Paper 3981-2019

Disclosure Control: Project Management Issues and Solutions

Peter Timusk, Centre for Special Business Projects, Statistics Canada

ABSTRACT

With many statistics, protecting the privacy of individual response data is legally mandated. Privacy of data is not a new topic in government statistics. In fact, much of the work involved in producing tables of statistical estimates can concern making sure data from individual respondents (be they businesses or persons) is kept confidential. The consideration of confidentiality needs to occur early in productions of tables of statistical estimates. The author explains some basics of statistical confidentiality calculations and explains why early consideration of confidentiality needs to happen in projects.

BRIEF INTRODUCTION TO CONFIDENTIALITY ANALYSIS

Information collected under the Statistics Canada Act cannot be disclosed in a way that exposes any one business's or any one respondent's data. A table of estimates, where one estimate is calculated from only one enterprise, would expose this enterprise's value, example their revenue. So aggregates are what is published, where there are more than a threshold number of respondents. An example might be that the minimum number of respondents must be greater than or equal to three.

DOMINANCE

At Statistics Canada dominance is identified with, our SAS based tool we call G-Confid and our in-house and secret rules in PROC SENSITIVITY. General software= G; Confidentiality = Confid. A big player dominates the estimate and can be guessed or other players in that estimate be guessed by subtracting the known dominant player. Known as dominance. Example: large public companies.

SECONDARY SUPPRESSION

- A cell in the tables needs to be suppressed (replaced with a letter 'X') if there is a low count or dominance.
- After suppression of a dominant or low count cell the linear nature of the tables of estimates may mean a second secondary suppression is still needed.
- Suppressing only one cell in a sum allows the other components of the sum to be subtracted from the total to reveal the supressed cell.
- Secondary suppression is calculated in G-Confid with an in-house macro that uses PROC OPTMODEL to minimize cost of suppression. It solves linear equations based on the table of estimates' dimensions.

Table 1

	Province	NL	PE	NS	NB	QC	ON	MB	SK	AB	BC	YT,NW,NU	Total
Industry													
11		10	78	86	127	7689	6790	1780	2499	1760	1430	3	22252
21		21	15	20	34	78	121	209	125	1209	276	650	2758
22		3	4	1	2	45	54	34	56	230	156	145	730
23		12	2	2	4	65	47	49	30	46	89	23	369
31		420	267	504	430	4029	4398	1987	2098	3879	1209	879	20100
44		816	390	2398	2398	18000	29909	2890	4002	2098	6789	2098	71788
Total		1282	756	3011	2995	29906	41319	6949	8810	9222	9949	3798	117997

Table 1. Example table with counts

Table 2

	Province	NL	PE	NS	NB	QC	ON	MB	SK	AB	BC	YT,NW,NU	Total
Industry													
11		10	78	86	127	7689	6790	1780	2499	1760	1430	3	22252
21		21	15	20	34	78	121	209	125	1209	276	650	2758
22		3	4	1	2	45	54	34	56	230	156	145	730
23		12	2	2	4	65	47	49	30	46	89	23	369
31		420	267	504	430	4029	4398	1987	2098	3879	1209	879	20100
44		816	390	2398	2398	18000	29909	2890	4002	2098	6789	2098	71788
Total		1282	756	3011	2995	29906	41319	6949	8810	9222	9949	3798	117997

Table 2. Example with counts showing low counts (red)

Table 3

	Province	NL	PE	NS	NB	QC	ON	MB	SK	AB	BC	YT,NW,NU	Total
Industry													
11		10	78	86	127	7689	6790	1780	2499	1760	1430	3	22252
21		21	15	20	34	78	121	209	125	1209	276	650	2758
22		3	4	X	X	45	54	34	56	230	156	145	730
23		12	X	X	4	65	47	49	30	46	89	23	369
31		420	267	504	430	4029	4398	1987	2098	3879	1209	879	20100
44		816	390	2398	2398	18000	29909	2890	4002	2098	6789	2098	71788
Total		1282	756	3011	2995	29906	41319	6949	8810	9222	9949	3798	117997

Table 3. Example with counts showing low counts and suppression less than 3 (red X)

Table 4

	Province	NL	PE	NS	NB	QC	ON	MB	SK	AB	BC	YT,NW,NU	Total
Industry													
11		10	78	86	127	7689	6790	1780	2499	1760	1430	3	22252
21		21	15	20	34	78	121	209	125	1209	276	650	2758
22		3	X	X	X	45	54	34	56	230	156	145	730
23		12	X	X	X	65	47	49	30	46	89	23	369
31		420	267	504	430	4029	4398	1987	2098	3879	1209	879	20100
44		816	390	2398	2398	18000	29909	2890	4002	2098	6789	2098	71788
Total		1282	756	3011	2995	29906	41319	6949	8810	9222	9949	3798	117997

Table 4. Example with counts showing suppression less than 3 (red X) and secondary suppression (brown X)

Table 5

	Regions	Atlantic	QC	ON	Prairies	BC and Territories	Total
Industry							
11		301	7689	6790	6039	1433	22252
21		90	78	121	1543	926	2758
22		10	45	54	320	301	730
23		20	65	47	125	112	369
31		1621	4029	4398	7964	2088	20100
44		6002	18000	29909	8990	8887	71788
Total		8044	29906	41319	24981	13747	117997

Table 5. Example with counts and aggregation or 'collapsing' of categories

Table 6Table 6. Example with counts and aggregation or 'collapsing' of categories and now counts are high enough (green)

	Regions	Atlantic	QC	ON	Prairies	BC and Territories	Total
Industry							
11		301	7689	6790	6039	1433	22252
21		90	78	121	1543	926	2758
22		10	45	54	320	301	730
23		20	65	47	125	112	369
31		1621	4029	4398	7964	2088	20100
44		6002	18000	29909	8990	8887	71788
Total		8044	29906	41319	24981	13747	117997

Table 6. Example with counts and aggregation or 'collapsing' of categories and now counts are high enough (green)

SUPPRESSION IS ONE APPROACH AND THERE ARE OTHERS

Suppression is used when publishing business statistics and survey results about businesses. The Canadian population census uses rounding of values. Making the data and results fuzzy by adding noise is also a protection of confidentiality method. In some publications of results, manipulation or weighting can cloud the exact data and allow enough protection.

G-CONFID A SAS BASED TOOL FOR HELP HERE.

We have a set of SAS macros and one SAS procedure we use in G-Confid. Here it helps locate the dominant, and low count, cells. Then the macros help identify the secondary suppression that is optimal in having the least amount of secondary suppression across the table while still protecting the sensitive data. There are various intermediate steps we can use to adjust the data or change the outcome pattern of suppression. We often need to transpose and re-code the survey data for input to the G-Confid system. We classify the data along the table dimensions; Prov='35'; NAICS2='11';, for example.

G-CONFID METHODOLOGISTS

Our internal teams like the G-Confid methodologists can support us in the tool use. The support they offer has to do with the use of our tool and how to conceptualize the relations between variables. Example: revenue is related to profit, so calculate the suppression pattern for both at the same time.

TABLE DESIGN AND PROJECT MANAGEMENT ISSUES

Counts in subdomains of dimensions need to be high enough to meet the thresholds. (Like variance and having a 'good' sample.) This is solved by designing the tables of estimates and the dimensions of interest before drawing a sample, so that under coverage is avoided and subdomain samples have enough respondents. While it may seem rather straightforward to add in more variables in a given analysis or table this may greatly increase the complexity and computational demands in using the G-Confid tools. The linear models become complex and unsolvable at some point when many dimensions are used with many levels of each dimension. Sometimes, after developing complex tabulations and suppression systems at a great cost, dimensions must be removed from the tables for less suppression of results. Often the G-Confid work is left to the last minute and then there is a pressure on the tool users. It is better to inform the clients early on of potential results and to keep them informed. Example: a simple frequency table performed at an early stage shows us low counts and would prevent time being wasted on developing and processing improper tables.

The understanding of the use of the tool can be demanding for some members of a given team. Describing and visualizing linear models or table dimensions can be confusing.

CONCLUSION

- G-Confid users need to be SAS programmers (or know how to use SAS programming).
- Recent concerns in some subject matters (inside and outside of governments) are increasing privacy awareness. Example personal data online.
- There is some academic study of 'disclosure control'; although this is a more obscure statistics subject.
- Some changes have been considered and are occurring at Statistics Canada concerning approaches used to protect data and still remain in accord with the Statistics Act.

ACKNOWLEDGMENTS

The author wishes to thank the G-Confid team for work over the years and his co-workers for review help with this paper. In particular, Alexander Davies helped a great deal.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Peter Timusk
Centre for Special Business Projects, Statistics Canada
peter.timusk@canada.ca