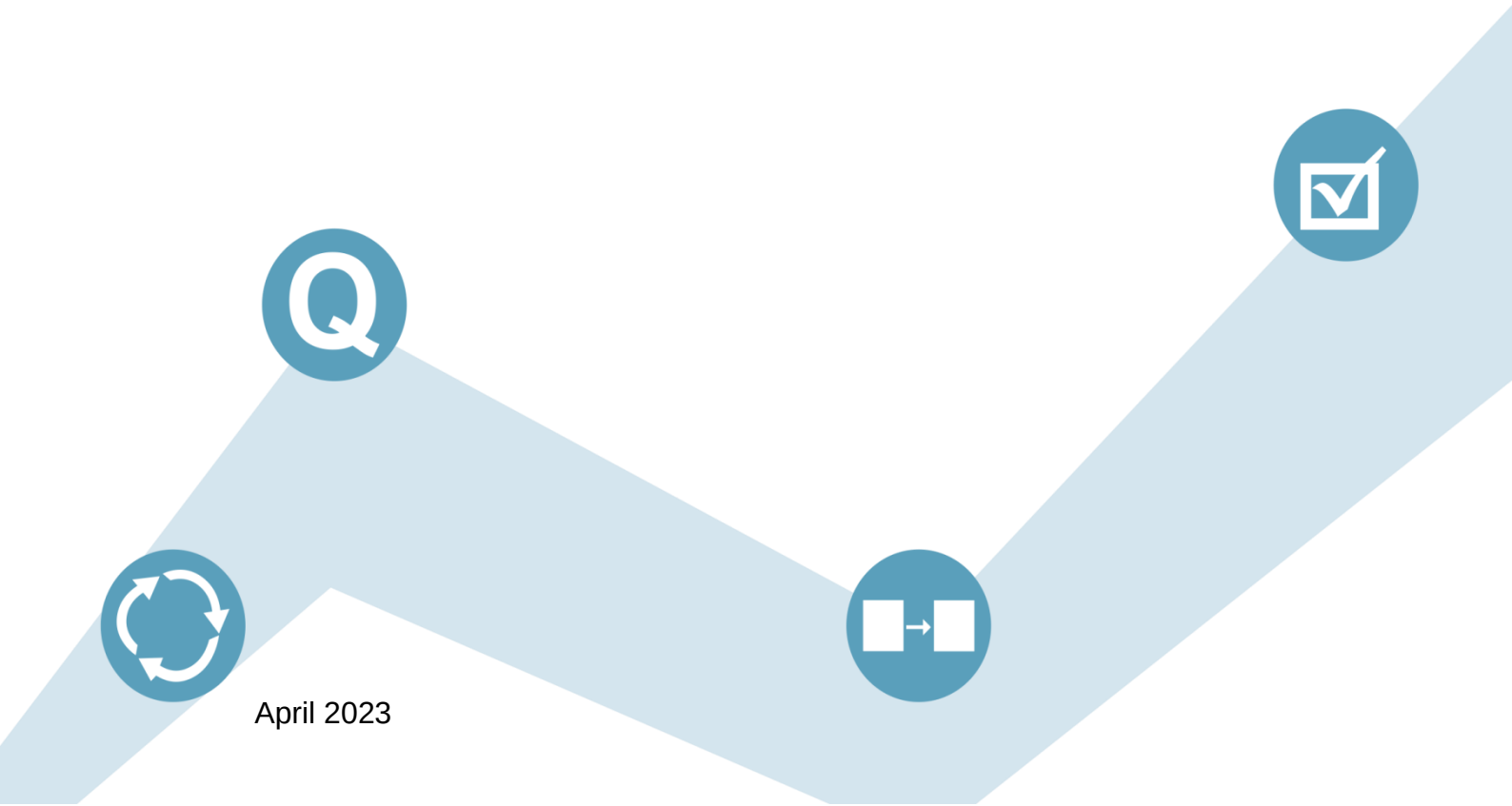
REPUBLIC OF SLOVENIA
STATISTICAL OFFICE

GENERAL METHODOLOGICAL EXPLANATION

Manca Golmajer

STATISTICAL DISCLOSURE CONTROL FOR TABULAR DATA



April 2023

Content

1 DESCRIPTION

Achieving strategic objectives and implementing the tasks of national statistics depend among other things on the trust of observation units (data providers), i.e. persons, households, enterprises, agricultural holdings and other organisations. This means that they will only trust us their data if they are certain that we provide full confidence, i.e. protect their identity, and confidentiality of provided information, and that the collected data will be used for statistical purposes only. To this end we are constantly informing them about this and explaining the procedures used. This is also the purpose of this methodological explanation.

A part of protection of data confidentiality is based on disseminated data not enabling direct identification (with the help of direct identifiers) or indirect identification (in any other way). Our office is obliged to do this by the national and European legislation. Since the mission of national statistics is to transmit and disseminate statistical results to as wide a circle of users as possible, but so that the risk of disclosure of information is minimal, methods of statistical disclosure control are needed to achieve that legislation is followed preserving a sufficient (the highest possible) information value of transmitted and disseminated statistical results.

We distinguish two types of methods for statistical disclosure control for tabular data:

- **Methods used on basic microdata:**
  - *Methods that change the data* (by using such a method the data should be changed as little as possible, but to such an extent that the risk of disclosure is minimal):
    - Rounding
    - CTA
    - Cell key method
  - *Methods that do not change the data* (data have a lower information value, but are still accurate):
    - Cell suppression
    - Aggregation
- **Methods for generating synthetic data** (data are obtained by taking into account a statistical model)

Methods are implemented in various tools (e.g. Tau-Argus, SAS-Tool, R (sdcTable, cellKey, ptable, sdcHierarchies packages)).

2 STATISTICAL DISCLOSURE CONTROL FOR TABULAR DATA

A table is a data structure that realises a mathematical structure of the matrix and distributes data in cells by rows and columns. A table shows aggregated data (various statistics) describing aggregate data of several units. Each statistic we intend to transmit can be represented as a 1 x 1 table. The table is defined by the explanatory variables (they specify the table) and by the response variable (it is tabulated). The number of explanatory variables determines the number of dimensions of the table.
In the procedure of statistical disclosure control for tabular data rules are determined for identifying primary sensitive cells (values shown in table cells). Which rules will be used depends on the type of table. There are three types of tables:

- **Frequency tables** (each table cell contains the number of units belonging to this cell)
- **Magnitude tables** (each table cell contains the sum of the values of a certain variable of units belonging to this cell)
- **Other tables** (table cells contain values of other types of statistics – shares, ratios, indices, etc.) that are usually related to magnitude tables and frequency tables

Rules for identifying primary sensitive cells are:

- **Threshold** (if the number of units is smaller than the threshold value, the cell is primary sensitive); the minimum value for a threshold should be 3
- **Dominance rule** (if n largest contributors contribute more than k% to the cell value, the cell is primary sensitive)

- **p%-rule** (if the unit with the second largest contribution estimates the largest contribution better than p% accurately, the cell is primary sensitive)

In frequency tables only the threshold is considered, while in magnitude tables and other tables the threshold and the p%-rule or the threshold and the dominance rule are considered. The rules for identifying primary sensitive cells are confidential and cannot be transmitted, since by disclosing parameters the protection of units would be diminished.

If certain units agree (we have to obtain their written consent) that their individual data can be released or if some of their data are public, the so-called request rule can be applied. This means that these units are assigned a special status and thus the loss of information in tables in which they appear is smaller.

Using a method for statistical disclosure control for tabular data, primary sensitive cells are always protected. Due to the links within the table, additional cells (called secondary protected cells) are often protected. If there are no links, secondary protection is not necessary.

Using cell suppression, we suppress the primary and secondary sensitive values by not releasing them, i.e. not entering them in the cell, but entering letter „z“ (confidential data). Loss of information is determined by the share of suppressed cells.

Using the cell key method, all the cells in the table are randomly changed regardless of whether they are primary sensitive or not. A random cell change can be positive, negative, or zero. A table protected by the cell key method is not additive in general.

In statistical disclosure control for tabular data one needs to be careful about the hierarchies of explanatory variables, the number of table dimensions and links between tables.

3 EXAMPLE OF STATISTICAL DISCLOSURE CONTROL FOR TABULAR DATA

Let’s say the frequency table has an explanatory variable for which it holds that Total = 1 + 2, 1 = 11 + 12, and an explanatory variable for which it holds that Total = A + B. Let’s say that the rule for identifying primary sensitive cells is threshold 3. So the table has only one primary sensitive cell, i.e. cell [11,A] with value 1. This primary sensitive cell can be protected in several ways.

In the case of aggregation, the number of primary sensitive cells is reduced by transforming the table: in our case eliminating both subcategories of category 1.

In the case of cell suppression, we decide to suppress in addition to the primary sensitive cell [11,A] also secondary sensitive cells [11,B], [12,A], [12,B]; in this way we prevent the user to disclose the real value of the primary sensitive cell with the help of simple mathematical functions.

In the case of the cell key method, the value of each cell is randomly changed. The resulting table is not additive, even though the unprotected table is.

Original table

	Total	A	B
Total	51	23	28
1	21	11	10
11	7	1	6
12	14	10	4
2	30	12	18

Primary protected table

	Total	A	B
Total	51	23	28
1	21	11	10
11	7	z	6
12	14	10	4
2	30	12	18

Final protected table

Example of protection by aggregation:

	Total	A	B
Total	51	23	28
1	21	11	10
2	30	12	18

Example of protection by cell suppression:

	Total	A	B
Total	51	23	28
1	21	11	10
11	7	z	z
12	14	z	z
2	30	12	18

		Example of protection by the cell key method:

	Total	A	B
Total	51	23	27
1	21	12	10
11	7	2	6
12	13	8	3
2	29	12	18