ESSnet on common tools and harmonised methodology for SDC in the ESS

Overview of methods for Statistical Disclosure Control of Census Frequency Tabular Data

Sarah Giessing
Statistisches Bundesamt, 65180 Wiesbaden, Germany
Email: Sarah.Giessing@destatis.de

1. Introduction

This overview is meant to briefly outline well known methods for disclosure controls for frequency tables, that is tables of counts (or percentages) where each cell value represents the number of respondents in that cell. It is mainly a summary / shortened version of chapter 5 of the ESSNET-SDC handbook (Hundepool et al, 2010) contributed by Jane Naylor (ONS). To some extent this paper presents just extracts from the handbook chapter (partly in the original phrasing). The interested reader is very much invited to look up the respective original sections and paragraphs of the handbook chapter for details, more worked out examples, further reading and literature.

Section 2 introduces typical disclosure risks resulting from publication of Census frequency counts data. Regarding protection methods, we distinguish pre- and post-tabular approaches. In section 3, some well known methods are briefly outlined with a short discussion of the pros and cons and occasional reference to software implementations. Of course data protection always has its costs in terms of information loss. Section 4 gives a first idea of the issues.

2. Disclosure Risks

Tables from Census publications contain counts of people or households with certain social characteristics. The disclosure risk situations described here primarily apply to tables produced from registration processes, administrative sources or Censuses, e.g. data sources with a complete coverage of the population or sub-population. Where frequency tables are derived from sample surveys, e.g. the counts in the table are weighted, some protection is provided by the sampling process. The sample a priori introduces uncertainty into the zero counts and other counts through sample error[1].

It should be noted that when determining unsafe cells one should take into account the variables that define the population within the table, as well as the variables defining the table. For example, a table may display income by region for males. Although sex does not define a row or column it defines the eligible population for the table and therefore must be considered as an identifying variable when thinking about the disclosive situations. Disclosure risk is usually due to the presence of cells with small counts (f.i. frequency 1) in a table (risk of "Identification", see below). Also the distribution and location of cells with count zero might create certain risks that individual respondent data could be revealed. Those risks are referred to as "Attribute Disclosure".

Disclosure risks are categorised based on how information is revealed. The most common types of disclosure risk in frequency tables are described below.

[1] Regarding protection of this type of data, see the remarks at the end of the paragraph on conventional rounding in sec. 3.2.

Identification

Identification as a disclosure risk involves finding yourself or another individual or group within a table. Many NSIs will not consider that self-identification alone poses a disclosure risk. This is most likely to occur where a cell has a small value, e.g. a 1, or where it becomes in effect a population of 1 through subtraction or deduction using other available information. For certain types of information, rareness or uniqueness may encourage others to link this information to other data sources and seek out the individual. An example might be a unique male immigrant in a certain municipality. If this individual receives social security, and if a table on social security status by age, sex and immigration status is produced, an individual with these properties (male, immigrant) should also appear in that table. Anybody (*viz.* any data collecting body) able to identify this individual on the basis of his age, sex and immigration status can then link the information from social security statistics to this individual, because of the information on its uniqueness. Considering this argument, the problem of unique combinations of certain variables is mainly a problem in the context of variables that tend to be “identifying”, e.g. known among acquaintances of an individual or present in many data collections. On the basis of considerations such as this, several NSIs choose to protect against identification disclosure.

Attribute Disclosure

Another disclosure risk involves learning a new attribute about an identifiable group (or individual) or learning a group (or the individual) does not have a particular attribute. This is termed group (or individual) attribute disclosure, and it can occur when all respondents fall into a subset of categories for a particular variable, i.e. where a row or column contains mostly zeros and a small number of cells that are non-zero. This type of disclosure is a much neglected threat to the disclosure protection of frequency tables. In the case of group attribute disclosure, it does not require individual identification. In order to protect against group attribute disclosure it is essential to introduce ambiguity in the zeros and ensure that all respondents do not fall into just one or a few categories. If the social security status table of the example in the section of individual disclosure were indeed part of the Census publication, it would be a typical example for individual attribute disclosure. Assume now, there are 6 male immigrants living in the municipality of the example, and they all receive social security. Then still anybody able to identify any of the 6 by their age, sex and immigration status properties learns from the table that each of those individuals receives social security. The social security attribute of the entire group (of 6 individuals) is disclosed.

Disclosure by Differencing and/or Linking

There is one type of disclosure by differencing that involves differencing of sub-population tables. Sub populations are specific groups which data may be subset into before a table is produced (e.g. a table of fertility may use a sub-population of females). *Differencing* can occur when a published table definition corresponds to a sub-population of another published table, resulting in the production of a new, previously unpublished table. If the total population is known and the subpopulation of females is gathered from another table, the number of males can be deduced.

Tables based on categorical variables which have been recoded in different ways may also result in this kind of differencing. Assume, for example, table A presents the number of people aged under 20 receiving social security, and table B presents the number of males under 25 receiving social security. Subtracting table A from table B result in a new, not directly published table C that contains the number of people aged between 20 and 25 receiving social security. As this is a much smaller population, it is more likely that there are “risky” cells, e.g. cells where there is a risk of individual or attribute disclosure. If only those tables foreseen for publication (table A and table B) are checked for disclosure risk, this problem will be overlooked.

Disclosure by *linking* can occur when published tables relating to the same base population are linked by common variables. These new linked tables were not published by the NSI (and hence not checked for disclosure risk) and therefore may reveal unsafe cell counts. Assume a table A presents number non-immigrants by age and sex on the municipality level. A corresponding table A’ for immigrants is not published because of disclosure risk. Assume also that a table B is released that presents number of school children aged 5 to 20 by type of school and sex (all children, not only non-immigrants). Assume both tables

(A and B) do not contain any “risky” cells. By adding up (across school-types) it is then easy to calculate the total number of 5 to 20 year olds by sex (say: table b’). By adding up across the respective categories of age, it is possible from table A to deduce the total number of non-immigrant 5 to 20 year olds by sex (table a’). Subtracting the figures of table a’ from table b’ leads to a new table presenting the number of 5 to 20 year old immigrants by sex. This is a relatively small population and for some municipalities this table might contain “risky” constellations. However, if only tables are checked/protected that are foreseen to be published, this kind of risk will not be discovered.

For more (and more elaborate) examples, see chapter 5 of the ESSNET SDC-handbook.

3. Protection Methods

This section provides a survey of some popular methods used to protect tables of counts. SDC methods can be divided into three categories: those that adjust the data before tables are designed (pre-tabular), those that determine the design of the table (table redesign) and those that modify the values in the table (post-tabular). Thinking of the above mentioned examples of disclosure by differencing and linking, it becomes obvious that table redesign alone provides relatively little disclosure control protection, especially in the context of a Census when usually a huge variety of tables is released.

3.1. Pre-tabular methods

Typical pre-tabular methods change the micro-data before the process of tabulation is started. An important advantage of the methods of this type is that all tables will be consistent and additive. Once a method has been implemented for a dataset, standard tabulation packages can be used to compute the tables. Those methods typically tend to protect well against risks of disclosure by differencing. They can therefore be used without problems within a flexible tabulation process.

Disadvantages of pre-tabular techniques might be that a high level of perturbation may be required in order to disguise all unsafe cells. Pre-tabular methods have the potential to distort distributions in the data, but the actual impact of this will depend on which method is used and how it is applied. Generally pre-tabular methods are not transparent to users of the frequency tables.

Record Swapping

In the context of protecting Population census tabular data, Record Swapping is probably the most popular in the class of pre-tabular methods. See (Zayatz, 2003) for an illustrative description. The idea is to swap values of variables of persons in households between pairs of households. Before swapping, households are grouped. Typical grouping variables are geography (on a higher level), household size and the distribution of persons in the household by age and sex. Swaps will only be carried out between household within the same group. In a swapping variant known as “targeted record swapping” the probability of certain types of households to be swapped will be increased: Households with a relative high risk that of contributing to table cells in ‘critical constellation’ regarding the disclosure risks discussed in section 2 will be selected for swapping at relatively high probabilities. The most typical variable to be swapped is geography at the lowest level. If, for example, the municipality level geography is a grouping variable, swapping means to swap households between areas within a municipality. This way, disclosure risk problems due to “critical constellations” at the municipality level and above will not be solved – a unique case in a municipality will remain unique. This has to be taken into account when designing a swapping method.

Implementing a swapping procedure with standard packages such as SAS will be relatively easy, compared to an implementation of other methods (for example the method SAFE below). However, working out the details of the method and defining suitable parameters will be a major effort.

SAFE

The method SAFE (Höhne, 2011) is an implementation of an algorithm that yields a controlled cell frequency perturbation. When a microdata set has been protected by this method, any table which can be computed on the basis of this microdata set will not contain any small cells, e.g. cells with frequency counts one or two. Because SAFE is a pre-tabular method, all tables computed from the perturbed microdata set protected by SAFE are fully consistent and additive. The method has been implemented in the software package SAFE of the State Statistical Institute Berlin-Brandenburg (Germany).

3.2. Post-tabular methods

Statistical disclosure control methods that modify cell values within tabular outputs are referred to as post-tabular methods. Such methods are generally clear and transparent to users, and are easier to understand and account for in analyses, than pre-tabular methods. However, post-tabular methods suffer the problem that each table must be individually protected, and it is necessary to ensure that the new protected table cannot be compared against any other existing outputs in such a way which may undo the protection that has been applied. Considering this, some post-tabular methods can be cumbersome to apply to large tables or large sets of tables. The main post-tabular methods include cell suppression, cell perturbation, and rounding.

Cell Suppression

In a first phase cells are identified that need protection. In typical implementations, this will be small count cells (e.g. one and twos). Those cells are suppressed from the publication, e.g. the cell value is replaced by some special character like '/'. Protecting the unsafe cells this way is called primary suppression, and to ensure these cannot be derived by subtractions from published marginal totals, additional cells are selected for secondary suppression.

Cell suppression cannot be unpicked provided secondary cell suppression is adequate and the same cells in any linked tables are also suppressed. Other advantages are that the method is easy to implement on unlinked tables and it is highly visible to users. The original counts in the data that are not selected for suppression are left unadjusted.

However cell suppression has serious disadvantages as a protection method for frequency tables, in particular information loss can be high if more than a few suppressions are required. Secondary suppression removes cell values which are not necessarily a disclosure risk, in order to protect other cells which are a risk. When disseminating a very large number of tables (typical in a census context), it is virtually impossible to ensure the consistency of suppressed cells, e.g. to ensure that same cells in linked tables are always suppressed. Moreover, usually cell suppression is implemented to select only small cells, but not zero cells as primary suppressions. In that case, it will not protect against the risk of group attribute disclosure. Suppressing all zero cells on the other hand is "too expensive" in terms of the loss of information. But developing rules to identify some of the zero cells as primary suppressions can be cumbersome. Software for cell suppression is implemented in the package τ-ARGUS (Hundepool et al., 2011).

Additive Noise

The Australian Bureau of Statistics (ABS) has developed a cell perturbation algorithm (Fraser and Wooton, 2006). The method is designed to protect tables by adding random noise to all cell values. The cells are adjusted in such a way that the same cell is perturbed in the same way even when it appears across different tables. This method adds sufficient 'noise' to each cell so if an intruder tried to gather information by differencing, they would not be able to obtain the real data.

The method consists of a two stage process;

Stage one adds the perturbations to the cell values. All microdata records are assigned a record key and when creating a table the record keys for all records contributing to each internal cell are summed. A function is applied to this sum to produce the cell key. A design transition probability matrix and the cell key are then used to determine the amount each cell should be perturbed. This means that the same cell is always perturbed in the same way. Table margins are perturbed independently using the same method.

Stage two can be implemented to add another perturbation to each cell (excluding the grand total) to restore table additivity. The stage two perturbations can be generated for instance using an iterative fitting algorithm which attempts to balance and minimise absolute distances to the stage one table. However, while consistency is maintained during the first stage of the perturbation process, it is lost when the additivity module is applied.

The method provides protection for flexible tables and can be used to produce perturbations for large high dimensional hierarchical tables. The method must be applied to each table separately. Because of the cell keys it requires a specialized tabulation procedure. I.e. if a standard tabulation package is used to produce the tables this packages would have to be manipulated in a certain way. The tabulation package SuperCross[2] offers this special feature.

Rounding

Rounding involves adjusting the values in all cells in a table to a specified base so as to create uncertainty about the real value for any cell. It adds a small, but acceptable, amount of distortion to the original data. Rounding is considered to be an effective method for protecting frequency tables, especially when there are many tables produced from one dataset. It provides protection to small frequencies and zero values (e.g. empty cells). The method is simple to implement, and for the user it is easy to understand as the data is visibly perturbed.

There are several alternative rounding methods including; conventional rounding, random rounding, and controlled rounding, which are outlined below. Each method is flexible in terms of the choice of the base for rounding, although common choices are 3 and 5. All rounded values (other than zeros) will then be integer multiples of 3 or 5, respectively.

When using ***conventional rounding***, each cell is rounded to the nearest multiple of the base. The marginal totals and table totals are rounded independently from the internal cells. The advantages of this method are that the table totals are rounded independently from the internal cells, and therefore consistent table totals will exist within the rounding base. Cells in different tables which represent the same records will always be the same. While this method does provide some confidentiality protection, it is considered less effective than controlled or random rounding. Tables are not additive (i.e. the rounded values of the 'inner' cells of a table row usually do not add up to the rounded value of the row total). In certain constellations the method can be easily 'unpicked' when differencing or linking tables. Assume a table of one row with two 'inner' cells and conventional rounding to base 3. Assume the rounded row total is 3. The two rounded cell values within the row are 0. Then each of the two original values within the row must be at most 1, because otherwise its rounded value would be 3 instead of 0. And both must be at least one, because otherwise the original row total would be less than two, meaning it would have been rounded down to zero. This means that we have been able in this instance to unpick the rounding.

For these reasons conventional rounding is not recommended as a disclosure control method for frequency tables based on a full census. Conventional rounding is sometimes used by NSIs for quality reasons, e.g. rounding data from sample surveys to emphasize the uncertain nature of the data. A typical rounding base in that context would be larger, e.g. 10 or 50. Then, together with sample error, conventional rounding will usually reduce disclosure risks to an acceptable level. The distinction between rounding performed for disclosure control reasons and rounding performed for quality reasons should always be made clear to users.

Random rounding shifts each cell to one of the two nearest base values in a random manner. Each cell value is rounded independently of other cells, and has a greater probability of being rounded to the nearest multiple of the rounding base. For example, with a base of 5, cell values of 6, 7, 8, or 9 could be rounded to either 5 or 10. Random rounding may considerably reduce the likelihood of constellations where the rounding can be unpicked, but it does not completely eliminate this risk.

Unlike other rounding methods, ***controlled rounding*** yields additive rounded tables. It is the statistical disclosure control method that is generally most effective for frequency tables. The method uses linear programming techniques to round cell values up or down by small amounts, and its strength over other methods is that additivity is maintained in the rounded table, (i.e. it ensures that the rounded values add up to the rounded totals and sub-totals shown in the table). This property not only permits the release of realistic

[2] See http://www.spacetimeresearch.com/supercross.html

tables which are as close as possible to the original table, but it also makes it impossible to reduce the protection by ‘unpicking’ the original values by exploiting the differences in the sums of the rounded values. Controlled rounding can be used to protect large tables although computing time may make it unsuitable for this purpose. Moreover, it will be difficult in practice to avoid that cells of different tables that are logically identical are rounded in the same way in each tables.

Whereas implementation of the other rounding methods is a relative easy task, controlled rounding involves the computation of complex mixed integer programming problems. An implementation is available in the package τ-ARGUS (Hundepool et al., 2011).

4. Information Loss

Each of the above protection methods modifies the original data in the table in order to reduce the disclosure risk from small cells (0’s, 1’s and 2’s). However, the process of reducing disclosure risk results in information loss. With cell suppression, NSIs typically count the number of primary and secondary suppressions to get an idea of information loss. For a perturbative protection method the distribution of the (relative) deviations between true and perturbed cell values will give a first idea.

A number of more elaborate quantitative information loss measures have been developed by Shlomo and Young (2005 & 2006) to determine the impact various statistical disclosure control (SDC) methods have on the original tables. Some of them are explained in the ESSNET SDC-handbook.

References

Fraser, B., Wooton, J. (2006). *A proposed method for confidentialising tabular output to protect against differencing*, in *Monographs of Official Statistics. Work session on Statistical Data Confidentiality*, Eurostat-Office for Official Publications of the European Communities, Luxembourg, 2006, pp. 299-302

Höhne, J., (2011) *SAFE – A method for anonymising the German Census*, Proceedings of the Joint UNECE/Eurostat WorkSession on Statistical Data Confidentiality (Tarragona, October 2011), available at http://www.unece.org/stats/documents/2011.10.confidentiality.html/wp.16.s.e.pdf

Hundepool, A., et al (2010) *ESSNET Handbook on Statistical Disclosure Control* http://neon.vb.cbs.nl/casc/handbook.htm

Hundepool A, van de Wetering A, Ramaswamy R, de Wolf P, Giessing S, Fischetti M, J.J.Salazar, Castro J and Lowthian P (2011) τ-ARGUS 3.5 user manual. Statistics Netherlands, Voorburg NL, available at http://neon.vb.cbs.nl/casc/tau.htm

Shlomo, N. and Young, C. (2005) *Information Loss Measures for Frequency Tables,* Proceeding of the Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality, Geneva.

Shlomo, N. and Young, C. (2006) *Statistical Disclosure Control Methods Through a Risk-Utility Framework,* PSD'2006 Privacy in Statistical Databases, Springer LNCS proceedings.

Zayatz, L. (2003) *Disclosure Limitation for Census 2000 Tabular Data*, Proceedings Joint UNECE/Eurostat WorkSession on Statistical Data Confidentiality (Luxembourg, 7-9 April 2003), available at http://www.unece.org/stats/documents/2003.04.confidentiality.html/wp.15.s.e.pdf