A MULTI-CONSTRAINT SAMPLE ALLOCATION FOR DETAILED NATIONAL ESTIMATES MADE TRICKIER BY PROVINCIAL BUY-INS

Darren Gray[1]

ABSTRACT

For the Survey of Innovation and Business Strategy (SIBS), the 2009 sample allocation was designed to satisfy precision requirements for three modules of questions, each of which targeted different industry and employment size group combinations at the national level. For the 2012 iteration of the survey, this already complex system of constraints was made more complicated when provincial specifications were added to the equation. A non-linear programming method was proposed to solve this complex problem. This paper will go over the methods used to ensure that all precision requirements were satisfied, while minimizing sample size and avoiding small stratum sizes.

KEY WORDS: Domain estimation, Non-linear programming, Precision requirements, Sample allocation.

RÉSUMÉ

Pour l'enquête sur l'innovation et les stratégies d'entreprise, l'allocation de l'échantillon en 2009 devait satisfaire les exigences de précision pour trois modules de questions ciblant chacun des combinaisons d'industries et groupes de taille d'emploi différentes au niveau national. Pour l'enquête de 2012, ce système de contraintes déjà complexe s'est complexifié davantage par l'ajout de demandes provinciales. Une méthode de programmation non-linéaire a été proposée pour résoudre ce problème. Cet article discutera des méthodes employées pour s'assurer que toutes les exigences soient satisfaites tout en minimisant la taille d'échantillon et en évitant les strates de petite taille.

MOTS CLÉS : Allocation d'échantillon; estimation par domaine; programmation non-linéaire; exigences de précision.

1. INTRODUCTION

1.1 Description of the Problem

The Survey of Innovation and Business Strategy (SIBS) is an enterprise-level survey designed to provide useful statistical information on strategic decisions, innovation activities and operational tactics used by Canadian firms. As the survey was developed for several organizations, the questionnaire consists of three distinct themes. Each theme corresponds to a subset of the questionnaire referred to as modules: the CORE module (business strategy) the GVC module (global value chains) and the SOI module (Survey of Innovation). The survey targets enterprises across fourteen sectors, with the bulk of interest focused within manufacturing. Only enterprises with at least 20 employees and $250,000 in revenue are targeted, resulting in a 2012 target population of 67,807 units.

For each module, a set of national-level domains were defined based on industry groups (using NAICS[2] classification codes) and employment groups (small, medium and large). For each domain a targeted precision (standard error) was pre-specified for proportion estimates. Both the domains and precision targets varied by module.

The first iteration of the survey was conducted in 2009. The sample was designed to meet the precision targets of each domain, resulting in a stratified simple random sample of 6,233 units. For the 2012 survey, certain provincial and regional organizations expressed interest in producing estimates on proportions at a geographic level, requiring a revised sample

[1] Darren Gray, Statistics Canada, Tunney's Pasture, Ottawa, ON K1A 0T6, darren.gray@statcan.gc.ca
[2] North American Industry Classification System

design. In this paper we will describe both the original and revised sample designs, and examine the effect the change had on the selected sample.

2. METHODOLOGY

2.1 Initial Sampling Strategy

In the early design stages, the 2012 iteration of SIBS was to follow the stratified simple random sampling strategy from 2009. Each of the three modules (CORE, GVC and SOI) required national estimates for different industrial groupings and at different levels of precision, depending on industry sector and enterprise size. These requirements are summarized in Table 1, given in terms of maximum standard error (SE) on estimated proportions.

Table 1 - Precision requirements by module

Module	Number of industrial groupings	Precision targets within industrial groupings	Total number of targeted domains[3]
CORE	64	- Census for large enterprises - 10% SE for small and medium enterprises combined	126
GVC	34	- 8% SE within manufacturing sector, by enterprise size - 10% SE elsewhere (all sizes combined)	90
SOI	44	- 8% SE within manufacturing sector - 10% SE elsewhere	44

An industry stratification variable was derived by creating non-overlapping groups of NAICS codes that could be used to derive the different industrial groupings that were required for each module. The target population was also stratified by enterprise size, defined as small (20-99 employees), medium (100-249 employees) and large (250 employees or more). This stratification allowed the flexibility to control the sample size at the level of detail required to support targeted precision for each module simultaneously.

Because of the complexity of the precision requirements, standard sampling tools available at Statistics Canada (such as the generalized sampling tool "GSAM") were not equipped to determine sample size and allocation. Instead, the following strategy was implemented:

- For each module m and domain d, a module domain sample size $n_{m,d}$ was calculated. Re-arranging equation 2.16 given by Lohr (1999), we can estimate the maximum domain sample size required to meet the domain precision requirements $SE_{m,d}$ (on a proportion of 50%) for domain size $N_{m,d}$ as follows:

$$n_{m,d} = \frac{4\,SE_{m,d}^2 + 1}{4\,SE_{m,d}^2 + {}^1/_{N_{m,d}}}$$

- The resulting domain sample size $n_{m,d}$ was then allocated proportionally to each contributing strata[4]. For each stratum h we thus obtained three sample sizes ($n_{CORE,h}$, $n_{GVC,h}$, $n_{SOI,h}$) – one for each module.
- Final stratum sample size n_h was then set to the maximum of these three values, and subsequently augmented to account for an estimated 50% response rate and a minimum sample size of 5 within each stratum.

Although not an optimized solution, this strategy provided a quick and simple allocation method expected to ensure all the precision requirements were met, and had proved effective in 2009. Following this strategy resulted in a preliminary sample size of 6,074 units.

[3] For some domains, the set of units within the targeted industrial grouping and employment size categories was empty – these were not counted amongst the total number of targeted domains.

[4] For a given module, each stratum contributed to no more than one targeted domain.

2.2 Addition of Geographic Domains

During the planning process, certain organizations expressed interest in obtaining estimates on proportions at the provincial and regional levels, for various industry and enterprise size domains. As the initial sampling strategy contained no geographic component, the resulting sample was not expected to produce precise estimates at the level of detail requested. To meet these requirements, three options were investigated:

- Selecting a supplementary (dual) sample and constructing an appropriate composite estimator;
- Incorporating the regions and provinces into the existing stratification, and following the initial strategy; and
- Relying solely on domain estimation by augmenting the original sample.

Each option came with complications. The first would require a complete redesign of the estimation process, and was deemed unfeasible considering time and budget constraints. The second resulted in a large number of very small strata; subsequently, the minimum stratum sample size requirement resulted in an impractical sample size. The same problem (large sample size) ruled out the third option as well. In the end, a combination of the second and third options, along with a completely revamped allocation strategy, was adopted.

2.3 Revised Sampling Strategy

Maintaining national estimates for the three survey modules remained a priority, and in light of this, provincial and regional estimates were only proposed for industry and size domains that would not require any additional stratification of the existing industry and size groupings. Based on a costing analysis[5], a standard error of 12% (on proportions of 50%) was targeted for provincial and regional domains at the NAICS 3 level, with no size consideration.

Four regions (Atlantic Canada, Quebec, Ontario and Alberta) opted to participate. The sampling design now had to satisfy precision requirements for a total of 385 overlapping domains, as shown in Table 2.

Table 2 – Revised precision requirements

Domain set	Number of industrial groupings	Precision targets (maximum standard error) within industrial groupings	Total number of targeted domains
CORE module	64	- Census for large enterprises - 10% for small and medium enterprises combined	126
GVC module	34	- 8% within manufacturing sector, by enterprise size - 10% elsewhere (all sizes combined)	90
SOI module	44	- 8% within manufacturing sector - 10% elsewhere	44
Geographic	42	- 12% by region	125

These new regions were incorporated into stratification. To decrease the resulting number of small strata (which can be problematic depending on non-response), certain strata were collapsed together with respect to enterprise size and geography. As a result, the allocation strategy now had to take into account domain estimation when considering precision requirements.

[5] For costing purposes, expected sample counts for each region were calculated from the initial sampling design.

2.4 Revised Allocation Strategy

The goal of the allocation strategy was to minimize overall sample size while meeting the precision requirements of all 385 domains of interest. The procedure follows that given by Demnati and Turmelle (2011) for Statistics Canada's Integrated Business Statistics Program. For each domain *d*, we begin by assuming a value of 50% for the proportions of interest, thereby ensuring that sampling variance V_d is at a maximum. For stratum *h*, let N_h and n_h be the stratum population and sample sizes respectively, let N_d be the domain population size and $N_{h,d}$ the number of units in domain *d* contained in stratum *h*. Then the problem reduces to minimizing overall sample size $\sum n_h$ with respect to the set of domain variance constraints $V_d \leq SE_d^2$ with SE_d the standard errors specified in Table 2 and

$$V_d = \sum_h N_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h}\right) S_{h,d}^2$$

$$S_{h,d}^2 = \frac{1}{(N_h - 1)} \frac{N_{h,d}}{4N_d^2}$$

To solve for the values n_h we re-write the above equation as

$$V_d = v_{0,d} + \sum_h \frac{v_{h,d}}{n_h}$$

where $v_{0,d} = -\sum_h N_h S_{h,d}^2$ and $v_{h,d} = N_h^2 S_{h,d}^2$.

For certain domains, precision demands required a census. Some small domains were also designated for a census, to avoid issues relating to small respondent counts. Any strata contributing to these domains were set as take-all, i.e. $n_h = N_h$. The remaining n_h were calculated using a non-linear programming solution (employing the Newton-Raphson method) so as to simultaneously minimize overall sample size and meet or exceed the precision targets for all domains.

The resulting values were increased to account for non-response and a minimum stratum sample size requirement of 5 units. For certain strata, this increase was not possible due to stratum population size. To investigate the effect of this limitation, expected standard error was calculated for each domain, based on a response rate of 50% and sampled proportion of 50%. For domains whose expected standard error fell short of the precision targets, contributing strata were set as take-all. The result was that for every domain, either a census was taken or the expected standard error met the precision targets of each module.

3. RESULTS

The revised sampling design and allocation was successful in meeting the sampling objectives. The design is expected to meet or exceed the new geographic domain precision targets (where possible) while maintaining the expected precision of the national domains. Additionally, the basic structure of the design (stratified simple random sampling) stayed consistent, minimizing the amount of work required during estimation and table production. The revised sample design resulted in a sample size of 7,818 units, an increase of 1,744. A detailed comparison of the two designs can be found in Table 3.

The effect of the revised sampling strategy on the regional samples is included in Table 4. Although domain estimates were added for only four regions (Atlantic Canada, Quebec, Ontario and Alberta), it was expected that the new sample design and allocation method would result in an increased sample size in the rest of Canada. It was hoped this increase would be minimal; in the end it accounted for less than 10% (162 of 1,744) of the total sample increase.

Table 3 Design comparison

	Number of targeted domains	Number of strata created	Sampled units		
			Take-all	Take-some	Total
Initial design (national estimates only)	260	320 (189 take-some)	1,676	4,398	6,074
Revised design (including geographic-level estimates)	385	900 (417 take-some)	2,514	5,304	7,818
Difference	125	580	838	906	1,744

Table 4 Regional Oversampling

	Atlantic Canada	Quebec	Ontario	Alberta	Rest of Canada	Total
Population	4,137	17,290	25,197	7,985	13,198	67,807
Initial sample[6]	311	1,643	2,548	593	979	6,074
Revised sample	787	2,073	2,714	1,103	1,141	7,818
Sample added	476	430	166	510	162	1,744

4. CONCLUDING REMARKS

The effectiveness of the new sampling design won't be fully known until the results of the 2012 survey have been collected and thoroughly analyzed. As with any survey, the precision of produced estimates will depend on a variety of factors. In particular, non-response will play an important role, and a response rate significantly below the expected 50% (within a given domain) could severely affect standard errors.

Efforts will be made to improve upon the sampling allocation approach in any future iteration of SIBS. Although the approach met the sampling objectives, the number of take-all units was relatively high, which can lead to high response burden for enterprises, especially in surveys repeated over time. In light of this, further investigation into the collapsing of strata should be investigated, to see if this could lead to a decrease in the number of take-all units while maintaining the expected precision in all domains of interest.

REFERENCES

Lohr, S.L. (1999). *Sampling: Design and Analysis*. Pacific Grove: Duxbury Press.

Demnati, A. & Turmelle, C. (2011). "Proposed Sampling and Estimation Methodology for the Integrated Business Statistics Program". *Advisory Committee on Statistical Methods*, Statistics Canada

[6] A sample was never drawn using the initial sample design. The regional figures are based on expected counts.