Statistical Disclosure Control for Frequency Tables

A thesis submitted to The University of Manchester
for the degree of
Doctor of Philosophy
in the Faculty of Humanities

2016

László Antal

School of Social Sciences

Contents

List of Figures

List of Tables

$\approx$ 55,000

Abstract

Statistical Disclosure Control for Frequency Tables

A thesis submitted to The University of Manchester for the degree of
Doctor of Philosophy
in the Faculty of Humanities

2016

László Antal

Disclosure risk assessment of statistical data, such as frequency tables, is a prerequisite for data dissemination. This thesis investigates the problem of disclosure risk assessment of frequency tables from the perspective of a statistical institute.

In the research reported here, disclosure risk is measured by a mathematical function designed for the data according to a disclosure risk scenario. Such functions are called disclosure risk measures. A disclosure risk measure is defined for frequency tables based on the entire population using information theory.

If the disclosure risk of a population based frequency table is high, a statistical institute will apply a statistical disclosure control (SDC) method possibly perturbing the table. It is known that the application of any SDC method lowers the disclosure risk. However, measuring the disclosure risk of the perturbed frequency table is a difficult problem. The disclosure risk measure proposed in the first paper of the thesis is also extended to assess the disclosure risk of perturbed frequency tables.

SDC methods can be applied to either the microdata from which the frequency table is generated or directly to the frequency table. The two classes of methods are called pre- and post-tabular methods accordingly. It is shown that the two classes are closely related and that the proposed disclosure risk measure can account for both methods.

In the second paper, the disclosure risk measure is extended to assess the disclosure risk of sample based frequency tables. Probabilistic models are used to estimate the population frequencies from sample frequencies which can then be used in the proposed disclosure risk measures.

In the final paper of the thesis, we investigate an application of building a flexible table generator where disclosure risk and data utility measures must be calculated on-the-fly. We show that the proposed disclosure risk measure and a related information loss measure are adaptable to these settings. An example implementation of the disclosure risk and data utility assessment using the proposed disclosure risk measure is given.

Declaration

No portion of the work referred to in the thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Copyright Statement

i. The author of this thesis (including any appendices and/or schedules to this thesis) owns certain copyright or related rights in it (the "Copyright") and s/he has given The University of Manchester certain rights to use such Copyright, including for administrative purposes.

ii. Copies of this thesis, either in full or in extracts and whether in hard or electronic copy, may be made only in accordance with the Copyright, Designs, and Patents Act 1988 (as amended) and regulations issued under it or, where appropriate, in accordance with licensing agreements which the University has from time to time. This page must form part of any such copies made.

iii. The ownership of certain Copyright, patents, designs, trade marks and other intellectual property (the "Intellectual Property") and any reproductions of copyright works in the thesis, for example graphs ad tables ("Reproductions"), which may be described in this thesis, may not be owned by the author and may not be owned by third parties. Such Intellectual Property and Reproductions cannot and must not be made available for use without the prior written permission of the owner(s) of the relevant Intellectual Property and/or Reproductions.

iv. Further information on the conditions under which disclosure, publication and commercialisation of this thesis, the Copyright and any Intellectual

Chapter 1

Introduction

1.1 Statistical Disclosure Control

The collection and dissemination of data serve various purposes, such as supporting the policy and decision making processes. The quality of such decision making effects all of our lives and therefore the need for data collection and dissemination is widely accepted as necessary component of a well-functioning society. Many organisations collect and treat data for statistical purposes. However, the confidential treatment of individual data is a requirement for every organisation that holds data. This requirement has emerged as a social need and is regulated by law in many countries. A high number of organisations collect and treat data for statistical purposes.

Statistical agencies produce outputs based on confidential data. Confidentiality involves not just excluding unauthorised people from data processing but also not disseminating data containing identifiable individual information. In order to maintain the individuals' confidentiality, statistical disclosure control (SDC) methods are applied to the data before dissemination. Statistical disclosure control aims at discovering the best possible methods to ensure confidentiality.

The social need for data protection stems from the belief that confidential

data might be used intentionally for harmful purposes. A hypothetical person who attempts to use disseminated data for detrimental purposes is called the intruder. The intruder scrutinises the data and attempts to reveal confidential parts of them. If the intruder is able to connect some parts of the released data to a certain individual, then a breach of confidentiality might happen. The risk inherent in the data might increase as the number of such possible individual-data matches rises. Statistical disclosure control methods limit the intruder's opportunity for matching individuals and data. In fact, SDC methods foster the dissemination of data since they eliminate or reduce the disclosure risk to an acceptable level before the data release.

In order to identify the risk in the data to be disseminated, a data protector has to assess the potential sources of risk. These sources come from assumptions about how an intruder attempts to reveal confidential data. The assumed methods of a potential intruder are called intruder scenarios. To prevent such intrusion, the data protector has to act as an intruder and detect possible vulnerabilities of the data. This detection should be followed by the elimination of the vulnerabilities.

The other, also very important, side of the coin is the utility of the released data compared to the data before SDC methods are applied. SDC methods distort the distribution of the data. The aim when employing SDC methods is to cause the least distortion, while guaranteeing confidentiality.

1.2 Tabular Data

Statistical disclosure control methods apply to microdata as well as to tabular data. This thesis deals with tabular data, especially with frequency tables. A cell in a table is determined by a category-combination of table-spanning variables.[1]

[1]To generate a frequency table we need to select variables of a microdata set and cross-classify them. We refer to the selected variables as table-spanning variables.

A frequency table can be based either on the data of the entire population or on sample data. The distinction is important as the evaluation of the disclosure risk in the two cases requires different methods.

In a population based table, for example a census table, each individual contributes to one cell, that is, the cells partition the population into equivalence classes. The cell value or cell frequency or cell count is the number of individuals contributing to the cell or, in other words, the size of the equivalence class.

Sample based tables do not include every individual of the population. Individuals not selected in the sample do not contribute to the cells, therefore a sample-based cell frequency will usually be lower than the (possibly unknown) population cell frequency. Often, a sample weight is associated with each responding unit in the sample and only a weighted count might be released.

Throughout the thesis we call a cell value small if it is 1 or 2. Identity disclosure happens when an individual can be identified in released data. An intruder might identify an individual without further knowledge if the cell value is 1. If the cell value is 2, then one of the contributors of the cell might identify the other without further knowledge. In such situations the first contributor can be considered as an intruder. If a cell value is at least 3, then the individuals might be identified if further information is available to the intruder. By further information we mean that the intruder has knowledge of cells where some of the individuals fall. In other words, the intruder might have his/her own frequency table on a subset of the population (or sample) and he/she can extract his/her own frequencies from the frequency table released by the statistical institute. There might be small cell values in the resulting frequency table. The chance of identity disclosure usually decreases as the frequency of a cell rises. However, an intruder might have extensive knowledge on certain datasets, therefore even high cell frequencies can lead to identity disclosure. Disclosive cells might be defined differently depending on the

data protector's assumptions about intruder knowledge.

Identity disclosure itself might not do much harm to the identified individual. However, identity disclosure often implies attribute disclosure. In case of attribute disclosure the intruder learns a new attribute, new information about an individual or a group of individuals. For example, a row in a frequency table might enable attribute disclosure if it has only one positive cell value and the rest of the cells are zeroes. The cells of the row and the categories of a table-spanning variable correspond, therefore the intruder will know that the individuals contributing to the row necessarily have only one particular category of the variable. Conversely, a row with the same total but equal cell frequencies is likely to be of lower disclosure risk. A disclosure risk measure should reflect the risk (and the lack of risk) of identity disclosure and attribute disclosure.

To measure disclosure risk in tabular data, a substantial part of this work employs the entropy to formulate a disclosure risk measure. The entropy of a $P = (p_1, p_2, \ldots, p_K)$ probability distribution is

$$H(P) = -\sum_{i=1}^{K} p_i \cdot \log p_i \ .$$

Entropy is suitable to capture the risk of attribute disclosure. It takes its extreme values when the P distribution is uniform or degenerate. This property reflects the risk (and the lack of risk) of attribute disclosure described above.

Therefore, in the case of population frequency tables the entropy (or an appropriately chosen function of the entropy) helps to measure the disclosure risk. However, entropy is defined for probability distributions and for count data it will first be necessary to adapt it accordingly and apply a transformation. One of the main points of this work is to identify how the entropy can capture the disclosure risk in population based frequency tables. Chapter 3 describes a possible approach to this problem.

If the disclosure risk measured by the entropy is high, a statistical disclosure control method will be applied to the data. This method can be perturbative, that is, some of the categories of some individuals are deliberately changed in order to introduce more uncertainty into the table. The extent of distortion caused by the SDC method requires the comparison of the original and the perturbed data. f-divergences (where f is an appropriately chosen function) can be employed in order to compare the probability distributions of the original and perturbed data. The f-divergence of the distributions, denoted by $P = (p_1, p_2, \ldots, p_K)$ and $Q = (q_1, q_2, \ldots, q_K)$, is given by

$$D_f(P \parallel Q) = \sum_{i=1}^{K} q_i \cdot f\left(\frac{p_i}{q_i}\right) \ .$$

Chapter 3 also introduces some well-known f-divergences and their suitability for measuring the utility of the perturbed data.

In Chapter 4 we focus on sample based tables. Although the disclosure risk of sample based frequency tables needs to be measured differently to population based tables, an entropy based approach can also be used. The difference from population based tables is that the intruder does not know whether a particular individual contributes to the table or not. Even if the intruder knows that a particular individual contributes to a sample based frequency table, the intruder might not be aware of individuals having the same or similar attributes, therefore attribute disclosure might not happen with absolute certainty. In other words, sampling brings more uncertainty from the intruder's point of view. This uncertainty should be reflected in the disclosure risk measure. Chapter 4 deals with the problem of measuring disclosure risk in sample based frequency tables.

If the disclosure risk measure is high and an SDC method is applied to the sample-based table, then the problem of measuring utility arises again. The situation resembles the problem of measuring the utility in population-based tables.

Chapter 5 presents an application based on the theoretical parts of Chapters 3 and 4. It is shown how a statistical institute can carry out disclosure risk assessment on-the-fly in the context of a flexible table generator. This is a web-based platform where users can define and generate their own tables of interest. There are several examples of these web-based platforms that are in use for census data, for example in the United States and in Australia.

Chapter 6 summarizes the results of the main chapters and discusses potential future work.

1.3 'Beyond 2011' Programme

The PhD studentship is sponsored by the 'Beyond 2011' programme of the Office for National Statistics (ONS). The programme aims at discovering viable alternatives to the traditional census based on complete enumeration. The programme intends to exploit administrative data sources, design-based and model-based estimates to produce census outputs. This approach requires fundamentally new SDC consideration compared to previous censuses. In particular, the Office for National Statistics identified attribute disclosure as the key risk of disclosure that should be considered in the disclosure risk assessment for UK Census tabular outputs. When assessing disclosure risk in the 2011 UK Census tables, heuristics were used to assess attribute disclosure based on multiple generations of tables from test data. These methods are described in Chapter 3. Therefore, the main driver for this thesis supported by the 'Beyond 2011' programme was to develop a quantitative disclosure risk measure for attribute disclosure to replace the heuristics that were developed for the 2011 Census.

Another aspect of the 'Beyond 2011' programme is the need to modernize data dissemination. The number of potential data users is rising and they desire instant access to the data. To meet these increasing

needs, statistical agencies are considering new, alternative channels of data dissemination. The ONS aims to promote online availability for the 2021 census outputs in order to ease data access. In case of tabular outputs the goal would be to enable the users to tailor and generate their own tables through web-based platforms. From a statistical disclosure point of view it demands the automated assessment of disclosure risk. This assessment should be fast and on-the-fly since it would delay the data access. Therefore, ideally the system of the web-based platform should employ formulae that are relatively easy to calculate and should capture the disclosure risk. If the risk exceeds a certain level set by the statistical institute, the output might either be denied or it may be perturbed, reassessed for disclosure risk and subsequently released to the data user.

This thesis aims to contribute to the problems mentioned above. It seeks new approaches to the disclosure risk measurement of census outputs, especially frequency tables. Such tables can be based either on whole population counts or on sample counts.

1.4 Alternative Format Thesis

The thesis is written using the alternative format, that is, it centres on three papers. Each of the main chapters, namely Chapters 3, 4 and 5, includes a paper. Each paper is the central part of its chapter. The papers present the essence of the chapter and they are suitable for publication on their own. However, they pertain to the same research area and together they constitute a body of work on a set of related statistical disclosure control problems. The papers were written concisely in order to make their publication possible. Therefore, additional ideas and details are given in the corresponding chapters.

The papers presented in Chapters 3, 4 and 5 were co-authored with my supervisors, Natalie Shlomo and Mark Elliot. Newly defined disclosure risk

measures appear in every paper and the measures were developed through the normal supervision process. The papers in Chapters 3 and 4 were primarily written by myself with comments provided by my co-authors. The application paper in Chapter 5 was led by Natalie Shlomo's work with contributions by myself and Mark Elliot.

1.5 Motivation

The motivation for this work is to define a disclosure risk measure that reflects the properties of attribute disclosure in frequency tables.

The aim of introducing the disclosure risk measure is to facilitate data releases by statistical institutes. The disclosure risk measure helps decide whether particular data can be disseminated and/or whether disclosure control needs to be applied prior to release.

The thesis takes the viewpoint of the statistical institute that has access to the original data and needs to assess disclosure risk of statistical outputs. The question addressed is how to quantify the disclosure risk in frequency tables.

The thesis contributes to the SDC literature with respect to the disclosure risk measurement using information theory. A central theoretical goal is to show that entropy and conditional entropy can be used to measure the disclosure risk.

Chapter 2

Some Aspects of Statistical Disclosure Control

This chapter highlights some aspects of statistical disclosure control that are important from the point of view of a statistical institute.

The main objective of a statistical institute is to disseminate useful data. Tabular data have always been a significant element of this dissemination function. Recently, the demand for microdata sets has also been increasing, mainly from researchers. Microdata offer more flexibility than tabular data because they include more variables than a given tabular output and serve as the basis of more sophisticated statistical analysis.

Regardless of the data format, statistical institutes have to protect against potential disclosure. Section 2.1 discusses disclosure, disclosure risk and disclosure risk measures in general. Section 2.2 introduces our notation. Some measures of disclosure risk are presented in Section 2.3. Section 2.4 describes SDC methods for tabular data. The list cannot be exhaustive because SDC is a live research area in itself and new SDC methods are always being investigated and introduced; the goal of section 2.4 is to give an overview of basic SDC methods. In Section 2.5 we briefly discuss the loss of information SDC methods cause.

2.1 Disclosure, Disclosure Risk and Disclosure Risk Measures

Disclosure occurs if an intruder learns the identity and/or some attributes of an individual or a group of individuals. By releasing data, statistical institutes expose the individuals contributing to the data to intruders' potential attacks. *Disclosure risk* is a term that is hard to define and may mean different things according to a disclosure risk scenario. Data protectors have recognized and collected certain permutations that lead to disclosure with high probability, for example, a cell value of 1 in a frequency table might imply identity disclosure. The disclosure risk of a particular microdata set or table becomes more severe as the number of such permutations increases. *Disclosure risk measures* quantify the disclosure risk with respect to a specified risk scenario. Disclosure risk measures, in fact, reflect the data custodian's understanding of disclosure risk and assign numerical values to statistical outputs, such as frequency tables. Disclosure risk measures are based on the list of potentially disclosive permutations, for example the number of cell values of size 1 in a table. However, such a list cannot be exhaustive, hence there is no universal or best disclosure risk measure. The need for disclosure risk measures lies in supporting good decision making whether particular data are safe to release or not. Without disclosure risk measures such decisions are subjective and subject to individual cognitive biases. Numerical values make the decision more objective and impartial. In other words, disclosure risk measures are mathematical functions. Such functions are applied to the statistical output to be released and they compress the disclosure risk into a numerical value.

Throughout the thesis the term 'disclosure risk measure' has two meanings. It is either a mathematical function as described above or the numerical value that such a mathematical function provides when applied to particular data.

Since disclosure risk and disclosure risk measure are closely related terms, sometimes, for ease of expression, we use them interchangeably.

2.2 Notation

Throughout the thesis we use the following notation.

- The number of cells in the frequency table(s) is K.
- The cells are denoted $C = \{c_1, c_2, \ldots, c_K\}$.
- The vector of frequencies of the original population based frequency table is $F = (F_1, F_2, \ldots, F_K)$.
- The sum of the frequencies is $N = \sum_{i=1}^{K} F_i$. It means implicitly that the population consists of N individuals.
- The set of individuals will be denoted $I = \{a_1, a_2, \ldots, a_N\}$.
- The probability distribution $P = (p_1, p_2, \ldots, p_K)$ will provide the probability that an individual falls into cell c_i in the original population based table.
- The vectors of frequencies of perturbed and sample based tables are $G = (G_1, G_2, \ldots, G_K)$ and $f = (f_1, f_2, \ldots, f_K)$ respectively.
- In few cases we will use the uniform distribution. If A is a (finite) set (for example C or I), then U_A is the uniform distribution on A. The power set of A is $\mathcal{P}(A) = \{B : B \subseteq A\}$, while the cardinality of A is $|A|$.

2.3 Measuring Disclosure Risk

2.3.1 Variables of Microdata Sets

Microdata sets are not the main topic of our discussion. However, tabular data are generated from microdata, and as we will discuss some SDC methods are applied to the originating microdata, therefore a brief discussion of microdata is warranted.

Microdata sets usually contain more information than tabular data, and therefore a data protector needs to act more carefully. A microdata set consists of data units, such as individuals, and variables that describe the attributes of those data units. Hundepool et al. (2012) produced a four-way categorization of variables, which with slight variation of terminology is widely used for SDC purposes.

The first category is directly *identifying variables*, for instance name or personal identification number. From a statistical disclosure control point of view it is essential to remove such variables before the data are disseminated.[1] The main purpose of statistical data release is to provide accurate information about the overall population. Data, including microdata, may serve as a basis of decision making and/or statistical analysis. Neither requires identifying variables, therefore the removal of such variables does not conflict with the interest of data users.

The second category of variables is called *quasi-identifiers*. Such variables are usually categorical and do not identify individuals directly. However, in combination with other variables they might allow an intruder to identify individuals.

[1] A directly identifying variable identifies the individuals without any further information. An intruder might identify an individual by name or personal identification number. However, there are direct identifiers that do not raise much disclosure risk. Such identifiers usually provide a unique number for each individual in the population/sample. They are defined by the statistical institute and not disseminated. Since an intruder does not know which number relates to which individual, breach of confidentiality is unlikely to happen. Such directly identifying variables can be used to link two or more different datasets within the statistical institute or by a trusted third party.

Non-sensitive variables are the third category. The disclosure of such variables does not have much detrimental effect. An example of a non-sensitive variable might be the number of books read in a month from a survey examining access to public libraries.

The fourth category, which consists of *sensitive variables*[2], is the biggest concern from a statistical disclosure control point of view. Individuals would like to keep sensitive variables, such as health information or income, secret. When providing statistical institutes with sensitive variables, individuals place their trust in the institutes. Disclosure of sensitive variables might result in the loss of individuals' trust and impact negatively on response rates.

The four categories listed above are not necessarily disjoint. For example, the 'gender' variable might be considered as a quasi-identifying and non-sensitive variable.

The overarching assumption of disclosure risk problem of microdata is that an intruder attempts to identify individuals using the quasi-identifiers. A data protector often considers disclosure scenarios. He/she tries to guess how an intruder would attack the microdata. The data protector assumes that the intruder might base the attack on quasi-identifying variables, called key variables. For example, the data protector often calculates frequencies given by the value-combinations of certain key variables in the microdata. If some frequencies are low, then the data protector needs to act accordingly.

[2]The categorization of a particular variable into the class of sensitive variables should depend on what the society considers as sensitive. The law of a certain country always tries to reflect the needs of the society. Sensitive variables, and in general, statistical confidentiality is also regulated by law. In the UK sensitive variables are regulated in the Data Protection Act 1998. At European level the European Data Protection Directive 1995 and the 831/2002 Commission Regulation are the most relevant legal acts to our topic.

2.3.2 Disclosure Risk Measures of Tabular Data

For tabular data Willenborg and de Waal (2001) assert that: 'Disclosure risk [measures] may be defined either for the whole table or separately for each cell into which the table is organized.' In both cases, SDC methods need to be applied if the disclosure risk measure exceeds a predefined threshold.

2.3.2.1 Disclosure Risk Measures of Table Cells

In understanding disclosure risk for tabular data we need to distinguish between frequency tables and magnitude tables. See Willenborg and de Waal (2001) for a discussion on the disclosure risk of both kinds of tables. Here we outline the essential issues.

2.3.2.1.1 Disclosure Risk Measures of Frequency Tables

A cell of a frequency table contains the number of individuals of the population (or sample) that possess the cell-defining attributes. Frequency tables are often generated and released from census data.

The most common disclosure risk measure for frequency tables is called a threshold rule or minimum frequency rule. The disclosure risk measure of a cell is high if the cell value is lower than a given threshold.

2.3.2.1.2 Disclosure Risk Measures of Magnitude Tables

Many disclosure risk measures have been introduced for table cells of magnitude tables. They are often called 'sensitivity rules'[3] (see Hundepool et al. (2012)).

A definition of magnitude tables can be found in Hundepool et al. (2012) as follows. 'In a magnitude table, each table cell value represents the sum of a particular response, across all respondents that belong to

[3]Here 'sensitivity' has a different meaning to that discussed for sensitive variables of microdata. Here it refers to cells that are potentially disclosive.

that cell. Magnitude tables are commonly used for business or economic data providing, for example turnover of all businesses of a particular industry within a region.'

For magnitude tables the (n, k)-dominance rule and the (p, q)-rule are two of the most important sensitivity measures. The former defines a cell sensitive if the n greatest contribution to the cell surpasses $k\%$ of the cell value. The latter also has two parameters. It assumes that a respondent's contribution to the cell can be estimated within $q\%$ prior to the data dissemination. The cell is considered sensitive if the respondent's contribution to the cell can be estimated within $p\%$ after the data release.

Oganian and Domingo-Ferrer (2003) introduced a conditional-entropy related disclosure risk measure. (See also Domingo-Ferrer et al. (2002).) The disclosure risk is defined for table cells and uses the conditional entropy. It takes the original cell value as well as the intruder's knowledge into account. The disclosure risk for a particular cell is

$$DR(\mathcal{X}) = \frac{1}{H(\mathcal{X}|\mathcal{Y}=y)} = \frac{1}{-\sum_x Pr(\mathcal{X}=x|\mathcal{Y}=y)\cdot \log_2 Pr(\mathcal{X}=x|\mathcal{Y}=y)}\,.$$

Here $\mathcal{X}$ represents a random variable that provides the cell value, while $\mathcal{Y}$ is the intruder's knowledge. The threshold rule, the (n, k)-dominance rule and the (p, q)-rule quantify the disclosure risk before an SDC method is applied to the table. $DR(\mathcal{X})$ may be employed after that.

The nature of $DR(\mathcal{X})$ is close to the disclosure risk measure we introduce in Chapter 3. Our disclosure risk measure is defined for whole frequency tables, not for table cells. We measure the disclosure risk before as well as after an SDC method is applied to the table.

2.3.2.2 Disclosure Risk Measures of Entire Tables

So far we have discussed disclosure risk measures that are defined for table cells. Such measures can easily be transformed into a disclosure risk measure for the entire table, for example by calculating the number/percentage

of cells that have higher cell-level disclosure risk measure than a given threshold. Below we discuss further disclosure risk measures defined for entire (frequency) tables in the SDC literature.

2.3.2.2.1 SDC Literature for Population Based Tables

There is limited literature that relates to the disclosure risk of whole frequency tables, and in particular based on information theory. Frank (1976) identified three types of potential scenarios for disclosure in frequency tables.

The first assumes that only one individual contributes to a row in a frequency table. In that case, the identification of the individual might imply disclosure of new attributes.

Only two individuals are in a certain row in the second scenario and they are not in the same cell. Each individual can then know the other's previously unknown attribute.

Also two cells are populated in the third scenario but only one of their values needs to be 1. The individual who is unique in the row can then know the attribute of the other contributors of the row.

These basic scenarios show some concepts of statistical disclosure control. The data protector assumes that a potential intruder may or may not contribute to the frequency table.

The concepts given by Frank (1976) are valid. However, a data protector also needs to think about the data environment. The data environment approach attempts to take into account all possible datasets available to a potential intruder. It is assumed that the intruder tries to link the available datasets to the disseminated data. For a reference on the data environment see Elliot et al. (2010).

An information-theoretical approach to disclosure risk assessment was introduced by Frank (1978). It is assumed that an intruder possesses information about a set of individuals, denoted by A. The $A \in \mathcal{P}(I)$ set, called the prior disclosure set, is random, there is a probability distribution

given on $\mathcal{P}(I)$. Denote the distribution by $\mathcal{D}_A$. The prior information allows the intruder to disclose (at least) the individuals of A after the release of the frequency table. If every individual in $I \setminus A$ falls into the same cell, then every individual of I is disclosed by definition, otherwise only the individuals of A. It defines a so-called posterior disclosure set, denoted by B.

$$B = \begin{cases} I & \text{if the individuals in } I \setminus A \text{ fall into the same cell,} \\ A & \text{otherwise.} \end{cases}$$

Since A is random, so is B. Denote the distribution of $B \in \mathcal{P}(I)$ by $\mathcal{D}_B$. It is proven that

$$0 \leq H(\mathcal{D}_A) - H(\mathcal{D}_B) \leq \log\left(1 + \sum_{i=1}^{K}(2^{F_i} - 1)\right) ,$$

where $H(\mathcal{D}_A)$ and $H(\mathcal{D}_B)$ are the respective entropies of $\mathcal{D}_A$ and $\mathcal{D}_B$. The $H(\mathcal{D}_A) - H(\mathcal{D}_B)$ difference is suggested as a potential disclosure risk measure. Under various assumptions about the $\mathcal{D}_A$ distribution, $H(\mathcal{D}_A) - H(\mathcal{D}_B)$ is calculated exactly.

The subtraction-attribution probability (SAP) method (see Smith and Elliot (2008)) is mainly based on the frequency table to be published and on the intruder's knowledge. The paper shows how zero cells in the frequency table can lead to disclosure and provides a disclosure risk measure.

Statistical institutes publish tables which by definition are the margins of larger tables. In effect the internal cells of the larger table have been suppressed. An important question is whether those internal cells can be recalculated from the released marginals. Even if an intruder cannot determine the frequencies exactly, a lower and an upper bound of an internal cell can be given by the so called Frechet bounds. The bounds depend on the number of the table-spanning variables. Assume that the table has k spanning variables. We follow the notation of Fienberg (1999)

and denote an internal cell by $n_{i_1 i_2 \ldots i_k}$. The marginal total over the jth variable is $n_{i_1 \ldots i_{j-1} + i_{j+1} \ldots i_k}$. If the summation is over more variables, then the corresponding indices are substituted for $'+'$ signs. It implies that the overall total is $n_{++\ldots+}$.

The bounds for a 2×2 two-dimensional table, where $i_1, i_2 \in \{1, 2\}$, are given as follows.

$$\min\{n_{i_1+}, n_{+i_2}\} \geq n_{i_1 i_2} \geq \max\{n_{i_1+} + n_{+i_2} - n_{++}, 0\}$$

More general bounds are given by Dobra and Fienberg (2000). They consider decomposable and reducible graphs, which correspond to the set of marginal tables, and provide sharper bounds if those special classes of marginal tables are given.

The so called "shuttle algorithm" was proposed by Buzzigoli and Giusti (1999) (see also Buzzigoli and Giusti (2006)). It starts with an upper bound and a lower bound for each of the cell entries, for example with the Frechet bounds. The bounds are based on ($(k-1)$-dimensional) marginal totals. In each step the algorithm computes new upper and lower bounds. The essence of the algorithm can be found in Buzzigoli and Giusti (2006) as follows.

- 'the upper bound of the generic element cannot be greater than the lowest difference between each of its marginals and the sum of the lower bounds of the other elements along the same dimension (row, column, etc.); to start the procedure the lower bounds are set to zero;'

- 'the lower bound of the generic element cannot be less than the highest positive difference between each of its marginals and the sum of the upper bounds of the other elements along the same dimension (row, column, etc.); if no difference is positive the lower bound remains zero;'

- 'if some of the lower bounds are greater than zero the previously computed upper bounds could obviously change; revised upper bounds imply the revision of the previously computed lower bounds and so on'.

The shuttle algorithm does not always provide sharp bounds. However, computationally it is effective providing results quickly.

2.3.2.2.2 SDC Literature for Sample Based Tables

Disclosure risk might arise if a particular individual can be identified from the released data. If a cell value in the sample-based table is 1, we will call the individual contributing to this cell a sample unique. Similarly, an individual is a population unique if there is no other individual having the same value combination of the table-spanning variables in the population. A population unique, if selected in the sample, is a sample unique. However, sample uniques are not necessarily population uniques.

Bethlehem et al. (1990) also considers sample uniques to be of the highest disclosure risk within sample microdata. They define the so-called 'key' as 'the set of variables used for identification'. They omit the cells that have 0 population counts. The table size is therefore the number of category-combinations that actually occur. Therefore the table size can be smaller than the total number of the categories. It means that the population frequencies (F_i) can only be positive but there might be zeroes among the sample frequencies (f_i). Their estimation of the number of population uniques is based on a Poisson-gamma model. The model can be formulated as follows.

$$F_i|\Pi_i = \pi_i \sim Po(N\pi_i) \quad \text{and} \quad \Pi_i \sim \text{Gamma}(\alpha, \beta) \ .$$

These assumptions mean that the marginal distribution of F_i, $i = 1, 2, \ldots, K$ is negative binomial.

$$Pr(F_i = x_i) = \frac{\Gamma(x_i + \alpha \cdot N)}{\Gamma(\alpha \cdot N) \cdot \Gamma(x_i + 1)} \cdot \frac{\beta^{x_i}}{(1+\beta)^{x_i + \alpha \cdot N}}$$

The α and β parameters satisfy the following equations.

$$E(\Pi_i) = \alpha \cdot \beta \qquad \text{and} \qquad var(\Pi_i) = \alpha \cdot \beta^2 \ .$$

Bethlehem et al. assume furthermore that

$$\sum_{i=1}^{K} E(\Pi_i) = 1 \ .$$

It implies that

$$K \cdot \alpha \cdot \beta = 1 \ .$$

The formula for the expected number of population uniques is given as

$$K \cdot Pr(F_i = 1) = N \cdot (1+\beta)^{-(1+\alpha \cdot N)} \ .$$

The α and β parameters can be estimated. Although it is computationally convenient, the Poisson-gamma model usually underestimates the number of population uniques, therefore its applicability is restricted.

Takemura (1999) introduced the Dirichlet-multinomial model. The probabilities of the cells are given by the

$$\pi = (\pi_1, \pi_2, \ldots, \pi_K)$$

vector. By assumption, the distribution of π is a Dirichlet distribution.

$$\pi \sim Dirichlet(\alpha_1, \alpha_2, \ldots, \alpha_K) \ .$$

Here $\alpha_1, \alpha_2, \ldots, \alpha_K$ are parameters and can be estimated using the sample frequencies ($f = (f_1, f_2, \ldots, f_K)$). The population frequencies follow a multinomial distribution.

$$F \sim Multinom(N, \pi_1, \pi_2, \ldots, \pi_K) \; .$$

Denote $\mathcal{A} = \sum_{i=1}^{K} \alpha_i$. The probability of a cell being a population unique, as given by Takemura, is

$$Pr(F_i = 1) = N \cdot \alpha_i \cdot \frac{\Gamma(\mathcal{A}) \cdot \Gamma(\mathcal{A} - \alpha_i + N - 1)}{\Gamma(\mathcal{A} + N) \cdot \Gamma(\mathcal{A} - \alpha_i)}$$

and the estimated number of population uniques is

$$N \cdot \frac{\Gamma(\mathcal{A})}{\Gamma(\mathcal{A} + N)} \cdot \sum_{i=1}^{K} \alpha_i \cdot \frac{\Gamma(\mathcal{A} - \alpha_i + N - 1)}{\Gamma(\mathcal{A} - \alpha_i)} \; .$$

While the Poisson-gamma model controls only the expected sum of the population frequencies, the Dirichlet-multinomial model provides exactly the required sum of population frequencies (N).

Skinner and Holmes (1993) provide a more general approach, substituting the gamma distribution of the Poisson-gamma model for an arbitrarily chosen g density function. The model assumes that F_i, $i = 1, 2, \ldots, K$ are independent Poisson distributed variables with respective rates λ_i, $i = 1, 2, \ldots, K$.

$$F_i | \lambda_i \sim Po(\lambda_i), \quad i = 1, 2, \ldots, K \tag{2.3.1}$$

Drawing the λ_i parameters from a log-normal distribution results in a better fit than the previously chosen gamma.

$$\log \lambda_i \sim N(\mu, \sigma^2) \tag{2.3.2}$$

or equivalently,

$$\log \lambda_i = \mu + \epsilon_i, \quad \text{where} \quad \epsilon_i \sim N(0, \sigma^2) \; . \tag{2.3.3}$$

The probability of being a population unique can be given by the following formula (Skinner and Holmes, 1998):

$$Pr(F_i = 1) = \frac{P_1}{\sum_{j=0}^{\infty} j P_j} \quad \text{where} \quad P_j = \int \frac{e^{-\lambda}\lambda^j}{j!} g(\lambda) d\lambda$$

Under Poisson sampling we can write (see Skinner and Holmes, 1998)

$$f_i|\lambda_i \sim Po(p\lambda_i) \quad \text{and} \quad F_i - f_i|\lambda_i \sim Po((1-p)\lambda_i) \quad i = 1, 2, \ldots, K \tag{2.3.4}$$

where f_i and $F_i - f_i$ are independent given λ_i. The probability of population uniqueness given sample uniqueness is also given in the article:

$$Pr(F_i = 1|f_i = 1) = \int \exp\left[-(1-p)\right] g(\lambda|f = 1) d\lambda \; ,$$

where $p = n/N$ is the sampling fraction and

$$g(\lambda_i|f_i = 1) = \frac{\lambda_i \cdot e^{-p\cdot\lambda_i} \cdot g(\lambda_i)}{\int \lambda \cdot e^{-p\cdot\lambda} g(\lambda) d\lambda} \; .$$

The (2.3.3) model is generalised as follows:

$$\log \lambda_i = \eta_i + \epsilon_i, \qquad \epsilon_i \sim N(0, \sigma^2) \; , \tag{2.3.5}$$

where $\eta_i = \mu_i + u_1 + u_2 + \ldots + u_l$, where u_j, $j = 1, 2, \ldots, l$ are main effects in a log-linear model. Under model (2.3.5)

$$g(\lambda_i) = (2\pi\sigma^2)^{-1/2} \lambda_i^{-1} \cdot \exp\left[-(\log \lambda_i - \eta_i)^2 / 2\sigma^2\right] \; .$$

In the article point estimates are given for μ and σ^2. The estimated values are denoted $\hat{\mu}_i$ and $\hat{\sigma}^2$. A disclosure risk measure given is

$$\hat{P}(F_i = 1 | f_i = 1) = \frac{\int \exp\left[-\lambda - (\log\lambda - \hat{\eta}_i)^2/2\hat{\sigma}^2\right] d\lambda}{\int \exp\left[-p\lambda - (\log\lambda - \hat{\eta}_i)^2/2\hat{\sigma}^2\right] d\lambda} \tag{2.3.6}$$

If ϵ_i disappears and λ_i is estimated by $\hat{\mu}_i/p$, then

$$\hat{P}(F_i = 1 | f_i = 1) = \exp\left[-(1-p)\hat{\mu}_i/p\right] \ . \tag{2.3.7}$$

The (2.3.6) disclosure risk measure provides better estimates of the true values than the (2.3.7) measure according to the article.

Although population uniqueness is an important indicator of disclosure risk, and the measures based on that are well-established, attention also has been paid to higher frequencies. Elamir (2004) fixes the λ_i values and provides the expected number of cells where $\{F_i = s, f_i = r\}$, $r \leq s$ under the (2.3.1) and (2.3.4) assumptions. Fixing the λ_i values is equivalent to using a conditional probability space. Therefore we can rewrite Elamir's result as follows.

$$E\left[\sum_{i=1}^{K} I\left(F_i = s, f_i = r\right) \Big| \lambda_1, \lambda_2, \ldots, \lambda_K\right] = \sum_{i=1}^{K} \frac{(p\lambda_i)^r [(1-p)\lambda_i]^{s-r} \exp(-\lambda_i)}{r! \cdot (s-r)!} \tag{2.3.8}$$

Elamir combines (2.3.8) and a log-linear model to estimate the unknown F_i frequencies. The log-linear model is applied to the expected sample frequencies, denoted by $\mu_i = E(f_i)$, $i = 1, 2, \ldots, K$. Under (2.3.4) we obtain $\mu_i = p \cdot \lambda_i$, $i = 1, 2, \ldots, K$. The log-linear model is

$$\log \mu_i = x_i^T \beta \ , \tag{2.3.9}$$

where the x_i vector contains the main effect and interaction terms selected in the model. The conclusion of Elamir's work is that the model selection

is important in the estimation of the cell frequencies. A complex model might provide an unstable estimation for (2.3.8), while a too simple model might not capture the variation of μ_i.

The (2.3.9) model also appears in Elamir and Skinner (2006). A possible refinement of the model is also given by adding a random effect to the model.

$$\log \mu_i = x_i^T \beta + \varepsilon_i \ . \tag{2.3.10}$$

The ε_i term can allow for overdispersion. They develop and evaluate a record-level disclosure risk measure for both log-linear models. The record level disclosure risk measure introduced in the article is

$\theta_j = Pr(\text{correct match}|\text{unique match, the record falls in cell } c_j) =$
$E\left(\frac{1}{F_j}\Big| f_j = 1\right)$

For the model under (2.3.10), $\omega_j = \exp(\varepsilon_j)$ is assumed to follow a gamma distribution as follows.

$$g(\omega_j; v) = \frac{v^v}{\Gamma(v)} \cdot \omega_j^{v-1} \exp(-v\omega_j)$$

(2.3.4) is a fundamental assumption in the article. It is shown that

$$E\left(\frac{1}{F_j}\Big|\lambda_j\right) = \frac{1}{(1-p)\lambda_j} \cdot [1 - \exp(-(1-p)\lambda_j)]$$

Under the (2.3.9) model λ_j is fixed, therefore the above formula provides an expression for θ_j. If we consider the (2.3.10) model, then

$$\theta_j = \frac{p \cdot (\exp(x_j^T\beta) + v)}{(1-p) \cdot (\exp(x_j^T\beta)) \cdot v} \left[1 - \left(\frac{\exp(x_j^T\beta + v)}{p^{-1} \cdot \exp(x_j^T\beta) + v}\right)^v\right]$$

By estimating the λ_j, v and $\exp(x_j^T\beta)$ values we can get an estimation of

θ_j for each model. There is no evidence that the (2.3.10) model performs better than (2.3.9) according to the article. However, (2.3.9) is easier to compute.

2.3.2.2.3 Our Contribution

This thesis focuses on measuring attribute disclosure in frequency tables. There are only few disclosure risk measures defined for entire tables. The measure we introduce in Chapter 3 quantifies the disclosure risk for such tables. The analysis shows that entropy and conditional entropy can be applied not just to cells but also to entire tables. It is shown that the disclosure risk can be measured for both population based and sample based tables by using information theoretical expressions.

2.4 Protection Against Disclosure

Once the data are recognised as having high disclosure risk, a statistical institute should not release them in their original form. However, withholding such data contravenes the main objective of the statistical institute, which is data dissemination. Therefore, statistical disclosure control methods are applied to the data. Such methods either alter the data or reduce the information content and therefore reduce the disclosure risk. SDC methods often target the elimination of problematic cases from the data, for example by changing or suppressing a cell value of 1 in a frequency table.

2.4.1 SDC Methods for Tabular Data

A statistical institute may employ either pre-tabular or post-tabular SDC methods to protect tabular data.

Pre-tabular methods alter the underlying microdata before tabulation takes place. Multiple tables may be generated from the same protected

microdata set. Advantages of pre-tabular methods are

- that they need to be carried out only once and
- table margins are consistent across different tables.

However, a pre-tabular method will modify at least one of the table-spanning variables and produce different tables from those generated from the original microdata thereby it requires much more suppression/perturbation in order to protect all possible tables. Pre-tabular methods are also less transparent to users and more difficult to account for in statistical analyses than post-tabular methods.

SDC methods applied directly to tables generated from original microdata are referred to as post-tabular methods. It means that the tables are protected after tabulation. Tables need to be protected separately from each other, therefore it demands more effort than pre-tabular methods. Table cells might not be consistent or additive across different tables. This is both an irritation for analysts and can also increase risk inadvertently as the inconsistencies could enable an intruder to 'unpick' the disclosure control.

However, post-tabular methods are more transparent than pre-tabular methods and provide better protection since all cells will generally be affected. Post-tabular methods include, for example, cell suppression or rounding of table cells.

2.4.1.1 Pre-tabular SDC Methods

Global recoding is one of the pre-tabular SDC methods. Two or more categories of a certain variable (or more variables) of the microdata set are combined into one category. Global recoding aims to increase the frequencies for some category-combinations since the modified variable is not as detailed as before.[4]

[4]In theory the modified variable should be considered as a different variable from the original one since the range of the original variable differs from that of the modified variable. However, the two variables are often called the same.

The post-randomization method (PRAM) (see Gouweleeuw et al. (1998)) is used to alter categorical variables. If a variable has L categories, then a $\boldsymbol{P}$ transition matrix of size $L \times L$ is defined prior to the perturbation. Each row and each column corresponds to a category. The (i, j)th component of the matrix is

$$p_{ij} = Pr(\text{perturbed cell value is } j | \text{original cell value is } i) \qquad (2.4.1)$$

The data are perturbed according to the $\boldsymbol{P}$ matrix. In practice the p_{ii}, $i = 1, 2, \ldots, L$, probabilities are close to 1 to ensure that the perturbed data remain close to the original values.

Record swapping is also an SDC method for microdata and is often used as pre-tabular method. It selects a small proportion of pairs of records within control strata and exchanges the values of a variable (or more variables) between paired records. Frequency tables may be generated from the perturbed microdata. Record swapping is often applied to geographical variables. Geography is usually a hierarchical variable. Record swapping is carried out usually at lower level in order to preserve the distribution of data at higher level. More details about record swapping can be found in Shlomo (2007).

2.4.1.2 Post-tabular SDC Methods

Cell suppression is one of the oldest methods for protecting tabular data. For each table cell a sensitivity rule, for example the minimum frequency rule or the (n, k)-dominance rule, provides a disclosure risk measure. Cells that are of high disclosure risk according to the selected sensitivity measure are subject to 'primary suppression'. It means that the values of such cells are withheld. However, tabular data are usually released with marginal totals. Totals and internal cells might allow an intruder to calculate the value of a primarily suppressed cell. Therefore secondary suppression needs to be carried out on the table. Secondarily

suppressed cells prevent an intruder from determining the exact cell value of primarily (and secondarily) suppressed cells. Although cell suppression is an adequate method to protect tabular data, sometimes the number of suppressed cells is high, thus so is the information loss.

To lessen the information loss controlled tabular adjustment (CTA) has been proposed, see Dandekar and Cox (2002). Sensitive table cells are determined as for cell suppression. However, cells are not suppressed but modified. In order to preserve the additivity of tables, additional non-sensitive cells are also adjusted. CTA leads to less information loss than cell suppression and results in analytically more useful tables but it is not transparent and bias may result.

The cell suppression problem can be approached by linear programming. Sensitive table cells are determined again by a sensitivity measure (or more sensitivity measures). The values of sensitive cells (and probably some additional cells) are modified. The linear program ensures that the modified values are 'sufficiently far' from the original cell value. The aim is to provide a table that is different to the original table, has an acceptable level of disclosure risk measure, provides minimum information loss and preserves the marginal totals. Papers that deal with this problem are for example Fischetti and Salazar-González (2000), Fischetti and Salazar-González (2003), Salazar-González (2006) and Hernández-García and Salazar-González (2014).

The above mentioned methods are more typically used for magnitude tables. In the case of many frequency tables generated from a microdata set such as a census, cell suppression is not often used due to the need to provide consistency across tables. For frequency tables, a more common approach is rounding. Rounding is also a post-tabular method. Deterministic rounding is the simplest version. A base b is selected and cell values are rounded to the closest multiple of b. Random rounding assigns a probability distribution to the set of multiples of b for each cell. The distribution depends on the cell value. The value of the cell

is changed to a multiple of b according to the probability distribution. Often the probability mass is concentrated on at most two multiples of b. For example, random rounding to base 3 can be carried out as follows. If a cell value is a multiple of 3, it remains unaltered. If the remainder is 1 or 2 when dividing the cell value by 3, then we round it to the closest or second closest multiple of 3 with probability 2/3 or 1/3 respectively. Different cells in the table, including marginal cells, are rounded independently. Deterministic and random rounding may not result in additive tables, that is, internal cells may not add up to the marginal total. A variation called controlled rounding uses mathematical optimization programming and ensures that rounded cells add up to the rounded marginal cells (see Cox (1987), Cox and George (1989)).

Barnardisation (see Hundepool et al. (2012)) modifies the non-zero cell values by adding or subtracting 1 from the cell value with probability $\frac{1-p}{2}$ and $\frac{1-p}{2}$ respectively. p is a parameter and equals the probability that a (non-zero) cell value remains unperturbed.

2.5 Information Loss

Information loss is an important aspect of statistical disclosure control. Undoubtedly, application of SDC methods distorts the data and reduces the information content, hence original and disclosure controlled data might lead a user to different results and/or conclusions.

Information loss can be measured by various information loss measures. They quantify the information loss by taking into account the distortion or suppression between the original and protected data.

The importance of information loss measures stems from the so-called R-U (risk-utility) confidentiality map, introduced by Duncan et al. (2001). A data protector always should use the 'best' available SDC method. It reduces the disclosure risk measure to a required level, whilst keeping the information loss minimal. Without an information loss measure the latter

criterion cannot be properly evaluated, SDC methods could be compared only on the basis of disclosure risk measures. Obviously, not releasing data should always provide the least disclosure risk, but this conflicts with the aim of functional data dissemination. Put another way, withholding data causes the greatest information loss, therefore it should be avoided. Therefore, information loss must be measured. Domingo-Ferrer and Torra (2001) and Shlomo and Young (2006) and references therein provide some examples of information loss measures depending on the types of outputs.

The terms 'information loss' and 'utility' need to be distinguished from each other since they are often used in the statistical disclosure literature. Information loss is discussed above. If perturbed data provide similar analytical results to original data, then the utility of the perturbed data is high, otherwise it is low.

Data utility is measured in Purdam and Elliot (2007). The paper investigates how SDC methods affect the results of some research projects. The utility of disclosure controlled data is gained by the comparison of the results based on the original and perturbed data. The paper shows that utility can be measured only after the research on perturbed data is done. However, a data protector needs to rely on information loss measures, since he/she cannot measure the utility.

Chapter 3

Disclosure Risk and Information Loss in Population Based Frequency Tables

3.1 Introduction

This chapter deals with disclosure risk and information loss measures that can be applied to population based frequency tables. Such tables are assumed to include every individual in the population. The measures ensure the assessment of disclosure risk and give some ideas about the information loss of the perturbed data.

Since we focus on population based tables, the counts in the tables are based on complete enumeration. Hereinafter we call the table before SDC methods are employed the 'original table'. The table after the application of SDC methods is the 'perturbed table'. The statistical agency naturally has the advantage of being aware of both the original and the perturbed tables over the data user. This fact is essential to determine the information loss and also very important in measuring the

disclosure risk. As we put ourselves into the agency's point of view, our approach also takes both the original and perturbed tables into account.

The use of information theoretical formulas in statistical disclosure control evolves naturally. After the application of SDC methods to the data, the information loss needs to be measured to inform data users about the extent of bias in the perturbed table. The information loss can be measured on the basis of the entropy as it is discussed for both microdata and tabular data in Willenborg and de Waal (2001). The comparability of (conditional) entropy across different techniques might verify its application according to Willenborg and de Waal. For microdata they determine the conditional entropy of the original data with respect to the perturbed data for several SDC techniques, such as global recoding and data swapping. Obviously, the higher the information loss, the less preferred the SDC method. However, some authors (Oganian and Domingo-Ferrer (2003), Oganian et al. (2004)) argue that entropy, as a utility measure, sometimes can be misleading and might not express the information loss properly. Oganian and Domingo-Ferrer provide an example when the conditional entropy does not give a proper information loss measure for rounded tabular data.

As the risk and utility framework (Duncan et al., 2001) is a crucial concept in statistical disclosure control, the investigation of the entropy as a disclosure risk measure also has undergone a development. Entropy-based disclosure risk measures have been defined at cell level for tabular data (Oganian and Domingo-Ferrer (2003); Oganian et al. (2004)). The difference between these approaches and our investigation is that we define the disclosure risk directly for the entire census frequency table. Our attitude is led by the desire to develop a risk measure that can be calculated with relative ease and provides an overview of the overall disclosure risk. Prior to the definition of our disclosure risk measure we review some information theoretical aspects in the following sections. Most of the measures can be formalised as f-divergences. The theoretical

concept of the f-divergence together with an allusion to an information loss measure can be found in Csiszár (1967). f-divergences can express the deviation of two probability distributions. Considering the measures as f-divergences allows us to point out connections between different measures and in a few cases the properties of a specific measure are derived from the general investigation of f-divergences.

3.2 Notation

Although in this chapter we deal with post-tabular perturbation methods, i.e. the perturbation is carried out on the F frequency table, we need to define random variables on the set of individuals. According to the general definition, random variables are measurable functions, that is, functions between measurable spaces. If Σ_A is a σ-algebra on an arbitrary (non-empty) set A, then the pair (A, Σ_A) is a measurable space. Below we need the $(I, \mathcal{P}(I))$ and $(C, \mathcal{P}(C))$ measurable spaces. The

$$X : (I, \mathcal{P}(I)) \to (C, \mathcal{P}(C)) \tag{3.2.1}$$

variable will indicate where the individuals fall in the original frequency table. For example, $X(a_k) = c_i$ means that individual a_k contributes to cell c_i. Since we do not distinguish between the individuals of I, it is reasonable to assign the same probability to each individual. It would ensure that the distribution of the X variable is $\frac{F}{N} = \left(\frac{F_1}{N}, \frac{F_2}{N}, \ldots, \frac{F_K}{N}\right)$. An arbitrary (A, Σ_A) measurable space becomes a measure space as an arbitrary measure, say μ, is attached to it. A μ measure is a function on the Σ_A (sigma-)algebra.

$$\mu : \Sigma_A \to \mathbb{R} \, .$$

Hence, the (A, Σ_A, μ) triple is a measure space.

In our case the individuals should have the same probability, therefore

a measure that is related to the U_I uniform distribution should be used. However, U_I is *not* a measure on $\mathcal{P}(I)$ since it is defined on I and not on $\mathcal{P}(I)$.

$$U_I : I \to \mathbb{R} ,$$
$$U_I(a_1) = U_I(a_2) = \ldots = U_I(a_N) = 1/N .$$

While $\mathcal{P}(I)$ is a (sigma-)algebra, I is not. However, U_I can easily be extended to $\mathcal{P}(I)$ by the following definition. For an arbitrary $B \subseteq I$ subset define $\mu_{U_I}(B) = |B|/N$. By this definition,

$$\mu_{U_I} : \mathcal{P}(I) \to \mathbb{R} .$$

$(I, \mathcal{P}(I), \mu_{U_I})$ is a measure space and the distribution of X is indeed F/N.

The frequencies of the perturbed table are denoted $G = (G_1, G_2, \ldots, G_K)$. The sum of the perturbed frequencies is $M = \sum_{i=1}^{K} G_i$.

The generation of a frequency table can be referred to as the categorization of individuals into cells. (In the situation above the categorization is given by X.) The perturbed frequency table (G) can be also considered as the categorization of (imaginary) individuals into C. The number of individuals must be equal to M. We denote these imaginary individuals by $J = \{b_1, b_2, \ldots, b_M\}$. The situation is very similar to that for 'real' individuals. However, for each real individual the category where the individual falls is assumed to be known. We assume that a post-tabular SDC method is applied to F. The categorization of the b_k, $k = 1, 2, \ldots, M$ individuals should be given by a similar random variable to (3.2.1).

$$(J, \mathcal{P}(J)) \to (C, \mathcal{P}(C)) .$$

We attach the U_J distribution to the $(J, \mathcal{P}(J))$ measurable space. The distribution of the random variable should be $\left(\frac{G_1}{M}, \frac{G_2}{M}, \ldots, \frac{G_K}{M}\right)$. However, there may be more than one random variables with this property; the

image of any specific (imaginary) individual and therefore the variable is not uniquely determined by its distribution. We denote the set of variables of distribution $\left(\frac{G_1}{M}, \frac{G_2}{M}, \ldots, \frac{G_K}{M}\right)$ by Ω_G.

$$\Omega_G = \{Y : (J, \mathcal{P}(J)) \to (C, \mathcal{P}(C)) : \quad G_l = |\{b_k \in J : Y(b_k) = c_l\}|, l = 1, 2, \ldots, K\} .$$

Let the elements of Ω_G be $Y_1, Y_2, \ldots, Y_{|\Omega_G|}$. If we need to select only one of these variables, then we will omit the subscript and refer to the variable as Y.

3.3 Information Theoretical Properties

The basis of the disclosure risk and information loss measures we intend to investigate is mainly mathematical. Before we consider these mathematical expressions as disclosure risk measures, the basic properties are worth mentioning. The field of information theory and related areas are covered comprehensively in Cover and Thomas (2006). Entropy is one of the most important formulae. It is defined for random variables and depends on their distribution. In order to maintain generality, the information theoretical definitions below are defined for arbitrary X and Y random variables. Their distributions are $P = (p_1, p_2, \ldots, p_K)$ and $Q = (q_1, q_2, \ldots, q_K)$, respectively. However, while considering the general definitions, one can think of the X and Y variables as defined in Section 3.2 and can identify P as F/N and Q as G/M. The sum determining the entropy of X is as follows.

$$H(X) = -\sum_{i=1}^{K} p_i \cdot \log p_i \ . \tag{3.3.1}$$

As it can be seen, $H(X)$ depends exclusively on the distribution of X. Therefore it does not lead to confusion if the argument of the $H(\cdot)$ function is a distribution. In our case $H(P)$ will represent the same value

as $H(X)$.

If $p_i = 0$ for a certain i, the respective term in the sum will be considered 0, since $\lim_{x\to 0} x \cdot \log x = 0$.

One can see easily that $H(X) \geq 0$, since $-p_i \cdot \log p_i \geq 0$. Entropy is equal to 0 if the probability mass is concentrated on one point. In any other case entropy cannot be 0, because $-p_i \cdot \log p_i > 0$ if $0 < p_i < 1$. On the other hand, $H(X) \leq \log K$, as it will be proven below. For the uniform distribution equality holds, $H(U_C) = \log K$.

There is a natural generalisation of the entropy of a single variable to multivariate random variables. The joint entropy of two single variables will be sufficient for us, although the definition can be easily extended to any finite number of variables. Note that in order to define the entropy of the (X, Y) bivariate random variable, we have to assume that $I = J$.

$$H(X,Y) = -\sum_{i=1}^{K}\sum_{j=1}^{K} Pr(X = c_i, Y = c_j) \cdot \log Pr(X = c_i, Y = c_j) \; .$$

The conditional entropy of two variables is defined as follows.

$$H(X|Y) = -\sum_{j=1}^{K} Pr(Y = c_j) \cdot \sum_{i=1}^{K} Pr(X = c_i|Y = c_j) \cdot \log Pr(X = c_i|Y = c_j) \; . \tag{3.3.2}$$

It is well-known that $H(X|Y) \leq H(X)$.

The connection between entropy, joint entropy and conditional entropy can be formulated as

$$H(X,Y) = H(Y) + H(X|Y) \; . \tag{3.3.3}$$

If X and Y are independent, the (3.3.3) relationship will reduce to $H(X,Y) = H(X) + H(Y)$.

The utility measures we intend to investigate are in most of the cases f-divergences. The concept of f-divergences can be found in Csiszár

(1967) and Csiszár and Shields (2004). To define an f-divergence we need two probability distributions (P and Q) and an $f : \mathbb{R}^+ \to \mathbb{R}$ convex function. For the sake of convenience we assume that $f(1) = 0$. The divergence between the two distributions determined by f is

$$D_f(P \parallel Q) = \sum_{i=1}^{K} q_i \cdot f\left(\frac{p_i}{q_i}\right) .$$

We also assume that $0 \cdot f\left(\frac{0}{0}\right) = 0$, $f(0) = \lim_{x \to 0} f(x)$, $0 \cdot f\left(\frac{c}{0}\right) = \lim_{x \to 0} x \cdot f\left(\frac{c}{x}\right)$. Using the convexity of f, we are able to derive an inequality that has several applications. Let $u_1, u_2, \ldots, u_K$ and $v_1, v_2, \ldots, v_K$ be positive real numbers with sums $u = \sum_{i=1}^{K} u_i$ and $v = \sum_{i=1}^{K} v_i$. Then the following inequality holds.

$$\sum_{i=1}^{K} v_i \cdot f\left(\frac{u_i}{v_i}\right) \geq v \cdot f\left(\frac{u}{v}\right) . \tag{3.3.4}$$

The proof of the above inequality is based on the convexity of f. According to Jensen's inequality $\sum_{i=1}^{K} \frac{v_i}{v} \cdot f\left(\frac{u_i}{v_i}\right) \geq f\left(\sum_{i=1}^{K} \frac{v_i}{v} \cdot \frac{u_i}{v_i}\right) = f\left(\frac{u}{v}\right)$. If f is strictly convex at $\frac{u}{v}$, then equality holds if and only if $u_i = \frac{u}{v} \cdot v_i$ for every i.

It follows immediately that $D_f(P \parallel Q) \geq 0$, since choosing $u_i = p_i$ and $v_i = q_i$ provides the inequality of

$$D_f(P \parallel Q) = \sum_{i=1}^{K} q_i \cdot f\left(\frac{p_i}{q_i}\right) \geq 1 \cdot f(1) = 0 .$$

Relative entropy or Kullback-Leibler divergence has a similar formulation to entropy but it serves the comparison of two distributions. The relative entropy of the P and Q distributions is defined as

$$D(P \parallel Q) = \sum_{i=1}^{K} p_i \cdot \log\left(\frac{p_i}{q_i}\right) .$$

Here $0 \cdot \log\left(\frac{0}{q_i}\right) = 0$, if $q_i > 0$. Relative entropy is considered infinity if $p_i > 0$ and $q_i = 0$ hold simultaneously for at least one i. It is easy to see that relative entropy is an f-divergence with $f(x) = x \cdot \log x$, that is, $D(P \parallel Q) = D_{x \cdot \log x}(P \parallel Q)$.

Relative entropy is not symmetric, as it can be seen from the formula. Triangle inequality also does not hold. Therefore $D(P \parallel Q)$ does not meet the criteria of distances. However, it is worth considering the relative entropy as how far two distributions are from each other. The second argument of the relative entropy (Q) can be taken as a reference value. The non-negativity of relative entropy follows from the non-negativity of f-divergences,

$$D(P \parallel Q) \geq 0 \ ,$$

with equality if and only if $p_i = q_i$ for all i.

With this inequality, $H(P) \leq \log K$ follows without much effort.

$$0 \leq D(P \parallel U_K) = \sum_{i=1}^{K} p_i \cdot \log\left(\frac{p_i}{1/K}\right) =$$
$$\sum_{i=1}^{K} p_i \cdot \log p_i + \sum_{i=1}^{K} p_i \cdot \log K = \log K - H(P) \ .$$

It is also obvious from the proof that equality holds if and only if the distribution is uniform.

Note that $D(Q \parallel P) = \sum_{i=1}^{K} q_i \cdot \log\left(\frac{q_i}{p_i}\right)$ is also an f-divergence with $f(x) = -\log x$.

To measure the distance between two distributions, L_p-norms also can be used. For an arbitrary vector $x = (x_1, x_2, \ldots, x_K) \in \mathbb{R}^K$ the L_p-norm $(1 \leq p < \infty)$ of x is defined as

$$\|x\|_p = \left(\sum_{i=1}^{K} |x_i|^p\right)^{\frac{1}{p}} \ .$$

Most importantly, L_2-norm is the Euclidean-norm. As p converges to infinity, $\|x\|_p$ tends to $\max_i |x_i|$. Therefore $\|x\|_\infty = \max_i |x_i|$ is referred to as the L_∞-norm of x. The distance of two distributions can be expressed as the L_p-norm of the difference: $\|P - Q\|_p$. If $p = 1$, this distance is equivalent to the f-divergence given by $f(x) = |x - 1|$. An L_p-norm produces a metric on $\mathbb{R}^K$, since

1. $\|x - y\|_p \geq 0$,
2. $\|x - y\|_p = 0 \iff x = y$,
3. $\|x - y\|_p = \|y - x\|_p$,

   trivially satisfy, and also

4. $\|x - y\|_p + \|y - z\|_p \geq \|x - z\|_p$

fulfils. The latter equation follows from the triangle inequality of the L_p-norm, which is proven in Serre (2002).

Let $p > 1$ and $f(x) = -\log x$. Selecting $u_i = p_i^p$ and $v_i = p_i$ results in the following inequality according to (3.3.4):

$$\sum_{i=1}^{K} p_i \cdot (-\log p_i^{p-1}) \geq -\log \sum_{i=1}^{K} p_i^p \ ,$$

or

$$H(P) \geq -\frac{1}{p-1} \cdot \log \|P\|_p^p = \frac{p}{p-1} \cdot \log \frac{1}{\|P\|_p} \ .$$

If $p \to \infty$, then we get

$$H(P) \geq \log \frac{1}{\max_i p_i} \ .$$

This inequality implies that the lower bound of $H(P)$ is higher as $\max_i p_i$ decreases. The lowest possible value for $\max_i p_i$ is $1/K$ (in case of a uniform distribution). $H(P)$ may decrease as $\max_i p_i$ increases.

The distance of $P = (p_1, p_2, \ldots, p_K)$ and U_C in L_2-norm is as follows.

$$\|P - U_C\|_2^2 = \sum_{i=1}^{K} (p_i - 1/K)^2 =$$
$$\sum_{i=1}^{K} p_i^2 - 2 \cdot \frac{1}{K} \sum_{i=1}^{K} p_i + K \cdot \left(\frac{1}{K}\right)^2 = \|P\|_2^2 - \frac{1}{K} \ .$$

The trivial consequence of this equation is $\|P\|_2^2 \geq \frac{1}{K}$. On the other hand,

$$\|P\|_2^2 = \sum_{i=1}^{K} p_i^2 \leq \sum_{i=1}^{K} p_i^2 + 2 \sum_{i<j} p_i \cdot p_j = \left(\sum_{i=1}^{K} p_i\right)^2 = \|P\|_1^2 = 1 \ .$$

For $P = (p_1, p_2, \ldots, p_K)$ and $Q = (q_1, q_2, \ldots, q_K)$ we denote $\sqrt{P} = (\sqrt{p_1}, \sqrt{p_2}, \ldots, \sqrt{p_K})$ and $\sqrt{Q} = (\sqrt{q_1}, \sqrt{q_2}, \ldots, \sqrt{q_K})$. They are not (necessarily) probability distributions, however, as vectors, their L_2-norms are 1.

One type of distance based on the L_2-norm is the Hellinger distance. This distance between P and Q is defined as

$$HD(P, Q) = \frac{1}{\sqrt{2}} \cdot \|\sqrt{P} - \sqrt{Q}\|_2 \ .$$

Obviously, $HD(P, Q) \geq 0$. On the other hand, $HD(P, Q) \leq 1$, since

$$HD(P, Q) = \frac{1}{\sqrt{2}} \cdot \sqrt{\sum_{i=1}^{K} (\sqrt{p_i} - \sqrt{q_i})^2} = \frac{1}{\sqrt{2}} \cdot \sqrt{\sum_{i=1}^{K} (p_i + q_i - 2 \cdot \sqrt{p_i \cdot q_i})} =$$
$$\frac{1}{\sqrt{2}} \cdot \sqrt{2 - 2 \cdot \sum_{i=1}^{K} \sqrt{p_i \cdot q_i}} = \sqrt{1 - \sum_{i=1}^{K} \sqrt{p_i \cdot q_i}} \leq 1 \ .$$

Suppose that $f(x) = 1 - \sqrt{x}$. Then $D_f(P \parallel Q) = \sum_{i=1}^{K} q_i \cdot f\left(\frac{p_i}{q_i}\right) = \sum_{i=1}^{K} q_i \cdot \left(1 - \sqrt{\frac{p_i}{q_i}}\right) = \sum_{i=1}^{K} (q_i - \sqrt{p_i \cdot q_i}) = 1 - \sum_{i=1}^{K} \sqrt{p_i q_i} = HD^2(P, Q)$. Therefore the Hellinger distance can also be considered as an f-divergence.

Assuming that $p_i > 0$ and $q_i > 0$ for all i, we can prove an inequality

between the Hellinger distance and the relative entropy. Applying the known $x - 1 \geq \log x$ $(x > 0)$ inequality to $\sqrt{\frac{q_i}{p_i}}$, we get

$$\sqrt{\frac{q_i}{p_i}} - 1 \geq \frac{1}{2} \cdot \log \frac{q_i}{p_i} \, ,$$

$$2 \cdot p_i \cdot \left(\sqrt{\frac{q_i}{p_i}} - 1 \right) \geq p_i \cdot \log \frac{q_i}{p_i} \, ,$$

$$2 \cdot \left(p_i - \sqrt{p_i \cdot q_i} \right) \leq p_i \cdot \log \frac{p_i}{q_i} \, .$$

Summing this inequality over i results in

$$\sum_{i=1}^{K} 2 \cdot \left(p_i - \sqrt{p_i \cdot q_i} \right) \leq \sum_{i=1}^{K} p_i \cdot \log \frac{p_i}{q_i} \, ,$$

which means

$$2 \cdot \left(1 - \sum_{i=1}^{K} \sqrt{p_i q_i} \right) \leq \sum_{i=1}^{K} p_i \log \frac{p_i}{q_i} \, ,$$

and thus

$$2 \cdot HD^2(P, Q) \leq D(P \parallel Q) \, .$$

Equality holds if and only if $\sqrt{\frac{q_i}{p_i}} - 1 = \frac{1}{2} \cdot \log \frac{q_i}{p_i}$ for every i. It means that $p_i = q_i$, i.e. $P = Q$. Since Hellinger distance is symmetric and the roles of p_i and q_i are commutable in the proof, we gain immediately a similar inequality.

$$2 \cdot HD^2(P, Q) \leq D(Q \parallel P) \, .$$

The distance of two distributions can be measured also by the chi-square statistic. Chi-square statistic is fundamental to test statistical hypothesis, since the independence of two random variables can be tested by this function. Chi-square statistic, like relative entropy, is not symmetric:

$$\chi(P,Q) = \sum_{i=1}^{K} \frac{(p_i - q_i)^2}{q_i} \ .$$

However, it is an f-divergence with $f(x) = (x-1)^2$, since $D_f(P \parallel Q) = \sum_{i=1}^{K} q_i \cdot f\left(\frac{p_i}{q_i}\right) = \sum_{i=1}^{K} q_i \cdot \left(\frac{p_i}{q_i} - 1\right)^2 = \sum_{i=1}^{K} \frac{(p_i - q_i)^2}{q_i}$.

The distance of an arbitrary P and U_C is as follows.

$$\chi(P,U_C) = \sum_{i=1}^{K} \frac{(p_i - 1/K)^2}{1/K} = K \cdot \sum_{i=1}^{K} \left(p_i - \frac{1}{K}\right)^2 = K \cdot \|P - U_K\|_2^2 =$$
$$K \cdot \left(\|P\|_2^2 - \frac{1}{K}\right) = K \cdot \|P\|_2^2 - 1$$

Chi-square statistic, as it can be found in Csiszár and Shields (2004), connects different f-divergences in the following sense. Assume that $q_i > 0$ for all i. If f is twice differentiable at $x = 1$ and $f''(1) > 0$, then

$$\frac{D_f(P \parallel Q)}{\chi(P,Q)} \to \frac{f''(1)}{2}$$

as $P \to Q$. (Here P converges pointwise to Q.)

The above convergence with $f(x) = x \cdot \log x$ yields that relative entropy and chi-square statistic can be "close" to each other: $D(P \parallel Q) \approx \frac{1}{2} \cdot \chi(P,Q)$. Also, with $f(x) = -\log x$ it follows that $D(Q \parallel P) \approx \frac{1}{2} \cdot \chi(P,Q)$. Finally, with $f(x) = 1 - \sqrt{x}$ we get that $HD^2(P,Q) \approx \frac{1}{8}\chi(P,Q)$ provided that P and Q are close enough.

The Hellinger distance has been used to measure the information loss, see for example Gomatam and Karr (2003), Gomatam et al. (2003), Gomatam et al. (2005) and Shlomo (2007). The motivation for its use in measuring

information loss is due to the following:

1. it is a metric (in the mathematical sense);
2. it provides larger impact when cell counts are small;
3. it can be calculated when the original count is zero which would not be the case when using relative distance metrics.

For these reasons, it is a good distance measure to assess deviations from original cell counts and the information loss resulting from perturbation. The Hellinger distance is a useful measure both to the statistical agency who needs to ensure that the perturbed data are fit for purpose as well as to the users who need an understanding of the impact of the perturbation in order to compensate for the distortion in their statistical analysis.

In the above paragraphs, the Hellinger distance has been formally defined as an information theoretic f-divergence and is part of the solution of using information theory for measuring both disclosure risk and information loss in frequency tables. Through the definition of the Hellinger distance as an f-divergence, the primary motivation for its use is now more transparent. The main advantage of the Hellinger distance over other f-divergences is its symmetry. A non-symmetric information loss measure might be more difficult to account for.

3.4 Formulae in the Context of Frequency Tables

This section describes how the above formulae can be analogously defined for population based frequency tables. The calculation of the disclosure risk and utility measures is easier if they are expressed by counts and there is no need to convert the counts into probability distributions. The division of the $(F_1, F_2, \ldots, F_K)$ cell counts by the N population size results in the $\left(\frac{F_1}{N}, \frac{F_2}{N}, \ldots, \frac{F_K}{N}\right)$ probability distribution, where the ith

probability shows the likelihood of an individual falling into the ith cell. The entropy of this distribution is

$$H\left(\frac{F}{N}\right) = -\sum_{i=1}^{K} \frac{F_i}{N} \cdot \log \frac{F_i}{N} = \frac{N \cdot \log N - \sum_{i=1}^{K} F_i \cdot \log F_i}{N} .$$

The relative entropy of the distributions of the original and perturbed frequency tables is

$$D\left(\frac{F}{N} \middle\| \frac{G}{M}\right) = \sum_{i=1}^{K} \frac{F_i}{N} \cdot \log\left(\frac{F_i/N}{G_i/M}\right) = \log \frac{M}{N} + \frac{1}{N} \cdot \sum_{i=1}^{K} F_i \cdot \log \frac{F_i}{G_i} .$$

The natural extension of the f-divergences is the application of the f function to frequency counts. The analogous formulae to relative entropy defined directly on the frequencies are

$$D(F \parallel G) = \sum_{i=1}^{K} F_i \cdot \log \frac{F_i}{G_i}$$

Unlike relative entropy defined on the distributions, $D(F\|G)$ can be negative when $N \neq M$. For example, if $0 < F_1 < G_1$ and $F_2 = G_2$, $F_3 = G_3$, $\ldots$, $F_K = G_K$, then $D(F\|G) < 0$. It can be seen that

$$D\left(\frac{F}{N} \middle\| \frac{G}{M}\right) = \log \frac{M}{N} + \frac{1}{N} \cdot D(F\|G) .$$

It implies that if $N = M$, then

$$D\left(\frac{F}{N} \middle\| \frac{G}{N}\right) = \frac{1}{N} \cdot D(F\|G) .$$

Therefore, if the perturbation method is unbiased, that is, the sum of the expected perturbed frequencies is equal to the original sum of frequencies, then the calculation of $D\left(\frac{F}{N} \middle\| \frac{G}{M}\right)$ and $D(F\|G)$ is basically the same.

The Hellinger distance of the distributions expressed by frequencies is

$$HD\left(\frac{F}{N},\frac{G}{M}\right)=\frac{1}{\sqrt{2}}\cdot\left\|\sqrt{\frac{F}{N}}-\sqrt{\frac{G}{M}}\right\|_2=\frac{1}{\sqrt{2NM}}\cdot\|\sqrt{MF}-\sqrt{NG}\|_2\ .$$

Being an f-divergence, the Hellinger distance also can be extended to frequencies without the division by the population size.

$$HD(F,G)=\frac{1}{\sqrt{2}}\cdot\|\sqrt{F}-\sqrt{G}\|_2\ .$$

where $\sqrt{F}=(\sqrt{F_1},\sqrt{F_2},\ldots,\sqrt{F_K})$ and $\sqrt{G}=(\sqrt{G_1},\sqrt{G_2},\ldots,\sqrt{G_K})$. We will use the above formula to measure the information loss.

3.5 Measuring Disclosure Risk Before Perturbation

3.5.1 The Disclosure Risk Measure

Having known the basic properties of the information theoretical expressions, we try to highlight how these formulas might be beneficial from statistical disclosure control point of view.

The proportion of small cells in the table is one of the indicators of disclosure risk. We consider a cell small, if its count is 1 or 2. Too many cells with small values in the frequency table may require the application of SDC methods.

Apart from the proportion of small cells, the average cell size (the number of contributors to the table divided by the number of cells, N/K) also measures the disclosure risk.

The entropy provides alternative opportunity to measure the disclosure risk. Small entropy in a row/column can indicate few non-zero cells in the row/column. The fewer the number of the non-zero cells, the more likely that disclosure occurs. However, relatively large entropy does not

exclude disclosure risk. Hence the entropy can be used as a risk measure but

$$1 - \frac{H(P)}{\log K} \tag{3.5.1}$$

can express the disclosure risk more effectively. We may recognise the relative entropy in this expression, as

$$D(P \parallel U_C) = \log K - H(P) = \log K \cdot \left(1 - \frac{H(P)}{\log K}\right) .$$

The $1 - \frac{H(P)}{\log K}$ formula is zero if P is the uniform distribution and is 1 if there is only one non-zero probability in P. In terms of frequencies, this risk measure can be expressed as $1 - \dfrac{N \cdot \log N - \sum_{i=1}^{K} F_i \cdot \log F_i}{N \cdot \log K}$. In most of the cases uniformly distributed frequencies in a row/column do not raise too much disclosure risk. On the other hand, if we know that the sum of a certain row/column is positive and entropy is zero (or equivalently, $1 - \frac{H(P)}{\log K} = 1$), then all the contributors of this row/column belong to one cell. Thus identity disclosure and group attribute disclosure are likely to occur. Entropy of a row/column in a frequency table can also be zero if there are only zeros in that particular row/column.

The disclosure risk measure we define in (3.5.2) is the convex combination of the proportion of the zeros in the table, the entropy-based term defined above and a third term depending on the N population size. We denote the set of cells with zero frequency in the original table by D. $\boldsymbol{w} = (w_1, w_2, w_3)$ is a vector of weights, $w_1, w_2, w_3 \geq 0$ and $w_1 + w_2 + w_3 = 1$.

$$R_1(F, \boldsymbol{w}) = w_1 \cdot \frac{|D|}{K} + w_2 \cdot \left(1 - \frac{N \cdot \log N - \sum_{i=1}^{K} F_i \cdot \log F_i}{N \cdot \log K}\right) - w_3 \cdot \frac{1}{\sqrt{N}} \cdot \log \frac{1}{e \cdot \sqrt{N}} \tag{3.5.2}$$

where e is the base of the natural logarithm. More justification for the

definition of the measure can be found in Section 3.9.

In the (3.5.2) formula we select three terms and take their convex combination. The entropy term is the most important disclosure risk measure because it measures attribute disclosure, however it is insufficient on its own because it does not take into account the number of zeros in the table and the population size of the table. The first and third terms in (3.5.2) support the entropy measure in that the first term accounts for the number of zeros and the third term the overall size of the table. Therefore, the convex combination reflects the properties we set out in the paper in Section 3.9.

The terms of the disclosure risk measure can also be evaluated separately from each other. In that case the

$$\left(\frac{|D|}{K}, 1 - \frac{H(X)}{\log K}, \frac{1}{\sqrt{N}} \cdot \log \frac{1}{e \cdot \sqrt{N}}\right)$$

triple is a disclosure risk measure. However, a statistical institute always has to decide whether a certain frequency table can be released. In order to decide about the release, the disclosure risk measure should be below a predefined threshold. In the case of the triple above, the statistical institute has to set three thresholds, one for each of the elements of the disclosure risk measure.

By choosing the weights of the three terms, a data protector can decide which term is the most important. We concentrate on attribute disclosure, therefore we propose that the highest weight be allocated to the second term. More discussion on determining the weights is in Section 3.5.2.

The three terms may be combined into other expressions by taking for example their geometric mean or root mean square. The geometric mean does not reflect the disclosure risk according to the properties stated in Section 3.9 since it can be zero if only one term is zero and the other terms are positive and may be large. The root mean square is a viable alternative as we show in Section 3.5.2.

We turn now to an elaboration of the third term of the disclosure risk measure in (3.5.2). This term, $-\frac{1}{\sqrt{N}} \cdot \log \frac{1}{e \cdot \sqrt{N}}$, is a monotonically decreasing function of N. It is in line with the fact that higher frequencies often imply lower disclosure risk than small frequencies. We also would like the term to be bounded by 0 and 1. There are other functions with these properties and below we list some of the functions that were investigated.

A natural choice would be $1/N$. However, this function decreases very quickly. If we consider

$$h_1(N, \varepsilon) = N^{-\varepsilon} \ ,$$

where $0 \leq \varepsilon \leq 1$, then we get a function that can decrease more slowly. To decrease the rate of decline even more we can use the

$$h_2(N, \varepsilon) = -N^{-\varepsilon} \cdot \log \left(e^{-1} \cdot N^{-\varepsilon}\right)$$

expression. The term used in the disclosure risk measure in (3.5.2) is $h_2(N, 0.5)$. The $\left(-\frac{1}{K} \sum_{i=1}^{K} \frac{1}{F_i} \log \frac{1}{e \cdot F_i}\right)$ function also naturally arises, provided that $F_i > 0$ for all i. The formula can be generalised as follows:

$$h_3(F, \varepsilon) = \left(-\frac{1}{K - |D|} \cdot \sum_{F_i > 0} \cdot \frac{1}{{F_i}^{\varepsilon}} \log \frac{1}{e \cdot {F_i}^{\varepsilon}}\right) \ .$$

The above functions are calculated for some small examples of tables in Appendix A.1.

Regarding the $h_3(N, \varepsilon)$ function, it obtains a value of 1 if the frequency table consists of all zeroes except for one and all ones: the $(0, 0, 0, \ldots, 0, 1)$ and $(1, 1, 1, \ldots, 1)$ frequency tables have the same $h_3(N, \varepsilon)$ value. This is a serious problem since it does not seem to be sensitive to the overall size of the population and we will not investigate this function further.

Regarding the two remaining functions evaluated at $\varepsilon = 0.1$: $h_1(N, 0.1)$ and $h_2(N, 0.1)$, these decrease slowly to zero (i.e., no risk) and tables of

large population sizes might have too large of a disclosure risk measure associated to them in spite of their low risk.

At the other extreme of $\varepsilon = 0.8$ or $\varepsilon = 1$, a $(1, 1, 1, 1, 1, 1, 1, 1, 1, 1)$ frequency table may not have its disclosure risk accurately reflected using $h_1(N, \varepsilon)$ since we obtain the following measures: $h_1(N, 0.8) = 0.1585$, $h_1(N, 1) = 0.1000$. It is clear that this value would be too low for a disclosure risk measure. On the other hand, $h_2(N, 0.8) = 0.4504$, $h_2(N, 1) = 0.3303$ and this more accurately reflects the disclosure risk in the table. In general, we see that the $h_2(N, \varepsilon)$ is preferable to $h_1(N, \varepsilon)$ since the former falls more slowly to zero (i.e., no risk).

Regarding the $h_2(N, \varepsilon)$ at different levels of ε, the disclosure risk is best reflected when ε is between 0.4 and 0.8 according to the data protector's preference. We choose $\varepsilon = 0.5$ because it is more straightforward to understand and calculate than other values of ε. This is the third term shown in (3.5.2).

3.5.2 The Choice of Weights Before Perturbation

The weights of the (3.5.2) disclosure risk measure can be chosen by the data protector. The weights can put emphasis on a selected term. However, even if the data protector opts for a term to be emphasised, the choice of the weights might remain problematic. Below we describe a method that attempts to find the highest possible disclosure risk. This method slightly changes the (3.5.2) risk measure. The modified risk measure differs from (3.5.2) in the weights. We continue to assume that $w_1, w_2, w_3 \geq 0$. However, now the weights will satisfy the following equation: $w_1^2 + w_2^2 + w_3^2 = 1$.

$$R_1^*(F, \boldsymbol{w}) = w_1 \cdot \frac{|D|}{K} + w_2 \cdot \left(1 - \frac{N \cdot \log N - \sum_{i=1}^{K} F_i \cdot \log F_i}{N \cdot \log K}\right) - w_3 \cdot \frac{1}{\sqrt{N}} \cdot \log \frac{1}{e \cdot \sqrt{N}} \,. \tag{3.5.3}$$

For the sake of simplicity we introduce

$$x_1 = \frac{|D|}{K} \ ,$$

$$x_2 = 1 - \frac{N \cdot \log N - \sum_{i=1}^{K} F_i \cdot \log F_i}{N \cdot \log K} \ ,$$

$$x_3 = -\frac{1}{\sqrt{N}} \cdot \log \frac{1}{e \cdot \sqrt{N}} \ ,$$

and the vector of these terms is

$$\boldsymbol{x} = (x_1, x_2, x_3) \ .$$

Here $0 \leq x_1 \leq 1$, $0 \leq x_2 \leq 1$ and $0 \leq x_3 \leq 1$.

Fix now the F table and consider the (3.5.3) risk measure as the function of $\boldsymbol{w}$. It means that we can allocate different weights to different tables. Fixing the table implies that $\boldsymbol{x}$ is constant. With the above notations

$$R_1^*(F, \boldsymbol{w}) = \boldsymbol{x} \cdot \boldsymbol{w} = x_1 \cdot w_1 + x_2 \cdot w_2 + x_3 \cdot w_3 \ .$$

According to the Cauchy-Schwarz inequality:

$$R_1^*(F, \boldsymbol{w}) = \boldsymbol{x} \cdot \boldsymbol{w} \leq \sqrt{\boldsymbol{x}^2} \cdot \sqrt{\boldsymbol{w}^2} =$$
$$\sqrt{x_1^2 + x_2^2 + x_3^2} \cdot \sqrt{w_1^2 + w_2^2 + w_3^2} = \sqrt{x_1^2 + x_2^2 + x_3^2} = \|\boldsymbol{x}\|_2 \ .$$

Equality holds if and only if

$$(w_1, w_2, w_3) = C \cdot (x_1, x_2, x_3)$$

with an appropriately chosen $C > 0$ constant. It is easy to see that this constant must be $C = \frac{1}{\sqrt{x_1^2+x_2^2+x_3^2}}$.

Hence we get the L_2-norm of $\boldsymbol{x}$ as a risk measure. While the (3.5.2) risk measure is bounded by 0 and 1, the bounds of (3.5.3) are 0 and $\sqrt{3}$. We prefer the risk measure to be bounded by 0 and 1, therefore one can use the following disclosure risk measure.

$$R_1^*(F) = \frac{\|\boldsymbol{x}\|_2}{\sqrt{3}} \tag{3.5.4}$$

Note that (3.5.4) depends only on the table, it is not necessary to choose weights.

3.6 Measuring Disclosure Risk After Perturbation

The disclosure risk in frequency tables should be measured not just before but also after perturbation. However, the perturbation method can make a significant difference to the disclosure risk assessment.

The expected disclosure risk measure of perturbed tables is lower than that of the original table since more uncertainty is introduced in the perturbed table. In the following sections we propose to modify the first and second terms of $R_1(F, \boldsymbol{w})$ in order to lower the disclosure risk. We assume that both the original and the perturbed tables are known when the disclosure risk is assessed.

3.6.1 Modifying the First Term of the Disclosure Risk Measure

Denote the set of cells of zero frequency in the perturbed table by E. We reduce $R_1(F, \boldsymbol{w})$ by replacing $\frac{|D|}{K}$ with

$$\left(\frac{|D|}{K}\right)^{\frac{|D\cup E|}{|D\cap E|}} . \tag{3.6.1}$$

If $D = \emptyset$ or $E = \emptyset$, then the first term of the disclosure risk measure will be considered 0. The (3.6.1) term is smaller than $\frac{|D|}{K}$ because $\frac{|D|}{K} \leq 1$ and $|D \cup E| \geq |D \cap E|$.

3.6.2 Modifying the Second Term of the Disclosure Risk Measure

3.6.2.1 The Modification

In order to reduce the entropy-based term of the disclosure risk measure, we will use the conditional entropy. The $H(X|Y) \leq H(X)$ relationship will help to decrease the second term of the disclosure risk measure.

In order to define the $H(X|Y)$ conditional entropy, we need to ensure that X and Y are defined on the same probability space, that is, $I = J$. It implies that $N = M$.

We need to express the $Pr(Y = c_j|X = c_i)$ conditional probabilities in order to determine the conditional entropy. These probabilities vary if we select different elements of Ω_G. The choice of the $Y \in \Omega_G$ variable is arbitrary. We can assume that an R_G probability distribution is given on Ω_G and we select the Y variable according to that. $R_G(Y)$ is the probability that Y is selected from Ω_G. Another option is as follows. Once the R_G distribution is given, the expectation of the $Pr(Y = c_j|X = c_i)$ probabilities can be calculated for every fixed i and j. For every i and j

we define a

$$Z_{ij}^G = Z_{ij} : (\Omega_G, \mathcal{P}(\Omega_G)) \to ([0,1], \mathcal{P}([0,1]))$$

random variable. By definition, $Z_{ij}(Y_l) = Pr(Y_l = c_j | X = c_i)$. The expectation is

$$E(Z_{ij}) = \sum_{l=1}^{|\Omega_G|} R_G(Y_l) \cdot Pr(Y_l = c_j | X = c_i) \; . \tag{3.6.2}$$

The first natural assumption is to not distinguish between the elements of Ω_G. It means that we assume a uniform distribution on Ω_G. Theorem 1 shows that in that case the (3.6.2) expectation depends mainly on the G frequency table.

Theorem 1. *If $R_G = U_{\Omega_G}$, then*

$$E(Z_{ij}) = \begin{cases} G_j/N & \text{if } F_i > 0 \; , \\ 0 & \text{if } F_i = 0 \; . \end{cases}$$

Proof. The proof can be found in the Appendices. □

As Willenborg and de Waal (2001) point out, the (3.3.2) formula can be rewritten as

$$H(X|Y) = -\sum_{i=1}^{K} \sum_{j=1}^{K} Pr(X = c_i) \cdot Pr(Y = c_j | X = c_i) \cdot \log \frac{Pr(X = c_i) \cdot Pr(Y = c_j | X = c_i)}{\sum_{k=1}^{K} Pr(X = c_k) \cdot Pr(Y = c_j | X = c_k)} \; . \tag{3.6.3}$$

Therefore in our case the conditional entropy, calculated on $E(Z_{ij})$, is

$$H(X|Y) = -\sum_{i=1}^{K}\sum_{j=1}^{K}\frac{F_i}{N}\cdot\frac{G_j}{N}\cdot\log\frac{\frac{F_i}{N}\cdot\frac{G_j}{N}}{\sum_{k=1}^{K}\frac{F_k}{N}\cdot\frac{G_j}{N}} = -\sum_{j=1}^{K}\frac{G_j}{N}\sum_{i=1}^{K}\frac{F_i}{N}\cdot\log\frac{F_i}{N} = H(X)\ .$$

This equation proves that the entropy cannot be lowered by the conditional entropy if $R_G = U_{\Omega_G}$. However, the risk measure should show difference between the original disclosure risk and the disclosure risk after perturbation. Therefore, we select a new R_G distribution and calculate the conditional entropy accordingly.

The new R_G assigns positive probability to a Y_l variable if the number of $a \in I$ individuals with $X(a) = Y_l(a)$ is maximal. Note that the F and G frequency tables are fixed. The criterion means that the highest number of individuals remain in the same cell 'after perturbation'. (Here by perturbation we mean the switch from X to Y. It represents a perturbation similar to pre-tabular methods.) If a $Y_l \in \Omega_G$ variable does not satisfy this criterion, then its probability is zero, $R_G(Y_l) = 0$. The positive probabilities probabilities are distributed uniformly.

For a fixed j consider the $X^{-1}(c_j) = \{a \in I : X(a) = c_j\}$ set of individuals. Obviously, $|X^{-1}(c_j)| = F_j$. The highest number of individuals in $X^{-1}(c_j)$ that can remain in the same c_j cell after perturbation is $\min(F_j, G_j)$. Therefore, the highest number of individuals in the population that can remain in the same cell is $\sum_{j=1}^{K}\min(F_j, G_j)$. Consequently, the new R_G distribution assigns positive probabilities to the variables of the following set, denoted by Ω_G^*.

$$\Omega_G^* = \left\{ Y_l \in \Omega_G : |\{a \in I : X(a) = Y_l(a)\}| = \sum_{j=1}^{K}\min(F_j, G_j) \right\}\ .$$

Our aim is to determine the (3.6.2) average using the new R_G distribution.

Theorem 2. *Assume that* $F \neq G$ *and*

$$R_G(Y_l) = \begin{cases} 1/|\Omega_G^*| & \text{if } Y_l \in \Omega_G^* \\ 0 & \text{if } Y_l \notin \Omega_G^* \end{cases}$$

Then

$$E(Z_{ij}) = \sum_{Y_l \in \Omega_G^*} R_G(Y_l) \cdot Pr(Y_l = c_j | X = c_i) =$$

$$\begin{cases} \dfrac{\min(F_i, G_i)}{F_i} & \text{if } i = j \text{ and } F_i > 0\ , \\ \dfrac{(F_i - \min(F_i, G_i)) \cdot (G_j - \min(F_j, G_j))}{F_i \cdot (N - \sum_{k=1}^{K} \min(F_k, G_k))} & \text{if } i \neq j \text{ and } F_i > 0\ , \\ 0 & \text{if } F_i = 0\ . \end{cases}$$

Proof. The proof can be found in the Appendices. □

We can calculate the conditional entropy again with the new $E(Z_{ij})$ values. The formula is given below. The proof can be found in the Appendices.

$$\begin{aligned} H(X|Y) = &- \sum_{i=1}^{K} \frac{\min(F_i, G_i)}{N} \cdot \log \frac{\min(F_i, G_i)}{G_i} - \\ &\sum_{i=1}^{K} \frac{F_i - \min(F_i, G_i)}{N} \cdot \log \frac{F_i - \min(F_i, G_i)}{N - \sum_{k=1}^{K} \min(F_k, G_k)} - \\ &\sum_{j=1}^{K} \frac{G_j - \min(F_j, G_j)}{N} \cdot \log \frac{G_j - \min(F_j, G_j)}{G_j}\ . \end{aligned} \tag{3.6.4}$$

3.6.2.1.1 An Example

To illustrate Theorems 1 and 2, consider the example below. Assume that $K = 3$ and the original and perturbed frequency tables are $F = (0, 2, 4)$ and $G = (0, 3, 3)$ respectively. Without loss of generality we can assume that $X(a_1) = X(a_2) = c_2$ and $X(a_3) = X(a_4) = X(a_5) = X(a_6) = c_3$. We can assume furthermore that $I = J$. The X and Y_l

Variable	Individuals (I)						
	a_1	a_2	a_3	a_4	a_5	a_6	Is the variable in Ω_G^*?
X	c_2	c_2	c_3	c_3	c_3	c_3	-
Y_1	c_2	c_2	c_2	c_3	c_3	c_3	yes
Y_2	c_2	c_2	c_3	c_2	c_3	c_3	yes
Y_3	c_2	c_2	c_3	c_3	c_2	c_3	yes
Y_4	c_2	c_2	c_3	c_3	c_3	c_2	yes
Y_5	c_2	c_3	c_2	c_2	c_3	c_3	no
Y_6	c_2	c_3	c_2	c_3	c_2	c_3	no
Y_7	c_2	c_3	c_2	c_3	c_3	c_2	no
Y_8	c_2	c_3	c_3	c_2	c_2	c_3	no
Y_9	c_2	c_3	c_3	c_2	c_3	c_2	no
Y_{10}	c_2	c_3	c_3	c_3	c_2	c_2	no
Y_{11}	c_3	c_2	c_2	c_2	c_3	c_3	no
Y_{12}	c_3	c_2	c_2	c_3	c_2	c_3	no
Y_{13}	c_3	c_2	c_2	c_3	c_3	c_2	no
Y_{14}	c_3	c_2	c_3	c_2	c_2	c_3	no
Y_{15}	c_3	c_2	c_3	c_2	c_3	c_2	no
Y_{16}	c_3	c_2	c_3	c_3	c_2	c_2	no
Y_{17}	c_3	c_3	c_2	c_2	c_2	c_3	no
Y_{18}	c_3	c_3	c_2	c_2	c_3	c_2	no
Y_{19}	c_3	c_3	c_2	c_3	c_2	c_2	no
Y_{20}	c_3	c_3	c_3	c_2	c_2	c_2	no

Table 3.6.1: Example: the X variable and Y_l variables

variables can be seen in Table 3.6.1.

Theorem 1 determines the expected conditional probability of $Pr(Y_l = c_j|X = c_i)$. For instance, in our example one of the conditional probabilities in the (3.6.2) average for $i = 2$ and $j = 2$ is

$$Pr(Y_1 = c_2|X = c_2) = \frac{Pr(Y_1 = c_2, X = c_2)}{Pr(X = c_2)} = \frac{\frac{|\{a_k \in I: X(a_k) = c_2, Y_1(a_k) = c_2\}|}{N}}{\frac{F_2}{N}} = \frac{|\{a_1, a_2\}|}{F_2} = \frac{2}{2} = 1\,.$$

The proof of Theorem 1 is based on the following idea, which is called 'double counting' in the literature. Assume that $i = 2$ and $j = 2$. The

point of Theorem 1 is to determine

$$E(Z_{22}) = \sum_{l=1}^{20} \frac{1}{20} \cdot Pr(Y_l = c_2 | X = c_2) = \frac{1}{20} \cdot \sum_{l=1}^{20} \frac{Pr(X = c_2, Y_l = c_2)}{Pr(X = c_2)} =$$

$$\frac{1}{20} \cdot \sum_{l=1}^{20} \frac{\frac{|\{a_k \in I : X(a_k) = c_2, Y_l(a_k) = c_2\}|}{N}}{\frac{F_2}{N}} =$$

$$\frac{1}{20 \cdot F_2} \sum_{l=1}^{20} |\{a_k \in I : X(a_k) = c_2, Y_l(a_k) = c_2\}| \ .$$

It implies that for each Y_l variable we need to count the number of individuals that fall into c_2 by both X and Y. Since only a_1 and a_2 are in c_2 by X, we need to count the number of c_2s in the first two columns of Table 3.6.1 (excluding the c_2s in the row of X). Each column has 10 c_2s, therefore there are 20 c_2s altogether. It means that $E(Z_{22}) = \frac{1}{20 \cdot 2} \cdot 20 = \frac{1}{2}$.

The proof of Theorem 2 is based on the same idea. However, the possible Y_l variables are limited, only Y_1, Y_2, Y_3 and Y_4 are taken into consideration in the example of Table 3.6.1.

3.6.2.2 Uneven Sums of Frequencies

In Section 3.6.2.1 we assumed that $I = J$ and $N = M$. In numerous cases the sum of the original frequencies is not equal to the sum of the perturbed frequencies, that is, $N \neq M$. In this section we extend our results to this situation.

We define first a new set of individuals, denoted by I'. This set will consist of $N \cdot M$ (imaginary) individuals. If $a \in I$, then I' will contain M 'copies' of a. We define the J' set similarly. If $b \in J$, the J' set of (imaginary) individuals will contain N individuals identical to b. It implies that the new frequency tables are $M \cdot F = (M \cdot F_1, M \cdot F_2, \ldots, M \cdot F_K)$ and $N \cdot G = (N \cdot G_1, N \cdot G_2, \ldots, N \cdot G_K)$. Note that $\sum_{i=1}^{K} M \cdot F_i = \sum_{j=1}^{K} N \cdot G_j = N \cdot M$ and the entropies of the new frequency tables are equal to those of the initial tables. It means that we can assume

that $I' = J'$ and calculate the conditional entropy as it is described in Section 3.6.2.1. Although the X, Y and Z_{ij} variables are different from those used in Section 3.6.2.1, we do not change the notation.

Theorems 1 and 2 can be rewritten as follows.

Theorem 3. *If $R_{N\cdot G} = U_{\Omega_{N\cdot G}}$, then*

$$E(Z_{ij}) = \sum_{l=1}^{|\Omega_{N\cdot G}|} R_{N\cdot G}(Y_l) \cdot Pr(Y_l = c_j | X = c_i) = \begin{cases} \dfrac{G_j}{M} & \text{if } F_i > 0\ , \\ 0 & \text{if } F_i = 0\ . \end{cases}$$

Proof. The proof is the same as that of Theorem 1. □

The (3.6.3) formula takes the following form.

$$H(X|Y) = -\sum_{i=1}^{K}\sum_{j=1}^{K} \frac{M\cdot F_i}{N\cdot M}\cdot\frac{N\cdot G_j}{N\cdot M}\cdot \log \frac{\frac{M\cdot F_i}{N\cdot M}\cdot\frac{N\cdot G_j}{N\cdot M}}{\sum_{k=1}^{K}\frac{M\cdot F_k}{N\cdot M}\cdot\frac{N\cdot G_j}{N\cdot M}} = -\sum_{j=1}^{K}\frac{G_j}{M}\cdot\sum_{i=1}^{K}\frac{F_i}{N}\cdot\log\frac{F_i}{N} = H(X)\ .$$

Theorem 4. *Assume that $F \neq G$ and*

$$R_{N\cdot G}(Y_l) = \begin{cases} 1/|\Omega^*_{N\cdot G}| & \text{if } Y_l \in \Omega^*_{N\cdot G}\ , \\ 0 & \text{if } Y_l \notin \Omega^*_{N\cdot G}\ . \end{cases}$$

Then

$$E(Z_{ij}) = \sum_{Y_l\in\Omega^*_{N\cdot G}} R_{N\cdot G}(Y_l)\cdot Pr(Y_l = c_j|X = c_i) =$$

$$\begin{cases} \dfrac{\min(M\cdot F_i, N\cdot G_i)}{M\cdot F_i} & \text{if } i = j \text{ and } F_i > 0\ , \\ \dfrac{(M\cdot F_i - \min(M\cdot F_i, N\cdot G_i))\cdot(N\cdot G_j - \min(M\cdot F_j, N\cdot G_j))}{M\cdot F_i\cdot(N\cdot M - \sum_{k=1}^{K}\min(M\cdot F_k, N\cdot G_k))} & \text{if } i \neq j \text{ and } F_i > 0\ , \\ 0 & \text{if } F_i = 0\ . \end{cases}$$

Proof. The proof is the same as that of Theorem 2. □

We can calculate the conditional entropy on $E(Z_{ij})$ again. The proof

of the following formula can be found in the Appendix.

$$H(X|Y) = -\sum_{i=1}^{K} \frac{\min(M \cdot F_i, N \cdot G_i)}{N \cdot M} \cdot \log \frac{\min(M \cdot F_i, N \cdot G_i)}{N \cdot G_i} -$$
$$\sum_{i=1}^{K} \frac{M \cdot F_i - \min(M \cdot F_i, N \cdot G_i)}{N \cdot M} \cdot \log \frac{M \cdot F_i - \min(M \cdot F_i, N \cdot G_i)}{N \cdot M - \sum_{k=1}^{K} \min(M \cdot F_k, N \cdot G_k)} -$$
$$\sum_{j=1}^{K} \frac{N \cdot G_j - \min(M \cdot F_j, N \cdot G_j)}{N \cdot M} \cdot \log \frac{N \cdot G_j - \min(M \cdot F_j, N \cdot G_j)}{N \cdot G_j} \tag{3.6.5}$$

This formula is the generalisation of (3.6.4).

3.6.2.3 The Perturbation Method

So far we have dealt with a fixed G perturbed frequency vector without taking the perturbation method into account. A post-tabular perturbation method assigns probabilities to potential perturbed tables given the F original table. In order to maintain generality, at this point we do not select a perturbation method, therefore the set of the potential perturbed tables consists of every integer vector of length K. We denote the set of potential perturbed tables by PG.

$$PG = \{G : G = (G_1, G_2, \ldots, G_K) \in \mathbb{Z}^K\} = \mathbb{Z}^K \ .$$

However, we assume that the perturbation method assigns non-zero probability to finite number of perturbed vectors. The probability assigned to the G table by the perturbation method will be denoted $T(G)$. It implies that $T(\cdot)$ provides a probability distribution on PG.

We have defined the Ω_G set for an arbitrarily chosen G perturbed table. Denote the (disjoint) union of these sets by Ω.

$$\Omega = \bigcup_{G \in PG} \Omega_G \ .$$

Assume that we have defined an $R_G : \Omega_G \to \mathbb{R}$ probability distribution for all G. It provides an $R : \Omega \to \mathbb{R}$ probability distribution as follows. If $Y \in \Omega$, then Y is an element of an Ω_G. By definition,

$$R(Y) = T(G) \cdot R_G(Y) \quad \text{if } Y \in \Omega_G \ .$$

R is a probability distribution on Ω, since

$$\sum_{Y \in \Omega} R(Y) = \sum_{G \in PG} \sum_{Y \in \Omega_G} T(G) \cdot R_G(Y) =$$
$$\sum_{G \in PG} T(G) \cdot \sum_{Y \in \Omega_G} R_G(Y) = \sum_{G \in PG} T(G) = 1 \ .$$

The Ω^* set can be defined as follows.

$$\Omega^* = \bigcup_{G \in PG} \Omega^*_G \ .$$

At this point it is inevitable to replace the M and Z_{ij} notations with M_G and Z^G_{ij}. It shows that they depend on the G perturbed table.

Theorems 3 and 4 can be extended as follows.

Theorem 5. *If $R_{N \cdot G} = U_{\Omega_{N \cdot G}}$ for all $G \in PG$, then*

$$\sum_{G \in PG} T(G) \cdot E(Z^G_{ij}) = \begin{cases} \sum_{G \in PG} T(G) \cdot \dfrac{G_j}{M_G} & \text{if } F_i > 0 \ , \\ 0 & \text{if } F_i = 0 \ . \end{cases}$$

Proof. This Theorem is the straightforward consequence of Theorem 3. □

In this case the conditional entropy might not be equal to the $H(X)$ entropy.

Theorem 6. *Assume that*

$$R_{N\cdot G}(Y_l) = \begin{cases} 1/|\Omega^*_{N\cdot G}| & \text{if } Y_l \in \Omega^*_{N\cdot G}\ , \\ 0 & \text{if } Y_l \notin \Omega^*_{N\cdot G}\ . \end{cases}$$

for all $G \in PG$. Then

$$\sum_{G\in PG} T(G)\cdot E(Z^G_{ij}) =$$

$$\begin{cases} \sum_{G\in PG} T(G)\cdot \dfrac{\min(M_G\cdot F_i, N\cdot G_i)}{M_G\cdot F_i} & \text{if } i=j \text{ and } F_i>0\ , \\ \sum_{G\in PG\setminus\{F\}} T(G)\cdot \dfrac{(M_GF_i-\min(M_GF_i,NG_i))\cdot(NG_j-\min(M_GF_j,NG_j))}{M_G\cdot F_i\cdot(N\cdot M_G-\sum_{k=1}^K \min(M_G\cdot F_k, N\cdot G_k))} & \\ & \text{if } i\neq j \text{ and } F_i>0\ , \\ 0 \qquad \text{if } F_i=0\ . & \end{cases}$$

Proof. It can be seen easily that $E(Z^F_{ii}) = 1$ and $E(Z^F_{ij}) = 0$ if $i \neq j$. Otherwise the proof can be derived from Theorem 4. □

The conditional entropy can be calculated on the formula given in Theorem 6.

3.6.2.4 The New Term of the Disclosure Risk Measure

To reduce the second term of the risk measure we exploit the $H(X|Y) \leq H(X)$ property. The risk measure before perturbation includes a $1-\frac{H(X)}{\log K}$ term. Similarly to this term, $1-\frac{H(X|Y)}{H(X)}$ is also bounded by 0 and 1. However, the latter expression might surpass $1-\frac{H(X)}{\log K}$. Thus, the product of the two terms, $\left(1-\frac{H(X|Y)}{H(X)}\right)\cdot\left(1-\frac{H(X)}{\log K}\right)$, will be included in our disclosure risk measure after perturbation.

3.6.3 The Disclosure Risk Measure After Perturbation

The risk measure we apply after perturbation also consists of three weighted terms.

$$R_2(F, G, \boldsymbol{w}) = w_1 \cdot \left(\frac{|D|}{K}\right)^{\frac{|D \cup E|}{|D \cap E|}} + w_2 \cdot \left(1 - \frac{H(X|Y)}{H(X)}\right) \cdot \left(1 - \frac{H(X)}{\log K}\right) - w_3 \cdot \frac{1}{\sqrt{N}} \cdot \log \frac{1}{e \cdot \sqrt{N}} \,. \quad (3.6.6)$$

3.6.4 The Choice of Weights After Perturbation

The choice of the weights after perturbation is a similar problem to that before perturbation. The method described in Section 3.5.2 can be applied to the (3.6.6) risk measure.

We introduce the following risk measure analogously to (3.5.3).

$$R_2^*(F, G, \boldsymbol{w}) = w_1 \cdot \left(\frac{|D|}{K}\right)^{\frac{|D \cup E|}{|D \cap E|}} + w_2 \cdot \left(1 - \frac{H(X|Y)}{H(X)}\right) \cdot \left(1 - \frac{H(X)}{\log K}\right) - w_3 \cdot \frac{1}{\sqrt{N}} \cdot \log \frac{1}{e \cdot \sqrt{N}} \,. \quad (3.6.7)$$

Here $w_1 \geq 0$, $w_2 \geq 0$, $w_3 \geq 0$ and $w_1^2 + w_2^2 + w_3^2 = 1$.

We expect the disclosure risk to be lower after perturbation than prior to that. Denote the first, second and third term of the (3.6.7) risk measure by y_1, y_2 and y_3 respectively.

$$y_1 = \left(\frac{|D|}{K}\right)^{\frac{|D \cup E|}{|D \cap E|}} \,,$$

$$y_2 = \left(1 - \frac{H(X|Y)}{H(X)}\right) \cdot \left(1 - \frac{H(X)}{\log K}\right) \,,$$

$$y_3 = -\frac{1}{\sqrt{N}} \cdot \log \frac{1}{e \cdot \sqrt{N}} \ .$$

Let the vector of these terms be $\boldsymbol{y} = (y_1, y_2, y_3)$. Define now the risk measure after perturbation as

$$R_2^*(F, G) = \frac{\|\boldsymbol{y}\|_2}{\sqrt{3}} \ .$$

Since $y_1 \leq x_1$, $y_2 \leq x_2$ and $y_3 = x_3$, the risk measure after perturbation is smaller than that before perturbation.

$$R_2^*(F, G) = \frac{\|\boldsymbol{y}\|_2}{\sqrt{3}} \leq \frac{\|\boldsymbol{x}\|_2}{\sqrt{3}} = R_1^*(F) \ .$$

Therefore our expectation that we should get a lower disclosure risk measure after perturbation is met.

3.6.5 A Note on the Disclosure Risk Measure and Pre-Tabular SDC Methods

The $R_2(F, G, \boldsymbol{w})$ disclosure risk measure is designed for post-tabular SDC methods. We assume that an R_G distribution is given on the Y_l, $l = 1, 2, \ldots, |\Omega_G|$ variables and the conditional entropy is calculated accordingly. This section is devoted to the relation of the conditional entropy, and therefore the disclosure risk measure to pre-tabular methods.

A pre-tabular method, by definition, modifies certain categories of few individuals. In other words, the original $X : I \to C$ categorization of individuals is changed to a $Y : I \to C$ variable. Note that in this case the variables share the same domain. The frequency table, generated from the perturbed microdata, may be released.

The original X categorization of individuals provides the $F = (F_1, F_2, \ldots, F_K)$ frequency table. Denote the frequency table that a Y variable provides by $G_Y = (G_{Y1}, G_{Y2} \ldots, G_{YK})$. It is now obvious that our way of thinking must be reversed with respect to post-tabular

methods. In case of post-tabular methods, the G perturbed table is given and we need to find the corresponding Y random variables. A pre-tabular method selects a Y random variable and the G_Y perturbed table depends on the variable.

Pre-tabular methods preserve the number of individuals of the microdata set, that is, $N = \sum_{i=1}^{K} F_i = \sum_{j=1}^{K} G_{Yj}$. (A post-tabular method might change the sum.) In order to maintain generality, we now do not exclude any random variable from the potential Y variables, denoted by Ω_{Pre}.

$$\Omega_{Pre} = \{Y | Y : I \to C \text{ is a random variable}\} .$$

Apparently, $\Omega_{Pre} \subseteq \Omega$, where Ω is defined in Section 3.6.2.3.

It is easy to see that for an arbitrarily given $G = (G_1, G_2 \ldots, G_K) \in \mathbb{Z}^K$ frequency vector, where $\sum_{j=1}^{K} G_j = N$, there exists at least one $Y \in \Omega_{Pre}$ that provides $G_Y = G$. Such variable is given for example by the following definitions.

$$\begin{aligned}
&Y(a_1) = Y(a_2) = \ldots = Y(a_{G_1}) = c_1, \\
&Y(a_{G_1+1}) = Y(a_{G_1+2}) = \ldots = Y(a_{G_1+G_2}) = c_2, \\
&\ldots, \\
&Y(a_{G_1+G_2+\ldots+G_{K-1}+1}) = Y(a_{G_1+G_2+\ldots+G_{K-1}+2}) = \ldots = Y(a_N) = c_K .
\end{aligned}$$

Therefore, the set of potential perturbed tables, denoted by PG_{Pre}, can be given as follows.

$$PG_{Pre} = \{G | G = (G_1, G_2 \ldots, G_K) \in \mathbb{Z}^K \text{ and } \sum_{j=1}^{K} G_j = N\} .$$

Obviously, $PG_{Pre} \subseteq PG$, where PG is a set introduced in Section 3.6.2.3.

A pre-tabular method, by definition, determines a probability distribution

on the Ω_{Pre} set. Denote it by R_{Pre}.

$$R_{Pre} : \Omega_{Pre} \to [0, 1] \ .$$

A data protector selects a $Y \in \Omega_{Pre}$ variable according to the R_{Pre} distribution and generates the G_Y frequency table. A particular pre-tabular method might not assign positive probability to each Y element of Ω_{Pre}. Consequently, there might be $G_Y \in PG_{Pre}$ tables that occur with zero probability after the application of the pre-tabular method.

If we select and fix now a particular $G \in PG_{Pre}$ frequency table, then we can follow a similar thought process to that we followed for post-tabular methods. We can define the set of variables that provides the fixed G table.

$$\Omega_{Pre,G} = \{Y \in \Omega_{Pre} | Y \text{ provides the } G \text{ frequency table}\} \ .$$

Apparently, Ω_{Pre} is the disjoint union of the above sets.

$$\Omega_{Pre} = \bigcup_{G \in PG_{Pre}} \Omega_{Pre,G} \ .$$

The R_{Pre} distribution naturally derives a distribution on each $\Omega_{Pre,G}$ set. It is basically a conditional distribution. Denote it by $R_{Pre,G} : \Omega_{Pre,G} \to [0, 1]$. By definition,

$$R_{Pre,G}(Y) = Pr(\text{we select } Y | Y \text{ is in } \Omega_{Pre,G}) = \frac{R_{Pre}(Y)}{\sum_{Y^* \in \Omega_{Pre,G}} R_{Pre}(Y^*)} \ .$$

Obviously, the objects in the above discussion of pre-tabular methods correspond to those in the description of post-tabular methods. The corresponding pairs are Ω_{Pre} and Ω, $\Omega_{Pre,G}$ and Ω_G, PG_{Pre} and PG, R_{Pre} and R, $R_{Pre,G}$ and R_G.

However, there is a major difference between the two situations. In case of a post-tabular method we do not know the R_G distribution on Ω_G,

therefore we need to make an assumption on it. Theorems 1, 2, 3, 4, 5 and 6 are all based on that assumption. A pre-tabular method, by its nature, assigns a probability to each $Y \in \Omega_{Pre}$ variable, therefore determines the $R_{Pre,G}$ distribution. This fact is easy to demonstrate on the PRAM method. The category of the a_i individual before perturbation is $X(a_i)$, $i = 1, , 2, \ldots, N$. Select and fix now a $Y \in \Omega_{Pre}$ variable. Assume that (before perturbation) the probability of changing $X(a_i)$ to $Y(a_i)$ as the result of the perturbation is $p_{a_i}(Y)$. Since the a_i individuals are perturbed independently from each other, the probability of getting exactly the Y variable is $R_{Pre,G}(Y) = \prod_{i=1}^{N} p_{a_i}(Y)$.

3.7 A Figure

Figure 3.1 attempts to clarify the relations of sets and variables used in the sections above. The main point of the figure is to show that in case of post–tabular methods we get the Y variables from the G perturbed table, while for pre-tabular methods the opposite is true.

3.8 Measuring Information Loss

The application of SDC methods evidently brings some information loss to the data. The role of information loss measures lies in expressing how much the loss is. In other words, how close the perturbed data are to the original data.

As mentioned in Section 3.3, we use the Hellinger distance to measure the utility of perturbed data.

In frequency tables it is not necessary to use the distributions when calculating the Hellinger distance. Similar formula can apply to the square root of the F and G frequencies instead of P and Q probabilities. In fact, Hellinger distance shows the magnitude of the cells since the difference between the square roots of two 'large' numbers are higher than in case of

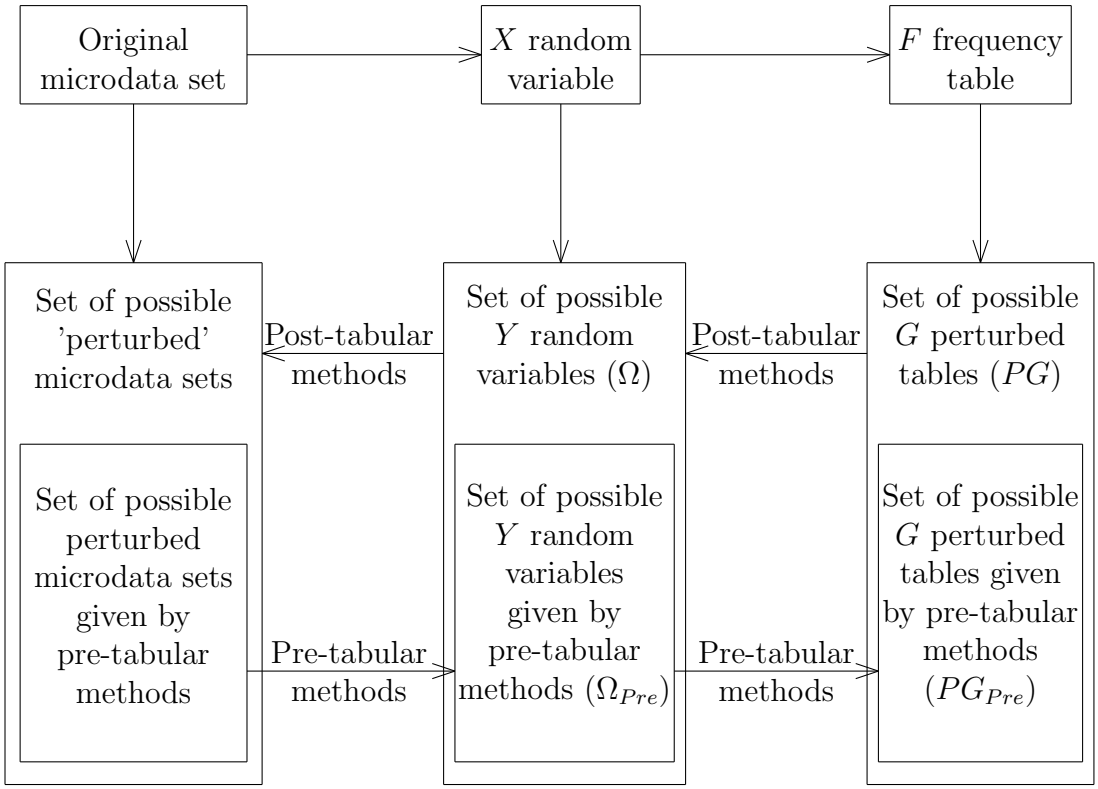

Figure 3.1: Relations of sets and variables

two 'small' numbers, even if these pairs have the same absolute difference.

$$HD(F,G) = \frac{1}{\sqrt{2}} \cdot \|\sqrt{F} - \sqrt{G}\|_2$$

Naturally, the lower bound is zero, while the upper bound of this distance of counts is $\sqrt{\frac{N+M}{2}}$ since

$$HD(F,G) = \frac{1}{\sqrt{2}} \cdot \|\sqrt{F} - \sqrt{G}\|_2 = \frac{1}{\sqrt{2}} \cdot \sqrt{\sum_{i=1}^{K} (\sqrt{F_i} - \sqrt{G_i})^2} =$$

$$\frac{1}{\sqrt{2}} \cdot \sqrt{\sum_{i=1}^{K} (F_i + G_i - 2 \cdot \sqrt{F_i \cdot G_i})} =$$

$$\frac{1}{\sqrt{2}} \cdot \sqrt{N + M - 2 \cdot \sum_{i=1}^{K} \sqrt{F_i \cdot G_i}} \leq \sqrt{\frac{N+M}{2}} \, .$$

If the perturbation method is unbiased, then the expected value of M is N. It means that the upper bound is approximately $\sqrt{N}$.

3.9 Paper: Measuring Disclosure Risk with Entropy in Population Based Frequency Tables

The paper below was submitted to the Privacy in Statistical Databases 2014 conference. It was published in the proceedings.

Measuring Disclosure Risk with Entropy in Population Based Frequency Tables

Laszlo Antal, Natalie Shlomo, and Mark Elliot

University of Manchester, UK
laszlo.antal@postgrad.manchester.ac.uk,
{natalie.shlomo,mark.elliot}@manchester.ac.uk

Abstract. Statistical agencies assess the risk of disclosure before releasing data. Unacceptably high disclosure risk will prevent a statistical agency from disseminating the data. The application of statistical disclosure control (SDC) methods aims to provide sufficient protection and make the data release possible. The disclosure risk of tabular data is typically quantified at the level of table cells. However, the evaluation of disclosure risk can require the assessment of the table as a whole, for example in the case of online flexible table generators. In this paper we use information theory to develop a disclosure risk measure for population-based frequency tables. The proposed disclosure risk measure quantifies the risk of attribute disclosure before and after an SDC method is applied. The new measure is compared to alternative disclosure risk measures developed at the Office for National Statistics.

Keywords: Information theory, attribute disclosure, conditional entropy.

1 Introduction

Statistical agencies follow strict confidentiality rules since releasing data always increases the risk of disclosure. They measure the risk of disclosure and apply statistical disclosure control (SDC) methods if the risk is unacceptably high. The subject of this paper is disclosure risk measurement in population-based frequency counts of tabular form.

Dislosure risk measures of tabular data usually express the risk at cell level. A regularly used disclosure risk measure for frequency counts is the so-called threshold rule. A cell is of high risk if the count does not exceed a certain value, for example 2.

The main objective of this paper is to measure the risk of attribute disclosure. Attribute disclosure happens if confidential information about an individual can be retrieved from the data. We use information theory to quantify the disclosure risk of population based frequency tables. The disclosure risk is expressed for the entire frequency table and for rows and columns of the frequency table. Single cells in themselves are not considered here. Information theory has been investigated in [5] to measure the disclosure risk of individual cells of magnitude tables. However, there has been no attempt to quantify the disclosure risk of an

J. Domingo-Ferrer (Ed.): PSD 2014, LNCS 8744, pp. 62–78, 2014.

entire (either frequency or magnitude) table by information theory. This paper provides a novel disclosure risk measure, which makes relatively quick disclosure risk assessment possible. The bases of the measure are entropy and conditional entropy.

Our aim is to develop a disclosure risk measure around the following properties.

Property 1A If only one cell is populated in the table, then the disclosure risk is high.

Property 1B Uniformly distributed frequencies imply low risk.

Property 2 Small cell values (i.e. ones and twos) are more disclosive than higher values. In general, the greater the cells, the lower the disclosure risk.

Property 3 Assume that two tables are given and there is only one cell populated in each table. The frequencies of the non-zero cells are equal. In this case we deem the table that has more cells (and therefore more zeroes) to be of higher disclosure risk.

Property 4 We would like the disclosure risk measure to be bounded by 0 and 1.

The motivation behind the properties is as follows. The risk of attribute disclosure is normally high if the population is concentrated in one cell, see [7]. It explains Property 1A. On the other hand, attribute disclosure is unlikely to occur if the frequencies are uniformly distributed, which drives Property 1B. The ground of Property 2 is the fact that revealing new information about a respondent becomes more difficult as the cell frequencies increase. The rationale behind Property 3 is that a table may be a more detailed version of another table, e.g. the breakdown of a table-spanning variable might be different in two tables. For example, if we replace super output area with output area, then the table will contain more detailed information. An intruder may obtain more information from more detailed tables. Property 4 is driven by the desire of comparing the disclosure risk of different tables.

Besides disclosure risk, information loss is also a crucial concept in statistical disclosure control. We use another information theory-related expression, Hellinger distance to measure the loss of information.

Although SDC methods provide protection to the data, a statistical agency might not be certain about the adequacy of the protection. Therefore, we assess the disclosure risk not just before but also after perturbation. The disclosure risk measures before and after perturbation are described in Section 2. Perturbation methods used for this study are outlined in Section 3. Section 4 discusses alternative disclosure measures that were used by the Office for National Statistics (ONS). Application of the theoretical results can be found in Section 5. A discussion closes our investigation in Section 6.

2 Disclosure Risk Measures and a Utility Measure

2.1 Before SDC Methods Are Applied

The most important information theoretical definition we use is entropy. Information theory is covered comprehensively in [2]. If X is a random variable with distribution $P = (p_1, p_2, \ldots, p_K)$, then the entropy of X is defined as

$$H(X) = -\sum_{i=1}^{K} p_i \cdot \log p_i \ . \tag{1}$$

Here log is the natural logarithm. If $p_i = 0$ for a certain i, then the respective term in the sum is considered 0.

Entropy is ideal to capture Properties 1A and 1B listed above since the value of entropy is 0 if and only if the P distribution can be written as $(0, \ldots, 0, 1, 0, \ldots, 0)$, and the value of entropy is maximal ($\log K$) if and only if P is uniform. Therefore, the expression $[1 - H(X)/\log K]$ exactly reflects Properties 1A, 1B and 4. However, entropy does not capture Properties 2 and 3 properly. The reason for this is given below.

The table of frequencies we investigate is denoted $F = (F_1, F_2, \ldots, F_K)$. The population size is $N = \sum_{i=1}^{K} F_i$. Consequently, the distribution of the table is

$$P = \left(\frac{F_1}{N}, \frac{F_2}{N}, \ldots, \frac{F_K}{N} \right) \ . \tag{2}$$

If we apply (1) to this distribution, we obtain

$$H(X) = \frac{N \cdot \log N - \sum_{i=1}^{K} F_i \cdot \log F_i}{N} \ .$$

Consider, for example, an $F = (F_1, F_2, F_3) = (0, 2, 4)$ frequency table. Then $P = \left(0, \frac{2}{6}, \frac{4}{6}\right)$ and $H(X) = \frac{6 \cdot \log 6 - 2 \cdot \log 2 - 4 \cdot \log 4}{6} = 0.6365$.

It can be seen from (1) and (2) that $H(X)$ depends only on the F_i/N, $i = 1, 2, \ldots, K$ ratios. Therefore, $[1 - H(X)/\log K]$ does not meet the expectations outlined in Properties 2 and 3. The entropy of F is the same as the entropy of $c \cdot F$, where $c > 1$ is a constant, contradicting Property 2. On the other hand, (1) shows that zeroes do not contribute to the value of entropy, therefore it does not reflect Property 3.

In order to compensate for Properties 2 and 3, the proportion of zeroes in the frequency table and an additional expression, based on N will be included in the disclosure risk measure. Denote the set of zeroes in the F table by D. The disclosure risk measure we define is a weighted average of three terms as follows. The weights are $\mathbf{w} = (w_1, w_2, w_3)$, where $w_1, w_2, w_3 \geq 0$ and $w_1 + w_2 + w_3 = 1$.

$$R_1(F, \mathbf{w}) = w_1 \cdot \frac{|D|}{K} + w_2 \cdot \left(1 - \frac{H(X)}{\log K}\right) - w_3 \cdot \frac{1}{\sqrt{N}} \cdot \log \frac{1}{e \cdot \sqrt{N}} \ . \tag{3}$$

Here e is the base of the natural logarithm. Each term is bounded by 0 and 1, and therefore so is the overall disclosure risk measure. The third term is a monotonically decreasing function of N, which reflects Property 2.

Considering the above example with $F = (0, 2, 4)$, we will obtain that $\frac{|D|}{K} = 0.3333$, $1 - \frac{H(X)}{\log K} = 1 - \frac{0.6365}{\log 3} = 0.4206$ and $-\frac{1}{\sqrt{N}} \cdot \log \frac{1}{e \cdot \sqrt{N}} = -\frac{1}{\sqrt{6}} \cdot \log \frac{1}{e \cdot \sqrt{6}} = 0.7740$.

If a frequency table consists of 1s only, that is, $F = (1, 1, \ldots, 1)$, then only the third term of (3) differs from 0. In this case the chance of attribute disclosure is low, since the number of zeroes is 0. The disclosure risk of $F = (1, 1, \ldots, 1)$ is also lower than that of $F = (10, 10, \ldots, 10)$, therefore monotonicity is maintained.

2.2 After SDC Methods Are Applied

The disclosure risk after SDC methods are applied to the table must also be assessed. The perturbed frequencies are denoted by $G = (G_1, G_2, \ldots, G_K)$ and their sum by $M = \sum_{j=1}^{K} G_j$. We assume that a statistical agency intends to release the G frequencies and withhold F. Therefore, we assume that $F \neq G$.

The disclosure risk after perturbation should be lower than that before perturbation, since an intruder has to encounter more uncertainty in G than in F. We adjust (3) in order to assess the disclosure risk after perturbation. The first and second terms of (3) are reduced in the new measure.

Denote the set of zeroes in G by E. The first term of (3) will be changed to

$$w_1 \cdot \left(\frac{|D|}{K} \right)^{\frac{|D \cup E|}{|D \cap E|}} .$$

If $D = \emptyset$ or $E = \emptyset$, then this term will be considered 0. This expression is not greater than $|D|/K$ and is still bounded by 0 and 1.

The second term of (3) will be multiplied by a factor, which depends on the conditional entropy. Assume that X and Y are two random variables with a common domain (I) and a common range ($C = \{c_1, c_2, \ldots, c_K\}$).

$$X : I \rightarrow C .$$

$$Y : I \rightarrow C .$$

The definition of the conditional entropy of X and Y is as follows.

$$H(X|Y) = -\sum_{j=1}^{K} Pr(Y = c_j) \cdot \sum_{i=1}^{K} Pr(X = c_i|Y = c_j) \cdot \log Pr(X = c_i|Y = c_j) .$$

In our case I is the set of individuals and C is the set of table cells (or categories). (Note that c_i is not the frequency of the cell.) X provides the categories where the individuals fall originally. Since we are dealing with the perturbation of frequency tables, the individuals and their categories might not be exactly followed after perturbation. Y should provide a similar categorisation to X after perturbation. More details about the Y variable can be found below.

It is well-known that $H(X|Y) \leq H(X)$. Roughly speaking, in our case X represents the original data and Y the perturbed data. $H(X|Y)$ expresses the uncertainty the X variable has if Y is known. Therefore we choose the second term of the disclosure risk measure after perturbation to be

$$w_2 \cdot \left(1 - \frac{H(X)}{\log K}\right) \cdot \frac{H(X|Y)}{H(X)} \ .$$

If $H(X) = 0$, then the second term of the disclosure risk measure is considered 0.

The conditional entropy can be rewritten using the $Pr(Y = c_j|X = c_i)$ probabilities, as it can be found in [8].

$$H(X|Y) = -\sum_{i=1}^{K}\sum_{j=1}^{K} Pr(X = c_i) \cdot Pr(Y = c_j|X = c_i) \cdot \log \frac{Pr(X = c_i) \cdot Pr(Y = c_j|X = c_i)}{\sum_{k=1}^{K} Pr(X = c_k) \cdot Pr(Y = c_j|X = c_k)} \ .$$

$Pr(X = c_i)$ in the above formula provides the probability that an individual falls in cell c_i in the original frequency table. It can be easily estimated by F_i/N.

The $Pr(Y = c_j|X = c_i)$ conditional probabilities will be expressed by $F = (F_1, F_2, \ldots, F_K)$ and $G = (G_1, G_2, \ldots, G_K)$. The formula we use is as follows.

$$Pr(Y = c_j|X = c_i) = \\ = \begin{cases} \dfrac{\min(M \cdot F_i, N \cdot G_i)}{M \cdot F_i} & \text{if } i = j \text{ and } F_i > 0 \ , \\ \dfrac{(M \cdot F_i - \min(M \cdot F_i, N \cdot G_i)) \cdot (N \cdot G_j - \min(M \cdot F_j, N \cdot G_j))}{M \cdot F_i \cdot (N \cdot M - \sum_{k=1}^{K} \min(M \cdot F_k, N \cdot G_k))} & \text{if } i \neq j \text{ and } F_i > 0 \ , \\ 0 & \text{if } F_i = 0 \ . \end{cases} \tag{4}$$

The complete justification for the (4) formula can be found in [1], we only outline the proof here.

The X random variable determines the cells where each individual falls in the original frequency table. Assume temporarily that $N = M$. The Y variable should provide the counterpart of X for the perturbed frequency table. It means that the individuals are recategorised in the perturbed table. However, the only requirement for Y is given by the $G = (G_1, G_2, \ldots, G_K)$ frequencies, the cell where a certain individual falls in the perturbed frequency table is not determined unambiguously. In case of a post-tabular SDC method, such as random rounding, Y is not (necessarily) uniquely defined. Different Y variables lead to different values of $Pr(Y = c_j|X = c_i)$. Instead of choosing one of the possible variables, we select a set of Y variables and calculate the average of the $Pr(Y = c_j|X = c_i)$ conditional probabilities in the set. If we took the average of the conditional probabilities over the entire set of possible Y variables, then the $H(X|Y)$ conditional entropy would not differ from $H(X)$. Consequently, the second term of the disclosure risk measure would not be lowered. Therefore, we take the average conditional probability of a narrower set of possible Y variables. In statistical disclosure control a general aim is to cause the least possible distortion to the data, therefore we select the Y variables that are as similar to X as possible. It means that an individual should fall in the same cell by X and Y, provided that the $G = (G_1, G_2, \ldots, G_K)$ frequencies allow that.

Table 1. Example: the X variable and possible Y variables

	Individuals (I)					
	1	2	3	4	5	6
X variable	c_2	c_2	c_3	c_3	c_3	c_3
First possible Y variable	c_2	c_2	c_2	c_3	c_3	c_3
Second possible Y variable	c_2	c_2	c_3	c_2	c_3	c_3
Third possible Y variable	c_2	c_2	c_3	c_3	c_2	c_3
Fourth possible Y variable	c_2	c_2	c_3	c_3	c_3	c_2
A Y variable not taken into account	c_3	c_2	c_2	c_2	c_3	c_3

Continuing with our example, where $F = (0, 2, 4)$, we can see that there are six individuals and three categories, that is, $C = \{c_1, c_2, c_3\}$. Assume that $G = (0, 3, 3)$. An X variable and possible Y variables are given in Table 1.

The variable in the last row of Table 1 is not taken into account when calculating the average conditional probability because the category of the first individual changes. It causes more distortion than necessary.

If $N \neq M$, then we can apply the same reasoning to the $M \cdot F = (M \cdot F_1, M \cdot F_2, \ldots, M \cdot F_K)$ and $N \cdot G = (N \cdot G_1, N \cdot G_2, \ldots, N \cdot G_K)$ frequency tables. The entropy of $M \cdot F$ is the same as that of F. The average conditional probability is given under (4).

To summarize, the disclosure risk measure after perturbation is

$$R_2(F, G, \mathbf{w}) = w_1 \cdot \left(\frac{|D|}{K}\right)^{\frac{|D \cup E|}{|D \cap E|}} + w_2 \cdot \left(1 - \frac{H(X)}{\log K}\right) \cdot \frac{H(X|Y)}{H(X)} - w_3 \cdot \frac{1}{\sqrt{N}} \cdot \log \frac{1}{e \cdot \sqrt{N}} \,. \tag{5}$$

2.3 A Utility Measure

Besides disclosure risk, information loss is also an important aspect of SDC. We measure that by a modified Hellinger distance, which is also related to information theory. Hellinger distance measures the divergence between two probability distributions, $P = (p_1, p_2, \ldots, p_K)$ and $Q = (q_1, q_2, \ldots, q_K)$. The definition of Hellinger distance is as follows.

$$HD(P, Q) = \frac{1}{\sqrt{2}} \cdot \sqrt{\sum_{i=1}^{K} (\sqrt{p_i} - \sqrt{q_i})^2} \,.$$

This expression is bounded by 0 and 1. We substitute P and Q for F and G respectively.

$$HD(F, G) = \frac{1}{\sqrt{2}} \cdot \sqrt{\sum_{i=1}^{K} (\sqrt{F_i} - \sqrt{G_i})^2} \,.$$

$HD(F, G)$ is the L_2-norm of the difference of $\sqrt{F} = (\sqrt{F_1}, \sqrt{F_2}, \ldots, \sqrt{F_K})$ and $\sqrt{G} = (\sqrt{G_1}, \sqrt{G_2}, \ldots, \sqrt{G_K})$ and therefore it is a metric. Hellinger distance

shows the magnitude of the cells since the difference between the square roots of two 'large' numbers are higher than in case of two 'small' numbers, even if these pairs have the same absolute difference. The lower bound of $HD(F,G)$ is 0, while the upper bound is $\sqrt{\frac{N+M}{2}}$.

In our example, where $F = (0,2,4)$ and $G = (0,3,3)$, the modified Hellinger distance is $HD(F,G) = \frac{1}{\sqrt{2}} \cdot \sqrt{(\sqrt{2}-\sqrt{3})^2 + (\sqrt{4}-\sqrt{3})^2} = 0.2940$.

3 Perturbation Methods

We place ourselves in the statistical agency's point of view and compare two perturbation methods. The perturbation methods we consider are random rounding to base 3 and record swapping.

Random rounding moves the frequencies to one of the multiples of 3 with certain probability structure. If a cell value is a multiple of 3, it remains unaltered. If the remainder is 1 or 2 when dividing the cell value by 3, then we round it to the closest or second closest multiple of 3 with probability 2/3 or 1/3 respectively. Different cells in the table, including marginal cells, are rounded independently. Random rounding may not result in additive tables, that is, the internal cells may not add up to the marginal total. In this paper we deal with internal cells only.

Record swapping is a pre-tabular method and as such, it is applied to the microdata. It selects some pairs of records and exchanges the values of a variable (or more variables) between paired records. Frequency tables may be generated from the perturbed microdata. However, if the table-spanning variables do not include at least one perturbed variable, then the frequency table generated from the perturbed microdata is the same as that generated from the original microdata. More details about record swapping can be found in [7]. In Section 5, we always include a perturbed variable in the table-spanning variables and consider the resulting $G = (G_1, G_2, \ldots, G_K)$ table as the perturbed frequency table. Although record swapping is a pre-tabular method and the Y variable and the $Pr(Y = c_j|X = c_i)$ conditional probabilities can be determined exactly, we use the (4) and (5) formulae to quantify the disclosure risk after perturbation. The reason for this is ease of computation since (4) and (5), and therefore the $H(X|Y)$ conditional entropy can be calculated on the F and G frequencies directly. There is no need to calculate the exact $Pr(Y = c_j|X = c_i)$ values, which can be computationally challenging since there are $K \times K$ such probabilities.

4 Alternative Disclosure Risk Measures

The Office for National Statistics (ONS) applied alternative disclosure risk measures to the 2001 UK census data in order to determine the best perturbation methods for tabular outputs of the 2011 UK census. The disclosure risk measures below were developed specifically for record swapping. The measure to express the degree of group attribute disclosure risk for rows was

$$GAD_1(F,G) = \frac{\sum I(\text{rows where all respondents fall into same category in the } F \text{ and } G \text{ tables})}{\sum I(\text{rows where all respondents fall into same category in the } F \text{ table})} .$$

Here $I(\cdot)$ is the indicator function. If a row in F has only one populated category, then it is counted as 1 in the numerator of $GAD_1(F,G)$ if the same category is the only populated cell in that row of G and the same individuals contribute to the category before and after perturbation.

The within group attribute disclosure for rows was measured by

$$WGAD_1(F,G) = \frac{\sum I\left(\begin{array}{c}\text{rows where all respondents fall into}\\ \text{same 2 categories in } F \text{ and } G \text{ (only 1 respondent in one)}\end{array}\right)}{\sum I\left(\begin{array}{c}\text{rows where all respondents fall into 2}\\ \text{categories in } F \text{ (only 1 respondent in one)}\end{array}\right)} .$$

The same features can be repeated as for $GAD_1(F,G)$.

The measures above may also be evaluated columnwise to obtain $GAD_2(F,G)$ and $WGAD_2(F,G)$.

$GAD_1(F,G)$ and $WGAD_1(F,G)$ express the proportion of rows where an intruder may correctly reveal a new attribute of an individual or more individuals. In case of $WGAD_1(F,G)$ the data protector assumes that the intruder may be the person who contributes to a cell with frequency 1.

Denote the set of cells having frequency 1 by D_1 and frequency 2 by D_2 in the original table, that is, $D_1 \cup D_2$ is the set of small cells in the table. The counterparts of these sets in the perturbed table are denoted E_1 and E_2 respectively. A third measure, which was also used by the ONS, is as follows.

$$DR(F,G) = \frac{|D_1 \cap E_1| + |D_2 \cap E_2|}{|E_1 \cup E_2|} .$$

The numerator is the number of small cells unchanged in the perturbed table, while the denominator is the number of small cells in the perturbed table. Therefore $DR(F,G)$ measures the proportion of small cells where the original and perturbed frequencies are equal.

The disclosure risk measures above were developed for the pre-tabular method of record swapping. In order to adapt them to post-tabular random rounding in our numerical study, we need to change the definitions slightly. In case of random rounding the individuals cannot be followed through in the microdata, therefore we cannot garantee that the same individuals contribute to a certain category before and after perturbation. Therefore, GAD will be changed as follows.

$$GAD_1^*(F,G) = \frac{\sum I(\text{rows where only one frequency is higher than 0 in the } F \text{ and } G \text{ tables})}{\sum I(\text{rows where only one frequency is higher than 0 in the } F \text{ table})} .$$

In the numerator the non-zero frequencies before and after perturbation are in the same category.

Similarly,

$$WGAD_1^*(F,G) = \frac{\sum I\left(\begin{array}{c}\text{rows where exactly two frequencies are higher}\\ \text{than 0 in the } F \text{ and } G \text{ tables and at least one of them is 1}\end{array}\right)}{\sum I\left(\begin{array}{c}\text{rows where exactly two frequencies are higher}\\ \text{than 0 in the } F \text{ table and at least one of them is 1}\end{array}\right)}.$$

The non-zero categories are the same in F and G if the row is counted in the numerator of $WGAD_1^*(F,G)$.

These measures can also be evaluated columnwise and we obtain $GAD_2^*(F,G)$ and $WGAD_2^*(F,G)$ respectively.

The idea behind $GAD_1^*(F,G)$ and $WGAD_1^*(F,G)$ is similar to $GAD_1(F,G)$ and $WGAD_1(F,G)$. An intruder might correctly reveal a new attribute of an individual or more individuals if the same (one or two) cells are populated in the original and perturbed tables.

5 Numerical Results

5.1 Numerical Results for $R_1(F, \mathbf{w})$ and $R_2(F, G, \mathbf{w})$

The data we use is an extract from the 2001 UK census tables. The table-spanning variables for various tables include age, sex, output area, country of birth, mode of travel, religion. In this paper only two-dimensional tables are considered.

We investigate the output area × country of birth, output area × mode of travel, output area × sex and output area × religion tables, where only 10 output areas are taken into account. The population size is $N = 2449$. In case of the output area × mode of travel table the population is restricted to individuals between 16 and 74 years of age. As can be seen, each table includes output area as a table-spanning variable. It coincides with the practice followed by the ONS, since the geographical variable in their frequency tables is normally output area and is the swapping variable for record swapping.

The entropy-based term is the core of the diclosure risk measure, therefore we assign high weight to that term in $R_1(F, \mathbf{w})$. We use $\mathbf{w} = (w_1, w_2, w_3) = (0.1, 0.8, 0.1)$.

The $R_2(F, G, \mathbf{w})$ disclosure risk depends on the G perturbed frequency table, therefore different perturbed tables provide different values of disclosure risk. In order to avoid an extreme value, we carry out the perturbation 1,000 times and take the average disclosure risk. This also reflects the perturbation method since more possible perturbed tables and their respective chance of being the outcome of the perturbation are taken into account. Random rounding and record swapping were carried out. Random rounding was applied to the frequency table, while record swapping to the output area variable of the microdata. 5% percent of the individuals were selected and paired with other individuals from distinct output areas, resulting in a total of 10% swapped individuals. The G frequency

table was generated on the perturbed microdata. The weights of $R_2(F, G, \mathbf{w})$ are unaltered compared to those before perturbation, $\mathbf{w} = (0.1, 0.8, 0.1)$.

An individual's attribute might be revealed using the rows or columns of a frequency table. Since the main point of this paper is measuring attribute disclosure, $R_1(F, \mathbf{w})$ and $R_2(F, G, \mathbf{w})$ are evaluated for each row, for each column and for the entire table. The F frequency tables, $R_1(F, \mathbf{w})$, $R_2(F, G, \mathbf{w})$ and $HD(F, G)$ for random rounding and record swapping can be found in the Appendix, see Tables 2, 3, 4 and 5.

The values of $R_1(F, \mathbf{w})$ reflect the Properties listed in Section 1 reasonably well. It can be observed that longer rows have higher disclosure risk. It might be attributed to the potentially higher number of zeroes in longer rows.

It can be seen that $R_2(F, G, \mathbf{w})$ is always substantially lower than $R_1(F, \mathbf{w})$.

The $R_2(F, G, \mathbf{w})$ disclosure risk measure of rows and columns for record swapping often shows slightly smaller values than for random rounding. This is attributed to the different methods of perturbation. While random rounding completely removes small cell values and frequencies that are not multiples of 3, and therefore it might change the distribution significantly, record swapping results in similar distribution to that of the original table. Record swapping also provides better information loss in numerous cases, especially in rows/columns where the majority of the counts is not higher than 10. Note that the disclosure risk of such rows/columns should not be low. Therefore, for rows and columns record swapping seems to be preferable to random rounding. However, the values of $R_2(F, G, \mathbf{w})$ for entire frequency tables are lower for random rounding compared to record swapping. On the other hand, the Hellinger distance is higher for random rounding compared to record swapping, reflecting higher information loss. The statistical agency must balance the disclosure risk against information loss.

5.2 Alternative Disclosure Risk Measures

For the alternative disclosure risk measures discussed in Section 4, random rounding and record swapping were carried out as described in the previous section. The frequency tables were perturbed 1,000 times and the average disclosure risk measures are shown in Table 6.

The value of $GAD^*(F, G)$ for random rounding is zero with one exception. The non-zero value is the result of column 5 of the output area $\times$ country of birth table. The value of $GAD^*(F, G)$ for that column is either 0 or 1 for each iteration.

In case of record swapping, $GAD(F, G)$ and $WGAD(F, G)$ are also zero with two exceptions. Each individual contributes to the same column before and after perturbation since only the output area variable is perturbed. (Consequently, each column has the same total before and after perturbation.) Therefore, columns 5 and 7 in the output area $\times$ country of birth table can be accounted for the two non-zero disclosure risk measures.

As it can be seen, $GAD(F, G)$ and $WGAD(F, G)$ are either 0 or 1 for each iteration. This fact might overestimate or underestimate the true risk. The disclosure risk measures defined under (3) and (5) provide more realistic measures.

6 Discussion

In this paper we have presented a new disclosure risk measure for population-based frequency tables. Information theoretical expressions, such as entropy and conditional entropy, are the focus of our investigation. We have demonstrated that they are able to quantify the risk of attribute disclosure both before and after the application of an SDC method.

The proposed disclosure risk measure can be applied to the entire frequency table and to rows and columns of the table. A statistical agency may set a threshold in order to decide whether a frequency table is safe to release or the application of an SDC method is required. We have used the Hellinger distance to measure the loss of information.

The entropy, the conditional entropy and therefore the whole disclosure risk measure can be expressed by the $F = (F_1, F_2, \ldots, F_K)$ original and $G = (G_1, G_2, \ldots, G_K)$ perturbed frequencies. This feature is particularly advantageous for post-tabular perturbation methods, where the category of a certain individual is not determined in the perturbed frequency table.

We compared our new disclosure risk measure with alternative disclosure risk measures. While $R_1(F, \mathbf{w})$ and $R_2(F, G, \mathbf{w})$ provide a disclosure risk measure for each row and column of the original and perturbed tables, $GAD(F, G)$ and $WGAD(F, G)$ use both F and G to evaluate the disclosure risk for entire tables. By applying $GAD(F, G)$ and $WGAD(F, G)$ the statistical agency automatically assumes that an SDC method should be applied to the frequency table. However, it is not always necessary. If $R_1(F, \mathbf{w})$ shows low disclosure risk, then the table might be released without perturbation. As we have seen, $GAD(F, G)$ and $WAGD(F, G)$ can show high disclosure risk if one row or column is of high risk and do not distinguish well between disclosure risk of different tables.

Although we have shown that $R_1(F, \mathbf{w})$ and $R_2(F, G, \mathbf{w})$ are preferable to $GAD(F, G)$ and $WGAD(F, G)$, further research is needed to reveal further properties of $R_1(F, \mathbf{w})$ and $R_2(F, G, \mathbf{w})$.

Acknowledgments. This work was funded by the ONS-ESRC PhD studentship (Ref. ES/J500161/1).

References

1. Antal, L., Shlomo, N., Elliot, M.: Measuring Disclosure Risk and Information Loss in Population Based Frequency Tables, `http://www.ccsr.ac.uk/publications/Measuring_Disclosure_Risk`
2. Cover, T.M., Thomas, J.A.: Elements of Information Theory, 2nd edn. Wiley, Hoboken (2006)
3. Domingo-Ferrer, J., Oganian, A., Torra, V.: Information-Theoretic Disclosure Risk Measures in Statistical Disclosure Control of Tabular Data. In: Proceedings of the 14th International Conference on Scientific and Statistical Database Management, Washington, pp. 227–231 (2002)

4. Duncan, G., Keller-McNulty, S., Stokes, S.: Disclosure Risk vs. Data Utility: the R-U Confidentiality Map. Technical Report LA-UR-01-6428, Statistical Sciences Group. Los Alamos National Laboratory, Los Alamos, N.M (2001)
5. Oganian, A., Domingo-Ferrer, J.: A Posteriori Disclosure Risk Measure for Tabular Data Based on Conditional Entropy. SORT-Statistics and Operations Research Transactions 27, 175–190 (2003)
6. Oganian, A., Domingo-Ferrer, J., Torra, V.: Internal Intrusion Scenarios in Inference Control of Tabular Databases. In: Information Processing and Management of Uncertainty in Knowledge-Based Systems (2004)
7. Shlomo, N.: Statistical Disclosure Control Methods for Census Frequency Tables. International Statistical Review 75, 199–217 (2007)
8. Willenborg, L., de Waal, T.: Elements of Statistical Disclosure Control. Lecture Notes in Statistics. Springer (2001)

Appendix

Table 2. Frequency table (F) and disclosure risk and utility measures: output area (10 output areas) $\times$ country of birth. The right lower corner shows the measures for the entire table, while the other measures are calculated rowwise/columnwise. 'Ran. Rou.' and 'Rec. Sw.' denote random rounding and record swapping respectively.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	$R_1(F,\mathbf{w})$	Ran. Rou. $R_2(F,G,\mathbf{w})$	Rec. Sw. Ran. Rou. $R_2(F,G,\mathbf{w})$	$HD(F,G)$	Rec. Sw. $HD(F,G)$
1	230	5	7	3	0	0	0	1	19	0	1	13	19	4	0	0	0	0.5847	0.1383	0.1183	1.0169	0.9030
2	154	1	2	8	0	5	0	1	5	0	3	1	11	1	4	3	0	0.5584	0.1720	0.1138	1.5053	0.7302
3	188	9	6	2	0	5	0	2	1	0	2	10	18	6	0	3	6	0.5125	0.0940	0.0981	1.2258	0.7001
4	278	3	1	1	0	1	0	1	8	0	0	3	8	8	0	0	0	0.7085	0.2142	0.1999	1.3384	0.9348
5	173	4	8	0	0	7	0	2	4	1	1	6	14	5	0	4	0	0.5472	0.1570	0.1094	1.2188	0.7532
6	161	10	1	2	0	1	1	1	9	1	1	7	13	0	7	3	1	0.5092	0.1497	0.0862	1.8428	0.8235
7	151	8	1	4	0	1	0	0	7	1	0	8	27	5	0	0	1	0.5422	0.1495	0.1133	1.3765	0.7425
8	208	7	1	3	1	9	1	0	2	6	1	5	29	3	1	2	1	0.5178	0.1249	0.0844	1.8226	0.7743
9	174	12	0	4	0	3	0	1	5	1	0	3	3	3	0	2	2	0.6095	0.1478	0.1326	1.2813	0.7545
10	171	5	0	4	0	4	0	1	8	0	0	6	10	8	2	3	1	0.5558	0.1538	0.1107	1.1956	0.7002
$R_1(F,\mathbf{w})$	0.0175	0.0874	0.2714	0.1435	0.9900	0.1744	0.7344	0.1844	0.1158	0.4915	0.3273	0.0930	0.0756	0.1309	0.5151	0.2185	0.3928	0.3242	-	-	-	-
Ran. Rou. $R_2(F,G,\mathbf{w})$	0.0110	0.0463	0.1295	0.0742	0.1893	0.0791	0.4139	0.1347	0.0538	0.3911	0.2192	0.0463	0.0322	0.0619	0.2747	0.1065	0.2733	-	0.0459	-	-	-
Rec. Sw. $R_2(F,G,\mathbf{w})$	0.0111	0.0440	0.0907	0.0660	0.1898	0.0671	0.1932	0.0879	0.0478	0.1687	0.1227	0.0454	0.0332	0.0592	0.1866	0.0916	0.1360	-	-	0.0469	-	-
Ran. Rou. $HD(F,G)$	0.1031	0.8224	1.4450	1.1257	0.6416	1.2123	0.9152	1.7978	1.0099	1.2946	1.4278	0.7659	0.3677	0.8118	0.9048	0.7980	1.4238	-	-	-	4.4800	-
Rec. Sw. $HD(F,G)$	0.2530	0.5382	0.6842	0.6123	0.1010	0.6995	0.2029	0.5109	0.5824	0.5384	0.5127	0.5483	0.5616	0.6330	0.6520	0.6166	0.5762	-	-	-	-	2.6307

Table 3. Frequency table (F) and disclosure risk and utility measures: output area (10 output areas) $\times$ mode of travel (age: 16-74). The right lower corner shows the measures for the entire table, while the other measures are calculated rowwise/columnwise. 'Ran. Rou.' and 'Rec. Sw.' denote random rounding and record swapping respectively.

	1	2	3	4	5	6	7	8	9	10	11	$R_1(F,\mathbf{w})$	Ran. Rou. $R_2(F,G,\mathbf{w})$	Rec. Sw. $R_2(F,G,\mathbf{w})$	Ran. Rou. $HD(F,G)$	Rec. Sw. $HD(F,G)$
1	8	11	7	1	44	5	3	3	47	0	126	0.3291	0.0533	0.0587	0.7576	0.5397
2	5	4	2	0	50	2	0	7	15	1	70	0.3670	0.0829	0.0792	1.1139	0.5386
3	3	1	8	0	18	5	0	8	26	0	122	0.4417	0.0916	0.1034	0.7442	0.5942
4	7	1	10	0	18	4	0	5	24	0	135	0.4536	0.0922	0.1064	0.7665	0.5979
5	3	2	4	0	17	6	0	2	26	0	107	0.4563	0.1000	0.1091	0.8289	0.5539
6	8	2	7	1	97	9	0	8	28	1	54	0.3157	0.0573	0.0674	1.1356	0.6101
7	5	0	8	0	29	2	1	3	14	0	88	0.4252	0.1019	0.0937	0.9364	0.5844
8	14	1	22	1	30	4	0	4	30	0	93	0.3214	0.0615	0.0674	0.9884	0.5623
9	10	0	8	1	23	2	0	0	17	2	78	0.3946	0.0996	0.0865	1.0818	0.6571
10	17	9	5	1	96	3	1	2	17	4	52	0.3003	0.0579	0.0617	1.1450	0.6392
$R_1(F,\mathbf{w})$	0.0850	0.2862	0.0944	0.3715	0.0927	0.0847	0.6206	0.1335	0.0474	0.5107	0.0309	0.2016	-	-	-	-
Ran. Rou. $R_2(F,G,\mathbf{w})$	0.0425	0.1363	0.0464	0.2868	0.0219	0.0536	0.4924	0.0655	0.0252	0.3283	0.0147	-	0.0295	-	-	-
Rec. Sw. $R_2(F,G,\mathbf{w})$	0.0415	0.0936	0.0427	0.1372	0.0262	0.0490	0.1996	0.0608	0.0258	0.1763	0.0153	-	-	0.0372	-	-
Ran. Rou. $HD(F,G)$	0.5190	1.4289	0.7948	1.4492	0.2165	1.1111	0.9151	0.9035	0.2844	1.1103	0.1221	-	-	-	3.1133	-
Rec. Sw. $HD(F,G)$	0.5292	0.6876	0.5249	0.3668	0.6663	0.4934	0.3480	0.5969	0.4666	0.4902	0.5434	-	-	-	-	1.9920

Table 4. Frequency table (F) and disclosure risk and utility measures: output area (10 output areas) $\times$ sex. The right lower corner shows the measures for the entire table, while the other measures are calculated rowwise/columnwise. 'Ran. Rou.' and 'Rec. Sw.' denote random rounding and record swapping respectively.

				Ran. Rou.	Rec. Sw.	Ran. Rou.	Rec. Sw.
	1	2	$R_1(F,\mathbf{w})$	$R_2(F,G,\mathbf{w})$	$R_2(F,G,\mathbf{w})$	$HD(F,G)$	$HD(F,G)$
1	161	141	0.0247	0.0222	0.0223	0.0376	0.1165
2	105	94	0.0276	0.0259	0.0260	0.0486	0.1259
3	142	116	0.0294	0.0237	0.0239	0.0612	0.1202
4	158	154	0.0220	0.0219	0.0219	0.0539	0.1213
5	139	90	0.0512	0.0252	0.0269	0.0398	0.1445
6	129	90	0.0434	0.0250	0.0265	0.0000	0.1292
7	107	107	0.0252	0.0252	0.0252	0.0660	0.1274
8	133	147	0.0243	0.0228	0.0229	0.0402	0.1343
9	98	115	0.0289	0.0253	0.0255	0.0666	0.1396
10	136	87	0.0529	0.0254	0.0273	0.0409	0.1432
$R_1(F,\mathbf{w})$	0.0170	0.0209	0.0150	-	-	-	-
Ran. Rou. $R_2(F,G,\mathbf{w})$	0.0127	0.0135	-	0.0100	-	-	-
Rec. Sw. $R_2(F,G,\mathbf{w})$	0.0128	0.0136	-	-	0.0100	-	-
Ran. Rou. $HD(F,G)$	0.1227	0.1032	-	-	-	0.1611	-
Rec. Sw. $HD(F,G)$	0.3452	0.3720	-	-	-	-	0.5076

Table 5. Frequency table (F) and disclosure risk and utility measures: output area (10 output areas) $\times$ religion. The right lower corner shows the measures for the entire table, while the other measures are calculated rowwise/columnwise. 'Ran. Rou.' and 'Rec. Sw.' denote random rounding and record swapping respectively.

	1	2	3	4	5	6	7	8	9	$R_1(F,\mathbf{w})$	Ran. Rou. $R_2(F,G,\mathbf{w})$	Rec. Sw. Ran. Rou. $R_2(F,G,\mathbf{w})$	$HD(F,G)$	Rec. Sw. $HD(F,G)$
1	181	0	0	1	17	1	1	83	18	0.4626	0.0634	0.0743	1.1356	0.5507
2	138	2	4	2	0	0	1	36	16	0.4973	0.1028	0.0906	1.0752	0.8832
3	130	0	0	0	22	4	1	61	40	0.3939	0.0742	0.0799	0.7054	0.4893
4	173	0	0	1	14	4	1	97	22	0.4403	0.0669	0.0710	0.9668	0.5323
5	142	2	5	0	15	6	1	37	21	0.3869	0.0568	0.0680	0.8948	0.4443
6	129	0	0	0	0	0	1	69	20	0.5460	0.1011	0.1195	0.6535	0.9705
7	118	2	0	2	24	9	1	38	20	0.3456	0.0562	0.0627	1.0288	0.4715
8	130	0	0	0	34	1	1	82	32	0.3974	0.0673	0.0808	0.9218	0.5466
9	148	3	0	0	0	2	1	38	21	0.5243	0.0894	0.1048	0.8468	0.8918
10	136	1	2	0	13	0	0	55	16	0.4692	0.0917	0.0908	0.8783	0.5312
$R_1(F,\mathbf{w})$	0.0152	0.3770	0.5763	0.4754	0.2029	0.2892	0.1166	0.0393	0.0404	0.2315	-	-	-	-
Ran. Rou. $R_2(F,G,\mathbf{w})$	0.0123	0.2235	0.2944	0.3282	0.0690	0.1304	0.0986	0.0178	0.0255	-	0.0327	-	-	-
Rec. Sw. $R_2(F,G,\mathbf{w})$	0.0124	0.1362	0.2085	0.1621	0.0508	0.1012	0.0766	0.0184	0.0259	-	-	0.0377	-	-
Ran. Rou. $HD(F,G)$	0.1176	1.1877	0.6261	1.2280	0.2664	1.1223	1.9488	0.1920	0.2832	-	-	-	2.9751	-
Rec. Sw. $HD(F,G)$	0.3553	0.4995	0.5804	0.3839	1.4187	0.7087	0.5697	0.4779	0.4832	-	-	-	-	2.2708

Table 6. Disclosure risk measures, $GAD(F,G)$, $WGAD(F,G)$ and $DR(F,G)$

		Random rounding			Record swapping		
Frequency table		$GAD^*(F,G)$	$WGAD^*(F,G)$	$DR(F,G)$	$GAD(F,G)$	$WGAD(F,G)$	$DR(F,G)$
output area	Rows	0	0	-	0	0	-
×	Columns	0.329	0	-	0.902	0.796	-
country of birth	Table	-	-	0	-	-	0.7009
output area	Rows	0	0	-	0	0	-
×	Columns	0	0	-	0	0	-
mode of travel	Table	-	-	0	-	-	0.7444
output area	Rows	0	0	-	0	0	-
×	Columns	0	0	-	0	0	-
sex	Table	-	-	0	-	-	0
output area	Rows	0	0	-	0	0	-
×	Columns	0	0	-	0	0	-
religion	Table	-	-	0	-	-	0.7004

Chapter 4

Disclosure Risk and Information Loss in Sample Based Tabular Data

4.1 Introduction

A census is always one of the most important data collections a statistical institute can conduct. A census may provide the most accurate and comprehensive data about the population. A complete enumeration delivers the broadest picture and takes every individual's characteristics into consideration. Therefore, numerous statistical institutes have gained experience in conducting censuses. However, statistical institutes tend to seek alternative solutions to the complete enumeration because there are other important factors to take into account. Data users expect the statistical institute to provide the data promptly. A complete enumeration might last long and also the cost of such enumeration might be high. This leads statistical institutes to investigate secondary or administrative data sources and conduction of surveys instead of complete enumeration. Their aim may be to substitute the traditional census for a new method based on administrative data and surveys. Before a decision is made about this

substitution, the institutes have to take the advantages and disadvantages of the old and new approaches into account. For example, they have to evaluate whether the change in the production of data can result in the same or higher quality and whether it reduces the cost significantly.

4.2 Notation for Sample Based Tabular Data

Tabular data are always important outputs of a census. The tabulation of the data might help to understand them better. Before the release of tabular data, the disclosure risk should be assessed. The disclosure risk of sample and population counts should be evaluated differently since sampling brings uncertainty to the data; a particular individual might or might not contribute to the sample-based table. In fact, sampling can be considered an SDC method.

We use the following notations. n individuals are sampled from the N individuals comprising the population, $n \leq N$. The sampled individuals are $I_S = \{b_1, b_2, \ldots, b_n\} \subseteq I$, where subscript S refers to the sample. The sampling fraction is denoted $p = n/N$. Sample frequencies are denoted by $f = (f_1, f_2, \ldots, f_K)$. It implies that $n = \sum_{i=1}^{K} f_i$. The set of cells having population frequency r is $C_r = \{c_i : F_i = r\}$. The analogous set in the sample-based table is $D_{S,r} = D_r = \{c_i : f_i = r\}$. The cardinalities of the sets are $N_r = |C_r|$ and $n_r = |D_r|$ respectively. Consequently, $N = \sum_{r=0}^{\infty} r \cdot N_r$ and $n = \sum_{r=0}^{\infty} r \cdot n_r$.

4.3 Disclosure Risk Measure for Sample Based Tables

We treat sampling as an SDC method. Therefore, the disclosure risk of a sample based frequency table can be given by (3.6.6), where G is substituted for f.

$$R_2(F, f, \boldsymbol{w}) = w_1 \cdot \left(\frac{|D|}{K}\right)^{\frac{|D\cup E|}{|D\cap E|}} + \\ w_2 \cdot \left(1 - \frac{H(X|Y)}{H(X)}\right) \cdot \left(1 - \frac{H(X)}{\log K}\right) - w_3 \cdot \frac{1}{\sqrt{N}} \cdot \log \frac{1}{e \cdot \sqrt{N}} \quad (4.3.1)$$

Here the Y variable is defined on I_S, while the form of X is the same as in (3.2.1).

$$Y : (I_S, \mathcal{P}(I_S)) \to (C, \mathcal{P}(C)) \, .$$

However, the situation is reversed compared to population based tables. In Chapter 3 we assumed that the individuals of the population are known and there are imaginary individuals in the perturbed frequency table. The X variable was known exactly, while Y was not.

In contrast with perturbed frequency tables, for sample based frequency tables $b_1, b_2, \ldots, b_n$ are not imaginary individuals and the categories where they fall are known. Now we assume that there is no information available on the individuals in $I \setminus I_S$. That is, in this chapter we assume that the Y variable is known exactly, while X is not. Therefore, we need to estimate the $(F_1, F_2, \ldots, F_K)$ population frequencies.

Although the F_i, $i = 1, 2, \ldots, K$ population frequencies are random variables and $N = \sum_{i=1}^{K} F_i$, we assume that the value of N is fixed and not random.

F_i, $i = 1, 2, \ldots, K$ is a random variable, meaning that $F_i : \Psi \to \mathbb{N}$ is a (measurable) function, where the Ψ domain can be given as

$$\Psi = \{\omega | \omega \text{ is a function}, \omega : \{a_1, a_2, \ldots, a_N\} \to \{c_1, c_2, \ldots, c_K\}\} \, .$$

With this notation

$$F_i(\omega) = |\{a \in I |\, \omega(a) = c_i\}| \, .$$

Consider the $\frac{F}{N} = \left(\frac{F_1}{N}, \frac{F_2}{N}, \ldots, \frac{F_K}{N}\right) : \Psi \to \mathbb{N}^K$ multivariate random variable. For a fixed $\omega \in \Psi$ the $\frac{F}{N}(\omega) = \left(\frac{F_1}{N}(\omega), \frac{F_2}{N}(\omega), \ldots, \frac{F_K}{N}(\omega)\right)$ vector is a probability distribution. If we denote $P_\omega = \frac{F}{N}(\omega)$, then $\mathcal{P} = \{P_\omega : \omega \in \Psi\}$ resembles a statistical model, where we need to find the 'true' ω given the sample.

4.4 Estimating the Population Frequencies

4.4.1 The Expectation of $F_i \cdot \log F_i$

Applying the (3.3.1) definition to the $\frac{F}{N}(\omega) = \left(\frac{F_1}{N}(\omega), \frac{F_2}{N}(\omega), \ldots, \frac{F_K}{N}(\omega)\right)$ distribution results in the following equation.

$$H(X) = -\sum_{i=1}^{K} \frac{F_i}{N}(\omega) \cdot \log\left(\frac{F_i}{N}(\omega)\right) = \frac{N \cdot \log N - \sum_{i=1}^{K} F_i(\omega) \cdot \log F_i(\omega)}{N} \,. \tag{4.4.1}$$

The estimation of $H(X) = H\left(\frac{F}{N}(\omega)\right)$ plays an important role in the estimation of $R_2(F, f, \boldsymbol{w})$. Since the N population size is assumed to be known, we need to estimate the $\sum_{i=1}^{K} F_i(\omega) \cdot \log F_i(\omega)$ term from the sample. The most natural approach is to substitute $F_i \cdot \log F_i$ for its expectation.

We know that under the (2.3.1) and (2.3.4) assumptions equation (2.3.8) satisfies. Below we provide a formula for $E(F_i \cdot \log F_i | f_i = r, \lambda_i)$ under the same assumptions.

For $x > 0$ introduce the $g(x) = x \cdot \log x$ function. Since $\lim_{x \to 0} g(x) = 0$, we can assume that $g(0) = 0$. With this notation we need to determine the expectation of $g(F_i)$ given f_i and λ_i.

$$E(g(F_i)|f_i = r, \lambda_i) = \sum_{s=0}^{\infty} Pr(g(F_i) = g(s)|f_i = r, \lambda_i) \cdot g(s) \,. \tag{4.4.2}$$

$g(0) = g(1) = 0$, thus the terms corresponding to $s = 0$ and $s = 1$ vanish in (4.4.2). The $g(x)$ function is strictly monotonically increasing on the $\left(\frac{1}{e}, \infty\right)$ interval. Hence,

$$Pr(g(F_i) = g(s)|f_i = r, \lambda_i) = Pr(F_i = s|f_i = r, \lambda_i) \ ,$$

provided that $s \geq 2$.

The conditional independence of f_i and $F_i - f_i$ implies that the conditional probability can be rewritten as follows.

$$Pr(F_i = s|f_i = r, \lambda_i) = \frac{Pr(F_i = s, f_i = r|\lambda_i)}{Pr(f_i = r|\lambda_i)} =$$
$$\frac{Pr(F_i - f_i = s - r, f_i = r|\lambda_i)}{Pr(f_i = r|\lambda_i)} = \frac{Pr(F_i - f_i = s - r|\lambda_i) \cdot Pr(f_i = r|\lambda_i)}{Pr(f_i = r|\lambda_i)} =$$
$$Pr(F_i - f_i = s - r|\lambda_i) \ .$$

According to (2.3.4) this probability is equal to

$$Pr(F_i - f_i = s - r|\lambda_i) = \frac{((1-p)\lambda_i)^{s-r}}{(s-r)!} \cdot e^{-(1-p)\lambda_i} \ .$$

Thus the (4.4.2) equation takes the following form.

$$E(F_i \cdot \log F_i|f_i = r, \lambda_i) = \sum_{s=r}^{\infty} \frac{((1-p)\lambda_i)^{s-r}}{(s-r)!} \cdot e^{-(1-p)\lambda_i} \cdot s \cdot \log s \ ,$$

or equivalently

$$E(F_i \cdot \log F_i|f_i = r, \lambda_i) = \sum_{s=0}^{\infty} \frac{((1-p)\lambda_i)^{s}}{s!} \cdot e^{-(1-p)\lambda_i} \cdot (s+r) \cdot \log (s+r) \ .$$

The applicability of this formula is restricted mainly due to the (2.3.1) and (2.3.4) assumptions. The value of $E(F_i \cdot \log F_i|f_i = r, \lambda_i)$ cannot be computed exactly by the formula, the sum of the above series can only be estimated by a finite $\sum_{s=0}^{L} \frac{((1-p)\lambda_i)^{s}}{s!} \cdot e^{-(1-p)\lambda_i} \cdot (s+r) \cdot \log (s+r)$

sum, where L can be chosen according to the required precision of the estimation.

4.5 Frequencies of Frequencies

4.5.1 Entropy and Frequencies of Frequencies

The entropy can be determined by the N_r, $r = 0, 1, 2, \ldots$ frequencies of frequencies as we will show below. In fact, $N_r : F(\Psi) \to \mathbb{N}$ is a function, where the domain is $F(\Psi) = \{F(\omega) |\, \omega \in \Psi\}$. The function is defined as

$$N_r(F(\omega)) = |\{i | F_i(\omega) = r,\ i = 1, 2, \ldots, K\}| \ .$$

Since $F = (F_1, F_2, \ldots, F_K)$ is a random variable, therefore so is N_r, $r = 0, 1, 2, \ldots$. However, we omit the rather cumbersome $N_r\left(F(\omega)\right)$ notation and write simply N_r because it will not lead to confusion. The (4.4.1) formula can be rewritten as follows:

$$H\left(X\right) = -\sum_{i=1}^{K} \frac{F_i}{N}(\omega) \cdot \log\left(\frac{F_i}{N}(\omega)\right) =$$
$$-\sum_{r=0}^{\infty} \sum_{i:\, c_i \in C_r} \frac{F_i}{N}(\omega) \cdot \log\left(\frac{F_i}{N}(\omega)\right) = -\sum_{r=0}^{\infty} N_r \cdot \frac{r}{\sum_{r=0}^{\infty} r N_r} \cdot \log \frac{r}{\sum_{r=0}^{\infty} r N_r} \ . \tag{4.5.1}$$

This equation ensures that the estimation of the N_r values provides an approximation of the entropy. Skinner and Shlomo (2012) provide an estimation for N_r. The fundamental assumption is (2.3.1) but λ_i has a gamma distribution with mean $E(\lambda_i) = \theta_1$ and variance $\mathrm{var}(\lambda_i) = \theta_1/\theta_2$. The compound distribution of the Poisson and gamma distributions is a negative binomial, therefore $F_i \sim NegBin(\frac{\theta_2}{\theta_1}, \theta_2)$. In order to introduce the result of the article mentioned above we need to follow its notations.

$$\hat{\mu}_1 = \frac{n}{K} \quad \text{and} \quad \hat{\mu}_2 = \frac{1}{K} \sum_{i=1}^{K} f_i \cdot (f_i - 1) \ ,$$

$$\hat{\theta}_1 = \frac{\hat{\mu}_1}{p} \quad \text{and} \quad \hat{\theta}_2 = p \cdot \frac{\hat{\mu}_1}{\hat{\mu}_2 - \hat{\mu}_1^2} \ .$$

The estimation of N_r is

$$\hat{N}_r = \frac{K \cdot \Gamma(r + \hat{\theta}_1 \cdot \hat{\theta}_2) \cdot \hat{\theta}_2^{\hat{\theta}_1 \cdot \hat{\theta}_2}}{r! \cdot \Gamma(\hat{\theta}_1 \cdot \hat{\theta}_2) \cdot (1 + \hat{\theta}_2)^{r + \hat{\theta}_1 \cdot \hat{\theta}_2}} \ , \tag{4.5.2}$$

where $\Gamma(\cdot)$ is the gamma function. This provides an estimation of the entropy through (4.5.1).

We can derive another natural estimator for N_r from (2.3.8) as follows.

$$\hat{N}_s = \sum_{r=0}^{s} E\left[\sum_{i=1}^{K} I\left(F_i = s, f_i = r\right) \Big| \lambda_1, \lambda_2, \ldots, \lambda_K \right] \ . \tag{4.5.3}$$

The sum expresses the expected number of cells where $F_i = s$. This estimation depends entirely on how the λ_i parameters and s are chosen, as it can be seen below.

$$\hat{N}_s = \sum_{r=0}^{s} \sum_{i=1}^{K} \frac{(p\lambda_i)^r [(1-p)\lambda_i]^{s-r} \cdot \exp(-\lambda_i)}{r! \cdot (s-r)!} =$$
$$\sum_{r=0}^{s} \frac{p^r \cdot (1-p)^{s-r}}{r! \cdot (s-r)!} \sum_{i=1}^{K} \lambda_i^s \cdot \exp(-\lambda_i) = \frac{(p + (1-p))^s}{s!} \sum_{i=1}^{K} \lambda_i^s \cdot \exp(-\lambda_i) =$$
$$\sum_{i=1}^{K} \frac{\lambda_i^s \cdot \exp(-\lambda_i)}{s!} = \sum_{i=1}^{K} Pr(F_i = s) \ .$$

The λ_i, $i = 1, 2, \ldots, K$ parameters should depend on the known f_i, $i = 1, 2, \ldots, K$ sample frequencies and the p sampling fraction. Skinner and Holmes (1998) estimate the λ_i parameters using a log-linear model similar to (2.3.10):

$$\log \lambda_i = x_i^T \beta + \varepsilon_i \ . \tag{4.5.4}$$

The ε_i term follows a normal distribution by assumption, $\varepsilon_i \sim N(0, \sigma^2)$.

The model includes main effect terms only. Thus, the $\hat{\mu}_i$ estimation of the expected sample frequencies is the product of n and the marginal proportions chosen according to the cell. The paper provides also the estimation of the σ parameter:

$$\hat{\sigma}^2 = \log \left\{ \frac{\sum_{i=1}^{K} \frac{f_i^2 - f_i}{\hat{\mu}_i^2}}{\sum_{i=1}^{K} \frac{f_i}{\hat{\mu}_i}} \right\} .$$

Finally, the estimation of $x_i^T \beta$ is given by

$$\log \frac{\hat{\mu}_i}{p \cdot \exp\left(\frac{\hat{\sigma}^2}{2}\right)} .$$

4.6 Estimating the Population Frequencies by Models

Surveys are easier and cheaper to conduct than a complete enumeration. Therefore, the estimation of population frequencies from samples is essential to gain a relatively precise picture of the population. The higher the sampling fraction, the closer the estimate can be.

A population based frequency table can be estimated from a sample based one by using sampling weights. Individuals selected in the sample represent those that are not selected. By using sampling weights we try to account for not selected individuals. However, this approach uses the attributes of selected individuals only. There might be a not observed individual whose attributes cannot be inferred. Consequently, the most difficult problem is the estimation of non-zero population frequencies that are zeroes in the sample based table. Models applied to sample based tables account for such frequencies. However, models might provide positive population estimates where the 'true' count is zero.

4.6.1 Log-linear Models

Log-linear models can be applied to frequency tables regardless of their dimensions. The number of potential log-linear models rises as the dimension of the table increases. Cell probabilities provided by log-linear models help to estimate the population frequencies and therefore the overall disclosure risk measure.

For the case of measuring disclosure risk, log-linear models can be used to estimate cell probabilities from sample frequencies and we are able to infer the population frequencies. Depending on the model, log-linear models will assign positive probabilities to cells where the sample frequencies are random zero and the population frequencies are positive and in addition, will assign a zero probability to population frequencies that are structural zeros. By applying log-linear models, we take into account the dependence between the table-spanning variables and model the contingency table for inference.

As mentioned in Section 2.3.2.2.2 and references mentioned therein, log-linear models appear in the SDC literature. For more discussion on using log-linear models in disclosure risk assessment, see also Skinner and Shlomo (2008).

Denote the estimated cell probabilities by $\hat{P} = (\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_K)$. The next sections outline the models on which we build some of our numerical results.

4.6.1.1 Multinomial Model

In this section we assume that the sample based table, as well as the population based table, follows a multinomial model.

$$f \sim Multinom(n; \hat{p}_1, \hat{p}_2, \ldots, \hat{p}_K)$$
$$F \sim Multinom(N; \hat{p}_1, \hat{p}_2, \ldots, \hat{p}_K)$$

As a consequence, the distribution of $F-f$ is also a multinomial distribution.

$$F - f \sim Multinom(N - n; \hat{p}_1, \hat{p}_2, \ldots, \hat{p}_K) \tag{4.6.1}$$

The advantage of this model is the fixed population size. The estimation of the population frequencies is straightforward as follows. $F-f$ can be generated as in (4.6.1) and the generated table can be added to f. N and n are assumed to be known.

4.6.1.2 Poisson Model

The Poisson model is formulated under (2.3.4). The λ_i, $i = 1, 2, \ldots, K$ parameters can be given as follows.

$$\lambda_i = N \cdot \hat{p}_i$$

The problem with the Poisson model is the not fixed sum of population frequencies (N). If we again generate the $F_i - f_i \sim Po(\lambda_i)$ frequencies and add them to the known f_i, then

$$\sum_{i=1}^{K} f_i + \sum_{i=1}^{K} (F_i - f_i) \sim \sum_{i=1}^{K} f_i + \sum_{i=1}^{K} Po(\lambda_i) \; .$$

The latter expression is not necessarily equal to the fixed N population size.

4.6.2 Pólya Urn Model

The Pólya urn model is discussed in Section 4.7.

4.7 Paper: Disclosure Risk Measurement with Entropy in Two-Dimensional Sample Based Frequency Tables

The paper below was submitted to the 'Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality 2015'.

Disclosure Risk Measurement with Entropy in Two-Dimensional Sample Based Frequency Tables

Laszlo Antal*, Natalie Shlomo*, Mark Elliot*

* University of Manchester, UK, email: laszlo.antal@postgrad.manchester.ac.uk, natalie.shlomo@manchester.ac.uk, mark.elliot@manchester.ac.uk

Abstract. We extend a disclosure risk measure defined for population based frequency tables to sample based frequency tables. The disclosure risk measure is based on information theoretical expressions, such as entropy and conditional entropy, that reflect the properties of attribute disclosure. To estimate the disclosure risk of a sample based frequency table we need to take into account the underlying population and therefore need both the population and sample frequencies. However, population frequencies might not be known and therefore they must be estimated from the sample. We consider two probabilistic models, a log-linear model and a so-called Pólya urn model, to estimate the population frequencies. Numerical results suggest that the Pólya urn model may be a feasible alternative to the log-linear model for estimating population frequencies and the disclosure risk measure.

1 Introduction

Statistical agencies measure the disclosure risk before releasing statistical outputs, such as frequency tables. This work discusses how information theoretical definitions, such as entropy and conditional entropy, can be employed to measure the disclosure risk in two-dimensional sample based frequency tables. A similar approach has been followed and a disclosure risk measure has been introduced in [1] for population based frequency tables. However, there has been no attempt to employ a similar disclosure risk measure to sample based tables. In this paper we show that the disclosure risk measure can be applied to two-dimensional sample based tables. The disclosure risk measure reflects the properties of attribute disclosure properly as set out in [1].

The population from which a sample is drawn may be known or unknown to the statistical agency. The disclosure risk assessment of a sample based table is more straightforward in the former case. If the population is unknown, the population frequencies can be estimated from the sample. We then use the estimated population based table to estimate the disclosure risk of the sample based table.

The outline of the paper is as follows. In Section 2 we introduce the notation we follow throughout the paper. Section 3 describes how the entropy and the conditional entropy can be applied to assess disclosure risk in tabular data. Section 4 presents the disclosure risk measure. Section 5 proposes two models for estimating the population frequencies and the disclosure risk measure when the population is unknown. A simulation study with numerical results can be found in Section 6, followed by a conclusion in Section 7.

2 Notation

The frequency tables we deal with have K cells. Table cells are denoted $C = \{c_1, c_2, \ldots, c_K\}$. The (potentially unknown) population based frequencies are $F = (F_1, F_2, \ldots, F_K)$, and their sample based counterparts are denoted $f = (f_1, f_2, \ldots, f_K)$. The population size and the sample size are $N = \sum_{i=1}^{K} F_i$ and $n = \sum_{i=1}^{K} f_i$, respectively. The set of individuals of the population is $I = \{a_1, a_2, \ldots, a_N\}$. The set of sampled individuals, denoted by $I_S = \{b_1, b_2, \ldots, b_n\}$, is a subset of the population, $I_S \subseteq I$.

In order to present our results, we need to introduce two random variables. The variables, X and Y, provide the classification of individuals into table cells for the whole population (X) and for the sampled individuals (Y).

$$X : I \to C \,,$$
$$Y : I_S \to C \,.$$

X is an extension of Y in the following sense. If we restrict X to I_S, then we will get Y, since $I_S \subseteq I$ and an individual in I_S is classified in the same table cell by X and Y. Note that X is not always known in practice.

Denote the distribution of X by $P = (p_1, p_2, \ldots, p_K) = \left(\frac{F_1}{N}, \frac{F_2}{N}, \ldots, \frac{F_K}{N}\right)$, while that of Y by $Q = (q_1, q_2, \ldots, q_K) = \left(\frac{f_1}{n}, \frac{f_2}{n}, \ldots, \frac{f_K}{n}\right)$.

Estimated population frequencies are referred to as $\hat{F} = (\hat{F}_1, \hat{F}_2, \ldots, \hat{F}_K)$.

3 Entropy and conditional entropy

The basis of the proposed disclosure risk measure is the entropy. The entropy of X is given as follows.

$$H(X) = -\sum_{i=1}^{K} Pr(X = c_i) \cdot \log Pr(X = c_i) = -\sum_{i=1}^{K} \frac{F_i}{N} \cdot \log \frac{F_i}{N} \qquad (1)$$

Note that $H(X)$ is never negative. It takes its maximum value if (and only if) F is uniform. The maximum value is $\log K$.

The entropy of Y may be defined similarly. Since (1) depends only on the F table, we sometimes refer to $H(X)$ as the entropy of F.

The conditional entropy of two variables also has an important role in our disclosure risk measure. Since the domain of X and that of Y are different, the conditional entropy of X and Y cannot be defined directly. In order to calculate the conditional entropy, we modify the variables.

First we define a new set of (imaginary) individuals, denoted by $\tilde{I}$ as follows. If we multiply F by n and f by N, then we get the $n \cdot F = (n \cdot F_1, n \cdot F_2, \ldots, n \cdot F_K)$ and $N \cdot f = (N \cdot f_1, N \cdot f_2, \ldots, N \cdot f_K)$ frequency tables. Note that the entropy of F is equal to that of $n \cdot F$ and the entropy of f is the same as that of $N \cdot f$. It is easy to see that $\sum_{i=1}^{K} n \cdot F_i = \sum_{i=1}^{K} N \cdot f_i = n \cdot N$. Therefore $n \cdot N$ imaginary individuals contribute to each table. We assume that the same imaginary individuals contribute to the two tables. This set of individuals is $\tilde{I}$. The two variables are

$$\tilde{X} : \tilde{I} \to C$$

and

$$\tilde{Y} : \tilde{I} \to C \; .$$

The conditional entropy, defined below, depends on the $Pr(\tilde{X} = c_i | \tilde{Y} = c_j)$ conditional probabilities. We have not defined the probabilities unambigously, since we have to define for each imaginary individual where the individual falls by both $\tilde{X}$ *and* $\tilde{Y}$.

We assume that $\tilde{X}$ and $\tilde{Y}$ are as 'similar' to each other as possible. This assumption means that the maximum possible number of individuals fall into the same category by $\tilde{X}$ and $\tilde{Y}$. For instance, if $n \cdot F_1 \leq N \cdot f_1$, then the $n \cdot F_1$ imaginary individuals that fall in c_1 by $\tilde{X}$ also fall in c_1 by $\tilde{Y}$. This assumption reduces the number of possible $(\tilde{X}, \tilde{Y})$ pairs. Instead of selecting one of the possible pairs, we use the average $Pr(\tilde{X} = c_i | \tilde{Y} = c_j)$ conditional probabilities over the possible pairs in order to define the conditional entropy in (2). More details can be found in [1].

We define the conditional entropy of X and Y as follows.

$$H(X|Y) = H(\tilde{X}|\tilde{Y}) =$$
$$- \sum_{j=1}^{K} Pr(\tilde{Y} = c_j) \cdot \sum_{i=1}^{K} Pr(\tilde{X} = c_i | \tilde{Y} = c_j) \cdot \log Pr(\tilde{X} = c_i | \tilde{Y} = c_j) \qquad (2)$$

The conditional entropy is always smaller or equal to the entropy, $H(\tilde{X}|\tilde{Y}) \leq H(\tilde{X}) = H(X)$.

4 The disclosure risk measure

4.1 The disclosure risk measure for population based frequency tables

The disclosure risk measure, which has been introduced for population based frequency tables, is a weighted average as follows.

$$R_1(F, \boldsymbol{w}) = w_1 \cdot \frac{|D|}{K} + w_2 \cdot \left(1 - \frac{H(X)}{\log K}\right) - w_3 \cdot \frac{1}{\sqrt{N}} \cdot \log \frac{1}{e \cdot \sqrt{N}} \tag{3}$$

Here D is the set of zeroes in the population based table, therefore $|D|/K$ is the proportion of zeroes. $\boldsymbol{w} = (w_1, w_2, w_3)$ is a vector of weights, $w_i \geq 0$, $i = 1, 2, 3$, $\sum_{i=1}^{3} w_i = 1$.

4.2 The disclosure risk measure for sample based frequency tables

While population based tables include every individual, only selected individuals contribute to sample based frequency tables. Sampling can be considered as a special statistical disclosure control (SDC) method. The smaller number of individuals in sample based tables ensures protection against attribute disclosure to a certain extent. An intruder faces more uncertainty in a sample based table than in a population based table. Zeroes in a sample based table seemingly increase the chance of attribute disclosure. However, a zero in a sample based table is not necessarily zero in the population based table.

A disclosure risk measure for sample based frequency tables (f) is as follows.

$$\begin{aligned} R_2(F, f, \boldsymbol{w}) = w_1 \cdot \left(\frac{|D|}{K}\right)^{\frac{|D \cup E|}{|D \cap E|}} + \\ w_2 \cdot \left(1 - \frac{H(X)}{\log K}\right) \cdot \left(1 - \frac{H(X|Y)}{H(X)}\right) - w_3 \cdot \frac{1}{\sqrt{N}} \cdot \log \frac{1}{e \cdot \sqrt{N}} \end{aligned} \tag{4}$$

Here E is the set of zeroes in the sample based table and e is the base of the natural logarithm. The above disclosure risk measure was developed for perturbed population based frequency tables. Since sampling can be considered as an SDC method, the formula can be applied directly to sample based tables. Note that the power of the first term reduces as follows since in our case $D \subseteq E$.

$$\frac{|D \cup E|}{|D \cap E|} = \frac{|E|}{|D|} .$$

We assume that the population size (N) is known to the statistical agency, therefore the third term of the above formula can be calculated with ease. Our aim is to estimate $H(X)$, $H(X|Y)$ and $|D|$ from the sample based table when the population frequencies are unknown. In this paper, this aim is achieved by estimating population frequencies. From the estimated population frequencies the above mentioned

quantities can be calculated as for known population based tables. The population frequencies may be estimated from the sample by probabilistic models.

5 Models to estimate population frequencies

We present numerical results in Section 6 using two modelling approaches for estimating population frequencies, a log-linear model approach and a so-called Pólya urn model approach. The results are derived from generated and real population based tables in order to assess the estimation error arising from the sampling in the first case and from the sampling and estimation of population parameters in the second case.

5.1 Log-linear model approach

A sample based frequency table may contain zero cells that have positive values in the population based table due to the random sampling. Therefore cell probabilities might not be reflected properly in a sample based table. Log-linear models can compensate for sample-based (random) zero cells and introduce positive cell probabilities by taking the table structure into account. On the other hand, log-linear models can also estimate positive cell probabilities when there should be a true population (structural) zero.

We apply a log-linear model to two-dimensional (sample based) frequency tables. In this situation we can only include main effects in the mode which will have the effect of estimating positive cell values even for those cells that are true (structural) zeroes in the population.

Denote the sum of row i by $n_{i\bullet}$ and that of column j by $n_{\bullet j}$. The expected cell count under the log-linear model is

$$\hat{\mu}_{ij} = \frac{n_{i\bullet} \cdot n_{\bullet j}}{n} \ .$$

Dividing the above formula by n provides (estimated) cell probabilities $\hat{p}_{ij} = \frac{\hat{\mu}_{ij}}{n}$.

5.2 Pólya urn model approach

The urn model has been employed in [8] to estimate population uniques in a frequency table. Now we use a similar model to estimate all population frequencies.

The model starts with positive sample based frequencies. The frequencies are represented by coloured balls in an urn. The urn contains $f_1 > 0$ balls of colour 1, $f_2 > 0$ balls of colour 2, etc. In addition to the coloured balls, θ black balls are also placed into the urn, where θ is a parameter to be estimated. In each step we draw a ball from the urn. If it is a coloured ball, then we replace it and add a new ball of the same colour to the urn. If the ball we draw is black, then the ball is replaced and another of a new colour is placed into the urn. The balls of new colours account for sample zeroes.

In our case there might be true zeroes in the sample based table, therefore we do not assume that all sample frequencies are positive. However, zeroes do not influence the estimated population frequencies.

The estimation of the θ parameter has an impact on the number of newly introduced frequencies. The number of zeroes in the population based table plays an important role in the estimation of θ and in our disclosure risk measure with respect to the first term in (4). A high θ might result in a large number of new frequencies, therefore the number of zeroes in the population based table might be underestimated. Similarly, a low θ might imply a high number of population zeroes. We determine θ according to the number of zeroes in the population based table.

First assume that $|D|$ is known. The number of cells that are zeroes in the sample based table but positive in the population based table is $|E| - |D|$. Denote W_z, $z = 1, 2, \ldots, N - n$, an indicator variable as follows.

$$W_z = \begin{cases} 1 & \text{if the } z\text{th draw is a black ball} \\ 0 & \text{if the } z\text{th draw is a coloured ball} \end{cases}$$

The expected number of new colours is $E(\sum_{z=1}^{N-n} W_z) = \sum_{z=1}^{N-n} E(W_z)$. The total number of balls before the zth draw is $n + \theta + z - 1$. Since the number of black balls is constant at θ, therefore $E(W_z) = \dfrac{\theta}{n + \theta + z - 1}$. We obtain θ by solving the following equation (numerically):

$$|E| - |D| = \sum_{z=1}^{N-n} \frac{\theta}{n + \theta + z - 1} \ . \tag{5}$$

Assume now that $|D|$ is unknown. In order to use (5), we need to estimate $|D|$ from the sample based table. Section 9.8 of [2] provides expected frequencies of frequencies. The expected number of zeroes is given by the following formula.

$$\widehat{|D|} = \sum_{i=1}^{K} (1 - p_i)^N \ ,$$

where p_i is the probability of cell c_i. We estimate p_i, $i = 1, 2, \ldots, K$ by applying an independent log-linear model to the sample based table.

Therefore, (5) can be rewritten as follows.

$$|E| - \widehat{|D|} = \sum_{z=1}^{N-n} \frac{\theta}{n + \theta + z - 1} \ . \tag{6}$$

We can solve (6) numerically to obtain the estimate $\hat{\theta}$.

6 Simulation study

In this section we present results of a simulation study to assess the estimation error of the disclosure risk measure in (4). We use a real population based table and a table that is generated according to known model parameters estimated from the real table. The aim is to assess the estimation error arising from sampling alone and the estimation error arising from both sampling and estimated model parameters.

6.1 Data

The dataset we used is an extract from the 2001 UK census data. The dataset consists of $N = 2449$ individuals of 10 selected output areas. The output area (10 output areas) $\times$ religion two-dimensional table has $K = 90$ cells. The frequencies are shown in Table 1.

181	0	0	1	17	1	1	83	18
138	2	4	2	0	0	1	36	16
130	0	0	0	22	4	1	61	40
173	0	0	1	14	4	1	97	22
142	2	5	0	15	6	1	37	21
129	0	0	0	0	0	1	69	20
118	2	0	2	24	9	1	38	20
130	0	0	0	34	1	1	82	32
148	3	0	0	0	2	1	38	21
136	1	2	0	13	0	0	55	16

Table 1: Original frequency table

To obtain the generated population table for assessing the log-linear model approach, we applied the log-linear model with main effects on Table 1. The estimated cell probabilities, denoted by $(\hat{p}_1^{sim}, \hat{p}_2^{sim}, \ldots, \hat{p}_K^{sim})$, were then used as the parameters of a multinomial distribution. We drew N individuals from $Multinom(N; \hat{p}_1^{sim}, \hat{p}_2^{sim}, \ldots, \hat{p}_K^{sim})$. When assessing the estimation error arising from sampling alone, we use these same parameters $(\hat{p}_1^{sim}, \hat{p}_2^{sim}, \ldots, \hat{p}_K^{sim})$ for estimating the disclosure risk measure.

To obtain the generated population table for assessing the Pólya urn model approach, we use θ given in (5) to generate the population based frequencies from the sample based frequencies.

6.2 Simulation method

For the simulation study, we drew 1000 simple random samples from the (original or generated) population using two sample fractions of 0.1 and 0.05. $R_2(F, f, \boldsymbol{w})$ can be calculated on the (original or generated) population based table for each of the sample based tables. The average of the $R_2(F, f, \boldsymbol{w})$ values is considered as the

'original disclosure risk'. For this simulation, we use the following weights in the disclosure risk measure: $\boldsymbol{w} = (0.1, 0.8, 0.1)$.

When population frequencies are assumed unknown, we need to estimate them from the sample based table. In the log-linear approach, for the case of the generated population table with known parameters, we estimate the population frequencies by drawing $N - n$ individuals from $Multinom(N - n; \hat{p}_1^{sim}, \hat{p}_2^{sim}, \ldots, \hat{p}_K^{sim})$ and adding these frequencies to the respective sample-based table. For the case of the real population table, we estimate the population frequencies by applying the log-linear model with main effects to the sample-based table (f). The resulting table provides estimated cell probabilities. Denote them by $(\hat{q}_1^S, \hat{q}_2^S, \ldots, \hat{q}_K^S)$, where the superscript S refers to the sample. $N - n$ individuals are drawn from $Multinom(N - n; \hat{q}_1^S, \hat{q}_2^S, \ldots, \hat{q}_K^S)$, and are then added to the sample-based table, thereby estimating the population frequencies.

In the Pólya urn approach, for the case of the generated population table we use θ given by (5) and for the case of the real population table, we estimate θ on each of the sample based tables as defined in (6).

The simulation is carried out as follows for each sample fraction and for each (original or generated) population. On each of the 1000 sample-based tables we estimate the population frequencies 1000 times. For each estimated population-based table ($\hat{F}$) of each sample-based table we obtain an estimated disclosure risk measure ($R_2(\hat{F}, f, \boldsymbol{w})$). Note that the overall number of the $R_2(\hat{F}, f, \boldsymbol{w})$ values is equal to $1000 \cdot 1000$. The average of the $R_2(\hat{F}, f, \boldsymbol{w})$ values is considered as the final 'estimated disclosure risk'.

6.3 Numerical results

Table 2 presents the results of the simulation study using both the generated and real population based tables and two sampling fractions 0.1 and 0.05. We compare the 'original disclosure risk' with the 'estimated disclosure risk'. The weights for the disclosure risk measure are $\boldsymbol{w} = (0.1, 0.8, 0.1)$.

Generated and real data		Original disc. risk $R_2(F, f, (0.1, 0.8, 0.1))$		Log-linear model $R_2(\hat{F}, f, (0.1, 0.8, 0.1))$		Pólya urn model $R_2(\hat{F}, f, (0.1, 0.8, 0.1))$	
	Sampling fr.	Mean	St. dev.	Mean	St. dev.	Mean	St. dev.
Generated table	0.1	0.1538	0.0043	0.1568	0.0039	-	-
(log-linear m.)	0.05	0.1427	0.0059	0.1416	0.0054	-	-
Generated table	0.1	0.1694	0.0049	-	-	0.1758	0.0053
(Pólya urn m.)	0.05	0.1535	0.0061	-	-	0.1640	0.0057
Real	0.1	0.1697	0.0048	0.1715	0.0173	0.1764	0.0186
table	0.05	0.1535	0.0061	0.1731	0.0254	0.1821	0.0283

Table 2: Results of disclosure risk measures on generated and real population based tables

For the log-linear model, using the generated population based table with known parameters under the log-linear model, we see that we can obtain close estimates to

the original disclosure risk measures when only sampling error is considered. The estimated disclosure risk measure based on the real population table is slightly higher than the original disclosure risk. The overestimation is worse for the smaller sample fraction.

The Pólya urn modelling approach provides only slightly less accurate estimates than the log-linear modelling approach but there appears to be overestimation both in the generated table and the real population table.

7 Conclusion

In this paper, we present an information theoretical based disclosure risk measure for two-dimensional sample based tables. Under the generated population based table with known parameters, the disclosure risk can be estimated accurately and therefore the estimation error arising from the sampling alone appears to be unbiased. However, the estimated disclosure risk for a real population based table where we need to account for the estimating of the parameters from the sample based table is less accurate. The Pólya urn model approach is a feasible alternative to the log-linear model approach. Further reseach needs to be carried out in order to provide a more accurate approximation of the disclosure risk using different size tables with varying sampling fractions and levels of random and true zero cells in the population. In addition, further research is needed to explore the estimation of disclosure risk in higher dimensional tables.

Acknowledgements

This work was funded by the ONS-ESRC PhD studentship (Ref. ES/J500161/1).

References

[1] Antal, L. and Shlomo, N. and Elliot, M (2014) "Measuring Disclosure Risk with Entropy in Population Based Frequency Tables", *Privacy in Statistical Databases*, 62–78, Springer.

[2] Bishop, Y. M., Fienberg, S. E. and Holland, P. W. (2007) "Discrete Multivariate Analysis: Theory and Practice", Springer.

[3] Bunge, J. and Fitzpatrick, M. (1993) "Estimating the Number of Species: a Review", *Journal of the American Statistical Association*, **88/421**, 364–373.

[4] Cover, T. M. and Thomas, J. A. (2006) "Elements of Information Theory", 2nd. ed., Wiley, Hoboken.

[5] Goodman, L. A. (1949) "On the Estimation of the Number of Classes in a Population", *The Annals of Mathematical Statistics*, **20/4**, 572–579.

[6] Haas, P. J., Naughton, J. F., Seshadri, S. and Stokes, L. (1995) "Sampling-Based Estimation of the Number of Distinct Values of an Attribute", *Proceedings of the 21th International Conference on Very Large Data Bases*, 311–322.

[7] Hoppe, F. M. (1984) "Pólya-Like Urns and the Ewens' Sampling Formula", *Journal of Mathematical Biology*, **20/1**, 91–94.

[8] Samuels, S. M. (1998) "A Bayesian Species-Sampling-Inspired Approach to the Uniques Problem in Microdata Disclosure Risk Assessment", *Journal of Official Statistics*, **14**, 373–384.

[9] Shlomo, N. (2007) "Statistical Disclosure Control Methods for Census Frequency Tables", *International Statistical Review*, **75**, 199–217.

[10] Skinner, C. J. and Shlomo, N. (2012) "Estimating Frequencies of Frequencies in Finite Populations", *Statistics & Probability Letters*, **82/12**, 2206–2212.

[11] Willenborg, L. and de Waal, T. (2001) "Elements of Statistical Disclosure Control", Lecture Notes in Statistics, Springer

4.8 Disclosure Risk in Three-Dimensional Tables

The disclosure risk measurement of three-dimensional tables does not differ much from that of two-dimensional tables. In order to assess the disclosure risk we need to estimate cell frequencies. Log-linear models can be applied again. However, there are more choices of the log-linear model. We consider models with only main effects as well as with two-way interactions included.

We considered the output area (10 output areas) $\times$ religion $\times$ sex table. First 1,000 samples were drawn from the original population. The sampling fractions we used are 0.1 and 0.05. A log-linear model was applied to each sample based table and each resulting table was divided by the sample size (n) in order to get estimated cell probabilities. We generated $N - n$ individuals 1,000 times from a multinomial distribution and another 1,000 times following the Pólya urn model. The parameters were given by the estimated cell probabilities. The frequency table based on $N - n$ individuals was added to the sample based table, thereby estimating the population frequencies. We generated 1,000 sample based tables, therefore there are 1,000 'original' $R_2(F, f, \boldsymbol{w})$ disclosure risk measures. The mean of the 1,000 values is considered as the final original disclosure risk measure. Since there are 1,000 estimated tables for each sample, we have $1,000 \cdot 1,000 = 10^6$ estimated population based frequency tables, and therefore the same number of 'estimated' $R_2(\hat{F}, f, \boldsymbol{w})$ disclosure risk measures. The final disclosure risk is the average of the 10^6 values. The weights we used are $\boldsymbol{w} = (0.1, 0.8, 0.1)$. The results are shown in Table 4.8.1. In the table, 'Log-linear model 1' refers to the log-linear model with main effects only, while in 'Log-linear model 2' also two-way interactions are included. The estimates given by 'Log-linear model 1' are closer to the original disclosure risk than the values given by 'Log-linear model 2'.

		Original disc. risk		Multinomial distr.		Pólya urn model	
		$R_2(F, f, (0.1, 0.8, 0.1))$		$R_2(\hat{F}, f, (0.1, 0.8, 0.1))$		$R_2(\hat{F}, f, (0.1, 0.8, 0.1))$	
	Sampling fr.	Mean	St. dev.	Mean	St. dev.	Mean	St. dev.
Log-linear	0.1	0.1542	0.0039	0.1499	0.0153	0.1491	0.0153
model 1	0.05	0.1327	0.0049	0.1412	0.0206	0.1412	0.0206
Log-linear	0.1	0.1542	0.0039	0.2264	0.0112	0.2260	0.0104
model 2	0.05	0.1327	0.0057	0.2169	0.0166	0.2451	0.0162

Table 4.8.1: Results of three dimensional tables

Chapter 5

Flexible Table Generators

5.1 Introduction

Statistical agencies seek new channels to disseminate data. One of the potential ways is online dissemination. A user can download tables prepared by statistical agencies. A much more flexible alternative is to allow users to tailor their own outputs. It serves the user needs much better than fixed outputs. Also, by collecting the outputs users generate, a statistical agency can easily follow what data are of most demand. Such information can help statistical agencies to improve on their services. Naturally, users desire swift access to the outputs. However, disclosure risk assessment cannot be neglected, even if outputs are available online. The process to protect data, described in the Section 5.2, is not different for online access but it has to be carried out automatically.

5.2 Dissemination of Data

A statistical agency regularly disseminates data. The process of dissemination follows the next steps for a particular dataset.

1. Select a disclosure risk measure.

2. Set a threshold (T_1) for the disclosure risk measure.

3. Measure the disclosure risk of the data.

4. If the disclosure risk measure is below T_1, then the data can be released. If not, then go to the next step.

5. Apply an SDC method with certain parameters to the data.

6. Set a threshold (T_2) for the perturbed data.

7. Measure the disclosure risk of the perturbed data.

8. If the disclosure risk measure is below T_2, then the perturbed data may be released.

9. If the data can be released, then select an information loss measure.

10. Set a threshold (T_3) for the information loss measure.

11. Measure the information loss.

12. If the information loss exceeds T_3, then the statistical agency might prefer to not disseminate the data.

13. Repeat steps 5 - 12 for more SDC methods and parameters.

14. Find the best SDC method and its best parameters.

15. Disseminate the data if possible.

In step 2 a threshold has to be set. This work does not deal with the problem of how to set it in detail. The threshold should depend on the disclosure risk measure. A statistical agency might evaluate a disclosure risk measure on many datasets and gain experience of data of high/low disclosure risk. This approach is based on our understanding of the term 'disclosure risk'. A data protector might 'feel' that a particular

dataset should/should not be released. The disclosure risk measure of such datasets should be below/above the threshold.

If the disclosure risk measure surpasses T_2 in step 8, then the data protector needs to select either other parameters of the SDC method or an entirely different SDC method.

Selecting a threshold (T_3) for the information loss measure in step 10 is similar to that for the disclosure risk measure. The data protector can decide whether the information loss is acceptable or not.

Steps 5 - 12 can be repeated for various SDC methods and parameters. The point is to find the best method and parameters. The 'best' SDC method and its 'best' parameters (see step 14) ideally provides the lowest disclosure risk measure and the lowest information loss. Such method and parameters might not exist. The situation resembles a search for Pareto optimality. An SDC method and its parameters are not optimal if other parameters or another SDC method and its parameters provide lower/not higher disclosure risk and not higher/lower information loss measure. We can call an SDC method and its parameters a (weak) Pareto optimum if it is impossible to improve both on the disclosure risk and the information loss. There might be more weak Pareto optimal SDC methods and parameters. For example, the comparison of two SDC methods may show that the first method provides lower disclosure risk and higher information loss, while the second method higher disclosure risk and lower information loss.

Disclosure risk measures always show a single numerical value on a certain dataset. However, since a disclosure risk measure focuses on certain aspects/problematic cases, a data protector might consider using more different disclosure risk measures at the same time. More disclosure risk measures will produce a vector of numerical values. A data protector can deem a dataset safe to release if the elements of the disclosure risk vector remain below a vector of thresholds. Setting the thresholds causes problems again.

Quick dissemination of data requires quick disclosure risk assessment. Steps 1 - 15 above can be made automatic. Section 5.3 discusses how frequency tables can be generated and released quickly.

5.3 Paper: Measuring Disclosure Risk and Data Utility for Flexible Table Generators

The paper below was submitted to the Journal of Official Statistics. It was published in 2015.

DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 31, No. 2, 2015, pp. 305–324, http://dx.doi.org/10.1515/JOS-2015-0019

Measuring Disclosure Risk and Data Utility for Flexible Table Generators

Natalie Shlomo[1], *Laszlo Antal*[1], *and Mark Elliot*[1]

Statistical agencies are making increased use of the internet to disseminate census tabular outputs through web-based flexible table-generating servers that allow users to define and generate their own tables. The key questions in the development of these servers are: (1) what data should be used to generate the tables, and (2) what statistical disclosure control (SDC) method should be applied. To generate flexible tables, the server has to be able to measure the disclosure risk in the final output table, apply the SDC method and then iteratively reassess the disclosure risk. SDC methods may be applied either to the underlying data used to generate the tables and/or to the final output table that is generated from original data. Besides assessing disclosure risk, the server should provide a measure of data utility by comparing the perturbed table to the original table. In this article, we examine aspects of the design and development of a flexible table-generating server for census tables and demonstrate a disclosure risk-data utility analysis for comparing SDC methods. We propose measures for disclosure risk and data utility that are based on information theory.

Key words: Statistical disclosure control; census tabular data; entropy; Hellinger distance.

1. Introduction

Driven by demand from policy makers and researchers for specialized and tailored census frequency tables, many statistical agencies are considering the development of a web-based software platform where users can generate tables of interest from underlying census microdata through a user-friendly interface. This platform is called a "flexible table-generating server". Users access the server via the internet and generate their preferred set of tables from predefined variables or categories using drop-down lists. These tables can then be downloaded to the personal computers of the users. The United States Census Bureau and the Australian Bureau of Statistics have developed such servers on their websites to disseminate census frequency tables.

When generating flexible tables, the server should be able to provide a measure of disclosure risk for the original table, apply a statistical disclosure control (SDC) method and then reassess disclosure risk and the impact on data utility following the SDC method. These steps must be carried out "on the fly" within the server for each generated output table. SDC is a set of statistical practices which aim to ensure that no individual population

[1] University of Manchester, Social Statistics, Humanities Bridgeford Street, Manchester M13 9PL, United Kingdom. Emails: natalie.shlomo@manchester.ac.uk, laszlo.antal@postgrad.manchester.ac.uk, and mark.elliot@manchester.ac.uk

Acknowledgments: The project is funded by the EU 7th framework infrastructure research grant: 262608, Data Without Boundaries (DwB) and the ONS-ESRC funded PhD studentship (Ref. ES/J500161/1).

unit can be reidentified from anonymised data nor any new information learnt about any specific individual (with certainty). SDC is an active research area. For reviews of this area, see Willenborg and de Waal (2001), Doyle et al. (2001), Duncan et al. (2011) and Hundepool et al. (2012).

There are two main types of disclosure risks in census frequency tables: identity disclosure, where small cell counts may lead to the identification of an individual in the population, and attribute disclosure, where new information may be learnt about an individual or group of individuals. Attribute disclosure in frequency tables occurs when rows or columns of a table contain (real) zeroes and only one or two cells are nonzero. This enables an "intruder" to first make an identification based on a margin total and subsequently reveal new information according to other variables spanning the table. Another type of disclosure risk that needs to be guarded against is disclosure by differencing. The differencing of tables generated through the server can lead to residual tables that are more susceptible to the above disclosure risks and even to the reconstruction of individual records. This is typically dealt with by applying perturbative methods of SDC, which raises the level of uncertainty of true counts in the tables and hence of the difference between counts across tables. After the table is protected, a data utility measure must also be calculated by comparing the perturbed table to the original table.

The need to measure disclosure risk "on the fly" for census frequency tables produced via a flexible table-generating server motivated the research and development of a new global disclosure risk measure. Until now, disclosure risk measures for tabular data have been defined at the cell level and not for the entire table. We propose a new disclosure risk measure based on information theory as shown in Antal et al. (2014) and also relate this theory to a data utility measure.

The key issues when developing a web-based flexible table generating server addressed in this article are: (1) what underlying data should be used in the background for generating the output tables, and (2) at what stage should the SDC method be applied. In addition, the article provides a comparison study of some common SDC methods which may be used to protect census tables within a flexible table-generating server and demonstrates how statistical agencies should undertake a disclosure risk-data utility analysis to inform decisions about SDC methods and their parameterization. In general, SDC methods employed by statistical agencies are often motivated by country-specific agendas and policy sensitivities and it is difficult to develop a universal best practice. However, one important distinction when considering SDC methods for flexible table-generating servers is that the outputs are defined by users and the amount of disclosure risk may vary in each output.

Section 2 presents aspects to consider in the design of a flexible-table generating server, including the underlying data for generating output tables and the stage when SDC methods may be applied. In Section 3, some common SDC methods for census frequency tables are described. Section 4 introduces a new global disclosure risk measure based on information theory and a related data utility measure that can be calculated "on the fly" for each output table generated in the server. In Section 5, a comparison study is carried out on generated census output tables from a flexible table-generating server. The comparison study will be informed by a disclosure risk-data utility analysis on the generated tables perturbed by the SDC methods described in Section 3 based on the

measures outlined in Section 4. A discussion and concluding remarks are presented in Section 6.

2. Designing a Flexible Table-Generating Server

In this section, we describe the design of an online flexible table-generating server and discuss the following issues: the underlying data that may be used as input to the server, the stage at which SDC methods can be applied, and preliminary SDC rules to determine *a priori* whether the requested table can be generated or not.

2.1. Underlying Input Data to the Server

The underlying data to use as input for a web-based flexible table-generating server can be based on the original microdata or disclosure-controlled microdata. The input data is largely determined by the source and content of the data as well as the SDC method that will be applied to the final output tables (if any). Microdata arising from social surveys with small sampling fractions have a lower disclosure risk than microdata arising from censuses containing whole population counts, and therefore are more appropriate for use in their original form. Output tables generated from survey microdata where only weighted counts are released are generally considered to be of low disclosure risk with no further need for an application of SDC methods. Census (and administrative data) containing whole populations and particularly those containing sensitive data, such as health statistics or business microdata, are more problematic. In microdata containing the whole population, individuals (or businesses) can easily be identified leading to the disclosure of attributes. In this case, the underlying input data should be protected prior to the generation of tables.

For a flexible table-generating server of census tables, one method for producing the underlying input data is to aggregate the microdata into a very large multi-dimensional frequency table, called a hypercube, where no data of individuals can be disseminated below the level of a cell value in the hypercube. For example, users may only be able to disseminate frequency counts of age in 5-year age bands and not counts for single years. This approach was taken by Eurostat for the dissemination of census tables from European Member States. A flexible table-generating server for European census tables is being developed through the European Census Hub Project. Each Member State is required to produce a set of predefined hypercubes containing their country's census counts: 19 hypercubes at the geography level of LAU2 and over 100 hypercubes at the geography level of NUTS2, cross-classified with as many as six other census variables in each hypercube. NUTS2 is a European subregional geography and LAU2 are small municipalities or equivalent. Researchers are able to use the considerable number of multidimensional hypercubes and their wealth of census data made available through the European Census Hub to generate tables of interest beyond what would have been available previously using standard table-extraction software. The flexible table-generating server will allow comparative tables across Member States and the combining of census data from multiple Member States. The hypercubes have the additional advantage that they provide some limited protection against disclosure risk since no data below the level of the cell values of the hypercube can be disseminated.

However, the hypercubes themselves still have considerable disclosure risk since they are very large and sparse with many zero and small cell counts. Therefore, there will still be the need to apply an SDC method to protect output tables generated from the flexible table-generating server.

2.2. Application of SDC Methods

SDC methods for protecting output tables generated from a flexible table-generating server can be applied either on the underlying input data so that all tables generated are deemed safe for dissemination (the pretabular SDC approach), or applied directly to the final output table generated from the original data (the post-tabular SDC approach) or a combination of both. Although sometimes neater and less resource intensive when data is from a single source, the pretabular SDC approach is problematic for the dissemination of European Census data for two reasons. Firstly, all Member States would have to agree on a common SDC method in order to provide consistent hypercubes across all Member States. For example, if one Member State employs a rounding method whilst another Member State employs cell suppression, there will be significant quality issues in a table that is generated based on both Member States' data. Secondly, when aggregating data which have been separately disclosure controlled, the effects of the SDC methods are compounded and the data may be overprotected. For example, aggregating cells that have already been rounded not only overprotects the data but also exacerbates the data utility impact by providing counts that are no longer rounded to the nearest base. With the second approach of protecting only the final tabular output, SDC methods are not compounded in this way. We investigate the pretabular and post-tabular approaches in the comparison study presented in Section 5.

2.3. Preliminary SDC Rules

The design of a web-based flexible table-generating server typically involves many *ad hoc* preliminary SDC rules which determine *a priori* if generated tables can be released or not. These SDC rules may include:

- Limiting the number of dimensions in the output tables.
- Ensuring consistent and nested categories of variables to avoid disclosure by differencing.
- Ensuring minimum population thresholds.
- Ensuring that the percentage of small cells is below a maximum threshold.
- Ensuring average cell size above a minimum threshold.

The steps in a flexible table-generating server are:

(1) Determine whether the table can be released according to the preliminary SDC rules.
(2) Calculate a disclosure risk measure to determine if an SDC method should be applied to the final output table.
(3) Apply the SDC method.

(4) Recalculate the disclosure risk measure to determine if the table is safe to generate; if yes proceed to Step 5, otherwise do not release the table.
(5) Output the final table with a measure of data utility.

According to the steps of a flexible table-generating server, it is clear that analytical expressions of disclosure risk and data utility that can be calculated "on the fly" within the server are necessary.

3. Statistical Disclosure Control Methods

In this section, we describe some common SDC methods which have been used to protect census frequency tables: a pretabular SDC method of record swapping is used in the United States and the United Kingdom, a post-tabular method of random rounding is used in New Zealand and Canada, and a post-tabular probabilistic perturbation mechanism has recently been implemented in Australia.

3.1. Record Swapping

Record swapping is based on the exchange of values of variable(s) between similar pairs of population units (often households). In order to minimize bias, pairs of population units are determined within strata defined by control variables. For example, when swapping households, control variables may include: a large geographical area, household size, and the age-sex distribution of individuals in the households. In addition, record swapping can be targeted to high-risk population units found in small cells of census tables. In a census context, geographical variables related to place of residence are often swapped. Swapping place of residence has the following properties: (1) it minimizes bias based on the assumption that place of residence is independent of other census target variables conditional on the control variables; (2) it provides more protection for census tables since place of residence is a highly visible variable which can be used to identify individuals; (3) it preserves marginal distributions within a larger geographical area. For more information on record swapping, see Dalenius and Reiss (1982), Fienberg and McIntyre (2005), and Shlomo (2007).

3.2. Semi-Controlled Random Rounding

A post-tabular method of SDC for census frequency tables is unbiased random rounding. Let *Floor(x)* be the largest multiple bk of the base b such that $bk < x$ for any value of x. In this case, $res(x) = x\text{-}Floor(x)$. For an unbiased rounding procedure, x is rounded up to $Floor(x) + b$ with probability $res(x)/b$ and rounded down to $Floor(x)$ with probability $(1 - (res(x)/b))$. If x is already a multiple of b, it remains unchanged.

In general, each cell is rounded independently in the table, that is, a random uniform number u between 0 and 1 is generated for each cell. If $u \leq (res(x)/b)$ then the entry is rounded up, otherwise it is rounded down. This ensures an unbiased rounding scheme, that is, the expectation of the rounding perturbation is zero. However, the realization of this stochastic process on a finite number of cells in a table will not ensure that the sum of the perturbations will exactly equal zero. To place some control in the random rounding procedure, we use a semi-controlled random rounding algorithm for selecting entries to round up or down as follows: first the expected number of entries of a given *res(x)* that are

to be rounded up is predetermined (for the entire table or for each row/column of the table). The expected number is rounded to the nearest integer. Based on this expected number, a random sample of entries is selected (without replacement) and rounded up. The other entries are rounded down. This procedure ensures that rounded internal cells aggregate to the controlled rounded total.

Due to the large number of perturbations under random rounding, margins are typically rounded separately from internal cells and tables are not additive. When using semicontrolled random rounding this alleviates some of the problems of nonadditivity since one of the margins and the overall total will be preserved. Another problem with random rounding is the consistency of the rounding across same cells that are generated in different tables. It is important to ensure that the cell value is rounded consistently, otherwise the true cell count can be learnt by generating many tables containing the same cell and observing the perturbation patterns. Fraser and Wooton (2005) propose the use of *microdata keys* which can solve the consistency problem. First, a random number (which they call a key) is defined for each record in the microdata. When building a census frequency table, records in the microdata are combined to form a cell defined by the spanning variables of the table. When these records are combined to a cell, their keys are also aggregated. This aggregated key serves as the seed for the rounding and therefore same cells will always have the same seed and result in consistent rounding.

Further research is needed to ensure both the additivity and consistency properties for random rounding. For simple tables of the type that would be generated in a flexible table-generating server, controlled rounding algorithms can be applied to ensure additivity on remaining totals without distorting the unbiasedness of the rounding (see Willenborg and De Waal 2001).

3.3. *Stochastic Perturbation*

A more general method than random rounding is stochastic perturbation, which involves perturbing the internal cells of a table using a probability transition matrix and is similar to the postrandomisation method that is used to perturb categorical variables in microdata (see Gouweleeuw et al. 1998). In this case, it is the cell counts in a table that are perturbed. More details can be found in Fraser and Wooton (2005) and Shlomo and Young (2008).

Let $\mathbf{P}$ be a $(L+1)\times(L+1)$ transition matrix containing conditional probabilities: $p_{ij} = P(\textit{perturbed cell value is } j \,|\, \textit{original cell value is } i)$ for cell values from 0 to L, where L is a cap on the cell values and any cell value above the cap will have the same perturbation probabilities. Let $\mathbf{t}$ be the vector of frequencies of the cell values where the last component would contain the number of cells above cap L and let $\mathbf{v}$ be the vector of relative frequencies: $\mathbf{v} = \mathbf{t}/\mathrm{K}$ where K is the number of cells in the table. In each cell of the table, the cell value i is changed or not changed according to the prescribed transition probabilities in matrix $\mathbf{P}$ and the result of a draw of a random multinomial variate u with parameters $p_{ij} j = 0, 1, \ldots, L$. If the jth value is selected, value i is moved to value j. When $i = j$, no change occurs.

Placing the condition of invariance on the probability transition matrix $\mathbf{P}$ (i.e., $\mathbf{tP} = \mathbf{t}$) means that the marginal distribution of the cell values are approximately preserved under

the perturbation. As described in the random rounding procedure in Subsection 3.2, in order to obtain the exact marginal distribution a similar strategy for selecting cell values to change can be carried out. For each cell value *i,* the expected number of cells that need to be changed to a different value *j* is calculated according to the probabilities in the transition matrix. We then randomly select (without replacement) the expected number of cells *i* and carry out the change to *j*.

To preserve exact additivity in the table, an iterative proportional fitting algorithm can be used to fit the margins of the table after the perturbation according to the original margins. This results in cell values that are not integers. Exact additivity with integer counts can be achieved for simple tables by controlled rounding to base 1 using Tau-Argus, for example (Salazar-Gonzalez et al. 2005). Cell values can also be rounded to their nearest integers resulting in "close" additivity because of the invariance property of the transition matrix. Finally, the use of microdata keys as described in Subsection 3.2 can also be adapted to this SDC method to ensure the consistent perturbation of same cells across different tables by fixing the seed for the perturbation.

4. Information Theory-Based Disclosure Risk and Data Utility Measures

For each output table generated, the flexible table-generating server must provide analytical expressions of disclosure risk and data utility that can be calculated "on the fly" within the server. As mentioned in Section 1, one of the major causes of disclosure risk in census frequency tables is attribute disclosure caused by rows/columns that have many zero cells and only one or two populated cells. A row/column with a uniform distribution of cell counts would have little attribute disclosure risk, whilst a degenerate distribution of cell counts would have high attribute disclosure risk. Moreover, a row/column with large counts would have less risk of reidentification compared to a row/column with small counts.

There is no single global-level disclosure risk measure for census frequency tables that measures attribute disclosure and identity disclosure. In planning for the 2011 UK Census, the Office for National Statistics assessed attribute disclosure by producing many census tables and calculating the proportion of those columns/rows where only one or two cells were populated and the rest of the cells were zero. They also provided a measure based on the proportion of small cells across the tables. These measures do not provide an accurate quantification of the disclosure risk for a specific table. To obtain an analytical expression of disclosure risk for the entire table (or row/columns), it is natural to use information theory, specifically the entropy.

4.1. An Information Theory Disclosure Risk Measure

As described in Antal et al. (2014), a disclosure risk measure for a census frequency table should have the following properties: (a) small cell values have higher disclosure risk than large values; (b) uniformly distributed frequencies imply low disclosure risk; (c) the more zero cells in the census table, the higher the disclosure risk; (d) the risk measure should be bounded by 0 and 1. Using information theory, we develop an analytical expression of disclosure risk that meets these properties.

Information theory is covered comprehensively in Cover and Thomas (2006). One of the most important measures is the entropy. Let X be a discrete random variable having a

distribution $P = (p_1, p_2, \ldots, p_K)$. The entropy is defined as:

$$H(X) = H(P) = -\sum_{i=1}^{K} p_i \cdot \log p_i$$

If $p_i = 0$ for a category i, the respective term in the sum will be considered 0, since $\lim_{x \to 0} x \log x = 0$. It follows that $H(P) \geq 0$, since $-p_i \cdot \log p_i \geq 0$ with $H(P) = 0$ *iff* the probability mass is concentrated on one point. Therefore, the smaller the entropy *H(P)*, the more likely that attribute disclosure can occur. Under the uniform distribution $U_K = ((1/K), (1/K), \ldots, (1/K))$, we obtain the maximum entropy: $H(U_K) = \log K$ and minimum attribute disclosure risk.

The entropy of the frequency vector in a table of size K, $F = (F_1, F_2, \ldots, F_K)$ where $\sum_{i=1}^{K} F_i = N$ is:

$$H(P) = H\left(\frac{F}{N}\right) = -\sum_{i=1}^{K} \frac{F_i}{N} \cdot \log \frac{F_i}{N} = \frac{N \cdot \log N - \sum_{i=1}^{K} F_i \cdot \log F_i}{N} \tag{1}$$

To produce a disclosure risk measure between 0 and 1, we define the risk measure as:

$$1 - \frac{H\left(\frac{F}{N}\right)}{\log K}. \tag{2}$$

The disclosure risk measure in (2) ensures property (b) since the term will tend to zero as the frequency distribution is more uniform, and ensures property (d) since the measure is bounded between 0 and 1. However, the disclosure risk measure does not take into account the magnitude of the cells counts or the number of zero cells in the table (or row/column of the table) and does not preserve properties (a) and (c). Therefore, an extended disclosure risk measure is proposed in (3) and is defined as a weighted average of three different terms, each term being a measure between 0 and 1.

$$\begin{aligned} R(F, w_1, w_2) = {} & w_1 \cdot \left[\frac{|A|}{K}\right] + w_2 \cdot \left[1 - \frac{N \cdot \log N - \sum_{i=1}^{K} F_i \cdot \log F_i}{N \cdot \log K}\right] \\ & - (1 - w_1 - w_2) \cdot \left[\frac{1}{\sqrt{N}} \cdot \log \frac{1}{e\sqrt{N}}\right] \end{aligned} \tag{3}$$

where A is the set of zeroes in the table and $|A|$ the number of zeros in the set, K, N and F as defined above and w_1, w_2 are arbitrary weights: $0 \leq w_1 + w_2 \leq 1$.

The first measure in (3) is the proportion of zeros which is relevant for attribute disclosure and property (c). The third measure in (3) allows us to differentiate between tables with different magnitudes and accounts for property (a). As the population size N gets larger in the table, the third measure tends to zero. The weights w_1 and w_2 should be chosen depending on the data protector's choice of how important each of the terms are in contributing to disclosure risk. Alternatively, one can avoid weights altogether by taking the L_2- norm (see Subsection 4.3) of the three terms of the risk measure in (3) as follows: $\left(\left(\left(\sum_{i=1}^{3} |x_i|^2\right)^{1/2}\right) \Big/ \sqrt{3}\right)$ where x_i represents term i, $i = 1, 2, 3$ in (3).

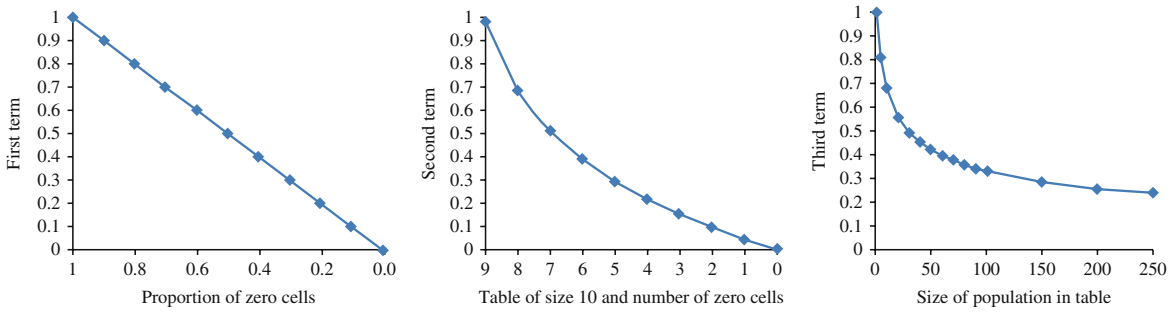

Fig. 1. The three components of the proposed disclosure risk measure in (3)

Figure 1 provides a graphical interpretation of each of the three terms of the proposed disclosure risk measure in (3). The figure on the left shows the first term of the disclosure risk measure as a function of the proportion of zero cells (although a table of all zeros would not be permissible in a flexible table-generating server). The figure in the middle shows the second term based on the entropy in (2) where we demonstrate with a table of ten cells and move from a uniform distribution to a degenerate distribution by accumulating zero cells and spreading the total to the remaining cells. The figure on the right shows the third term of the disclosure risk measure as the size of the population of the table increases.

The final disclosure risk measure (3) is an analytical expression and can be calculated "on the fly" in the flexible table-generating server without the need to see the generated table beforehand. In order to emphasize the risk of identity disclosure arising from small counts (ones and twos), we split the entropy measure as shown in (2) into two parts, small counts up to six and larger counts of seven and more, and provide different weights for each part. For the comparison study in Section 5, the following weights were chosen: $w_1 = 0.1$, $w_{2Part1} = 0.7$, $w_{2Part2} = 0.1$ and $w_3 = 0.1$ where the largest weight is attributed to the entropy term based on small counts. These weights were motivated by the empirical work carried out at the Office for National Statistics on SDC methods for the 2011 UK census tabular outputs, where attribute disclosure and small counts were of the highest concern.

4.2. Modifying the Disclosure Risk Measure After Perturbation

The disclosure risk measure in (3) does not take into account the application of SDC methods and therefore needs to be modified to reflect the uncertainty that is introduced into the counts of the table. Random rounding, for example, eliminates cells of size one and two by introducing more cells of size zero and three in the table, and seemingly increases the risk of attribute disclosure. However, these additional cells of size zero and three are not true counts and the risk of attribute disclosure should decrease. The disclosure risk as measured by the entropy in (2) (and the second term in (3)) does not reflect this uncertainty on whether the cell count is a true value or not. Therefore, we introduce an additional property for the disclosure risk measure following on from those defined in Subsection 4.1: (e) the disclosure risk measure following the application of an SDC method must be less than the original disclosure risk measure. In order to ensure property (e), we propose to modify the first two terms of the disclosure risk measure in (3) after the application of an SDC method as follows:

Modifying the First Term in (3):

The first term in (3) based on the proportion of zero cells can be generalized to compare the number of zero cells in the original and perturbed table. From (3), A is the set of zero cells in the original table and $|A|$ is the number of zero cells in the set. Similarly, let B be the set of zero cells in the perturbed table and $|B|$ the number of zero cells in the set. Denote $A \cup B$ as the union of the sets of zero cells and $A \cap B$ as the intersection of the sets of zero cells in the original and perturbed table. The revised first term in (3), which takes into account that nonzero cells may have been perturbed into zero cells and vice versa, is defined as: $(|A|/K)^{|A \cup B|/|A \cap B|}$. If there are no zero cells in the original table and hence $A \cap B = 0$, then the first term in (3) will remain equal to 0 following perturbation. For example, assume in a table there is a fraction of 0.10 zero cells and following perturbation a fraction of 0.20 zero cells and all original zero cells remain as zero in the perturbed table. In this case, the power term will be 2 and the risk measure following perturbation is reduced to 0.01 from the original 0.10. The modification of the first term in (3) is always less than the original term if nonzero cells are perturbed to zero cells and vice versa, and thus property (e) is ensured.

Modifying the Second Term in (3):

Assume that the possible values in the table are: $0, 1, 2, \ldots, L$ and the frequency of frequencies of these values is denoted by: $(n_0, n_1, n_2, \ldots, n_L)$. The table is perturbed according to a probability transition matrix (for example, the probability transition matrix $\mathbf{P}$ defined in Subsection 3.3). Let the frequency of frequencies of the perturbed values be denoted by: $(n'_0, n'_1, n'_2, \ldots, n'_L)$. For an observed perturbed value j, $j = 0, 1, \ldots, L$, the expected total from the cells of value j can be estimated by the proportion of the original values of j that are not changed: $(j \cdot n_j) \cdot p_{jj}$ and the proportion of other values i, $i \neq j$ that are changed to value j: $\sum_{i \neq j} (i \cdot n_i) \cdot p_{ij}$, so the expected total from cells of value j after perturbation is: $\sum_{i=0}^{L} (i \cdot n_i) \cdot p_{ij}$.

To reflect the uncertainty of the counts in the perturbed table, we replace the observed perturbed cells of value j by the expected total from cells of value j distributed evenly across all cells having the perturbed value j: $\left(\left(\sum_{i=0}^{L} (i \cdot n_i) \cdot p_{ij} \right) / (n'_j) \right)$. As an example, assume the SDC method of random rounding to base 3. We replace the zero cells in the perturbed table with: $[0 \cdot n_0 + 1 \cdot n_1 \cdot (2/3) + 2 \cdot n_2 \cdot (1/3)]/n'_0$. This reflects the fact that zero cells in the perturbed table are not true zeroes; rather, a proportion of them arise from the perturbation of cells of values one and two to zero cells under the probability mechanism, and it is unknown which zero cells are true zero cells and which zero cells are a result of the perturbation. Similarly, for the perturbed cell values of size three, we replace these with the term: $[1 \cdot n_1 \cdot (1/3) + 2 \cdot n_2 \cdot (2/3) + 3 \cdot n_3 + 4 \cdot n_4 \cdot (2/3) + 5 \cdot n_5 \cdot (1/3)]/n'_3$.

For the pretabular method of record swapping, we use a probability transition matrix applied at the cell level of the table for calculating the expectations as explained above, although it is possible that a perturbed table will be equal to the original table if the swapping variable is not involved in generating the table. The expected total from cells of value j in the table after record swapping is: $\sum_{i=0}^{L} (i \cdot n_i) \cdot p_{ij}$, where p_{ij} is a probability

transition matrix with the swap rate on the diagonal and all off-diagonals have equal probability constrained to the sum of the row probabilities being equal to 1. This means that we assume that every cell in the table can be perturbed according to the swap rate and reflects the assumption that an intruder would not know which variables were swapped.

The modification of the entropy term in (2) replaces observed perturbed counts with their expectations according to the probability transition matrix. In particular, true zero cells which did not contribute to the entropy in the original table are now replaced by their expected values. This should lead to a more even distribution of cell counts in the calculation of the entropy and to a general reduction in the disclosure risk measure in (2) following perturbation. As a final adjustment and to further guarantee property (e), we multiply the resulting entropy-based disclosure risk measures in (2) by a multiplier based on the average of the diagonal probabilities of the probability transition matrix. This multiplier reflects a global level of uncertainty introduced into the perturbed cell counts.

4.3. An Information Theory Data Utility Measure

To assess the distance between two distributions, we use the L_2-norm which, when applied to the difference of two vectors, preserves the properties of a distance metric (non-negativity, coincidence axiom, symmetry and triangle inequality). Measuring the distance infers that the smaller the distance, the more information is left in the table. For an arbitrary vector $x = (x_1, x_2, \ldots, x_K)$ the L_2-norm of x is defined as:

$$\| x \|_2 = \left(\sum_{i=1}^{K} |x_i|^2 \right)^{1/2}.$$

Let $P = (p_1, p_2, \ldots, p_K)$ be the original probability distribution of cell counts and $Q = (q_1, q_2, \ldots, q_K)$ the perturbed probability distribution of cell counts. Define: $\sqrt{P} = (\sqrt{p_1}, \sqrt{p_2}, \ldots, \sqrt{p_K})$ and $\sqrt{Q} = (\sqrt{q_1}, \sqrt{q_2}, \ldots, \sqrt{q_K})$. These are not (necessarily) probability distributions but have the property that as vectors, their L_2- norms are 1.

The Hellinger Distance is defined as the L_2 -norm:

$$HD(P, Q) = \frac{1}{\sqrt{2}} \cdot \left\| \sqrt{P} - \sqrt{Q} \right\|_2$$

and is bounded by 0 and 1.

In the case of frequency distributions from census tables, where $F = (F_1, F_2, \ldots, F_K)$ is the vector of original counts and $G = (G_1, G_2, \ldots, G_K)$ is the vector of perturbed counts, and $\sum_{i=1}^{K} F_i = N$ and $\sum_{i=1}^{K} G_i = M$, the Hellinger distance is defined as:

$$HD(F, G) = \frac{1}{\sqrt{2}} \cdot \left\| \sqrt{F} - \sqrt{G} \right\|_2 = \frac{1}{\sqrt{2}} \cdot \sqrt{\sum_{i=1}^{K} \left(\sqrt{F_i} - \sqrt{G_i} \right)^2} \tag{4}$$

The Hellinger distance is grounded in Information Theory and takes into account the magnitude of the cells since the difference between square roots of two "large" numbers is smaller than the difference between square roots of two "small" numbers, even if these pairs have the same absolute difference. Naturally, while the lower bound remains zero,

the upper bound of this distance metric changes:

$$HD(F,G) = \frac{1}{\sqrt{2}} \cdot \sqrt{\sum_{i=1}^{K} \left(\sqrt{F_i} - \sqrt{G_i}\right)^2} = \frac{1}{\sqrt{2}} \cdot \sqrt{\sum_{i=1}^{K} \left(F_i + G_i - 2 \cdot \sqrt{F_i \cdot G_i}\right)}$$

$$= \frac{1}{\sqrt{2}} \cdot \sqrt{N + M - 2 \cdot \sum_{i=1}^{K} \sqrt{F_i \cdot G_i}} \leq \sqrt{\frac{N+M}{2}}.$$

Since the SDC methods described in Section 3 produce approximately the same overall population total N due to controlled methods of perturbation, the Hellinger distance is bounded by 0 and $\sqrt{N}$. For the comparison study in Section 5, we use the expression of $1 - (HD(F,G)/\sqrt{N})$ as the data utility measure, which is bounded between 0 and 1, 0 representing low utility and 1 representing high utility.

5. A Comparison Study

In this section we present a flexible table-generating server for census tables where we proceed with the European Census Hub approach of defining a large hypercube as the underlying data input to the server. We compare the application of SDC methods described in Section 3 to four generated output tables and examine the properties of the disclosure risk and data utility measures presented in Section 4.

5.1. Preparing the Underlying Hypercube and Generating Output Tables

For the comparison study, we generate a hypercube with an underlying population of size 1,500,000 individuals for two NUTS2 regions. The variables defining the hypercube follow one of Eurostat's specifications for a hypercube required by all Member States as follows:

- NUTS2 Region – 2 regions
- Gender – 2 categories
- Banded age groups – 21 categories
- Current employment status – 5 categories
- Occupation – 13 categories
- Educational attainment – 9 categories
- Country of citizenship – 5 categories

From the UK Census 2001, cell proportions from published tables were calculated and cross-classified using iterative proportional fitting. We then multiplied the proportions by our population size of 1,500,000 individuals to produce the final hypercube. The hypercube used in the comparison study has 245,700 cells. The distribution of cell counts is skewed with a large proportion of zero cells as seen in Table 1.

The distribution of cell counts in the hypercube as shown in Table 1 was comparable to the hypercube based on real census data produced by the United Kingdom according to the above specification.

Table 1. Distribution of cell counts in the generated hypercube

Cell value	Number of cells	Percentage of cells
0	226,939	92.4
1	4,028	1.6
2	2,112	0.9
3–5	2,964	1.2
6–8	1,664	0.7
9–10	720	0.3
11 and over	7,273	3.0
Total	245,700	100.0

In the flexible table-generating server of our comparison study, we apply a set of preliminary SDC rules for generating tables and allow a maximum of three dimensions with one additional variable to define the population of the table. Four different-size output tables are generated from the input hypercube as follows (number of categories of each variable are in parenthesis):

(1) Selected population: NUTS2 = 1, table spanned by: Banded age group (21) * Educational Attainment (9) * Occupation (13).
(2) Selected population: NUTS2 = 2, table spanned by: Gender (2) * Banded age group (21) * Country of citizenship (5)
(3) Selected population: Gender = 1, table spanned by: Current activity status (5) * Occupation (13) * Educational attainment (9)
(4) Selected population: Banded age group = 10, table spanned by: NUTS2 (2) * Occupation (13) * Educational attainment (9)

Table 2 contains details of the four generated output tables that are used in the comparison study: the total size of the population, the number of cells and the average cell size in each table as well as the distribution of cell counts.

Table 2. Details of four generated tables to be used in the comparison study

Details	Table 1	Table 2	Table 3	Table 4
Total Population	854,539	645,461	736,355	96,656
Number of cells	2,457	210	585	234
Average cell size	347.8	3,073.6	1,258.7	413.1
Number of	%	%	%	%
Zeroes	1,534 (62.4)	49 (23.3)	275 (47.0)	84 (35.9)
Ones	44 (1.8)	14 (6.7)	16 (2.7)	9 (3.9)
Twos	35 (1.4)	2 (1.0)	9 (1.5)	4 (1.7)
Threes	27 (1.1)	5 (2.4)	3 (0.5)	6 (2.6)
Fours	20 (0.8)	4 (1.9)	9 (1.5)	1 (0.4)
Fives	17 (0.7)	1 (0.5)	5 (0.9)	4 (1.7)
Sixes and over	780 (31.8)	135 (64.3)	268 (45.8)	126 (53.9)

From Table 2, output Table 1 is the largest table with the largest proportion of zero cells. Output Tables 2 and 4 are similar in the number of cells but the size of the population is considerably smaller in output Table 4, resulting in a larger proportion of zero cells and a smaller proportion of cells of value one. Output Table 3 is a midsize table. It is clear from the small cell counts and many zero cells that the generated output tables require the application of SDC methods in the flexible table-generating server.

In the comparison study we provide an example of how a statistical agency might go about assessing different SDC methods for a flexible table-generating server of census tables through disclosure risk and data utility measures. In the pretabular approach of protecting the input hypercube prior to generating tables, we apply three SDC methods as follows:

- Full random rounding of the hypercube to base 3 semicontrolled to the two NUTS2 totals.
- Random record swapping carried out by first constructing microdata of individuals from the hypercube where each cell is duplicated to the number of individuals in the cell. A random sample of five percent of individuals is selected in each NUTS2 region, then randomly paired with individuals in the opposite NUTS2 region and their geographical variables swapped. This produced a total swap rate of ten percent of individuals having their NUTS2 regions swapped. Following the record-swapping procedure, the hypercube is reconstructed.
- Stochastic perturbation on the hypercube is based on an invariant probability transition matrix with controls in the overall totals of the two NUTS2 regions. The perturbation is carried out on cells of values in the range 0–10; all cells above a value of 10 have the same probabilities of perturbation depending on their residual value to base 5. The probability transition matrix for each NUTS2 region used in this study is presented in Table 3.

The pretabular disclosure-controlled hypercubes are used as input to the flexible table-generating server and the four output tables generated under each SDC method. The comparison results also include the case where a post-tabular SDC method of semicontrolled random rounding to base 3 is applied directly to the four output tables that are generated from the original unperturbed hypercube. The SDC methods are compared through the disclosure risk and data utility measures described in Section 4.

5.2. *Results of the Comparison Study*

To compare the pretabular SDC methods applied to the original hypercube (record swapping, semicontrolled random rounding and stochastic perturbation), we first assess the impact of the perturbation on the small cells in the generated output tables. Table 4 presents the number of small cells of size 1 and 2 in the original hypercube and in each of the four output tables defined in Subsection 5.1, and the percentage of those cells that were altered under the SDC methods. Record swapping generally provided the least number of small cells perturbed except for output Table 4, where the swapped variable NUTS2 is used as a spanning variable of the table. Output Table 3 did not include the swapped NUTS2 variable and hence all cells in the table contain original cell counts. Random rounding eliminates all small cells of size 1 and 2 and provides more protection compared

Table 3. Probability transition matrix used to perturb the hypercube in each NUTS2

		Perturbed Counts											Residual of count to base 5 is:				
Cell Value		0	1	2	3	4	5	6	7	8	9	10	1	2	3	4	5
Original Counts	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	1	0	0.5000	0.1250	0.1250	0.1250	0.1250	0	0	0	0	0	0	0	0	0	0
	2	0	0.1250	0.5000	0.1250	0.1250	0.1250	0	0	0	0	0	0	0	0	0	0
	3	0	0.0167	0.0167	0.9000	0.0167	0.0167	0.0167	0.0167	0	0	0	0	0	0	0	0
	4	0	0.0167	0.0167	0.0167	0.9000	0.0167	0.0167	0.0167	0	0	0	0	0	0	0	0
	5	0	0	0.0167	0.0167	0.0167	0.9000	0.0167	0.0167	0.0167	0	0	0	0	0	0	0
	6	0	0	0	0.0167	0.0167	0.0167	0.9000	0.0167	0.0167	0.0167	0	0	0	0	0	0
	7	0	0	0	0	0.0167	0.0167	0.0167	0.9000	0.0167	0.0167	0.0167	0	0	0	0	0
	8	0	0	0	0	0	0.0167	0.0167	0.0167	0.9000	0.0167	0.0167	0.0167	0	0	0	0
	9	0	0	0	0	0	0	0.0167	0.0167	0.0167	0.9000	0.0167	0.0167	0.0167	0	0	0
	10	0	0	0	0	0	0	0	0.0167	0.0167	0.0167	0.9000	0.0167	0.0167	0.0167	0	0
Residual of count	1	0	0	0	0	0	0	0	0	0.0167	0.0167	0.0167	0.9000	0.0167	0.0167	0.0167	0
to base 5 is:	2	0	0	0	0	0	0	0	0	0	0.0167	0.0167	0.0167	0.9000	0.0167	0.0167	0.0167
	3	0	0	0	0	0	0	0	0	0	0.0167	0.0167	0.0167	0.0167	0.9000	0.0167	0.0167
	4	0	0	0	0	0	0	0	0	0	0.0167	0.0167	0.0167	0.0167	0.0167	0.9000	0.0167
	5	0	0	0	0	0	0	0	0	0	0.0167	0.0167	0.0167	0.0167	0.0167	0.0167	0.9000

Table 4. Number of small cells of size 1 and 2 in original hypercube and generated tables, and percentage of those cells that were perturbed

	Original hypercube	Table 1	Table 2	Table 3	Table 4
Number of cells of size 1 and 2	6140	79	16	25	13
Percentage perturbed:					
Record swapping	26.9	15.2	12.5	0	30.8
Stochastic perturbation	33.2	29.1	25.0	36.0	23.1
Random rounding	100	100	100	100	100

to record swapping and the stochastic perturbation. It is well known, however, that random rounding has the risk of being able to reveal original cell values, especially when the sum of rounded cells does not add up to the rounded marginal totals. However, ensuring the consistency of the rounding across same cells in different tables and controlling some of the marginal totals lowers the risk of being able to reveal original cell values.

Table 5 presents the disclosure risk measure in (3) and the Hellinger distance in (4) for the output tables defined in Subsection 5.1 generated on the pretabular disclosure-controlled hypercubes according to the SDC methods: record swapping, semicontrolled random rounding and stochastic perturbation. In addition, we report the measures for the case where the SDC method of semicontrolled random rounding is applied directly to the output tables that were generated from the original hypercube to compare the pretabular and post-tabular approach for this SDC method.

To modify the second term in the disclosure risk measure in (3) following the SDC methods as described in Subsection 4.2, we used the following multipliers: for record swapping, the average diagonal probability of the probability transition matrix is 0.9; for the stochastic perturbation, the average diagonal probability of the probability transition matrix is 0.75 for the small counts and 0.9 for the large counts; for the random rounding to base 3, we use the multiplier of 0.33.

From Table 5, we see that the disclosure risk measures are all smaller for the perturbed tables compared to the original tables, even for the case of record swapping in output Table 3 where the perturbed table is identical to the original table since the perturbed NUTS2 variable was not included as a spanning variable of the table. The utility measures are all high, showing that all SDC methods can provide tables that are fit for purpose for users.

In general, it is clear that the method of record swapping when applied to the input hypercube did little to reduce disclosure risk in the final output tables in the comparison study. However, the disclosure risk measure is always slightly smaller than the disclosure risk measure of the original table to reflect the uncertainty in the table based on the assumption that an intruder cannot be certain which variables were swapped. The data utility measure based on the Hellinger distance for output Table 3 under record swapping is 1.00, since the perturbed table is equal to the original table. The data utility measure under record swapping was low for the two output Tables 1 and 2 where the perturbed

Table 5. Disclosure risk and data utility (Hellinger distance) for the generated tables

	Disclosure risk $R(F, w_1, w_2)$ in (3)	Data utility $1 - (HD(F, G)/\sqrt{N})$ in (4)
Table 1		
Original	0.318	-
Perturbed input		
Record swapping:	0.282	0.988
Semicontrolled random rounding	0.137	0.991
Stochastic perturbation	0.239	0.995
Perturbed output:		
Semicontrolled random rounding	0.135	0.993
Table 2		
Original	0.248	-
Perturbed input:		
Record swapping	0.191	0.972
Semicontrolled random rounding	0.070	0.996
Stochastic perturbation	0.210	0.998
Perturbed output:		
Semicontrolled random rounding	0.072	0.996
Table 3		
Original	0.339	-
Perturbed input:		
Record swapping	0.295	1.000
Semicontrolled random rounding	0.130	0.994
Stochastic perturbation	0.254	0.996
Perturbed Output:		
Semicontrolled random rounding	0.127	0.996
Table 4		
Original	0.298	-
Perturbed input:		
Record swapping	0.271	0.987
Semicontrolled random rounding	0.105	0.991
Stochastic perturbation	0.229	0.994
Perturbed output:		
Semicontrolled random rounding	0.105	0.992

NUTS2 variable was used to select the population for these tables. The data utility measure under record swapping for output Table 4 was slightly higher, since in this case NUTS2 was a variable spanning the table and hence did not change the overall total of the table.

The stochastic perturbation carried out on the input hypercube outperformed record swapping with smaller disclosure risk measures and higher data utility measures (except for output Table 3). The stochastic perturbation has a higher disclosure risk compared to semicontrolled random rounding, since a large percentage of small cells are unchanged by the perturbation, but it has higher data utility.

The semicontrolled random rounding outperformed all other methods with respect to the lowest disclosure risk, since there are no small cells in the tables and attribute disclosure risk is reduced by the introduction of random zeros. However, the data utility measure based on the Hellinger distance was slightly lower compared to the stochastic perturbation method as mentioned above. There was little difference between the disclosure risk measures comparing the pretabular semicontrolled random rounding on the input hypercube to the post-tabular semicontrolled random rounding applied directly to the output tables generated from the original hypercube. However, there is an increase in the data utility measure when applying the post-tabular semicontrolled random rounding, especially for the large output Table 1 and midsize output Table 3.

Figure 2 presents a disclosure risk-data utility map of the four generated tables where RS is record swapping, SP is the stochastic perturbation, RR is the semicontrolled random rounding on the input hypercube and RRP is the semicontrolled random rounding applied directly to the generated output tables. The data utility measure is the Hellinger distance in (4). The upper right-hand quadrant of the map represents high disclosure risk and high utility and the lower left-hand quadrant represents low disclosure risk and low data utility.

The statistical agency needs to decide on a tolerable disclosure risk threshold above which they are not prepared to release a table. As an example, the disclosure risk-data utility map shows that for a tolerable disclosure risk threshold of up to 15 percent, the

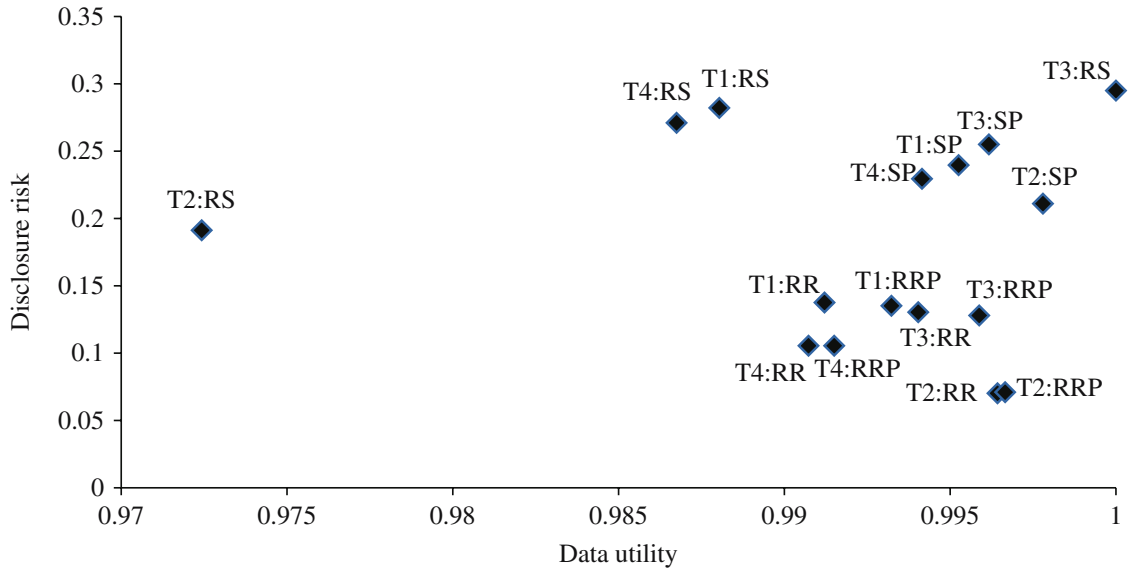

Fig. 2. Disclosure risk – data utility map for generated tables (output Table 1 (T1) to output Table 4 (T4)): RS – record swapping, SP – stochastic perturbation, RR – semicontrolled random rounding on input hypercube, RRP – semicontrolled random rounding on generated tables

output tables where semicontrolled random rounding was applied directly to tables that were generated from the original hypercube have the highest data utility as they are on the farthest right-hand side of the map.

6. Concluding Remarks

In this article, we have compared pretabular SDC methods applied to a large hypercube (record swapping, stochastic perturbation and semicontrolled random rounding) and a semicontrolled random rounding applied directly to output tables generated from the original hypercube. For the pretabular SDC methods, record swapping had little impact on reducing disclosure risk and also had lower data utility. Semicontrolled random rounding offered more protection as all cell values in the table not a multiple of base b are perturbed, and by preserving the consistency of cells across tables, it is more difficult to undo the rounding to reveal original cell values. The stochastic perturbation had the best overall data utility, but entailed higher disclosure risks compared to the semicontrolled random rounding. Finally, we have seen that the post-tabular SDC method of semicontrolled random rounding applied directly to the generated output tables produced nearly the same amount of disclosure risk reduction as the pretabular semicontrolled random rounding applied to the input hypercube, but had a higher level of data utility.

The aim of the comparison study was not primarily to evaluate specific SDC methods or indeed determine their optimum parameterization, but rather to demonstrate how such a disclosure risk and data utility analysis should be carried out by a statistical agency when disseminating census data. To this end, we have proposed new global measures of disclosure risk and data utility based on information theory that are particularly suited for assessing disclosure risk arising from attribute and identity disclosure in census frequency tables and can easily be embedded in a web-based flexible table-generating server. The proposed modifications to the disclosure risk measure following the application of an SDC method show that we can reflect the level of uncertainty added to the tables and therefore reduce the disclosure risk. Further research is needed to refine and improve post-tabular SDC methods whilst preserving additivity and consistency of user-defined tables. More extensive empirical studies are needed that involve real data and the testing of SDC methods across their respective parameter spaces.

Another key aspect of the SDC problem in a flexible table-generating server is the management of users and governance processes. The server can be freely available on the statistical agency's website for all users or restricted via licensing and passwords to only approved users. For the former case, it is clear that SDC rules and methods would have to be highly protective to guard against the fact that users can query the same table multiple times in an attempt to undo SDC methods and reveal original cell counts. Clearly, perturbative SDC methods, preserving the additivity and consistency of same cells across different tables, and high thresholds for dissemination would be required. For the latter case, less protection would be needed, allowing for higher-quality data, but protocols would then need to be in place to handle multiple overlapping queries from the same user, the management of users and their expectations.

7. References

Antal, L., N. Shlomo, and M. Elliot. 2014. "Measuring Disclosure Risk with Entropy in Population Based Frequency Tables." In *PSD'2014 Privacy in Statistical Databases*, edited by J. Domingo-Ferrer, 62–78. Berlin: Springer.

Cover, T.M. and J.A. Thomas. 2006. *Elements of Information Theory*, 2nd ed. New York: Wiley.

Dalenius, T. and S.P. Reiss. 1982. "Data Swapping: A Technique for Disclosure Control." *Journal of Statistical Planning and Inference* 7: 73–85.

Doyle, P., J.I. Lane, J.M.M. Theeuwes, and L. Zayatz. 2001. *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*. Amsterdam: Elsevier Science B.V.

Duncan, G., M.J. Elliot, and J.J. Salazar. 2011. *Statistical Confidentiality: Principles and Practice*. New York: Springer.

Fienberg, S.E. and J. McIntyre. 2005. "Data Swapping: Variations on a Theme by Dalenius and Reiss." *Journal of Official Statistics* 9: 383–406.

Fraser, B. and J. Wooton. 2005. "A Proposed Method for Confidentialising Tabular Output to Protect Against Differencing." Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality, Geneva, November 9–11. Available at: www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2005/wp.35.e.pdf (accessed April 2015).

Gouweleeuw, J., P. Kooiman, L.C.R.J. Willenborg, and P.P. De Wolf. 1998. "Post Randomisation for Statistical Disclosure Control: Theory and Implementation." *Journal of Official Statistics* 14: 463–478.

Hundepool, A., J. Domingo-Ferrer, L. Franconi, S. Giessing, E. Schulte Nordholt, K. Spicer, and P.P. de Wolf. 2012. *Statistical Disclosure Control*. Chichester: John Wiley & Sons.

Salazar-Gonzalez, J.J., C. Bycroft, and A.T. Staggemeier. 2005. "Controlled Rounding Implementation." Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality, Geneva, November 9–11. Available at: www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2005/wp.36.pdf (accessed April 2015).

Shlomo, N. 2007. "Statistical Disclosure Control Methods for Census Frequency Tables." *International Statistical Review* 75: 199–217. Doi: http://dx.doi.org/10.1111/j.1751-5823.2007.00010.x.

Shlomo, N. and C. Young. 2008. "Invariant Post-tabular Protection of Census Frequency Counts." In *In PSD'2008 Privacy in Statistical Databases*, edited by J. Domingo-Ferrer and Y. Saygin, 77–89. Berlin: Springer.

Willenborg, L.C.R.J. and T. de Waal. 2001. *Elements of Statistical Disclosure Control*. New York: Springer.

Received July 2013
Revised October 2014
Accepted November 2014

Chapter 6

Discussion

This chapter discusses the findings of the thesis and their consequences. It aims to: provide a quick overview of the results we described in detail in the preceding chapters; point out what has not been achieved; draw a conclusion and provide potential problems for further research.

Section 6.1 provides a short summary of the main findings. Section 6.2 discusses how our results relate to the SDC literature. Section 6.3 describes problems stemming from our research. The concluding notes in Section 6.4 close the thesis.

6.1 Summary

The objective of a statistical institute is to disseminate data collected from a sample or population of units. However, data also can be potentially disclosive. For example, an intruder might be able to match some parts of the disseminated data with a particular individual or a group of individuals based on prior knowledge or the availability of a publicly available dataset with shared quasi-identifiers. Statistical disclosure control methods are applied to the data before dissemination in order to prevent a potential intruder from matching the data to individuals.

A statistical institute must decide whether datasets can be released

without providing an opportunity for an intruder to breach confidentiality. Disclosure risk measures can be used to assess the risk of disclosure. Such measures, when applied to the data to be released, provide a numerical value and quantify the degree of disclosure risk. If the value is high, the data cannot be released. Statistical disclosure control (SDC) methods are applied to protect the data. They vary according to the data. SDC methods for microdata differ from those for tabular data because the released data are more detailed and therefore have more disclosure risk. For tabular data, pre-tabular or post-tabular SDC methods may be used. As their names suggest, pre-tabular methods are applied to microdata before tabulation, while post-tabular methods alter tables generated from the original microdata. Data might be released after an SDC method is applied to them. However, disclosure controlled data differ from the original data, therefore an information loss measure needs to be employed to quantify the degradation in the data.

The main point of Chapter 3 is to define a disclosure risk measure that fulfils the five properties listed in the introduction of the paper presented in Section 3.9. The $R_1(F, \boldsymbol{w})$ and $R_2(F, G, \boldsymbol{w})$ disclosure risk measures possess the properties before and after perturbation, respectively. They are defined as the weighted average of three terms which cover the properties. The weights allow a data protector to emphasize the property he/she considers the most important. The $R_1(F, \boldsymbol{w})$ and $R_2(F, G, \boldsymbol{w})$ measures can improve on the GAD and $WGAD$ measures defined in the paper presented in Section 3.9.

The core term of the disclosure risk measure depends on information theoretical expressions, such as the entropy and the conditional entropy. Chapter 3 also presents the main theoretical background of the disclosure risk measure, in particular the method to define the conditional entropy. The conditional entropy sheds light on the relationship of pre-tabular and post-tabular SDC methods. More discussion on pre- and post-tabular SDC methods can be found in Section 6.3.

The disclosure risk assessment of population based tables is more straightforward than that of sample based tables. The main difference is in the sets of individuals contributing to the tables. While population based tables include every individual, sample based tables include only a random selection of the population and generally the underlying population frequencies are unknown and need to be estimated. This means that an intruder gains more information from a population based table and does not need to have prior knowledge about a particular individual's inclusion into the table.

The estimation of population based frequencies from sample based frequencies is the main point of the paper presented in Chapter 4. The disclosure risk measure can be calculated on the estimated population frequencies. While in Chapter 3 the X random variable is given and the Y variable is to be determined, in Chapter 4 the Y variable is given and the X variable is not. X provides the cell where the individuals fall originally, while Y shows the cells the individuals contribute to after perturbation.

The paper presented in Chapter 5 provides an application of the proposed disclosure risk measure and associated information loss measure. It shows how a statistical institute should carry out a disclosure risk assessment by optimising the SDC paradigm of minimum disclosure risk according to a threshold and minimizing information loss for an automatic web-based flexible table generating server. The disclosure risk and information loss measures can be calculated automatically on-the-fly. A user requests a table and an iterative process is carried out within the server of assessing disclosure risk, applying SDC methods and reassessing disclosure risk and information loss until the table can be released to the user. The technical implementation of a flexible table generator tool that is able to assess and release data without human interaction is not covered in the thesis.

6.2 Relation to SDC Literature

The disclosure risk measure defined in Chapter 3 is based on properties listed in the introduction of the paper presented in Section 3.9. Properties 1A, 1B and 2 are considered true more widely and are often used in the SDC literature. Properties 3 and 4 are technical requirements. The role of any disclosure risk measure is in the decision about data releases. Our disclosure risk measure is not an exception. Tabular data can be released if $R_1(F, \boldsymbol{w})$ or $R_2(F, G, \boldsymbol{w})$ is low according to a threshold that would be determined by the statistical institute.

The $R_2(F, G, \boldsymbol{w})$ disclosure risk measure is developed for post-tabular SDC methods, such as CTA and various forms of rounding. However, it is not compatible with cell suppression because the disclosure risk measure assumes that the cell values of a perturbed table are not suppressed but altered. In the case of cell suppression, the un-suppressed cell values are not modified and the other cells are blanked out. CTA takes the cell suppression a step further in that it imputes a value for the blanked out cells. Cell suppression and CTA are not considered in the thesis since these methods are more common in magnitude tables and are generally not applied to frequency tables. However, since CTA provides an analytically similar perturbed table to the original frequency table, the $R_2(F, G, \boldsymbol{w})$ disclosure risk measure would not be greater than $R_1(F, \boldsymbol{w})$. Any form of rounding changes the 'structure' of the frequencies since only multiples of the rounding base can appear in a rounded frequency table. It does not change the fact that the perturbed table has lower or equal disclosure risk measure to the original frequency table. Therefore the value of the disclosure risk measure on a rounded frequency table can also be different from that of a perturbed table that has similar frequencies to the original table.

A commonly used sensitivity measure defined for frequency tables is the threshold rule. This is a measure that is defined at the cell level. The comparison of the $R_1(F, \boldsymbol{w})$ and $R_2(F, G, \boldsymbol{w})$ disclosure risk

measures to the threshold rule is difficult since they are defined at the row/column/table level and assess different scenarios of disclosure risk. According to the threshold rule, a perturbed frequency table cannot be released if a certain proportion of cell values is lower than the prescribed minimum. The threshold rule aims to minimise the risk of identification. Rounding for example might eliminate all cell values below the threshold but CTA may still have small values remaining in the tables. On the other hand, the risk measures $R_1(F, \boldsymbol{w})$ and $R_2(F, G, \boldsymbol{w})$ assess the distributions in the table and show how close those distributions are to the degenerate distribution (only one non-zero cell value in the row/column/table). This provides a measure for attribute disclosure.

Frank (1978) also defined a disclosure risk measure based on the entropy, see Section 2.3.2.2.1. However, his assumptions about a prior disclosure set and an intruder's knowledge are slightly unrealistic and simplify the situation in our opinion. Contrary to Frank's approach, we adopt the viewpoint of a statistical institute. The disclosure risk is also measured from the perspective of a statistical institute. Potential intruders and their knowledge are not the focus of the thesis. However, disclosure risk scenarios discussed in Frank (1976) are accepted and used in the thesis.

Our proposed disclosure risk measure can be applied to a set of internal cells as well as to marginal tables. However, it cannot measure the disclosure risk of single cells, in contrast with the information theory based disclosure risk measure in Oganian and Domingo-Ferrer (2003).

The SAP method (Smith and Elliot (2008)), contrary to the approach laid out in this thesis, adopts the intruder's perspective. Cell values of zero, however, play an important role in both approaches. One advantage of the approach described here over SAP is computational tractability. SAP is very computationally intensive and is not feasible for the types of on-the-fly output I consider in Chapter 5. SAP also focuses on the risk of a single zero being discoverable in the table and does not take account

of accumulative risk from multiple vulnerabilities unlike the measures developed here.

Our proposed disclosure risk measure is defined mainly for attribute disclosure. Small cell values, either in population based or in sample based tables, increase the value of our disclosure risk measure. However, the overall disclosure risk measure might not be high, even if there are some small cell frequencies in the table. Therefore, the presence of a small cell might not imply the need to apply an SDC method immediately. This viewpoint is in line with the UK Office for National Statistics who deemed attribute disclosure in the form of degenerate distributions as potentially more disclosive than small internal cell values in a table.

While the SDC literature for sample based tables (and microdata) treats the estimation of population uniques as an important problem, the main concern in the thesis is the 'estimation of zeroes' since they determine potentially degenerate distributions in the table and lead to attribute disclosure. Population uniques increase the value of our disclosure risk measure but they do not necessarily have to be eliminated in a perturbed table according to the UK perspective. The number of zeroes is a key term in our proposed disclosure risk measure. In fact, it is not just the number of zero cells but their location in the table that is important since it influences the $H(X|Y)$ conditional entropy in the disclosure risk measure.

In summary, SDC approaches vary depending on the disclosure risk scenarios, type of data, type of disclosure risk and whether we take the statistical institute or intruder perspective. This thesis focuses on the disclosure risk of attribute disclosure for frequency tables from a statistical institute perspective and focuses on both population based tables and sample based tables.

6.3 Future Work

The disclosure risk measure we introduce has its limitations. It is based on some defined desirable properties and those properties are open to criticism. The list of properties in the introduction of the paper presented in Section 3.9 might be extended or narrowed and the disclosure risk measure might be changed accordingly.

Below we address some topics for future work.

- Topic 1

  The conditional entropy is calculated on the $E(Z_{ij})$ average, where $E(Z_{ij})$ is the expectation of the $Pr(Y = c_j|X = c_i)$ values, see Section 3.6.2.1. The R_G distribution (see Section 3.6.2.1) on the set of potential Y variables (Ω_G, see Section 3.2) is crucial in determining $E(Z_{ij})$. Theorems 1, 2, 3 and 4 in Sections 3.6.2.1 and 3.6.2.2 can be proven with relative ease because R_G is chosen as a simple uniform distribution (U_{Ω_G} or $U_{\Omega^*_G}$). Further work needs to be carried out to find other R_G distributions that may express the disclosure risk more accurately. However, a more complex distribution might not provide a closed formula for $E(Z_{ij})$, therefore $E(Z_{ij})$ might have to be calculated numerically.

- Topic 2

  In Section 2.4.1 we discussed the advantages and drawbacks of pre- and post-tabular SDC methods. A post-tabular method can be given by the original frequency table F and the set of the following conditional probabilities.

  $$Pr(\text{the perturbed frequency table is } G|\text{the original frequency table is } F)\ , \tag{6.3.1}$$

  where $G \in PG = \{G : G = (G_1, G_2, \ldots, G_K) \in \mathbb{Z}^K\}$. Although PG is countably infinite, in practice only a finite number of the above

probabilities differ from zero for an arbitrarily selected post-tabular method.

A pre-tabular method is based on the $Pr(X = c_i|Y = c_j)$ conditional probabilities directly. Indirectly, it provides the probabilities in (6.3.1), therefore a pre-tabular method always implies a post-tabular method (given the original microdata set and the table-spanning variables). However, it cannot be reversed, not every post-tabular method has a corresponding pre-tabular method.

Figure 3.1 in Section 3.7 shows that pre- and post-tabular methods are not as far from each other as it seems. Further work needs to be carried out to prove whether there are post-tabular methods that are 'equivalent' to pre-tabular methods, that is, whether the (6.3.1) conditional probabilities are the same for a pre-tabular and a post-tabular SDC method. If there are such methods, then, in principle, a statistical institute could benefit from the advantages of both pre- and post-tabular methods discussed in Section 2.4.1. Below we describe some initial thoughts to link pre-tabular and post-tabular SDC methods.

A pre-tabular method, such as PRAM, defines the $p_{ij} = Pr(Y = c_j|X = c_i)$ conditional probabilities. PRAM perturbs the individuals independently from each other. A perturbed frequency table is then generated from the perturbed microdata. Since the number of individuals is not changed by the perturbation, we obtain $\sum_{i=1}^{K} F_i = \sum_{j=1}^{K} G_j$. By definition, $|\{a \in I : X(a) = c_i\}| = F_i$. Consider the number of individuals that originally fall in c_i and after applying PRAM they contribute to c_j. Denote the number by F_i^j.

$$F_i^j = |\{a \in I : X(a) = c_i, Y(a) = c_j\}| .$$

Now, $\sum_{j=1}^{K} F_i^j = F_i$. On the other hand, $\sum_{i=1}^{K} F_i^j = G_j$. We first fix G. In this case, the F_i^j values define an (adjacency) matrix, where two marginal tables are F and G. The following equation can be proven,

provided that the individuals are perturbed independently from each other.

$$Pr(\text{perturbed frequency table is } G|\text{original frequency table is } F) = \sum_{\text{Potential } (F_i^j) \text{ matrices}} \prod_{i=1}^{K} \frac{F_i!}{F_i^1! \cdot F_i^2! \cdot \ldots \cdot F_i^K!} \cdot p_{i1}^{F_i^1} \cdot p_{i2}^{F_i^2} \cdot \ldots \cdot p_{iK}^{F_i^K} \ . \quad (6.3.2)$$

Here by 'potential (F_i^j) matrices' we mean matrices where F and G are marginal tables, that is, the sums of rows provide F and the sums of columns G. The $\frac{F_i^j}{F_i}$ proportion should be close to p_{ij}, since p_{ij} provides the probability that an individual that originally falls into c_i contributes to c_j after PRAM is applied. Therefore,

$$F_i^j \approx p_{ij} \cdot F_i \ . \quad (6.3.3)$$

Consider now the situation where the G perturbed table is *not* fixed. Still, F is one of the marginal tables, $F_i = \sum_{j=1}^{K} F_i^j$, but the 'other' marginal table is not fixed. Therefore, any partition $F_i = F_i^1 + F_i^2 + \cdots + F_i^K$, $i = 1, 2, \ldots, K$ is admissible. Consider a partition for each F_i, $i = 1, 2, \ldots, K$. They still provide a matrix and the sums of the columns provide the G perturbed table. The p_{ij} probabilities define a probability distribution on the set of the (F_i^j) matrices. The distribution provides the $Pr(\text{perturbed frequency table is } G|\text{original frequency table is } F)$ probability for each G table.

Clearly, (6.3.2) assumes that the frequencies follow a multinomial distribution. (6.3.3) provides an alternative idea. Assume that the F_i^j frequency is the result of a draw from a Poisson distribution as follows.

$$F_i^j \sim Po(p_{ij} \cdot F_i) \ .$$

The F_i^j, $i, j = 1, 2, \ldots, K$ frequencies are, by assumption, independent

from each other. Since $\sum_{j=1}^{K} p_{ij} = 1$ and the sum of independent Poisson-distributed numbers follows a Poisson distribution, therefore

$$\sum_{j=1}^{K} F_i^j \sim Po(F_i) \ .$$

It implies that the sum of the F_i^j frequencies over j might not be equal to F_i. However, $\sum_{j=1}^{K} E(F_i^j) = F_i$. Summing the F_i^j frequencies 'columnwise' provides the G_j frequencies. We can exploit again that the convolution of independent Poisson distributions is a Poisson distribution.

$$G_j = \sum_{i=1}^{K} F_i^j \sim Po\left(\sum_{i=1}^{K} p_{ij} \cdot F_i\right) \ .$$

The above distribution of G_j provides the $Pr(G_j = l)$, $l = 0, 1, \ldots$ probabilities. The product of the probabilities gives the probability that the 'perturbation' results in the G frequency table.

A post-tabular method is often defined at cell level and the cells are perturbed independently from each other, see for example random rounding. Therefore often

$$Pr(\text{perturbed frequency table is } G | \text{original frequency table is } F) = \prod_{j=1}^{K} Pr(j\text{th perturbed cell value is } G_j | j\text{th original cell value is } F_j) \ . \tag{6.3.4}$$

If (6.3.2) and (6.3.4) were linked, then a pre-tabular method and a post-tabular method could be interconnected.

- Topic 3

  The disclosure risk assessment for sample based tables requires the estimation of population frequencies. We estimated the frequencies by probabilistic models. The disclosure risk measure is sensitive to the

(estimated) population frequencies, therefore the model is important. Further work can be carried out to find other models that provide good estimates of the population frequencies and thereby the disclosure risk measure. The impact of having imprecise estimated population frequencies on the value of the disclosure risk measure is not exactly clear. The difference between the true and estimated values stems from the difference between the first terms and the difference between the second terms of the disclosure risk measure. (We assume that the third term is equal for the two cases.) The absolute difference between the true and estimated first terms (without the w_1 weight) is

$$\left| \left(\frac{|D|}{K} \right)^{\frac{|D \cup E|}{|D \cap E|}} - \left(\frac{|\hat{D}|}{K} \right)^{\frac{|\hat{D} \cup E|}{|\hat{D} \cap E|}} \right| . \tag{6.3.5}$$

It can be seen that the estimated zero cells have a significant influence on the first term. As we know, the estimation of zero cells from a sample based table is always a difficult problem. Here we also need to 'estimate' the place of zero cells in the table since the $|\hat{D} \cup E|$ and $|\hat{D} \cap E|$ expressions depend on where the zeroes are. Therefore the (6.3.5) difference is hard to estimate.

Regarding the second term of the disclosure risk measure, an upper bound of the difference between the true and estimated second terms can be given as follows.

$$\left| \left(1 - \frac{H(X)}{\log K} \right) \cdot \left(1 - \frac{H(X|Y)}{H(X)} \right) - \left(1 - \frac{\widehat{H(X)}}{\log K} \right) \cdot \left(1 - \frac{\widehat{H(X|Y)}}{\widehat{H(X)}} \right) \right| \leq$$

$$\frac{\varepsilon_1 + \varepsilon_2}{\widehat{H(X)}} + \left| \frac{\varepsilon_1 \cdot (H(X) - H(X|Y))}{H(X) \cdot \widehat{H(X)}} \right| . \tag{6.3.6}$$

Here $\varepsilon_1 = |H(X) - \widehat{H(X)}|$ and $\varepsilon_2 = |H(X|Y) - \widehat{H(X|Y)}|$. The proof of

the above formula can be found in the Appendix.

If we assume that the $H(X)$ entropy and the $H(X|Y)$ conditional entropy are estimated 'well' by $\widehat{H(X)}$ and $\widehat{H(X|Y)}$, that is, ε_1 and ε_2 are small, then the above upper bound depends mostly on $H(X) - H(X|Y)$. The $I(X;Y) = H(X) - H(X|Y)$ expression is known as the mutual information of X and Y. According to Cover and Thomas (2006), 'The mutual information $I(X;Y)$ is a measure of the dependence between two random variables. It is symmetric in X and Y and always non-negative and is equal to zero if and only if X and Y are independent.' The mutual information can also be described as the relative entropy between the joint distribution of X and Y and the distribution given by the $Pr(X = c_i) \cdot Pr(Y = c_j)$, $i, j = 1, 2, \ldots, K$ products. (The set of the latter products obviously defines a distribution on $X \times Y$.)

$$I(X;Y) = \sum_{i=1}^{K} \sum_{j=1}^{K} Pr(X = c_i, Y = c_j) \cdot \log \frac{Pr(X = c_i, Y = c_j)}{Pr(X = c_i) \cdot Pr(Y = c_j)} \, .$$

The value of (6.3.6) is small if $I(X;Y) = H(X) - H(X|Y)$ is small. However, the mutual information is zero if the variables are independent. Since we get the Y variable by 'perturbing the X variable', they would not be independent. In fact, the correlation between the two variables would be strong. However, (6.3.6) might still be sufficiently low since $\left|\frac{H(X)-H(X|Y)}{H(X)}\right| = \frac{I(X;Y)}{H(X)} \leq 1$. The key point is to keep the ε_1 and ε_2 values low. Further work should be carried out to estimate ε_1 and ε_2.

The third term of the disclosure risk measure depends on the population size (N) only. This term is not sensitive to small cell values. A table that has many ones and twos and some high frequencies can provide the same value on the third term as a uniformly distributed table of medium size cell values. A potential refinement of the disclosure risk measure might improve on the third term.

- Topic 4

$R_1(F, \boldsymbol{w})$ and $R_2(F, G, \boldsymbol{w})$ do not take the structure of the table into account. We only need a vector of original and a vector of perturbed frequencies to calculate the above quantities. It is important, however, that the cells of the original vector and the cells of the perturbed vector correspond. Further work needs to be carried out to define a disclosure risk measure that takes the table structure into account.

6.4 Concluding Remarks

Any statistical institute or other agency responsible for releasing data must assess the disclosure risk of the data that it is intending to disseminate and therefore must have at its disposal adequate tools for assessing that risk. No disclosure risk measure can be used in every disclosure risk scenario, therefore the choice of the measure is important. Whilst developing a new disclosure risk measure for a given disclosure risk scenario, appropriate SDC methods or combination of SDC methods need to also be developed leading to new insight and experiences. It is only through taking a methodological approach to disclosure risk assessment that statistical institutes can fully ensure the protection of released statistical data. The definition of the new disclosure risk measure outlined in this thesis is a step in the right direction.

Appendix A

Appendices

A.1 Numerical Values for the Third Term of the Disclosure Risk Measure

Numerical results for the $h_1(N, \varepsilon)$, $h_2(N, \varepsilon)$ and $h_3(F, \varepsilon)$ functions can be found below.

F	$h_1(F,0.1)$	$h_2(F,0.1)$	$h_3(F,0.1)$	$h_1(F,0.3)$	$h_2(F,0.3)$	$h_3(F,0.3)$	$h_1(F,0.5)$	$h_2(F,0.5)$	$h_3(F,0.5)$
(1,1)	0.9330	0.9977	1.0000	0.8123	0.9812	1.0000	0.7071	0.9522	1.0000
(2,2)	0.8706	0.9912	0.9977	0.6598	0.9341	0.9812	0.5000	0.8466	0.9522
(5,5)	0.7943	0.9772	0.9884	0.5012	0.8474	0.9150	0.3162	0.6803	0.8071
(10,10)	0.7411	0.9632	0.9772	0.4071	0.7730	0.8474	0.2236	0.5585	0.6803
(20,20)	0.6915	0.9466	0.9632	0.3307	0.6966	0.7730	0.1581	0.4497	0.5585
(50,50)	0.6310	0.9215	0.9408	0.2512	0.5982	0.6722	0.1000	0.3303	0.4180
(75,75)	0.6059	0.9095	0.9297	0.2224	0.5568	0.6285	0.0816	0.2862	0.3647
(100,100)	0.5887	0.9006	0.9215	0.2040	0.5283	0.5982	0.0707	0.2580	0.3303
(0,1)	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
(0,2)	0.9330	0.9977	0.9977	0.8123	0.9812	0.9812	0.7071	0.9522	0.9522
(0,5)	0.8513	0.9884	0.9884	0.6170	0.9150	0.9150	0.4472	0.8071	0.8071
(0,10)	0.7943	0.9772	0.9772	0.5012	0.8474	0.8474	0.3162	0.6803	0.6803
(0,20)	0.7411	0.9632	0.9632	0.4071	0.7730	0.7730	0.2236	0.5585	0.5585
(0,50)	0.6762	0.9408	0.9408	0.3092	0.6722	0.6722	0.1414	0.4180	0.4180
(0,75)	0.6494	0.9297	0.9297	0.2738	0.6285	0.6285	0.1155	0.3647	0.3647
(0,100)	0.6310	0.9215	0.9215	0.2512	0.5982	0.5982	0.1000	0.3303	0.3303
(1,1,1,1,1,1,1,1,1,1)	0.7943	0.9772	1.0000	0.5012	0.8474	1.0000	0.3162	0.6803	1.0000
(2,2,2,2,2,2,2,2,2,2)	0.7411	0.9632	0.9977	0.4071	0.7730	0.9812	0.2236	0.5585	0.9522
(5,5,5,5,5,5,5,5,5,5)	0.6762	0.9408	0.9884	0.3092	0.6722	0.9150	0.1414	0.4180	0.8071
(10,10,10,10,10,10,10,10,10,10)	0.6310	0.9215	0.9772	0.2512	0.5982	0.8474	0.1000	0.3303	0.6803
(20,20,20,20,20,20,20,20,20,20)	0.5887	0.9006	0.9632	0.2040	0.5283	0.7730	0.0707	0.2580	0.5585
(50,50,50,50,50,50,50,50,50,50)	0.5372	0.8710	0.9408	0.1550	0.4440	0.6722	0.0447	0.1837	0.4180
(75,75,75,75,75,75,75,75,75,75)	0.5158	0.8573	0.9297	0.1372	0.4098	0.6285	0.0365	0.1574	0.3647
(100,100,100,100,100,100,100,100,100,100)	0.5012	0.8474	0.9215	0.1259	0.3868	0.5982	0.0316	0.1408	0.3303
(0,0,0,0,0,0,0,0,0,1)	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
(0,0,0,0,0,0,0,0,0,2)	0.9330	0.9977	0.9977	0.8123	0.9812	0.9812	0.7071	0.9522	0.9522
(0,0,0,0,0,0,0,0,0,5)	0.8513	0.9884	0.9884	0.6170	0.9150	0.9150	0.4472	0.8071	0.8071
(0,0,0,0,0,0,0,0,0,10)	0.7943	0.9772	0.9772	0.5012	0.8474	0.8474	0.3162	0.6803	0.6803
(0,0,0,0,0,0,0,0,0,20)	0.7411	0.9632	0.9632	0.4071	0.7730	0.7730	0.2236	0.5585	0.5585
(0,0,0,0,0,0,0,0,0,50)	0.6762	0.9408	0.9408	0.3092	0.6722	0.6722	0.1414	0.4180	0.4180
(0,0,0,0,0,0,0,0,0,75)	0.6494	0.9297	0.9297	0.2738	0.6285	0.6285	0.1155	0.3647	0.3647
(0,0,0,0,0,0,0,0,0,100)	0.6310	0.9215	0.9215	0.2512	0.5982	0.5982	0.1000	0.3303	0.3303
(0,0,0,0,0,1,1,1,2,3)	0.8123	0.9812	0.9984	0.5359	0.8702	0.9875	0.3536	0.7212	0.9693
(0,0,0,0,0,2,2,2,4,6)	0.7579	0.9680	0.9940	0.4353	0.7973	0.9552	0.2500	0.5966	0.8954
(0,0,0,0,0,5,5,5,10,15)	0.6915	0.9466	0.9823	0.3307	0.6966	0.8793	0.1581	0.4497	0.7419
(0,0,0,0,0,10,10,10,20,30)	0.6452	0.9279	0.9697	0.2686	0.6217	0.8087	0.1118	0.3568	0.6185
(0,0,0,0,0,20,20,20,40,60)	0.6020	0.9075	0.9544	0.2182	0.5503	0.7336	0.0791	0.2797	0.5038
(0,0,0,0,0,50,50,50,100,150)	0.5493	0.8784	0.9307	0.1657	0.4636	0.6343	0.0500	0.1998	0.3741
(0,0,0,0,0,75,75,75,150,225)	0.5275	0.8649	0.9191	0.1467	0.4284	0.5918	0.0408	0.1714	0.3255
(0,0,0,0,0,100,100,100,200,300)	0.5125	0.8551	0.9106	0.1346	0.4046	0.5626	0.0354	0.1535	0.2942

Table A.1.1: Values of the $h_1(F,\varepsilon)$, $h_2(F,\varepsilon)$ and $h_3(F,\varepsilon)$ functions on various F frequency tables ($\varepsilon = 0.1, 0.3, 0.5$)

F	$h_1(F,0.8)$	$h_2(F,0.8)$	$h_3(F,0.8)$	$h_1(F,1)$	$h_2(F,1)$	$h_3(F,1)$
(1,1)	0.5743	0.8928	1.0000	0.5000	0.8466	1.0000
(2,2)	0.3299	0.6957	0.8928	0.2500	0.5966	0.8466
(5,5)	0.1585	0.4504	0.6312	0.1000	0.3303	0.5219
(10,10)	0.0910	0.3092	0.4504	0.0500	0.1998	0.3303
(20,20)	0.0523	0.2066	0.3092	0.0250	0.1172	0.1998
(50,50)	0.0251	0.1177	0.1806	0.0100	0.0561	0.0982
(75,75)	0.0182	0.0910	0.1408	0.0067	0.0401	0.0709
(100,100)	0.0144	0.0756	0.1177	0.0050	0.0315	0.0561
(0,1)	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
(0,2)	0.5743	0.8928	0.8928	0.5000	0.8466	0.8466
(0,5)	0.2759	0.6312	0.6312	0.2000	0.5219	0.5219
(0,10)	0.1585	0.4504	0.4504	0.1000	0.3303	0.3303
(0,20)	0.0910	0.3092	0.3092	0.0500	0.1998	0.1998
(0,50)	0.0437	0.1806	0.1806	0.0200	0.0982	0.0982
(0,75)	0.0316	0.1408	0.1408	0.0133	0.0709	0.0709
(0,100)	0.0251	0.1177	0.1177	0.0100	0.0561	0.0561
(1,1,1,1,1,1,1,1,1,1)	0.1585	0.4504	1.0000	0.1000	0.3303	1.0000
(2,2,2,2,2,2,2,2,2,2)	0.0910	0.3092	0.8928	0.0500	0.1998	0.8466
(5,5,5,5,5,5,5,5,5,5)	0.0437	0.1806	0.6312	0.0200	0.0982	0.5219
(10,10,10,10,10,10,10,10,10,10)	0.0251	0.1177	0.4504	0.0100	0.0561	0.3303
(20,20,20,20,20,20,20,20,20,20)	0.0144	0.0756	0.3092	0.0050	0.0315	0.1998
(50,50,50,50,50,50,50,50,50,50)	0.0069	0.0414	0.1806	0.0020	0.0144	0.0982
(75,75,75,75,75,75,75,75,75,75)	0.0050	0.0316	0.1408	0.0013	0.0102	0.0709
(100,100,100,100,100,100,100,100,100,100)	0.0040	0.0260	0.1177	0.0010	0.0079	0.0561
(0,0,0,0,0,0,0,0,0,1)	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
(0,0,0,0,0,0,0,0,0,2)	0.5743	0.8928	0.8928	0.5000	0.8466	0.8466
(0,0,0,0,0,0,0,0,0,5)	0.2759	0.6312	0.6312	0.2000	0.5219	0.5219
(0,0,0,0,0,0,0,0,0,10)	0.1585	0.4504	0.4504	0.1000	0.3303	0.3303
(0,0,0,0,0,0,0,0,0,20)	0.0910	0.3092	0.3092	0.0500	0.1998	0.1998
(0,0,0,0,0,0,0,0,0,50)	0.0437	0.1806	0.1806	0.0200	0.0982	0.0982
(0,0,0,0,0,0,0,0,0,75)	0.0316	0.1408	0.1408	0.0133	0.0709	0.0709
(0,0,0,0,0,0,0,0,0,100)	0.0251	0.1177	0.1177	0.0100	0.0561	0.0561
(0,0,0,0,0,1,1,1,2,3)	0.1895	0.5046	0.9346	0.1250	0.3849	0.9092
(0,0,0,0,0,2,2,2,4,6)	0.1088	0.3502	0.7909	0.0625	0.2358	0.7203
(0,0,0,0,0,5,5,5,10,15)	0.0523	0.2066	0.5414	0.0250	0.1172	0.4286
(0,0,0,0,0,10,10,10,20,30)	0.0300	0.1353	0.3811	0.0125	0.0673	0.2675
(0,0,0,0,0,20,20,20,40,60)	0.0172	0.0873	0.2591	0.0062	0.0380	0.1603
(0,0,0,0,0,50,50,50,100,150)	0.0083	0.0480	0.1501	0.0025	0.0175	0.0782
(0,0,0,0,0,100,75,75,150,225)	0.0060	0.0366	0.1167	0.0017	0.0123	0.0563
(0,0,0,0,0,100,100,100,200,300)	0.0048	0.0302	0.0973	0.0012	0.0096	0.0444

Table A.1.2: Values of the $h_1(F,\varepsilon)$, $h_2(F,\varepsilon)$ and $h_3(F,\varepsilon)$ functions on various F frequency tables ($\varepsilon = 0.8, 1$)

A.2 Proof of Theorem 1

Proof of Theorem 1. If $F_i = 0$, then $Pr(Y_l = c_j | X = c_i) = 0$ for all $l = 1, 2, \ldots, |\Omega_G|$, therefore $E(Z_{ij}) = 0$.

Assume now that $F_i > 0$.

$$E(Z_{ij}) = \sum_{l=1}^{|\Omega_G|} R_G(Y_l) \cdot Pr(Y_l = c_j | X = c_i) = \frac{1}{|\Omega_G|} \cdot \sum_{l=1}^{|\Omega_G|} Pr(Y_l = c_j | X = c_i) =$$

$$\frac{1}{|\Omega_G|} \cdot \sum_{l=1}^{|\Omega_G|} \frac{Pr(X = c_i, Y_l = c_j)}{Pr(X = c_i)} = \frac{N}{F_i \cdot |\Omega_G|} \cdot \sum_{l=1}^{|\Omega_G|} Pr(X = c_i, Y_l = c_j) =$$

$$\frac{N}{F_i \cdot |\Omega_G|} \cdot \sum_{l=1}^{|\Omega_G|} \frac{|\{a \in I : X(a) = c_i, Y_l(a) = c_j\}|}{N} =$$

$$\frac{1}{F_i \cdot |\Omega_G|} \cdot \sum_{l=1}^{|\Omega_G|} |\{a \in I : X(a) = c_i, Y_l(a) = c_j\}| \tag{A.2.1}$$

Note that

$$\sum_{l=1}^{|\Omega_G|} |\{a \in I : X(a) = c_i, Y_l(a) = c_j\}| = |\{(a, Y_l) \in I \times \Omega_G : X(a) = c_i, Y_l(a) = c_j\}| \tag{A.2.2}$$

We need to determine the cardinality of Ω_G. We have $\binom{N}{G_1}$ choices to select the individuals that fall into cell c_1. Once one of the choices is fixed, we have $\binom{N-G_1}{G_2}$ choices to select the individuals that belong to cell c_2, etc. Therefore

$$|\Omega_G| = \binom{N}{G_1} \cdot \binom{N - G_1}{G_2} \cdot \binom{N - G_1 - G_2}{G_3} \cdot \ldots \cdot \binom{N - \sum_{i=1}^{K-1} G_i}{G_K} =$$

$$\frac{N!}{G_1! \cdot (N - G_1)!} \cdot \frac{(N - G_1)!}{G_2! \cdot (N - G_1 - G_2)!} \cdot \frac{(N - G_1 - G_2)!}{G_3! \cdot (N - \sum_{i=1}^{3} G_i)!} \cdot \ldots \cdot \frac{\left(N - \sum_{i=1}^{K-1} G_i\right)!}{G_K! \cdot (N - \sum_{i=1}^{K} G_i)!} \cdot$$

We get the following formula as the cardinality of Ω_G.

$$|\Omega_G| = \frac{N!}{G_1! \cdot G_2! \cdot \ldots \cdot G_K!} \; . \tag{A.2.3}$$

Since $F_i > 0$, we can choose an $a_0 \in I$ individual such that $X(a_0) = c_i$. If $G_j = 0$, then the number of $Y \in \Omega_G$ variables with $Y(a_0) = c_j$ is 0. If $G_j > 0$, then the same number is

$$|\{Y \in \Omega_G : Y(a_0) = c_j\}| = \frac{(N-1)!}{G_1! \cdot \ldots \cdot G_{j-1}! \cdot (G_j - 1)! \cdot G_{j+1}! \cdot \ldots \cdot G_K!} \; .$$

The proof of this formula is the same as for the cardinality of Ω_G. For both $G_j = 0$ and $G_j > 0$ we can write

$$|\{Y \in \Omega_G : Y(a_0) = c_j\}| = \frac{G_j}{N} \cdot |\Omega_G| \; . \tag{A.2.4}$$

There are F_i choices to select the a_0 individual with the $X(a_0) = c_i$ property, therefore

$$|\{(a, Y_l) \in I \times \Omega_G : X(a) = c_i, Y_l(a) = c_j\}| = F_i \cdot \frac{G_j}{N} \cdot |\Omega_G| \; . \tag{A.2.5}$$

Equations (A.2.1), (A.2.2) and (A.2.5) prove the theorem. □

A.3 Proof of Theorem 2

Proof of Theorem 2. If $F_i = 0$, then $Pr(Y_l = c_j | X = c_i) = 0$ for all $l = 1, 2, \ldots, |\Omega_G^*|$, therefore $E(Z_{ij}) = 0$.

Assume now that $F_i > 0$. Similarly to the proof of Theorem 1, we get again that

$$E(Z_{ij}) = \frac{1}{F_i \cdot |\Omega_G^*|} \cdot \sum_{Y_l \in \Omega_G^*} |\{a \in I : X(a) = c_i, Y_l(a) = c_j\}| =$$
$$\frac{1}{F_i \cdot |\Omega_G^*|} \cdot |\{(a, Y_l) \in I \times \Omega_G^* : X(a) = c_i, Y_l(a) = c_j\}| \; . \tag{A.3.1}$$

Our next aim is to determine $|\Omega_G^*|$. First we need to choose $\sum_{k=1}^{K} \min(F_k, G_k)$ individuals that remain in the same cell. The number of the choices is

$$\binom{F_1}{\min(F_1, G_1)} \cdot \binom{F_2}{\min(F_2, G_2)} \cdot \ldots \cdot \binom{F_K}{\min(F_K, G_K)}.$$

If we fix one of the choices, we will know that the selected individuals remain in their cells after perturbation. However, the cells of the not selected individuals are not derived. Similarly to (A.2.3), there are

$$\frac{(N - \sum_{k=1}^{K} \min(F_k, G_k))!}{\prod_{j=1}^{K} (G_j - \min(F_j, G_j))!}$$

possible variables on the not selected individuals. Therefore the cardinality of Ω_G^* is

$$|\Omega_G^*| = \frac{(N - \sum_{k=1}^{K} \min(F_k, G_k))!}{\prod_{j=1}^{K} (G_j - \min(F_j, G_j))!} \cdot \prod_{i=1}^{K} \binom{F_i}{\min(F_i, G_i)}. \qquad \text{(A.3.2)}$$

Since $F_i > 0$, we can select an $a_0 \in I$ individual such that $X(a_0) = c_i$.

If $i = j$ and $G_i = 0$, then $|\{Y_l \in \Omega_G^* : Y_l(a_0) = c_i\}| = 0$. If $i = j$ and $G_i > 0$, then, similarly to (A.3.2),

$$|\{Y_l \in \Omega_G^* : Y_l(a_0) = c_i\}| =$$
$$\frac{\left(N - 1 - \min(F_i - 1, G_i - 1) - \sum_{k \neq i} \min(F_k, G_k))\right)!}{(G_i - 1 - \min(F_i - 1, G_i - 1))! \cdot \prod_{k \neq i} (G_k - \min(F_k, G_k))!} \cdot$$
$$\binom{F_i - 1}{\min(F_i - 1, G_i - 1)} \cdot \prod_{k \neq i} \binom{F_k}{\min(F_k, G_k)} =$$
$$\frac{\left(N - \sum_{k=1}^{K} \min(F_k, G_k)\right)!}{\prod_{k=1}^{K} (G_k - \min(F_k, G_k))!} \cdot \binom{F_i - 1}{\min(F_i, G_i) - 1} \cdot \prod_{k \neq i} \binom{F_k}{\min(F_k, G_k)}.$$

There are F_i choices to fix the a_0 individual. Therefore

$$|\{(a, Y_l) \in I \times \Omega_G^* : X(a) = c_i, Y_l(a) = c_i\}| = F_i \cdot |\{Y_l \in \Omega_G^* : Y_l(a) = c_i\}| =$$
$$F_i \cdot \frac{\left(N - \sum_{k=1}^K \min(F_k, G_k)\right)!}{\prod_{k=1}^K (G_k - \min(F_k, G_k))!} \cdot \binom{F_i - 1}{\min(F_i, G_i) - 1} \cdot \prod_{k \neq i} \binom{F_k}{\min(F_k, G_k)} =$$
$$\min(F_i, G_i) \cdot \frac{\left(N - \sum_{k=1}^K \min(F_k, G_k)\right)!}{\prod_{k=1}^K (G_k - \min(F_k, G_k))!} \cdot \prod_{k=1}^K \binom{F_k}{\min(F_k, G_k)} . \tag{A.3.3}$$

The latter equation also fulfils if $i = j$ and $G_i = 0$.

Combine this result with (A.3.1) and (A.3.2). We get

$$E(Z_{ii}) = \frac{\min(F_i, G_i)}{F_i} .$$

Assume now that $i \neq j$. We will show that if $F_i \leq G_i$ or $F_j \geq G_j$, then $|\{(a, Y_l) \in I \times \Omega_G^* : X(a) = c_i, Y_l(a) = c_j\}| = 0$.

If $F_i \leq G_i$, then $X(a) = c_i$ implies that $Y_l(a) = c_i$ since $Y_l \in \Omega_G^*$. On the other hand, if $F_j \geq G_j$, then $Y_l(a) = c_j$ implies that $X(a) = c_j$. Since $i \neq j$, it indeed follows that $|\{(a, Y_l) \in I \times \Omega_G^* : X(a) = c_i, Y_l(a) = c_j\}| = 0$. Therefore we can assume now that $F_i > G_i$ and $F_j < G_j$.

We need to determine the $|\{Y_l \in \Omega_G^* : Y_l(a_0) = c_j\}|$ frequency.

$$|\{Y_l \in \Omega_G^* : Y_l(a_0) = c_j\}| =$$
$$\frac{\left(N - 1 - \min(F_i - 1, G_i) - \min(F_j, G_j - 1) - \sum_{k \neq i, k \neq j} \min(F_k, G_k)\right)!}{(G_i - \min(F_i - 1, G_i))! \cdot (G_j - 1 - \min(F_j, G_j - 1))! \cdot \prod_{k \neq i, k \neq j} (G_k - \min(F_k, G_k))!} \cdot$$
$$\binom{F_i - 1}{\min(F_i - 1, G_i)} \cdot \binom{F_j}{\min(F_j, G_j - 1)} \cdot \prod_{k \neq i, k \neq j} \binom{F_k}{\min(F_k, G_k)} . \tag{A.3.4}$$

$F_i > G_i$ and $G_j > F_j$ yield that $\min(F_i - 1, G_i) = \min(F_i, G_i) = G_i$ and

$\min(F_j, G_j - 1) = \min(F_j, G_j) = F_j$. Therefore (A.3.4) can be written as

$$|\{Y_l \in \Omega_G^* : Y_l(a_0) = c_j\}| =$$
$$\frac{\left(N - 1 - \sum_{k=1}^{K} \min(F_k, G_k))\right)!}{(G_j - 1 - \min(F_j, G_j))! \cdot \prod_{k \neq j} (G_k - \min(F_k, G_k))!} \cdot$$
$$\binom{F_i - 1}{\min(F_i, G_i)} \cdot \binom{F_j}{\min(F_j, G_j)} \cdot \prod_{k \neq i, k \neq j} \binom{F_k}{\min(F_k, G_k)} =$$
$$\frac{G_j - \min(F_j, G_j)}{N - \sum_{k=1}^{K} \min(F_k, G_k)} \cdot \frac{\left(N - \sum_{k=1}^{K} \min(F_k, G_k))\right)!}{\prod_{k=1}^{K} (G_k - \min(F_k, G_k))!} \cdot$$
$$\frac{F_i - \min(F_i, G_i)}{F_i} \cdot \prod_{k=1}^{K} \binom{F_k}{\min(F_k, G_k)}.$$

We have F_i choices for the a_0 individual, therefore

$$|\{(a, Y_l) \in I \times \Omega_G^* : X(a) = c_i, Y_l(a) = c_j\}| = F_i \cdot |\{Y_l \in \Omega_G^* : Y_l(a) = c_i\}| =$$
$$\frac{(F_i - \min(F_i, G_i)) \cdot (G_j - \min(F_j, G_j))}{(N - \sum_{k=1}^{K} \min(F_k, G_k))} \cdot$$
$$\frac{\left(N - \sum_{k=1}^{K} \min(F_k, G_k))\right)!}{\prod_{k=1}^{K} (G_k - \min(F_k, G_k))!} \cdot \prod_{k=1}^{K} \binom{F_k}{\min(F_k, G_k)}. \tag{A.3.5}$$

Combine this result with (A.3.1) and (A.3.2). We get

$$E(Z_{ij}) = \frac{(F_i - \min(F_i, G_i)) \cdot (G_j - \min(F_j, G_j))}{F_i \cdot (N - \sum_{k=1}^{K} \min(F_k, G_k))}.$$

This proves the theorem. □

A.4 Proof of (3.6.4)

Proof of (3.6.4).

$$H(X|Y) = -\sum_{i=1}^{K} \frac{F_i}{N} \cdot \frac{\min(F_i, G_i)}{F_i} \cdot$$

$$\log \frac{\frac{F_i}{N} \cdot \frac{\min(F_i, G_i)}{F_i}}{\frac{F_i}{N} \cdot \frac{\min(F_i, G_i)}{F_i} + \sum_{k \neq i} \frac{F_k}{N} \cdot \frac{(F_k - \min(F_k, G_k)) \cdot (G_i - \min(F_i, G_i))}{F_k \cdot \left(N - \sum_{m=1}^{K} \min(F_m, G_m)\right)}} -$$

$$\sum_{i=1}^{K} \sum_{j \neq i} \frac{F_i}{N} \cdot \frac{(F_i - \min(F_i, G_i)) \cdot (G_j - \min(F_j, G_j))}{F_i \cdot (N - \sum_{k=1}^{K} \min(F_k, G_k))} \cdot$$

$$\log \frac{\frac{F_i}{N} \cdot \frac{(F_i - \min(F_i, G_i)) \cdot (G_j - \min(F_j, G_j))}{F_i \cdot (N - \sum_{k=1}^{K} \min(F_k, G_k))}}{\frac{F_j}{N} \cdot \frac{\min(F_j, G_j)}{F_j} + \sum_{k \neq j} \frac{F_k}{N} \cdot \frac{(F_k - \min(F_k, G_k)) \cdot (G_j - \min(F_j, G_j))}{F_k \cdot (N - \sum_{m=1}^{K} \min(F_m, G_m))}} =$$

$$-\sum_{i=1}^{K} \frac{\min(F_i, G_i)}{N} \cdot \log \frac{\min(F_i, G_i)}{\min(F_i, G_i) + \frac{\sum_{k=1}^{K} (F_k - \min(F_k, G_k)) \cdot (G_i - \min(F_i, G_i))}{\left(N - \sum_{m=1}^{K} \min(F_m, G_m)\right)}} -$$

$$\sum_{i=1}^{K} \sum_{j \neq i} \frac{(F_i - \min(F_i, G_i)) \cdot (G_j - \min(F_j, G_j))}{N \cdot (N - \sum_{k=1}^{K} \min(F_k, G_k))} \cdot$$

$$\log \frac{(F_i - \min(F_i, G_i)) \cdot (G_j - \min(F_j, G_j))}{\min(F_j, G_j) \cdot (N - \sum_{k=1}^{K} \min(F_k, G_k)) + \sum_{k=1}^{K} (F_k - \min(F_k, G_k)) \cdot (G_j - \min(F_j, G_j))} =$$

$$-\sum_{i=1}^{K} \frac{\min(F_i, G_i)}{N} \cdot \log \frac{\min(F_i, G_i)}{G_i} -$$

$$\sum_{i=1}^{K} \sum_{j \neq i} \frac{(F_i - \min(F_i, G_i)) \cdot (G_j - \min(F_j, G_j))}{N \cdot (N - \sum_{k=1}^{K} \min(F_k, G_k))} \cdot \log \frac{(F_i - \min(F_i, G_i)) \cdot (G_j - \min(F_j, G_j))}{G_j \cdot (N - \sum_{k=1}^{K} \min(F_k, G_k))} =$$

$$-\sum_{i=1}^{K} \frac{\min(F_i, G_i)}{N} \cdot \log \frac{\min(F_i, G_i)}{G_i} -$$

$$\sum_{i=1}^{K} \sum_{j=1}^{K} \frac{(F_i - \min(F_i, G_i)) \cdot (G_j - \min(F_j, G_j))}{N \cdot (N - \sum_{k=1}^{K} \min(F_k, G_k))} \cdot \log \frac{(F_i - \min(F_i, G_i)) \cdot (G_j - \min(F_j, G_j))}{G_j \cdot (N - \sum_{k=1}^{K} \min(F_k, G_k))} =$$

$$-\sum_{i=1}^{K}\frac{\min(F_i,G_i)}{N}\cdot\log\frac{\min(F_i,G_i)}{G_i}-$$
$$\sum_{j=1}^{K}\frac{G_j-\min(F_j,G_j)}{N-\sum_{k=1}^{K}\min(F_k,G_k)}\cdot\sum_{i=1}^{K}\frac{F_i-\min(F_i,G_i)}{N}\cdot\log\frac{F_i-\min(F_i,G_i)}{N-\sum_{k=1}^{K}\min(F_k,G_k)}-$$
$$\sum_{i=1}^{K}\frac{F_i-\min(F_i,G_i)}{N-\sum_{k=1}^{K}\min(F_k,G_k)}\cdot\sum_{j=1}^{K}\frac{G_j-\min(F_j,G_j)}{N}\cdot\log\frac{G_j-\min(F_j,G_j)}{G_j}$$

This completes the proof of the formula. □

A.5 Proof of (3.6.5)

Proof of (3.6.5).

$$H(X|Y)=-\sum_{i=1}^{K}\frac{\min(M\cdot F_i,N\cdot G_i)}{N\cdot M}\cdot\log\frac{\min(M\cdot F_i,N\cdot G_i)}{N\cdot G_i}-$$
$$\sum_{i=1}^{K}\sum_{j=1}^{K}\frac{(M\cdot F_i-\min(M\cdot F_i,N\cdot G_i))\cdot(N\cdot G_j-\min(M\cdot F_j,N\cdot G_j))}{N\cdot M\cdot(N\cdot M-\sum_{k=1}^{K}\min(M\cdot F_k,N\cdot G_k))}\cdot$$
$$\log\frac{(M\cdot F_i-\min(M\cdot F_i,N\cdot G_i))\cdot(N\cdot G_j-\min(M\cdot F_j,N\cdot G_j))}{N\cdot G_j\cdot(N\cdot M-\sum_{k=1}^{K}\min(M\cdot F_k,N\cdot G_k))}=$$
$$-\sum_{i=1}^{K}\frac{\min(M\cdot F_i,N\cdot G_i)}{N\cdot M}\cdot\log\frac{\min(M\cdot F_i,N\cdot G_i)}{N\cdot G_i}-$$
$$\sum_{j=1}^{K}\frac{N\cdot G_j-\min(M\cdot F_j,N\cdot G_j)}{N\cdot M-\sum_{k=1}^{K}\min(M\cdot F_k,N\cdot G_k)}\cdot$$
$$\sum_{i=1}^{K}\frac{M\cdot F_i-\min(M\cdot F_i,N\cdot G_i)}{N\cdot M}\cdot\log\frac{M\cdot F_i-\min(M\cdot F_i,N\cdot G_i)}{N\cdot M-\sum_{k=1}^{K}\min(M\cdot F_k,N\cdot G_k)}-$$
$$\sum_{i=1}^{K}\frac{M\cdot F_i-\min(M\cdot F_i,N\cdot G_i)}{N\cdot M-\sum_{k=1}^{K}\min(M\cdot F_k,N\cdot G_k)}\cdot$$
$$\sum_{j=1}^{K}\frac{N\cdot G_j-\min(M\cdot F_j,N\cdot G_j)}{N\cdot M}\cdot\log\frac{N\cdot G_j-\min(M\cdot F_j,N\cdot G_j)}{N\cdot G_j}$$

This completes the proof of the formula. □

A.6 Proof of (6.3.6)

Proof of (6.3.6).

$$\left|\left(1-\frac{H(X)}{\log K}\right)\cdot\left(1-\frac{H(X|Y)}{H(X)}\right)-\left(1-\frac{\widehat{H(X)}}{\log K}\right)\cdot\left(1-\frac{\widehat{H(X|Y)}}{\widehat{H(X)}}\right)\right| =$$
$$\left|\left(\frac{\widehat{H(X)}}{\log K}-\frac{H(X)}{\log K}\right)+\left(\frac{\widehat{H(X|Y)}}{\widehat{H(X)}}-\frac{H(X|Y)}{H(X)}\right)+\left(\frac{H(X|Y)}{\log K}-\frac{\widehat{H(X|Y)}}{\log K}\right)\right| \leq$$
$$\frac{|H(X)-\widehat{H(X)}|}{\log K}+\left|\frac{H(X|Y)}{H(X)}-\frac{\widehat{H(X|Y)}}{\widehat{H(X)}}\right|+\frac{|H(X|Y)-\widehat{H(X|Y)}|}{\log K}\,.$$

Denote $\varepsilon_1 = |H(X)-\widehat{H(X)}|$ and $\varepsilon_2 = |H(X|Y)-\widehat{H(X|Y)}|$. We can assume that $\widehat{H(X)} \geq \widehat{H(X|Y)} \geq 0$.

$$\left|\frac{H(X|Y)}{H(X)}-\frac{\widehat{H(X|Y)}}{\widehat{H(X)}}\right| = \left|\frac{H(X|Y)\cdot\widehat{H(X)}-H(X)\cdot\widehat{H(X|Y)}}{H(X)\cdot\widehat{H(X)}}\right| =$$
$$\left|\frac{H(X|Y)\cdot\widehat{H(X)}-H(X)\cdot\widehat{H(X)}+H(X)\cdot\widehat{H(X)}-H(X)\cdot\widehat{H(X|Y)}}{H(X)\cdot\widehat{H(X)}}\right| =$$
$$\left|\frac{\widehat{H(X)}\cdot(H(X|Y)-H(X))+H(X)\cdot\left(\widehat{H(X)}-\widehat{H(X|Y)}\right)}{H(X)\cdot\widehat{H(X)}}\right| =$$
$$\left|\frac{H(X)\cdot\left(\widehat{H(X)}-\widehat{H(X|Y)}\right)-\widehat{H(X)}\cdot(H(X)-H(X|Y))}{H(X)\cdot\widehat{H(X)}}\right|$$

From the definitions of ε_1 and ε_2 we get $\widehat{H(X)} \leq H(X)+\varepsilon_1$ and $-\widehat{H(X|Y)} \leq \varepsilon_2 - H(X|Y)$. Therefore $\widehat{H(X)}-\widehat{H(X|Y)} \leq \varepsilon_1+\varepsilon_2+$

$H(X) - H(X|Y)$ and

$$\left| \frac{H(X) \cdot \left(\widehat{H(X)} - \widehat{H(X|Y)} \right) - \widehat{H(X)} \cdot (H(X) - H(X|Y))}{H(X) \cdot \widehat{H(X)}} \right| \leq$$

$$\left| \frac{H(X) \cdot (\varepsilon_1 + \varepsilon_2 + H(X) - H(X|Y)) - \widehat{H(X)} \cdot (H(X) - H(X|Y))}{H(X) \cdot \widehat{H(X)}} \right| =$$

$$\left| \frac{H(X) \cdot (\varepsilon_1 + \varepsilon_2) + (H(X) - \widehat{H(X)}) \cdot (H(X) - H(X|Y))}{H(X) \cdot \widehat{H(X)}} \right| \leq$$

$$\frac{\varepsilon_1 + \varepsilon_2}{\widehat{H(X)}} + \left| \frac{\varepsilon_1 \cdot (H(X) - H(X|Y))}{H(X) \cdot \widehat{H(X)}} \right| \tag{A.6.1}$$

□

Bibliography

Aigner, M., Ziegler, G. M., and Hofmann, K. H. (2014). *Proofs from the Book*. Springer.

Barata, J. C. A. and Hussein, M. S. (2012). The Moore-Penrose Pseudoinverse: A Tutorial Review of the Theory. *Brazilian Journal of Physics*, 42:146–165.

Bethlehem, J. G., Keller, W. J., and Pannekoek, J. (1990). Disclosure Control of Microdata. *Journal of the American Statistical Association*, 85(409):38–45.

Buzzigoli, L. and Giusti, A. (1999). An Algorithm to Calculate the Lower and Upper Bounds of the Elements of an Array Given its Marginals. *Statistical DATA Protection (SDP'98) Proceedings, Luxembourg, Eurostat*, pages 131–147.

Buzzigoli, L. and Giusti, A. (2006). From Marginals to Array Structure with the Shuttle Algorithm. *Journal of Symbolic Data Analysis*, 4:1–14.

Cover, T. M. and Thomas, J. A. (2006). *Elements of Information Theory*. Wiley, 2nd edition.

Cox, L. H. (1987). A Constructive Procedure for Unbiased Controlled Rounding. *Journal of the American Statistical Association*, 82(398):520–524.

Cox, L. H. and George, J. A. (1989). Controlled Rounding for Tables with Subtotals. *Annals of Operations Research*, 20(1):141–157.

Csiszár, I. (1967). Information-Type Measures of Difference of Probability Distributions and Indirect Observations. *Studia Scientiarum Mathematicarum Hungarica*, 2:299–318.

Csiszár, I. and Shields, P. (2004). *Information Theory And Statistics: A Tutorial*, volume 1 of *Foundations and Trends in Communications and Information Theory*. Now Publishers Incorporated.

Dandekar, R. A. and Cox, L. H. (2002). Synthetic Tabular Data : An Alternative to Complementary Cell Suppression. Unpublished manuscript.

Dobra, A. and Fienberg, S. E. (2000). Bounds for Cell Entries in Contingency Tables Given Marginal Totals and Decomposable Graphs. *Proceedings of the National Academy of Sciences*, 97(22):11885–11892.

Domingo-Ferrer, J. (2014). *Privacy in Statistical Databases: UNESCO Chair in Data Privacy, International Conference, PSD 2014, Ibiza, Spain, September 17-19, 2014. Proceedings*, volume 8744. Springer.

Domingo-Ferrer, J., Oganian, A., and Torra, V. (2002). Information-Theoretic Disclosure Risk Measures in Statistical Disclosure Control of Tabular Data. In *Proceedings of the 14th International Conference on Scientific and Statistical Database Management*, SSDBM '02, pages 227–231, Washington, DC, USA.

Domingo-Ferrer, J. and Torra, V. (2001). Disclosure Control Methods and Information Loss for Microdata. *Confidentiality, disclosure, and data access: theory and practical applications for statistical agencies*, pages 91–110.

Duncan, G., Keller-McNulty, S., and Stokes, S. (2001). Disclosure Risk vs. Data Utility: the R-U Confidentiality Map. Technical Report

LA-UR-01-6428, Statistical Sciences Group, Los Alamos, N.M.: Los Alamos National Laboratory.

Elamir, E. A. H. (2004). Analysis of Re-identification Risk Based on Log-Linear Models. In *Privacy in Statistical Databases*, pages 273–281. Springer.

Elamir, E. A. H. and Skinner, C. (2006). Record Level Measures of Disclosure Risk for Survey Microdata. *Journal of Official Statistics*, 22(3):525–539.

Elliot, M., Lomax, S., Mackey, E., and Purdam, K. (2010). Data Environment Analysis and the Key Variable Mapping System. In *Privacy in Statistical Databases*, pages 138–147. Springer.

Fienberg, S. E. (1999). Fréchet and Bonferroni Bounds for Multi-way Tables of Counts with Applications to Disclosure Limitation. In *Statistical Data Protection (SDP98) Proceedings*, pages 115–129.

Fischetti, M. and Salazar-González, J. J. (2000). Models and Algorithms for Optimizing Cell Suppression in Tabular Data with Linear Constraints. *Journal of the American Statistical Association*, 95(451):916–928.

Fischetti, M. and Salazar-González, J.-J. (2003). Partial Cell Suppression: A New Methodology for Statistical Disclosure Control. *Statistics and Computing*, 13(1):13–21.

Forster, J. J. and Webb, E. L. (2007). Bayesian Disclosure Risk Assessment: Predicting Small Frequencies in Contingency Tables. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 56(5):551–570.

Frank, O. (1976). Individual Disclosures from Frequency Tables. *Personal Integrity and the Need for Data in the Social Sciences. Swedish Council for Social Science Research*, pages 175–187.

Frank, O. (1978). An Application of Information Theory to the Problem of Statistical Disclosure. *Journal of Statistical Planning and Inference*, 2(2):143–152.

Gomatam, S. and Karr, A. F. (2003). Distortion Measures for Categorical Data Swapping. Technical Report Number 131, National Institute of Statistical Sciences.

Gomatam, S., Karr, A. F., and Sanil, A. (2003). A Risk-Utility Framework for Categorical Data Swapping. Technical Report Number 132, National Institute of Statistical Sciences.

Gomatam, S., Karr, A. F., and Sanil, A. P. (2005). Data Swapping as a Decision Problem. *Journal of Official Statistics*, 21(4):635.

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

Hernández-García, M.-S. and Salazar-González, J.-J. (2014). Enhanced Controlled Tabular Adjustment. *Computers & Operations Research*, 43:61–67.

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Nordholt, E. S., Spicer, K., and De Wolf, P.-P. (2012). *Statistical Disclosure Control*. John Wiley & Sons.

Oganian, A. and Domingo-Ferrer, J. (2003). A Posteriori Disclosure Risk Measure for Tabular Data Based on Conditional Entropy. *SORT*, 2:175.

Oganian, A., Domingo-Ferrer, J., and Torra, V. (2004). Insider Disclosure in Inference Control of Tabular Databases. volume 2 of *IPMU 2004. Information processing and management of uncertainty in knowledge-based systems*, pages 2007–2014. Università la Sapienza.

Purdam, K. and Elliot, M. (2007). A Case Study of the Impact of Statistical Disclosure Control on Data Quality in the Individual UK Samples of Anonymised Records. *Environment and Planning A*, 39(5):1101–1118.

Salazar-González, J.-J. (2006). Controlled Rounding and Cell Perturbation: Statistical Disclosure Limitation Methods for Tabular Data. *Mathematical Programming*, 105(2-3):583–603.

Serre, D. (2002). *Matrices: Theory and Applications*. Graduate Texts in Mathematics. Springer Verlag Gmbh.

Shlomo, N. (2007). Statistical Disclosure Control Methods for Census Frequency Tables. *International Statistical Review*, 75(2):199–217.

Shlomo, N. and Young, C. (2006). Statistical disclosure control methods through a risk-utility framework. In *Proceedings of the 2006 CENEX-SDC project international conference on Privacy in Statistical Databases*, PSD'06, pages 68–81, Berlin, Heidelberg. Springer-Verlag.

Skinner, C. J. and Holmes, D. J. (1993). Modelling Population Uniqueness. In *Proceedings of the International Seminar on Confidentiality*, pages 175–199, Dublin.

Skinner, C. J. and Holmes, D. J. (1998). Estimating the Re-identification Risk Per Record in Microdata. *Journal of Official Statistics*, 14(4):361–372.

Skinner, C. J. and Shlomo, N. (2008). Assessing Identification Risk in Survey Microdata Using Log-Linear Models. *Journal of the American Statistical Association*, 103(483):989–1001.

Skinner, C. J. and Shlomo, N. (2012). Estimating Frequencies of Frequencies in Finite Populations. *Statistics & Probability Letters*, 82(12):2206–2212.

Smith, D. and Elliot, M. J. (2008). A Measure of Disclosure Risk for Tables of Counts. *Transactions on Data Privacy*, 1(1):34–52.

Takemura, A. (1999). Some Superpopulation Models for Estimating the Number of Population Uniques. In *Proceedings of the Conference on Statistical Data Protection*, pages 45–58. Citeseer.

Willenborg, L. and de Waal, T. (2001). *Elements of Statistical Disclosure Control*. Lecture notes in statistics. Springer.