Department for Applied Statistics
Johannes Kepler University Linz

IFAS Research Paper Series
2011-55

A Family of Methods for Statistical Disclosure Control

Andreas Quatember

June 2011

1 Introduction

There is a continuously increasing demand by the public for access to microdata files from surveys conducted by the official statistics- or other institutions or agencies in all kinds of fields such as education, employment, or public health. Because such data may contain sensitive information on natural or legal persons, such as information on poverty, alcoholism, tax morality, or bank rating, the release of such files is subject to the laws of data protection. Disclosure happens, if the release of data allows an intruder to connect the surveyed information to certain population units. To protect the survey units against disclosure it might not suffice to delete the variables, which are directly linked to entities, such as name, address or an artificial identification number like the social security number. Some of the units might still be identifiable by the rest of their records if the individuals own rare values of certain variables such as a very large income for instance, or other rare combinations of variables. For this reason, methods of statistical disclosure control (SDC), that make impossible linking sensitive information to individuals with certainty have to be applied before data can be handed out to the public. The purpose of these SDC methods is to manipulate variables in a way that enhances privacy protection and it is still possible to estimate the unknown parameters of interest. In the literature, several such methods are discussed (cf. for instance, Winkler 2004, or Matthews and Harel 2011).

One of these techniques artificially introduces "missings" instead of sensitive variable-values into the complete data file. In the relevant literature, this approach is called *suppression of data* (cf. for instance, Willenborg and de Waal 1996, p.77ff). After suppressing data in this way, the estimation of the parameters under study (like population or stratum means) from the available cases is as problematic as it would be in the presence of "real" nonresponse, where the missing data mechanism is not "completely at random" (cf. Little and Rubin 2002, p.12).

In Section 2 of this paper, a family of methods for SDC is defined and embedded in the existing SDC terminology, combining SDC with the ideas of data imputation from the field of analysing missing data. In Section 3, three examples of methods, all of them belonging to this "blended family" of techniques, are discussed showing how different procedures belonging to this family can be.

2 A Family of SDC Methods

A whole family of methods for SDC "masks" a sensitive or identifying variable y in a data file. These methods consist of three consecutive steps: At the first one, the C-step, an additional variable is created by "cloning" the original variable y. Then, in the S-step, the idea of data suppression is applied locally or even globally to the y-clone. This means that the values of the cloned variable are set to missing for a group of, or for all survey units. In the concluding I-step, a data imputation method is applied to these artificial missings. For this purpose, in addition to information on available auxiliary variables $\mathbf{x}$ the original sensitive variable y can also be used for the imputations in contrast to a real nonresponse case. After the missing values have been replaced by imputed values, $\widehat{y}$, variable y can be deleted. Henceforward, its masked substitute z has to serve as the basis for the estimation of the parameters of the sensitive variable y in the publishable data file.

For instance, let

$$t = \sum\nolimits_U y_k, \tag{1}$$

the total of variable y in population U be the parameter of interest. With the original variable y the unbiased Horvitz-Thompson estimator of t in any probability sample s with sample size n

$$\widehat{t}_{HT} = \sum\nolimits_s y_k \cdot d_k, \tag{2}$$

where the design weight d_k is the reciprocal of the sample inclusion probability of a survey unit k ($k = 1, 2, ..., n$). When the "cloning-suppression-imputation" (CSI) procedure described above is applied as the SDC technique, after the C-step within the S-step the sample s is partitioned into an artificial "response set" $r \subset s$ of size n_r and an artificial "missing set" $m = s - r$ of size n_m ($s = r \cup m$, $r \cap m = \emptyset$, $n = n_r + n_m$). Then, after the I-step, the estimation of t has to be done by

$$\widehat{t} = \sum\nolimits_s z_k \cdot d_k, \tag{3}$$

with

$$z_k = \begin{cases} y_k & \text{if unit } k \in r, \\ \widehat{y}_k & \text{otherwise.} \end{cases}$$

Hence, $\widehat{t}$ can be decomposed into

$$\widehat{t} = \sum\nolimits_r y_k \cdot d_k + \sum\nolimits_m \widehat{y}_k \cdot d_k. \tag{4}$$

In this masking process, both the quality of the estimation and the degree of data protection do – although inversely – depend on the imputation method applied in the I-step and the size of the missing set m. In principle, all imputation methods known from the missing data literature (see for instance, Little and Rubin 2002, p.59ff) can be used in this context. Multivariate relationships between surveyed variables may be maintained in most cases, when an efficient masking process can simultaneously be applied to different variables.

When it comes to statistical disclosure control, some variants of the described CSI method are already in use. As an example, randomly interchanging the values of the sensitive variable of two different groups of the same size is called *data swapping* (cf. Dalenius and Reiss 1982). Within the CSI framework, the basic version of this method can be described in the following way: After the cloning process, the data of two different groups are suppressed. For both groups a random imputation of data from the other group without replacement is applied at the I-step. This preserves the privacy of the units belonging to these groups. So the real data remain in the data set, but some of them assigned to other survey units. This does not affect the estimation quality of statistics for the distribution of y unless s is a "non-self-weighting sample" (cf. for instance, Lohr 2010, p.287f).

The term *micro-aggregation of data* (cf. Defays and Anwar 1998) refers to a strategy where sensitive values of a quantitative variable are – generally spoken – substituted by aggregates such as means, medians, modes, or some other measures. Within the CSI framework the imputation methods used after the suppression step are, for instance, overall or (with the help of the auxiliary variables $\mathbf{x}$ and y) class mean, median, or mode

imputation. After the micro-aggregation simple univariate statistics of y such as its total may still be calculated from z, but the estimation of the variance will understate the true variance of the estimator. Compared to overall aggregate imputation the imputation of aggregates within classes (of $\mathbf{x}$ and/or y) will surely help in increasing the quality of the estimation of such parameters on the basis of the new variable z.

The *addition of noise* (cf. Fuller 1993) is another example of a SDC procedure belonging to our family of methods. Herein, in the imputation step, random errors are added to y to create the publishable variable z. This I-step can be seen as an application of stochastic regression imputation. The estimation of univariate parameters is without a problem, if the suppression mechanism is completely at random. If it is only "at random" (cf. Little and Rubin 2022, p.12), conditional stochastic regression imputation can be applied. However, variance estimates calculated from z instead of y may be too small unless the number of suppressed values is negligible since they do not account for imputation uncertainty. Here, multiple imputation may be helpful (cf. for instance, Rubin 1987).

As another example of techniques of that family *global recoding* and *top* and/or *bottom coding* can be mentioned (see for instance, Willenborg and de Waal 2001, p.27f). Herein, the cloned data are globally or locally suppressed. The concluding I-step of the CSI procedure uses only the original variable y as auxiliary information and transforms its values, on the one hand, into large(r) intervals and, on the other hand, it limits the extreme values of y to an upper and/or lower bound. This means a loss of information (and efficiency), which does not affect the estimation of parameters when robust estimators with respect to outliers are calculated such as the median.

Rubin (1993) proposed the use of *multiple imputation* in the SDC context. When it is applied as an imputation method within the I-step of the CSI framework to replace all or a certain part of the y-clone, so called "partially synthetic datasets" are generated (cf. Drechsler et al. 2008, p.1007). All stochastic imputation methods may be applied. For the estimation of the interesting parameters the multiple imputation framework can be used (cf. Rubin 1987, p.76) with a modification for the estimation of the variance because the "nonresponse mechanism" is not stochastic in our case (cf. Reiter 2003, p.5f).

Evidently, no matter what method is applied, the recipient of the data has to pay for the survey units' privacy protection by a loss of accuracy. But when reasonable imputation procedures are used it may work better than just to suppress data.

3 Three Examples of the CSI method

In this section, three special cases of the CSI method are presented to show how different methods belonging to this family can be. The first one is carried out by the respondents rather than the agency *during* the survey, whereas the second one uses a randomization device as the imputation algorithm *after* the data collection. In the third example, hot deck imputation with replacement is used at the I-step of the masking process to reduce the respondents' disclosure risk.

3.1 Techniques of Randomized Response

Techniques of randomized response were originally presented as methods to reduce nonresponse and untruthful answering when sensitive questions such as on drug use, domes-

tic violence or tax evasion are asked in a survey. Warner (1965) published the pioneering work in this field for the estimation of the relative size of subpopulations having the sensitive attributes. Since then, various randomized response techniques with different randomization devices have been proposed for qualitative as well as quantitative variables (for a review, see for instance, Tracy and Mangat 1996; for standardization of different techniques, see, Quatember 2009). Some of them use different random devices for the question selection depending on the respondent's possession or nonpossession of a certain attribute (cf. Singh and Chen 2009).

The central element of all of these methods is that survey units do not have to answer with certainty the sensitive question but can choose the one to be answered randomly from two or more questions. This does not enable the data collector to identify the question, on which the respondents have given the answer, although the given answers do still allow estimating the univariate parameter under study. In this way, the idea is to reduce the individual's fear of an embarrassing "outing" to make sure that the responding person is willing to cooperate.

Warner (1971) was also the first to indicate that these techniques are applicable as methods of SDC applied *during* the data collection (cf. ibd., p.887). In this case, the survey units already perform the masking process on their own at the survey's design stage. At the I-step of the CSI strategy described in section 2, the value y_k of the original variable y of a sampling unit k is replaced by the respondent's answer on the randomly selected question. Within the randomization device, the original variable value y_k may serve as auxiliary information. Anyhow, the user of the microdata file has to be informed about the details of the masking and the estimation process.

3.2 A Post Randomization Method

Procedures, where a randomization mechanism is applied on a variable *after* the data collection in order to reduce the risk of disclosure, are called *post randomization methods* (see: Gouweleeuw et al. 1998). In this section, the method is applied to a dichotomous variable to estimate π_A, the relative size of a subgroup A of U ($A \subseteq U$). Its theory is extended to any probability sampling design and, additionally, it allows individually differing privacy protection levels for the sample units k.

For this purpose, in the context of the CSI method after global suppression let the imputed value $\widehat{y}_k$, conditioned on the original variable value y_k as auxiliary information, be

$$\widehat{y}_k|(y_k = 1) = \begin{cases} 1 & \text{with probability } p_k \\ 0 & \text{with probability } 1 - p_k \end{cases}$$

($0 \leq p_k \leq 1$) and

$$\widehat{y}_k|(y_k = 0) = \begin{cases} 1 & \text{with probability } 1 - q_k \\ 0 & \text{with probability } q_k \end{cases}$$

($0 \leq q_k \leq 1$). Besides information on y itself, also other auxiliary variables $\mathbf{x}$ can be incorporated in the I-step of the masking process. For instance, this can be done by assigning different probabilities p_k and/or q_k to men and women. Therefore, these conditional probabilities p_k and q_k can be seen as the *individual masking parameters* for sample unit k in this scheme. A data protector such as a national statistical agency should

be able to decide reasonably on these parameters with regard to the privacy protection needed for a survey unit k.

Without loss of generality, let us furthermore assume, that the two categories of y are coded in a way, that the variable value $y_k = 1$ is at least as worthy of protection as $y_k = 0$. If variable y is absolutely nonsensitive for a survey unit k, no privacy protection is needed. Therefore, in this case both the masking parameters p_k and q_k should equal 1 (or 0 respectively, which we ignore subsequently). For a variable, of which only $y_k = 1$ is sensitive but not $y_k = 0$, for the masking parameters $p_k = 1$ and $0 < q_k < 1$ applies. Moreover, if both possible values of y are sensitive (not necessarily equally sensitive) $0 < q_k \leq p_k < 1$ applies.

After the deletion of the original variable y under study at the end of the CSI process, the publishable variable z has to serve as the basis for the estimation of π_A: The probability of $z_k = 1$ is given by

$$P(z_k = 1|y_k) = p_k \cdot y_k + (1 - q_k) \cdot (1 - y_k) = (p_k + q_k - 1) \cdot y_k + 1 - q_k. \quad (5)$$

Hence, the following theorem applies.

Theorem:

(a) For any probability sampling design with design weights d_k

$$\widehat{\pi}_A = \frac{1}{N} \cdot \sum\nolimits_s \widehat{y}_k \cdot d_k = \frac{1}{N} \cdot \sum\nolimits_s \frac{z_k + q_k - 1}{p_k + q_k - 1} \cdot d_k \quad (6)$$

($p_k \neq 1 - q_k \; \forall \; k \in s = m$) is an unbiased moment estimator of parameter π_A of a dichotomous variable y based on the masked variable z.

(b) The variance of $\widehat{\pi}_A$ is given by

$$\begin{aligned} V(\widehat{\pi}_A) &= \frac{1}{N^2} \cdot \left(V\left(\sum\nolimits_s y_k \cdot d_k\right) + \sum\nolimits_U \frac{q_k \cdot (1 - q_k)}{(p_k + q_k - 1)^2} \cdot d_k + \right. \\ &\quad \left. + \sum\nolimits_U \frac{q_k - p_k}{p_k + q_k - 1} \cdot y_k \cdot d_k \right). \end{aligned} \quad (7)$$

(c) $V(\widehat{\pi}_A)$ is unbiasedly estimated by

$$\begin{aligned} \widehat{V}(\widehat{\pi}_A) &= \frac{1}{N^2} \cdot \left(\widehat{V}\left(\sum\nolimits_s y_k \cdot d_k\right) + \sum\nolimits_U \frac{q_k \cdot (1 - q_k)}{(p_k + q_k - 1)^2} \cdot d_k + \right. \\ &\quad \left. + \sum\nolimits_s \frac{q_k - p_k}{p_k + q_k - 1} \cdot \frac{z_k + q_k - 1}{p_k + q_k - 1} \cdot d_k^2 \right). \end{aligned} \quad (8)$$

For the proofs, see Appendix. $V(\sum_s y_k \cdot d_k)$ refers to the variance of the Horvitz-Thompson estimator for the true total t for any probability sampling design. $\widehat{V}(\sum_s y_k \cdot d_k)$ is an unbiased estimator of this variance. In $\widehat{V}(\sum_s y_k \cdot d_k)$ estimator $\widehat{\pi}_A$ is inserted, where an estimator of π_A is needed (see, the example below). The other two summands within the outer brackets of (7) and (8) can be seen as the price that has to be paid by the recipient of the data for the data protection provided by this masking scheme and an estimator of this price. Note, the moment estimator $\widehat{\pi}_A$ may result in a value smaller than 0 or larger than 1. For instance, a negative estimator for π_A may occur, if π_A and n are small. In such

cases, the ML estimator of π_A is simply 0. Compared to the moment estimator the ML estimator is slightly biased, but its MSE is smaller than the variance of $\widehat{\pi}_A$ (cf. Gouweleeuw et al. 1998, p.470).

Formulae (6) to (8) have to be supplied along with the data file in order to enable the recipient to calculate point estimators or confidence intervals and to carry out tests of hypothesis. As an example, let population U be partitioned into H strata U_h of sizes N_h ($h = 1, 2, ..., H$; $\sum N_h = N$). For a stratified simple random sample (STSI) of the population, a simple random sample s_h of size n_h is selected from each stratum U_h without replacement using design weights $d_k = N_h/n_h \; \forall \; k \in U_h$ ($\sum n_h = n$). If the masking parameters are chosen constant for all survey units in sample s_h from stratum h ($p_k = p_h$ and $q_k = q_h \; \forall \; k \in s_h$), estimator (6) is given by

$$\widehat{\pi}_{A,STSI} = \sum\nolimits_h \frac{N_h}{N} \cdot \underbrace{\frac{1}{n_h} \cdot \sum\nolimits_{s_h} \frac{z_k + q_h - 1}{p_h + q_h - 1}}_{\equiv \widehat{\pi}_{A,h}}, \tag{9}$$

where $\widehat{\pi}_{A,h}$ is the unbiased estimator of the proportion $\pi_{A,h}$ of elements belonging to group A in the h-th stratum.

The theoretical variance of (9) is given by

$$\begin{aligned} V(\widehat{\pi}_{A,STSI}) \;\; = \;\; & \sum\nolimits_h \left(\frac{N_h}{N}\right)^2 \cdot \left[\frac{\pi_{A,h} \cdot (1 - \pi_{A,h})}{n_h} \cdot \frac{N_h - n_h}{N_h - 1} + \right. \\ & \left. + \frac{1}{n_h} \cdot \left(\frac{q_h \cdot (1 - q_h)}{(p_h + q_h - 1)^2} + \frac{q_h - p_h}{p_h + q_h - 1} \cdot \pi_{A,h}\right)\right]. \end{aligned} \tag{10}$$

The term

$$\sum\nolimits_h \left(\frac{N_h}{N}\right)^2 \cdot \frac{1}{n_h} \cdot \left(\frac{q_h \cdot (1 - q_h)}{(p_h + q_h - 1)^2} + \frac{q_h - p_h}{p_h + q_h - 1} \cdot \pi_{A,h}\right)$$

is the additional uncertainty caused by the masking of variable y to protect the respondents' privacy. (10) is unbiasedly estimated by

$$\begin{aligned} \widehat{V}(\widehat{\pi}_{A,STSI}) \;\; = \;\; & \sum\nolimits_h \left(\frac{N_h}{N}\right)^2 \cdot \left[\frac{\widehat{\pi}_{A,h} \cdot (1 - \widehat{\pi}_{A,h})}{n_h - 1} \cdot \frac{N_h - n_h}{N_h} + \right. \\ & \left. + \frac{1}{n_h} \cdot \left(\frac{q_h \cdot (1 - q_h)}{(p_h + q_h - 1)^2} + \frac{q_h - p_h}{p_h + q_h - 1} \cdot \widehat{\pi}_{A,h}\right)\right]. \end{aligned} \tag{11}$$

3.3 Hot Deck Random Imputation with Replacement

Another CSI method to reduce the risk of disclosure for survey units when data are handed out to a third party can be described in the following way. After the cloning of the original variable y in a dataset conducted by a probability sampling design P, a subset m with n_m elements is randomly set to missing. This mimics the nonresponse mechanism "missing completely at random" (cf. Little and Rubin 2002, p.12). Hence, the proportion n_m/n, which can be called the artificial nonresponse rate of this procedure, determines the level of privacy protection that is incorporated in the data for the sensitive variable y. Let the I-step of this method replace the artificial missings by hot deck random imputation with replacement (HD) of values taken randomly from the n_r remaining values in the response set r. If the total t of y in the population is the parameter under study, then the estimation

is done by (4). For a general probability sampling design P and for a imputation method I, the statistical properties of $\widehat{t}$ can be derived from

$$E(\widehat{t}) = E_P[E_I(\widehat{t}|r)] \tag{12}$$

and

$$V(\widehat{t}) = V_P[E_I(\widehat{t}|r)] + E_P[V_I(\widehat{t}|r)]. \tag{13}$$

(cf. Little and Rubin 2002, p.67). E_P and V_P denote expected value and variance over the probability sampling design P, whereas E_I and V_I denote the respective measures over the imputation method I.

In a non-self-weighting sample s selected from the population U by an unequal probability sampling design P with design weights d_k not being equal for all survey units of U for imputation method HD the expectation $E(\widehat{t})$ does not equal t, because $E_{HD}(\widehat{t}|r) \neq \widehat{t}_{HT}$. But for self-weighting samples with $d_k = d$ for all survey units the expectation over the imputation method HD is given by

$$E_{HD}(\widehat{t}|r) = d \cdot \left[\sum\nolimits_r y_k + E\left(\sum\nolimits_m \widehat{y}_k\right)\right] = \widehat{t}. \tag{14}$$

This yields $E_P(\widehat{t}) = t$. For the variance $V(\widehat{t})$ to be calculated according to (13) – as an example – let's think of a stratified simple random sample (STSI) with hot deck imputation with replacement (HD,h) applied as imputation method at the I-step of the process within each stratum h ($h = 1, 2, ..., H$). Then, according to (13) the variance of the unbiased estimator $\widehat{t}_h$ of the total t_h within the h-th of H strata is derived in the following way:

$$\begin{aligned} V_{HD,h}(\widehat{t}_h|r_h) &= \left(\frac{N_h}{n_h}\right)^2 \cdot V_{HD,h}(n_r \cdot \overline{y}_{r_h} + (n - n_{r_h}) \cdot \overline{\widehat{y}}_{m_h}) \\ &= \left(\frac{N_h}{n_h}\right)^2 \cdot (n_h - n_{r_h})^2 \cdot V_{HD,h}(\overline{\widehat{y}}_{m_h}) \end{aligned}$$

with r_h, the response set in stratum h with n_{r_h} elements. Furthermore, $\overline{y}_{r_h}$ denotes the mean value of y in r_h and $\overline{\widehat{y}}_{m_h}$ the mean value of the imputed values $\widehat{y}$ in the missing set m_h of stratum h. The variance of the n_{m_h} imputed values is given by

$$V_{HD,h}(\overline{\widehat{y}}_{m_h}|r_h) = \frac{S^2_{r_h}}{n_h - n_{r_h}} \cdot \frac{n_{r_h} - 1}{n_{r_h}},$$

where

$$S^2_{r_h} = \frac{1}{n_{r_h - 1}} \cdot \sum\nolimits_{r_h} (y_k - \overline{y}_{r_h})^2.$$

Hence,

$$V_{HD,h}(\widehat{t}_h|r_h) = N_h^2 \cdot \left(1 - \frac{n_{r_h}}{n_h}\right) \cdot \frac{n_{r_h} - 1}{n_{r_h}} \cdot \frac{S^2_{r_h}}{n_h}$$

and

$$E_{STSI}(V_{HD,h}(\widehat{t}_h|r_h)) = \sum\nolimits_h N_h^2 \cdot \left(1 - \frac{n_{r_h}}{n_h}\right) \cdot \frac{n_{r_h} - 1}{n_{r_h}} \cdot \frac{S^2_{r_h}}{n_h}.$$

From

$$V_{STSI}(E_{HD}(\widehat{t}|r)) = \sum\nolimits_h N_h^2 \cdot \left(1 - \frac{n_{r_h}}{N_h}\right) \cdot \frac{S_h^2}{n_{r_h}}$$

with

$$S_h^2 = \frac{1}{N_h - 1} \cdot \sum\nolimits_{U_h} (y_k - \overline{y}_{U_h})^2.$$

it can be derived that

$$V(\widehat{t}) = \sum\nolimits_h N_h^2 \cdot \left(1 - \frac{n_h}{N_h}\right) \cdot \frac{S_h^2}{n_h} + \sum\nolimits_h N_h^2 \cdot \left(\frac{n_h - 1}{n_{r_h}} - \frac{n_{r_h} - 1}{n_h}\right) \cdot \frac{S_h^2}{n_h}. \qquad (15)$$

In (15) the first component of the two summands is the sample variance of the original data, whereas the second can be interpreted as the price that has to be paid for the reduction of disclosure risk in terms of accuracy. This increase in variance is only negligible for n_{r_h} close to n_h.

4 Summary

On the one hand, data providers like national statistical offices try hard to achieve high data quality in sample- or population surveys. On the other hand, the laws of data protection make it necessary to apply techniques of statistical disclosure control (SDC) to distort surveyed information before it can be delivered to secondary analysts outside the agency.

A three-step process characterizes a whole family of masking schemes. The three consecutive steps consist of the cloning of the sensitive variable (C-step), data suppression within the clone (S-step) and the use of imputation methods to fill in for the artificially generated missings (I-step). Well-known methods like data swapping, micro-aggregation of data or addition of noise do belong to this family. The idea of the definition of the CSI family is to incorporate the wide field of imputation methods for SDC. In the paper, three examples of members of this family have been presented. These show how different such methods, following the CSI strategy, can be. All techniques of randomized response belong to the family as well as well as post randomization methods. The third example uses the simplest imputation technique at the I-step.

5 Appendix: Proof of the Theorem

Proof of Theorem (a):

For a sampling design P and a post randomization mechanism R determining the masking parameters p_k and q_k for all sampled units we have

$$\begin{aligned} E(\widehat{\pi}_A) &= \frac{1}{N} \cdot E_P \left[E_R \left(\sum\nolimits_s \left(\frac{z_k + q_k - 1}{p_k + q_k - 1} \cdot d_k \right) \middle| s \right) \right] \\ &= \frac{1}{N} \cdot E_P \left(\sum\nolimits_s d_k \cdot E_R \left(\frac{z_k + q_k - 1}{p_k + q_k - 1} \middle| s \right) \right) \\ &= \frac{1}{N} \cdot E_P \left(\sum\nolimits_s y_k \cdot d_k \right) = \frac{1}{N} \cdot \sum\nolimits_U y_k = \pi_A. \end{aligned}$$

Proof of Theorem (b):

For sampling design P and post randomization mechanism R the variance of estimator $\widehat{\pi}_A$ (6) is given by

$$V(\widehat{\pi}_A) = V_P(E_R(\widehat{\pi}_A|s)) + E_P(V_R(\widehat{\pi}_A|s)).$$

Then

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

Let the sample inclusion indicator be

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

The covariance $C_R(\frac{z_k+q_k-1}{p_k+q_k-1}, \frac{z_l+q_l-1}{p_l+q_l-1}\Big| s)$ is equal to 0 ($\forall\ k \neq l$). Because $E_P(I_k^2) = E_P(I_k) = 1/d_k$ applies, the second summand of $V(\widehat{\pi}_A)$ is given by

$$\begin{aligned} E_P(V_R(\widehat{\pi}_A|s)) &= E_P\left[\frac{1}{N^2} \cdot V_R\left(\sum\nolimits_U I_k \cdot \frac{z_k+q_k-1}{p_k+q_k-1} \cdot d_k \Big| s\right)\right] \\ &= E_P\left[\frac{1}{N^2} \cdot \sum\nolimits_U I_k^2 \cdot d_k^2 \cdot V_R\left(\frac{z_k+q_k-1}{p_k+q_k-1}\right)\right] \\ &= \frac{1}{N^2} \cdot \sum\nolimits_U V_R\left(\frac{z_k+q_k-1}{p_k+q_k-1}\right) \cdot d_k. \end{aligned}$$

Thus,

$$V_R\left(\frac{z_k+q_k-1}{p_k+q_k-1}\right) = \frac{1}{(p_k+q_k-1)^2} \cdot V_R(z_k)$$

and because of $y_k^2 = y_k$ it follows

$$\begin{aligned} V_R(z_k) &= 1 - q_k + (p_k+q_k-1) \cdot y_k - (1 - q_k + (p_k+q_k-1) \cdot y_k)^2 \\ &= (1 - q_k + (p_k+q_k-1) \cdot y_k) \cdot (q_k - (p_k+q_k-1) \cdot y_k) \\ &= (1-q_k) \cdot q_k + (p_k+q_k-1) \cdot (q_k - p_k) \cdot y_k. \end{aligned}$$

Furthermore this yields

$$\begin{aligned} E_P(V_R(\widehat{\pi}_A|s)) &= \frac{1}{N^2} \cdot \left(\sum\nolimits_U \frac{(1-q_k) \cdot q_k}{(p_k-q_k-1)^2} \cdot d_k + \right. \\ &\quad \left. + \sum\nolimits_U \frac{q_k - p_k}{p_k+q_k-1} \cdot y_k \cdot d_k\right), \end{aligned}$$

which completes the proof.

Proof of Theorem (c):

Per definitionem $\widehat{V}_P(\sum_s y_k \cdot d_k)$ is an unbiased estimator of $V_P(\sum_s y_k \cdot d_k)$. Furthermore

$$\sum\nolimits_s \frac{q_k - p_k}{p_k+q_k-1} \cdot \frac{z_k+q_k-1}{p_k+q_k-1} \cdot d_k^2$$

is an unbiased estimator of

$$\sum\nolimits_U \frac{q_k - p_k}{p_k+q_k-1} \cdot y_k \cdot d_k.$$

Therefore (8) is an unbiased estimator of (7).

References

Dalenius, T. and Reiss, S. P. (1982). Data-Swapping: A Technique for Disclosure Control. Journal of Statistical Planning and Inference, 6, 73–85.

Defays, D. and Anwar, M. N. (1998). Masking Microdata Using Micro-Aggregation. Journal of Official Statistics, 14 (4), 449–461.

Drechsler, J., Bender, S., and Rässler, S. (2008). Comparing Fully and Partially Synthetic Datasets for Statistical Disclosure Control in the German IAB Establishment Panel. Transactions on Data Privacy, 1, 1002–1050.

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.

Little, R. J. A. and Rubin, D. B. (2002). Statistical Analysis with Missing Data. 2^{nd} edition. Hoboken: Wiley & Sons.

Lohr, S. L. (2010). Sampling: Design and Analysis. 2^{nd} edition. Brooks/Cole: Boston.

Matthews, G. J. and Harel, O. (2011). Data confidentiality: A review of methods for statistical disclosure limitation and methods for assessing privacy. Statistics Surveys, 5, 1–29.

Quatember, A. (2009). A Standardization of Randomized Response Strategies. Survey Methodology, 35 (2), 143–152.

Reiter, J. P. (2003). Inference for Partially Synthetic, Public Use Microdata Sets. Survey Methodology, 29 (2), 181–188.

Rubin, D. B. (1987). Multiple Imputation for Nonresponse in Surveys. New York: Wiley & Sons.

Rubin, D. B. (1993). Discussion: Statistical Disclosure Limitation. Journal of Official Statistics, 9 (2), 461–468.

Singh, S. and Chen, C. C. (2009). Utilization of higher order moments of scrambling variables in randomized response sampling. Journal of Statistical Planning and Inference, 139, 3377–3380.

Tracy, D.S. and Mangat, N.S. (1996). Some Developments in Randomized Response Sampling during the last Decade – A Follow Up of Review by Chaudhuri and Mukerjee. Journal of Applied Statistical Science, 4 (2/3), 147–158.

van den Hout, A. and van der Heijden, P. G. M. (2002). Randomized Response, Statistical Disclosure Control and Misclassification: a Review. International Statistical Review, 70 (2), 269–288.

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.

Willenborg, L. and de Waal, T. (2001). Elements of Statistical Disclosure Control. New York: Springer.

Winkler, W. E. (2004). Masking and Re-identification Methods for Public-use Microdata: Overview and Research Problems. Statistical Research Division, U.S. Bureau of the Census, Research Report Series (Statistics # 2004-06).