Sample Coordination and Response Burden for Business Surveys: Methodology and Practice of the Procedure Implemented at INSEE

Emmanuel Gros[1] (emmanuel.gros@insee.fr), Ronan Le Gleut[1] (ronan.le-gleut@insee.fr)

Keywords: Sampling coordination, permanent random numbers, coordination function, response burden, stratified sampling.

1. Introduction

The public statistical system carries out each year a significant number of businesses and establishments surveys. The objective of the negative coordination of samples is to foster, when selecting a sample, the selection of businesses that have not already been selected in recent surveys, while preserving the unbiasedness of the samples. This coordination contributes to reduce the statistical burden of small businesses – large businesses, from a certain threshold, are systematically surveyed in most surveys.

This paper presents the new sampling coordination method currently used at Insee. This method, using Permanent Random Numbers (PRN) assigned to each unit, is based on the notion of coordination function, defined for each unit and each new drawing, which transforms permanent random numbers.

2. Methods

We present here the main principles of the method, detailed in [1], limited to the case of stratified simple random sampling. This method was proposed by C. Hesse in 2001 in [2], and studied by P. Ardilly in 2009 in [3].

2.1. A PRN method resting on the concept of coordination functions

The concept of coordination function plays an essential role in the method.

> A coordination function g is a measurable function from [0,1] onto itself, which preserves uniform probability: if P is the uniform probability on [0,1], then the image probability P^g is P. It means that for any interval I = [a, b[included in [0,1] :
>
> $$P\left[g^{-1}(I)\right] \overset{\text{def}}{=} P^g(I) = P(I) = b - a$$
>
> The length of the inverse image of any interval under g equals the length of this interval: a coordination function preserves the length of intervals – or union of intervals – by inverse image.

Each unit k of the population is given a permanent random number ω_k, drawn according to the uniform distribution on the interval [0,1[. The drawings of the ω_k are mutually independent.

We consider a sequence of surveys t = 1, 2,…(t refers to the date and the number of the survey), and we denote by S_t the sample corresponding to survey t. Suppose that one has defined for each unit k a “wisely chosen” coordination function (see 2.2.) $g_{k,t}$ which changes at each survey t.

[1] INSEE – France's National Institute for Statistics and Economic Studies – 18 boulevard Adolphe Pinard, 75675 Paris Cedex 14, FRANCE

The drawing of the sample S_t by stratified simple random sampling is done by selecting, within each stratum (h,t) of size $N_{(h,t)}$, the $n_{(h,t)}$ units associated with the $n_{(h,t)}$ smallest numbers $g_{k,t}(\omega_k),\ k = 1 \ldots N_{(h,t)}$.

Proof

> The $N_{(h,t)}$ random numbers (ω_k) associated to the $N_{(h,t)}$ units of the stratum have been independently selected according to the uniform probability on [0,1], denoted P. Since we have $P^{g_{k,t}} = P$ for each k, the N numbers $g_{k,t}(\omega_k)$ are also independently selected according to P. Then, using a well-known result, the $n_{(h,t)}$ smallest values $g_{k,t}(\omega_k)$ give a simple random sample of size $n_{(h,t)}$ in the stratum.

2.2. Construction of a coordination function from the cumulative response burden

❶ Response burden and coordination function: Let Ω denote the vector of random numbers ω_k given to the population units k, and $c_{k,t}$ be the response burden of a questioned business k at survey t. The cumulative burden for unit k is a random variable, function of Ω, equal to:

$$C_{k,t}(\Omega) = \sum_{u \le t} c_{k,u} . I_{k \in S_u}(\Omega) \quad (1)$$

We wish to define, for each unit k, a coordination function $g_{k,t}$ based on $C_{k,t-1}$, the cumulative burden of unit k until survey t-1. To meet the objective of negative coordination – to draw as a priority, for a given sample selection, units that have had the lowest response burden during the recent period – and taking into account the selection scheme of the units – the higher the probability for the unit to be selected the smaller the number $g_{k,t}(\omega_k)$ –, a desirable property for any coordination function is the following:

$$C_{k,t-1}(\Omega^{(1)}) < C_{k,t-1}(\Omega^{(2)}) \Rightarrow g_{k,t}(\omega_k^{(1)}) \le g_{k,t}(\omega_k^{(2)})$$

where $\omega_k^{(i)}$ (i=1,2) denotes the k^{th} component of vector $\Omega^{(\mathbf{i})}$. This condition is not easy to handle, because the function $C_{k,t-1}(\Omega)$ is a function of vector Ω: it depends not only on the random number ω_k given to unit k, but on all the other random numbers $\omega_1 \ldots \omega_N$. We will see on ❷ how we can replace this function by a function $C^e_{k,t-1}(\omega_k)$ which depends only on ω_k. The desirable property for any coordination function $g_{k,t}$ will become :

$$C^e_{k,t\text{-}1}(\omega_k^{(1)}) < C^e_{k,t\text{-}1}(\omega_k^{(2)}) \Rightarrow g_{k,t}(\omega_k^{(1)}) \le g_{k,t}(\omega_k^{(2)}) \quad (2)$$

❷ The expected cumulative response burden: We need to replace the cumulative burden function $C_{k,t}$, function of vector Ω, by an approximate cumulative burden function $C^e_{k,t}$, which should be a function of ω_k close to $C_{k,t}$. The best approximation of the indicator function $I_{k \in S_u}(\Omega)$ depending only on ω_k, in the L2-norm sense, is its conditional expectation given ω_k:

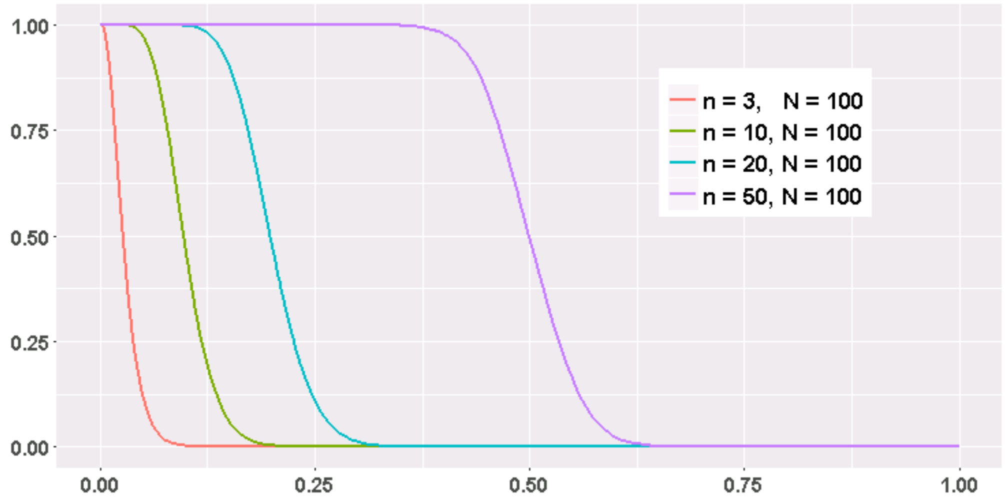
$$I^a_{k \in S_u}(\omega) = E\left(I_{k \in S_u}(\Omega) \mid \omega_k = \omega\right) = P\left(k \in S_u \mid \omega_k = \omega\right)$$

If we suppose that the coordination functions are bijective[2] functions, we can write:

$$I^a_{k \in S_u}(\omega) = P\left(k \in S_u \mid g_{k,u}(\omega_k) = g_{k,u}(\omega\ )\right) = b_{k,u}\left(g_{k,u}(\omega\ )\right)$$

[2] This property is satisfied with the method described here, but it is not an intrinsic property of a coordination function.

where $1 - b_{k,u}(x)$ is the cumulative distribution function of a beta distribution with parameters n and N-n. The next graph shows the shape of the b(x) function for some values of n and N.



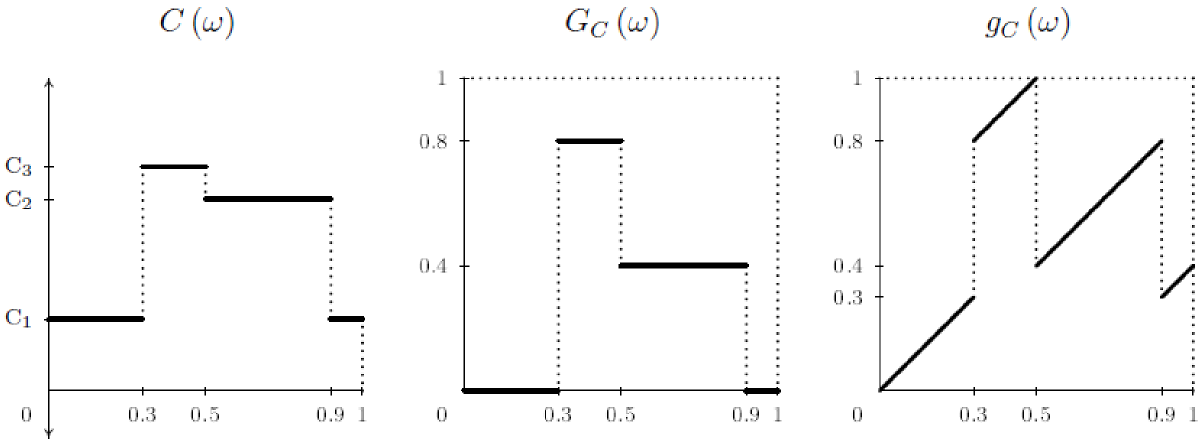
This function is then approximated by a step function (see <u>❹</u>) to construct an approximate expected cumulative burden function from formula (1):

$$C^{e}_{k,t}(\omega) = \sum_{u \leq t} c_{k,u} . I^{a}_{k \in S_u}(\omega)$$

<u>❸ Construction of a coordination function:</u> For the sake of simplicity, we omit the subscripts k and t. Let us define the function $G_C = F_C(C)$, with F_C the cumulative distribution function of C:

$$\forall \omega \in [0,1], G_C(\omega) = P\left(u \middle| C(u) < C(\omega)\right)$$

We can show that the range of G_C is included in [0,1], and that G_C satisfies (2), but is not a coordination function if C has "levels", that is subsets of [0,1] where C is constant (G_C has then the same levels). However, we can construct a bijective coordination function on [0,1] g_C equal to G_C outside the levels and composed of line segments having a slope equal to 1 on the levels of G_C, as illustrated in the next figure, where C is a step function, with 4 levels:



❹ Approximation by step functions: As mentioned in ❷, we need to approximate the indicator functions $I^{a}_{k\in S_u}=b_{k,u}$ by step functions that can be easily "computed". We simplify the shape of this function as follow:

- We divide the interval [0,1[into L equal subintervals $I_\ell=\left[\frac{\ell-1}{L};\frac{\ell}{L}\right[$, ℓ = 1…L[3]

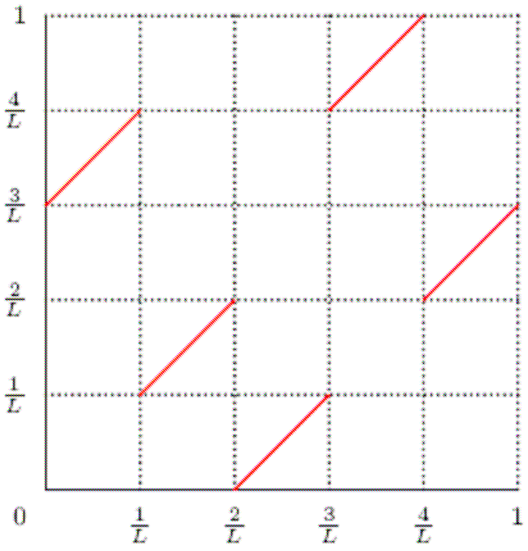
- We replace the approximate indicator function $b_{k,u}$ by a piecewise linear function $\tilde{b}_{k,u}$ which takes the same values as $b_{k,u}$ at the endpoints of the intervals I_ℓ.

- We compute the the average value $\beta_{k,u}(\ell)$ of $\tilde{b}_{k,u}$ on each interval I_ℓ.

- We define the function $\beta_{k,u}$ as : $\forall\omega\in I_\ell\quad \beta_{k,u}(\omega)=\beta_{k,u}(\ell)$. $\beta_{k,u}$ is an approximation of the approximate indicator function $I^{a}_{k\in S_u}$ by a piecewise constant function.

Finally, the cumulative burden function $C_{k,t}$ is replaced by the approximated expected cumulative burden function $C^{e}_{k,t}(\omega)=\sum_{u\le t}c_{k,u}.\beta_{k,u}\left(g_{k,u}(\omega\ )\right)$, which is a step function, constant on each I_ℓ. So we are in the same context as in the example presented in ❸: from the function $C^{e}_{k,t}(\omega)$, we construct a "G" function, also constant on each I_ℓ, and then a coordination function g which looks like in the opposite example with L=5. It is entirely defined by a permutation σ on {1,2,3...,L}, according to the following formula:

$$\forall\,\omega\in\left[\frac{\ell-1}{L};\frac{\ell}{L}\right[\qquad g_\sigma(\omega)=\frac{\sigma(\ell)-1}{L}+(\omega-\frac{\ell-1}{L})$$



The only remaining issue is so the definition of the permutation σ. To do this, we go back to the fundamental property (2) of the coordination function: the smaller is the criterion (here the cumulative response burden), the smaller is the value of the coordination function g_σ. Now, on $\left[\frac{\ell-1}{L};\frac{\ell}{L}\right[$, the smaller is $\sigma(\ell)$, the smaller is g_σ. So, we will arrange the values $\sigma(\ell)$ exactly in the same order as the values of the approximate expected cumulative burden function $C^{e}_{k,t}(\ell)$:

$$C^{e}_{k,t}(\ell_1)\le C^{e}_{k,t}(\ell_2)\le\ldots\le C^{e}_{k,t}(\ell_L)\Leftrightarrow\sigma(\ell_1)\le\sigma(\ell_2)\le\ldots\le\sigma(\ell_L)$$

where ℓ_i is the identifier of the ith standardized interval.

[3] L being a "large enough" integer (at least greater than 50).

As σ has to be a permutation, and therefore bijective, we add the following additional constraint: if $C^{e}_{k,t}(p) = C^{e}_{k,t}(q)$ and $p < q$, then $\sigma(p) < \sigma(q)$. *In fine*, this means imposing strict inequalities in the ranking of the $\sigma(\ell)$, which leads to $\sigma(\ell_i) = i$ and completely defines the permutation σ.

2.3. Sample coordination between surveys based on different kind of units

The methods allows the coordination of samples relating to surveys based on different kind of units, for example legal units and local units. This "multi-level" coordination is obtained by defining a permanent link between the legal unit and one of its local units – the head office for example – and by assigning to this "principal local unit" the same permanent random number as the legal unit – the PRN of other local units being drawn according to the uniform distribution on the interval [0,1[. So, the response burden of principal local units can be taken into account in the cumulative response burden of legal units for the drawing of legal units samples, and reciprocally, the response burden of legal units can be taken into account in the cumulative response burden of principal local units for the drawing of local units samples.

3. RESULTS

A simulation study has been conduct to assess the properties of this coordination method. 20 legal units samples and 8 local units samples have been drawn with the multi-level procedure describe in 2., each sample being coordinated with the whole of past samples, with $c_{k,t}$=1 for all units k and all samples t. We compare the results, in terms of distribution of legal units response burden, with those, on the one hand of a sequence of 28 independent drawings, and on the other hand of the "level by level" coordinated drawing of the 20 legal units samples and independently the coordinated drawing of the 8 local units samples. The following table shows the high efficiency of the multi-level coordination procedure.

Cumulative response burden of legal units, except take-all stratum	Frequency according to the sampling scheme			Differences between drawings:		
	Independant drawings	"Level by level" coordinated drawings	Multi-level coordinated drawings	Independant *versus* "level by level" coordinated	"level by level" *versus* multi-level coordinated	Independant versus multi-level coordinated
0	4 670 676	4 651 954	4 634 250	-18 722	-17 704	-36 426
1	410 016	439 355	474 286	29 339	34 931	64 270
2	40 095	34 824	18 230	-5 271	-16 594	-21 865
3	8 072	4 679	4 125	-3 393	-554	-3 947
4	2 142	813	737	-1 329	-76	-1 405
5	578	93	92	-485	-1	-486
6	121	5	2	-116	-3	-119
7	20	0	1	-20	1	-19
8	3	0	0	-3	0	-3

4. CONCLUSIONS

The sampling coordination method presented in this paper proves, via many simulations studies conducted on simulated as well as real data, to be very efficient – providing significant gains in terms of response burden allocation over the population units – as well as outstandingly robust vis-à-vis sampling design parameters. It is used operationally at INSEE since the end of 2013.

REFERENCES

[1] Guggemos, F. And Sautory, O. (2012). *Sampling Coordination of Business Surveys Conducted by INSEE*. ICES-IV, Montréal.

[2] Hesse, C. (2001). *Généralisation des tirages aléatoires à numéros aléatoires permanents, ou la méthode JALES+*. INSEE working paper E0101.

[3] Ardilly, P. (2009). *Présentation de la méthode JALES+ conçue par Christian Hesse*. Internal INSEE working paper.