Sampling coordination of business surveys conducted by INSEE

Fabien Guggemos, Olivier Sautory, INSEE, Direction des statistiques d'entreprises

Summary

A new method of coordination, which generalizes the current technique based on permutations of random numbers, is experimented at Insee. The random number is replaced by a "coordination function", which is a function which transforms the random numbers, and has the characteristic of preserving uniform probability. This function changes with each selection, depending of the desired type of coordination : negative coordination for separate samples, controlling of the rotation rate when updating panels. This method takes into account the cumulative response burden over several samples, and can be used with Poisson sampling and stratified simple random sampling.

Several national statistics agencies use the *permanent random numbers* technique for the sampling coordination of business surveys. Each unit k of the population (including the new units) is independently assigned a number ω_k, selected according to the uniform distribution in the interval [0,1[. In *Poisson sampling*, we define a starting point d, between 0 and 1, and we select the units whose number ω_k belongs to the interval [d, d + π_k[, where π_k is the probability of inclusion of the unit k. In *simple random sampling* (SRS) of size n (without replacement), we define a starting point d, between 0 and 1, and we select the n units with the lowest ω_k superior to d. For a *stratified simple random sampling* (SSRS), we define a starting point d_h in each stratum h.

The *constant shift method* can be used when we have J panels indexed with j = 1…J. To ensure that there is some degree of separation between these panels, we select different starting points d_j which we shift periodically by the same quantity c. The starting point for panel j at the date a is therefore :

$$d_{j,a} = d_{j,1} + (a-1)\,c,\ a \geq 1$$

1. Definition of a coordination function

A coordination function g is a measurable function from [0, 1[to 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[: $P\left[g^{-1}(I)\right] \overset{\text{def}}{=} P^g(I) = P(I) = b - a$

Each unit k in the sampling frame is given a permanent random number ω_k, and a coordination function that changes at each sampling : $g_{k,t}$ is the coordination function for sampling t = 1, 2, …The selection of the units is performed in the following way (we omit the index t) :

- Poisson sampling

We select the units k such that $g_k(\omega_k) \in [0, \pi_k[$, where π_k is the probability of inclusion of the unit k. We have : $P(k \in \text{Sample}) = P(g_k(\omega_k) \in [0,\pi_k[) = P^{g_k}([0,\pi_k[) = P(\omega_k \in [0,\pi_k[) = \pi_k$, and the drawings are independent.

- SSRS

Within a stratum, we select the n units k associated with the n smallest numbers $g_k(\omega_k)$. Since the n numbers (ω_k) are drawn independently from the uniform distribution on [0, 1[, and since $P^{g_k} = P$, then the n numbers $g_k(\omega_k)$ are drawn independently from the uniform distribution on [0, 1[, and the n smallest numbers $g_k(\omega_k)$ give a simple random sample of size n in the stratum.

Example 1 : the constant shift method

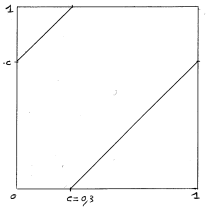
Let $d_1 = 0$ et $d_2 = c$.
We can define the coordination functions in the following way : $\forall k \quad g_{k,1}(\omega) = \omega \qquad g_{k,2}(\omega) = \omega - c \pmod{1}$

Then we have :

$k \in S_1 \Leftrightarrow \omega_k \in [0, \pi_{k,1}[\;\Leftrightarrow\; g_{k,1}(\omega_k) \in [0, \pi_{k,1}[$

$k \in S_2 \Leftrightarrow \omega_k \in [c, c + \pi_{k,2}[\;\Leftrightarrow\; g_{k,2}(\omega_k) \in [0, \pi_{k,2}[$

For c=0.3, $g_{k,2}$ has the following shape :



Example 2 : a method which controls the rotation rate

Let us assume that we select, then update, a panel without considering negative coordination with other samples. In this case, we can establish the rotation rate individually, to the extent permitted by the probabilities of inclusion. Let t be the index of a sample resulting from an update. And let $\pi_{k,t}$ be the probability of inclusion in that sample. The rotation rate $r_{k,t}$ between samples t-1 and t is

$$r_{k,t} = P(k \notin S_t | k \in S_{t-1})$$

If we perform Poisson sampling, we can choose for $r_{k,t}$ any value in the interval :

$$\left[\mathrm{Max}(0, \frac{\pi_{k,t-1} - \pi_{k,t}}{\pi_{k,t-1}}) \,;\, \mathrm{Min}(1, \frac{1 - \pi_{k,t}}{\pi_{k,t-1}}) \right]$$

We use the following coordination functions :

$g_{k,1}(\omega_k) = \omega_k$

$g_{k,2}(\omega_k) = \omega_k - r_{k,2}.\pi_{k,1}$

Then the rotation rate is : $r_{k,2}.\pi_{k,1} / \pi_{k,1} = r_{k,2}$

More generally, at step t, the coordination function is : $g_{k,t}(\omega_k) = \omega_k - \sum_{u=2}^{t} r_{k,u}.\pi_{k,u}$

2. A step by step procedure reflecting response burdens

The general idea is to choose, as a priority for each selection, the units that have had the lowest response burden during the recent period. With this method, a response burden is defined for each unit, corresponding to each survey. Each sampling consists in selecting the units in each stratum that have the lowest cumulative burdens. If the subsequent selections used the same stratification, this method would not pose a problem, and would be trivial. The problem is that the stratification is subject to change from on selection to the next. We note in this case that permutations by increasing cumulative burdens should occur within intersections of strata, so that unbiased, or at least consistent, estimators can be obtained. For this reason, a solution is to define every year "building blocks", i.e. fixed strata intersections from which all strata for the year can be established through a union. Another problem occurs when units change strata in connection with annual updates.

We shall implement the same idea, but without establishing building blocks, using the coordination function. Samples are selected in a certain order, indexed with t. This operation involves either separate sampling, or the first selection of a panel, or updating of a panel.

$\boldsymbol{\omega}$ = (… ω_k…) = vector of random numbers given to the population units k.

Let $I_{k,t}(\boldsymbol{\omega})$ be an indicator function, equal to 1 if the values in $\boldsymbol{\omega}$ lead to select the unit k in the sampling t, and 0 otherwise :

$k \in S_t \Leftrightarrow I_{k,t}(\boldsymbol{\omega}) = 1$ (the inclusion of k in S_t depends only on the vector $\boldsymbol{\omega}$)

Let $\gamma_{k,t}$ be the response burden of a questioned enterprise k at survey t.

The effective burden is a random variable $\gamma_{k,t}(\boldsymbol{\omega}) = \gamma_{k,t}\, I_{k,t}(\boldsymbol{\omega})$

The cumulative burden for unit k is a function of $\boldsymbol{\omega}$ equal to $\Gamma_{k,t}(\boldsymbol{\omega}) = \sum_{u \le t} \gamma_{k,u} . I_{k,u}(\boldsymbol{\omega})$

We wish to define the coordination function $g_{k,t}$ for the selection of sample S_t using $\Gamma_{k,t-1}$. In particular, for a separate sample, for any k, we want that :

$$\Gamma_{k,t-1}(\boldsymbol{\omega_1}) < \Gamma_{k,t-1}(\boldsymbol{\omega_2}) \Rightarrow g_{k,t}(\omega_{k,1}) < g_{k,t}(\omega_{k,2})$$

We meet two difficulties :

1. to substitute for $\Gamma_{k,t}(\boldsymbol{\omega})$ a function of ω_k only, denoted $\Gamma'_{k,t}(\omega_k)$, that closely approximates $\Gamma_{k,t}(\boldsymbol{\omega})$.

For a Poisson sampling : $I_{k,t}(\boldsymbol{\omega})$ depends only on ω_k (indicator function of an interval of length π_k), and can be denoted $I_{k,t}(\omega_k)$. Therefore, $\Gamma_{k,t}(\boldsymbol{\omega})$ depends only on ω_k , and can be denoted $\Gamma'_{k,t}(\omega_k)$.

For a SSRS, $I_{k,t}(\boldsymbol{\omega})$ depends on all coordinates of the vector $\boldsymbol{\omega}$, but "primarily" on coordinate ω_k : if we select the ***n*** units among ***N*** with the ***n*** smallest values ω_j, it will be equal to 1 for values of ω_k near to 0, regardless of the values of the other coordinates.
Accordingly, we will be able to replace $I_{k,t}(\boldsymbol{\omega})$ with an approximation $I'_{k,t}(\omega_k)$ which depends only on ω_k, and therefore to replace $\Gamma_{k,t}(\boldsymbol{\omega})$ with an approximation $\Gamma'_{k,t}(\omega_k)$ which depends only on ω_k.

2. So we get a one-dimension problem : to define the coordination function $g_{k,t}$ such that :

$$\Gamma'_{k,t-1}(\omega_{k,1}) < \Gamma'_{k,t-1}(\omega_{k,2}) \Rightarrow g_{k,t}(\omega_{k,1}) < g_{k,t}(\omega_{k,2})$$

Thus an enterprise having a smaller cumulative burden will have a smaller number $g_{k,t}(\omega_k)$ and a greater probability of being selected.

If sample t is a panel update, the problem is somewhat more complicated as it involves, in addition to $\Gamma_{k,t-1}$, the coordination function $g_{k,u}$, where u is the index for updating the previous panel. In other words, several criteria are involved in determining $g_{k,t}$.

The solution to this type of problem is provided in the following section.

3. Construction of a coordination function

Let us consider the problem theoretically, and somewhat more generally in this section.

Let $C_{k,t}$ (ω_k) be a criterion such that the smaller is the criterion, the larger is the probability of selection for unit k at sampling t.
We drop the subscripts k and t. So ω is now a simple real number.
C is supposed to be a bounded measurable function : $\omega \in [0,1[\to C(\omega) \in IR$

We wish to associate to this criterion a coordination function g such that :

$$C(\omega_1) < C(\omega_2) \Rightarrow g(\omega_1) < g(\omega_2) \quad (1)$$

Let P^C be the image probability of P under C, F_C the distribution function of C. The coordination function is built from $G_C = F_C(C)$. Considering the definitions of P^C and F_C, we can write :

$$G_C(\omega) = P^C(]-\infty, C(\omega)[) = P(C^{-1}]-\infty, C(\omega)[) = P(u | C(u) < C(\omega))$$

The way to derive g from G_C depends on whether or C has *levels*.

<u>Definition</u> : we call a *level* of criterion C any inverse image of a real number y such that $P^C(y) = P(A) > 0$
In other words, C has *levels* when horizontal line segments form part of the graph of C.

<u>Properties of G_C</u>

- The range of function G_C is in [0,1[
- G_C has the same *levels* as C
- G_C verifies implication (1)
- For every y in the range of G_C, we have $P(u|G_C(u) < y) = y$
- If C has no *level*, G_C is a coordination function.

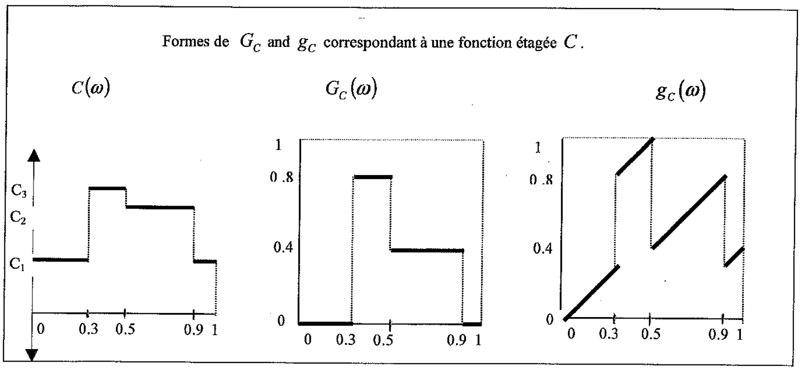
<u>Example 1</u>
C is a strictly increasing function. Then $G_C(\omega) = \omega$ is a coordination function.

<u>Example 2</u>
$C(\omega) = \omega\,(1-\omega)$. We find :
$F_C(x) = 1 - 2\,(0.25 - x)^{1/2}$

$$
\begin{aligned}
G_C(\omega) &= 2\,\omega && \text{if } \omega \in [0, 0.5[\\
&= 2 - 2\,\omega && \text{if } \omega \in [0.5, 1[
\end{aligned}
$$

If C has at least one level, the range of function G_C is strictly included in [0,1[. We have to deduce from G_C another function, denoted g_C, such that the range of g_C is equal to [0,1[. This can be obtained in the following way.
Let A be a level of C (and G_C) : if $\omega \in A$, $G_C(\omega)$ = constant = y. We denote B the largest interval [y,t[such that $G_C^{-1}(B) = A$.
The probability $P^G(B)$ is concentrated at point y, and we have P(B) = P(A). We obtain a uniform distribution of this probability in B by adding to $G_C(\omega)$ the linear function with slope 1 which transforms A into B.
If A = [a,b[, for any $\omega \in A$, we define $g_C(\omega) = \omega - y + a$.
More generally, if we have several levels A_i : $\quad g_C(\omega) = G_C(\omega) + \sum_i 1_{A_i}(\omega) \cdot \int 1_{A_i \cap [0,\omega]}(u)\,du \quad (2)$

<u>Example 3</u>
C is a step function, i.e. the function C contains only levels A_i, i=1…I. Function G_C has exactly the same levels. Values $y_i = G_C(A_i)$ are ranked in the same order as values $x_i = C(A_i)$. A value y_i equals the sum of the lengths of the levels A_i whose values are less than x_i. The associated coordination function g is one-to-one and is composed of line segments having a slope of 1. The shape of these two functions is shown in next figure.


Formes de G_C and g_C correspondant à une fonction étagée C.

Coordination function with several criteria.

If criterion C has levels A_i, we can introduce secondary criteria C_i corresponding to each level. In this case, we can define a coordination function that verifies the following conditions, in addition to condition (1) :

$$\forall\ \omega_1, \omega_2 \in A_i\ \ C_i(\omega_1) < C_i(\omega_2) \Rightarrow g(\omega_1) < g(\omega_2)$$

Details are not given in this paper.

4. Application to Poisson sampling

Initialization : $\Gamma_{k,0}(\omega_k) = 0 \quad g_{k,1}(\omega_k) = \omega_k$

$I_{k,1}(\omega_k) = \mathbb{1}_{[0,\pi_{k,1}[}(\omega_k) \quad \Gamma_{k,1}(\omega_k) = \gamma_{k,1}\ \mathbb{1}_{[0,\pi_{k,1}[}(\omega_k)$

For sample S_t, we choose a coordination function $g_{k,t}$ associated to each unit k. Then :

$$k \in S_t \Leftrightarrow g_{k,t}(\omega_k) \in [0, \pi_{k,t}[$$

We define : $A_{k,t} = g_{k,t}^{-1}[0, \pi_{k,t}[$, and the indicator function is : $I_{k,t}(\omega_k) = \mathbb{1}_{A_{k,t}}(\omega_k)$

Sampling

Every sampling follows from the coordination function, and therefore from the desired criteria.

For a separate sample t or for the first selection t of a panel, we use for the selection of the sample S_t : criterion $C_{k,t}(\omega_k)$ = cumulative burden $\Gamma_{k,t-1}(\omega_k)$, and then deduce $g_{k,t}$ (formula 2 supra).

For updating a panel : S_u = sample corresponding to the latest update ($u \leq t - 1$). We denote $A_{k,u} = g_{k,u}^{-1}[0, \pi_{k,u}[$

We use as a first-stage criterion in the calculation of the coordination function any decreasing function of the indicator function of $A_{k,u}$. For example : $C_{k,t}(\omega_k) = 1$ if $\omega_k \in A_{k,u}$ (i.e. $k \in S_u$)

$= 2$ if $\omega_k \notin A_{k,u}$ (i.e. $k \notin S_u$)

If we wish to take into account past burdens, we can use the cumulative burden $\Gamma_{k,t-1}$ as a secondary criterion. This leads to a certain coordination function $g_{k,t}$, but without rotation.

If we wish individual rotation rate $r_{k,u}$, the coordination function becomes $g'_{k,t} = g_{k,t} - r_{k,u}\, \pi_{k,u}$.

5. Application to SSRS

Details are not given in this paper.

1. Calculation of the approximate $I'_{k,t}(\omega_k)$, by its conditional expectation :

$I'_{k,t}(\omega_k) = E(I_{k,t}(\Omega) | \Omega = \Omega_k)$, which can be written as a function $b_{k,t}(g_{k,t}(\omega_k))$
where $\boldsymbol{\Omega} = (\Omega_1 \ldots \Omega_k \ldots \Omega_N)$ is a random vector from which we have a realization $\boldsymbol{\omega}$.

2. Approximation by step functions.

$I'_{k,t}$ and the cumulative burden function are no longer step functions. These functions are approximated by step functions, which are constant over predefined intervals, obtained by dividing [0, 1[in L equal subintervals.

Then, the procedures are similar to those used in Poisson sampling.