Disclosure Control - Measuring the disclosure risk

A breach of confidentiality occurs when a statistical unit is re-identified and the values of sensitive variables are disclosed. Several approaches have been proposed to measure the disclosure, i.e., re-identifcation risk, but none of them has been universally accepted as the best method.

Disclosure Control - Measuring the disclosure risk

A quantitative measure of the risk, however, is necessary. Since the disclosure risk cannot be reduced to zero, such a measure would mean adopting a threshold rule to establish whether the release of a dataset is safe. Mathematical measures of the re-identification risk can be classified as:

Disclosure Control - Measuring the disclosure risk

Mathematical measures of the re-identification risk can be classified as:
Individual measures, which measure the risk per record. It is typically expressed by means of the probability of correctly re-identifying a unit, or by means of the uniqueness and rareness in the sample or population.

Statistical Disclosure Control - Measuring the disclosure risk

Global measures, which measure the risk for the entire file. It is typically expressed by means of the expected number of correct re-identifications. Global measures of risk can be derived by synthesizing individual measures. The advantage of an individual risk measure is that only those records appearing unsafe for a given risk threshold need to be locally protected, while a global measure involves the protection of the entire file.

Statistical Disclosure Control - Measuring the disclosure risk

- Let K be the number of combinations in the population P that is obtained by cross-tabulating a given set of key variables. Denote by k, k=1, ,K a combination of values observed on a sampled unit.
- Each combination k has its own re-identification risk.
- All records characterized by the same combination k share the same re-identification risk.

Statistical Disclosure Control - Measuring the disclosure risk

Let fk be the frequency count of the records in the sample presenting the same combination k of key variables, and let Fk be the frequency count relative to the same combination k in the population P.

Statistical Disclosure Control - Measuring the disclosure risk

In the following example, we assume that three variables are potential identifiers: sex (M=Male, F=Female), age, and marital status (M=Married; N=Never Married). The file contains 2,500 observations.