Statistics Canada Statistique Canada

Disclosure control that accounts for survey realities: assessing the risk using G-Confid

www.statcan.gc.ca

Telling Canada's story in numbers

Peter Wright
Methodology Branch

Presented to the UNECE work session on confidentiality

21 September 2017

Canada

Presentation

- Overview of G-Confid
- Modernizing G-Confid to treat
  - waivers
  - negative values
  - estimation weights
- *PTN*, a new framework
- Future directions

Overview of G-Confid version 1.06

- Generalized system programmed in SAS™
- Created at Statistics Canada (closed source)
- Licence is free since 2015
- Serves to ensure the protection of tabular magnitude data mainly for business surveys
  - **PROC SENSITIVITY** identifies sensitive cells
  - The **SUPPRESS** macro protects sensitive cells using iterative linear programming
- See Rondeau and Fillion (2011) for information

Overview of G-Confid: Proc Sensitivity

- Sensitivity rule (PQ or P%, NK, custom-made rule)
- Traditional linear sensitivity measure

$$S = \sum_{r=1} \alpha_r x_r \text{ where } 1 \geq \alpha_1 \geq \cdots \geq -1$$

where the α_i values represent the rule.

- Example (x_1 is the *target* and x_2 is the *suspect*):

$$\text{Let } \alpha_1 = \frac{p}{q} \qquad \alpha_2 = 0 \qquad \alpha_3 = \alpha_4 = \ldots = -1$$

$$\text{Then } S = \frac{p}{q} x_1 - 0 x_2 - \sum_{r \geq 3} x_r$$

Modernizing G-Confid

- Initiatives must:
  - Improve the assessment of confidentiality
  - Respect increasing demands to publish more data
  - Reduce user burden
- Improvements involving three specific aspects of survey sampling:
  - Adjusting sensitivity in the presence of **waivers**
  - Processing **negative values** (mixed-sign variables)
  - Making use of **estimation weights**

Modernizing G-Confid: waivers

A *waiver* is a signed record of the respondent granting permission to publish its data. Waivers greatly help Statistics Canada to publish more data.

Old way:

1. G-Confid calculates the sensitivity, ignoring waivers.
2. (manual check) If the top two contributors supplied waivers then manually recode the sensitivity to zero.

Modernizing G-Confid: waivers

New way, **PQ or P% rule**:

1. If the largest contributor (or *target*) supplied a waiver. G-Confid changes its role to that of the *suspect*, and the second largest contributor becomes the *target.*

2. G-Confid calculates $S = \left(\frac{p}{q}\right) x_2 - \sum_{r \geq 3} x_r$

(second-largest contributor)

Modernizing G-Confid: waivers

New way, **NK rule**: convert to the PQ rule.

G-Confid calculates:

1. the sensitivity ignoring waivers

$$S_0 = \left(\frac{100-k}{k}\right)(x_1 + \cdots) - \sum_{r>N} x_r$$

2. the relative protection offered to the largest contributor

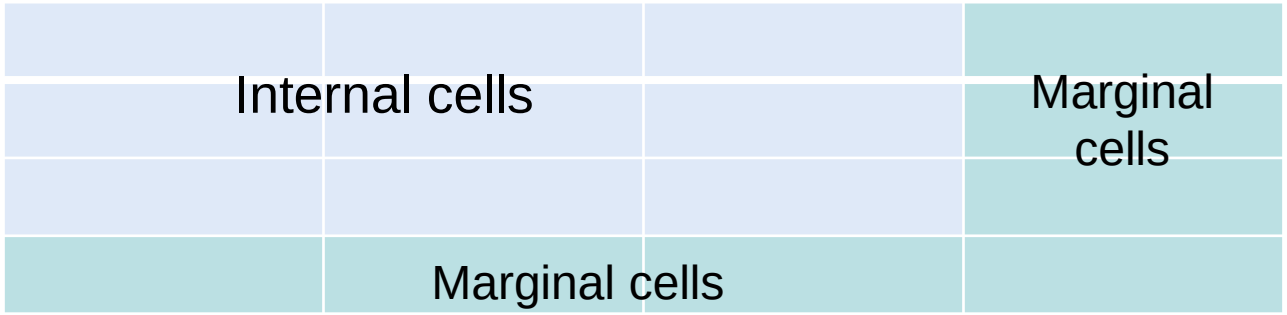
$$\left(\frac{p}{q}\right) = \frac{1}{x_1}\left(S_0 + \sum_{r>N} x_r\right)$$

3. the sensitivity to protect the largest contributor without a waiver (*target t*)

$$S = \left(\frac{p}{q}\right)x_t - \sum_{r \notin t, x_1} x_r$$

Modernizing G-Confid: negative values

- Several solutions have been proposed; see for example the FCSM (2005), Giessing (2008), and Daalmans and de Waal (2010)
- Starting with version 1.07, G-Confid users can choose from three options:
  - Absolute values at the level of the internal cells
  - Absolute values for all cells, including marginal cells
  - Use of a proxy variable (see next slide)



Modernizing G-Confid: negative values

Tambay and Fillion (2013): assess the variable Z_i for sensitivity where

X_i is a mixed-sign variable (e.g., profit)

Y_i is a non-negative size variable (e.g., gross revenue)

δ is a parameter defined on $0 \le \delta \le 1$

$$Z_i = \max\left\{\left|X_r\right|, \delta Y_r\right\}$$

Modernizing G-Confid: negative values

Examples of generating $Z_i = \max\left\{\left|X_r\right|, \delta Y_r\right\}$

Let δ = 0.05 (chosen by the G-Confid user)

Enterprise	Profit	Gross Revenue	Z
ABC Co.	-20	600	max(|-20| , 0.05×600) = 30
LMN Inc.	-50	600	max(|-50| , 0.05×600) = 50
XYZ Ltd.	30	1500	max(|30| , 0.05×1500) = 75

$$S = \left(\frac{p}{q}\right)(75) - 30$$

Modernizing G-Confid: weights

- No obvious way to include weights when using the traditional linear sensitivity measure
- Old solution:

Assess the sensitivity among contributions with weight < 3

Subtract if weight ≥ 3

$$S = \left(\frac{p}{q}\right) x_{1|WEIGHT<3} - \sum_{r\geq 3 \,|\, WEIGHT<3} x_r - \sum_{WEIGHT\geq 3} x_r$$

- Scalar value x_i captures limited information

PTN, a new framework

Gray (2016) proposed the Precision Threshold and Noise (*PTN*) framework.

Each contributor is represented by a vector:

- **Precision Threshold (*PT*)**: the degree of protection that must be accorded to the contribution
- **Self-noise (*SN*):** the amount of protection provided by a suspect's own contribution
- **Noise (*N*):** the amount of protection offered by a contributor that is neither target nor suspect

PTN, a new framework (cont.)

PQ rule without weights

PT(r)	***SN(r)***	***N(r)***
px_r	0	qx_r

PQ rule with weights (assuming $w_r \geq 1$)

PT(r)	***SN(r)***	***N(r)***
$px_r - f(x_r, w_r)$	$q(w_r - 1)x_r$	$qw_r x_r$

where f increases as $w_r \uparrow$

(We can rescale by $\frac{1}{q}$)

PTN, a new framework (cont.)

Steps (using a PQ rule):

1. For the r^{th} contributor to a cell, calculate $PT(r)$, $SN(r)$ and $N(r)$.
2. For every combination* of (t,s), calculate

$$S(t,s) = PT(t) - SN(s) - \sum_{r \neq t,s}^{n} N(r)$$

*Gray proposed a search-limiting algorithm

3. Identify the maximum value

$$S_{CELL} = \max\{S(t,s) | t \neq s\}$$

PTN, a new framework (cont.)

PTN permits multiple aspects of survey sampling to be represented.

PQ rule with weights, without or with waivers

PT(r)	***SN(r)***	***N(r)***
$px_r - f(x_r, w_r)$ no waiver	$q(w_r - 1)x_r$	$qw_r x_r$
0 if waiver		

where f increases as $w_r \uparrow$

Future directions

- Releasing v1.07 of G-Confid in October 2017
- Seeking innovative and user-centric approaches
  - to improve the assessment of sensitivity
    - to consider other aspects of survey sampling
  - to generate a more efficient suppression pattern
- Collaborating on methods and implementation
- Developing Random Tabular Adjustment
  - Stinner (Statistical Society of Canada, 2017)
  - Bayesian approach
  - Risk-utility assessment

For further information please contact

Pour plus d'information veuillez contacter

Peter.Wright2@Canada.ca

To order **G-Confid**, please contact

Pour commander **G-Confid** veuillez contacter

G-Confid@Statcan.gc.ca