Suppression or perturbation?

Wim Kloek (Eurostat)

Joint UNECE/Eurostat Expert Meeting on Statistical Data Confidentiality

2 December 2021

Purpose

- Compare two specific ways to protect confidential data in tables: suppression and perturbation
- Suppression is still the dominant protection method
- There are several alternative perturbation methods. We will focus on the cell key method, as the method is in use. 1. Integrated in table building tools, which allows users to define their own tables. 2. In the European context, the method is also recommended for the protection of the census 2021.

Suppression

- On the basis of certain rules (threshold, dominance) the confidential cells are established; these cell will be suppressed
- Additional cells have to be suppressed to avoid calculating back the confidential cells by differencing (secondary confidentiality)
- There are usually several solutions for the choice of the secondary suppressions. The value should be high enough to actually protect the confidential cell. Otherwise you would like to avoid suppressing higher level aggregates.

Perturbation: cell key method

- Developed by the Australian Bureau of Statistics (ABS)
- The method was originally developed for the census, and we describe here the version with additive perturbation used for frequency counts
- The basis of the method is microdata. All records in the file get a random key. The keys in a cell get summed to the cell key. The perturbation is determined by the looking up the combination of the cell key and the cell value in a table of random perturbations

Features of the cell key method

- Consistent perturbation: each time the same set of records are in one cell, the same perturbation will be applied
- The perturbation adds pre-defined variance to the cells; the variance can be shared with user, without endangering the protection
- Column and row totals are perturbed in the same way as individual cells (this produces some lack of additivity)
- To protect the method, also values that do no require protection are perturbed

Thou shalt not perturb

- Especially from the research environment comes the message that it is not the task of statistical offices to ‘manipulate’ data. Data should be reported as unprocessed as possible. There is no trust in the competence the statistical authorities or in the neutrality of the methods
- Perturbation is not the only way statistical authorities ‘manipulate’ data that can potentially create biases: detection and correction of errors in the data, modelling for non-response
- The cell key methods produces unbiased perturbation
- Statistical offices should be transparent on their ‘manipulations’

Problems with suppression

- Incomplete tables that are difficult to analyse as a whole
- Difficult to manage over linked tables (in the standard table set and in tables on request)
- Difficult to manage over time (need to freeze suppression patterns)
- If an intruder knows (approximately) one of the suppressed values, other values can be calculated back (domino effect)

Problems with cell-key perturbation

The level of perturbation (what is the usefulness of the data?)

How large is the perturbation compared to sampling and non-sampling errors? Statistical offices tend to not be very transparent on non-sampling errors. It is true, that they are often difficult (or expensive) to measure, but not reporting them sends the wrong message. Is this mainly an issue of communication and statistical awareness?

The additivity of the data is lost

The lack of additivity feels to some as a serious indicator of bad quality. Lack of additivity also occurs in rounding, and there a simple footnote seems to be sufficient. For perturbation the problem is perceived as worse (trust).

Problems with cell-key perturbation (2)

All cells get perturbed

With the cell-key method all cells get perturbed, also cells that do not need any protection. Obviously, the relative perturbation is much higher in the small counts. Perturbing the whole table is required to protect the protection method and to guarantee the unbaisedness of the result.

Perturbed means safe?

The user may think to identify someone in the data, even if this identification is far from certain.

Areas of application

Full counts and near full counts (weights close to 1)

Relatively small perturbations will protect small counts. The cell-key method allows detailed and flexible production of tables.

Frequency counts from sample surveys

Sample surveys are inherently already a bit protected, assuming that response knowledge is not widely available. The perturbation can be done on the unweighted numbers, and this perturbation can be smaller than in full counts (but after applying the weights it will still be higher). No need to have integers for the perturbation, the results are rounded to integers anyway.

Application in magnitude tables

- Protection of magnitude tables, especially in business statistics, are problematic, as due to skewed distributions, the most important data are usually also primarily at risk of disclosure.
- In the suppression approach waivers are practically the only way out. This is a very expensive method (depending on the scope and requirements for renewal of the waiver).
- A promising multiplicative extension of the cell-key method exists. It consists in multiplying the largest contributor in the cell with a random factor. This approach is directly linked to the p% rule for detecting primary confidentiality due to dominance. Further experience with this approach has to establish in which situations the information value of the tables remain sufficient to publish.

Suppression or perturbation?

- In the domain of frequency count tables, both approaches come with advantage and disadvantages. These advantages and disadvantages are of a different nature and cannot be easily compared in a mathematical way. The choice depends on the user preferences for certain characteristics: additivity, flexibility of producing additional tables. In the end also practical arguments (ease to implement in the production process) can play a role.
- Non-sampling errors are a natural protection against disclosure. More research into non-sampling errors and more transparency could change the user attitude towards perturbation. Such transparency is not self-evident, as it seems to undermine the trust in the data.
- Further research is required in the potential of the multiplicative cell-key method for magnitude tables.