Protection of frequency tables – current work at Statistics Sweden

Karin Andersson
Ingegerd Jansson
Karin Kraft

name.familyname@scb.se

Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality
Helsinki 2015





Background

- Data collected for official statistics are by the main rule confidential
- Exceptions from the legislation make it possible to publish statistics and to use data for research purposes, if it can be guaranteed that a disclosure will not cause harm or damage to an individual, household or establishment
- Since 2007 moving towards standardization: common methodology and common IT-tools
  - Handbook
  - Training
  - Common IT-tool mainly for magnitude data
  - Implementation in a number of surveys
  - Solution for Census 2011
  - Ongoing project: method for frequency tables from registers

ABS Census TableBuilder Protection Method

- Fraser & Wooton (2005), Leaver (2009), Marley & Leaver (2001), Thompson et al (2013)
- Implemented by ABS in their online tool
- Perturbative method –noise is added to the original cell values
- Protects against differencing
- Each cell is perturbed in the same way every time it is requested: consistency between tables
- Hierarchies or other complex tables or different choices of background variables doesn´t matter
- All cells are protected separately: sums of protected cell values might not equal the corresponding protected margins

ABS TableBuilder Protection Method

Original micro data:
N objects with observations on V variables

	Y_1	…	Y_V	Record key
1				
2				
3				
.				
.				
.				
N				

Each object is randomly assigned a permanent numeric value: record key.

When a table is requested, record keys are combined to create cell-level keys, one for each cell *u* in the requested table.

ABS TableBuilder Protection Method

Look-up table: 256 rows and a predefined number of columns (≤ 30)

	1	2	…	30
1	p_{11}	p_{12}		
2	p_{21}			
3				
.				
.				
.				
256				

The rows of the look-up table are row indices corresponding to cell-level keys.

The columns are original cell counts *u* (unperturbed).

The p_{ij} are perturbation values that will be added to the original cell values, resulting in protected cell values $c=u+p_{ij}$

ABS TableBuilder Protection Method

p is determined by maximizing the entropy subject to a number of constraints

$$-\sum_{p\in\Pi_i} P_i(p \mid u)\log[P_i(p \mid u)]$$

The entropy is maximized subject to

- $P_i(p|u)$ is nonnegative and their sum is equal to 1,
- *p* will not add bias to the cells, $E[p] = 0$,
- the confidentialized values *c* = *u* + *p* cannot be negative,
- the set of available perturbation values,
- *c* cannot take specified values,
- the variance of *p* cannot exceed a specified threshold, $Var[p] \leq v_p$

Tests

- Find a good set of parameters for the ABS method
- Compare with other methods and evaluate on risk and utility loss

- Four datasets from the Total Population Register:
  - County (21), Age (6), Civil status (7)
  - Parish (1371), Age (6), Civil status (7)
  - County (21), Age (6), Country of Birth (216)
  - Parish (1371), Age (6), Country of Birth (216)

Choice of parameter values

- Sets of available perturbation values:
  - -1,…,1
  - -3,…,3
  - -5,…,5
  - (-10,…,10)
- $c \geq 0$ or $c \neq 1,2$
- Maximum variance of the perturbation
  - 1, (3), 5, 10, 15, 30, (50)

Risk and utility loss

- Risk measure 1: variance of noise

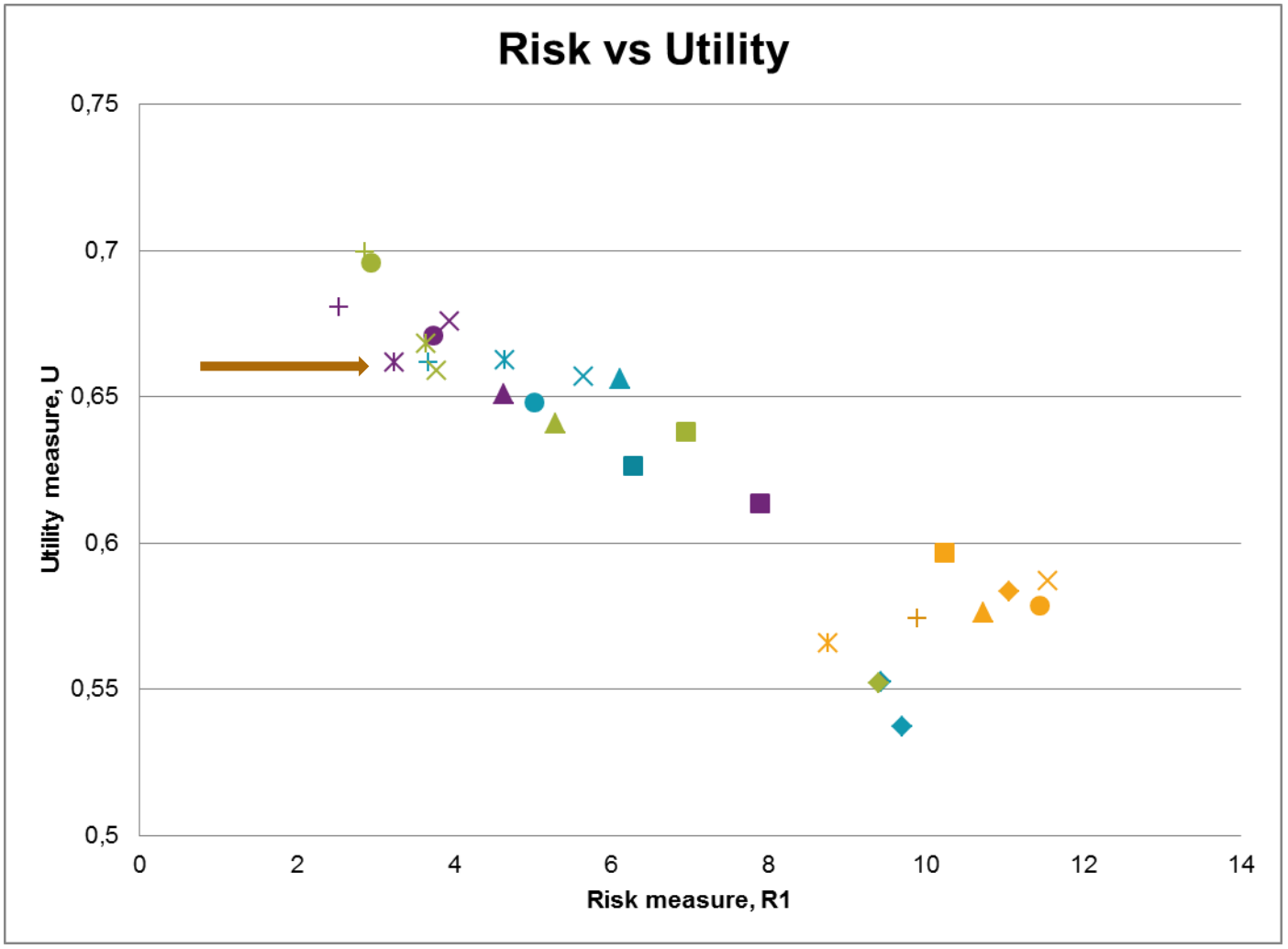
$$\boldsymbol{v}(\boldsymbol{t}) = \left(\frac{1}{v_1(\boldsymbol{t})}, \frac{1}{v_2(\boldsymbol{t})}, \ldots, \frac{1}{v_M(\boldsymbol{t})} \right)$$

$$R_1(\boldsymbol{t}) = \|\boldsymbol{v}(\boldsymbol{t})\|_2$$

- Risk measure 2: $R_2(\boldsymbol{t})$, percentage of unchanged cells
- Utility loss: Hellinger distance (Shlomo 2007)

Statistiska centralbyrån Statistics Sweden

Example of graph



Comparing methods

- Perturbations -5,…,5, variance ≤ 15, $c \geq 0$
- Perturbations -5,…,5, variance ≤ 10, $c \neq 1,2$
- Deterministic rounding to two digits
- Deterministic rounding to base 5

- (Probabilistic rounding of only small values:1,2)

Results

Data set	Measure	-5,…,5, 15,c≥0	-5,…,5, 10,c≠1,2	Rounding base 5	Rounding 2 digits
1	U	0,66	0,70	0,87	0,96
	R_1	3,2	2,9	-	-
	R_2	13,1	8,7	19,2	4,1
2	U	0,16	0,22	0,78	0,87
	R_1	19,5	20,5	-	-
	R_2	10,6	8,8	18,1	7,7
3	U	0,17	0,31	0,82	0,90
	R_1	11,1	13,5	-	-
	R_2	12,6	7,9	15,2	5,8
4	U	0,07	0,19	0,90	0,96
	R_1	20,7	20,5	-	-
	R_2	18,0	5,3	8,1	2,4

Implementation

- Defining risk
- Production
  - On appropriation
  - On commission
- Consequences for users
- Future work