Statistical Office in Poznań

Statistical disclosure control for hypercubes with data from the National Population and Housing Census 2021

Kamil Wilak[1,2], Andrzej Młodak[1,3],
Tomasz Józefowski[1,2], Tomasz Klimanek[1,2]

[1]Statistical Office in Poznań
[2]Poznań University of Economics and Business
[3]Calisia University

July 2-4, 2024

Outline of the presentation

Introduction

Introduction

- The preparation of data from the National Population and Housing Census conducted in 2021 in the form of hypercubes (specific multi-dimensional tables) not only meets the needs of potential users of statistical data but also meets the requirement to supply the resources of the Statistical Office of the European Union (Eurostat), specified in the relevant legal regulations of the European Commission.
- Due to the significant detail of the information contained in hypercubes developed according to EU assumptions, it was necessary to use the Statistical Disclosure Control (SDC) methods aimed at effectively protecting statistical confidentiality by minimizing the risk of identifying an entity while maximizing the usefulness of the shared data.

Hypercubes

Hypercubes

- total population – persons (103)
- private households (3)
- families (3)
- conventional dwellings (2)
- occupied conventional dwellings (7)
- living quarters (1)

} 119 hybercubes

Hypercubes

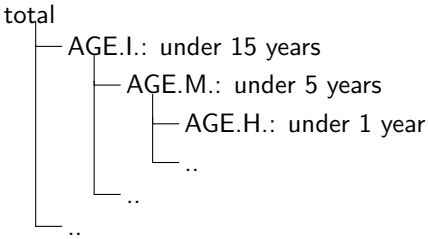
Hypercube group 1 „Marital status of people in households"

No	Number of cells	Breakdowns					
1.1	3072	GEO.N.	SEX.	AGE.H.	LMS.H.		
1.2	3840	GEO.N.	SEX.	AGE.H.		HST.H.	
1.3	4608	GEO.N.	SEX.	AGE.H.			FST.H.
1.4	249	GEO.N.	SEX.		LMS.H.	HST.H.	

GEO.N – Place of usual residence
SEX – Sex
AGE.H – Age
LMS.H – Legal marital status
HST.H – Household status
FST.H – Family status

Hypercubes

Hierarchy of the AGE variable



Cell Key Method

Cell Key Method

Procedure

1. Assign each record a random number (record key – Rkey).

Record	Rkey
1	0,03729499
2	0,15868515
3	0,00846373
4	0,66297515
5	0,83186228
...	...
N	0,97542859

Cell Key Method

Procedure

2. Create a frequency table. For each cell, sum record keys and take the modulo to get the cell key.

Age \ Sex	Total	Female	Male
Total	.	.	.
under 15 years	.	.	4
15 to 29 years	.	.	.
30 to 49 years	.	.	.
50 to 64 years	.	.	.
65 to 84 years	.	.	.
85 years and over	.	.	.

Record	Rkey
2	0,15868515
4	0,66297515
56	0,30777595
72	0,77265550

Sum Rkey =	1,90209175
Cell key =	0,90209175

Cell Key Method

Procedure

3. Set the parameters.

- `D` – perturbation parameter for maximum noise (scalar integer)
- `V` – perturbation parameter for variance (scalar double)
- `js` – threshold value for blocking of small frequencies (i.e. the perturbation will not produce positive cell values that are equal to or smaller than the threshold value). (scalar integer)
- `pstay` – optional parameter to set the probability ($0 < p < 1$) of an original frequency to remain unperturbed: NA (default) no preset probability (i.e. produces the maximum entropy solution)

Cell Key Method

Procedure

4. Use perturbation table to get perturbation value from cell value and cell key ($\texttt{D} = 8, \quad \texttt{V} = 3, \quad \texttt{js} = 2, \quad \texttt{pstay} = \texttt{NA}$).

Count	Cell key	Noise
4	(0.00000000, 0.08411495)	-4
4	(0.08411495, 0.36322993)	-1
4	(0.36322993, 0.64234490)	0
4	(0.64234490, 0.83157663)	+1
4	(0.83157663, 0.93583431)	+2
4	(0.93583431, 0.98044632)	+3
4	(0.98044632, 0.99527235)	+4
4	(0.99527235, 0.99909902)	+5
4	(0.99909902, 0.99986613)	+6
4	(0.99986613, 0.99998557)	+7
4	(0.99998557, 1.00000000)	+8

Cell Key Method

Procedure

5. Apply the chosen perturbation to the cell.

Age \ Sex	Total	Female	Male
Total	.	.	.
under 15 years	.	.	4 + 2 = 6
15 to 29 years	.	.	.
30 to 49 years	.	.	.
50 to 64 years	.	.	.
65 to 84 years	.	.	.
85 years and over	.	.	.

Cell Key Method

Properties

Consistency (+)

If a particular cell appears in more than one table, it is always perturbed in the same way.

Non-additivity (-)

Since noise is applied independently to inner and marginal aggregates, those often do not add up exactly.

Results

Results

Set of parameters:

- $\texttt{D} \in \{8, ?, ?, ?\}$
- $\texttt{V} \in \{3, ?\}$
- $\texttt{js} = 2$
- $\texttt{pstay} \in \{\texttt{NA}, ?, ?, ?\}$

} 32 sets

To discourage attempts of inferring true values, the central parameters (here: noise variance `V` and noise maximum `D`) are generally not published.

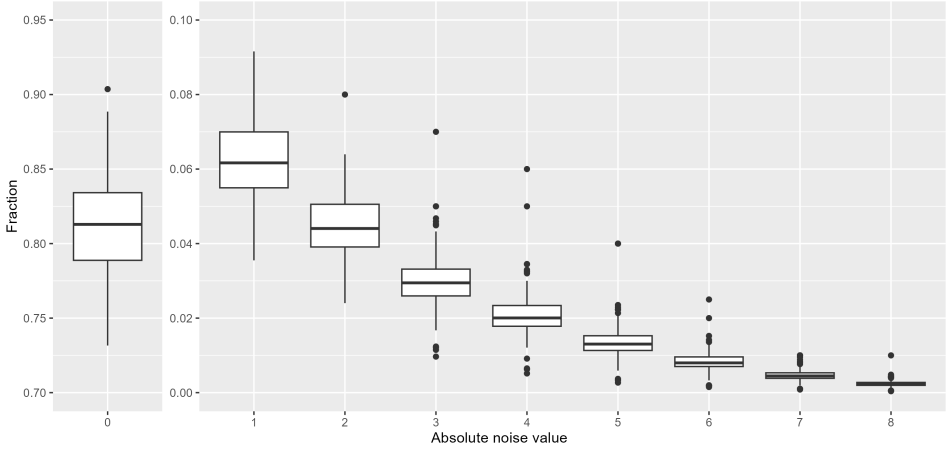
Results

R packages:

- `cellKey`: Consistent Perturbation of Statistical Frequency- And Magnitude Tables
- `ptable`: Generation of Perturbation Tables for the Cell-Key Method
- `sdcHierarchies`: Create and (Interactively) Modify Nested Hierarchies

Results

Distribution of absolute noise values



Literature

Literature

Eurostat (2019). EU legislation on the 2021 population and housing censuses. Explonatory notes. Publications Office of the European Union, Luxemburg.

Fraser, B. and Wooton, J. (2005). A proposed method for confidentialising tabular output to protect against differencing. In *Joint UNECE / Eurostat Work Session on Statistical Data Confidentiality, Geneva, Switzerland, November 9-11*.

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Lenz, R., Naylor, J., Nordholt, E. S., Seri, G., and deWolf, P.-P. (2024). *Handbook on Statistical Disclosure Control, Version 2.0. ESSNet.*

Schulte Nordholt, E., Golmajer, M., de Vries, M., de Wolf, P.-P., Tent, R., Giessing, S., van de Laar, R., Krol, N., and Bach, F. (2024). Guidelines for Statistical Disclosure Control methods for census and demographics data. STACE project WP2 Deliverable D2.11.

Thank you for your attention