STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE

COMMISSION OF THE EUROPEAN COMMUNITIES

CONFERENCE OF EUROPEAN STATISTICIANS

EUROSTAT

Joint ECE/Eurostat Work Session on Statistical Data Confidentiality
(Skopje, The former Yugoslav Republic of Macedonia, 14-16 March 2001)

Working Paper No. 42
English only

Topic I: Application of statistical disclosure control methodology and software in business statistics and social and demographic statistics

A COMPARATIVE TEST FOR SEVERAL THRESHOLD VALUES IN FREQUENCY TABLES: A TAU-ARGUS PERFORMANCE EXAMPLE

Contributed paper

Submitted by the Basque Institute of Statistics (EUSTAT), Spain [1]

I. INTRODUCTION

1. The election of a suitable threshold value (n) in table protection is one of the main points of discussion for a statistical agency, when fixing generic rules to preserve confidentiality. Nowadays, different criteria are applied in statistical offices and institutes. These cell limit values vary from 3 to 5, in some cases. In Spain, the most expanded rule is not to publish frequency cells under 3 when at least one of the crossed variable is non-public and the data refers to a geographical area less than certain size [1].

2. However, the final decision is based on the safety requirements imposed by the statistical office. But, which is the best choice? What added quality we provide to our data if we increase this threshold value? Is it worth in terms of information loss?

3. In EUSTAT, we have been testing τ-Argus software for table protection and we have used it to test several limit values for the parameters in the sensitive rules that the package applies. In this performance example we have checked two frequency tables from the Census survey in the Basque Country. Different threshold values were imposed for each table considering the same cost variable.

4. We will focus on the variation of the number of total suppressions needed to properly protect each table. This number is supposed to rise if we increase the cell limit. We will also distinguish between the primary and the secondary suppressions in order to compare their increases.

II. VARIABLE SPECIFICATION AND TABLE DESCRIPTION

5. The τ-Argus input file must be a fixed format ASCII file. In our case, we start from a record file containing individuals by rows and variables by columns. Microdata refer to the last Census survey in the Basque Country, made in 1996. The total number of records is 2,098,055.

6. A metadata file containing variable information is also needed but it can be interactively specified during the Argus session. In a previous step, we have selected some variables of our interest and included them in the input file. Variables are specified as follows:

[1] Prepared by Marta Mas.

- *Place of Residence*, divided in 250 municipalities;
- *Age*, originally coded by year but recoded later in three groups (≤19, 20-64, ≥65);
- *Residential Situation*, divided in two categories: present and absent residents;
- *Sex*;
- *Municipality Size*, used as a cost variable.

7. The election of the cost variable is based on the way Argus uses it. The software allows the data protector to guide the suppression process by associating weights with the cell items in a table. These weights measure how important the cell value is. The higher is the weight the less likely for the cell to be suppressed. If we consider the municipality size as a weight (cost) variable, the procedure will prefer to suppress cells pertaining to small areas where the disclosure risk is higher. This supposes an added protection to the table, taking into account the high level of geographical detail we consider in this case (250 municipalities).

8. Tables to be constructed and protected by τ-Argus are the following:

- *Place of Residence* x *Age* (three groups) x *Sex*

- *Place of Residence* x *Residential Situation* x *Sex*

9. Both are frequency tables as they represent population in their cell items and both generate a "low" number of unsafe cells for common values of the threshold. Finding an optimal pattern of suppressions in terms of information loss, is a complex programming problem. Argus can solve it in more or less computational time, depending on the number of primary sensitive cells to protect. It is desirable to launch the suppression process with the minimum number of primary suppressions. This factor should make the computational work easier and faster, considering that we are going to check different cell limits for each table.

III. SENSITIVE RULES AND PARAMETER VALUES

10. In the case of magnitude tables, Argus provides an extended sensitive rule to detect the unsafe cells, based on the dominant contributions to a cell (N, p-rule) [2]. In addition, it allows to fix a cell limit value (n) which represents the minimum number of records contributing to a cell. Cell items under this threshold value are considered sensitive. This last option is the only criteria we are going to apply to the cells in a frequency table, as it is our case.

11. The objective is to test different values for the threshold in the same table. Of course the range of suitable values is not very wide. If we consider $n \leq 1$, only unitary cells are considered unsafe, which is not enough protection in many cases. On the other hand, if we impose $n \leq 5$, of course it is not easy to identify any individual, but the cost in terms of information loss is higher. Therefore, anything between 2 and 5 can be considered acceptable but, what is meant by "acceptable"?. It is necessary to find a balanced value which provides the required protection and keeps the maximum amount of information.

Example 1

12. We run τ-Argus through the first table specification: *Place of Residence* x *Age* (three groups) x *Sex*. The variable *Age* has been recoded in three groups before the suppression process. For each threshold value we stare at the number of primary suppressions (sensitive cells) and secondary ones (those needed to protect the primaries) after the suppression process. The information about the process and the suppressions made are saved in a report file that the program provides with the following contents:

Produced 19:11:29 on 14/12/2000

The input file with metadata is C:\Censo\Datos\Area1.rda (20:16:20 on 14/11/2000)

The input file with microdata is C:\Censo\Datos\Area1.asc (11:30:27 on 1/8/2000)
The table was saved in C:\Censo\Datos\Ejemplo1.ttb as follows:

Table:

MUNRv1 x EDAD3P x SEXOP : frequency

Cost variable for cells: TMUNR

The cell frequency limit 5 was applied;
the safety range for each cell was [70%, 130%].

There are 5 primary and 3 secondary suppressions in the elementary cells.
There are 0 primary and 6 secondary suppressions in the 2-dimensional marginals.
There are 0 primary and 0 secondary suppressions in the 1-dimensional marginals.
The general total was not suppressed.

"EDAD3P" has been recoded as follows:
1: 1- 19
2: 20- 64
3: 65-101

13. The information referring to the number of suppressions has been summarised in this table below, for all the values tested for *Example 1* table:

Example 1.

Place of residence x age(3 groups) x sex		
Threshold value	**Type of suppression**	**Number of suppressions**
n <= 1	Primary	0
	Secondary	0
	Total	**0**
n <= 2	Primary	0
	Secondary	0
	Total	**0**
n <= 3	Primary	1
	Secondary	3
	Total	**4**
n <= 4	Primary	3
	Secondary	11
	Total	**14**
n <= 5	Primary	5
	Secondary	9
	Total	**14**
n <= 6	Primary	8
	Secondary	12
	Total	**20**

14. As was expected, the more the primary suppressions rise, the more secondary ones we need to protect the table. However, primary suppressions themselves provide additional protection. Thus, in this example, in case of n<=4, eleven secondary suppressions are needed to protect three sensitive cells, while only nine secondary ones protect five primary suppressions in case of n<=5. In this example, the election of the threshold value as either 4 or 5 gives the same results in terms of total suppressions.

Example 2

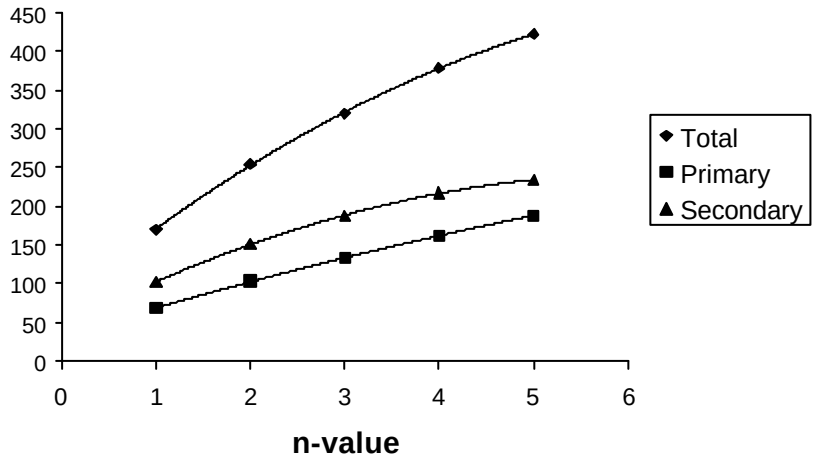
15. We applied the same process followed for the first table to this second example: *Place of Residence* x *Residential Situation* x *Sex*. These are the results obtained:

Example 2.

Place of residence x residential situation x sex		
Threshold value	**Type of suppression**	**Number of suppressions**
n <= 1	Primary	68
	Secondary	102
	Total	**170**
n <= 2	Primary	103
	Secondary	151
	Total	**254**
n <= 3	Primary	133
	Secondary	187
	Total	**320**
n <= 4	Primary	161
	Secondary	217
	Total	**378**
n <= 5	Primary	188
	Secondary	234
	Total	**422**

16. This table in *Example 2* could be treated as well by means of a previous recoding of the variable "*Residential situation*". In case n≤1, this characteristic generates all the primary suppressions in a unique sensitive category (*absent residents*), therefore we could avoid to suppress any cell by aggregating this group.

17. Nevertheless, our interest resides now in the variation of the number of suppressions as the threshold value changes. The following graphic shows the increasing of the suppressions in the *Example 2*:



18. The number of secondary suppressions tends to approximate the primary ones for higher values of n. This situation is going to soften the total number of primary suppressions. Of course, the range of suitable values for the threshold limit n, is not very large as it has no sense to consider, in practice, values higher than 5. Nevertheless, it could be interesting to see both: how the growing number of total suppressions behaves as we increase the threshold value, and also the consequences for the primary and secondary suppressions.

19. As we have seen, in most cases, we prefer a higher protection level assuming certain added cost in terms of information loss. However, this cost does not differ very much if we consider consecutive values for the threshold. Thus, we could apply a "solid" threshold value that provides the required protection and improves the confidence of our respondents, at the same time.

References

[1] Garín, A. Urrutia J. "Keeping statistical secrecy: basic elements in a data protection system". OFISTAT seminar. (October 2000).

[2] Hundepool, A. Willenborg, L. Tau-Argus. Version 2.0 User`s Manual (December 1998).