Statistical Disclosure Control - Concepts

(1) A **microdata file** is a dataset that holds information collected on individual units; examples of units include people, households or enterprises.
For each unit, a set of variables is recorded and available in the dataset.

(2) A disclosure risk occurs if an unacceptably narrow estimation of a respondents confidential information is possible or if exact disclosure is possible with a high level of confidence.

Categorization of Variables

In accordance with disclosure risks, variables can be classied into three groups, which are not necessarily disjunctive.

1. Direct Identifiers

- Direct Identifiers are variables that precisely identify statistical units. (i.e. Primary Keys)
- For example, social insurance numbers, names of companies or persons and addresses are direct identifiers.

2. Key Variables

- Key variables are a set of variables that, when considered together, can be used to identify individual units.
- For example, it may be possible to identify individuals by using a combination of variables such as gender, age, region and occupation.
- Other examples of key variables are income, health status, nationality or political preferences.
- Key variables are also called **implicit identifiers** or **quasi-identifiers**.
- When discussing SDC methods, it is preferable to distinguish between categorical and continuous key variables based on the scale of the corresponding variables.
- **Non-identifying variables** are variables that are not direct identifiers or key variables.

Identity disclosure:

- In this case, the intruder associates an individual with a released data record that contains sensitive information, i.e. linkage with ex- ternal available data is possible.
- Identity disclosure is possible through direct identifiers, rare combinations of values in the key variables and exact knowledge of continuous key variable values in external databases.
- For the latter, extreme data values (e.g., extremely high turnover values for an enterprise) lead to high re-identification risks, i.e. it is likely that responends with ex- treme data values are disclosed.

Attribute disclosure:

- In this case, the intruder is able to determine some characteristics of an individual based on information available in the released data.
- For example, if all people aged 56 to GO who identify their race as black in region 12345 are unemployed, the intruder can determine the value of the variable labor status.

Inferential disclosure:

- In this case, the intruder, though with some uncertainty, can predict the value of some characteristics of an individual more accu- rately with the released data.

Statistical Disclosure Control - Concepts

- Removing and encrypting are only acceptable ways for disclosure control of identifier values.
- For another types of attributes we can apply masking methods.
- They can in turn be divided on two groups depending on their effect on the original data .

Perturbative methods. The original microfile is distorted, but in such a way that a difference between values of statistical rates of masking and original data would be acceptable.

Non-perturbative methods. They protect data without altering them. Non-perturbative methods base on principles of suppression and generalization (recoding).

- Suppression is a removing some data from original set.
- Recoding is a data enlargement.
- In case of continuous data (for example, age) non-perturbative methods can transform it to interval, fuzzy or categorical datatype.

Sometimes methods generating synthetic data or combined methods are also applied. Such methods are not acceptable at microdata preparation, because if we dont know the aim of the follow-up analysis, we can not generate adequate set of **synthetic data**.