LA COORDINATION D'ÉCHANTILLONS D'ENQUÊTES AUPRÈS DES ENTREPRISES MISE EN PLACE À L'INSEE

Fabien GUGGEMOS (*), Olivier SAUTORY (**)

(*) Insee, Département des statistiques de court-terme
(**) Insee, Unité méthodologie statistique entreprises

Introduction

Le système statistique public réalise chaque année un nombre important d'enquêtes auprès des entreprises et des établissements. L'objectif de la coordination négative d'échantillons est de favoriser, lors du tirage d'un échantillon, la sélection d'entreprises n'ayant pas déjà été sélectionnées lors d'enquêtes récentes, tout en conservant le caractère sans biais des échantillons. Elle s'inscrit donc dans une démarche de réduction de la charge statistique imposée aux *petites* entreprises – les *grandes* entreprises, à partir d'un certain seuil, étant systématiquement enquêtées dans la plupart des enquêtes. La méthode actuellement utilisée à l'Insee permet de fournir des échantillons d'entreprises (ou des échantillons d'établissements) coordonnés négativement deux à deux. Une méthode plus générale, permettant la prise en compte de la charge cumulée des entreprises et la coordination globale d'un ensemble d'enquêtes, fait actuellement l'objet de travaux méthodologiques au sein de l'Unité méthodologie statistique - entreprises (UMS-E). Cette méthode a été proposée par Ch. Hesse et étudiée par Pascal Ardilly (voir [1] et [2]).

Ce document expose, dans une première partie, cette nouvelle méthodologie de tirages d'échantillons coordonnés négativement, en cours d'expérimentation. Dans une deuxième partie, seront présentés les résultats de premières simulations permettant de tester la méthode.

1. La méthode actuellement utilisée à l'Insee

1.1. Quelques rappels sur les plans de sondage des enquêtes auprès des entreprises

Les échantillons des enquêtes auprès des entreprises (appelées également unités dans ce qui suit) sont le plus souvent tirés selon des plans de sondage aléatoires simples stratifiés. La population d'entreprises correspondant au champ de l'enquête est découpée en strates construites à partir de caractéristiques connues des entreprises, en général disponibles dans le répertoire Sirene (activité, taille, localisation géographique…). Des taux de sondage, définis par le responsable d'enquête, sont appliqués dans chaque strate, conduisant à tirer n_h unités dans la strate h. L'algorithme de tirage généralement utilisé consiste à attribuer à chaque unité un nombre aléatoire compris entre 0 et 1, puis à sélectionner les n_h unités possédant les plus petits nombres aléatoires de chaque strate h.

Les grandes entreprises, en raison de leur poids économique, sont soumises à des taux de sondage beaucoup plus élevés que les petites. Dans la quasi-totalité des enquêtes auprès d'entreprises, une partie de la population, généralement composée de très grandes entreprises, est même intégrée d'office à l'échantillon, constituant ainsi une strate exhaustive.

1.2. La méthode actuellement pratiquée à l'UMS-E : permutation des nombres aléatoires

Afin de répartir au mieux la charge d'enquête entre les entreprises, l'UMS-E utilise actuellement une méthode mise au point il y a une vingtaine d'année par Franck Cotton et Christian Hesse (voir [3]), qui permet de coordonner des échantillons deux à deux.

Elle consiste à échanger les nombres aléatoires entre les unités d'une même strate dans la base de sondage de l'enquête 1, de façon à ce que les unités sélectionnées lors du premier tirage récupèrent les plus grands nombres aléatoires de leur strate d'origine. Ces permutations, réalisées strate par strate, attribuent ainsi à chaque unité de la base un nombre aléatoire "coordonné".

Pour procéder au tirage de l'échantillon de l'enquête 2, on affecte à toutes les unités de la base de sondage de l'enquête 2 :

- le nombre aléatoire coordonné, si l'unité était présente dans la base de sondage de l'enquête 1 ;
- un nombre aléatoire généré dans le cas contraire.

On sélectionne alors les unités ayant les plus petits nombres aléatoires dans chaque nouvelle strate de la base de sondage de l'enquête 2.

2. Fonction de coordination – Sélection des échantillons

La nouvelle méthodologie de tirages d'échantillons coordonnés repose comme la méthode actuelle sur l'utilisation de nombres aléatoires attribués aux unités. Mais contrairement à la méthode actuelle, fondée sur des permutations opérées sur ces nombres, ces nombres seront attribués une fois pour toutes aux unités, et ce sont des transformations de ces nombres qui seront mises en œuvre pour obtenir la coordination souhaitée. Ces transformations se feront par des fonctions ayant des propriétés particulières, appelées fonctions de coordination. Ce concept de fonction de coordination joue un rôle essentiel dans la méthode.

2.1. Définition d'une fonction de coordination

Une fonction de coordination g est une application mesurable de [0,1[dans [0,1[qui conserve la loi uniforme : si P est la loi de probabilité uniforme sur [0,1[, alors la loi image P^g est encore égale à la probabilité uniforme P.

Cela signifie que pour tout intervalle I = [a,b[inclus dans [0,1[, on a :

$$P\left[g^{-1}(I)\right] \overset{\text{def}}{=} P^g(I) = P(I) = b - a$$

La longueur de l'image réciproque par g d'un intervalle est égale à la longueur de cet intervalle : une fonction de coordination a donc pour propriété de conserver la longueur des intervalles - et des réunions d'intervalles - par image réciproque.

2.2. Sélection des échantillons

On considère une succession d'enquêtes t = 1, 2,…(t désigne à la fois la date et le numéro de l'enquête), et on note S_t l'échantillon correspondant à l'enquête t.
On attribue à chaque unité k de la base de sondage un nombre aléatoire permanent ω_k, tiré dans la loi de probabilité uniforme sur [0,1[. Les tirages des ω_k sont indépendants les uns des autres. On va définir pour chaque unité k une fonction de coordination qui change à chaque enquête : $g_{k,t}$ est la fonction de coordination de l'unité k pour l'enquête t. La sélection de l'échantillon S_t va dépendre des valeurs des nombres aléatoires ainsi transformés $g_{k,t}(\omega_k)$.
On va examiner les deux modes de tirage couramment utilisés pour la sélection d'échantillons d'enquêtes auprès des entreprises : le tirage de Poisson et le tirage aléatoire simple stratifié.

2.2.1. Tirage de Poisson

Nous présentons ce mode de tirage, très simple à mettre en œuvre, bien qu'il ne soit pas utilisé à l'Insee. Rappelons-en le principe : on attribue à chaque unité k de la base de sondage une probabilité d'inclusion π_k et chaque unité est tirée indépendamment des autres avec la probabilité π_k (ce qui a pour conséquence que l'échantillon est de taille aléatoire).

Dans le cas présent : pour sélectionner l'échantillon S_t, on sélectionne les unités k telles que $g_{k,t}(\omega_k) \in [0,\pi_{k,t}[$, où $\pi_{k,t}$ désigne la probabilité d'inclusion de l'unité k pour l'enquête t.

On a alors : $P(k \in S_t) = P\left(g_{k,t}(\omega_k) \in [0,\pi_{k,t}[\right) = P^{g_{k,t}}\left([0,\pi_{k,t}[\right) = P\left(\omega_k \in [0,\pi_{k,t}[\right) = \pi_{k,t}$

Les probabilités d'inclusion sont bien respectées, et les tirages des unités sont indépendants car les ω_k le sont.

2.2.2. Tirage aléatoire simple stratifié

Rappelons le principe de ce mode de tirage : on partitionne la base de sondage en strates, et dans chaque strate h on tire n_h unités selon un sondage aléatoire simple sans remise.

Dans le cas présent : à la date t, la base de sondage est découpée en strates (h,t). À l'intérieur de la strate (h,t), de taille $N_{(h,t)}$, on sélectionne les $n_{(h,t)}$ unités correspondant aux $n_{(h,t)}$ plus petites valeurs $g_{i,t}(\omega_i),\ i = 1...N_{(h,t)}$.

On notera par la suite, pour alléger les notations, N et n les tailles de la strate et de l'échantillon dans la strate.

On a donc : $k \in S_t \Leftrightarrow g_{k,t}(\omega_k) \in E_n\left(g_{i,t}(\omega)\right)$

en notant $E_n\left(g_{i,t}(\omega)\right)$ l'ensemble des n plus petites valeurs $g_{i,t}(\omega_i)$.

Les N nombres aléatoires (ω_i) associés aux N unités de la strate ont été tirés indépendamment dans la loi de probabilité uniforme sur [0,1[, notée P. Comme on a $P^{g_{i,t}} = P$ pour tout i, les N nombres $g_{i,t}(\omega_i)$ sont eux-mêmes tirés indépendamment dans la loi P. Par conséquent, selon l'argument classique, les n plus petites valeurs $g_{i,t}(\omega_i)$ donnent bien un échantillon aléatoire simple de taille n dans la strate.

3. Une procédure de sélection pas à pas

3.1. Charge de réponse cumulée et fonction de coordination

Le principe général de la coordination négative est de sélectionner en priorité, pour un tirage donné, les unités qui ont eu la plus faible charge de réponse dans un passé plus ou moins récent.

On note **ω** = (… ω_k…) le vecteur des nombres aléatoires attribués aux unités k de la population.

On note $I_{k,t}(\boldsymbol{\omega})$ l'indicatrice d'appartenance de l'unité k à l'échantillon S_t , égale à 1 si les valeurs dans **ω** conduisent à sélectionner l'unité k, et 0 sinon :

$$k \in S_t \Leftrightarrow I_{k,t}(\boldsymbol{\omega}) = 1$$

(l'inclusion de k dans l'échantillon S_t ne dépend que du vecteur **ω**).

Il s'agit d'une variable aléatoire, dépendant du vecteur **ω**.

On note $\gamma_{k,t}$ la charge de réponse d'une unité k pour l'enquête t (que l'on supposera souvent identique pour toutes les unités pour une enquête donnée).

La charge de réponse effective est donc une variable aléatoire $\gamma_{k,t}(\boldsymbol{\omega}) = \gamma_{k,t} . I_{k,t}(\boldsymbol{\omega})$

La charge de réponse cumulée pour une unité k est une fonction de $\boldsymbol{\omega}$ égale à:

$$\Gamma_{k,t}(\boldsymbol{\omega}) = \sum_{u \le t} \gamma_{k,u} . I_{k,u}(\boldsymbol{\omega}) \quad (1)$$

Pour répondre à l'objectif de coordination négative, au moment de la sélection de l'échantillon S_t, on souhaite définir, pour chaque unité k, une fonction de coordination $g_{k,t}$ fondée sur $\Gamma_{k,t-1}$, i.e. la charge cumulée de l'unité k jusqu'à l'enquête t-1. Compte tenu du mode de sélection des unités choisi, à savoir que la probabilité qu'une unité soit sélectionnée est d'autant plus élevée que $g_{k,t}(\omega_k)$ est petit, une propriété souhaitée pour les fonctions de coordination est la suivante :

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

où $\omega_{k,i}$ (i=1,2) désigne la $k^{ème}$ composante du vecteur $\boldsymbol{\omega_i}$.

Cette condition n'est pas facile à manipuler, car la charge cumulée $\Gamma_{k,t-1}(\boldsymbol{\omega})$ est une fonction du vecteur $\boldsymbol{\omega}$, i.e. non seulement du nombre aléatoire ω_k associé à l'unité k, mais de tous les autres nombres aléatoires. Nous verrons plus loin comment on peut la remplacer par une fonction $\Gamma'_{k,t-1}(\omega_k)$ qui dépende uniquement de ω_k.

La propriété attendue pour une fonction de coordination $g_{k,t}$ s'écrit alors :

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

3.2. Construction d'une fonction de coordination

3.2.1. Le principe de la construction d'une fonction de coordination à partir d'un critère

Considérons le problème d'une façon plus générale. On suppose que l'on dispose d'un critère $C_{k,t}(\omega_k)$, tel que plus petite est la valeur de ce critère, plus grande doit être la probabilité pour l'unité k d'être sélectionnée dans l'échantillon S_t. Ce critère pourra ainsi représenter une charge, mais pas uniquement. Il peut être vu comme un simple critère de tri.

On omet les indices k et t, pour simplifier les notations. Ainsi ω désigne un réel compris entre 0 et 1.

On suppose que C est une fonction mesurable bornée : $\omega \in [0,1[\rightarrow C(\omega) \in \mathbb{R}$

On veut lui associer une fonction de coordination g telle que :

$$C(\omega_1) < C(\omega_2) \Rightarrow g(\omega_1) < g(\omega_2) \quad (2)$$

On note P^C la loi de probabilité image par C de la loi de probabilité uniforme P sur [0,1[, et F_C la fonction de répartition de C. On définit la fonction $G_C = F_C(C)$. On peut écrire :

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

Avant de voir les propriétés de la fonction G_C, définissons la notion de *palier*.

Définition d'un palier

On appelle *palier* du critère C tout sous-ensemble A de [0,1[, avec P(A) > 0, tel qu'il existe un réel x vérifiant $C^{-1}(\{x\})=A$. En d'autres termes, cela signifie qu'il existe des segments de droite horizontaux dans le graphe de C.

Propriétés de G_C

On montre que :

- G_C a une image incluse dans [0,1[
- G_C possède les mêmes paliers que C (à un ensemble de mesure nulle près)
- G_C vérifie la propriété (2)
- Pour tout y dans l'image de G_C, on a $F_{G_c}(y) = P(u \mid G_C(u) < y) = y$, en notant F_{G_c} la fonction de répartition de G_C.
- Si C n'a pas de palier, l'image de G_C est exactement [0,1[, (à un ensemble de mesure nulle près), et G_C est alors une fonction de coordination.

Quand la fonction C a au moins un palier, l'image de G_C est strictement incluse dans [0,1[, il faut modifier G_C pour obtenir une fonction de coordination, notée g_C, d'image égale à [0,1[. Ceci peut se faire de la façon suivante.

On impose $g_C(\omega) = G_C(\omega)$ pour tout ω n'appartenant pas à un palier.

Considérons un palier A de C (et donc de G_C) : si $\omega \in A$, $G_C(\omega)$ = constant = y.

On note B le plus grand intervalle de la forme [y,t[tel que $G_C^{-1}(B) = A$. Sur le graphe de C, B est l'intervalle de l'axe des ordonnées qui matérialise le "saut" entre le palier A et le premier palier situé au-dessus. On montre que P(B) = P(A), i.e. les segments A et B ont même longueur.

La probabilité $P^G(B)$ est concentrée sur y : pour obtenir une probabilité uniforme dans B, on considère la fonction affine de pente 1 qui transforme A en B :

si A = [a,b[, pour tout $\omega \in A$, on définit $g_C(\omega) = \omega - y + a$.

Ceci peut se réécrire : $g_C(\omega) = G_C(\omega) + \int 1_{A \cap [0,\omega]}(u)\,du$

Plus généralement, si C a plusieurs paliers A_i, on définit :

$$g_C(\omega) = G_C(\omega) + \sum_i 1_{A_i}(\omega) \cdot \int 1_{A_i \cap [0,\omega]}(u)\,du$$

3.2.2. Un exemple de construction d'une fonction de coordination à partir d'un critère étagé

Le critère C est une fonction étagée, contenant 4 paliers :

- A_1 = [0, 0.3[$x_1=C(A_1) = C_1$
- A_2 = [0.3; 0.5[$x_2=C(A_2) = C_3$
- A_3 = [0.5; 0.9[$x_3=C(A_3) = C_2$
- A_4 = [0.9; 1[, $x_4=C(A_4) = C_1$.

On sait que la fonction G_C a les mêmes 4 paliers que C.

Les valeurs $y_i = G_C(A_i)$ sont classées dans le même ordre que les valeurs x_i.

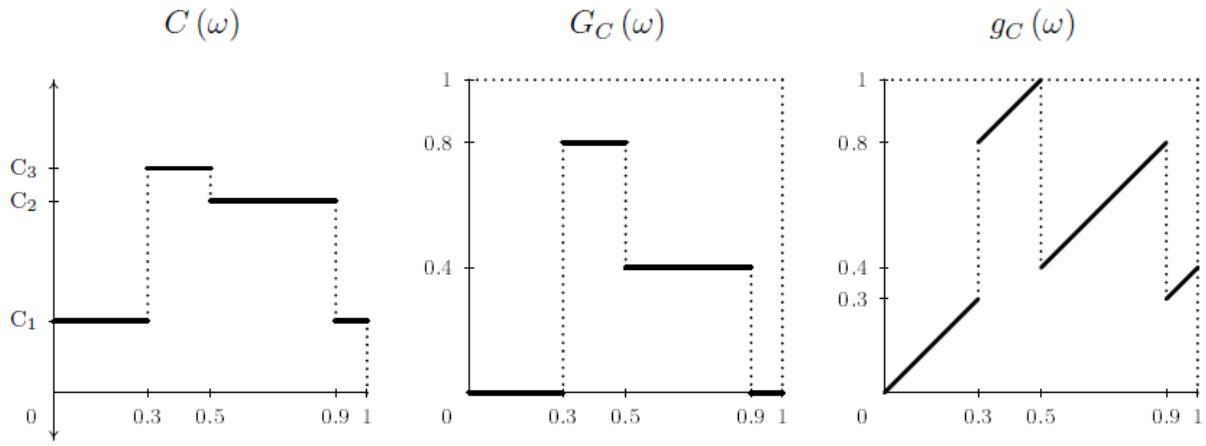
Une valeur y_i est égale à la somme des longueurs des paliers A_j tels que $x_j = C(A_j) < x_i$:

- y_1 (comme y_4) vaut 0 car aucun palier A_j ne vérifie $x_j = C(A_j) < C_1$ (=C4) ;
- y_2 est égale à la somme des longueurs des paliers A_1 (0.3), A_3 (0.4) et A_4 (0.1), soit 0.8 ;
- y_3 est égale à la somme des longueurs des paliers A_1 (0.3) et A_4 (0.1), soit 0.4.

La fonction de coordination g_C est une fonction bijective sur [0,1[, formée de segments de droite de pente égale à 1.

La figure suivante présente les graphes de ces 3 fonctions.

On peut noter que c'est le classement des C_i (ici le fait que $C_1 < C_2 < C_3$), et non leurs valeurs en tant que telles, qui déterminent la forme des fonctions G_C et g_C .



4. Application au tirage de Poisson

On rappelle que, avec ce mode de tirage, on sélectionne une unité k dans l'échantillon S_t si le nombre aléatoire $g_{k,t}(\omega_k) \in [0, \pi_{k,t}[$, où $\pi_{k,t}$ désigne la probabilité d'inclusion de l'unité k. Par conséquent, l'indicatrice $I_{k,t}(\boldsymbol{\omega})$ définie au § 3.1. est une fonction de ω_k seulement.

$$k \in S_t \Leftrightarrow I_{k,t}(\omega_k) = 1$$

Sélection de l'échantillon S_1

On initialise la charge à 0 : $\Gamma_{k,0}(\omega_k) = 0 \quad \forall k$.

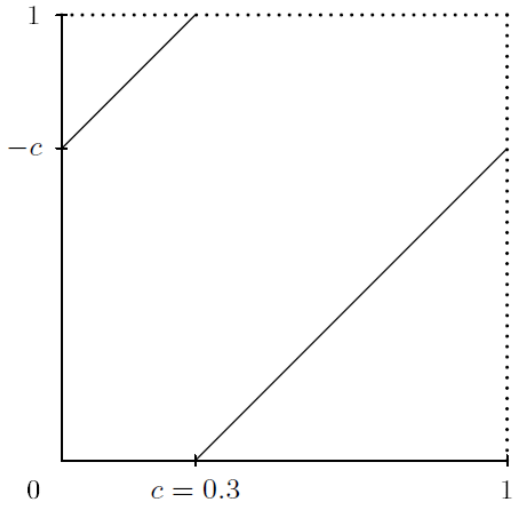
Il n'y a aucune coordination à réaliser : $k \in S_1 \Leftrightarrow \omega_k \in [0, \pi_{k,1}[$, ce qui revient à prendre comme fonction de coordination l'identité sur [0,1[: $g_{k,1}(\omega_k) = \omega_k \quad \forall k$.

La fonction indicatrice est égale à $I_{k,1}(\omega_k) = \mathbb{1}_{[0,\pi_{k,1}[}(\omega_k)$, elle ne dépend que de ω_k (et non de $\boldsymbol{\omega}$). C'est une fonction étagée, valant 1 sur l'intervalle $[0, \pi_{k,1}[$, et 0 sur l'intervalle $[\pi_{k,1}, 1[$.

La fonction de charge "cumulée" est définie par : $\Gamma_{k,1}(\omega_k) = \gamma_{k,1}\ \mathbb{1}_{[0,\pi_{k,1}[}(\omega_k)$, où $\gamma_{k,1}$ désigne la charge de réponse pour l'unité k ; la charge cumulée est elle-même une fonction de ω_k. C'est également une fonction étagée, valant $\gamma_{k,1}$ sur l'intervalle $[0, \pi_{k,1}[$, et 0 sur l'intervalle $[\pi_{k,1}, 1[$.

Sélection de l'échantillon S_2

On utilise cette fonction de charge cumulée $\Gamma_{k,1}$ comme "critère" (au sens du § 3.2.1.) pour construire une fonction de coordination $g_{k,2}$ pour le tirage du 2ème échantillon S_2. Ce critère a deux paliers, la fonction de coordination qui en résulte est particulièrement simple. Il est important de noter que l'on peut construire plusieurs fonctions de coordination qui satisfont à la condition (2) du § 3.2.1. Le graphe ci-dessous montre l'allure d'une fonction de coordination correspondant à $\pi_{k,1} = c = 0.3$.



Chacun des deux segments de droite de pente égale à 1 pourrait être remplacé par un segment de pente égale à −1, défini sur le même intervalle.

La sélection de l'échantillon S_2 s'effectue alors comme suit : $k \in S_2 \Leftrightarrow g_{k,2}(\omega_k) \in [0, \pi_{k,2}[$

Si on pose $A_{k,2} = g_{k,2}^{-1}[0, \pi_{k,2}[$, l'indicatrice s'écrit : $I_{k,2}(\omega_k) = \mathbb{1}_{A_{k,2}}(\omega_k)$, et la charge cumulée vaut : $\Gamma_{k,2}(\omega_k) = \Gamma_{k,1}(\omega_k) + \gamma_{k,2}\, \mathbb{1}_{A_{k,2}}(\omega_k)$, où $\gamma_{k,2}$ désigne la charge de réponse pour l'unité k lors de la deuxième enquête. L'indicatrice et la charge cumulée sont des fonctions de ω_k.

Sélection de l'échantillon S_t

Plus généralement, pour la sélection de l'échantillon S_t, on construit une fonction de coordination $g_{k,t}$ à partir de la fonction de charge cumulée $\Gamma_{k,t-1}$. Et on a

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

$$I_{k,t}(\omega_k) = \mathbb{1}_{A_{k,t}}(\omega_k) \quad \text{avec} \quad A_{k,t} = g_{k,t}^{-1}[0, \pi_{k,t}[$$

$$\Gamma_{k,t}(\omega_k) = \Gamma_{k,t-1}(\omega_k) + \gamma_{k,t}\, \mathbb{1}_{A_{k,t}}(\omega_k)$$

Les indicatrices étant à chaque étape des fonctions étagées, il en est de même des fonctions de charge cumulées, et on est toujours dans le cadre de la détermination d'une fonction de coordination à partir d'un critère ne présentant que des paliers… même si le nombre de paliers s'accroît bien entendu avec le nombre d'échantillons que l'on coordonne ! On ajoute en effet un palier à chaque sélection d'un échantillon. On peut limiter le nombre de discontinuités en fixant un nombre maximum d'échantillons à coordonner : les charges correspondant aux tirages les plus anciens sont progressivement éliminées du calcul de la charge cumulée.

La mise en œuvre informatique de la méthode ne pose pas de difficulté majeure, puisque toutes les fonctions indicatrices et charges cumulées sont des fonctions étagées, et que les fonctions de coordination sont donc des fonctions affines par morceaux, de pente 1 (ou −1).

5. Application au tirage aléatoire simple stratifié

On rappelle que, avec ce mode de tirage, on sélectionne une unité k dans l'échantillon S_t si le nombre aléatoire $g_{k,t}(\omega_k)$ figure parmi les n plus petits nombres $g_{i,t}(\omega_i)$ associés à toutes les unités i de la base de sondage[1]. L'inclusion de k dans S_t dépend donc de l'ensemble des nombres aléatoires ω_i de toutes les unités i de la base de sondage, et la fonction indicatrice $I_{k,t}$, de même que la charge cumulée $\Gamma_{k,t}$, sont donc des fonctions du vecteur **ω**. Il est donc nécessaire de remplacer l'indicatrice $I_{k,t}$ par une indicatrice approchée $I'_{k,t}$, qui sera une fonction de ω_k qui lui sera proche.

5.1. L'indicatrice approchée

On note $\Omega = (\Omega_1, \Omega_2, ..., \Omega_N)$ le vecteur aléatoire dont la réalisation est le vecteur $\omega = (\omega_1, \omega_2, ..., \omega_N)$ composé des N nombres aléatoires associés aux unités de la base de sondage. La meilleure approximation possible de la fonction indicatrice qui ne dépende que de ω_k, au sens de la norme L2, est son espérance conditionnelle :

$$I'_{k,t}(\omega_k) = E\left(I_{k,t}(\Omega) \middle| \Omega_k = \omega_k\right) = P\left(k \in S_t \middle| \Omega_k = \omega_k\right)$$

En supposant les fonctions de coordination bijectives[2], on peut écrire :

$$I'_{k,t}(\omega_k) = P\left(k \in S_t \middle| \Omega_k = \omega_k\right) = P\left(k \in S_t \middle| g_{k,t}(\Omega_k) = g_{k,t}(\omega_k)\right) = b_{k,t}\left(g_{k,t}(\omega_k)\right)$$

en notant $b_{k,t}(x) = P\left(k \in S_t \middle| g_{k,t}(\Omega_k) = x\right)$.

Les $g_{i,t}(\omega_i)$ étant des nombres aléatoires indépendants tirés de la loi uniforme sur [0,1[, $b_{k,t}(x)$ est égal à la probabilité que parmi les N-1 nombres aléatoires $g_{i,t}(\omega_i)$ (i ≠ k), il y en ait au plus n-1 qui soient inférieurs à x.

Autrement dit, l'événement $\left\{k \in S_t \middle| g_{k,t}(\Omega_k) = x\right\}$ équivaut à l'événement suivant :
$\{X_1, X_2, ..., X_{N-1}$ désignant N−1 variables aléatoires indépendantes tirées dans une loi uniforme sur $[0,1[$, la $n^{ème}$ statistique d'ordre $X_{(n)}$ est plus grande que $x\}$.

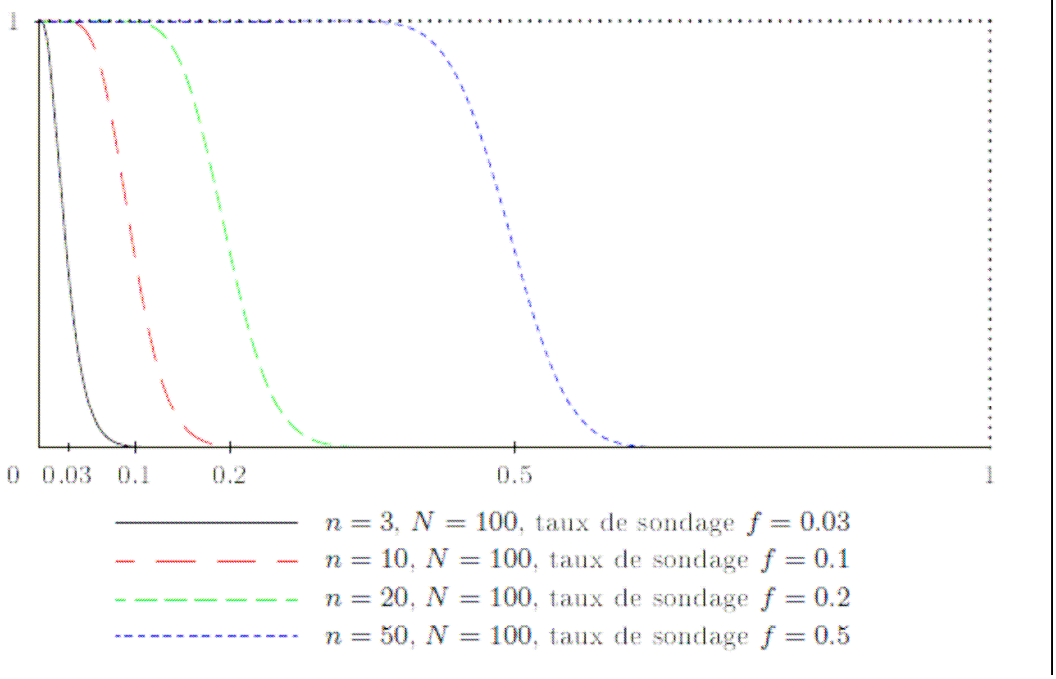
Or, résultat bien connu de la théorie des probabilités, cette statistique d'ordre suit une loi beta de paramètres n et N−n+1. On a donc :

$$b_{k,t}(x) = 1 - P\left(X_{(n)} \le x\right) = 1 - \frac{1}{B(p,q)} \int_{u=0}^{u=x} u^{p-1}(1-u)^{q-1} du$$

avec p=n, q = N−n+1 , et $B(p,q) = \frac{(p-1)!(q-1)!}{(p+q-1)!}$

On peut voir dans les graphiques ci-dessous la forme de la fonction b(x) pour quelques valeurs de n et N.

[1] On rappelle que l'on omet l'indice de strate.
[2] Ce qui est vérifié dans la méthode présentée ici, mais qui n'est pas une propriété intrinsèque d'une fonction de coordination.



Une fonction b(x) a l'allure suivante : une première partie "presque horizontale" proche de 1, correspondant à une sélection "presque certaine" de l'unité dans l'échantillon, une troisième partie "presque horizontale" proche de 0, correspondant à une non-sélection "presque certaine" de l'unité dans l'échantillon. Entre les deux, une partie décroissante "à forte pente", correspondant à un intervalle sur l'axe des abscisses plus ou moins long, à peu près centré sur la valeur n/N [3], égale aux taux de sondage. C'est donc autour de cette valeur qu'il y a la plus grande incertitude sur la sélection ou non de l'unité dans l'échantillon.

5.2. La fonction de charge espérée

Le remplacement de la fonction indicatrice par une indicatrice approchée, qui est son espérance conditionnellement à ω_k, fait que la fonction de charge cumulée est elle-même remplacée, dans l'expression (1) du §3.1., par une charge cumulée espérée $\Gamma'_{k,t}$, compte tenu de ω_k. Pour que l'algorithme fonctionne correctement, i.e. conduise à des échantillons sans biais, il est nécessaire d'utiliser cette charge espérée, et non la charge réelle, fondée sur les inclusions observées de l'unité k dans les différents échantillons :

$$\Gamma_{k,t} = \sum_{u=1}^{t} \gamma_{k,u}\ \mathbf{1}(k \in S_u)$$

On peut noter que, dans le cas de tirages de Poisson, c'est bien la charge réelle qui intervient dans l'algorithme.

5.3. Approximation par des fonctions étagées

Les fonctions indicatrices approchées $I'_{k,t}$ et les fonctions de charge cumulée espérée ne sont pas des fonctions étagées, ni même des fonctions que l'on peut "calculer" facilement. On va introduire deux types de simplifications.

On commence par diviser l'intervalle [0,1[en L intervalles de longueurs égales, $\left[\frac{\ell-1}{L};\frac{\ell}{L}\right[$ $\ell = 1 \ldots L$, L étant un nombre entier "assez élevé" (au moins supérieur à 100).

[3] Le point d'inflexion de la courbe a pour abscisse n−1 / N−1

1. Pour simplifier la forme de la fonction indicatrice approchée $b_{k,t}$, on la remplace par une fonction affine par morceaux $\tilde{b}_{k,t}$ prenant les mêmes valeurs que $b_{k,t}$ aux extrémités des intervalles :

$$\forall x \in \left[\frac{\ell-1}{L};\frac{\ell}{L}\right[\qquad \tilde{b}_{k,t}(x) = L \cdot \left(b_{k,t}\left(\frac{\ell}{L}\right) - b_{k,t}\left(\frac{\ell-1}{L}\right)\right) \cdot \left(x - \frac{\ell}{L}\right) + b_{k,t}\left(\frac{\ell}{L}\right)$$

2. Pour obtenir une fonction étagée, on calcule la valeur moyenne de $\tilde{b}_{k,t}$ sur chaque intervalle :

$$\beta_{k,t}(\ell) = \frac{1}{1/L} \cdot \int_{\frac{\ell-1}{L}}^{\frac{\ell}{L}} \tilde{b}_{k,t}\left(g_{k,t}(\omega)\right) d\omega$$

On définit alors la fonction $\beta_{k,t}$ par : $\forall \omega \in \left[\frac{\ell-1}{L};\frac{\ell}{L}\right[\quad \beta_{k,t}(\omega) = \beta_{k,t}(\ell)$, qui est donc une approximation de la fonction indicatrice approchée : au final, on approxime ainsi la fonction indicatrice approchée par une fonction constante par morceaux)

5.4. Construction d'une fonction de coordination

L'algorithme de sélection pas à pas, qui va être décrit plus loin, est similaire à celui présenté dans le cas de tirages de Poisson. Il fera en particulier intervenir les fonctions de charge cumulée espérées $\Gamma'_{k,t}$, qui seront de la forme (1) (§3.1.), mais où les indicatrices seront remplacées par les indicatrices approchées $\beta_{k,t}$. Comme ces dernières, les fonctions de charge cumulée seront donc des fonctions étagées, constantes sur chacun des intervalles $\left[\frac{\ell-1}{L};\frac{\ell}{L}\right[$.

Il s'agit donc de construire une fonction de coordination à partir d'un critère C constant par morceaux :

$$\forall \omega \in \left[\frac{\ell-1}{L};\frac{\ell}{L}\right[\quad C(\omega) = C_\ell$$

On est donc dans le même cas de figure que celui présenté dans l'exemple du § 3.2.2. On déduit du critère C la fonction G_C, que l'on notera G. Cette fonction G est également une fonction constante par morceaux :

$$\forall \omega \in \left[\frac{\ell-1}{L};\frac{\ell}{L}\right[\quad G(\omega) = G_\ell$$

La valeur G_ℓ est égale à longueur du sous-ensemble de [0,1[sur lequel $C(\omega) < C_\ell$: ce sous-ensemble est nécessairement une réunion d'intervalles de longueur $1/L$, donc G_ℓ est un multiple de $1/L$:

$$G_\ell = \lambda(\ell)/L.$$

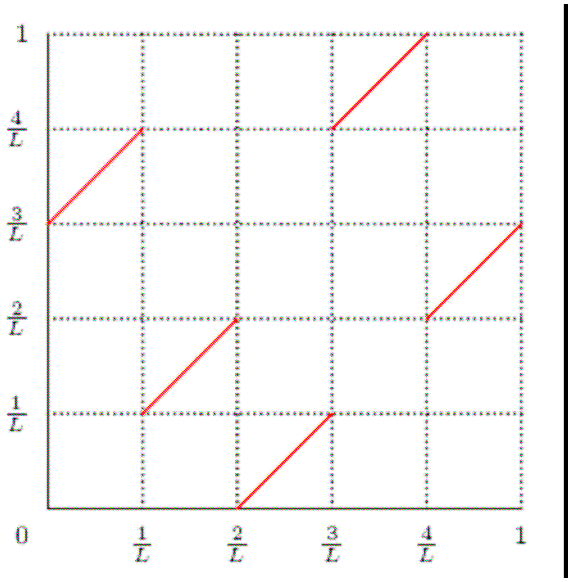
La représentation de la fonction G se fait, comme d'habitude, dans le carré [0,1[×[0,1[, que l'on va ici subdiviser en L^2 petits carrés définis par les intervalles de longueur 1/L sur l'axe des abscisses et l'axe des ordonnées.

Supposons que les valeurs C_ℓ, et donc les valeurs G_ℓ, soient toutes distinctes. Les paliers sont donc tous de longueur 1/L. Le graphe de la fonction G est alors constitué des "bases" de L de ces petits carrés, sachant qu'un unique carré est concerné dans chaque "ligne" et dans chaque "colonne" du "quadrillage" formé par les L^2 petits carrés.

λ est une fonction de ℓ à valeurs dans $\{0,1,2,...,L-1\}$, telle que deux valeurs distinctes de ℓ donnent lieu à deux valeurs $\lambda(\ell)$ différentes. Par conséquent, il existe une permutation σ sur $\{1,2,3,...,L\}$ telle que $\lambda(\ell) = \sigma(\ell) - 1$.

On en déduit la (plus précisément "une") fonction de coordination en considérant les diagonales principales de ces L petits carrés.

Le graphe suivant présente une fonction de coordination de ce type, avec L = 5.



Lorsqu'il existe des paliers de longueur égale à c/L, c>1, la construction de la fonction de coordination g est similaire : elle s'opère sur un regroupement de c intervalles, mais au final g a la même forme que dans le cas de paliers de longueur égale à 1/L. Elle est donc complètement définie par une permutation σ sur $\{1,2,3,...,L\}$. Son expression est la suivante :

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

que l'on peut réécrire :

$$g_\sigma(\omega) = \frac{\sigma(L\cdot(\omega - \omega \bmod \frac{1}{L}) + 1) - 1}{L} + \omega \bmod \frac{1}{L}$$

en notant y mod x le reste de la division euclidienne de y par x.

5.5. La sélection des échantillons

Sélection de l'échantillon S_1

On initialise la charge à 0 : $\Gamma_{k,0}(\omega_k) = 0 \quad \forall k$.

Il n'y a aucune coordination à réaliser : $k \in S_1 \Leftrightarrow \omega_k \in E_n(\boldsymbol{\omega})$ [4], ce qui revient à prendre comme fonction de coordination l'identité sur [0,1[: $g_{k,1}(\omega_k) = \omega_k \quad \forall k$.

La charge réelle vaut $\Gamma_{k,1} = \gamma_{k,1}\,\mathbf{1}(k \in S_1)$, en notant $\gamma_{k,1}$ la charge de réponse de l'unité k pour l'enquête 1.

Mais la fonction de charge espérée, fonction de ω_k uniquement, utilise la fonction indicatrice approchée $\beta_{k,1}$:

$$\Gamma'_{k,1}(\omega_k) = \gamma_{k,1}\,\beta_{k,1}(\omega_k)$$

[4] $E_n(\omega)$ désigne l'ensemble des n plus petites valeurs ω_i.

Conséquence de la forme de la fonction $\beta_{k,1}$, commentée plus haut, les fonctions de charge réelle et espérée coïncideront en général sur l'intervalle [0,1[, sauf sur un voisinage de la valeur n/N ; ce sont les unités de nombre aléatoire proche de n/N pour lesquelles l'appartenance à l'échantillon est *a priori* la plus incertaine.

Sélection de l'échantillon S_2

On utilise la fonction de charge espérée $\Gamma'_{k,1}$ comme "critère" (au sens du § 3.2.1.) pour construire une fonction de coordination $g_{k,2}$ pour le tirage du 2ème échantillon S_2. Cette fonction de coordination repose, comme on l'a vu plus haut, sur une permutation σ_2 de $\{1,2,3,...,L\}$.

La sélection de l'échantillon S_2 s'effectue alors comme suit : $k \in S_2 \Leftrightarrow g_{k,2}(\omega_k) \in E_n(g_{k,2}(\boldsymbol{\omega}))$

La charge cumulée espérée vaut :
$\Gamma'_{k,2}(\omega_k) = \Gamma'_{k,1}(\omega_k) + \gamma_{k,2}\ \beta_{k,2}(\omega_k)$, où $\gamma_{k,2}$ désigne la charge de réponse pour l'unité k, et $\beta_{k,2}$ l'indicatrice approchée correspondant à ce tirage.

Sélection de l'échantillon S_t

Plus généralement, pour la sélection de l'échantillon S_t, on construit une fonction de coordination $g_{k,t}$, définie par une permutation σ_t de $\{1,2,3,...,L\}$, à partir de la fonction de charge cumulée espérée $\Gamma'_{k,t-1}$. Et on a

$$k \in S_t \Leftrightarrow g_{k,t}(\omega_k) \in E_n(g_{i,t}(\boldsymbol{\omega}))$$

$$\Gamma'_{k,t}(\omega_k) = \Gamma'_{k,t-1}(\omega_k) + \gamma_{k,t}\ \beta_{k,t}(\omega_k)$$

6. Simulations et résultats

L'objet de cette partie est de présenter les résultats de simulations effectuées pour tester la méthode de tirage d'échantillons coordonnés présentée précédemment. L'ensemble des simulations ont été réalisées à l'aide d'un prototype développé sous la forme de macro-programmes SAS V9. Dans un premier temps, on illustrera sur un cas relativement simple le comportement de la procédure de tirage en justifiant les formes prises à chaque étape par les fonctions de coordination et de charge cumulée espérée. Dans un second temps, on exposera une série de résultats quantitatifs permettant de dresser une première évaluation de l'efficacité de la méthode, en fonction des paramètres caractérisant le plan de sondage ou la population d'unités statistiques.

6.1. Illustration des différentes étapes de la procédure de tirage dans un cas simple

On se propose à présent d'expliciter le fonctionnement de la méthode de tirage d'échantillons coordonnés à l'aide d'un exemple. Pour ce faire, on considère une population de taille N = 100 dont toutes les unités ont initialement une charge de réponse initiale nulle. Sur cette population vont être tirés successivement 20 échantillons d'enquêtes, par sondage aléatoire simple coordonné. Plus précisément, le plan de sondage testé est le suivant : chaque échantillon est de taille 25, avec une charge de réponse égale à 1, à l'exception des échantillons n°3 et n°15, de taille 50 et de charge de réponse égale à 3, ainsi que des échantillons n°10 et n°11, de taille 10 et de charge de réponse égale à 2. À l'issue des 20 tirages, la charge de réponse cumulée moyenne par unité s'élève donc à

$$E(C_{Finale}) \quad = \quad 16*1*(25/100) + 2*3*(50/100) + 2*2*(10/100) \quad = \quad 7{,}4$$

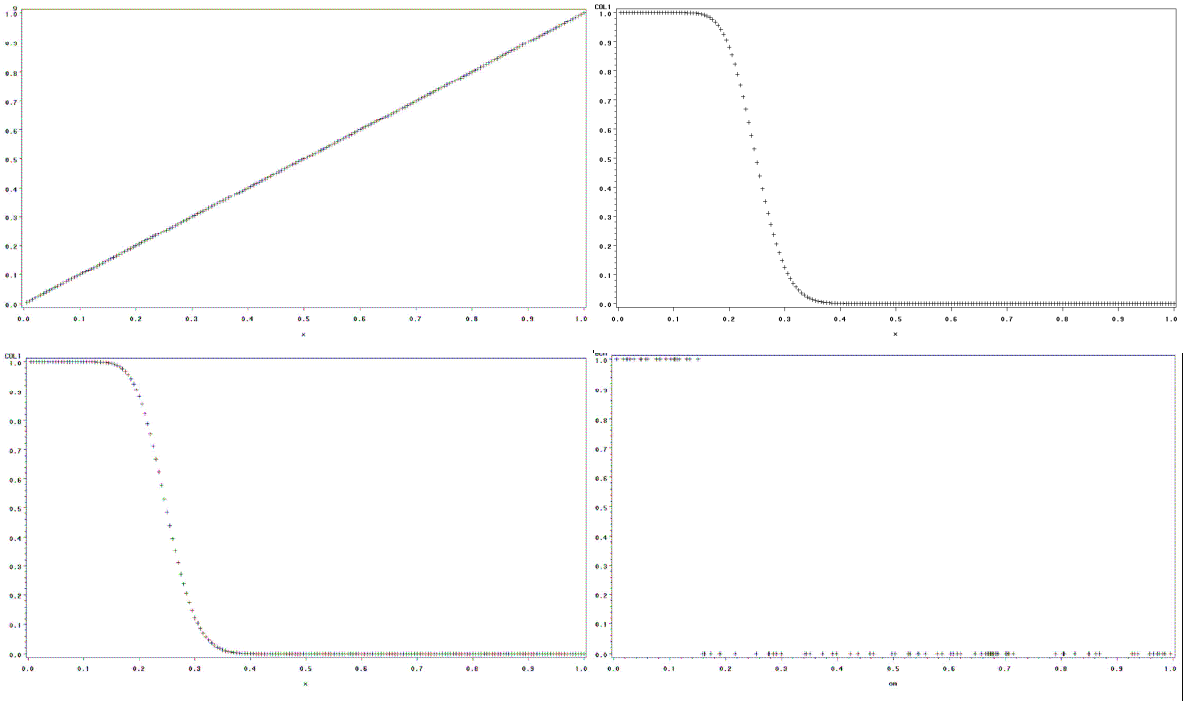
Le cas traité dans cet exemple est le plus simple qui puisse être, puisque la population d'intérêt dans laquelle on échantillonne ne subit aucune évolution d'un tirage à l'autre (pas d'apparition de nouvelles unités, pas de disparition d'unités, pas d'unités changeant de strates puisque dans cet exemple la

population n'est jamais stratifiée, etc.). Comme les unités ont en outre la même charge de réponse initiale, il en résulte finalement qu'**à chaque tirage leurs fonctions de coordination sont toutes égales, de même que leurs espérances d'appartenir à l'échantillon et leurs charges de réponse cumulées espérées après tirage**. Dans la suite, on présente, pour les quatre premiers tirages ainsi que pour le dernier, les graphiques de ces trois fonctions, complétés par un quatrième signalant les unités, identifiées par leur numéro aléatoire ω_k respectif, qui sont finalement réellement échantillonnées. Ces quatre graphiques sont présentés pour chaque tirage considéré de la façon suivante :

Fonction de coordination	Indicatrice ***espérée*** d'appartenance à l'échantillon
Charge de réponse cumulée ***espérée*** après tirage	Indicatrice ***réelle*** d'appartenance à l'échantillon

On se propose à présent de les commenter…

1ère ETAPE : Tirage n°1 taille échantillon = 25 charge de réponse = 1



Pour le premier tirage, les unités ayant une charge initiale nulle, aucune coordination n'est à effectuer. Le tirage aléatoire simple de l'échantillon de taille *n* = 25 suit alors une procédure « classique » consistant à retenir les *n* unités ayant le plus petit numéro aléatoire. C'est ce que l'on observe sur le graphique en bas à droite (sur ce graphique, chacune des 100 unités de la population est représentée par un point d'abscisse son numéro aléatoire ω_k et d'ordonnée valant 1 ou 0 selon qu'elle est effectivement échantillonnée ou non).

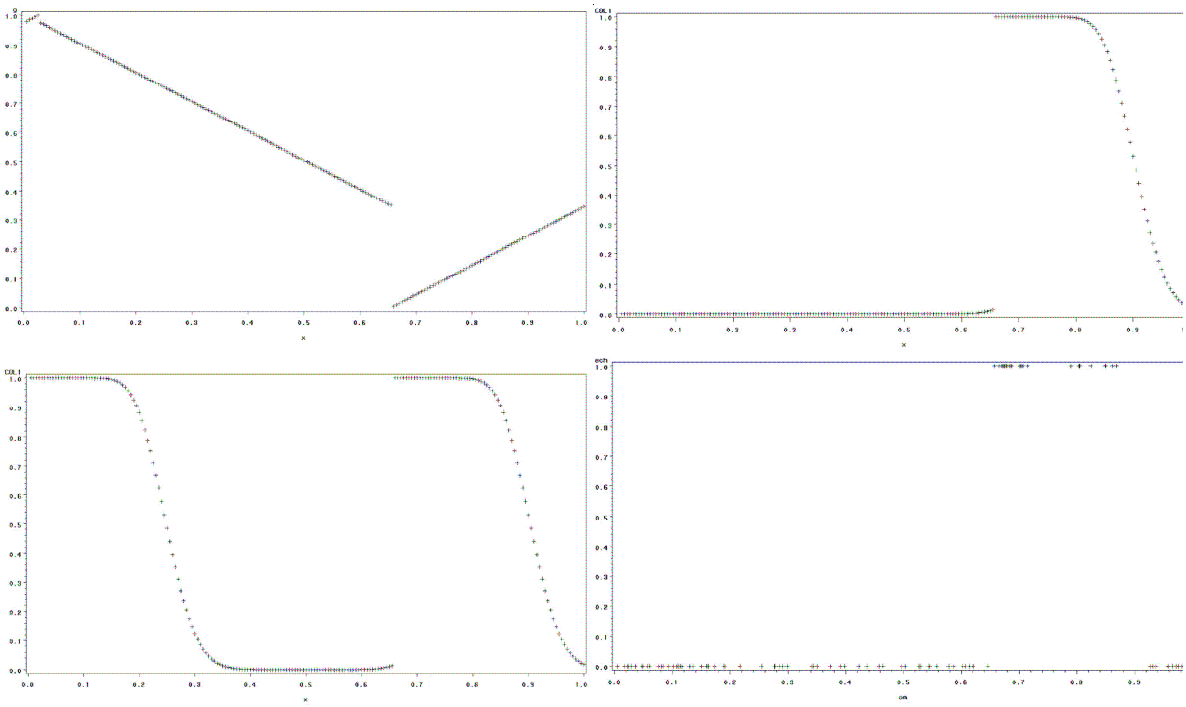
En d'autres termes, on a sélectionné les *n* unités ayant les plus petits numéros aléatoires transformés $g_k(\omega_k)$, où $g_k(\omega_k) = \omega_k$. La fonction de coordination g_k est donc simplement l'identité sur l'intervalle [0 ; 1], comme on peut le vérifier sur le graphique en haut à gauche.

L'espérance d'appartenance d'une unité k au premier échantillon, en fonction de la valeur de son propre numéro aléatoire ω_k, peut alors se calculer explicitement. C'est en effet la probabilité que, parmi les N-1 autres unités, il y en ait au plus *n*-1 dont le numéro aléatoire soit inférieur à ω_k. Autrement dit, c'est la probabilité que ω_k soit inférieur à la *n*-ième statistique d'ordre $\Omega_{-k}^{(n)}$ définie sur les N-1 numéros aléatoires des unités de la population autres que k :

$I_k^{'}(\omega_k) = E(I_k(\Omega)|\Omega_k = \omega_k) = P(\Omega_{-k}^{(n)} \geq \omega_k)$. Ces derniers étant indépendants de loi uniforme standard, leur *n*-ième statistique d'ordre suit une loi beta de paramètres *n* et N-*n*+1. L'espérance d'appartenance d'une unité k au premier échantillon, $I_k^{'}(\omega_k)$, apparaît alors simplement comme la fonction de survie de cette loi beta, fonction dont on peut aisément reconnaître la forme caractéristique sur le graphique en haut à droite. Le taux de sondage étant dans le cas présent égal à ¼, il n'est d'ailleurs pas surprenant d'observer que ce sont précisément les unités dont le numéro aléatoire est proche de ¼ pour lesquelles le fait d'appartenir à l'échantillon est incertain *a priori*.

Enfin, la charge de réponse liée à la première enquête étant unitaire, la charge de réponse cumulée espérée après le premier tirage est simplement égale à l'espérance d'être échantillonné, d'où le fait que les graphiques en haut à droite et en bas à gauche soient identiques !

2ème ETAPE : Tirage n°2 taille échantillon = 25 charge de réponse = 1



Rappelons que le second tirage va consister à retenir les *n* unités de plus petits numéros aléatoires transformés $g_k(\omega_k)$. La fonction de coordination g_k est donc naturellement construite de sorte qu'elle attribue un numéro aléatoire transformé d'autant plus petit que la charge cumulée espérée sur le passé pour l'unité k est faible : on privilégie la sélection des unités qui ont eu très peu de chances d'être sélectionnées la première fois. On a vu qu'à l'issue de la première étape, cette charge cumulée espérée est une fonction décroissante sur l'intervalle [0;1] (fonction de survie d'une loi beta). En toute logique, on devrait donc construire une fonction de coordination décroissante sur cet intervalle, qui, devant en outre préserver la probabilité uniforme, n'est autre que $g_k(\omega_k) = 1 - \omega_k$. Les unités sélectionnées seraient donc celles ayant les plus grands numéros aléatoires, c'est-à-dire celles qui ont eu les plus faibles probabilités d'être sélectionnées lors du premier tirage.

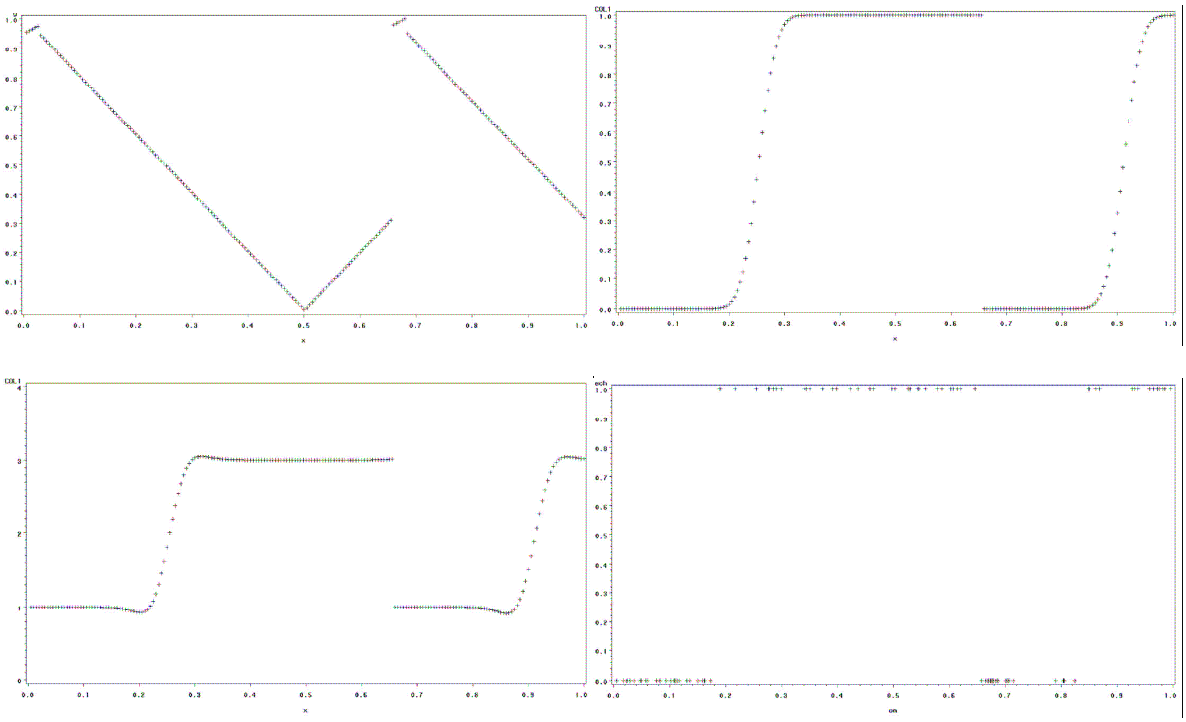
Cependant, le graphique en haut à gauche pour le second tirage présente visiblement une fonction de coordination légèrement différente. Cela est dû au fait que le programme informatique utilisé arrondit les nombres « trop proches » de 0 ou 1 à ces entiers. Plutôt que de considérer la charge cumulée espérée après premier tirage comme fonction décroissante sur tout l'intervalle [0;1], le programme informatique l'estime constante égale à 1 sur un intervalle proche de [0;0,3] et constante égale à 0 sur un intervalle proche de [0,66;1]. Comme cela a été mentionné précédemment dans ce papier, on traite le cas des paliers (domaines de numéros aléatoires pour lesquels la charge cumulée espérée est constante) en construisant sur ceux-ci une fonction de coordination croissante, sans que cela n'altère

de quelque façon que ce soit les propriétés théoriques de la méthode de coordination d'échantillons. C'est exactement ce que l'on observe et ce qui justifie la forme de la fonction de coordination pour le second tirage (en haut à gauche).

Plutôt que de privilégier la sélection des unités aux plus grands numéros aléatoires, on va donc favoriser celle des unités dont le numéro aléatoire est proche de et supérieur à 0,66. La coordination négative entre les deux premiers échantillons est tout aussi correctement atteinte dans la mesure où ces unités sont précisément celles pour lesquelles la probabilité d'être tirée dans le premier échantillon était suffisamment faible pour être arrondie à 0 par le programme informatique. L'observation des graphes de droite confirme clairement que les unités sélectionnées, en espérance comme dans les faits, lors du deuxième tirage sont celles que l'on vient de mentionner.

La charge de réponse pour la seconde enquête étant à nouveau unitaire, la charge de réponse cumulée espérée après second tirage (graphique en bas à gauche pour le second tirage) est donc simplement la somme de celle obtenue après le premier tirage (graphique en bas à gauche pour le premier tirage) et de l'indicatrice espérée d'appartenance au second échantillon (graphique en haut à droite pour le deuxième tirage).

3ème ETAPE : Tirage n°3 taille échantillon = 50 charge de réponse = 3



Pour ce troisième tirage, le principe demeure le même. La charge de réponse cumulée espérée après le deuxième tirage est la plus faible pour les numéros aléatoires proches de 0,5. La fonction de coordination pour le troisième tirage sera donc proche de 0 pour ces numéros aléatoires, favorisant de fait leur sélection dans l'échantillon. On laissera le lecteur se convaincre par lui-même que les valeurs prises par la fonction de coordination présentée ici sont d'autant plus faibles (resp. élevées) que la charge cumulée espérée après second tirage exposée à l'étape précédente est faible (resp. élevée). Sa forme affine par morceaux, où les pentes sont toutes égales à ±1, garantit en outre à la fonction de coordination qu'elle préserve bien la probabilité uniforme (Ceci étant dû au fait que, contrairement aux apparences, la fonction de coordination figurant en haut à gauche pour le troisième tirage est bel et bien bijective, aucun point des différents segments de « droite » figurant sur le graphique n'ayant la même ordonnée. L'impression de « non-bijectivité » émanant du graphique est en quelque sorte une illusion d'optique due au fait que le pas choisi pour discrétiser l'intervalle [0;1] est très petit).

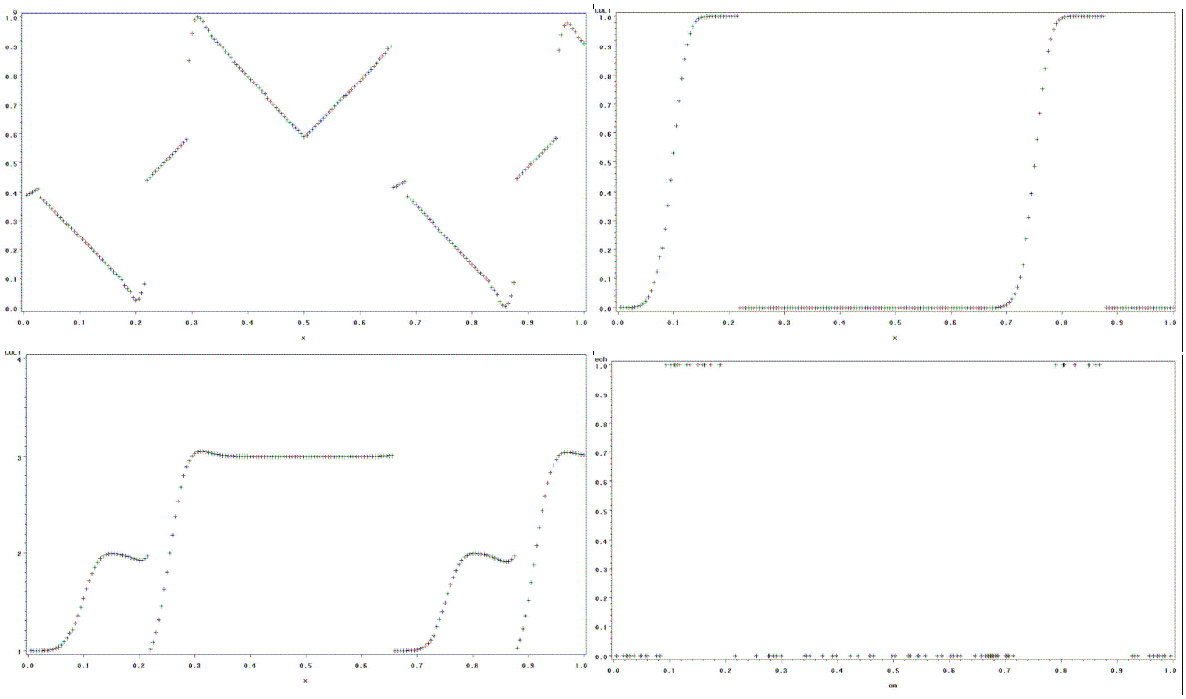
Le taux de sondage étant cette fois-ci de ½, le domaine de numéros aléatoires pour lesquels les unités ont de très fortes chances d'être sélectionnées recouvre à présent environ la moitié de

l'intervalle [0 ;1] (mathématiquement, c'est un borélien de mesure de Lebesgue proche de ½). Les unités sélectionnées étant par construction celles ayant eu de très faibles probabilités d'être dans les précédents échantillons, leur charge de réponse cumulée espérée après le troisième tirage est naturellement environ égale à la charge de réponse pour la troisième enquête, c'est-à dire 3. C'est bien l'ordonnée des paliers supérieurs que l'on observe sur le graphique en bas à gauche.

Sur ce même graphique, on peut observer des légers « creux » ou « bosses » lors des transitions entre paliers. Par exemple, la charge de réponse cumulée espérée après le troisième tirage semble être légèrement supérieure à 3 pour les unités de numéros aléatoires voisins de 0,3. De fait, ces unités sont celles sélectionnées presque sûrement pour le troisième échantillon, mais dont la probabilité d'être sélectionnées dans l'un des deux précédents échantillons (en l'occurrence dans le premier pour le cas examiné ici) s'avérait très faible tout en étant non négligeable.

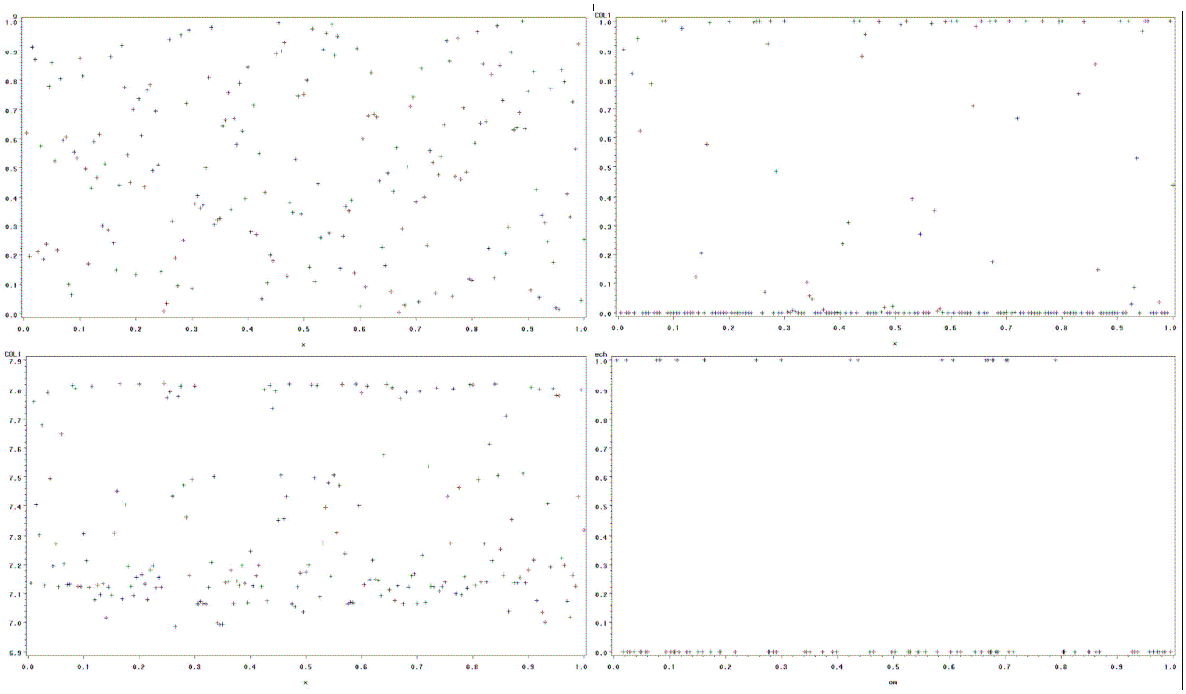
Les trois premiers tirages ayant été effectués avec des taux de sondage respectifs de ¼, ¼ et ½, une parfaite coordination négative des échantillons garantirait que les trois premiers échantillons forment une partition de la population : chaque unité appartiendrait à un et un seul de ces échantillons. **Le phénomène des « creux » et des « bosses » décrit précédemment met en évidence le fait qu'une méthode de coordination probabiliste ne peut pas garantir cela avec certitude ; c'est le prix à payer pour obtenir des échantillons sur lesquels on puisse mener des inférences sans biais.** Il est possible que dans l'exemple étudié ici, quelques unités se retrouvent simultanément dans le premier et le troisième échantillons (unités de la « bosse » observée pour les numéros aléatoires proches de 0,3) ou dans le second et le troisième échantillons (unités de la « bosse » observée pour les numéros aléatoires proches de 0,95), tandis qu'en contrepartie, quelques autres n'apparaissent dans aucun des trois échantillons (unités des « creux » observés pour les numéros aléatoires proches de 0,21 et 0,87)

4ème ETAPE : Tirage n°4 taille échantillon = 25 charge de réponse = 1



Dès le quatrième tirage, la forme de la fonction de coordination commence à être relativement complexe. Sans surprise, les unités correspondants aux « creux » observés précédemment sur la fonction de charge de réponse cumulée espérée après le troisième tirage sont celles qui vont être sélectionnées en priorité dans le quatrième échantillon (valeurs les plus faibles pour la fonction de coordination en haut à gauche), tandis que celles correspondants aux « bosses » se voient affectées des plus fortes valeurs de la fonction de coordination. A la suite de ce quatrième tirage, la charge cumulée espérée s'étend de 1 à une valeur légèrement supérieure à 3.

DERNIERE ETAPE : Tirage n°20 taille échantillo n = 25 charge de réponse = 1



Même dans le cas le plus simple possible, tirages aléatoires simples non stratifiés sur une population invariante dans le temps, les graphiques de la 20ème et dernière étape de la simulation mettent en évidence l'aspect particulièrement complexe des fonctions sur lesquelles s'appuie la méthode de coordination d'échantillons présentée dans ce papier. Le qualificatif « complexe » est bien sûr à entendre dans le sens où, au-delà de quelques étapes, il est pour ainsi dire impossible de calculer analytiquement les fonctions de coordination permettant ensuite de sélectionner les échantillons ; en revanche, l'implémentation informatique d'un tel calcul ne pose aucune difficulté.

Après les 20 tirages, on observera que la charge de réponse cumulée espérée finale (en bas à gauche) est toujours comprise entre 6,9 et 7,9, à rapprocher de la charge moyenne par unité annoncée au début de cette partie, égale à 7,4. Ces chiffres peuvent être mis en perspective avec les charges réelles, observées a posteriori, i.e. après que les tirages ont été réalisés :

Charge de réponse réelle, observée après les 20 tirages	Nombre d'unités
4	1
6	7
7	57
8	23
9	9
10	3

6.2. Evaluation empirique de la méthode sur données simulées

6.2.1. Indicateurs de performance de la méthode

Le prototype développé en SAS a été conçu pour pouvoir sélectionner dans une population donnée un échantillon selon un tirage aléatoire simple stratifié (TASST, plan de sondage usuel pour les enquêtes auprès des entreprises de l'Insee), coordonné avec un ensemble paramétré d'enquêtes du passé dont les échantillons ont tous été tirés selon la même procédure. La population peut connaître des évolutions démographiques (le champ des unités d'intérêt peut évoluer d'une enquête à l'autre) ; la stratification de la population peut également différer totalement d'une enquête à l'autre permettant de traiter notamment le cas particulier où la stratification est la même pour deux enquêtes, mais où certaines unités changent de strates entre les deux enquêtes.

L'objet des quelques simulations dont on expose les résultats dans la suite de cette partie est d'évaluer l'efficacité de la méthode de coordination, en fonction des divers paramètres caractérisant le plan de sondage ou la population d'unités statistiques.

Une simulation donnée va consister à effectuer 20 tirages coordonnés et 20 tirages indépendants (pour 20 enquêtes) au sein d'une population puis de confronter les échantillons obtenus de ces deux façons. Afin de prendre en compte les phénomènes démographiques et la variabilité des paramètres liés aux enquêtes et aux plans de sondage, on suit plus précisément la procédure suivante :

Pour une simulation :

- On définit une population d'intérêt U, dite intertemporelle, de taille N (100, 1000 ou 10000 selon les cas)
- Chacune des 20 enquêtes E_i, i = 1…20, se voit affectée aléatoirement d'un taux de couverture ς_i, compris entre deux bornes paramétrables. Le champ de E_i, c'est-à-dire l'ensemble des unités susceptibles d'être sélectionnées dans l'échantillon de l'enquête E_i, est alors constitué d'environ $\varsigma_i N$ unités choisies aléatoirement dans la population intertemporelle U. De la sorte, les champs des différentes enquêtes ne se recouvrent pas nécessairement parfaitement, simulant ainsi aléatoirement les phénomènes de naissances, décès d'unités, d'unités sortant ou entrant dans le champ,etc. Le paramètre $1-\varsigma_i$ sera appelé, certes un peu abusivement, démographie de la population liée à l'enquête E_i.
- Pour chacune des 20 enquêtes, une stratification est établie aléatoirement pour l'ensemble des unités dans le champ de l'enquête, un paramètre permettant de contrôler la taille moyenne - et donc leur nombre - des strates souhaitées.
- Pour chacune des 20 enquêtes et chacune de leurs strates, on choisit aléatoirement des taux de sondage, compris entre deux bornes paramétrables.
- Il reste enfin à définir le plan de coordination. Autrement dit, pour chaque enquête E_i, il s'agit de définir avec quelles enquêtes E_k, k<i, du passé l'enquête E_i doit être coordonnée et quelles charges de réponse $C_{k,i}$ leur sont attribuées.

Les 20 tirages d'échantillons indépendants et les 20 tirages d'échantillons coordonnés peuvent alors être effectués. Pour le cas indépendant comme pour le cas coordonné, on peut alors déterminer la **répartition (ou distribution) des unités de la population en fonction du nombre d'échantillons dans lesquels elles ont été sélectionnées (i.e. la distribution sur la population de la variable « nombre de sélections »)** : combien d'unités ont été sélectionnées une fois, combien ont été sélectionnées deux fois, etc. Cela conduit à dresser les deux histogrammes de la Figure 1 (correspondant à la simulation de la dernière ligne du Tableau 6 ci-après).

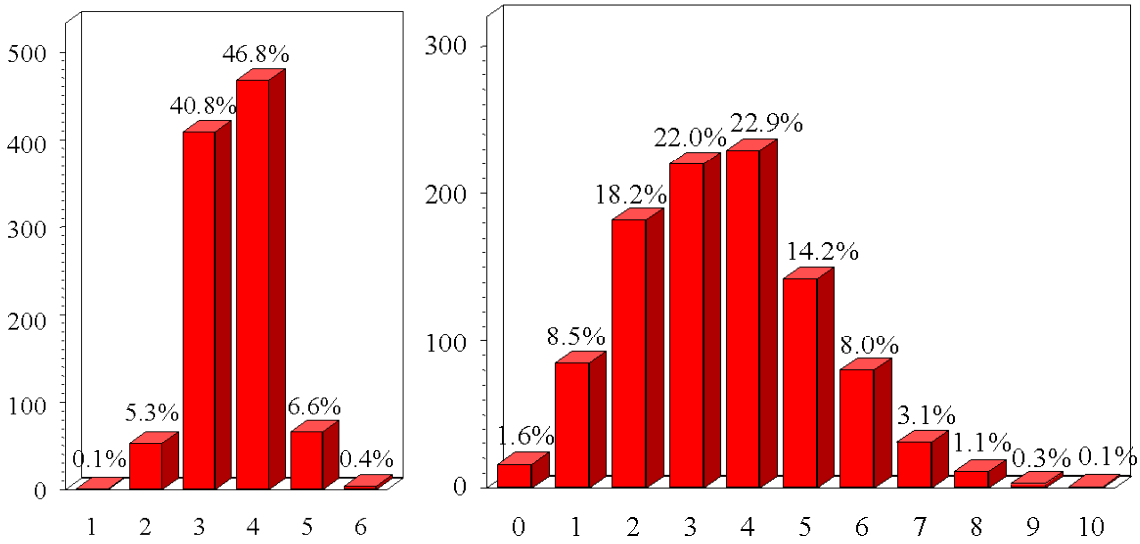


Figure 1 : *Histogrammes de répartition (ou distribution) des unités de la population en fonction du nombre d'échantillons dans lesquels elles ont été sélectionnées. A gauche, cas des des tirages coordonnés. A droite, cas des tirages indépendants. La population et les plans de sondages testés sont ceux correspondant aux résultats de la dernière ligne du* **Tableau 6** *ci-après.*

Les deux histogrammes de la Figure 1 sont emblématiques puisqu'ils mettent en évidence une situation où la coordination est très efficace.

En effet, compte tenu du fait que les tailles d'échantillons sont exactement les mêmes selon que les tirages sont réalisés indépendamment ou de façon coordonnée, le nombre moyen d'échantillons auquels appartient une unité quelconque de la population est le même dans les deux cas. Autrement dit, les moyennes des distributions représentées par les deux histogrammes de la Figure 1 sont nécessairement égales.

La qualité de la méthode de coordination va donc s'apprécier sur les moments d'ordre supérieurs de ces distributions. En l'occurrence, on observe ici très clairement que la distribution sur la population de la variable « nombre de sélections » se concentre beaucoup plus fortement autour de sa moyenne lorsque les tirages sont coordonnés. Cela signifie exactement que la méthode de coordination répartit autant que possible la charge de réponse sur l'ensemble des unités de la population. Dans le cas de la Figure 1, sachant qu'en moyenne une unité est sélectionnée 3,55 fois sur l'ensemble des 20 enquêtes, la méthode de coordination conduit à la sélection de plus de 87% des unités dans trois ou quatre échantillons seulement ; seules 6,4% des unités sont sélectionnées cinq fois et 0,4% six fois. Lorsque les tirages sont indépendants, de telles performances sont loin d'être atteintes, quelques unités étant même sélectionnées huit, neuf voire dix fois. La qualité de la coordination se mesure donc quantitativement au fait que l'écart-type de la distribution de l'histogramme de gauche de la Figure 1 est beaucoup plus faible que celui de droite.

Aussi définit-on l'indicateur principal de performance de la méthode de coordination de la façon suivante :

$$R_\sigma = \frac{\sigma_{Coord}}{\sigma_{Indep}}$$

où σ_{Coord} et σ_{Indep} sont les écarts-types des distributions sur la population de la variable « nombre de sélections », lorsque les tirages sont respectivement coordonnés et indépendants. Un ratio R_σ inférieur à 1 signifie que l'on répartit mieux la charge de réponse sur les entreprises en réalisant des tirages coordonnés, la méthode de coordination étant d'autant plus efficace que R_σ est proche de 0. L'avantage de cet indicateur est que les

valeurs numériques qu'il prend sont directement interprétables : lorsqu'il est inférieur à 1, la coordination d'échantillons permet de réduire de (1-R_σ)% la dispersion des charges sur les unités de la population.

Par ailleurs, rappelons que le but premier de la coordination d'échantillons est avant tout d'éviter des situations où des unités auraient une charge de réponse trop élevée (répartir de façon homogène la charge sur toutes les unités de la population n'est pas un but en soi, mais un moyen d'atteindre cet objectif). On peut donc s'attendre à ce que, dans le cas où les tirages sont coordonnés, la distribution sur la population de la variable « nombre de sélections » s'étale moins vers la droite, c'est-à-dire vers les valeurs supérieures à la moyenne : une bonne méthode de coordination doit éviter le tirage trop fréquent d'une même unité.

Ceci amène à définir un indicateur secondaire de performance de la méthode de coordination :

$$\Delta_s = S_{Coord} - S_{Indep}$$

où S_{Coord} et S_{Indep} sont les skewness (coefficients d'asymétrie) des distributions sur la population de la variable « nombre de sélections », lorsque les tirages sont respectivement coordonnés et indépendants. L'étalement de la distribution vers la droite étant espéré plus faible dans le cas de tirages coordonnés, une bonne coordination des échantillons devrait se traduire par une valeur négative de Δ_S. La valeur numérique de ce second indicateur est en revanche plus délicate à interpréter, aussi insistons-nous sur le fait que c'est davantage son signe qui mérite d'être examiné.

6.2.2. Impact des paramètres du plan de sondage et des caractéristiques de la population sur la qualité de la procédure de tirage

6.2.2.1. Impact du pas de discrétisation

La méthode de coordination des échantillons repose sur le calcul, pour chaque unité, de différentes fonctions (charge cumulée espérée, indicatrice espérée d'appartenance à l'échantillon, fonction de coordination) qui sont définies sur l'intervalle continu [0 ;1]. Sur le plan opérationnel, i.e. pour les calculs pratiques, seul un nombre discret fini de valeurs de ces fonctions peuvent être calculées. Dans cette première sous-partie, on analyse l'efficacité de la coordination en fonction du nombre L de sous-intervalles partitionnant [0 ;1] sur chacun desquels les différentes fonctions sont assimilées à des constantes ou des fonctions affines.

Deux types de simulations ont été effectuées : dans le premier cas, sur une population intertemporelle de taille 100, sans aucune stratification (Tableau 1), dans le second cas sur une population intertemporelle de taille 1000 où les plans de sondage pour les tirages successifs sont stratifiés (Tableau 2).

Les tableaux recensent, pour chaque simulation réalisée, les caractéristiques (moyenne, écart-type, skewness) des deux distributions sur la population de la variable « nombre de sélections », celle dans le cas de tirages indépendants et celle dans le cas de tirages coordonnés. On rappelle, comme cela a été expliqué en partie 6.2.1, que ces deux distributions ont même moyenne. Figurent également dans ces tableaux les valeurs prises par les deux indicateurs de performance explicités précédemment.

	Moyenne	σ_{Coord}	σ_{Indep}	S_{Coord}	S_{Indep}	R_σ	Δ_S
L = 10	3,190	0,873	1,482	0,362	0,274	***0,589***	***0,088***
L = 50	3,190	0,761	1,426	0,224	0,169	***0,534***	***0,055***
L = 100	3,190	0,849	1,686	0,128	0,843	***0,504***	***-0,715***
L = 200	3,190	0,734	1,862	0,153	0,320	***0,394***	***-0,167***
L = 500	3,190	0,873	1,680	-0,289	0,555	***0,519***	***-0,844***
L = 1000	3,190	0,734	1,587	0,465	0,315	***0,463***	***0,150***

Tableau 1 : ***Impact du pas de discrétisation de l'intervalle [0 ;1], découpé en L sous-intervalles***
Population non stratifiée, de taille N = 100.
Démographie de l'ordre de 14% (de 10% à 18%).
Taux de sondage de l'ordre de 20% (de 15% à 25%)
Coordination avec tout le passé (20 enquêtes au total), charge constante

	Moyenne	σ_{Coord}	σ_{Indep}	S_{Coord}	S_{Indep}	R_σ	Δ_S
L = 10	3,514	0,947	1,711	0,140	0,299	***0,554***	***-0,158***
L = 50	3,514	0,689	1,736	0,152	0,515	***0,397***	***-0,363***
L = 100	3,514	0,683	1,737	0,120	0,351	***0,393***	***-0,232***
L = 200	3,514	0,692	1,684	0,132	0,340	***0,411***	***-0,208***
L = 500	3,514	0,673	1,748	0,236	0,456	***0,385***	***-0,220***
L = 1000	3,514	0,677	1,705	0,260	0,392	***0,397***	***-0,132***

Tableau 2 : ***Impact du pas de discrétisation de l'intervalle [0 ;1], découpé en L sous-intervalles***
Population stratifiée, de taille N = 1000. Taille moyenne des strates : 200
Démographie de l'ordre de 14% (de 10% à 18%).
Taux de sondage de l'ordre de 20% (de 15% à 25%)
Coordination avec tout le passé (20 enquêtes au total), charge constante

Globalement, les résultats sont très encourageants, puisque la méthode de coordination semble très efficace, R_σ étant systématiquement nettement inférieur à 1 - y compris, de façon assez surprenante, lorsque l'approximation des calculs de fonctions est très grossière (L = 10). **La dispersion des charges de réponse sur les unités de la population est en général réduite de 50 à 60% en coordonnant les tirages !**

L'indicateur secondaire Δ_S est lui-même très souvent négatif, comme souhaité, excepté quelques cas du tableau 1, pour lesquels il faut garder en mémoire que l'on travaille sur une population de taille extrêmement faible (N = 100). Même si l'impact du pas de discrétisation de l'intervalle [0 ;1] sur la qualité de la coordination semble relativement limité à la lecture de ces résultats, des simulations complémentaires ont établi qu'il était tout de même préférable de choisir un paramètre L d'ordre de grandeur au moins 100.

6.2.2.2. Impact du taux de sondage

On teste à présent la méthode de coordination pour différentes valeurs du taux de sondage, identique pour les 20 enquêtes successives. Les résultats sont consignés dans le Tableau 3.

	Moyenne	σ_{Coord}	σ_{Indep}	S_{Coord}	S_{Indep}	R_σ	Δ_S
f = 0,01	0,113	0,317	0,329	2,448	2,781	***0,962***	***-0,332***
f = 0,05	0,839	0,462	0,879	-0,427	0,967	***0,526***	***-1,393***
f = 0,10	1,738	0,572	1,279	-0,186	0,691	***0,447***	***-0,877***
f = 0,20	3,569	0,701	1,714	0,251	0,495	***0,409***	***-0,244***
f = 0,40	7,105	0,865	2,176	-0,065	0,248	***0,398***	***-0,313***
f = 0,60	10,728	1,073	2,322	-0,173	-0,006	***0,462***	***-0,168***
f = 0,80	14,175	1,282	2,091	-0,390	-0,170	***0,613***	***-0,219***
f = 0,95	17,129	1,337	1,535	-0,388	-0,487	***0,871***	***0,100***
f = 0,99	17,733	1,427	1,432	-0,554	-0,538	***0,997***	***-0,016***

Tableau 3 : Impact du taux de sondage f
Discrétisation de l'intervalle [0 ;1] : L = 100
Population stratifiée, de taille N = 1000. Taille moyenne des strates : 150
Démographie de l'ordre de 10% (de 6% à 14%).
Coordination avec tout le passé (20 enquêtes au total), charge constante

Les résultats de ce second lot de simulations confirment les premières conclusions, à savoir que la méthode de coordination utilisée répartit efficacement la charge de réponse sur l'ensemble des unités de la population et évite en priorité d'affecter à des unités une charge de réponse trop supérieure à la moyenne : l'indicateur principal R_σ est systématiquement inférieur à 1, l'indicateur secondaire Δ_S négatif (excepté dans un cas).

Par ailleurs, conformément à l'intuition que l'on peut en avoir, la méthode de coordination n'apporte que très peu d'amélioration dans le cas où le taux de sondage est très faible ou très élevé. Lorsque le taux de sondage est très proche de 0 (respectivement 1), la coordination n'a plus réellement d'utilité puisque les unités, même dans le cas de tirages indépendants, ont une probabilité extrêmement faible d'être échantillonnées à deux reprises (respectivement d'être absentes de l'un des échantillons). Dans les cas limites f = 0 et f = 1, coordonner n'a évidemment plus aucun intérêt, le ratio R_σ sera égal à 1 et la différence Δ_S égale à 0, puisque les distributions sous-jacentes dans les cas coordonnés et indépendants seront parfaitement identiques.

Sur les simulations du Tableau 3, on observe que **la coordination permet en outre de réaliser des gains de plus de 50% en termes de répartition de la charge sur une plage de taux de sondage assez large**, en l'occurrence au moins pour des taux de sondage compris entre 0,05 et 0,60 ; la méthode demeure encore très performante pour un taux de 0,8, le gain avoisinant les 40% (le ratio R_σ étant encore proche de 0,6). Signalons cependant que la vitesse avec laquelle l'efficacité de la cooordination d'échantillons diminue lorsque le taux de sondage s'approche de 1 ou 0 dépend aussi du nombre d'enquêtes avec lesquelles on coordonne (voir la partie 6.2.2.4).

6.2.2.3. Robustesse de la coordination vis-à-vis de paramètres très volatils : population très mouvante dans le temps, taux de sondage très variables d'une enquête à l'autre

Quelques tests ont été effectués sur des populations de taille plus importante (population intertemporelle de taille 10000), mais en simulant de très fortes évolutions démographiques (taux de couverture variant entre 10% et 90% selon les enquêtes, i.e. démographie variant dans le même

intervalle) et en choisissant des taux de sondages très variables d'une enquête à l'autre (variant de 1% à 99%).

Deux principaux tests ont été réalisés, correspondant aux tableaux ci-dessous ; pour le premier, les échantillons sont coordonnés avec ceux de toutes les enquêtes précédentes ; pour le second seulement avec ceux des cinq enquêtes précédentes (avec tout le passé pour les cinq premières enquêtes).

La coordination demeure globalement très efficace, même si les gains sont moins substantiels que ceux observés dans les simulations précédentes : Δ_S est toujours négatif, R_σ demeure inférieur à 1, avec cependant des valeurs plus élevées qu'auparavant. **La méthode de coordination semble relativement robuste vis-à-vis de la volatilité des paramètres caractéristiques de la population et du plan de sondage.**

	Moyenne	**σ_{Coord}**	**σ_{Indep}**	**S_{Coord}**	**S_{Indep}**	**R_σ**	**Δ_S**
-	3,906	1,153	1,689	0,132	0,287	***0,683***	***-0,156***

Tableau 4 : Grande variabilité dans les caractéristiques de la population et les taux de sondage : Démographie variant entre 10% et 90%, taux de sondage variant entre 1% et 99%
Discrétisation de l'intervalle [0 ;1] : L = 100
Population stratifiée, de taille N = 10000. Taille moyenne des strates : 400
Coordination avec tout le passé (20 enquêtes au total), charge constante

	Moyenne	**σ_{Coord}**	**σ_{Indep}**	**S_{Coord}**	**S_{Indep}**	**R_σ**	**Δ_S**
-	5,704	1,371	1,810	-0,002	0,187	***0,757***	***-0,188***

Tableau 5 : Grande variabilité dans les caractéristiques de la population et les taux de sondage : Démographie variant entre 10% et 0,90%, taux de sondage variant entre 1% et 99%
Discrétisation de l'intervalle [0 ;1] : L = 100
Population stratifiée, de taille N = 10000. Taille moyenne des strates : 400
Coordination avec les cinq enquêtes précédentes (20 enquêtes au total), charge constante

6.2.2.4. Impact de la persistance et de la variabilité de la charge de réponse

Dans le Tableau 6 sont présentés les résultats de simulations où l'on teste l'effet lié à la persistance de la charge. Une ligne du tableau est caractérisée par la valeur du paramètre NCoord : pour la simulation correspondante, le tirage d'un échantillon est alors coordonné avec les échantillons des NCoord enquêtes précédentes (ou avec tout le passé pour les NCoord premières enquêtes).

Les résultats du Tableau 6 soulignent la nécessité de coordonner avec un nombre suffisamment important d'enquêtes du passé ; en coordonnant systématiquement avec au moins les quatre enquêtes précédentes, on obtient des performances similaires à celles observées lors des précédentes simulations, avec un indicateur R_σ nettement inférieur à 1, oscillant entre 0,4 et 0,5 et un indicateur Δ_S demeurant négatif.

Lorsque l'on coordonne avec trop peu d'enquêtes précédentes, la méthode de coordination est contre-productive car elle a plutot tendance à concentrer la charge de réponse sur certaines unités seulement.

Considérons le cas de la première simulation où chaque échantillon est simplement coordonné avec celui de la précédente enquête. Lors du premier tirage seront sélectionnées les unités à très faible numéro aléatoire ; l'application de la méthode de coordination fait que ces mêmes unités auront les probabilités les plus faibles (en fait quasi-nulles) d'être sélectionnées lors du second tirage ; mais

alors, comme le troisième échantillon n'est coordonné qu'avec le deuxième, elles auront de nouveau les plus fortes probabilités (proches de 1) d'être sélectionnées dans le troisième échantillon.

Ce raisonnement peut être itéré et met en évidence un phénomène pervers, puisque contraire aux objectifs de coordination souhaités : la coordination avec un nombre trop restreint d'enquêtes engendre des **« bassins d'attraction »** - domaines de numéros aléatoires inclus dans [0 ;1] pour lesquels les unités vont avoir périodiquement de très fortes probabilités de sélection conditionnelles (1 fois toutes les (NCoord+1) enquêtes) - ainsi que des **« bassins de répulsion »**, domaines de numéros aléatoires inclus dans [0 ;1] pour lesquels les unités auront systématiquement des probabilités de sélection conditionnelles très faibles.

Dans le cas particulier d'une population ne connaissant pas d'évolutions démographiques et où le taux de sondage constant d'une enquête à l'autre est significativement inférieur à 1/(NCoord+1), l'existence de ces bassins d'attraction et de répulsion peut être démontrée assez facilement sur le plan théorique. Dans le cas général, **ceci suggère au passage de choisir le nombre *NCoord* d'enquêtes passées avec lesquelles coordonner selon une règle du type suivant,**

$$NCoord > \frac{1}{\bar{f}} - 1$$

où $\bar{f}$ est le taux de sondage moyen sur les différentes enquêtes concernées (Noter que pour les simulations du Tableau 6, le taux de sondage moyen est égal à 1/5, ce qui conduit à NCoord > 4).

Ces bassins d'attraction et de répulsion expliquent la dégradation qu'apporte la coordination des tirages d'échantillons dans les deux premières lignes du Tableau 6, le ratio R_σ y étant supéreur à 1 (et même à 2 dans le premier cas). En examinant la distribution sur la population, non présentée dans ce papier, de la variable « nombre de sélections » lorsque la coordination ne s'effectue qu'avec l'enquête précédente, on observe d'ailleurs une proportion particulièrement élevée d'unités sélectionnées à 10 reprises, c'est-à-dire une fois sur deux.

	Moyenne	σ_{Coord}	σ_{Indep}	S_{Coord}	S_{Indep}	R_σ	Δ_S
NCoord = 1	3,557	3,905	1,704	0,481	0,308	***2,292***	***0,173***
NCoord = 2	3,557	2,306	1,723	0,025	0,424	***1,338***	***-0,400***
NCoord = 3	3,557	1,387	1,756	-0,612	0,403	***0,790***	***-1,015***
NCoord = 4	3,557	0,835	1,676	-0,506	0,491	***0,498***	***-0,998***
NCoord = 5	3,557	0,816	1,718	0,110	0,266	***0,475***	***-0,156***
NCoord = 7	3,557	0,859	1,722	0,037	0,379	***0,499***	***-0,343***
NCoord = 9	3,557	0,748	1,723	-0,310	0,388	***0,434***	***-0,698***
NCoord = 12	3,557	0,777	1,647	0,272	0,388	***0,472***	***-0,116***
NCoord = 15	3,557	0,787	1,670	-0,226	0,333	***0,471***	***-0,559***
NCoord = 19	3,557	0,718	1,674	0,027	0,369	***0,429***	***-0,342***

Tableau 6 : Impact de la persistance de la charge, caractérisée par le nombre NCoord des enquêtes précédentes avec lesquelles le tirage pour une enquête donnée est coordonné
Discrétisation de l'intervalle [0 ;1] : L = 100
Population stratifiée, de taille N = 1000. Taille moyenne des strates : 150
Démographie de l'ordre de 10% (de 8% à 12%).
Taux de sondage de l'ordre de 20% (de 18% à 22%)
Charge constante

Dans les trois derniers jeux de simulations présentés ci-après, on s'intéresse cette fois-ci à l'impact de la charge attribuée aux différentes enquêtes du passé lorsque l'on tire un échantillon.

Dans les Tableau 7 et Tableau 8, la charge de réponse attribuée à chaque enquête est choisie aléatoirement parmi plusieurs valeurs ; dans le cas le plus extrême (dernière ligne des deux tableaux), on autorise la charge de réponse de certaines des 20 enquêtes à être jusqu'à six fois supérieures à celles d'autres enquêtes. La différence entre les deux tableaux réside dans le fait que, pour le premier, la coordination se fait avec tout le passé tandis que, pour le second, les échantillons sont seulement coordonnés avec les cinq précédents.

Enfin, le Tableau 9 présente les résultats obtenus lorsque l'on coordonne également avec les seules cinq enquêtes précédentes, mais en considérant des profils de charge de réponse particuliers : charge constante, charge décroissant linéairement avec l'antériorité de l'enquête (pour le tirage de l'échantillon de l'enquête E_i, on attribue une charge $C_{k,i} = 1 - \left[(i-k-1)/5\right]$ à l'enquête E_k, pour k compris entre i-5 et i-1), et enfin charge décroissant exponentiellement avec l'antériorité de l'enquête (pour le tirage de l'échantillon de l'enquête E_i, on attribue une charge $C_{k,i} = a^{-(i-k-1)}$ à l'enquête E_k, pour k compris entre i-5 et i-1, le paramètre *a* désignant la base d'exponentiation). Des profils de charge de réponse décroissant avec l'antériorité de l'enquête permettent en effet de simuler les cas où l'on privilégie la coordination avec les échantillons les plus récents : on accepte plus facilement qu'une unité soit rééchantillonnée si la dernière fois qu'elle a été tirée correspond à une enquête déjà assez ancienne.

L'ensemble des résultats de ces trois jeux de simulation fournissent des résultats similaires, confirmant la nette efficacité de la méthode de coordination : grosso modo un coefficient R_σ compris entre 0,4 et 0,5 (soit 50% à 60% de gains en termes de répartition des unités en fonction de leur fréquence d'échantillonnage), un coefficient Δ_S toujours négatif à une exception près.

Il est en revanche difficile de détecter à partir de ces trois tableaux de résultats une quelconque influence de la variation de la charge de réponse sur les statistiques et indicateurs de performance présentés. En réalité, comme les charges de réponses ne sont plus identiques d'une enquête à l'autre pour ces dernières simulations, la charge cumulée de réponse d'une unité ne peut plus être assimilée au nombre de fois où l'unité a été sélectionnée dans un échantillon. **Ainsi, on pourrait préférer calculer ici les statistiques et indicateurs de performance relatifs aux distributions sur la population de la variable « charge de réponse cumulée », et non de la variable « nombre de sélections ». Cela n'est pas présenté dans ce papier, cependant les résultats obtenus de la sorte confirment encore davantage l'efficacité de la méthode de coordination.**

	Moyenne	**σ_{Coord}**	**σ_{Indep}**	**S_{Coord}**	**S_{Indep}**	**R_σ**	**Δ_S**
charge = 1 (constante)	3,547	0,727	1,723	0,259	0,387	***0,422***	***-0,128***
charge = 1 ou 2	3,547	0,795	1,709	0,013	0,347	***0,465***	***-0,334***
charge = 1, 2 ou 3	3,547	0,804	1,744	0,026	0,342	***0,461***	***-0,315***
charge = 1, 2, 3 ou 4	3,547	0,862	1,680	0,122	0,401	***0,513***	***-0,278***
charge = 1, 2, 3, 4, 5 ou 6	3,547	0,822	1,725	0,011	0,282	***0,477***	***-0,270***

Tableau 7 : Impact de la charge de réponse, variable d'une enquête à l'autre
Discrétisation de l'intervalle [0 ;1] : L = 100
Population stratifiée, de taille N = 1000. Taille moyenne des strates : 150
Démographie de l'ordre de 10% (de 8% à 12%).
Taux de sondage de l'ordre de 20% (de 18% à 22%)
Coordination avec tout le passé (20 enquêtes au total)

	Moyenne	σ_{Coord}	σ_{Indep}	S_{Coord}	S_{Indep}	R_σ	Δ_S
charge = 1 (constante)	3,366	0,810	1,632	0,077	0,334	***0,497***	***-0,257***
charge = 1 ou 2	3,366	0,800	1,722	0,126	0,534	***0,465***	***-0,408***
charge = 1, 2 ou 3	3,366	0,865	1,703	0,157	0,439	***0,508***	***-0,282***
charge = 1, 2, 3 ou 4	3,366	0,803	1,683	-0,361	0,448	***0,477***	***-0,808***
charge = 1, 2, 3, 4, 5 ou 6	3,366	0,841	1,673	-0,085	0,348	***0,503***	***-0,433***

Tableau 8 : Impact de la charge de réponse, variable d'une enquête à l'autre
Discrétisation de l'intervalle [0 ;1] : L = 100
Population stratifiée, de taille N = 1000. Taille moyenne des strates : 150
Démographie de l'ordre de 10% (de 8% à 12%).
Taux de sondage de l'ordre de 20% (de 18% à 22%)
Coordination avec les cinq enquêtes précédentes (20 enquêtes au total)

	Moyenne	σ_{Coord}	σ_{Indep}	S_{Coord}	S_{Indep}	R_σ	Δ_S
charge = constante	3,598	0,829	1,706	0,283	0,215	***0,486***	***0,068***
charge = décrois, linéaire	3,598	0,797	1,746	-0,579	0,340	***0,456***	***-0,919***
charge =décrois., Exponent., (base 1,2)	3,598	0,809	1,711	-0,223	0,418	***0,473***	***-0,641***
charge =décrois., Exponent., (base 1,5)	3,598	0,798	1,667	-0,453	0,261	***0,479***	***-0,714***
charge =décrois., Exponent., (base 2)	3,598	0,803	1,726	-0,333	0,479	***0,465***	***-0,812***

Tableau 9 : Impact de la charge de réponse, variable d'une enquête à l'autre
Discrétisation de l'intervalle [0 ;1] : L = 100
Population stratifiée, de taille N = 1000. Taille moyenne des strates : 150
Démographie de l'ordre de 10% (de 8% à 12%).
Taux de sondage de l'ordre de 20% (de 18% à 22%)
Coordination avec les cinq enquêtes précédentes (20 enquêtes au total)

6.2.2.5. Conclusion générale

Globalement, l'ensemble des résultats numériques plaident en faveur de la méthode de coordination présentée dans ce papier, en fournissant des gains conséquents et aisément quantifiables quant à la répartition de la charge de réponse sur les différentes unités de la population. L'efficacité de la méthode semble résister à des situations très instables où les populations d'intérêt peuvent connaître de fortes évolutions, ou lorsque les stratifications sont indépendantes et les taux de sondage très variables d'une enquête à l'autre. Si, sans surprise, la méthode devient moins performante lorsque les taux de sondage tendent vers 0 ou 1, il convient néanmoins de s'assurer de coordonner avec suffisamment d'enquêtes passées afin d'éviter les phénomènes de bassins d'attraction ou de répulsion. Une règle pour choisir le nombre d'enquêtes du passé avec lesquelles coordonner a été suggérée.

Les simulations pourront par la suite être réalisées en plus grand nombre et sur populations de tailles plus importantes, avant d'envisager une industrialisation du procédé à l'Insee. Reste également à affiner la méthode de coordination lorsque l'on souhaite en outre tirer des panels.

Bibliographie

[1] Christian Hesse, « Généralisation des tirages aléatoires à numéros aléatoires permanents, ou la méthode JALES+ », document de travail Insee E0101, 2001.

[2] Pascal Ardilly, « Présentation de la méthode JALES+ conçue par Christian Hesse », document de travail interne Insee, 2009.

[3] Franck Cotton et Christian Hesse, « *Tirages coordonnés d'échantillons* », document de travail Insee E9206, 1992.