Méthodes d'échantillonnage pour l'enquête annuelle d'entreprises

Frank COTTON et Christian HESSE
Insee

La coordination d'échantillons est un objectif souvent recherché par le statisticien. Selon les cas, il s'agit d'éviter qu'une unité soit sélectionnée dans deux échantillons différents, par souci d'équité par exemple, ou au contraire de favoriser l'apparition de ce genre d'événement (rafraîchissement d'un panel, extension d'un échantillon,...).

Ce papier décrit rapidement, dans une première partie, une formalisation simple de la notion de coordination, ainsi qu'une méthode générale, reposant sur l'utilisation de sondages conditionnels, pour contrôler le recouvrement d'échantillons quelconques, en l'absence de contraintes sur les probabilités d'inclusion du second ordre (sondages de Poisson, par exemple).

En pratique, les gestionnaires d'enqûetes souhaitent souvent que soient appliquées des conditions du second ordre, soit pour limiter les redressements a posteriori, soit parce qu'ils veulent prévoir la taille de l'échantillon. Ces conditions concernent alors le sondage marginal, et non le sondage conditionnel. Dans la seconde partie, on donne une idée des problèmes qui surgissent lorsque des conditions du second ordre sont imposées : il faut alors se contenter d'une coordination approchée des échantillons. C'est en particulier ce qui est fait pour l'échantillonnage des EAE, pour lequel la coordination est obtenue en pratique par une technique d'attribution aux unités de numéros aléatoires qui est détaillée dans la troisième partie. La quatrième partie explore diverses possibilités additionnelles de cette technique.

I Approche générale

Soit E un ensemble d'unités. Pour formaliser la notion de coordination de deux échantillons s_1 et s_2 tirés dans E, considérons le plan de sondage joint de s_1 et s_2. Ce plan est caractérisé par une loi de probabilité P^{12} sur $\mathcal{P}(E)\times\mathcal{P}(E)$, où $\mathcal{P}(E)$ est l'ensemble des parties de E. On a donc :

$$\sum_{s_1, s_2} P^{12}(s_1, s_2) = 1$$

Les plans de sondage de s_1 et s_2 sont définis par les deux lois marginales P^1 et P^2 induites par P^{12} sur $\mathcal{P}(E)$.
Enfin, on note $P^{2/1}$ la loi conditionnelle de s_2 sachant s_1. Donc, $P^{2/1}(s_2 \mid s_1)$ est la probabilité d'obtenir l'échantillon s_2 au second tirage, sachant que s_1 a été tiré au premier.
De la même façon, pour une unité i de E, on notera respectivement π_i^{12}, π_i^1, π_i^2 et $\pi_i^{2/1}(s_1)$ les probabilités jointe, marginales et conditionnelle d'inclusion du premier ordre. Par exemple :

$$\pi_i^{2/1}(s_1) = \sum_{s_2 \ni i} P^{2/1}(s_2 \mid s_1)$$

En général, l'utilisateur spécifie les probabilités d'inclusion π_i^1 et π_i^2 pour chaque échantillon. Eventuellement, il peut aussi indiquer, pour chaque tirage, des spécifications de stratification et de taille d'échantillon. Ceci se traduit par des contraintes sur les probabilités du second ordre $\pi_{i,j}^1$ ou $\pi_{i,j}^2$ que nous laissons de côté pour l'instant.

Les deux échantillons sont indépendants si $P^{2/1}$ ne dépend pas de s_1, c'est à dire si :

$$(\forall i \in E)\ (\forall s_1 \in \mathcal{P}(E)) \qquad P^{2/1}(s_2 \mid s_1) = P^2(s_2)$$

Ou encore :

$$(\forall i \in E)\ (\forall s_1 \in \mathcal{P}(E)) \qquad \pi_i^{2/1}(s_1) = \pi_i^2$$

Ce qui implique :

$$(\forall i \in E) \qquad \pi_i^{12} = \pi_i^1 \pi_i^2$$

Au contraire, on dira que les échantillons sont coordonnés si la probabilité conditionnelle dépend effectivement de s_1. Pour l'unité i, on dira que la coordination est négative si $\pi_i^{12} < \pi_i^1 \pi_i^2$ et positive si $\pi_i^{12} > \pi_i^1 \pi_i^2$.

On peut écrire :

$$\pi_i^2 = \sum_{s_1} \sum_{s_2 \ni i} p^1(s_1) p^{2/1}(s_2 | s_1) = \pi_i^1 g_i + (1 - \pi_i^1) h_i$$

où $g_i = \dfrac{\sum_{s_1 \ni i} p^1(s_1) \pi_i^{2/1}(s_1)}{\sum_{s_1 \ni i} p^1(s_1)} = \dfrac{\pi_i^{12}}{\pi_i^1}$ (respectivement h_i, sommations sur $s_1 \not\ni i$) est l'espérance de $\pi_i^{2/1}(s_1)$ selon la loi de s_1 conditionnelle à $i \in s_1$ (respectivement $i \notin s_1$), c'est à dire la probabilité conditionnelle d'inclusion moyenne de i dans s_2 sachant que i appartient à s_1 (resp. n'appartient pas à s_1). La coordination entre s_1 et s_2 pour l'unité i est négative si $h_i > g_i$ et positive si $h_i < g_i$. L'équivalence entre cette définition et la précédente est immédiate.

On tire de l'égalité précédente :

$$\pi_i^{12} = g_i \pi_i^1 = \pi_i^2 - h_i (1 - \pi_i^1)$$

En conséquence, si l'on pose :

$$\begin{cases} \pi_i^- = \max(0, \pi_i^1 + \pi_i^2 - 1) \\ \pi_i^+ = \min(\pi_i^1, \pi_i^2) \end{cases}$$

il vient :

$$\pi_i^- \leqslant \pi_i^{12} \leqslant \pi_i^+$$

Ces limites de π_i^{12} peuvent être atteintes dans le cadre d'un plan conditionnel très simple. Considérons le cas où $P^{2/1}$ ne dépend de s_1 que par le biais de son indicatrice ϵ^{s_1}. Pour une unité i donnée, $\pi_i^{2/1}(s_1)$ ne dépend alors que du fait que i a été sélectionné dans s_1 ou non. On peut écrire :

$$\pi_i^{2/1}(s_1) = g_i \epsilon_i^{s_1} + h_i (1 - \epsilon_i^{s_1})$$

Supposons que π_i^1 et π_i^2 soient fixés. L'objectif de coordination entre s_1 et s_2 conduit à maximiser ou minimiser :

$$\pi_i^{12} = g_i \pi_i^1 = \pi_i^2 - h_i (1 - \pi_i^1)$$

• $\pi_i^{12} = \pi_i^-$ (coordination négative maximale) s'obtient pour :

$$\begin{cases} g_i = 0,\ h_i = \dfrac{\pi_i^2}{1 - \pi_i^1} & \text{si } \pi_i^1 + \pi_i^2 \leqslant 1 \\ g_i = \dfrac{\pi_i^1 + \pi_i^2 - 1}{\pi_i^1},\ h_i = 1 & \text{si } \pi_i^1 + \pi_i^2 > 1 \end{cases}$$

ce que l'on peut également écrire :

$$g_i = \frac{\pi_i^-}{\pi_i^1} = g^-(\pi_i^1, \pi_i^2),\ h_i = \frac{\pi_i^2 - \pi_i^-}{1 - \pi_i^1} = h^-(\pi_i^1, \pi_i^2)$$

• $\pi_i^{12} = \pi_i^+$ (coordination positive maximale) s'obtient pour :

$$\begin{cases} g_i = 1,\ h_i = \dfrac{\pi_i^2 - \pi_i^1}{1 - \pi_i^1} & \text{si } \pi_i^1 \leqslant \pi_i^2 \\ g_i = \dfrac{\pi_i^2}{\pi_i^1},\ h_i = 0 & \text{si } \pi_i^1 > \pi_i^2 \end{cases}$$

ou encore :

$$g_i = \frac{\pi_i^+}{\pi_i^1} = g^+(\pi_i^1, \pi_i^2),\ h_i = \frac{\pi_i^2 - \pi_i^+}{1 - \pi_i^1} = h^+(\pi_i^1, \pi_i^2)$$

Pour donner un exemple, imaginons que l'on désire une coordination négative maximale entre s_1 et s_2. s_1 est déjà tiré, et on connaît π_i^1 et π_i^2 pour tout i. La coordination recherchée s'obtient comme suit pour l'unité i :

• si i a été tirée dans s_1, on la sélectionne pour s_2 avec une probabilité $\dfrac{\pi_i^-}{\pi_i^1}$,

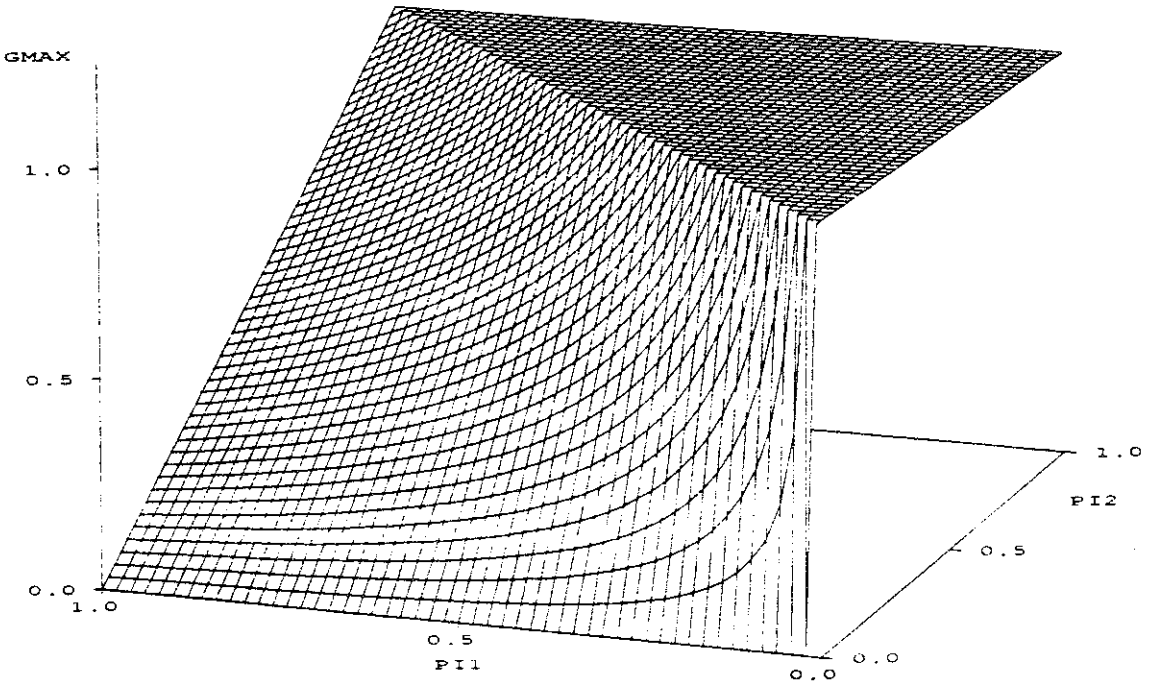
• si i n'a pas été tirée dans s_1, on la sélectionne pour s_2 avec une probabilité $\dfrac{\pi_i^2 - \pi_i^-}{1 - \pi_i^1}$

En déconditionnant, il est clair que l'on retrouve une probabilité marginale de sélection de i égale à π_i^2.
Le raisonnement est identique, en remplaçant π_i^- par π_i^+, lorsque c'est une coordination positive maximale qui est recherchée (voir figure 1). Plus généralement, en remplaçant π_i^- par une quantité π_i comprise entre π_i^- et π_i^+, on obtient des effets intermédiaires de coordination. Pour $\pi_i = \pi_i^1 \pi_i^2 = \pi_i^0$, la coordination est nulle (voir figure 2).

Naturellement, il existe de nombreuses relations de symétrie entre g^+, g^-, h^+ et h^-. Par exemple, il est équivalent de coordonner positivement s_1 avec s_2 et de coordonner négativement le complémentaire de s_1 avec s_2. On montre facilement :

$$(\forall (x,y)\in[0,1]^2)\ \begin{cases} g^{\pm}(1-x,1-y) = 1 - h^{\pm}(x,y) \\ g^{\pm}(1-x,y) = h^{\mp}(x,y) \\ g^{\pm}(x,1-y) = 1 - g^{\mp}(x,y) \\ h^{\pm}(x,1-y) = 1 - h^{\mp}(x,y) \end{cases}$$

VARIATION DE G SELON PI_1 ET PI_2 POUR UN RECOUVREMENT MAXIMUM



VARIATION DE H SELON PI_1 ET PI_2 POUR UN RECOUVREMENT MAXIMUM

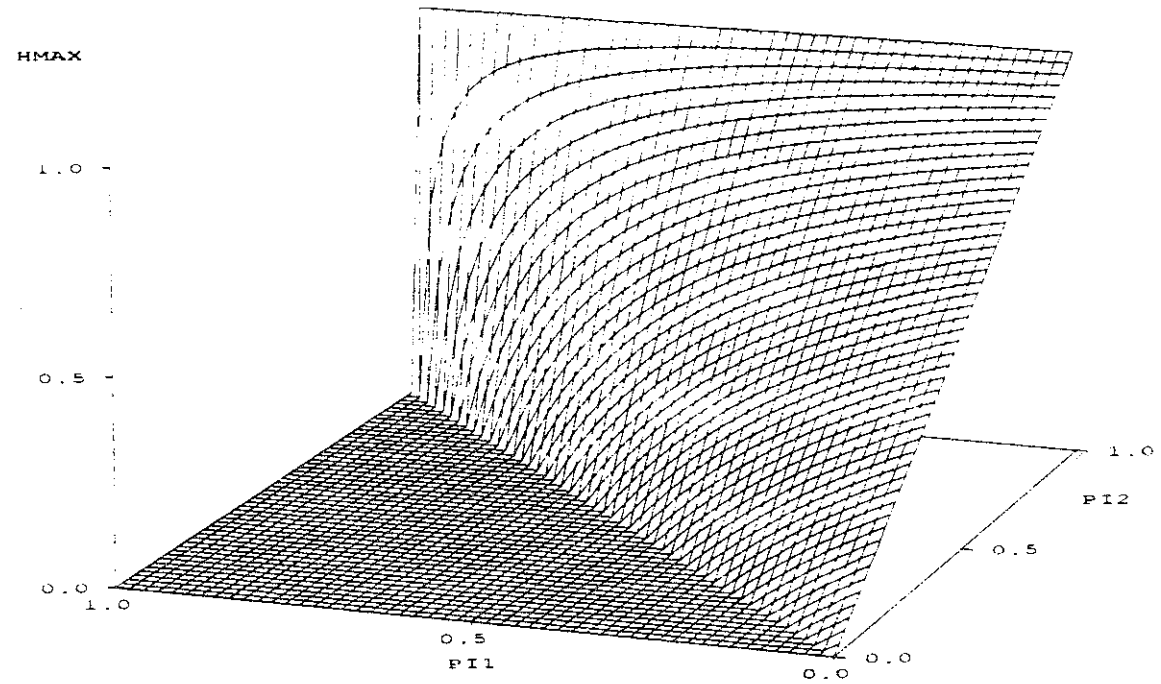
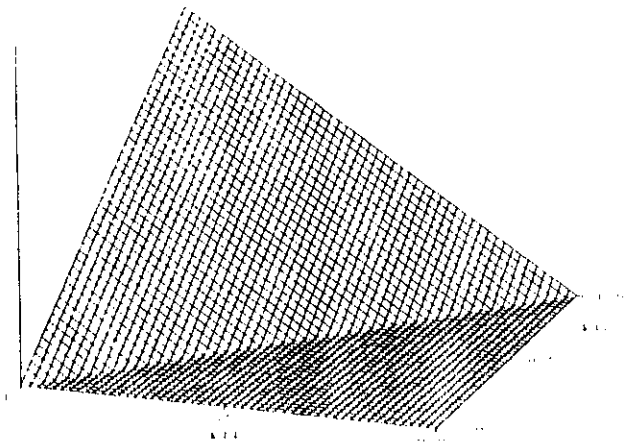


FIGURE 1

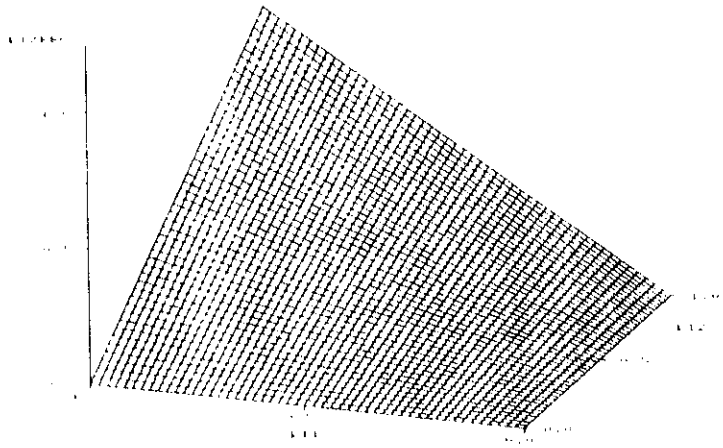
F. Cotton et C. Hesse, Insee Méthodes n° 29-30-31

CHOIX DE PI SELON LE TYPE DE COORDINATION DESIREE

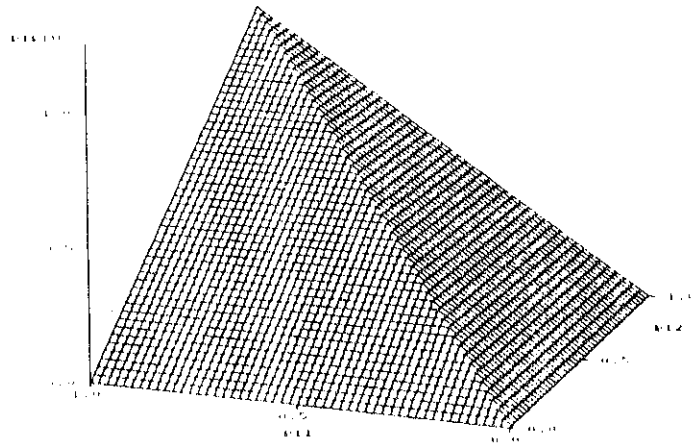
COORDINATION NEGATIVE MAXIMALE



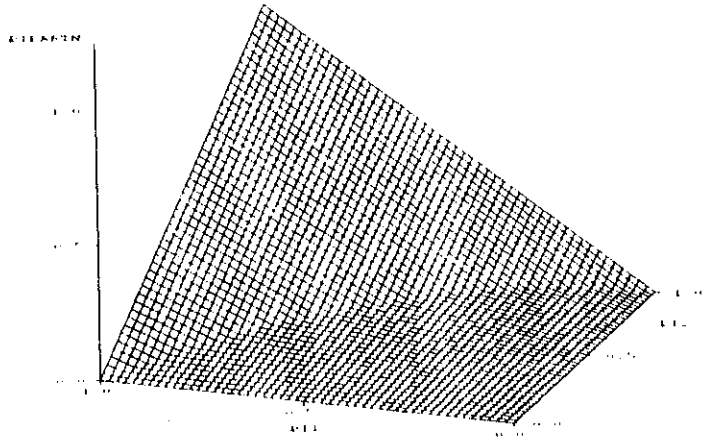
COORDINATION NULLE



COORDINATION POSITIVE MAXIMALE



COORDINATION NEGATIVE PARTIELLE



COORDINATION POSITIVE PARTIELLE

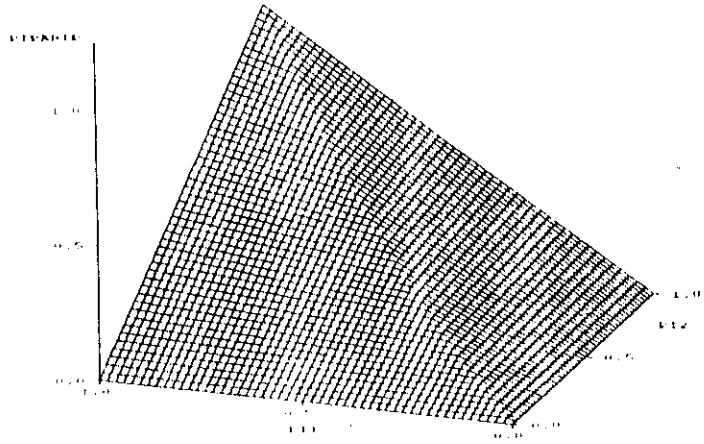


FIGURE 2

II Cas d'échantillons stratifiés

La méthode précédente se généralise sans difficulté pour permettre la coordination de plus de deux échantillons. Elle permet de coordonner parfaitement des sondages de Poisson : les sondages conditionnels sont alors eux-mêmes des sondages de Poisson. Cependant, puisqu'elle raisonne sur le plan de sondage conditionnel, cette méthode peut rendre difficile le respect de conditions sur les probabilités marginales d'inclusion du second ordre, par exemple des contraintes sur la taille de l'échantillon. Illustrons ce point sur quelques exemples simples :

• E est de taille N. s_1 et s_2 sont deux sondages aléatoires simples de taux f^1 et f^2. On désire donc tirer respectivement $f^1 N$ et $f^2 N$ (qui sont supposés entiers) unités, et on recherche une coordination négative maximale.

Si $f^1 + f^2 \leqslant 1$, on a $g^- = 0$ et $h^- = \dfrac{f^2}{1 - f^1}$. Le tirage conditionnel revient donc à un sondage aléatoire simple sur le complémentaire de s_1 au taux h^- : $f^2 N$ unités sont sélectionnées.

Si $f^1 + f^2 > 1$, on $g^- = \dfrac{f^1 + f^2 - 1}{f^1}$ et $h^- = 1$. Le tirage conditionnel revient alors à un sondage aléatoire simple sur s_1 au taux g^- : $(f^1 + f^2 - 1)N$ unités sont sélectionnées.
Le raisonnement est similaire pour une coordination positive maximale.
On voit donc que dans ce cas, les contraintes de taille d'échantillon sur s_2 se traduisent simplement sur le sondage conditionnel : la méthode permet de coordonner deux sondages aléatoires simples.

• E possède 4 unités. s_1 est un sondage aléatoire simple de taille 2, qui a déjà été tiré. s_2 est un sondage stratifié dont les deux strates possèdent chacune deux unités et sont sondées au même taux de 1/2 : une unité est sélectionnée dans chaque strate.
Les deux échantillons étant de taille 2, une coordination négative maximale conduit à un recouvrement nul. C'est d'ailleurs ce que donne la méthode précédente, puisque l'on calcule dans ce cas que $g^- = 0$ et $h^- = 1$. Néanmoins, ce résultat ne peut être atteint, dans le cadre des contraintes indiquées, si les deux unités tirées par s_1 se trouvent dans la même strate de s_2 : il faut se contenter d'une coordination partielle, ou renoncer à la stratification a priori de s_2 (ou retirer s_1).
Il apparaît donc ici que l'on ne peut obtenir de coordination parfaite pour deux échantillons stratifiés selon des critères différents, ou si des mises à jour des critères de stratification

ont affecté les unités entre s_1 et s_2.

• le premier exemple ci-dessus pourrait conduire à penser que la coordination parfaite peut être obtenue dans le cas de deux échantillons stratifiés selon les mêmes critéres : ce n'est pas le cas. Considérons le cas où E possède 6 unités, réparties en deux strates ς^a et ς^b de trois unités chacune. s_1 et s_2 sont deux sondages identiques, stratifiés selon ς^a et ς^b, de taux $f^a = \frac{2}{3}$ et $f^b = \frac{1}{3}$. Au total, quatre unités seront sondées dans ς^a et deux dans ς^b : une unité au moins de ς^a sera commune à s_1 et s_2, alors qu'une unité au moins de ς^b ne sera pas tirée. On ne peut atteindre la coordination négative maximale, théoriquement possible entre deux sondages de taille 3, sans "délester" la strate saturée au profit de celle qui ne l'est pas, c'est à dire en levant les contraintes de taille des sondages par strate.

On voit sur ces exemples que, sauf cas particuliers (sondages aléatoires simples, sondages proportionnels stratifiés selon les mêmes critères), il y a contradiction entre coordination parfaite et contraintes a priori sur les tailles d'échantillons, ou plus généralement sur les probabilités d'inclusion du second ordre.

Pour les enquêtes annuelles d'entreprises (EAE), dont les échantillons sont tirés par l'INSEE, le choix a été fait de se contenter d'une coordination approchée, et d'utiliser des sondages stratifiés à tailles fixes dans chaque strate. On décrit ci-après une méthode de coordination par attribution de numéros aléatoires aux unités qui permet, dans ce contexte plus simple que le cadre général décrit en I, de réaliser pratiquement la coordination sans passer par l'intermédiaire de sondages conditionnels.

III Coordination par attribution de numéros aléatoires

L'utilisation de numéros aléatoires pour le tirage d'échantillons est classique. Par exemple, pour un sondage de Poisson, la base de sondage est parcourue séquentiellement, et l'unité i est traitée comme suit :

- on tire un numéro aléatoire α_i uniforme sur]0,1[,
- l'unité i est sélectionnée si $\alpha_i \leqslant \pi_i$, où π_i est sa probabilité d'inclusion

De tels algorithmes séquentiels existent pour d'autres types de sondage, comme par exemple les sondages aléatoires simples ou stratifiés à tailles fixes.

Lorsque l'on vise à la coordination d'échantillons, l'idée la plus simple est de commencer par attribuer un α_i à chaque unité de la base avant d'effectuer le tirage, et de garder le même pour plusieurs tirages. α_i sert alors au moment du tirage lui-même, mais il peut également permettre de garder la mémoire de ce tirage.

Considérons pour commencer le tirage d'un seul échantillon stratifié. La base de sondage est constituée de H strates, $\{h\}_{h=1,..H}$ étant un système d'indexation des strates. n_h estla taille souhaitée de l'échantillon dans la strate h, taille entière que l'on suppose déterminée préalablement par un processus d'allocation. A chaque unité i de la base est associé un numéro aléatoire α_i tiré suivant une loi donnée de fonction de répartition F. Pour effectuer le tirage dans la strate h, on range les unités suivant les α_i croissants, et on sélectionne les n_h premières. Naturellement, on peut également sélectionner les n_h dernières, ou les n_h à partir d'un certain rang.

Supposons maintenant que l'on désire effectuer deux tirages s_1 et s_2 dans la base, suivant la même stratification, et que l'on souhaite en outre coordonner ces tirages d'une certaine manière. Par exemple, on cherche à minimiser le recouvrement entre s_1 et s_2 (coordination négative), ou au contraire on veut qu'un nombre maximum d'unités soient communes aux deux échantillons (coordination positive). La méthode précédente offre deux possibilités simples pour réaliser cette coordination.

Tout d'abord, on peut jouer sur la "fenêtre d'interrogation" (c'est à dire la position de la séquence d'unités retenues) entre les deux tirages. Plaçons-nous dans la strate h ordonnée suivant les α_i, où l'on a sélectionné pour s_1 les unités de rang compris entre 1 et n_{1h}. Pour maximiser le recouvrement, on choisira pour

s_2 les unités de rang compris entre 1 et n_{2h} (on ne modifie donc pas dans ce cas l'origine de la fenêtre d'interrogation). Inversement, si l'on sélectionne les unités de rang $n_{1h} + 1$ à $n_{1h} + n_{2h}$ (modulo N_h, taille de la population dans la strate), on minimise le recouvrement entre s_1 et s_2. Toutes les possibilités intermédiaires (interrogation de $d + 1$ à $d + n_{2h}$, $0 < d < n_{1h}$) sont possibles, et permettent de contrôler le recouvrement.
Une technique de ce type est utilisée à Statistics Sweden pour le tirage des principales enquêtes économiques.

Une seconde possibilité, logiquement équivalente à la précédente mais parfois plus simple à mettre en oeuvre, consiste à maintenir fixe l'origine de la fenêtre d'interrogation, mais à modifier la répartition des unités dans la strate en changeant leurs numéros aléatoires α_i. Considérons de nouveau la strate h après tirage de s_1 et supposons, pour fixer les idées, que les α_i soient distribués entre 0 et 1. On peut rejeter en fin de classement les unités sélectionnées pour s_1 (les n_{1h} premières) en ajoutant 1 à leur numéro aléatoire et en réordonnant la strate selon les α_i ainsi modifiés. Si l'on tire ensuite pour s_2 les n_{2h} premières unités, on minimise le recouvrement entre les deux tirages. Bien entendu, il est là encore possible de contrôler plus précisément ce recouvrement en ne modifiant que les numéros d'une partie des unités sélectionnées pour s_1.
Une méthode analogue a été développée à l'institut statistique des Pays-Bas. Dans cette méthode, les α_i sont initialement uniformes sur $]0,1[$. A chaque enquête s tirée dans la base est associée une charge statistique $\gamma_{s,h}$ qui peut dépendre de la strate. Lorsque l'unité i est tirée, son numéro aléatoire α_i est modifié en $\alpha_i + \gamma_{s,h}$, où h est la strate qui contient i. Ceci permet un étalement de la charge statistique entre les unités.

L'un des avantages de cette seconde méthode avec renumérotation des unités est qu'elle s'adapte assez aisément au cas où les critères de stratification de s_1 et s_2 sont différents, au prix de certaines contraintes sur la répartition des α_i et sur la technique de renumérotation.

Voyons tout d'abord quels sont les problèmes qui se posent dans ce cas de stratifications différentes si l'on applique la méthode avec renumérotation sans précaution. Imaginons pour simplifier que l'on recherche une coordination négative maximale entre s_1 et s_2. Après tirage de s_1, on a donc, comme indiqué ci-dessus, ajouté 1 aux numéros des unités sélectionnées.

La stratification de s_2 s'obtient en découpant les strates de s_1 et en recomposant différemment les sous-strates ainsi obtenues. L'opération de découpage ne soulève pas de difficulté : par

exemple, lorsqu'une strate de s_2 est une sous-strate de s_1, elle est composée d'unités homogènes au regard de s_1 et s_2, ordonnées aléatoirement à ceci près que les unités déjà tirées pour s_1 sont en fin de classement; le tirage des n_{2h} premières unités fournit donc la coordination attendue. En revanche, l'opération de recomposition pose problème : l'ordre des unités dans la strate de s_2 dépend de leurs caractéristiques au regard de s_1. Plus précisément, le problème résulte du fait que la distribution des unités dans une strate de s_1 après renumérotation est fonction du taux de sondage utilisé pour s_1 dans cette strate.
La figure 3 illustre ce fait sur un exemple simple. On considère deux strates ς^a et ς^b de s_1, qui sont fondues en une strate unique de s_2. ς^a contient dix unités, figurées par des étoiles à l'emplacement de leur numéro aléatoire, et est sondée par s_1 au taux $\frac{1}{2}$. ς^b contient également dix unités (losanges) et est sondée au taux $\frac{1}{10}$. La partie gauche de la figure montre ς^a et ς^b avant renumérotation : les cinq premières unités de ς^a et la première unité de ς^b ont été sélectionnées par s_1 et vont être renumérotées. Sur la partie droite de la figure, on montre les strates après la renumérotation, c'est à dire après que les numéros des unités tirées ait été augmenté de 1. Lorsque l'on mélange ς^a et ς^b pour obtenir la strate de s_2 (en bas à droite de la figure), on constate que les unités provenant de ς^b sont sur-représentées au début de la strate, ce qui compromet le tirage de s_2.

Ce genre d'ennuis peut être évité en choisissant une technique de renumérotation qui laisse invariante la fonction de répartition des α_i. Ainsi, si les α_i sont distribués indépendamment et uniformément sur]0,1[, on peut procéder par permutation des numéros aléatoires entre les unités.
Nous proposons une technique de renumérotation linéaire des unités, plus commode à mettre en pratique, qui se traduit dans la strate h par les opérations suivantes (figure □) :
• pour $i = n_h + 1, .. N_h$ (unités non tirées dans s_1), α_i est changé en $\alpha_i^1 = \alpha_i - \alpha_{n_h}$,
• pour $i = 1, ..n_h$ (unités tirées dans s_1), α_i est changé en $\alpha_i^1 = \alpha_i - \alpha_{n_h} + \alpha_{N_h}$.
Donc, en notant α le vecteur des α_i et α^1 celui des α_i^1 :

$$\alpha^1 = A\alpha$$

où A est une matrice bloc-diagonale composée de sous-matrices du type :

$$A_{(N_h, N_h)} = \begin{pmatrix} 1 & 0 & \cdots & 0 & -1 & 0 & \cdots & 0 & 1 \\ 0 & 1 & \cdots & 0 & -1 & 0 & \cdots & 0 & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & -1 & 0 & \cdots & 0 & 1 \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 1 \\ 0 & 0 & \cdots & 0 & -1 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & -1 & 0 & \cdots & 1 & 0 \\ 0 & 0 & \cdots & 0 & -1 & 0 & \cdots & 0 & 1 \end{pmatrix}$$

A est donc clairement de jacobien égal à 1 en valeur absolue, et laisse donc inchangée la distribution jointe des numéros aléatoires.

On peut donc ainsi coordonner négativement des tirages stratifiés suivant des critères différents. Il est clair là aussi qu'il est possible d'obtenir des effets de coordination divers en aménageant légèrement la technique de renumérotation. Dans le cas de la renumérotation linéaire, si l'on désire un taux de recouvrement $\tau_h = 1 - \rho_h$ dans la strate h, il suffit de remplacer dans les formules précédentes le "pivot" α_{n_h} de la renumérotation par α_{r_h}, où r_h est un entier qui approxime $\rho_h n_h$.

MELANGE DE STRATES APRES RENUMEROTATION QUELCONQUE

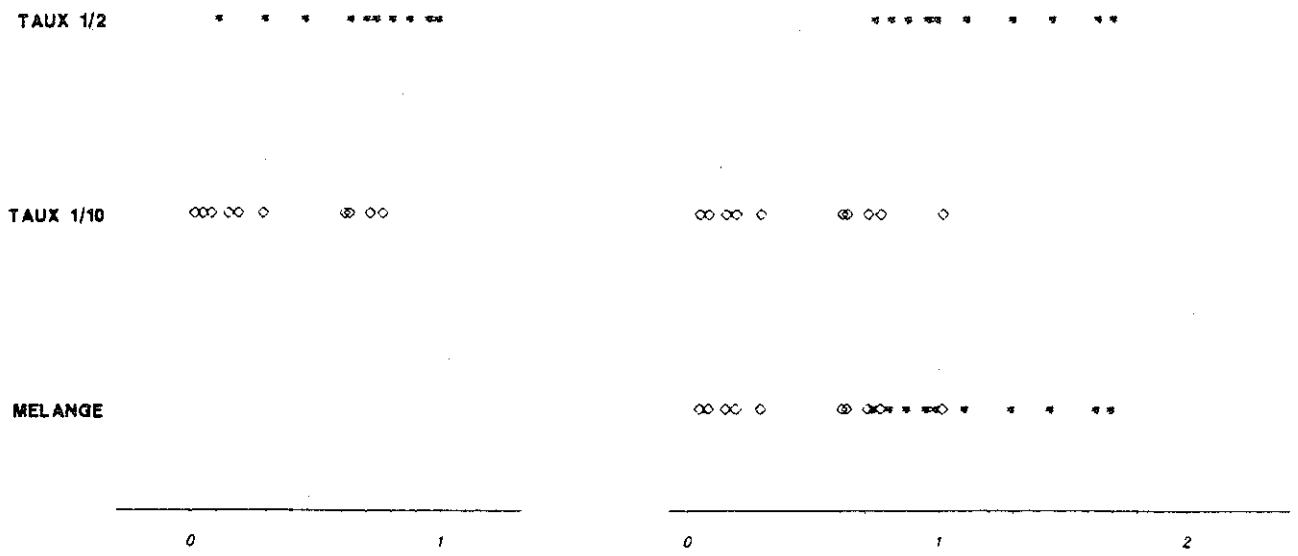


FIGURE 3

RENUMEROTATION LINEAIRE DES UNITES APRES TIRAGE

Les n unités de la strate sont rangées selon leur numéro aléatoire : $\alpha_1 < \ldots < \alpha_n$
Si les i premières unités sont tirées, les numéros sont recalculés comme suit :
- pour j <= i (unités tirées), α_j est changé en $\alpha_j - \alpha_n - \alpha_i$
- pour j > i, α_j est changé en $\alpha_j - \alpha_i$

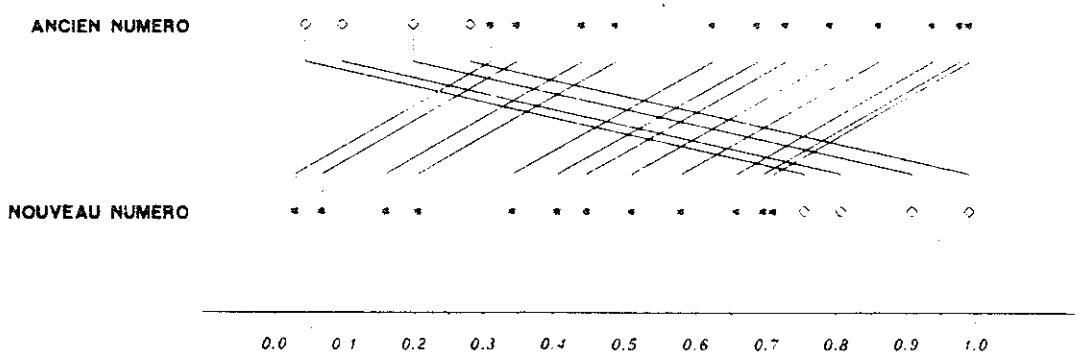


FIGURE 4

IV Une autre application de la méthode avec renumérotation

On vient de voir qu'il était possible de coordonner simplement des tirages d'échantillons, même en cas de changement de stratification (et donc, par exemple, de mise à jour de la base) entre deux tirages.
Plus généralement, on peut réaliser de multiples effets de coordination entre échantillons par simple manipulation des numéros aléatoires associés aux unités, à condition de ne pas modifier la distribution théorique de ces numéros.

Illustrons ce point dans une autre situation : considérons une base de sondage contenant deux types liés d'unités, par exemple des entreprises et leurs établissements. Il peut parfois être intéressant de coordonner des tirages portant sur ces deux types d'unités, notamment dans une optique de répartition de la "charge statistique" sur les enquêtés.

Supposons qu'un tirage s_1 a été effectué sur les entreprises en utilisant les méthodes précédentes. On souhaite maintenant procéder à un tirage s_2 sur les établissements, mais en évitant si possible de sélectionner des établissements appartenant à des entreprises tirées par s_1. De telles entreprises ont, après renumérotation linéaire, des α_i plutôt élevés. Il faut que leurs établissements aient également des numéros aléatoires élevés pour ne pas être sélectionnés au début de leur strate par s_2. L'idée est donc de relier le numéro de l'entreprise et le plus petit des numéros de ses établissements.

Si au contraire s_1 portait sur les établissements, on peut chercher, dans un tirage s_2 d'entreprises, à éviter celles dont des établissements ont été sélectionnés par s_1. On souhaite en conséquence, lorsque l'un des numéros d'établissements est élevé (parce qu'il a été tiré, et que la renumérotation l'a rejeté vers la fin de sa strate), que le numéro de l'entreprise le soit aussi, ce qui minimisera ses chances d'être sélectionnée. On peut donc penser ici à un lien entre le numéro aléatoire de l'entreprise et le plus grand des numéros de ses établissements.

On peut établir entre le numéro α^{ent} d'une entreprise et les numéros $\alpha^{éta}{}_j$ de ses n établissements des liens qui préservent l'uniformité des distributions. On utilise pour cela la propriété que si X est une variable aléatoire réelle de fonction de répartition F, F(X) est uniforme entre 0 et 1. Si les numéros des établissements sont donnés, on peut donc par exemple poser :

$$\alpha^{ent} = \left(\max(\alpha^{éta_1}, \alpha^{éta_2}, \cdots, \alpha^{éta_n})\right)^n$$

ou :

$$\alpha^{ent} = 1 - \left(1 - \min(\alpha^{éta_1}, \alpha^{éta_2}, \cdots, \alpha^{éta_n})\right)^n$$

Dans les deux cas, il est aisé de vérifier que α^{ent} est alors uniforme sur]0,1[si les $\alpha^{éta_j}$ le sont.

On peut également maintenir ces liens à travers des renumérotations linéaires portant sur les entreprises ou sur les établissements.
Si un tirage d'établissements modifie les $\alpha^{éta_j}$, α^{ent} peut être recalculé suivant l'une des formules ci-dessus. Comme on l'a vu au début de ce paragraphe, le choix de la formule (lien par le min ou par le max) dépend des effets de coordination que l'on désire obtenir entre échantillons d'entreprises et échantillons d'établissements.
Si α^{ent} est modifié par un tirage d'entreprises, on peut inverser les formules précédentes pour recalculer le min ou le max des numéros d'établissements. Les autres $\alpha^{éta_j}$ sont tirés au hasard sur]0,max[ou]min,1[selon le cas, ou répartis au prorata de leur distribution avant renumérotation.

Les figures 5 et 6 présentent des simulations simples qui permettent de visualiser les effets de coordination obtenus par les techniques présentées ci-dessus.
Dans ces simulations, on considère une population de 5 000 entreprises de deux établissements. Les établissements se voient affecter un numéro tiré au hasard entre 0 et 1. Le numéro aléatoire de l'entreprise est calculé en fonction du min ou du max des numéros d'établissements.
Dans la figure 5, on procède sur les établissements à un sondage aléatoire simple puis à une renumérotation linéaire. Le numéro de l'entreprise est ensuite recalculé par le min ou le max des nouveaux numéros d'établissements. Le titre "lien min-max", à titre d'exemple, indique que le numéro d'entreprise a été calculé par le min avant le tirage d'établissements, puis recalculé par le max après ce tirage. Les quatre graphiques correspondent aux quatre types de liens possibles. On trace sur chaque graphique le nouveau numéro de l'entreprise en fonction de l'ancien. Chaque point correspond à une entreprise. On a distingué les zones correspondant aux cas où aucun établissement (Ⓐ), un établissement (Ⓤ) ou les deux établissements (Ⓓ)de l'entreprise ont été sélectionnés.
On constate que le lien min-max fournit une coordination négative

efficace, puisque le numéro des entreprises dont au moins un établissement est tiré augmente fortement, et que leur probabilité de sélection dans le prochain échantillon d'entreprise diminue.
max-max donne également une bonne coordination, bien qu'il puisse arriver que des entreprises dont un établissement a été tiré voient leur numéro baisser, mais cela n'arrive qu'en fin de strate, c'est à dire en général hors de la fenêtre d'interrogation du prochain échantillon.
max-max et min-max assurent pratiquement qu'aucune entreprise dont les deux établissements ont été tirés ne sera sélectionnée de sitôt. En revanche, certaines entreprises dont un seul établissement a été tiré voient leur numéro diminuer, même en début de strate.

Dans la figure 6, c'est un tirage d'entreprises qui est effectué. Les entreprises sont donc renumérotées, et les numéros de leurs deux établissements recalculés. Les graphiques tracent, pour chaque type de lien, l'évolution du numéro d'un des deux établissements de l'entreprise (choisi au hasard). On distingue deux zones de points séparées sur chaque graphique, qui correspondent d'une part aux établissements dont l'entreprise a été tirée (𝕆), d'autre part aux établissements dont l'entreprise n'a pas été tirée (ℕ).
Dans ce cas, min-min et max-min procurent la meilleure coordination négative : les établissements dont l'entreprise a été tirée voient leur numéro augmenter, c'est à dire leur probabilité de sélection dans le prochain tirage d'établissement diminuer.

EFFET SUR LE NUMERO DE L'ENTREPRISE D'UN TIRAGE D'ETABLISSEMENTS (TAUX 1/5)

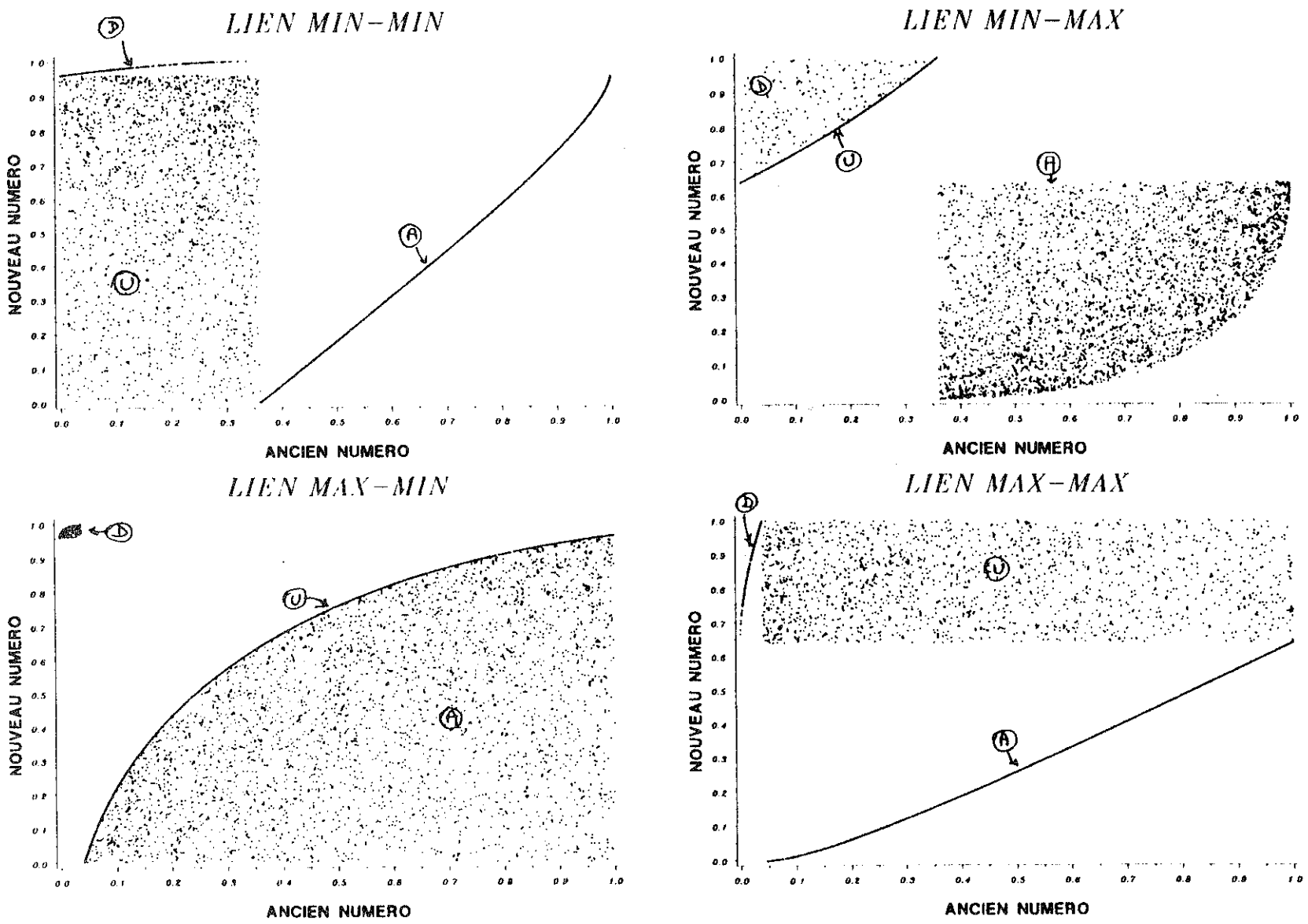
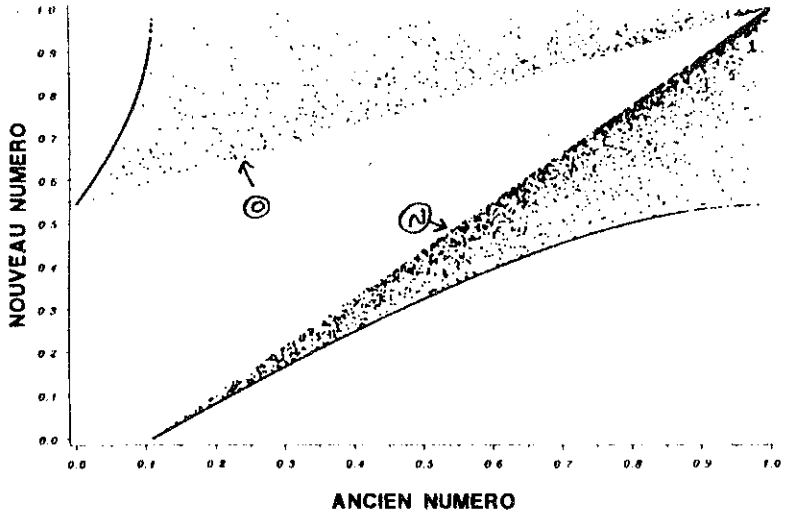


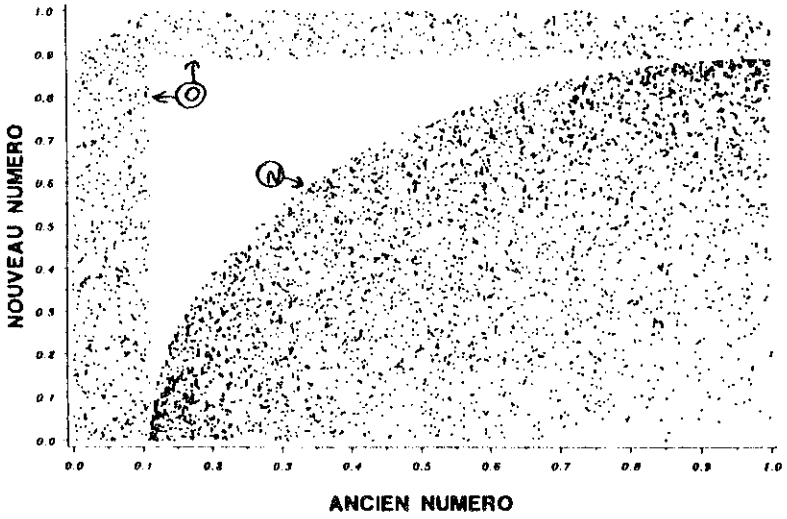
FIGURE 5 : SIMULATIONS SUR 5000 ENTREPRISES DE DEUX ETABLISSEMENTS

EFFET SUR LE NUMERO DE L'ETABLISSEMENT D'UN TIRAGE D'ENTREPRISES (TAUX 1/5)

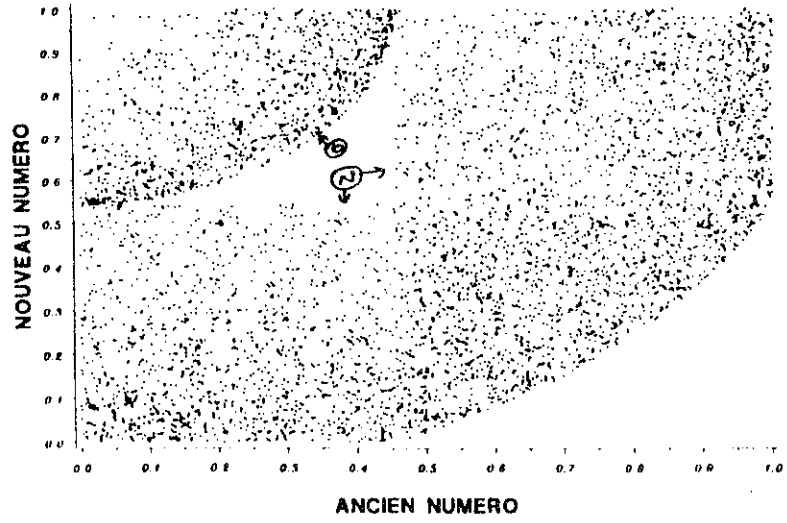
LIEN MIN–MIN



LIEN MIN–MAX



LIEN MAX–MIN



LIEN MAX–MAX

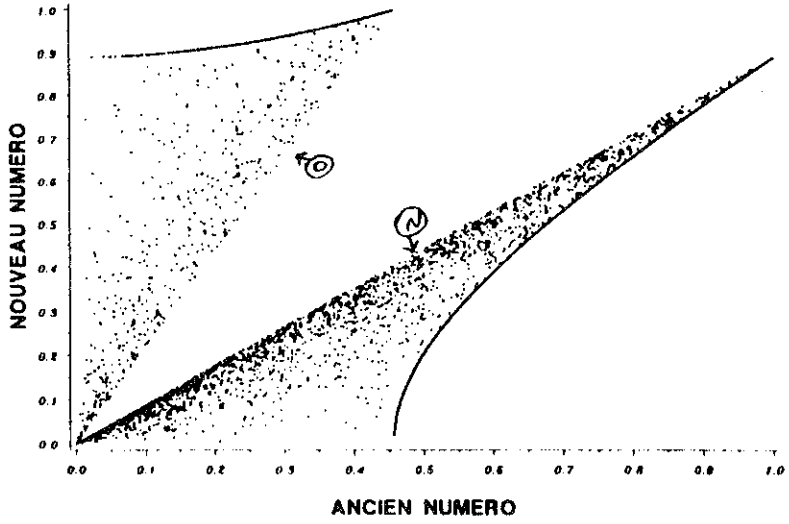


FIGURE 6 : SIMULATIONS SUR 5000 ENTREPRISES DE DEUX ETABLISSEMENTS