OIC ACCREDITATION CERTIFICATION PROGRAMME FOR OFFICIAL STATISTICS

Statistical Confidentiality and Disclosure Control

TEXTBOOK

ORGANISATION OF ISLAMIC COOPERATION

STATISTICAL ECONOMIC AND SOCIAL RESEARCH AND TRAINING CENTRE FOR ISLAMIC COUNTRIES

SESRIC

OIC ACCREDITATION CERTIFICATION PROGRAMME FOR OFFICIAL STATISTICS

Statistical Confidentiality and Disclosure Control

TEXTBOOK

TEMPLATE

Afsaneh Yazdani

ORGANISATION OF ISLAMIC COOPERATION

STATISTICAL ECONOMIC AND SOCIAL RESEARCH
AND TRAINING CENTRE FOR ISLAMIC COUNTRIES

TEMPLATE

© 2015 The Statistical, Economic and Social Research and Training Centre for Islamic Countries (SESRIC)

Kudüs Cad. No: 9, Diplomatik Site, 06450 Oran, Ankara – Turkey

Telephone +90 – 312 – 468 6172

Internet www.sesric.org

E-mail statistics@sesric.org

The material presented in this publication is copyrighted. The authors give the permission to view, copy download, and print the material presented that these materials are not going to be reused, on whatsoever condition, for commercial purposes. For permission to reproduce or reprint any part of this publication, please send a request with complete information to the Publication Department of SESRIC.

All queries on rights and licenses should be addressed to the Statistics Department, SESRIC, at the aforementioned address.

DISCLAIMER: Any views or opinions presented in this document are solely those of the author(s) and do not reflect the views of SESRIC.

ISBN: xxx-xxx-xxxx-xx-x

Cover design by Publication Department, SESRIC.

For additional information, contact Statistics Department, SESRIC.

CONTENTS

TEMPLATE

ACRONYMS

DL	Data Laboratory
NSO	National Statistical Organization
PUF	Public Use File
SDC	Statistical Disclosure Control

TEMPLATE

ACKNOWLEDGEMENT

This textbook on Statistical Confidentiality and Disclosure Control is prepared jointly by the Statistical Centre of Iran in Tehran – Islamic Republic of Iran and the Statistical, Economic and Social Research and Training Centre for Islamic Countries (SESRIC) under the OIC Accreditation and Certification Programme for Official Statisticians (OIC-CPOS) supported by Islamic Development Bank Group (IDB).

The author wishes to thank Mr. Arman Bidarbakhtnia for being a great help, and Mr. Hamidreza Navvabpour, Mr. Hassan Ranji, Mrs. Naeemeh Abi, and Mr. Mohmmad Bordbar for their generous support in preparing the textbook. Mr. Aliakbar Maleki's contribution in editing the text is greatly appreciated. The work would not have been accomplished without the moral and technical support received from Mr. Gholamreza Izadi.

TEMPLATE

UNIT 1

INTRODUCTION

The main objective of this unit is to introduce the basic concepts and definitions related to Statistical Disclosure Control (SDC). The attempt is to provide some answers to specific questions such as "Why statistical confidentiality matters?", "Why the National Statistical Offices (NSOs) try to protect confidentiality?", "What is the conflict between access to data and confidentiality?" and "Why the NSOs have little tendency to present more detailed data?".

1.1. Concepts and Definitions

Learning Objectives

– To understand basic concepts and definitions related to Statistical Confidentiality and Disclosure

A number of basic SDC concepts that one should know prior to any further study on statistical confidentiality and disclosure are presented in this section.

Confidentiality is a quality or condition accorded to information as an obligation not to transmit that information to any unauthorized party (Fienberg, 2005). Regardless of the context in which the term is used, confidentiality is a promise that the receiver and current holder of the information make to the provider of the information, remarking that the information will be:

- Exclusively reserved for intended purposes, and
- Used only by authorized individuals (Duncan et al., 2011)

Disclosure occurs when from a released data, in the form of a table or microdata, the identity of an individual is recognized or sensitive information about an individual is revealed. Three types of disclosure have been introduced in the literature (FCSM, 2005):

- *Identity Disclosure* or *Re-identification* occurs when the identity of an individual including a person, a household, a business enterprise, etc., is recognized via the disseminated data; i.e. a particular record within a set of data is exactly associated with a particular population unit (Hundepool et al., 2010).

TEMPLATE

- *Attribute Disclosure* or *Attribute Re-identification* occurs when sensitive information about an individual is revealed via a released data. In this case, the identity of the individual is not necessarily disclosed.
- *Inferential Disclosure* or *Approximate Disclosure* occurs when a disseminated data is in a form that makes it possible to determine the value of a sensitive variable about an individual, with high confidence. In some texts, inferential disclosure is also classified as the attribute disclosure.

Disclosure Risk or *Risk of Disclosure* is a function of the probability of identifying an individual (Groves et al., 2009), which may also lead to the disclosure of a sensitive information about an individual, by using released data (Hundepool et al., 2010).

Identifying variable or *key variable* is a variable that can be used, alone or in combination with other variables, to re-identify an individual (Willenborg and De Waal, 1996). Identifying variables may be divided into two groups:

- *Direct, formal, or explicit identifier*, such as social security number, is a variable or a set of variables which is structurally unique for every population unit (Hundepool et al., 2010), by which the individuals can be re-identified directly (Duncan et al., 2011).
- *Quasi-identifiers, indirect identifiers* such as age, sex and education, is a variable or a set of variables within a data set that is not structurally unique but might be empirically unique (Hundepool et al., 2010) and can be used in re-identification of individuals, with some degree of confidence.

Intruder, *Data Intruder* and *Data Snooper* are all synonyms that refer to an individual, group, or organization who attempts to identify an individual (including a person, a household, a business enterprise, etc.) within a data set and/or to discover sensitive information about a given individual, usually through a statistical linkage process (Duncan et al. 2011).

Macrodata or *Tabular data* is aggregate information on individuals presented in tables (Hundepool et al., 2010). Tables are the well-known and the most traditional data products of the statistical agencies. There are two types of tabular outputs (Hundepool et al., 2010):

- *Magnitude Tables*, where each cell's value represents a summary statistics of values for a particular characteristic across all individuals that belong to that cell, e.g. total (or average) income of the female employees within a company.

TEMPLATE

- *Frequency Tables*, where each cell's value represents the number of individuals that fall into that cell, e.g. number of single mothers in a region.

Microdata or Unit Record Data is relatively new and recent data products of the NSOs. Microdata consists of records with information about each single individual including a person, a household, a business enterprise, etc. In other words, each record contains the values of a number of variables for each individual entity (Willenborg and De Waal, 1996).

Privacy is closely linked to, but yet distinct from confidentiality. Privacy is an individual's freedom from excessive intrusion in the quest for information and an individual's ability to choose the extent and circumstances under which his or her beliefs, behaviour, opinions, and attitudes will be shared with or withheld from others (Duncan et al., 1993). Breach of confidentiality may harm the individual's privacy.

Sensitive Variable also referred to as *Confidential Variable* is a variable, apart from the key variables, whose values can represent characteristics (such as income or mental health) that an individual may not like to be revealed. Specifying which variables are sensitive is quite a vague issue, depending on personal taste, public opinion and cultural background (Willenborg and De Waal, 1996).

Statistical Disclosure Control (SDC), which is synonym to *Statistical Confidentiality*, or *Disclosure Limitation*, is a body of principles, concepts, and procedures that permit confidentiality to be afforded to data, while still permitting its use for statistical purposes (Duncan et al., 2011). In practice, disclosure control decisions are a trade-off between the utility and disclosure risk (Hundepool et al., 2010)
There are two types of SDC method, "*Perturbative*" and "*Non-Perturbative*" (Hundepool et al., 2010):

- *Perturbative Methods* falsify the data before publication by introducing an element of error deliberately for confidentiality reasons.
- *Non-Perturbative Methods* reduce the amount of information released, by suppression or aggregation of data.

Utility, sometimes called *Data Utility* means the value of a given data release as an analytical resource. This comprises the data's analytical completeness and its analytical validity (Hundepool et al., 2010).

TEMPLATE

User, or *Data User* is any person using a data set legally, with a good intention, and typically for statistical purposes.

1.2. Why is statistical confidentiality a concern?

Learning Objectives

– To understand the NSO's concerns about confidentiality

Protecting confidentiality has always been the NSOs' concern, but it has received more attention recently. In last decades, world has experienced a jump in technological advances. More statistical information is now collected in different fields and massive amounts of data are available. There has been a rapid growth in using computers and software. Many researchers are now capable of performing complicated data analyses themselves. Therefore, enquiries for more and more detailed data has increased. At the same time, societies' attitude towards human rights and privacy has greatly changed, and that forces the governments to seek special arrangements for handling privacy issues. Thus, today, resolving tension between protecting data and providing data is really a serious challenge for the NSOs. In this complicated situation, three motivations push the NSOs to preserve confidentiality.

The first motive for maintaining confidentiality comes from the NSO's moral obligations towards public. The NSO must respect the trust of respondents, take care of their privacy, and keep them away from any harm that may root from the information they have provided. The NSO should prevent violating the ethical norms.

The second motive is underlying in the desire of the NSO to gain cooperation of respondents and to obtain more accurate data. The respondents who believe that their information will remain confidential are more likely to participate in the survey and accurately report their private information. While any doubt about the confidentiality may reduce the willingness of potential respondents to cooperate in the survey and can affect the quality of responses.

The last motive is the obligation imposed on the NSO by the national law and regulations as well as international commitments. Society's force on governments has led to establishment of legal settings for safeguarding the privacy and the NSO is bound to observe these legal constraints (Duncan et al., 2011). Moreover, as unanimously endorsed by the General Assembly of the United Nations in January 2014, principle 6 of the Fundamental Principles of Official Statistics[1] postulates that "*Individual data collected by statistical agencies for statistical*

[1] http://unstats.un.org/unsd/dnss/gp/FP-Rev2013-E.pdf

TEMPLATE

compilation, whether they refer to natural or legal persons, are to be strictly confidential and used exclusively for statistical purposes".

1.3. Trade-off between data access and confidentiality

Learning Objectives

- To be able to explain:
  a. The conflict between data access and confidentiality
  b. Why more detailed data is required?
  c. Why the NSOs are unwilling in presenting more detailed data?

As discussed, advancements in technology have given rise to a new dilemma. Research society has become aware of the benefits of statistical information and has acquired the capability of extracting knowledge out of the statistical data, at the same time, members of general public are increasingly expressing their concerns about the privacy. The NSOs are now facing a conflicting situation; increasing demand for statistics and heightened confidentiality concerns, a contrast between the public's and individual's benefits. In other words, mandate to provide data against commitment to restrict access to it and maintain privacy and confidentiality. Deciding on what should be released and what should be protected is a serious challenge for the NSOs who are missioned to provide high quality data products and services, and meanwhile, are seriously expected to be conscious about the information that can be tied to individuals.

Modern societies call for accurate information in order to build a scientific understanding of the world around and to take realistic snapshots of society that lead to more effective decision making and policy analyses. As a response to this demand, broad spectrum of data users has stepped into existence who seeks information for empirical analyses. They require accurate, detailed and comprehensive data; and frequently request for geographically specific, hierarchical, longitudinal, or even individual-level data.

The NSOs, as national data hubs, should feel responsible in providing both traditional and new clients (the research community) with their required data and statistical services. If research community cannot access the relevant data from the official sources, they may try to find other alternatives that may not meet quality criteria and will incur additional costs, and response burden, too. Moreover, utilizing official data in research can add value to the data, shed more light on the quality of data, and return more benefit to the public. Extensive use of available data sources for producing statistics can reduce duplications and put national statistical systems in a better position to face budget limitations (UNECE, 2007).

TEMPLATE

Providing research community with microdata, though beneficial to the NSOs and the public, is not without difficulties. As data becomes more detailed, the risk of disclosure increases and preserving the confidentiality becomes a more complicated and demanding job. Consequences of even one confidentiality breach can be very destructive for the NSO's reputation; it can cause a severe breakdown in the public trust and result in significant reduction of cooperation. Preserving confidentiality of data is imperative for building public trust. To balance competing forces of confidentiality and access, and conflicting interests of different stakeholders (respondents and research community), the NSOs are now obliged to seek creative solutions.

In providing access to more disaggregated data, apart from the confidentiality issue, the NSOs have to be concerned about the quality of their statistics. Often, particularly when the data is collected through sample surveys, quality of statistical products are satisfactory only at the aggregate level and user demands for more disaggregated data may not meet quality requirements. In the other hand, giving access to microdata and producing disaggregated statistics is not free of cost for the NSOs. It requires creating and documenting microdata files, providing meta-data, creating access tools and safeguards, supporting and authorising enquiries made by clients, technical support of new users, etc. These imposed expenses are not normally provided within budget, or by the clients (UNECE, 2007). Therefore, pricing strategy is required by the NSOs that is affordable by the users and, at the same time, facilitates safe dissemination of microdata and disaggregation statistics.

1.4. Core Principles of Privacy Protection

Learning Objectives

- To comprehend core principles of privacy protection, and access to microdata
- To be able to explain legislation on microdata release

As mentioned before, many NSOs are now under specific or general legal constraints to protect the confidentiality. All principles addressing data access should observe the sixth Fundamental Principle of Official Statistics of United Nations on confidentiality. This section presents core principles of privacy protection based on the Fair Information Practice, and core principles for microdata access based on the UNECE's guideline. Both of them are inspired by the sixth Fundamental Principle and aim to provide more detailed guideline for the NSOs. The last part of the section will address the issues related to legislation on microdata release.

Core principles of privacy protection

Duncan et al. (2011), refer to five core principles of privacy protection which have been built

TEMPLATE

mostly based on the statement of Fair Information Practices[2] and have evolved through the works of various groups, particularly governmental agencies in United States, Canada and Europe. These five principles are as follows:

1. *Notification*

   Respondents should be given notice before any personal information is collected from them. Generally, notification includes a description of the statistical agency collecting the data, the purposes of the survey, expected duration of the interview, a description of procedures, potential recipients of the data, nature of data collected, whether participation is voluntary or obligatory, the consequences of a refusal to participate in the survey, steps taken to ensure the confidentiality, integrity and quality of data, etc.

2. *Consent*

   As much as it is possible and practicable, potential respondents should be able to choose to participate or not.

3. *Respondents access*

   A respondent has access when he or she can view their own data, and can contest accuracy and completeness of the data.

4. *Data integrity*

   Data should be accurate and secure. Security requires both managerial and technical measures to protect the data against loss and unauthorized access.

5. *Enforcement*

   Fair Information Practice acquires its power only through adequate enforcement mechanisms. These mechanisms may comprise arrangements for external audit to verify compliance, or legislations providing civil or criminal penalties for violation of Fair Information Practice.

Core principles for microdata access

According to the UNECE (2007), principles to be used in managing the confidentiality of microdata, are as follows:

Principle 1: It is appropriate for microdata collected for official statistical purposes to be used for statistical analysis to support research as long as confidentiality is protected.

Principle 2: Microdata should only be made available for the statistical purposes.

[2] Organization for Economic Cooperation and Development (http://oecdprivacy.org)

TEMPLATE

Principle 3: Provision of microdata should be consistent with the legal and other necessary arrangements that ensure that confidentiality of the released microdata is protected.

Principle 4: The procedures for researcher’s access to microdata, as well as the uses and users of microdata, should be transparent and publicly available.

Principle 1 does not constitute an obligation for the NSO to provide microdata to research community. It only explains as long as the NSO can maintain confidentiality, and there is no other concern (like quality of microdata), it is up to the NSO’s management to decide whether to provide microdata to specific users or not.

Principle 2 emphasizes on distinction between the statistical and administrative uses of data. Statistical use aims to derive statistics about a group of individuals while in the administrative use, the aim is to derive information about a particular individual which could potentially violates the confidentiality.

Principle 3 is explaining that legal arrangement to protect confidentiality should be in place before any microdata is released. These legal arrangements have to be complemented by administrative technical measures to regulate access to microdata and to ensure that individual data cannot be disclosed.

Principle 4 is important to increase public confidence that microdata is being used appropriately. The NSO’s decisions about providing access to microdata should be transparent, and inform members of public on how and to whom microdata will be released.

Legislation on Microdata Release

According to UNECE (2007), as highlighted by Principle 3 in previous section, a legislation supporting microdata release should essentially exist. This legislation needs to cover different aspects including the conditions of data release, and the consequences of breaching the conditions. It should also determine what can and cannot be done, and for what purposes the microdata can be used. Existence of such legislation provides:

- Public confidence in the arrangements
- Mutual understanding between the NSOs and researchers on the arrangements
- Greater consistency in the way research proposals are treated
- A basis for dealing with breaches

The legislation can be available in primary legislation, law or any other form of legal authorisation. The details may be better suited to regulations, rules, etc. However, the NSO should have some legal authority to permit release of microdata even in an anonymised form.

TEMPLATE

UNIT 2

PROTECTING TABULAR DATA

Tabular data are the most common products of statistical agencies that are aggregate information on individuals within the target sub-populations (each cell in statistical table is one desirable sub-population). Even though these tables generally do not contain individual's information, there are situations when information about an individual can be revealed. Therefore, tables need to be examined carefully and should go through SDC process in order to be immune from disclosure. This unit addresses the issues related to SDC of statistical tables, which are applicable to both the traditional paper tables and the modern tabulations through on-line query systems.

Tables are of different structures, and each type should be treated differently in SDC process. The structures of tabular data are introduced in section 2.1. The first step in SDC process for tables is determining the risky cells, i.e. the cells that can potentially disclose information on the individuals. Section 2.2 discusses some methods that one can use to identify the cells at disclosure risk. Finally, section 2.3 presents methods which can be used to control disclosure in tables.

2.1. Structure of Tabular Data

Learning Objectives

– To introduce different types of tables and enable reader to distinguish between the tables of frequency count data and the tables of magnitude data

Each entry in a statistical table represents the aggregate value of a "quantity of interest" over all individuals belonging to a unique cell (FCSM, 2005). Traditionally, tables have margins; i.e. they include cells containing row totals, column totals and grand total. Tables can be divided into two types, with respect to the "quantity of interest":

- *Frequency Tables:* Quantity of interest measures membership of individuals in a cell (a sub-population).
- *Magnitude Tables*: Quantity of interest measures something other than membership (such as average)

That means, the *frequency tables* (*tables of frequency count data*) only present number or percent of individuals within a cell, while the *magnitude tables* (*tables of magnitude data*) can

TEMPLATE

present any summary statistics of the quantity of interest, such as mean, or sum. Thus, a frequency table always contains non-negative integer values but the magnitude table can have decimal or fractional, negative or positive values. Tables 2.1 and 2.2 respectively represent examples of a frequency and a magnitude table, based on hypothetical data.

Table 2.1. Number of inhabitants by region and sex

Region	Male	Female	Total
Eastern Provinces	115,983	113,591	229,574
Northern Provinces	94,157	95,101	189,258
Southern Provinces	173,488	170,941	344,429
Western Provinces	59,491	63,987	123,478
Total	443,119	443,620	886,739

Table 2.2. Average household income by region and sex of head of household

Region	Male	Female	Total
Eastern Provinces	78	70	148
Northern Provinces	51	53	104
Southern Provinces	82	76	158
Western Provinces	69	67	136
Total	280	266	546

According to FCSM (2005) and Duncan et al. (2011), tables can also be classified with respect to their structure (rather than quantity of interest in the content):

- *Multi-dimensional Table*, is a kind of table formed by more than two categorical variables. Tables 2.1 and 2.2 both are formed by two categorical variables and are two-dimensional tables. If another categorical variable, such as age group, be added to each table, then they would change to three-dimensional tables. In practice, there may be higher-dimensional tables. The first two dimensions are called row and column, but the next more variables are called layers or pages.
- *Hierarchical Table* is a kind of table that one of the variables forming it, has a hierarchical structure; i.e. the variable has several values and each value is decomposable into other values. Geographical coding is a common example of a hierarchical variable, which can be decomposed into provinces and then into sub-provincial divisions and so on.
- *Linked Table* is a table in which several tables may be combined through the linkage provided by some common cells.

TEMPLATE

Regardless of the type, a statistical table is a collection of numbers (the internal and marginal cell values) and a collection of linear equations. These equations specify the structural inter-relations of cell values.

2.2. Assessing the disclosure risk for tables

Learning Objectives

– To learn how tables are at risk of disclosure and what are the rules for determining risky cells

Since tables do not generally contain individual's information, one may ask, "Why should the NSOs be concerned about the statistical disclosure via statistical tables?" In fact, cell values of each table are compiled by summarizing data collected on specific variables. When distribution of a variable is highly skewed, it may result in tables where small numbers of individuals contribute to a single cell. Tables on business enterprise information are good examples of this kind. When fewer individuals fall into a cell, the information on their characteristics are more likely to be revealed by an intruder. In extreme cases, presence of zero value in table shows that none of the individuals in the population possesses that specific characteristic, which turns the cell with zero value to a potentially disclosing cell.

A further problem that can arise only with tables is that of table linkage. This occurs when an intruder can combine few tables based on some common variables to disclose information about individuals. Following is an illustration presented by Duncan et al. (2011) on how three non-disclosing tables can be linked based on common variables and yield a disclosing table:

TEMPLATE

Table 2.3

Var2	**Var1**	
	A	**B**
C	3	9
D	2	2

Table 2.4

Var3	**Var1**	
	A	**B**
E	1	10
F	4	1

Table 2.5

Var3	**Var2**	
	C	**D**
E	8	3
F	4	1

Table 2.6

	Var1 and Var2			
Var3	**A,C**	**A,D**	**B,C**	**B,D**
E	0	1	8	2
F	3	1	1	0

Considering above discussions, tables are too at the risk of disclosure, and have to go through SDC process. A table, in addition to the values, contains linear equations that link the marginal values to internal values. This makes the problem of protecting table complicated.

The first step in SDC process for tables is determining the risky cells, i.e. the cells whose original values reveal information on the individuals. Available methods are generally based on identifying cells with small counts in frequency tables, and cells that contain dominant individuals in magnitude tables. As stated in Duncan et al. (2011), FCSM (2005), and Willenborg and De Waal (1996), the methods for identifying risky cells are:

- *Dominance rule or* (*n*, *k*)-*Rule*: In a magnitude table, a cell is declared as risky if a small number (*n*) of individuals contribute a large percentage (*k*) to the total cell value. The NSOs should choose n and k based on their desired strictness. These values should never be published, and must be kept confidential. The linear measure for (*n*, *k*)-*rule* is given by:

  $$S^{(n,k)}(X) = \sum_{i=1}^{n} x_i - \frac{k}{100-k} \sum_{i=n+1}^{N} x_i$$

  For a given cell *X*, values of the *N* individuals who contribute to that cell are arranged in descending order ($x_1 \geq x_2 \geq \cdots \geq x_N \geq 0$). The cell is risky if $S^{(n,k)}(X) > 0$. Note that if $N \leq n$, then $S^{(n,k)}(X) > 0$, so the cell *X* is risky for any value of *k*.

  In application, the individuals who belong to a cell are sorted by descending order of their values. If the largest "*n*" individuals contribute at least "*k*%" of the total value of the cell, then the cell is classified as risky. For example, consider the cell value is 78, while five individuals contribute with the values 24, 19, 17, 10, 8. Then given $n = 3$ and $k = 75$, this cell is identified as risky since the sum of three largest individuals equals to 60 which is about 77% of total value.

- *p-Percent Rule*: In a magnitude table, a cell is identified as risky if any contributing individual value to that cell can be estimated by other individuals with accuracy of more than *p*% of its actual value for a pre-specified value of *p*. The NSO should keep the value of *p*, confidential. The linear measure for *p-Percent rule* is given by:

  $$S^{p\%}(X) = x_1 - \frac{100}{p}[T - T_c - x_1]$$

  Where, for a given cell *X* which is a sub-population of *N* individuals, x_1 is the largest value within the cell, T is the total value of the cell, and T_c is the total value of a coalition; i.e. a group of individuals who decide to pool their data together to estimate the largest value of the cell. The cell is identified as risky if $S^{p\%}(X) > 0$. Note that if $N < 3$, then $S^{(p\%)}(X) = x_1 > 0$, so the cell is risky for any value of p and c (number of individuals in the coalition group).

As an example of application of the rule, consider the values 62, 52, 15, 10, and 4 contribute to a cell value. Then in the simple case where $T_c = x_2 = 52$ (i.e. the coalition group consist of only one individual, the respondent with the second largest value who attempts to estimate the largest value), and with $p = 50$, the cell will be identified as risky because $S^{50\%}(X) = 62 - 2[143 - 52 - 62] = 4 > 0$. It can be understood that *p-Percent* rule is trying to protect the largest value from an approximate disclosure by a coalition of other individuals (say companies), especially when respondent with the second largest value is part of the coalition.

- *Prior/Posterior Ambiguity Rule or q/p rule*: As an extension to the *p-Percent rule*, it is assumed that the coalition group may have a prior knowledge (*q*) about an individual value (normally the largest value) of the desired variable, and gain more knowledge (*p*) after release of the table. The rule, used for magnitude table, is then constructed based on measuring the relative size of prior and posterior knowledge. Meaning that, prior to releasing of a table, the contribution of an individual to a cell is estimated by coalition group with accuracy of *q%* of its actual value, after the table is released, this estimate would be within *p%* (with $p < q < 100$) of actual value. As in the *p-Percent rule*, here the desirable rule is the one which allows no one can estimate the largest value (x_1) in the cell with accuracy of *p%* of its actual value. If *q/p* ratio is small, then the information gain from released table is large and the cell is declared as risky. The NSOs should determine *p* and *q* in a way, which yields to low *q/p*. A *q/p rule* for a coalition of size *c* is given by:

$$S^{pq}(X) = x_1 - \frac{q}{p}\sum_{i=c+2}^{N} x_i$$

Where, for a given cell *X* that is a sub-population of *N* individuals contributing to that cell, x_1 is the largest value of the cell. The cell is determined to be risky if $S^{pq}(X) > 0$. It can be easily seen that *q/p rule* is the same as *p-Percent rule* when $\frac{q}{p} = \frac{100}{p}$. Note that if $N < 3$, then $S^{pq}(X) = x_1 > 0$, so the cell *X* is risky for any value of *q/p*.

As an example of application of the rule, consider values 40, 20, 11, 6, 2 contributing to a

TEMPLATE

cell, then the cell is risky with $p = 25, q = 50$, and coalition of size one, because $S^{pq}(X) = 40 - 2(19) = 2 > 0$.

- *Threshold Rule or n- Rule*: This is the common rule for frequency tables. An arbitrary small value, usually 3, is specified as the threshold, then all the cells with that size or less are classified as risky.

- *Subtraction-Attribution Probability (SAP)*: This measure was proposed by Smith and Elliot (2008). It is the probability of an intruder being able to recover one or more zeros in a table, given specified knowledge about the population. SAP measure can be applied, equally to perturbed or unperturbed tables, and efficiently deals with the linked tables problems. This technique is a well theoretically grounded algorithm but further work needs to be done to integrate it into the practice. For more details, refer to Duncan et al. (2011).

2.3. Statistical Disclosure Control methods for tables

Learning Objectives

– To learn perturbative and non-perturbative SDC methods used for magnitude and frequency tables

After risky cell is identified, an appropriate SDC method should be applied to treat the table prior to releasing. There are two types of approaches to concealing sensitive information in tables: *Perturbative* and *Non-perturbative* methods. *Perturbative* methods falsify the data while *non-perturbative* methods just reduce the amount of information released. Obvious mechanisms to reduce released information, among others, are "reducing the number of variables" or "reducing level of detail" within a table.

Beside the protection methods used after tabulation, one can utilize pre-tabular methods. It implies that SDC methods can be applied to the original data prior to tabulation. In this way, all the tables generated based on that original data would be fully protected.

The selection of a SDC technique depends on whether the cell values of a table represent frequency or magnitude measures. This section presents the commonly used techniques, as introduced in Duncan et al. (2011) and FCSM (2005).

- *Table Redesign* also called *Table Restructure* or *Global Recoding* is the oldest technique to protect frequency and magnitude tables that have too many risky cells. In this method, rows or columns containing risky cells are simply combined. Since merging the cells may lead to

TEMPLATE

loss of information, this method is not recommended for tables where the variables have very few numbers of categories. This method is not feasible where the table layout is fixed, due to the NSO's previous publications of similar results. Table 2.7 shows reconstruction of a table, using hypothetical data.

Table 2.7. Number of students by education level of their fathers and sex
Example of Table Redesign

Original Table

Education Level	**Sex**		**Total**
	Male	**Female**	
Low	2	0	2
Medium	1	4	5
High	12	10	22
Very High	1	2	3
Total	16	16	32

→

Redesigned Table

Education Level	**Sex**		**Total**
	Male	**Female**	
Low & Medium	3	4	7
High & Very High	13	12	25
Total	16	16	32

- *Cell Suppression* is one of the most common techniques for protecting risky cells, in both frequency and magnitude tables. In this method, risky information is protected through hiding (suppressing) the values of few cells by replacing them by a specified symbol, say an asterisk. In cell suppression method, initially the risky cells are suppressed; this is named *primary suppression*. Since marginal values are available, primary suppression is not sufficient to obtain a safe table and additional cells must be suppressed. The second phase is called *secondary suppressions* or *complementary suppressions*. Determining the secondary suppression without yielding to an unacceptable loss of information is a very complicated task. It sometimes requires applying sophisticated mathematical methods of linear programming that is out of the scope of the present discussion. For more details, see Duncan et al. (2011). Table 2.8 is a fictitious example for cell suppression (FSCM, 2005).
  An administrative way to avoid cell suppression, used by a number of statistical agencies, is to obtain written permission, or *informed consent* to publish a sensitive cell, from the respondents who contribute to the cell.

- *Random Rounding* is a perturbative method for frequency tables. In this method, all table cell values are randomly rounded up or down to the nearest multiple of a base that is equal to the specified threshold (say "3"). For example, when threshold is "3", the cell value that is "2" can be rounded up to "3" or rounded down to "0" using a random basis. The problem with this method is that rounding is done separately for each cell. Thus after random

TEMPLATE

rounding is applied, the row and column cells do not necessarily add to the published marginal totals. Table 2.9 shows a possible result (FCSM, 2005).

Table 2.8. Number of delinquent children by county and education level of household head
Example of Cell Suppression

Original Table

County	Education Level of Household Head				
	Low	Medium	High	Very High	Total
Alpha	15	1*	3*	1*	20
Beta	20	10	10	15	55
Gamma	3*	10	10	2*	25
Delta	12	14	7	2*	35
Total	50	35	30	20	135

Reference. FCSM (2005)

→

After Suppression

County	Education Level of Household Head				
	Low	Medium	High	Very High	Total
Alpha	15	*	*	*	20
Beta	20	10	10	15	55
Gamma	*	*	10	*	25
Delta	*	14	*	*	35
Total	50	35	30	20	135

Table 2.9. Number of delinquent children by county and education level of household head
Example of Random Rounding

Original Table

County	Education Level of Household Head				
	Low	Medium	High	Very High	Total
Alpha	15	1*	3*	1*	20
Beta	20	10	10	15	55
Gamma	3*	10	10	2*	25
Delta	12	14	7	2*	35
Total	50	35	30	20	135

Reference. FCSM (2005)

→

After Suppression

County	Education Level of Household Head				
	Low	Medium	High	Very High	Total
Alpha	15	0	0	0	20
Beta	20	10	10	15	55
Gamma	5	10	10	0	25
Delta	15	15	10	0	35
Total	50	35	30	20	135

- *Controlled Rounding* is a perturbative method for frequency tables that is developed to solve the additivity problem of random rounding method. In this method, linear programming methods are used to round the risky cell values in a way that they would add up to the

TEMPLATE

published marginal totals. Table 2.10 illustrates controlled rounding where sum of the cell values in each row and column are constrained to equal the marginal totals (FCSM, 2005).

- *Controlled Tabular Adjustment* is a perturbative method for both frequency and magnitude tables. In this method, in each risky cell, the original value is replaced by a safe value that is in a *sufficient distance* away from the original value; then marginal values are minimally adjusted to ensure additivity. A *sufficient distance* from the original value is a value that should be added to the risky cell value in order to make it a non-risky cell.
  Table 2.11 shows an example of controlled tabular adjustment (FCSM, 2005). In this example, threshold is '3' and sufficient distance changes the original value by either '1' or '2'. Here, the risky cells are firstly arranged in descending order from the most to the least risky$(2, 2, 1, 1)$. Then the first cell is changed at random to '0' or '3', by alternatively subtracting '2' or adding '1'. Subsequent adjustments will be implemented with alternative signs. Finally, the marginal values are re-computed to account for the changes imposed on the internal cells.

Table 2.10. Number of delinquent children by county and education level of household head
Example of Controlled Rounding

Original Table

County	**Education Level of Household Head**				
	Low	**Medium**	**High**	**Very High**	**Total**
Alpha	15	**1***	**3***	**1***	20
Beta	20	10	10	15	55
Gamma	**3***	10	10	**2***	25
Delta	12	14	7	**2***	35
Total	50	35	30	20	135

Reference. FCSM (2005)

After Suppression

County	**Education Level of Household Head**				
	Low	**Medium**	**High**	**Very High**	**Total**
Alpha	15	0	5	0	20
Beta	20	10	10	15	55
Gamma	5	10	10	0	25
Delta	10	15	5	5	35
Total	50	35	30	20	135

- *Pre-tabular methods* are perturbative methods that can be applied to the table's underlying microdata prior to tabulation in order to assure that any table generated from that data is fully protected. Applying pre-tabular methods provides simplicity in protecting tables especially for the tabulations through on-line query systems. These perturbative methods are exactly

TEMPLATE

the ones applied for protecting microdata and will be introduced in the next unit.

Table 2.11. Number of delinquent children by county and education level of household head
Example of Controlled Tabular Adjustment

Original Table

County	**Education Level of Household Head**				
	Low	**Medium**	**High**	**Very High**	**Total**
Alpha	15	**1***	**3***	**1***	20
Beta	20	10	10	15	55
Gamma	**3***	10	10	**2***	25
Delta	12	14	7	**2***	35
Total	50	35	30	20	135

→

After Suppression

County	**Education Level of Household Head**				
	Low	**Medium**	**High**	**Very High**	**Total**
Alpha	15	**1*-1=0**	3	**1*+2=3**	**20+1=21**
Beta	20	10	10	15	55
Gamma	3	10	10	**2*-2=0**	**25-2=23**
Delta	12	14	7	**2*+1=3**	**35+1=36**
Total	50	**35-1=34**	30	**20+1=21**	135

Reference. FCSM (2005)

TEMPLATE

UNIT 3

PROTECTING MICRODATA

The data collected by national statistical agencies, mainly through surveys and administrative sources, are considered to be public goods and data producers have to make every effort to facilitate the best utilization of the data in order to extract knowledge. This requires an open access policy for providing data to variety of users (including research community) along with an easy and affordable access to the microdata for statistical purposes. However, access to microdata can increase the risk of individuals' sensitive information disclosure.

There is therefore a need to balance the conflicting issues of "providing researchers with microdata" and "preserving the data confidentiality". This unit intends to discuss several existing methods that can protect microdata against disclosure and yet maintain the data utility.

Section 3.1 addresses the key characteristics and uses of microdata. Section 3.2 discusses how the microdata is at risk and how the risk can be assessed. The last section presents approaches to control the risk of disclosure and protect microdata.

3.1. Key characteristics and uses of microdata

Learning Objectives

– To understand the microdata, its content, characteristics, and uses

Microdata is in fact the raw material based on which statistical agencies construct their outputs, such as tables and graphs. This raw material can be obtained from a sample survey, a census, or even some administrative sources. Microdata is traditionally organized in a database, where each record contains information about one single individual. Thus, for each population unit represented in microdata, there is a record containing the values of multiple attributes. As an example, the microdata of the students of a class may consist of the attributes like name, sex, date and place of birth, place of residence, height, weight, scores, their parents' education level, their family income, etc. A microdata that contains all the variables for the whole population is called *full table* (Duncan et al., 2011).

TEMPLATE

Microdata Content

According to Willenborg and De Waal (1996), variables contained in a microdata can be subdivided into two kinds, *identifying* and *non-identifying* variables. Identifying variables are the variables that their value, alone or in combination with values of other variables, can lead to re-identification of an individual. Identifying variables can be divided into "*direct identifiers*" (such as social security number), and "*indirect identifiers*" (such as age, place of birth, and education). A single *direct identifier* surely leads to an immediate and absolute re-identification; thus, it should never be published. A single *indirect identifier* cannot yield re-identification, but a rare combination of few indirect identifiers makes re-identification possible. Thus in SDC process, rare combinations should be detected and removed. To simplify the detection process, firstly the indirect identifiers who participate in a combination should be classified based on the extent to which they are identifying. For classifying these variables, one can use the following characteristics in addition to common sense and experience:

- *Rareness*: The variable has a value that may occur rarely in the population, such as nationality, which may refer to very few individuals from a specific country.
- *Visibility*: The variable has a value that is known or can be ascertained easily, such as sex, and also place of residence.
- *Traceability*: Based on the variable's value the individual can be traced and located easily, such as place of residence, or other regional variables.

More details about rareness and handling it in SDC process will be described in the next sections.

In a microdata, there may be other variables, as follows:

- *Sensitive variable* is a variable whose value can represent characteristics that an individual does not like to be revealed, such as criminal past or mental health. Note that a sensitive variable may be an identifying variable, as well.
- *Household variable* is a variable whose value is the same for all members of the same household, such as household size or characteristics of household head.
- *Sampling weight* is a variable that contains sampling weighting scheme.

Each of the variables mentioned above, can in some way contribute to disclosure, and thus deserve special consideration through SDC process. (Section 3.3. discusses this in more details)

Microdata Uses

The use of microdata has evolved over time due to the technological advancements that have greatly changed the landscape of statistical services. In recent decades, role of microdata has

TEMPLATE

shifted from a mere raw material for statistical production to an important statistical product. Today, microdata is an essential factor in analysing the complex world around and finding solutions to multidimensional and complex issues. It enables researchers to conduct more sophisticated data analyses and apply advanced statistical methods (such as modelling) to extract information from data. Microdata with geographical and temporal details can provide a better understanding of many spatio-temporal phenomena. Hierarchical microdata (e.g. members of households or employees of enterprises) allow multi-level analyses that take account of variability at each level.

The NSOs can no longer avoid releasing microdata only for confidentiality reasons; instead, they have to find solutions to provide microdata under confidentiality constraints. Applying the SDC methods for protecting individual information are good alternatives to the restriction of access to microdata. Moreover, it gives the NSOs a flexibility to classify the users, and impose different levels of protection on microdata for different users; for example, the microdata provided to research community can be less protected than the public used files (PUFs).

3.2. Assessing the disclosure risk for microdata

Learning Objectives

– To identify how microdata is at risk of disclosure and become familiar with the risk metrics

Containing many attributes (often of high sensitivity) at individual level, makes microdata a good target for intruders, armed with computational capabilities and with access to other relevant external data sources. Intruders desire to enhance and enrich their data in hand, by gaining further information about specific individuals through matching. *Matching* means an attempt to link one or more records in one dataset (identifying dataset), with the records in another data set (target dataset), by using a common set of identifying variables. The term *match* refers to a successful matching attempt. Figure 3.1 illustrates how disclosure can occur through a successful matching (Duncan et al., 2011).

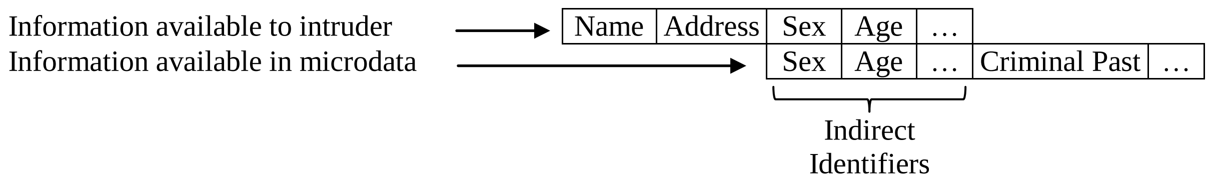

Figure 3.1. An illustration of successful matching leading to disclosure

TEMPLATE

To protect microdata from the risk of disclosure, the NSOs should have a risk management plan. The first step is to find out how a disclosure event might happen, by answering questions like who are the intruders, what are they trying to achieve, and what is available to them. Answers to such questions will form the disclosure scenario[3]. In a disclosure scenario, the plausible risk channels for a data intrusion are classified, and the output is a specification of a set of identifying variables that are likely to be held by intruders. This establishes the basis for defining the risk of disclosure.

Disclosure Risk Metrics

According to Hundepool et al. (2010), intuitively, an individual is at risk of identification when it cannot be confused with several other individuals, and thus can be singled out from the rest. The basic idea behind the most of risk measures is *rareness* (or *uniqueness*) of an individual either in sample or in population. While, in some of risk measures, this idea is extended to measuring "*the extent to which matching is possible*". The concept of *rareness*, though seems quite simple, is difficult to be expressed by means of statistical methodology.

There are many different methods and tools for microdata disclosure risk assessment used by statistical agencies. However, there is not yet agreement on which method is the best under given circumstances, neither is international standards on the subject matter. The followings are dominating approaches towards microdata disclosure risk assessment mentioned in Duncan et al. (2011) and Hundepool et al. (2010).

- *Threshold*: Early on, disclosure risk measures often used population thresholds. Based on the type of the disclosure risk measure and desired level of protection, a threshold for the population size is set. The microdata is released if the risk is below the threshold, and is sent for SDC treatment if the risk is above it. For example, it is decided that each cross-classification of two identifying variables, should contain values of at least 2000 individuals. Threshold rules are simple to understand and implement but they do not consider what other techniques an intruder may apply for data disclosure. Thus, more sophisticated measures are required to ensure a fully protected microdata is being released.

- *File-Level Risk Metrics* measure the average risk across the whole microdata file. Since they give a useful overview of the whole file risk, the NSOs have found them useful in determining how much the SDC treatments have been beneficial. File-level risk metrics may be briefly described as follows.

[3] For more details, refer to Duncan et al. (2011).

TEMPLATE

- *Population uniqueness*: As mentioned before, *uniqueness* (or *rareness*) is an important aspect in understanding disclosure risk for microdata. When an individual has unique values on a set of identifiers within a given population, then that individual is said to be a *population unique*. For example, if set of identifiers include age, sex, occupation and marital status, a cross-classification of these variables may yield to cells containing one or a small group of records. In this case, a 40-year-old female elementary school teacher who is widowed may be uniquely identified in the population, or there might be very few numbers of records with the same characteristics. When an individual is known to be unique in a certain population and there is a possibility of identifying that individual in a given microdata set, the chance of identification disclosure is considered high. The proportion of such individuals in a given population is the level of *population uniqueness*. This metric is simple to implement and understand but it is not always possible to compute. Mainly because determining *population uniqueness* requires access to population data, which is, except in the case of census and some administrative sources, rarely possible. The microdata often comes from sample surveys or administrative data without a full coverage of the total population, and uniqueness of an individual in the dataset does not necessarily lead to *population uniqueness*.

- *The Proportion of Sample Unique Individuals that are Population Unique*: To determine this measure, one should compute the probability of an individual being the *population unique* on a set of identifiers given that it is *sample unique* in a sample dataset on the same set of variables. This conditional probability is considered to be more useful metric of disclosure risk than the level of *population uniqueness*, because this metric is sensitive to changes in sampling fraction in a monotonic way. Yet, the problem of access to the population data remains with this method. In addition, the sampling fraction is not directly incorporated into the estimation of conditional probability, while the chance of an individual being selected in the sample has a direct impact on the probability of uniqueness in the population and consequently on the risk of disclosure.

- *Skinner and Elliot Method*: The alternative approach, developed by Skinner and Elliot (2002), overcomes the problems associated to the previous two methods. It focuses on the conditional probability of a correct match given a unique match. Their approach is a development of the method proposed by Elliot (2000).

  Elliot (2000) observed that an intruder attack can be mimicked by:

  1) Removing a record from a target microdata set at random with the probability equal to

TEMPLATE

the sampling fraction

2) Conditionally copying the record back into the microdata set with the same probability as the original sampling fraction

3) Matching the removed record against the microdata (based on a given set of identifiers)

The possible outcome of the process is shown in Table 3.1. The table shows all possible outcomes of implementing three steps on a record, based on the uniqueness of the record in the sample; i.e. removing the record from the microdata and copying it back into the data with a probability equal to the sampling fraction and finally matching the removed record. There are two critical cells, which are marked in the table. The first is when the removed record is sample unique and copied back into the file yielding a correct match, and the second shows a situation in which removed record is one of the two records with the same set of identifiers and not copied back into the file resulting in a false match.

Table 3.1. Possible outcome of Elliot's proposed process

Record is:	Copied back into the file	Not copied back
Sample unique	**Correct unique match***	Non-match
One of a sample pair	Multiple match including the correct match	**False unique match***
One of a larger equivalence class	Multiple match including the correct match	False multiple match

Reference. Elliot (2000)

Accordingly, the probability of a correct match given a unique match can be estimated through sums of the number of records with sample frequencies equal to '1' or '2' and the given sampling fraction. Assume that there is J unique values obtained from cross-classification of the set of identifiers (refer to the example under *population uniqueness*). For each value of j $(j = 1,2,..,J)$, there may be one or more records in the population. F_j and f_j are respectively the frequency of j within the population and the sample. For instance, if j refers to the 20 year old married male medical doctor living in rural area (combination of age, sex, marital status, occupation, area) and there are two people with the same characteristics in the population and only one in the sample, then $F_j = 2$ and $f_j = 1$. Then the conditional probability of a correct match given a unique match can be estimated as:

$$P(cm|um) = \frac{\sum_j I(f_j = 1)}{\sum_j F_j I(f_j = 1)}$$

Where, $I(A)$ is an indicator variable that $I(A) = 1$ if the condition A is correct (e.g. $f_j = 1$ means that frequency of j in the sample is "1") and $I(A) = 0$, otherwise. Notice that the denominator is the number of possible selected records from the population for which there is a unique match in the sample.

It is now worth to return to the first two methods and estimate probabilities with the same logic and notations. Then $P(PU)$ is the probability of *population uniqueness*, in which N is the size of population, and $P(PU|SU)$ is the conditional probability of an individual being *population unique* given that it is *sample unique* on the same set of variables:

$$P(PU) = \frac{\sum_j I(F_j = 1)}{N}$$

$$P(PU|SU) = \frac{\sum_j I(F_j = 1 \; and \; f_j = 1)}{\sum_j I(f_j = 1)}$$

All three proposed probabilities depend on the population frequencies, F_j, which makes calculation of the probabilities impossible since the population units are not available in practice. Skinner and Elliot (2002) examined characteristics of $P(cm|um)$ under different sampling designs from which sample microdata was drawn and proposed probability estimations under each design. In its simplest case, when all population units are independently sampled with common probability of π in a Bernoulli sampling, the sample frequencies, f_j, are independently binomially distributed. Under this design, Skinner and Elliot (2002) proposed the following estimation for $P(cm|um)$:

$$\hat{P}(cm|um) = \frac{\pi \sum_j I(f_j = 1)}{[\pi \sum_j I\big(f_j = 1\big) + 2(1 - \pi) \sum_j I\big(f_j = 2\big)]}$$

Two important points about this estimator are that (a) it no longer depends on the population and can be obtained only from sample microdata; and (b) it includes not only frequencies for single record identifier j values, but also j values with double records. One should remember that this estimation is under a simple sampling design with

TEMPLATE

common selection probability π that may not be always the case in practice. Extension of this method for other sampling designs is discussed in Skinner and Elliot (2002).

With the file-level risk metrics being mentioned, it should be noted that though risk analysis at the file level is useful, it only provides a partial measure of identification risk, and thus record-level metrics are required.

- *Record-Level Risk Metrics* measure the disclosure risk for the records. Intuitively, records that are *unique* (or *rare*) have high disclosure risk, and naturally, more attention should be paid to them. Nonetheless, the problem is that usually microdata is only a sample of a population and not the whole population. Therefore, a record that is unique in the sample microdata is not necessarily unique in the population. To develop this intuition further, consider a cross-classification of identifiers, where each cell is cross product of the categories of identifiers. As introduces earlier, let F_j be the number of individuals in the population that belong to cell j, and let f_j be the given sample frequency of this cell. Then the probability of re-identification of an individual in cell j is $P = 1/F_j$. Usually, population frequencies are not available, and therefore this probability should be estimated. Followings are few available methods that may be used for estimating *P*:

  – *Probability Modelling Approaches*: In this class of approaches, there are two main methods to infer population frequency F_j from the sample frequency f_j (a) *Poisson Model* which is based on the assumption that $F_j|f_j$ has a Poisson distribution; and (b) *Argus Model* which is based on the assumption that $F_j|f_j$ has a Negative Binomial distribution. In both methods, the individual risk measures can be aggregated to obtain a global risk measure for the entire file[4].

  – *Special Uniqueness* means that a record that is sample unique on a set of identifiers is also unique on a subset of those identifiers. Empirical work has shown that special uniques are more likely to be population unique than the random uniques. Special uniques can be classified according to the size and the number of the smallest subset of identifiers that defines the record as unique, known as minimal sample uniques (MSU). In the Special Uniques Detection Algorithm (SUDA), (Elliot et al., 2005) all MSUs are found for each record on all possible subsets of the identifiers, where the maximum size

[4] For more details, refer to Duncan et al. (2011) or Hundepool et al. (2010).

TEMPLATE

of the subsets (*m*) is specified by the user. SUDA grades and sorts all the records within a microdata according to the level of risk. The method assigns a "per record matching probability" to a sample unique based on the number and size of minimal uniques[5].

- *Record Linkage Techniques*: The risk assessment approaches mentioned above, are all applicable when the identifiers within a microdata are categorical variables. If identifiers are continuous variables, none of them can be used. For continuous identifiers, the concept of *rareness* transforms to the concept of "*rareness in the neighbourhood*". The *record linkage* techniques provide a method for measuring this new concept.
  In record linkage technique, an estimate of the probability of re-identification is obtained by attempting to link a record in a second data set with a record in the microdata to be released. The number of matches gives an estimation of the number of records at risk. Accordingly, disclosure risk is defined as the proportion of matches among the total number of records in microdata. In record linkage techniques, one can assess disclosure risk by linking the microdata to "*an external data set*" or "*the Pre-SDC version of the same microdata*". The main types of record linkage used to measure re-identification risk are "*Distance-based record linkage*", and "*Probabilistic record linkage*". Further details on the records linkage are beyond the scope of this text and the reader can refer to the Duncan et al. (2011) and Hundepool et al. (2010) for more details.

3.3. Statistical Disclosure Control methods for microdata

Learning Objectives

– To learn perturbative and non-perturbative SDC methods used for microdata

After assessing the risk of disclosure, an appropriate SDC method should be applied to treat the microdata distinguished to be at disclosure risk prior to its release. The purpose of SDC process is to prevent disclosure of individuals' information and at the same time maintaining the utility of data. In other words, the output should be a microdata that does not permit the intruders to discover the individuals' information through linking, while allowing the users to do their desired analyses and get similar results that would be obtained from analyses of the original microdata. This section, based on the works done by Hundepool (2010), FCSM (2005), and Willenborg and De Waal (1996), presents the commonly used SDC methods for microdata.

TEMPLATE

In general, SDC methods available for masking microdata can be divided into two categories:

- *Perturbative masking* methods that distort microdata set before release by replacing the unique combinations of values available in original dataset by new unique combinations in the perturbed dataset. This deliberately made confusion is beneficial for protecting confidentiality. To maintain the data utility, the perturbation should be in a way that the computed statistics from the original and the perturbed dataset do not differ significantly.

  Perturbative masking methods are mostly special cases of matrix masking; i.e. if the original microdata set is X, then the masked microdata set (Z) is computed as $Z = AXB + C$, where A is a record-transforming mask and C is a displacing mask or noise (Duncan and Pearson, 1991)
- *Non-Perturbative masking* methods do not alter data but produce partial suppression or reductions of detail in the original dataset.

Choosing an appropriate masking method depends on the type of variable on which the method will be applied. There are two types of variables considered in the SDC process:

- *Continuous variable*: A variable is called continuous if it is numerical and arithmetic operation can be performed on it (e.g. age and income). In the process of SDC, a continuous variable has the advantage of being masked by using arithmetic operations. However, its numerical nature makes each combination of its values likely to be unique.
- *Categorical variable:* A variable is called categorical if it takes values that may seem numerical but are not numbers in nature and arithmetic operations on it makes no sense (e.g. sex and education level). Most of the indirect identifiers are categorical. In the process of SDC, a categorical variable has the advantage that its value range is limited; but incapability of using arithmetic operations leads to inconvenience.

Table 3.2 lists common perturbative and non-perturbative methods, appropriate to mask continuous and categorical variables.

- *Noise addition* is a perturbative method mostly suitable for continuous variable. In this class of methods, a random noise is added to (sometimes multiplied by) the data. Main noise algorithms mentioned in literature are as follows:
  - Masking by uncorrelated noise addition
  - Masking by correlated noise addition
  - Masking by noise addition and linear transformation
  - Masking by noise addition and non-linear transformation

TEMPLATE

Table 3.2. Masking methods for microdata for different types of data

Type of variable / Method	Continuous	Categorical
Perturbative Masking	Noise Addition Micro-aggregation Swapping/Rank Swapping Rounding Re-sampling	Swapping PRAM MASSC
Non-Perturbative Masking	Global Recoding Top and Bottom Coding	Global Recoding Top and Bottom Coding Local Suppression Sampling

In practice, only the first two options and rarely the third one are used. The last algorithm is not a very practical, but it is the only one that can be applied for categorical variable too.

The noise is typically continuous, with mean zero and constant variance. One main challenge regarding the constant variance is that the small values of the variable will be strongly perturbed, while the large values remain less perturbed. As an example, in the business microdata, the large enterprises that are at higher risk of disclosure, will be weakly perturbed and remain at risk. A possible way out of this challenge is to use *Multiplicative Noise* approach, especially for highly skewed variables. For further details on this approach, refer to Hundepool et al. (2010).

- *Micro-aggregation* (also called *blurring*) is aggregating across small groups of individuals and replacing one individual's original value with the group average. There are many possible ways to implement this perturbative method. Groups of records for averaging may be formed by matching on other variables or by sorting the variable of interest. The number of records in a group (whose data will be averaged) may be fixed or random. The average associated with a particular group may be assigned to all members of a group, or only to the middle member. In case that there are more than one continuous variables at risk, the grouping may be the same or different for each of the variables. As an example, to blur the income of individuals, one can first group the records based on province, sex, and education level. Then for all the individuals who belong to one single group, their income is replaced by the average income in that specified group.

TEMPLATE

- *Swapping* refers to transforming a dataset by exchanging values of variables among records. This perturbative method involves selecting a sample of the records, finding a match in the dataset based on a set of categorical variables then swapping all other variables. For example, to protect the variable household income, records of the households who are living in different provinces but have been matched based on sex and education level of the household head, can be swapped. *Targeted Swapping* involves the records with high risk. Swapping offers the opportunity of preserving some statistics through swapping operation by forcing agreement between the swapped pairs on the variables that contribute to statistics of interest.

  Rank Swapping permits using continuous variables to define pairs of records to be swapped. In this method, the pairs are those records that are close to each other based on a list that has been sorted by the continuous variable. Those records with close ranks on the sorted variable form the potential pairs for swapping.

- *Rounding* is a perturbative method that replaces the original values with rounded values, and it is suitable for continuous variables. Rounded values are chosen from a set of rounding points defining a *rounding set*. Rounding is usually performed on one variable at a time (univariate rounding), although multivariate rounding is also possible.

- *Re-sampling* is a perturbative method in which t independent samples, say $S_1, \ldots, S_t$, are taken from the values of the original variable, say X. All the samples are sorted using the same ranking criterion. Then the masked variable Z is built as $\bar{x}_1, \ldots, \bar{x}_n$, where n is the number of records and $\bar{x}_j$ is the average of the j^{th} ranked values in $S_1, \ldots, S_t$.

- *PRAM (Post-Randomization Method)* is a probabilistic, perturbative method for categorical variables. In the masked dataset, the values of some categorical variables for certain records in the original file are changed to a different value according to a prescribed probability mechanism, namely a Markov matrix. The Markov approach makes PRAM very general, because it encompasses noise addition, data suppression, and data recoding. The PRAM matrix contains a row for each possible value for each variable to be protected. This, rules out using the method for continuous variables.

TEMPLATE

- *MASSC* is a perturbative masking method whose acronym summarizes its four steps: (a) Micro Agglomeration, (b) Substitution, (c) Sub-sampling, and (d) Calibration. The purpose of these four steps are briefly as follows, for more details, refer to Hundepool et al. (2010):
  a) Micro agglomeration is applied to partition the original dataset into risk strata (groups of records that are at a similar risk of disclosure). These strata are formed by using the key variables, i.e. the indirect identifiers in the records. The idea is that those records with rare combinations of indirect identifiers are at a higher risk.
  b) Optimal probabilistic substitution is then used to perturb the original data. (i.e. substitution is governed by a Markov matrix like in PRAM.
  c) Optimal probabilistic sub-sampling is used to suppress some variables or even entire records (i.e. variables and/or records are suppressed with a certain probabilities set as parameters).
  d) Optimal sampling weight calibration is used to preserve estimates for outcome variables in the treated dataset whose accuracy is critical for the intended data use.

- *Sampling* is a non-perturbative method, in which instead of releasing the original microdata file, a sample *S* of the original set of records is published. Sampling is suitable for categorical variables. This method is often used for releasing the microdata of population censuses.

- *Global Recoding* is a non-perturbative method which can be used for both categorical and continuous variables. For categorical variable, several categories are combined to form new less-specific categories; e.g., elementary school teachers and high school teachers can be recoded to teacher. For continuous variables, the values are recoded into intervals; e.g. incomes are presented in 10,000 dollar intervals.

- *Top and Bottom Coding*, as a non-perturbative method, is a special case of global recoding, where the top values (those above a certain threshold) are lumped together to form a new category. The same is done for the bottom values where those below a certain threshold are combined together.

- *Local suppression*, is a non-perturbative method, in which certain values of individual variables are suppressed. For example, if in cross-classification of few identifiers such as sex, occupation and marital status, one or more values are considered to be rare (a female elementary school teacher who is widowed), the value of one of these identifier variables

TEMPLATE

(say marital status) can be replaced by a missing value.

Beside the methods mentioned above, there are certain rules that should be followed to protect the microdata:

- Direct identifiers should never be released, and number of other identifiers should be limited.
- Indirect identifiers should be subdivided from "most identifying", "more identifying" and "identifying". To do this, one can use *rareness*, *visibility*, and *traceability* criteria. Any identifier that meets at least two of these criteria, can be classified as *most identifying*, and so on. Apart from these criteria, one has to rely on common sense and experience. Then the data should be checked so that any combination of three identifiers that are $most\ identifying \times more\ identifying \times identifying$ should not occur below a specified threshold in the population.
- Geographic detail should be limited.
- Number and detailed breakdown of categories within each variable should be limited.
- Very sensitive variables are not recommended to be released.
- Households should be published in a way to prevent re-grouping of the household.
- Sampling weight can provide additional identifying information to intruder, thus they should be omitted or be perturbed by adding noise.
- It is recommended that microdata on business enterprises, due to the fact they are more at risk of disclosure, not to be released publicly.
- It is recommended that only those microdata sets that are sufficiently outdated be released.
- Longitudinal microdata sets are more at risk because the same individuals are surveyed several times. Special attention should be paid in publicly releasing them.

TEMPLATE

UNIT 4

ADMINISTRATIVE PROCEDURES FOR DISCLOSURE CONTROL

To provide safe and yet useful data, statistical agencies have two alternatives, "*restricting data*" by applying the SDC methods which have been introduced in previous units, or "*restricting access*" by granting the data only to specific users, under specific restrictions. These will help in struggling with deliberate attempts to breach confidentiality. However, confidentiality can be violated inadvertently, too. Thus special provisions should be made for protecting confidentiality against all kinds of threats. This unit addresses administrative procedures for safeguarding the confidentiality. Section 4.1 addresses administrative arrangements to prevent inadvertent breach of confidentiality and section 4.2 presents the issues related to restricting the access.

4.1. Administrative arrangements within statistical agencies

Learning Objectives

– To learn about the appropriate actions that should be taken by the NSOs to prevent breach of confidentiality

According to Groves et al. (2009), a variety of reasons can cause breach of confidentiality. The most common is simple *carelessness*; such as forgetting to remove direct identifiers from the questionnaires or data files, or leaving the cabinets containing the sensitive information unlocked. To manage the case, the NSOs should raise the awareness and consciousness of their employees. This can be done through providing them guidelines for appropriate behaviour, and ensuring that these are observed systematically. It is also beneficial to develop a formal written pledge that their employees are obliged to sign as a condition of their employment, and as a requirement of continued collaboration. Under this pledge, the employees undertake to be conscious of not allowing any information to be revealed to any unauthorised party. Enforcement of this pledge may involve additional legal penalties.

The security guideline can contain a set of workplace rules. These rules may emphasize on the points such as:

- Direct identifiers should be removed.
- Materials containing sensitive information should be protected with a great deal of care.

TEMPLATE

- Questionnaires containing sensitive information should be stored in the locked cabinets.
- The sensitive electronic files should be encrypted.
- The sensitive and non-sensitive files on the hard disks should be kept separately.
- Sensitive information should never be transmitted through emails.
- Sensitive files should be erased in a way that cannot be recovered.

Less common, but potentially more serious threat to confidentiality are *legal demands* for identified data. For this case, the NSOs should have a legal safeguard to protect them from being compelled to disclose the individual's information.

4.2. Components of providing restricted access

Learning Objectives

– To become familiar with the restricted access procedures

Previous sections, focused on protecting confidentiality through SDC procedures, which restrict the released data. There are also administrative procedures that restrict the access to data. These procedures have the potential advantage for specific users who can obtain data that is richer than the SDC masked data. According to Duncan et al. (2011), in the restricted access procedures, data users agree to abide by specified conditions governing the access and use of the confidential data. Their research outputs are reviewed by the NSO's staff in order to assure no confidential information is revealed. The restricted access approach can be divided into four components, the "who", "where", "what", and "how" controls. These controls are presented below.

Who can have access?

The NSO should determine the *safe* or *trusted* individual researchers or organizations. Restricted access conditions often stipulate that researchers must have specified credentials to get access to data. For judging a potential user, the NSO can consider following typical qualification:

a) Capability of doing research of scientific merit that is intended to be published
b) Proposed research cannot be accomplished with the Public Use Files (PUFs) and really requires the data sought
c) User must be associated with some institution that can assure compliance with the NSO data access requirements
d) Proposed research must be beneficial for the society and be in line with the NSO's mission;
e) Pass some form of security screening.

TEMPLATE

Where can access be obtained?

In order to better control how data users may interact with the data, the NSOs prefer that users access the data only at special sites, the *Data Laboratories* (*DL*). DL is a physically secure environment that permits use of confidential files. In DL, usually users should analyse the data on a secure computer and they are not allowed to take in mobile data storage devices (such as CDs, or USB flash drivers). Users are allowed to take away certain analytical output after being checked by the NSO's experts, with respect to avoiding disclosure.

For users, these conditions are not desirable due to different reasons such as difficulty to travel to the DL, limited working hours of the DL, inadequate and unfamiliar computing facilities available in DL, and unavailability of internet access in the DL. However, some of the NSOs have established a network of DLs throughout the country for more convenience.

What analysis is permitted?

Some of the NSOs only permit the analyses that are beneficial to the programme of the NSO. However, as a widespread requirement, users must promise that they will not carry out analyses that would result in any disclosure.

Modes of access

Following are the modes of access commonly used by the NSOs for disseminating data used outside of their organizational boundaries:

- *Free Access* is used for publishing aggregate data that is intended for public-use. It imposes no restrictions on who access the data or on what is done with the data. The dissemination medium of such data can be paper-based or web-based.
- *Delivered Access* is a more restricted form of access, in which user requests and applies for access to the data. Then the data are delivered to the user through some physical medium such as CDs or through an internet portal. Usually, the user should specify what the data are to be used for. In addition, the user is required to agree to some specified conditions as a license for data access. The range of data sets covered by this mode is very wide, ranging from PUFs to confidential data sets that might be given only to a few trusted researchers under highly restrictive licensing conditions, which may even include heavy penalties on violating the conditions.
- *Virtual Access* (or *Remote Access Facilities*) as a virtual secure environment that permits remote use of confidential data is now widely regarded as the future of data access. It combines the advantages of DL with much of desirability from the users' point of view. Its advantage over delivered access is that firstly the output is checked before being delivered to

TEMPLATE

the user. This enables the NSO to spot any data abuse. Secondly, there is no possibility of a user/intruder directly linking the accessed data set to an external source.

There are two variants on virtual access theme:

- *Direct Virtual Access* uses virtual remote network interfaces to allow the users to view, interrogate, manipulate and analyse the data as if it was on their own computer.
- *Analysis Servers* do not allow direct access to a data set, but permits the user to interrogate it. In such system, data can be analysed but cannot be viewed. The analysis server returns the output, after being checked by the NSO staff with respect to disclosure issues. Its advantage over direct virtual is that the user cannot see the data, thus there will be no risk of disclosure but for the user it would be more difficult to explore the data.

The exposure of confidential information to the risk of disclosure is not always caused by the evil intentions from the intruders. Data integration, through record linkage, is now one of the most effective modes of exploiting microdata from variety of sources (surveys, statistical registers, and recently emerged Big Data). Thanks to IT developments, data analysts/scientists are now in a good condition to conduct multipurpose and multivariate data analysis in facing complex scientific and development questions. Respondents' concerns about their confidential information remain in place with such applications of data. Hence, the NSOs, as the national data guardians should feel responsible to build the public trust and maintain the confidentiality of respondents' data. The NSOs must play an active, rather than passive, role to create an environment within which: (a) users can trust that their information is kept confidential, and their privacy is highly respected; and (b) their responses to surveys and their available records in other sources of data contribute into creating better societies, at the lowest public cost. In this regard, many advanced statistical systems have initiated open data access projects through which the NSOs (as the data holders) take the role and responsibility of integrating multiple microdata and provide integrated data to the users via variety of access modes as discussed in this unit. Good examples of integrating multiple data for official statistics are centralized administrative registers (in Nordic countries)[5], data-sharing hub (in New Zealand)[6], and data integration projects (in Australia)[7]. Only through similar effective microdata dissemination policies and infrastructure, data can be used as a strategic resource for decision making, while public trust in official statistics is maintained highly.

[5] http://www.unece.org/fileadmin/DAM/stats/publications/Register_based_statistics_in_Nordic_countries.pdf
[6] http://www.stats.govt.nz/browse_for_stats/snapshots-of-nz/integrated-data-infrastructure.aspx
[7] http://www.nss.gov.au/nss/home.nsf/pages/Open%20Data%20Initiatives

TEMPLATE

UNIT 5

INTRODUCTION TO SOFTWARE ARGUS

ARGUS packages have been developed to assist disclosure control and to produce safe microdata or tables. These packages consist of μ-ARGUS for disclosure control in microdata and τ–ARGUS to process the tables. The recent version of these software packages has been developed by the Statistics Netherlands due to a continuous European cooperation effort in the CASC[8] project. The packages together with the user's manuals can be downloaded from http://neon.vb.cbs.nl/casc. Since a complete and detailed guide to use the software can be found easily in the user's manuals, this unit will only provide basic information required to start working with the two software packages as the basis for further self-studies.

5.1. Disclosure control in microdata with μ-ARGUS

Learning Objectives

– To become familiar with data management, and SDC methods for microdata in μ-ARGUS

A complete tour on using μ-ARGUS can be found in its user's manual. This section aims to provide a brief explanation on this software, in particular, for preparing data, metadata, key variables and variable combinations.

Step 1. Getting microdata

To start disclosure control in μ-ARGUS, one should first read the target microdata file from "open microdata" tab under toolbar or file menu. Only ASCII files can be read as dataset, and the file must contain records, each in one row and variables in columns. μ-ARGUS can read files with "fixed format", "free format with tab or character separated (for example comma)", and "free format with variable name in the first line".

Step 2. Specify Metadata

After getting data into μ-ARGUS, it is required to define metadata. Metadata is the structure of the data set including the additional SDC-specific information. Metadata can be loaded along with data file in open microdata window through "Specify metadata" window as illustrated in figure 5.1. If the metadata file is an ASCII file, this window will be automatically completed and

[8] Computational Aspects of Statistical Confidentiality

TEMPLATE

properties and attributes of the microdata will be shown in the relevant tabs. Otherwise, the information should be manually inserted in the window fields.

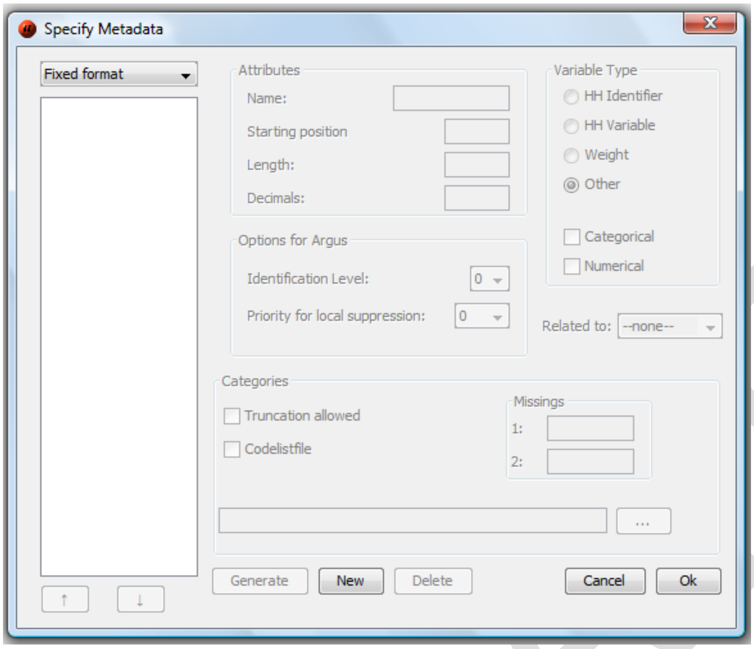

Figure 5.1. "Specify Metadata" window in µ-ARGUS

The elements of "Specify Metadata" window are as follows:

- In the left pane of the window, the format of the data file has to be defined. Then available variables of the data set should be set one by one by clicking the "New Tab" in the bottom of the window.
- "Attributes" pane contains the starting position, length and decimals for each record of the variables in data file.
- Under the "Variable Type" pane, the characteristics of variables should be specified:
  - HH Identifier: The variable contains unique identifier of households.
  - HH Variable: The variable that by nature has the same value for all members in a single household.
  - Weight: The variable contains sampling weights.
  - Categorical: The variable's type is categorical variable, and it can be used as a spanning variable in a table. For µ-ARGUS this variable can be defined as identifying.
  - Numerical: The variable's type is numerical variable, and it can be used for top/bottom coding, micro-aggregation and rounding.
- The identification level is an option to determine the extent to which the variable is identifying. By this, one can easily generate the set of tables to be inspected in the disclosure

TEMPLATE

control process.

– 0: variable is non-identifying
– 1: variable is most identifying.
– 2: the variable is more identifying.
– 3: the variable is identifying.

- The priority for local suppression: When at the end of a µ-ARGUS run the remaining unsafe combinations are suppressed by local suppression, one of the options is to use the user-supplied priority weights to select the variables that will be suppressed. These priority weights are specified in this tab.
- The codelists of the variables used to span the tables is always generated by µ-ARGUS itself. However, the user can supply the name of a codelist file. The labels in this codelist file are then used when displaying information on this variable.
- One or two missing values can be specified per variable. Missing values play a specific role in the SDC-process, as missing values will be imputed when local suppression is applied. Note that the weight variable cannot have a missing value, but categorical variables need to have at least one missing value.

Some predefined microdata with metadata can be found in data folder in µ-ARGUS Package. As an example, demodata.asc is the sample data that can be loaded to see how to specify metadata.

Step 3. Specify Combinations

When the metadata is ready, the set of combinations to be inspected by µ-ARGUS can be specified. These tables can be specified either manually or by using one of the two rules to generate this set. To select the tables manually, one should select the variables in the left pane and move them to the middle pane (with the "→" button). If the table is ready and the appropriate threshold has been specified, the table can be added to the set of tables in the right pane with the "→" button. Each table can have a different threshold.

- In the "automatic specification of tables", there are two options (a) using the identification level and (b) all tables up to a given dimension. When the set of tables is constructed using the generator, it is still possible to make adjustments. Certain tables can be added or omitted.
- If the risk model is used, a table for this model can be selected by pressing the button "Set table for risk model". A restriction is that there cannot be an overlap between the tables used for the classical threshold method and the new risk-model. Mixing the basic model and the new risk approach makes no sense. Therefore, the overlapping tables will be removed automatically.

TEMPLATE

Figure 5.2 shows the example of “Specify Combinations” window for demodata microdata. In this window, combinations of key variables are specified automatically using identification level and threshold ‘1’.

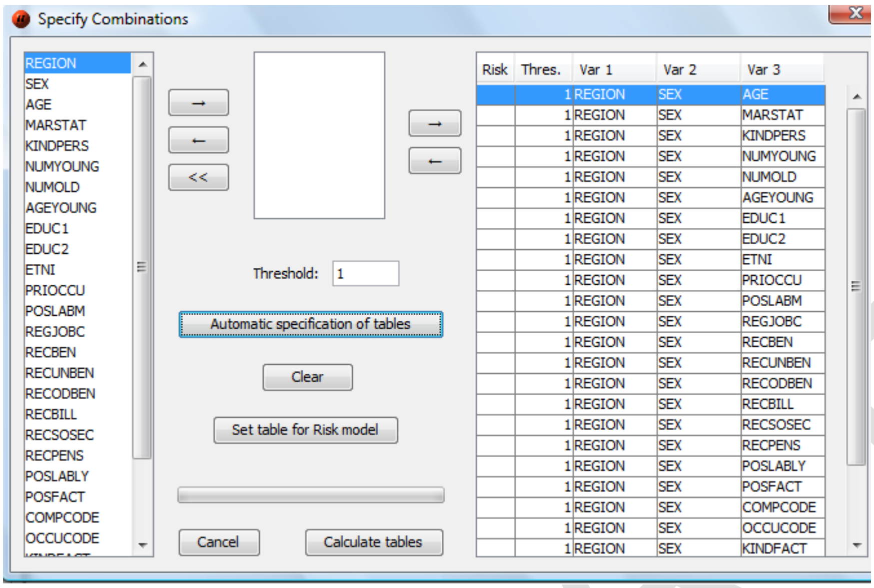

Figure 5.2. “Specify Combinations” Window (for demodata microdata)

When desirable combinations and threshold specified, one should click on “Calculate Tables” tab. After the progress is completed, number of unsafe combinations will be shown in the main window of μ-ARGUS.

Step 4. Selection and application of protection methods

After specification of combinations and calculating tables, disclosure control starts. In the *Modify* , disclosure control methods such as Global Recode, PRAM specification and Numerical Rank Swapping are available. If risk model is selected in specification of combinations, Individual or Household Risk Specification section are also activated.

Step 5. Output

After all disclosure control processing, the output protected files should be produced. These files can be produced via output menu in μ-ARGUS.

5.2. Disclosure control in tables with τ-ARGUS

Learning Objectives

– To become familiar with data management, and SDC methods for tables in τ-ARGUS

TEMPLATE

τ–ARGUS software is designed to produce safe tables. As for the μ-ARGUS, there is a user's manual describing how to work with the τ–ARGUS, but this sections just attempts to give a brief introduction on how to start working with this software.

Step 1. Getting data

The software can read microdata as well as tables. Reading microdata by the software is very similar to μ-ARGUS except that here SPSS files can also be uploaded.

Step 2. Specify metadata

Metadata is required for both tables and microdata. Figure 5.3 shows the window for specification of metadata for tau_testW.asc which can be found in data folder of downloaded software.

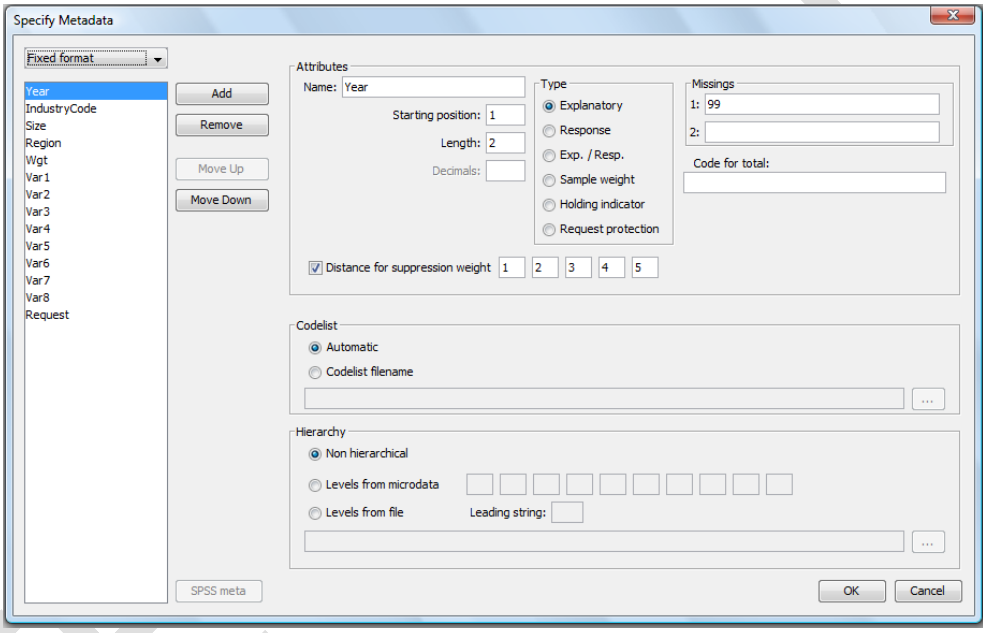

Figure 5.3. "Specify Metadata" window in τ–ARGUS

The key elements of this window are the definitions for each variable. Most variables will be defined as one of the following:

- Explanatory Variable: The variable that is used as a categorical (spanning) variable when defining a table.
- Response Variable: The numerical variable that is used as a cell item in a table.
- Weight variable: The variable contains sampling weighting scheme.

Other important features of this window are as follows:

- Codelist: τ-ARGUS will always automatically build the codelists for the explanatory variables from the datafile.

- Missing values: Gives information on the missing values that are attached to codelist. Two distinct missing value indicators can be set.
- Hierarchical codes: The hierarchy can be derived from the digits of the individual codes in the data file, or from a specified file containing the hierarchical structure.

Step 3. Specify Tables

When the metadata file is ready, the tables to be protected can be specified. This is achieved via “Specify Tables” window. Figure 5.4 shows the window of specification of tables.

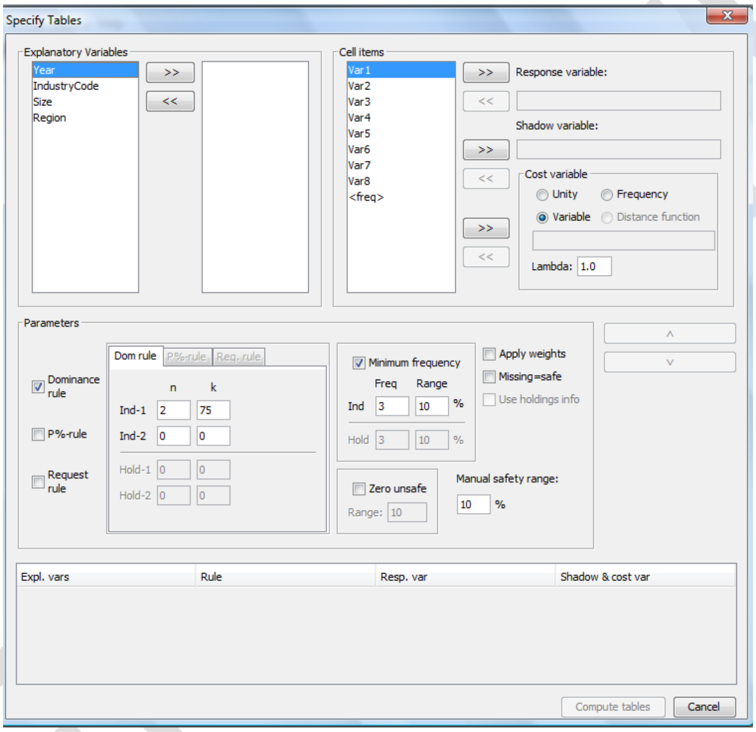

Figure 5.4. “Specify Tables” window in τ–ARGUS

The key elements of this window are as follows.

- Explanatory variables: On the left is the listbox with the explanatory variables. Clicking on “>>” moves the selected variables to the next box in which the selected explanatory variables can be seen. From the box on the left containing explanatory variables, the variables that will be used as the row/column of a 2-way table, can be selected. Up to six explanatory variables can be selected to create a table, but higher dimensions will restrict the options to process a table.
- Cell items: The “cell items” box contains the variables, which were declared as “response variables” in the metafile. By using the “>>” button they can be moved to the “response variable” box to be used in the defined table.

- Response variable: Any variable in the cell items box can be chosen as the response variable. In addition, the implicit variable, <freq>, can be used for making a frequency table.
- Shadow variable: The shadow variable is the variable that is used to apply the safety rule. By default, this is the response variable.
- Cost variable: This variable describes the cost of each cell. These are the costs that are minimized when the pattern of secondary suppressed cells is calculated. By default, this is the response variable but other choices are also possible.
- Weight: If the data file has a sample weight, specified in the metadata file, the table can be computed, considering this weight.
- The safety rule: In this window, the left side of the window allows the type of rule to be selected. This is usually either the dominance rule or *p-Percent* rule, along with the necessary parameter values.
- Creating the Table: After setting all required information, clicking on  sends table specifications to the table box and pressing the “Compute tables” button will invoke τ-ARGUS to actually compute the requested tables.

Step 4. Disclosure control processes

When the table (tables) has been generated, the main-window of τ-ARGUS will show the (first) table. Figure 5.5 illustrates an example of the window which contains result of computing table for sample data when explanatory variables are region and size, using *p-Percent* rule with $p = 10$, $q = 100$ and $N = 1$.

By clicking on any cell of the tables, value of the cell, its status (safe or unsafe) and some other specification of that cell will be displayed on the left side of the window. Also in this window, changing the status of any cells, recoding or suppressing the cell value can be done, as well.

Step 5. Save the safe table

When the table is safe, then it may be saved to the hard disk of the computer. The user has six options for output file format such as csv file, text file and csv for a pivot table.

TEMPLATE

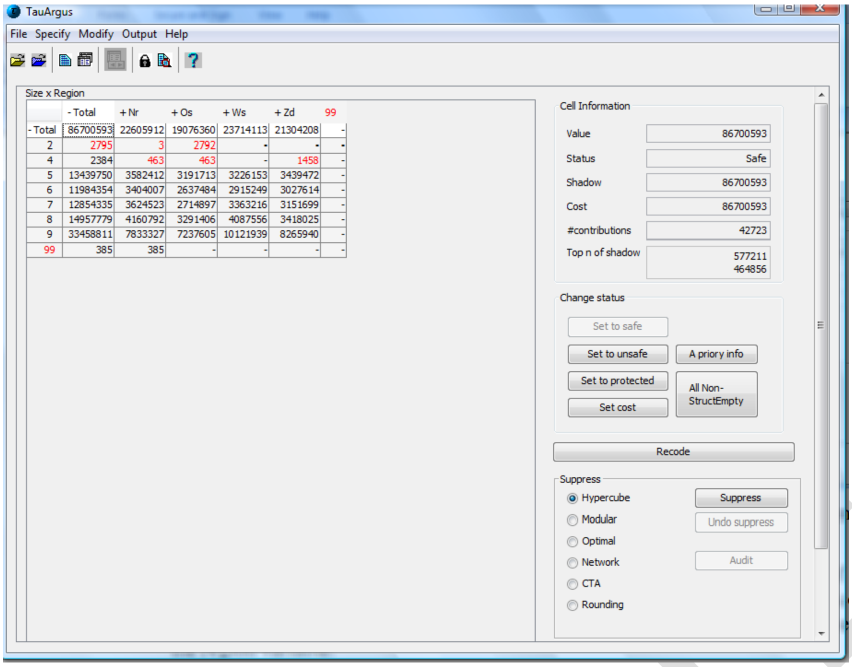

Figure 5.5- Illustration of the result of compute tables for tau_testW data

TEMPLATE

GLOSSARY[9]

Argus
Two software packages for Statistical Disclosure Control are called Argus. μ-Argus is a specialized software tool for the protection of microdata and τ-Argus is a specialized for tabular data.

Blurring
Synonym to Micro-aggregation

Confidentiality
Confidentiality is a quality or condition accorded to information as an obligation not to transmit that information to any unauthorized party.

Controlled rounding
In this method, linear programming methods are used to round the risky cell values in a way that they would add up to the published marginal totals.

Controlled tabular adjustment
In this method, in each risky cell, the original value is replaced by a safe value that is in a sufficient distance away from the original value; then marginal values are minimally adjusted to ensure additivity.

Data laboratory
Data laboratory is a physically secure environment that permits use of confidential files.

Disclosure
Disclosure occurs when from a released data, in the form of a table or microdata, the identity of an individual (including a person, a household, a business enterprise, etc.) is recognized or sensitive information about an individual is revealed.

Disclosure risk
Disclosure risk occurs if an unacceptably narrow estimation of a respondent's confidential information is possible.

Dominance rule
In a magnitude table, a cell is declared as risky if a small number (n) of individuals contribute a large percentage (k) to the total cell value.

File-level risk metrics
File-Level Risk Metrics measure the average risk across the whole microdata file

Frequency table
Frequency table is a table where each cell's value represents the number of individuals that fall into that cell.

[9] "The Glossary on Statistical Disclosure Control" is available at http://neon.vb.cbs.nl/casc/glossary.htm.

TEMPLATE

Global recoding
Global recoding is a non-perturbative where several values or categories are combined to form new less-specific categories.

Household variable
Household variable is a variable whose value is the same for all members of the same household.

Identifier
Synonym to identifying variable

Identifying variable
Identifying variable is a variable that can be used, alone or in combination with other variables, to re-identify an individual.

Intruder
Intruder refers to an individual, group, or organization who attempts to identify an individual (including a person, a household, a business enterprise, etc.) within a data set and/or to discover sensitive information about a given individual, usually through a statistical linkage process.

Key variable
Synonym to identifying variable

Local suppression method
Local suppression is a non-perturbative method, in which certain values of individual variables are suppressed.

MAASC
MASSC is a perturbative masking method whose acronym summarizes its four steps: (a) Micro Agglomeration, (b) Substitution, (c) Sub-sampling, and (d) Calibration.

Macrodata
Synonym to tabular data

Magnitude table
Magnitude table is a table where each cell's value represents a summary statistics of values for a particular characteristic across all individuals that belong to that cell.

Micro-aggregation method
Micro-aggregation is aggregating across small groups of individuals and replacing one individual's original value with the group average.

Microdata
Microdata consists of records with information about each single individual (including a person, a household, a business enterprise, etc.).

Noise addition method
Noise addition is a perturbative method that a random noise is added to (sometimes multiplied by) the data.

Non-Perturbative methods
Non-perturbative methods reduce the amount of information released.

TEMPLATE

(*n*, *k*)-rule
Synonym to dominance rule

***n*-rule**
Synonym to threshold rule

Perturbative methods
Perturbative methods falsify the data before publication by introducing an element of error deliberately for confidentiality reasons.

Population unique
When an individual has unique values on a set of identifiers within a given population, then that individual is said to be a population unique

***p*-Percent rule**
In a magnitude table, a cell is identified as risky if any contributing individual value to that cell can be estimated by other individuals with accuracy of more than p-Percent of its actual value for a pre-specified value of p.

PRAM
PRAM (Post-Randomization Method) is a probabilistic, perturbative method for categorical variables, that in the masked dataset, the values of some categorical variables for certain records in the original file are changed to a different value according to a prescribed probability mechanism, namely a Markov matrix.

Pre-tabular methods
Pre-tabular methods are perturbative methods that can be applied to the table's underlying microdata prior to tabulation in order to assure that any table generated from that data is fully protected.

Prior/Posterior ambiguity rule
Assuming that a coalition group may have a prior knowledge (*q*) about an individual value of the desired variable, and gain more knowledge (*p*) after release of the table; with this rule (which is suitable for magnitude table) a cell is identified as risky if the information gain is unacceptable.

Privacy
Privacy is an individual's freedom from excessive intrusion in the quest for information and an individual's ability to choose the extent and circumstances under which his or her beliefs, behavior, opinions, and attitudes will be shared with or withheld from others.

***q/p* rule**
Synonym to prior/posterior ambiguity rule

Random rounding
In this method, all table cell values are randomly rounded up or down to the nearest multiple of a base that is equal to the specified threshold.

Rareness
Rareness is the characteristic of a variable that has a value that may occur rarely in the population.

TEMPLATE

Record-level risk metrics
Record-Level Risk Metrics measure the disclosure risk for the records.

Remote access facilities
Synonym to virtual access

Rounding
Rounding is a perturbative method that replaces the original values with rounded values,

Sampling method
Sampling is a non-perturbative method, in which instead of releasing the original microdata file, a sample of the original set of records is published.

Sensitive variable
Sensitive variable is a variable whose values can represent characteristics that an individual may not like to be revealed.

Special uniqueness
Special uniqueness means that a record that is sample unique on a set of identifiers is also unique on a subset of those identifiers.

Statistical Disclosure Control
Statistical Disclosure Control is a body of principles, concepts, and procedures that permit confidentiality to be afforded to data, while still permitting its use for statistical purposes

Suppression
In this method, risky information is protected through hiding (suppressing) the values of few cells by replacing them by a specified symbol.

Swapping
Swapping refers to transforming a dataset by exchanging values of variables among records.

Tabular data
Tabular data is aggregate information on individuals presented in tables

Threshold rule
Threshold rule is a common rule for frequency tables, which defines a cell as risky if the frequency is less than a specified number.

Top and bottom coding method
Top and bottom coding, as a non-perturbative method, is a special case of global recoding, where the top values (those above a certain threshold) are lumped together to form a new category. The same is done for the bottom values where those below a certain threshold are combined together.

Traceability
Traceability is the characteristic of a variable that based on its values an individual can be traced and located easily.

Utility
Utility means the value of a given data release as an analytical resource. This comprises the data’s analytical completeness and its analytical validity

TEMPLATE

Virtual access
Virtual access is a virtual secure environment that permits remote use of confidential data.

Visibility
Visibility is the characteristic of a variable that has a value that is known or can be ascertained easily.

TEMPLATE

READING LIST

Duncan, G.T., Elliot, M., Salzar-González, J.J. (2011). *Statistical Confidentiality: Principles and Practice*, Springer, New York.

Federal Committee on Statistical Methodology (2005). Statistical policy working paper 22: report on statistical disclosure limitation methodology, 2nd ed., Office of Management and Budget, Washington, DC.

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Lenz, R., Naylor, J. Nordholt, E.S., Seri, G. and De Wolf, P.P. (2010). *Handbook on Statistical Disclosure Control*. ESSNet SDC.

United Nations Economic Commission for Europe Conference of European Statisticians (2007). *Managing Statistical Confidentiality & Microdata Access: Principles and Guidelines of Good Practices*, Sales No. E.07.II.E.7.

TEMPLATE

REFERENCES

Duncan, G.T., Jabine, T.B., de Wolf, V.A. (1993). *Panel on Confidentiality and Data Access, Committee on National Statistics, Commission on Behavioral and Social Sciences and Education, National Research Council and the Social Science Research Council, Private Lives and Public Policies: Confidentiality and Accessibility of Government Statistics*, National Academy of Sciences. Washington, DC.

Duncan, G.T., Elliot, M., Salzar-González, J.J. (2011). *Statistical Confidentiality: Principles and Practice*, Springer, New York.

Duncan, G.T., Pearson, R. W. (1991). Enhancing access to microdata while protecting confidentiality: prospects for the future. *Statistical Science*, **6**, 219-239.

Elliot, M. J. (2000). A new approach to the measurement of statistical disclosure risk. *Risk Manage. Int. J.* **2** (4), 39-48.

Elliot, M. J., Manning, A., Mayes, K., Gurd, J., Bane, M. (2005). SUDA: A program for detecting special uniques. *Proceedings of the UNECE/Eurostat Work Session on Statistical Data Confidentiality*. Geneva.

Federal Committee on Statistical Methodology (2005). Statistical policy working paper 22: report on statistical disclosure limitation methodology, 2nd ed., Office of Management and Budget, Washington, DC.

Fienberg, S.E. (2005). Confidentiality and disclosure limitation. In *Encyclopedia of Social Measurement*, K. Kempf-Leonard, eds. Elsevier, New York, pp. 463–469.

Groves, R.M., Fowler, F.J. Jr., Couper, M.P. Lepkowski, J.M., Singer, E. and Tourangeau, R. (2009). *Survey Methodology*, 2nd ed. Wiley, New York.

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Lenz, R., Naylor, J. Nordholt, E.S., Seri, G. and De Wolf, P.P. (2010). *Handbook on Statistical Disclosure Control*. ESSNet SDC.

Skinner, C. J., Elliot, M. J. (2002). A measure of disclosure risk for microdata, *J. R. Stat. Soc. Ser.* B **64** (4), 855-867.

Smith, D., Elliot, M. J. (2008). A measure of disclosure risk for tables of counts. *Trans. Data Priv.* **1** (1), 34-52.

Willenborg, L., De Waal, T. (1996). *Statistical Disclosure Control in Practice*. Springer, New York.

United Nations Economic Commission for Europe Conference of European Statisticians (2007). *Managing Statistical Confidentiality & Microdata Access: Principles and Guidelines of Good Practices*, Sales No. E.07.II.E.7.

Statistics Netherlands (2014). μ-ARGUS Version 5.1 User's Manual. Available on http://neon.vb.cbs.nl/casc/Software/MUmanual5.1.pdf.

Statistics Netherlands (2014). τ-ARGUS Version 4.1 User's Manual. Available on http://neon.vb.cbs.nl/casc/Software/TauManualV4.1.pdf.