



Introduction to Statistical Disclosure Control

University of Pisa - EUROSTAT

EMOS Learning Materials
Service Contract n. 2019.0249
between Eurostat and the
University of Pisa, Italy

Basic concepts

- ▶ Confidentiality and Data protection are fundamental principles of Official Statistics.
- ▶ National Statistical Institutes (NSIs) publish trusted and high quality statistical output which must be as detailed as possible.
- ▶ At the same time NSIs are obliged to protect the confidentiality of the information provided by the respondents.
- ▶ **Statistical Disclosure Control (SDC)** seeks to protect statistical data in such a way that they can be released without disclosing information on individual entities

Basic concepts

- ▶ SDC is a complex problem that intersects very different subjects.
- ▶ Consequently it requires the interaction between specific skills of the various sectors involved (administrative, legal, IT).
- ▶ A data dissemination strategy offers many different statistical outputs covering a range of different topics for many types of output:
 1. Tabular Data (which is just one example of aggregate output; other examples are graphs and parameters)
 2. Microdata Data
 3. Online Dynamic Databases

Basic concepts

- ▶ Different outputs require different approaches to SDC and different mixture of tools. As a consequence it is important to know:
 1. type of data (*full population or sample survey*),
 2. sample design,
 3. an assessment of quality of data (*the level of non-response and coverage of the data*),
 4. variables and whether they are categorical or continuous,
 5. type of outputs (*microdata, magnitude or frequency tables*).
- ▶ The needs of users is a fundamental concept.
- ▶ The actions carried out on data to protect against disclosure attacks, lead to a loss of information content in the data.

Definitions

- ▶ **Confidential / Sensitive variable:** a variable which is present in the released data but is unknown to an intruder before the release of the data. This variable has values that a respondent would not wish to be revealed.
- ▶ **Disclosure:** A disclosure occurs when a person or organisation recognises something that they did not know already about another person or organisation, via released data.
- ▶ **Disclosure Risk:** the risk that a given form of disclosure arise when a given data product is released. Two types of risk:
 1. **Identity disclosure:** it occurs with the association of a respondent's identity with a disseminated data record or cell containing confidential information (Duncan et al. (2001)),
 2. **Attribute disclosure:** it occurs with the association of either an attribute value in the disseminated data or an estimated attribute value based on the disseminated data with the respondent (Duncan et al. (2001)).

Definitions

- **Statistical Disclosure Control:** SDC techniques can be defined as the set of methods to reduce the risk of disclosing information on individuals, businesses or other organisations. SDC methods minimise the risk of disclosure to an acceptable level while releasing as much information as possible.

There are two types of SDC methods:

1. **Perturbative methods:** they change the data before publication by introducing an element of error purposely for confidentiality reasons,
2. **Non-Perturbative methods:** they reduce the amount of information released by suppression or aggregation of data.

Definitions

- ▶ **Statistical Integrity:** data maintains its statistical integrity if the validity of statistical analyses of the data is not systematically biased by the SDC technique.
- ▶ **Risk and Utility:** data producers (such as NSIs) should aim to determine optimal SDC methods and solutions that minimise disclosure risk while maximizing the utility of the data.

Definitions

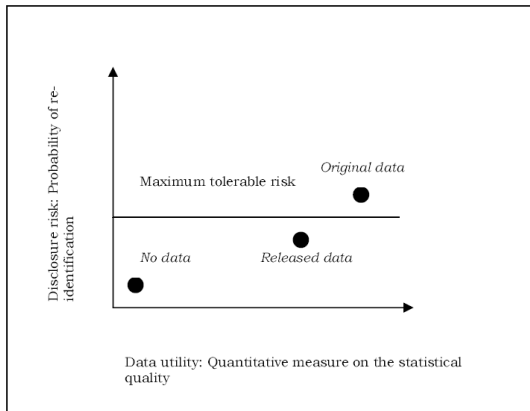


Figure 1: Risk and Utility Plot (Duncan et al.(2001))

MICRODATA

Types of outputs: Microdata

- ▶ Microdata consist of a series of records, each containing information on an individual unit such as a person, a firm, an institution.
- ▶ Microdata in their simplest form may be represented as a single data matrix, where the rows correspond to the units and the columns to the variables.
- ▶ Microdata can be classified as:
 1. **Simple Microdata**
 2. **Complex Microdata**

Roadmap to releasing a microdata file

Stage of disclosure process	Analyses to be carried out / problem to be addressed ↓ <i>Results expected</i>
1. Why is confidentiality protection needed	Does the data refer to individuals or legal entity? ↓ <i>We need to protect the statistical unit</i>
2. What are the key characteristics and use of the data	Analysis of the type/structure of the data ↓ <i>Clear vision of which units need protections</i>
	Analysis of survey methodology ↓ <i>Type of sampling frame, sample/complete enumeration of strata, further analysis of survey methodology, calibration</i>
	Analysis of NSI objectives ↓ <i>Type of release (PUF, MFR), dissemination policies, peculiarities of the phenomenon, coherence between multiple releases (PUF and MFR), coherence with released tables and on-line databases, etc.</i>
	Analysis of user needs ↓ <i>Priorities for variables, type of analysis, etc.</i>
	Analysis of the questionnaire ↓ <i>List of variables to be removed, variables to be included, some ideas of level of details of structural variables</i>
3. Disclosure risk	Disclosure scenario ↓ <i>List of identifying variables</i>
	Definition of risk ↓ Risk assessment
	↓ <i>If the risk is deemed too high need of disclosure limitation methods</i>
4. Disclosure limitation methods	Analysis of type of data involved, NSI policies and users needs ↓ <i>Identification of a disclosure limitation method</i>
	Information loss analysis
5. Implementation	Choice of software, parameters and thresholds for different methods

Figure 2: Roadmap to releasing a microdata file (Eurostat, 2018)

Disclosure Risks for Microdata

- ▶ Suppose the microdata to be released and consist of a standard rectangular data matrix

$$\begin{bmatrix} x_{11} & \cdots & x_{1m} \\ \vdots & \cdots & \vdots \\ x_{nm} & \cdots & x_{nm} \end{bmatrix}$$

- ▶ Suppose that the file has already been anonymized by removing direct identifiers, but geographical variables may remain.
- ▶ The file may also include variables related to the survey design.
- ▶ The rows of the matrix correspond to n sample units and will be referred to as **records**.
- ▶ The columns correspond to m variables so that x_{ij} is the value of the j^{th} variable for the i^{th} unit.

Disclosure Risks for Microdata

- ▶ The variables in a microdata file can be classified into four groups
 1. **Identifiers:** variables which identify the respondent without degree of ambiguity, *i.e. passport number, social security number*;
 2. **Key variables:** variables which identify the respondent with some degree of ambiguity, *i.e. gender, age* (microdata and prior information);
 3. **Confidential variables:** variables which contain sensitive information of the respondent, *i.e. salary, religion, political affiliation* (microdata);
 4. **Non-confidential outcome variable:** all the other variables. of disclosure risk)
- ▶ **Target Units:** units of which information is to be disclosed.
- ▶ An **identity** of a unit is a label which is publicly recognizable and unique in the population. A requirement of a target unit is that it be identifiable, that is that it has an associated identity.

Predictive Disclosure

- ▶ **Predictive Disclosure** occurs if an intruder is able to use the microdata to gain information about the target unit.
- ▶ Microdata are released only after taking out directly identifying variables but other variables in the microdata can be used as indirect identifying variables.
- ▶ If the identifying variables are categorical then the cross classification of these variables defines a key variable.
- ▶ The disclosure risk is a function of these identifying variables in the sample alone or in the sample and in the population.

Predictive Disclosure

- ▶ The **Disclosure Risk Scenarios** are the scenarios of possible information known to the intruder.
- ▶ The **Identifying variables** are determined on the basis of the disclosure risk scenarios.
- ▶ The other variables in the data are **Confidential variables** and represent the data not to be disclosed.

Predictive Disclosure

- ▶ For **microdata containing censuses** or registers **the disclosure risk is known** as we have all identifying variables available for the whole population.
- ▶ For **microdata containing samples** the population is unknown or partially known through marginal distribution. Therefore, **probabilistic modelling are used to estimate disclosure risk at population level based on information available at the sample.**

Predictive Disclosure

- ▶ As previously said, if the identifying variables are categorical the risk is cast in term of the cells of the contingency table built by cross-tabulating the identifying variables, **the keys**.
- ▶ All the records in the same cell have the same value of risk.
- ▶ Disclosure risk measures can be classified in:
 - ▶ **Risk based on the keys in the sample** (categorical identifying variable)
 - ▶ **Risk based on the keys in the population** (categorical identifying variable)
 - ▶ **Risk based on Record Linkage** (categorical and continuous identifying variable)

Predictive Disclosure

- ▶ **Risk based on the keys in the sample:** if the combination of indirect identifiers leads to small number of records (below threshold).
- ▶ **Risk based on the keys in the population:** the risk of a unit is determined by its combination of scores on the identifying variables within the population or its probability of re-identification. A unit is at risk if such quantity is above a given threshold. This quantity is unknown in the population as a consequence it may be estimated through a modelling process.
- ▶ **Risk based on Record Linkage:** if the variables are continuous, the keys are no longer applicable.

Disclosure Risk Scenarios (DRS)

- ▶ A scenario describes:
 1. which is the information potentially available to the intruder (**External Archive (EA)**)
 2. how the intruder would use such information to identify an individual
- ▶ The DRS is based on the assumption that the EA available to the intruder is an individual microdata archive: for each individual directly identifying variables and other variables are available. Some of these other variables are also in the microdata files that needs to be protected.
- ▶ **The intruder could use this overlapping information to match identifier to a record in the microdata files.**
- ▶ In this scenario **the matching variables are the identifying variables.**

Re-Identification

- ▶ **Spontaneous recognition:** the intruder might rely on personal knowledge about once or a few target of individuals and spontaneously recognise a sampled individual. This scenario is called **Nosy Neighbour scenario**.
- ▶ **Re-identification via record linkage:** the intruder has access to a public register and he tries to match the information provided by this EA with the microdata file in order to identify surveyed units. The intruder's chance of identifying a unit depends of EA main characteristics (e.g. completeness, accuracy, data classification)

Microdata Protection methods

- ▶ Assume that the identifiers and key variables have been removed in the original microdata file.
- ▶ If V is the original microdata file, the goal of the SDC methods is to release a protected microdata file V' such that:
 1. Disclosure Risk is low, i.e. the risk that a user or an intruder can use V' to determine confidential variables on a specific individual among those in V is low;
 2. User analyses on V' and on V yield the same or at least similar results.

Microdata Protection methods

- ▶ Microdata protection methods can generate the protected microdata V'
 1. by **masking original data**, that is generating V' a modified version of the original microdata V ;
 2. by **generating synthetic data V'** that preserve some statistical properties of the original data V .
- ▶ Masking original data involves use of **Masking methods**:
 1. **Perturbative Masking**;
 2. **Non Perturbative Masking**.

Microdata Protection methods

- ▶ **Perturbative Masking:** The microdata file is perturbed before the publication. Original unique combinations of score do not appear in the released file, which contains a modified version of them.
- ▶ **Non Perturbative Masking:** Data is not masked. These methods produce a partial suppression or reduction of details in the original microdata file.

Perturbative Masking

- ▶ Special case of Perturbative masking are:
 - ▶ Noise addition,
 - ▶ Rank swapping,
 - ▶ Microaggregation,
 - ▶ Post-Randomization.
- ▶ Suppose V is the original dataset. The masked microdata set Z is computed as:

$$Z = AXB + C$$

A is a record transforming mask

B is a variable transforming mask

C is a displacing mask or noise.

Perturbative Methods and Type of Data

<i>Method</i>	<i>Continuous Data</i>	<i>Categorical Data</i>
<i>Noise addition</i>	YES	NO
<i>Microaggregation</i>	YES	YES
<i>Rank swapping</i>	YES	YES
<i>Rounding</i>	YES	NO
<i>Resampling</i>	YES	NO
<i>PRAM</i>	NO	YES
<i>MASSC</i>	NO	YES

Table 1: Perturbative methods by Data Type.

TABULAR DATA

Types of Tabular Data: Single Tables

- ▶ Microdata are the basis for tabular data which could be obtained from microdata by a process called **aggregation**.
- ▶ For the aggregation process it is necessary to define the **spanning variables** and possibly a **response variable**.
- ▶ The spanning variables define a **cross-classification**.
- ▶ Each combination of values in a cross-classification defines a **cell** in the table.
- ▶ The number of spanning variables used in the definition of a table is called the **dimension of the table**.

Types of Tabular Data: Single Tables

Notation:

- ▶ y_i value of response variable Y for record i
- ▶ w_i weight for record i
- ▶ $t_c = \sum_{i \in C} w_i y_i$ cell total of cell C

Classification:

- ▶ **Frequency count table:**
If $w_i = 1$ and $y_i = 1$ for each record i ,
- ▶ **Weighted frequency count table:**
If $w_i \neq 1$ for at least one record i and $y_i = 1$ for each record i ,
- ▶ **Magnitude table:**
If $w_i = 1$ for each record i and $y_i \neq 1$ for at least one record i ,
- ▶ **Weighted magnitude table:**
If $w_i \neq 1$ for at least one record i and $y_i \neq 1$ for at least one record i .

Types of Tabular Data: Marginal Tables

- ▶ In practice, a single table is hardly released on its own. Along with it another set of tables is often published, consisting of one or more of its **marginal tables**.
- ▶ If T is a table then a **marginal table** of T is a table that can be obtained from T by aggregating over one or more spanning variables of T .
- ▶ Marginal tables introduce dependencies in the data, because they imply linear constraints.
- ▶ These dependencies are responsible for major complications in the disclosure protection of tabular data.

Disclosure Risk for Tabular Data

- ▶ Disclosure risk may be defined either for the whole table or separately for each cell into which the table is organized.
- ▶ A threshold may be specified as the maximum value below which the disclosure risk is deemed acceptable.
- ▶ If the disclosure risk exceeds the threshold it is necessary to use some form of SDC technique and the table (or the cell) is called sensitive. In the next slides cell sensitiveness is considered.
- ▶ The objective of disclosure risk assessment will then be to determine which cells of a table are sensitive: a table containing sensitive cells may not be published.

Disclosure Risk for Magnitude Tables

- Example: Consider the following table:

	legal advice offices	notary's offices	patent bureaus	Total
2 to 4 employees businesses	45	90	1	136
revenue (mln Euros)	8	38	×	×
5 to 9 employees businesses	10	429	3	442
revenue (mln Euros)	6	353	×	×
10 and more employees businesses	7	223	10	240
revenue (mln Euros)	14	393	82	489

Figure 3: Businesses with main activity legal services (Willendorg & De Waal, 2001)

Disclosure Risk for Magnitude Tables

- ▶ Figure 3 reports an example of **Magnitude Table**.
- ▶ The table consists of 9 internal cells together with 3 marginal totals.
- ▶ Each cell X contains two entries:
 1. $N(X)$ the number of businesses,
 2. $T(X) = \sum_{i=1}^{N(X)} x_i$, the total sum of the magnitudes x_i for all businesses i falling into that cella X .
 3. For example, the cell X consisting of all legal advice offices with 2 to 4 employees consists of $N(X) = 45$ businesses for which the total revenue is 8 million Euros.
- ▶ Cells denoted by \times indicate that their values have been suppressed.

Disclosure Risk for Magnitude Tables

- ▶ Refer to the units i , upon which tables of magnitude data are based, as businesses.
- ▶ The businesses are ordered as:

$$x_1 \geq x_2 \geq \dots \geq x_{N(X)} > 0$$

- ▶ Suppose $N(X)$ businesses cover the whole population in the cell.
- ▶ There is an intruder who want to use published table to disclose information about the individual businesses who contribute data to the table.
- ▶ Suppose that the identities of the $N(X)$ businesses in each cell X are known to the intruder, although their x_i values will generally not be known prior to data release.

Disclosure Risk for Magnitude Tables

- ▶ In particular, assume that the variables classifying the cells of the table, for example number of employees, are potentially knowable to the intruder before release of the table. For example, assume that the 3 patent bureaus with 5-9 employees are known to the intruder and suppose that two of these bureaus work together as a coalition to disclose information about the third.
- ▶ The total revenue of this category is not published because the coalition of the two patent bureaus would be able to calculate the revenue of the third bureau in this category exactly.
- ▶ → Although we are concerned about the potential disclosure of the values x_i , of individual businesses i , the assessment of the disclosure risk of a table cannot be made at any level finer than the cells of the table.

Disclosure Risk for Magnitude Tables

- ▶ The **disclosure risk** of a cell X , or the *sensitivity* of X , is denoted by $S(X)$. We shall say that the cell is *sensitive* (or **disclosive**) if

$$S(X) > d, \text{ for a given threshold value } d$$

- ▶ The sensitivity measure $S(X)$ should summarize the risk of disclosure of each of the values x_i , across all businesses i falling into cell X .
- ▶ The degree of disclosure about a given value x_i may be measured by the **accuracy** p with which x_i , could be inferred by the intruder.

Disclosure Risk for Magnitude Tables

Sensitivity measures $S(X)$

- ▶ The different definitions of sensitivity provide a basis for SDC techniques.
- ▶ If a cell is sensitive then its value should not be released.
- ▶ It should be suppressed or modified by a SDC technique in some way.

Disclosure Risk for Magnitude Tables

Sensitivity measures $S(X)$

► **Linear Sensitivity Measures:**

$$S(X) = \sum_{i=1}^{N(X)} \alpha_i x_i.$$

If the weight α_j and x_j are non-increasing, then $S(X)$ is **sub-additive**, i.e. the union of two non-sensitive cells is always non-sensitive.

Disclosure Risk for Magnitude Tables

Sensitivity measures $S(X)$

► (n, k) **Dominance Rule:**

$$S_n(X) = \frac{\sum_{i=1}^n x_i}{\sum_{i=1}^{N(X)} x_i}.$$

The parameter k is the threshold value; the parameter n is usually chosen one larger than the maximum size of (imagined) coalitions of respondents. The choice of both n and k depends on the desired level of protection.

Disclosure Risk for Magnitude Tables

Sensitivity measures $S(X)$

► Prior-posterior Rule:

$$S_q(X) = -\frac{q}{x_1} \sum_{i=3}^{N(X)} x_i$$

p is threshold value. It is assumed that, prior to the publication of the table, every respondent can estimate the contribution of each other respondent to within q percent. A cell is considered sensitive if someone can estimate the contribution of an individual respondent to that cell to within p percent after (i.e. posterior to) publication of the table. A cell is said to be sensitive if $S_q(X) > -p$. Otherwise it is not sensitive.

Disclosure Risk for Frequency Count Tables

Sensitivity measures $S(X)$

- ▶ In a frequency count table the cell value, $N(X)$, is equal to the number of respondents who possess the properties defining that cell.
- ▶ Frequency count tables are formed by cross-classifying a number of categorical variables and one may view the frequency count table as carrying equivalent information to a set of microdata on these same variables.
- ▶ One may, in principle, therefore assess the disclosure risk in a frequency count table in the same way as for microdata.

References

- ▶ Duncan, G.T., Kelly-McNulty, S.A. & Stokes, S.L. (2001). Disclosure risk vs. data utility: the R-U confidentiality map. Technical Report No. 121, National Institute of Statistical Sciences, North Carolina
- ▶ Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Lenz, R., Longhurst, J., ... & Wolf, P. (2010). Handbook on statistical disclosure control. ESSnet on Statistical Disclosure Control. Available online at <https://bit.ly/3o52MDE>
- ▶ Templ, M. (2017). Statistical disclosure control for micro-data. Cham: Springer.
- ▶ Willenborg, L., & De Waal, T. (2012). Elements of statistical disclosure control (Vol. 155). Springer Science & Business Media.