q)) \cdot y_k + 1 - q \tag{9}$$

Therefore the following theorem applies:

Theorem: For a probability sampling design P with design weight d_k

$$\widehat{\pi}_{y=1} = \frac{1}{N} \cdot \sum\nolimits_s \frac{z_k - (1-q)}{p-(1-q)} \cdot d_k \tag{10}$$

$(p \neq 1-q)$ is an unbiased estimator of parameter $\pi_{y=1}$. The variance of $\widehat{\pi}_{y=1}$ is given by

$$\begin{aligned} V_P(\widehat{\pi}_{y=1}) &= \frac{1}{N^2} \cdot \left(V_P\left(\sum\nolimits_s y_k \cdot d_k\right) + \frac{q \cdot (1-q)}{(p-(1-q))^2} \cdot \sum\nolimits_U d_k + \right. \\ &\quad \left. + \frac{1 - 2 \cdot (1-q) - (p-(1-q))}{(p-(1-q))} \cdot \sum\nolimits_U y_k \cdot d_k \right). \end{aligned} \tag{11}$$

This variance is unbiasedly estimated by

$$\begin{aligned} V_P(\widehat{\pi}_A) &= \frac{1}{N^2} \cdot \left(\widehat{V}_P\left(\sum\nolimits_s y_k \cdot d_k\right) + \frac{q \cdot (1-q)}{(p-(1-q))^2} \cdot \sum\nolimits_U d_k + \right. \\ &\quad \left. + \frac{1 - 2 \cdot (1-q) - (p-(1-q))}{(p-(1-q))} \cdot \sum\nolimits_s \frac{z_k - (1-q)}{p-(1-q)} \cdot d_k^2 \right). \end{aligned} \tag{12}$$

$V_P(\sum_s y_k \cdot d_k)$ refers to the variance of the Horvitz-Thompson estimator for the total $\sum_U y_k$ for a probability sampling design P. $\widehat{V}_P(\sum_s y_k \cdot d_k)$ is an unbiased estimator of this variance. The other two summands within the outer brackets can now be seen as the price that has to be paid in terms of accuracy for data privacy protection.

For simple random sampling without replacement estimator (10) is given by

$$\widehat{\pi}_{y=1} = \frac{\sum_s z_k/n - (1-q)}{p-(1-q)}. \tag{13}$$

And for this sampling method the variance of (13) is given by

$$\begin{aligned} V(\widehat{\pi}_{y=1}) &= \frac{\pi_{y=1} \cdot (1-\pi_{y=1})}{n} \cdot \frac{N-n}{N-1} + \frac{1}{n} \cdot \left(\frac{q \cdot (1-q)}{(p-(1-q))^2} + \right. \\ &\quad \left. + \frac{1 - 2 \cdot (1-q) - (p-(1-q))}{p-(1-q)} \cdot \pi_{y=1} \right). \end{aligned} \tag{14}$$

$V(\widehat{\pi}_{y=1})$ is unbiasedly estimated by

$$\begin{aligned} \widehat{V}(\widehat{\pi}_{y=1}) &= \frac{\widehat{\pi}_{y=1} \cdot (1-\widehat{\pi}_{y=1})}{n-1} \cdot \frac{N-n}{N} + \frac{1}{n} \cdot \left(\frac{q \cdot (1-q)}{(p-(1-q))^2} + \right. \\ &\quad \left. + \frac{1 - 2 \cdot (1-q) - (p-(1-q))}{p-(1-q)} \cdot \widehat{\pi}_{y=1} \right). \end{aligned} \tag{15}$$

4 2×2-Ccontingency Tables

Often besides the estimation of one-dimensional parameters also two-dimensional contingency tables are of interest for data analysts. In the case of masking at least one of the two original variables y_1 and y_2 included, such a table should be reconstructable from the masked variables in the microdata file. The sample proportions ρ_{ij} of the combinations $y_1 = i$ and $y_2 = j$ in a simple random sample are unbiased estimators of the two-dimensional $(y_1 \times y_2)$-population proportions π_{ij} and $\rho_{y_1=i} = \sum_{j=1,0} \rho_{ij}$ and $\rho_{y_2=j} = \sum_{i=1,0} \rho_{ij}$ of their marginal proportions $\pi_{y_1=i}$ and $\pi_{y_2=j}$ respectively $(i = 1, 0$ and $j = 1, 0)$. When y_1 is masked by z_1 before the release of data in the way described above and $z_2 = y_2$, the two-dimensional $(z_1 \times y_2)$-sample proportions τ_{ij} with marginals $\tau_{z_1=i} = \sum_{j=1,0} \tau_{ij}$ and $\rho_{y_2=j} = \sum_{i=1,0} \rho_{ij}$ are observed $(i = 1, 0$ and $j = 1, 0)$. The two marginal proportions $\rho_{y_2=j}$ of variable y_2 can be observed directly from the unmasked data vector of y_2 $(j = 1, 0)$. The unknown marginal proportions $\rho_{y_1=i}$ $(i = 1, 0)$ can be estimated unbiasedly by $\widehat{\pi}_{y_1=1}$ (10).

The expectation of τ_{11} is

$$\begin{aligned} E(\tau_{11}) &= \rho_{11} \cdot p + (\rho_{y_2=1} - \rho_{11}) \cdot (1 - q) \\ &= \rho_{11} \cdot (p - (1 - q)) + \rho_{y_2=1} \cdot (1 - q). \end{aligned}$$

Then the unobserved sample proportion ρ_{11} can be reconstructed unbiasedly by

$$\widehat{\rho}_{11} = \frac{\tau_{11} - \rho_{y_2=1} \cdot (1 - q)}{p - (1 - q)}. \tag{16}$$

The other original sample proportions of $(y_1 \times y_2)$ can be unbiasedly estimated by

$$\widehat{\rho}_{10} = \widehat{\pi}_{y_1=1} - \widehat{\rho}_{11}, \tag{17}$$

$$\widehat{\rho}_{01} = \rho_{y_2=1} - \widehat{\rho}_{11} \tag{18}$$

and

$$\widehat{\rho}_{00} = 1 - \rho_{y_2=1} - \widehat{\rho}_{10}. \tag{19}$$

If the variables y_1 and y_2 are both masked by the same masking procedure we can reconstruct the original $(y_1 \times y_2)$- from the $(z_1 \times z_2)$-table by substituting the marginal proportion $\rho_{y_2=1}$ in (16), (18) and (19) by its unbiased estimator $\widehat{\pi}_{y_2=1}$ (10).

5 Summary

The presented masking scheme allows to protect data for publication at a desired level. The price to pay for publishable microdata files is one of accuracy of survey results. Two-dimensional sample-tables for the original variables can be reconstructed from the masked variables. The next step should be to allow individually differing levels of privacy protection for the survey units. This makes it possible for instance to fix different data protection levels for different strata of survey units as for example for different municipalities.

Appendix: Proof of the Theorem

With the sampling design P and the masking mechanism M we have

$$\begin{aligned} E(\widehat{\pi}_{y=1}) &= \frac{1}{N} \cdot E_P \left[E_M \left(\sum\nolimits_s \left(\frac{z_k - (1-q)}{p \cdot (1-q)} \cdot d_k \right) | s \right) \right] \\ &= \frac{1}{N} \cdot E_P \left(\sum\nolimits_s y_k \cdot d_k \right) = \frac{1}{N} \cdot \sum\nolimits_U y_k = \pi_{y=1}. \end{aligned}$$

The variance of estimator (10) is given by

$$V(\widehat{\pi}_{y=1}) = V_P(E_M(\widehat{\pi}_{y=1}|s)) + E_P(V_M(\widehat{\pi}_{y=1}|s)).$$

Then

$$V_P(E_M(\widehat{\pi}_{y=1}|s)) = \frac{1}{N^2} \cdot V_P \left(\sum\nolimits_s y_k \cdot d_k \right).$$

Let the sample inclusion indicator be

$$I_k = \begin{cases} 1 & \text{if unit } k \in s, \\ 0 & \text{otherwise.} \end{cases}$$

The covariance $C_M(\frac{z_k-(1-q)}{p\cdot(1-q)}, \frac{z_l-(1-q)}{p\cdot(1-q)}|s)$ is equal to 0 $\forall\ k \neq l$ and therefore for the second summand of $V(\widehat{\pi}_{y=1})$

$$\begin{aligned} E_P(V_M(\widehat{\pi}_{y=1}|s)) &= E_P \left[\frac{1}{N^2} \cdot V_M \left(\sum\nolimits_U I_k \cdot \frac{z_k - (1-q)}{p \cdot (1-q)} \cdot d_k | s \right) \right] \\ &= E_P \left[\frac{1}{N^2} \cdot \sum\nolimits_U I_k^2 \cdot d_k^2 \cdot V_M \left(\frac{z_k - (1-q)}{p \cdot (1-q)} \right) \right] \\ &= \frac{1}{N^2} \cdot \sum\nolimits_U V_M \left(\frac{z_k - (1-q)}{p \cdot (1-q)} \right) \cdot d_k. \end{aligned}$$

applies. We can write

$$V_M \left(\frac{z_k - (1-q)}{p \cdot (1-q)} \right) = \frac{1}{(p - (1-q))^2} \cdot V_M(z_k)$$

and because of $y_k^2 = y_k$

$$\begin{aligned} V_M(z_k) &= 1 - q + (p - (1-q)) \cdot y_k - (1 - q + (p - (1-q)) \cdot y_k)^2 \\ &= (1 - q + (p - (1-q)) \cdot y_k) \cdot (q - (p - (1-q)) \cdot y_k) \\ &= (1-q) \cdot q + (p - (1-q)) \cdot (1 - 2 \cdot (1-q) - (p - (1-q))) \cdot y_k. \end{aligned}$$

This yields

$$\begin{aligned} E_P(V_M(\widehat{\pi}_{y=1}|s)) &= \frac{1}{N^2} \cdot \left(\frac{(1-q) \cdot q}{(p - (1-q))^2} \cdot \sum\nolimits_U d_k + \right. \\ &\quad \left. + \frac{1 - 2 \cdot (1-q) - (p - (1-q))}{p - (1-q)} \cdot \sum\nolimits_U y_k \cdot d_k \right), \end{aligned}$$

which completes the proof.

References

Domingo-Ferrer, J. and Mateo-Sanz, J. M. (2002). Practical Data-Oriented Microaggregation for Statistical Disclosure Control. *IEEE Transactions on Knowledge and Data Engineering*, 14 (1), 189–201.

Fuller, W. A. (1993). Masking Procedures for Microdata Disclosure Limitation. *Journal of Official Statistics*, 9 (2), 383–406.

Gouweleeuw, J. M., Kooiman, P., Willenborg, L. C. R. J., and de Wolf, P.-P. (1998). Post Randomization for Statistical Disclosure Control: Theory and Implementation. *Journal of Official Statistics*, 14(4), 463–478.

Greenberg, B. G., Abul-Ela, A.-L. A., Simmons, W. R., and Horvitz, D. G. (1969). The Unrelated Question Randomized Response Model: Theoretical Framework. *Journal of the American Statistical Association*, 64, 520–539.

Katzoff, M. J., and Kim, J. J. (2006). Masking for Discrete Variables. *Proceedings of Statistics Canada Symposium 2006: Methodological Issues in Measuring Population Health.*

Leysieffer, F. W. and Warner, S. L. (1976). Respondent Jeopardy and Optimal Designs in Randomized Response Models. *Journal of the American Statistical Association*, 71, 649–656.

Quatember, A. (2009). A Standardization of Randomized Response Strategies. *Survey Methodology* (to appear).

Srndal, C.-E., Swensson, B., and Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer.

Warner, S. L. (1965). Randomized Response: A Survey Technique for Eliminating Evasive 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. and de Waal, T. (1996). *Statistical Disclosure Control in Practice*. New York: Springer.