Tirages coordonnées et permutations

Paul-André SALAMIN

Office fédéral de la statistique (NEUCHATEL), Service de méthodes statistiques

Introduction

Le but de la coordination d'échantillons est de gérer un recouvrement, maximal ou minimal, de plusieurs échantillons tirés successivement à l'intérieur d'une population U. Pour effectuer une coordination, la sélection d'un nouvel échantillon dépendra donc des échantillons précédemment tirés. On vise à obtenir un recouvrement des échantillons plus fort ou plus faible que celui fournit par des tirages indépendants. Pour des tirages indépendants, le chevauchement des échantillons est aléatoire et la charge, c'est-à-dire le nombre de fois qu'une unité doit répondre à une enquête, n'est pas répartie de manière équitable. L'objet est donc de contrôler le recouvrement de plusieurs échantillons tirés successivement afin de limiter ou au contraire d'augmenter le nombre d'unités communes à tous les échantillons sélectionnés. On crée une dépendance entre les échantillons en tenant compte des sondages déjà effectués.

On distingue deux types de coordination :

—La *coordination positive* qui cherche à retenir le plus longtemps possible les mêmes unités dans des échantillons successifs, i.e. à maximiser le nombre d'unités communes aux deux échantillons en forçant leur chevauchement.

—La *coordination négative* qui vise à éviter au maximum le chevauchement des échantillons, i.e. à minimiser le nombre d'unités communes aux échantillons, afin de rendre plus uniforme la charge de réponse des unités.

Cet article donne une description des algorithmes de tirages coordonnés qui les rend accessibles au calcul et qui permet d'en étudier les propriétés.

Dans la section 1, on introduit le formalisme dans le cadre bien connu d'un tirage aléatoire simple (TAS). On donne la traduction d'un algorithme de tirage classique et, comme tout premier exemple d'application, on calcule les probabilités d'inclusion. Dans la section 2, on examine trois algorithmes pour la coordination négative d'une suite de tirages aléatoires simples. On en donne une description unifiée, qui permet de montrer que deux des algorithmes sont en fait équivalents. Dans la section 3, on aborde le tirage aléatoire simple stratifié (TASST). Comme pour le TAS, on donne la traduction dans notre formalisme d'un algorithme de tirage classique. On s'intéresse ensuite à un algorithme pour la coordination négative d'une suite de TASST et on donne une démonstration de sa validité. Dans la section 4, on considère la coordination positive d'un TAS et d'un TASST et on calcule de l'espérance de la taille du recouvrement des deux échantillons.

1. Sondage aléatoire simple

Un algorithme fréquemment utilisé pour tirer un échantillon aléatoire simple de taille n dans une population de taille N consiste à générer, pour chaque unité $k \in U$ de la population, des nombres aléatoires ω_k indépendants et de loi uniforme sur l'intervalle $(0,1)$, à ordonner les unités de la population par nombres aléatoires ω_k croissants et à sélectionner dans l'échantillon $S \subset U$ les n premières unités.

On remarque que sélectionner des unités qui ont reçu de petits nombres aléatoires revient à sélectionner des unités qui ont des rangs bas dans la suite $\omega = (\omega_k, k \in U)$. Comme les nombres aléatoires *i.i.d.* $\omega_k, k \in U$, sont générés à partir d'une loi ayant une fonction de distribution continue, il suit que les rangs des unités $k \in U$ dans la suite ω sont de distribution uniforme sur S_U,le groupe des permutations de U. On arrive ainsi à une version équivalente de l'algorithme qui consiste à choisir une permutation aléatoire $\sigma \in S_U$ telle que $P(\sigma) = 1/\mathrm{N!}$ et à sélectionner l'échantillon $s = \sigma^{-1}\{1,\ldots,n\}$.

Finalement, nous allons voir un échantillon $S \subset U$ comme un vecteur, ou une application, $I \in \{0,1\}^U$, où $I_k = 1$ si $k \in S$ et $I_k = 0$ autrement. Par exemple, si $U = \{1,2,3,4,5\}$, alors $S = \{1,3,5\}$ et $I = (10101)$ représentent le même échantillon. Dans ce cadre, l'algorithme de sélection prend la forme suivante : on choisit une permutation aléatoire $\sigma \in S_U$ telle que $P(\sigma) = 1/\mathrm{N!}$ et on obtient l'échantillon par $I = a \circ \sigma$ où a est le vecteur d'allocation

$$a = (1_n, 0_{N-n}) = (\underbrace{1\ldots1}_{n}\underbrace{0\ldots0}_{N-n}) \in \{0,1\}^U.$$

Par exemple, avec $\sigma = 24351$ et $n = 3$ on trouve

$$I = a \circ \sigma = \begin{pmatrix} 12345 \\ 11100 \end{pmatrix}\begin{pmatrix} 12345 \\ 24351 \end{pmatrix} = \begin{pmatrix} 12345 \\ 10101 \end{pmatrix},$$

ce qui correspond à l'échantillon $S = \{1,3,5\}$ obtenu par $S = \sigma^{-1}\{1,2,3\}$.

La représentation $I = a \circ \sigma$ d'un échantillon aléatoire simple ramène le calcul des probabilités d'inclusion à des sommations sur l'ensemble des permutations de U. Par exemple, on a pour les probabilités d'inclusion de premier ordre

$$\pi_k = \sum_S I_k(S) P(S) = \sum_{\sigma \in S_U} a(\sigma(k)) P(\sigma).$$

Pour calculer cette dernière somme, on peut utiliser la décomposition $S_U = \bigcup_{i \in U} (ki) S_k$, où (ki) est la transposition qui échange les unités k, i et qui laisse les autres unités fixes et $S_k = \{\sigma \in S_U ; \sigma(k) = k\}$. Alors

$$\pi_k = \sum_{i \in U} \sum_{\sigma \in (ki) S_k} a(\sigma(k)) P(\sigma).$$

Maintenant, si $\sigma \in (ki)S_k$ alors $\sigma(k) = 1$. Comme $P(\sigma) = 1/\mathrm{N!}$ et $|S_k| = (N-1)!$, nous avons

$$\pi_k = \frac{1}{N}\sum_{i \in U} a(i) = \bar{a}.$$

Finalement, puisque $a = (1_n, 0_{N-n})$, on trouve $\pi_k = n/N$. Le calcul des probabilités d'inclusion d'ordre deux est en tous points similaire. On a

$$\pi_{kl} = \sum_{\sigma \in S_U} a(o(k))\, a(\sigma(l))\, P(\sigma) = \frac{1}{N(N-1)} \sum_{i \neq j \in U} a_i\, a_j = \frac{n(n-1)}{N(N-1)}.$$

2. Trois algorithmes de coordination

On considère dans cette section une suite de tirages aléatoires simples (TAS) dans une population U aux temps T =(1 ,...,t,...,T). On se donne

— des tailles d'échantillons $(n_t, t \in T)$,

— des vecteurs de nombres aléatoires $(\omega_t, t \in T)$,avec $\omega_t = (\omega_{kt}, k \in U)$,

— des vecteurs de charges $(c_t, t \in T)$,avec $c_t = (c_{kt}, k \in U)$ et c_{kt} = charge de l'unité k lors du tirage t.

Le tirage de l'échantillon S_t est effectué à l'aide du vecteur de nombres aléatoires ω_t. Soit $I_t = (I_{kt}, k \in U)$ le vecteur représentant l'échantillon S_t. La charge due à l'échantillon S_t est définie comme $c_t I_t = (c_{kt} I_{kt}, k \in U)$. La charge cumulée due aux tirages jusqu'au temps t est alors donnée par $b_t = \sum_{v \leq t} c_v I_v$. Comme on l'a vu dans la section 1, on peut remplacer les vecteurs de nombres aléatoires ω_t par des permutations aléatoires σ_t. Le tirage de l'échantillon au temps t peut alors s'écrire $I_t = a_t \circ \sigma_t$, où $a_t = (1_{n_t}, 0_{N-n_t})$, et le vecteur de charge cumulée est donné par $b_t = \sum_{v \leq t} c_v (a_v \circ \sigma_v)$. Ainsi, pour une unité $k \in U$, nous avons la charge cumulée $b_t(k) = \sum_{v \leq t} c_v(k)\, (a_v(\sigma_v(k)))$.

Pour contrôler la répartition de la charge, σ_{t+1} doit dépendre des tirages précédant : $\sigma_{t+1} = \psi(\sigma_v, n_v; v \leq t)$. Pour que les échantillons marginaux correspondent à des TAS, on doit avoir $P(\sigma_t) = 1/N!$ pour tout $t \in T$.

On considère trois algorithmes (EDS, CH, RIV) pour la *coordination négative* d'une suite de tirages aléatoires simples. On donne tout d'abord la définition standard des algorithmes, qui tous trois utilisent des nombres aléatoires permanents. On traduit ensuite ces définitions en utilisant le formalisme introduit dans la section 1. Ceci amène à une description unifiée des trois algorithmes considérés qui sont ensuite comparés dans la section 2.4.

2.1 Algorithme EDS

L'algorithme EDS (De Ree 1999) utilise des nombres aléatoires permanents et des tris de la population par charge cumulée et nombres aléatoires permanents.

ALGORITHME EDS (TAS)

1. **Initialisation** : Nombres aléatoires permanents $\omega = (\omega_k, k \in U)$ où $\omega_k \overset{i.i.d}{\sim} \text{Unif}(0,1), k \in U$ (sans doublons : $\omega_k \neq \omega_l$ si $k \neq l$). Tailles d'échantillons $(n_t, t \in T)$. Charges d'enquêtes $(c_t, t \in T)$. On pose $c_t = 1_N$ pour tout $t \in T$. Charge cumulée initiale $b_0 = 0_N$.

2. Tirage $t \geq 1$

(a) **Sélection de l'échantillon**

i. On ordonne les unités de la population par charge cumulée croissante et, en cas d'égalité de la charge cumulée, par nombre aléatoire ω croissant.

ii. On obtient S_t en sélectionnant les n_t premières unités.

(b) **Actualiser la charge cumulée**

$$b_t(k) = \begin{cases} b_{t-1}(k) + 1 & si\ k \in S_t \\ b_{t-1}(k) & si\ k \in U \setminus S \end{cases}$$

3. On pose $t := t+1$ et on répète l'étape 2.

On peut reformuler l'algorithme EDS en utilisant les rangs par rapport à une partition, cf. l'annexe A, pour décrire l'opération "ordonner les unités de la population par charge cumulée et nombre aléatoire permanent". Soit ω le vecteur de nombres aléatoires permanents. Pour le tirage t, on doit ordonner les unités de la population par charge cumulée b t -1 et nombres aléatoires permanents ω. Comme $U = \{1, \ldots, N\}$ nous avons la transformation

$$id \mapsto id \circ R(b_{t-1}, \omega)^{-1} = R(b_{t-1}, \omega)^{-1},$$

qui met les unités qui ont une charge cumulée faible et des nombres aléatoires permanents petits en début de liste. La sélection de l'échantillon s'effectue ensuite par $I_t = a_t \circ R(b_{t-1}, \omega)$. On voit que la suite des permutations de sélection de l'algorithme EDS est donnée par $\sigma_1 = R(\omega)$ et $\sigma_{t+1} = R(b_t, \omega) = R(b_t, \sigma_1)$.

Comme $b_t = \sum_{v \leq t} I_v = \sum_{v \leq t} (a_v \circ \sigma_v)$ nous avons

$$\sigma_{t+1} = R(b_t, \sigma_1) = R(a_1 \circ \sigma_1 + a_2 \circ \sigma_2 + \cdots + a_t \circ \sigma_t, \sigma_1).$$

Nous sommes ainsi arrivés à la description suivante de l'algorithme EDS.

1. Choisir une permutation aléatoire initiale σ_1 de loi uniforme sur S_U.

Charge cumulée initiale $b_0 = 0_N$.

2. Tirage $t \geq 1$

(a) Sélection de l'échantillon $I_t = a_t \circ \sigma_t$ où $a_t = (1_n, 0_{N-n_t})$.

(b) Actualisation de la charge cumulée $b_t = b_{t-1} + I_t$.

(c) $\sigma_{t+1} = R(b_t, \sigma_1)$.

3. On pose $t := t+1$ et on répète l'étape 2.

On considère comme exemple le tirage coordonné de 3 échantillons aléatoires simples de tailles $n_1 = n_2 = n_3 = 3$ dans une population de taille $N = 5$. Les nombres aléatoires permanents sont $\omega = (0.5, 0.2, 0.4, 0.1, 0.3)$. La table ci-dessous montre le déroulement de l'algorithme utilisant les nombres aléatoires permanents et des tris de la population par charge cumulée et nombres aléatoires. On obtient la suite d'échantillons $S_1 = \{2, 4, 5\}$, $S_2 = \{1, 3, 4\}$ et $S_3 = \{2, 3, 5\}$.

Tirage 1

k	ω	b_0	b_1
4	0.1	0	1
2	0.2	0	1
5	0.3	0	1
3	0.4	0	0
1	0.5	0	0

Tirage 2

k	ω	b_1	b_2
3	0.4	0	1
1	0.5	0	1
4	0.1	1	2
2	0.2	1	1
5	0.3	1	1

Tirage 3

k	ω	b_2	b_3
2	0.2	1	2
5	0.3	1	2
3	0.4	1	2
1	0.5	1	1
4	0.1	2	2

De manière équivalente, on peut effectuer les tirages avec une suite de permutations. La permutation initiale est donnée par $\sigma_1 = R(\omega) = 52413$.

1. Tirage $t = 1$

$$I_1 = a_1 \circ \sigma_1 = \begin{pmatrix} 12345 \\ 11100 \end{pmatrix} \begin{pmatrix} 12345 \\ 52413 \end{pmatrix} \begin{pmatrix} 12345 \\ 01011 \end{pmatrix} \quad \leftrightarrow \quad S_1 = \{2, 4, 5\}.$$

Charge cumulée $b_1 = I_1$. La permutation pour le tirage $t = 2$ est donnée par

$$\sigma_2 = R(b_1, \sigma_1) = \begin{pmatrix} 12345 \\ 24135 \end{pmatrix}.$$

¶2. Tirage $t = 2$

$$I_2 = a_2 \circ \sigma_2 = \begin{pmatrix} 12345 \\ 11100 \end{pmatrix} \begin{pmatrix} 12345 \\ 24135 \end{pmatrix} = \begin{pmatrix} 12345 \\ 10110 \end{pmatrix} \quad \leftrightarrow \quad S_2 = \{1, 3, 4\}.$$

Charge cumulée $b_2 = I_1 + I_2 = (11121)$. La permutation pour le tirage $t = 3$ est donnée par

$$\sigma_3 = R(b_2, \sigma_1) = \begin{pmatrix} 12345 \\ 41352 \end{pmatrix}.$$

3. Tirage $t = 3$

$$I_3 = a_3 \circ \sigma_3 = \begin{pmatrix} 12345 \\ 11100 \end{pmatrix} \begin{pmatrix} 12345 \\ 41352 \end{pmatrix} = \begin{pmatrix} 12345 \\ 01101 \end{pmatrix} \quad \leftrightarrow \quad S3 = \{2, 3, 5\}$$

La table ci-dessous donne le détail du calcul de la séquence des permutations de sélection.

k	σ_1	b_1	σ_2	I_2	b_2	σ_1	σ_3	I_3
1	5	0	2	1	1	5	4	0
2	2	1	4	0	1	2	1	1
3	4	0	1	1	1	4	3	1
4	1	1	3	1	2	1	5	0
5	3	1	5	0	1	3	2	1

2.2 Algorithme CH

L'algorithme CH (Cotton et Hesse 1992) est basé sur une réattribution des nombres aléatoires permanents aux unités de la population.

ALGORITHME CH (TAS)

1. Initialisation : Nombres aléatoires initiaux $\omega_1 = (\omega_{k1},\ k \in U)$ où $\omega_{k1} \overset{i.i.d}{\sim} \text{Unif}(0,1), k \in U$ (sans doublons : $\omega_{k1} \neq \omega_{l1}$ si $k \neq l$). Tailles d'échantillons $(n_t, t \in T)$.

2. Tirage $t \geq 1$

(a) **Sélection de l'échantillon**

i. On ordonne les unités de la population par nombre aléatoire ω_t croissant.

ii. On obtient S_t en sélectionnant les n_t premières unités.

(b) **Renumérotation** : On détermine ω_{t+1} par une réattribution des numéros aléatoires aux unités de la population, qui respecte les rangs de ω_t dans S_t et $U \setminus S_t$, et qui associe

i. aux unités de S_t les n_t plus grands nombres aléatoires, et

ii. aux unités de $U \setminus S_t$ les $N - n_t$ plus petits nombres aléatoires.

3. On pose $t := t+1$ et on répète l'étape 2.

On considère comme exemple le tirage coordonné de 3 échantillons aléatoires simples de tailles $n_1 = n_2 = n_3 = 3$ dans une population de taille $N = 5$. Les nombres aléatoires initiaux sont $\omega = (0.5, 0.2, 0.4, 0.1, 0.3)$. La table ci-dessous montre le déroulement de l'algorithme utilisant les réattributions des nombres aléatoires initiaux aux unités de la population. On obtient la suite d'échantillons $S_1 = \{2, 4, 5\}$, $S_2 = \{1, 3, 4\}$ et $S_3 = \{2, 3, 5\}$.

Tirage 1

k	ω_1	I_1
1	0.5	0
2	0.2	1
3	0.4	0
4	0.1	1
5	0.3	1

Tirage 2

k	ω_2	I_2
1	0.2	1
2	0.4	0
3	0.1	1
4	0.3	1
5	0.5	0

Tirage 3

k	ω_3	I_3
1	0.4	0
2	0.1	1
3	0.3	1
4	0.5	0
5	0.2	1

On examine la procédure de renumérotation plus en détail. Nous commençons par la définition donnée par Cotton et Hesse (1992).

1. Soit $\omega = (\omega_k, k \in U)$ un vecteur de nombres aléatoires (sans doublons).
On détermine la statistique d'ordre de ω, que l'on note α .Si $\sigma = R(\omega)$ alors
$\alpha = \omega \circ \sigma^{-1}$.

2. On définit $\alpha \mapsto \beta$ par

$$\beta(k) = \begin{cases} \alpha(k+N-n) & si\ 1 \leq k \leq n \\ \alpha(k-n) & si\ n+1 \leq k \leq N \end{cases}$$

L'application $\alpha \mapsto \beta$ effectue (i) le transfert des n plus grands nombres aléatoires sur $\{1, \ldots, n\}$ et (ii) le transfert des $N-n$ plus petits nombres aléatoires sur $\{n+1, \ldots, N\}$.

3. On rétablit les liens avec les unités de la population en effectuant $\beta \circ \sigma$.

On peut décrire la procédure de renumérotation entièrement à l'aide de permutations. Soit en effet $\kappa = (12 \cdots N)$ la permutation cyclique fondamentale sur U. Alors $\beta(k) = \alpha(\kappa^{N-n}(k))$. La transformation β peut donc s'écrire $\beta = \alpha \circ \kappa^{N-n} = \alpha \circ \kappa^{-n}$, puisque $\kappa^N = id$. Nous pouvons donc écrire la renumérotation comme

$$\omega \mapsto \omega \circ R(\omega)^{-1} \kappa^{-n} R(\omega).$$

Si on passe aux permutations aléatoires grâce à l'application rang, la renumérotation $t \to t+1$ prend alors la forme

$$\sigma_{t+1} = R(\omega_{t+1}) = R(\omega_t)\, \sigma_t^{-1} \kappa^{-n_t} \sigma_t = \kappa^{-n_t} \sigma_t$$

Avec $m_t = \sum_{v=1}^{t} n_v$ le nombre d'unités tirées jusqu'au temps t ,nous avons $\sigma_{t+1} = \kappa^{m_t} \sigma_1$.

Nous pouvons aussi décrire le passage de σ_t à σ_{t+1} à l'aide des rangs par rapport à une partition. Si on reprend pour σ_t la description verbale de la procédure de renumérotation, nous avons les unités de S_t qui reçoivent les n_t plus grands rangs et les unités de $U \setminus \mathrm{S_t}$ qui reçoivent les $N - n_t$ plus petits rangs, tout en respectant les rangs de σ_t dans S_t et $U \setminus \mathrm{S_t}$. Comme $I_t = 1$ si $k \in S_t$ et si $k \in U \setminus \mathrm{S_t}$, nous voyons que $\sigma_{t+1} = R(I_t, \sigma_t)$. Comme $I_t = a_t \circ \sigma_t$ nous avons $\sigma_{t+1} = R(a_t \circ \sigma_t, \sigma_t) = R(a_t, id)\, \sigma_t$. On montre par un calcul direct que $R(a_t, id) = \kappa^{-n_t}$ et on retrouve bien $\sigma_{t+1} = \kappa^{-n_t} \sigma_t$. On arrive ainsi à la formulation suivante de l'algorithme HC.

1. Choisir une permutation aléatoire initiale σ_1 de loi uniforme sur S_U.

2. Tirage $t \geq 1$
   (a) Sélection de l'échantillon $I_t = a_t \circ \sigma_t$ où $a_t = (1_{n_t}, 0_{N-n_t})$.
   (b) Renumérotation $\sigma_{t+1} = R(I_t, \sigma_t)$.

3. On pose $t := t+1$ et on répète l'étape 2.

2.3 Algorithme RIV

L'algorithme RIV (Rivière 2001) utilise des réarrangements des nombres aléatoires permanents, qui tiennent compte de la charge cumulée. Etant donné un vecteur de nombres aléatoires x , un *réarrangement basé sur un vecteur de coûts c* est une application $y = \xi[c](x)$ obtenue par une permutation des composantes de x telle que

$$y_i \leq y_j \overset{def}{\Leftrightarrow} (c_i < c_j) \text{ ou } (c_i = c_j \text{ et } x_i \leq x_j).$$

ALGORITHME RIV (TAS)

1. **Initialisation** : Nombres aléatoires initiaux $\omega_1 = (\omega_{k1},\ k \in U)$ où $\omega_{k1} \overset{i.i.d}{\sim} \text{Unif}(0,1)$, $k \in U$ (sans doublons : $\omega_{k1} \neq \omega_{l1}$ si $k \neq l$). Tailles d'échantillons $(n_t, t \in T)$. Charges d'enquêtes $(c_t, t \in T)$. Charge cumulée initiale $b_0 = 0_N$.

2. Tirage $t \geq 1$

(a) **Sélection de l'échantillon**

i. On ordonne les unités de la population par nombre aléatoire ω_t croissant.

ii. On obtient S_t en sélectionnant les n_t premières unités.

(b) **Renumérotation**

i. On actualise la charge cumulée $b_t = b_{t-1} + c_t I_t$.

ii. On effectue un réarrangement de ω_t basé sur le charge cumulée $b_t : \omega_{t+1} = \xi[b_t](\omega_t)$

3. On pose $t := t+1$ et on répète l'étape 2.

On considère comme exemple le tirage coordonné de 3 échantillons aléatoires simples de tailles $n_1 = n_2 = n_3 = 3$ dans une population de taille $N = 5$. Les nombres aléatoires initiaux sont $\omega = (0.5, 0.2, 0.4, 0.1, 0.3)$. On prend une charge égale à 1 pour chaque unité et pour chaque tirage. La table ci-dessous montre le déroulement de l'algorithme utilisant les réarrangements des nombres aléatoires par rapport à la charge cumulée. On obtient la suite d'échantillons $S_1 = \{2, 4, 5\}$, $S_2 = \{1, 3, 4\}$ et $S_3 = \{2, 3, 5\}$.

Tirage 1				**Tirage 2**				**Tirage 3**			
k	ω_1	I_1	b_1	k	ω_2	I_2	b_2	k	ω_3	I_3	b_3
1	0.5	0	0	1	0.2	1	1	1	0.2	1	2
2	0.2	1	1	2	0.4	0	1	2	0.3	1	2
3	0.4	0	0	3	0.1	1	1	3	0.1	1	2
4	0.1	1	1	4	0.3	1	2	4	0.5	0	2
5	0.3	1	1	5	0.5	0	1	5	0.4	0	1

On va examiner plus en détails l'application $y = \xi[c](x)$ de réarrangement basé sur un vecteur de coûts. Par définition, $y = \xi[c](x)$ est obtenu par permutation des composantes de x. Il existe onc une permutation $\kappa = \kappa(x, c)$ telle que $y = x \circ \kappa$. On en déduit que les rangs de y sont reliés aux rangs de x par $R(y) = R(x)\kappa$. Maintenant, la permutation κ est telle que

$$y_i \leq y_j \text{ si et seulement si } (c_i < c_j) \text{ ou } (c_i = c_j \text{ et } x_i \leq x_j).$$

En intervertissant les indices i et j, on voit que cette définition est équivalente à

$$y_i \geq y_j \text{ si et seulement si } (c_i > c_j) \text{ ou } (c_i = c_j \text{ et } x_i \geq x_j).$$

Il suit que les rangs de y sont égaux aux rangs de x par rapport au vecteur c puisque

$$R(y)(k) = \sum_{l \in U} (y_k \geq y_l) = \sum_{l \in U} \left[(c_k > c_l) + (c_k = c_l)(x_k \geq x_l) \right] = R(c, x)(k).$$

Nous avons ainsi montré que $R(c, x) = R(y) = R(x)\kappa$, d'où l'on tire $\kappa = R(x)^{-1} R(c, x)$. Ainsi, si x est un vecteur dont toutes les composantes sont différentes, alors

$y = \xi[c](x) = x \circ R(x)^{-1} R(c, x)$. Le réarrangement de ω_t basé sur la charge cumulée b_t peut donc s'écrire

$$\omega_{t+1} = \xi[b_t](\omega_t) = \omega_t \circ R(\omega_t)^{-1} R(b_t, \omega_t)$$

Si on passe aux permutations aléatoires grâce à l'application rang, le réarrangement basé sur la charge cumulée b_t prend la forme

$$\begin{aligned}\sigma_{t+1} &= R(\omega_{t+1}) = R(\omega_t \circ R(\omega_t)^{-1}\ R(b_t, \omega_t)) = R(\omega_t)\ R(\omega_t)^{-1}\ R(b_t, \omega_t) \\ &= R(b_t, \omega_t) = R(b_t, \sigma_t)\end{aligned}$$

On arrive ainsi à la formulation suivante de l'algorithme RIV.

1. Choisir une permutation aléatoire initiale σ_1 de loi uniforme sur S_U.
Charge cumulée initiale $b_0 = 0_N$.

2. Tirage $t \geq 1$

   (a) Sélection de l'échantillon $I_t = a_t \circ \sigma_t$ où $a_t = (1_{n_t}, 0_{N-n_t})$.

   (b) Actualisation de la charge cumulée $b_t = b_{t-1} + c_t I_t$.

   (c) Renumérotation $\sigma_{t+1} = R(b_t, \sigma_t)$.

3. On pose $t := t+1$ et on répète l'étape 2.

2.4 Comparaison des algorithmes EDS, CH et RIV

Nous pouvons décrire les algorithmes EDS, CH et RIV de manière uniforme.

1. Tailles d'échantillons $(n_t, t \in T)$. Charges d'enquêtes $(c_t, t \in T)$. Charge cumulée initiale $b_0 = 0_N$. Choisir une permutation initiale σ_1 de loi uniforme sur S_U.

2. Tirage $t \geq 1$

   (a) Sélection de l'échantillon par $I_t = a_t \circ \sigma_t$ où $a_t = (1_{n_t}, 0_{N-n_t})$.

   (b) Actualisation de la charge cumulée $b_t = b_{t-1} + c_t I_t$ ($c_t = 1_N$ pour EDS).

   (c) Renumérotation

$$\sigma_{t+1} = \begin{cases} R(b_t, \sigma_1) & \text{EDS} \\ R(I_t, \sigma_t) & \text{CH} \\ R(b_t, \sigma_t) & \text{RIV} \end{cases}$$

3. On pose $t := t+1$ et on répète l'étape 2.

Les algorithmes EDS et CH, bien qu'en apparence différents, sont en fait équivalents. Pour l'algorithme EDS on a

$$\sigma_2 = R(b_1, \sigma_1) = R(I_1, \sigma_1) = R(a_1 \circ \sigma_1, \sigma_1) = R(a_1, id)\,\sigma_1 = \kappa^{n-1}\sigma_1.$$

De même on obtient

$$\sigma_3 = R(I_1 + I_2, \sigma_1) = R(a_1 + a_2 \circ \kappa^{n-1}, id)\,\sigma_1 = \kappa^{-(n_1+n_2)}\sigma_1.$$

On montre par induction que $\sigma_{t+1} = \kappa^{m_t}\,\sigma_1$, où $m_t = \sum_{v=1}^{t} n_v$ est la taille d'échantillon cumulée jusqu'au temps t. Pour l'algorithme CH,

$$\sigma_{t+1} = R(I_t, \sigma_t) = R(a_t \circ \sigma_t, \sigma_t) = R(a_t, id)\,\sigma_t = \kappa^{-n_t}\sigma_t,$$

d'où il suit que $\sigma_{t+1} = \kappa^{-m_t}\,\sigma_1,$ ce qui démontre l'équivalence des deux algorithmes.

Ces résultats montrent que les algorithmes CH et EDS effectuent bien des TAS, la transformation $\sigma_{t+1} = \kappa^{-n_t}\,\sigma_t$ étant clairement bijective. On voit aussi que le taux de recouvrement de deux échantillons consécutifs est optimal pour une coordination négative. De plus le *temps hors échantillon*, c'est-à-dire le nombre d'unités tirées avant qu'une unité donnée soit sélectionnée à nouveau, est également optimal, car il est égal à N.

Bien que la démonstration soit plus compliquée, on peut aussi montrer que l'algorithme RIV effectue bien des TAS et que le taux de recouvrement de deux échantillons consécutifs est optimal pour une coordination négative. On peut se demander si l'algorithme RIV est équivalent aux algorithmes CH et EDS. Ce n'est pas le cas si on prend une charge égale à 1 pour chaque unité et pour chaque tirage, comme le montre l'exemple suivant. On considère une population de taille $N = 6$ dans laquelle on tire des échantillons de tailles $n_t = 2$. On prend $\sigma_1 = Id$ comme permutation initiale. Si la charge est donnée par $c_t = 1_N$ pour tout $t \in T$, alors on obtient une suite de permutations sélectionnantes de période 6, cf. la table ci-dessous.

σ_1	b_1	σ_2	b_2	σ_3	b_3	σ_4	b_4	σ_5	b_5	σ_6	b_6
1	1	5	1	5	1	5	1	3	1	1	2
2	1	6	1	6	1	6	1	4	1	2	2
3	0	1	1	3	1	3	1	1	2	3	2
4	0	2	1	4	1	4	1	2	2	4	2
5	0	3	0	1	1	1	2	5	2	5	2
6	0	4	0	2	1	2	2	6	2	6	2

La suite des échantillons tirés 12, 34, 56, 56, 34, 12,... est différente de celle obtenue par les algorithmes EDS ou CH. Si on prend des charges strictement croissantes, par exemple $c_t = t1_N$ pour $t \in \{1, 2, \ldots\}$, alors l'algorithme RIV est équivalent aux algorithmes EDS ou CH. Dans notre exemple, les permutations sélectionnantes sont de périodes 3, comme pour les algorithmes EDS ou CH, cf. la table ci-dessous. La suite des échantillons tirés est 12, 34, 56, 12, 34, 56,... ce qui correspond à un temps hors échantillon optimal.

σ_1	b_1	σ_2	b_2	σ_3	b_3	σ_4
1	1	5	1	3	1	1
2	1	6	1	4	1	2
3	0	1	2	5	2	3
4	0	2	2	6	2	4
5	0	3	0	1	3	5
6	0	4	0	2	3	6

3. Sondage stratifié

3.1 Algorithme de tirage

On considère une population $U = \{1,\dots,k,\dots,N\}$ de taille N et une stratification de $U = \bigcup_{h\in H} U_h$ en H strates U_h de tailles N_h. Pour tirer un échantillon stratifié, on peut générer des nombres aléatoires ω_k indépendants et de loi uniforme sur l'intervalle $(0,1)$ puis, au sein de chaque strate, ordonner les unités de la population par nombres aléatoires ω_k croissants et sélectionner l'échantillon S_h en prenant les n_h premières unités dans la strate U_h.

Nous allons voir une stratification de la population U comme une application surjective $\zeta : U \to H.$ Les strates sont alors $U_h = \zeta^{-1}(h).$ On note $S_\zeta = \{\sigma \in S_U ; \zeta \circ \sigma = \zeta\}$ le sous-groupe de S_U qui laisse la stratification ζ invariante. Le vecteur d'allocation est donné par $a = ((1_{n_h}, 0_{N_h - n_h}), h \in H)$. Le tirage d'un échantillon stratifié peut alors être effectué par $I = a \circ R(\zeta,\sigma)$ où $R(\zeta,\sigma)$ sont les rangs de la permutation σ par rapport à la stratification ζ, cf. l'annexe A. On peut montrer que $R(\zeta,\sigma) = R(\zeta, id)\alpha$,où $\alpha \in S_\zeta$.

Nous avons alors

$$I = a \circ R(\zeta,\sigma) = (a \circ R(\zeta, id)) \circ \alpha = a_\zeta \circ \alpha,$$

ce qui permet de montrer que I est bien un échantillon stratifié si σ est de loi uniforme sur S_U

Nous considérons comme exemple une population de taille $N = 6$ avec la stratification $\zeta = (122112)$. Nous avons ainsi les deux strates $U_1 = \{1,4,5\}$ et $U_2 = \{2,3,6\}$ de tailles $N_1 = N_2 = 3$. Nous voulons tirer dans les strates des échantillons de tailles $n_1 = n_2 = 2$. Le vecteur d'allocation est alors donné par $a = (110110)$. Prenons $\sigma = 354621$ comme permutation aléatoire pour la sélection de l'échantillon. Dans la strate $U_1 = \{1,4,5\}$ nous avons les rangs 362 et nous tirons donc l'échantillon $S_1 = \{1,5\}$. De même, nous avons les rangs 541 dans la strate $U_2 = \{2,3,6\}$ et nous tirons l'échantillon $S_2 = \{3,6\}$. De manière équivalente, on calcule $R(\zeta,\sigma) = 265314$ et on tire l'échantillon par

$$I = a \circ R(\zeta,\sigma) = \begin{pmatrix} 123456 \\ 110110 \end{pmatrix} \begin{pmatrix} 123456 \\ 265314 \end{pmatrix} = \begin{pmatrix} 123456 \\ 101011 \end{pmatrix}$$

ce qui correspond bien à $S_1 = \{1,5\}$ et $S_2 = \{3,6\}$.

3.2 Algorithme CH pour le sondage stratifié

On considère dans cette section une suite de tirages aléatoires simples stratifiés (TASST) dans une population U aux temps $T = \{1,\dots,t,\dots,T\}$.

On se donne

- des stratifications $(\zeta_t, t \in T)$,
- des allocations $(a_t, t \in T)$ où $a_t = ((1_{n_{ht}}, 0_{N_{ht} - n_{ht}}), h \in H),$

- des vecteurs de nombres aléatoires $(\omega_t, t \in T)$,avec $\omega_t = (\omega_{kt}, k \in U)$.

Le tirage de l'échantillon $S_t \subseteq U$ est effectué à l'aide du vecteur de nombres aléatoires ω_t. De manière équivalente, le tirage de l'échantillon au temps t peut se faire par $I_t = a_t \circ R(\zeta_t, \sigma_t)$. Pour contrôler la répartition de la charge, σ_{t+1} doit dépendre des tirages précédants : $\sigma_{t+1} = \psi(\sigma_v, \zeta_v, a_v; v \le t)$. Pour que les échantillons marginaux correspondent à des TASST, on doit avoir $P(\sigma_t) = 1/N!$ pour tout $t \in T$.

L'algorithme CH pour la coordination négative d'une suite de TASST revient à effectuer par strate une réattribution des nombres aléatoires permanents aux unités de la population.

ALGORITHME CH (TASST)

1. **Initialisation** : Nombres aléatoires initiaux $\omega_1 = (\omega_{k1},\ k \in U)$ où $\omega_{k1} \overset{i.i.d}{\sim} \text{Unif}(0,1)$, $k \in U$ (sans doublons : $\omega_{k1} \neq \omega_{l1}$ si $k \neq l$). Stratifications $(\zeta_t, t \in T)$. Allocations $(a_t, t \in T)$.

2. Tirage $t \ge 1$

(a) **Sélection de l'échantillon**

i. On ordonne les unités de la population par nombre aléatoire ω_t croissant au sein de chacune des strates.

ii. On obtient S_t en sélectionnant les n_{ht} premières unités dans chacune des strates $h \in H_t$.

(b) Renumérotation : On détermine ω_{t+1} par une réattribution des numéros aléatoires aux unités de la population, qui respecte les rangs de ω_t dans S_{ht} et $U \setminus S_{ht}$ et qui associe

i. aux unités de S_{ht} les n_{ht} plus grands nombres aléatoires, et

ii. aux unités de $U \setminus S_{ht}$ les $N - n_{ht}$ plus petits nombres aléatoires.

3. On pose $t := t+1$ et on répète l'étape 2.

Nous considérons comme exemple une population de taille $N = 9$ avec la stratification $\zeta = (111122222)$. Nous avons ainsi les deux strates $U_1 = \{1,2,3,4\}$ et $U_2 = \{5,6,7,8,9\}$ de tailles $N_1 = 4$ et $N_2 = 5$. Nous voulons tirer dans les strates des échantillons de tailles $n_1 = 2$ et $n_2 = 3$. La table ci-dessous montre la réattribution des nombres aléatoires après le tirage du premier échantillon.

k	ζ	ω_1	I_1	ω_2
1	1	0.4	1	0.8
2	1	0.3	1	0.5
3	1	0.8	0	0.4
4	1	0.5	0	0.3
5	2	0.2	1	0.7
6	2	0.9	0	0.2
7	2	0.1	1	0.6
8	2	0.6	1	0.9
9	2	0.7	0	0.1

On examine la procédure de renumérotation plus en détail. La réattribution des numéros aléatoires se fait par strate comme pour le sondage aléatoire simple, c'est-à-dire essentiellement par $R(\zeta \wedge I, R(\zeta,\sigma))$, où $\zeta \wedge I$ est la partition du U obtenue en prenant l'intersection de la stratification ζ avec l'échantillon I . .Avec, par exemple, $\sigma = R(\omega_1) = 438529167$ on obtient

k	ζ	I	σ	$R(\zeta,\sigma)$	$R(\zeta \wedge I, R(\zeta,\sigma))$
1	1	1	4	2	4
2	1	1	3	1	3
3	1	0	8	4	2
4	1	0	5	3	1
5	2	1	2	6	8
6	2	0	9	9	6
7	2	1	1	5	7
8	2	1	6	7	9
9	2	0	7	8	5

On remarque que puisque $\zeta \wedge I$ est une partition plus fine que ζ, alors $R(\zeta \wedge I,\sigma) = R(\zeta \wedge I, R(\zeta,\sigma))$. La transformation $\sigma \mapsto R(\zeta \wedge I,\sigma)$ a brisé les liens entre les unités de la population et leurs rangs. On peut rétablir ces liens en utilisant la permutation $\sigma R(\zeta,\sigma)^{-1}$. La procédure de renumérotation est donc donnée par

$$\sigma \mapsto \sigma R(\zeta,\sigma)^{-1} \quad R(\zeta \wedge I,\sigma).$$

Dans notre exemple cela donne bien le résultat voulu :

$$\begin{pmatrix}123456789\\438529167\end{pmatrix}\begin{pmatrix}214369578\\123456789\end{pmatrix}\begin{pmatrix}123456789\\432186795\end{pmatrix}=\begin{pmatrix}123456789\\854372691\end{pmatrix}=R(\omega_2)$$

On arrive ainsi à la formulation suivante de l'algorithme CH pour la coordination négative de TASST.

1. Choisir une permutation aléatoire initiale σ_1 de loi uniforme sur S_U .

2. Tirage $t \geq 1$
   (a) Sélection de l'échantillon $I_t = a_t \circ R(\zeta_t,\sigma_t)$.
   (b) Renumérotation $\sigma_{t+1} = \sigma_t R(\zeta_t,\sigma_t)^{-1} R(\zeta_t \wedge I_t,\sigma_t)$.

3. On pose $t := t+1$ et on répète l'étape 2.

Dans le reste de cette section, on démontre que les échantillons fournis par l'algorithme CH sont bien des TASST. Comme σ_1 est de loi uniforme sur S_U, il suffit de montrer que la transformation

$$\sigma \mapsto \psi_{\zeta,a}(\sigma) = \sigma R(\zeta,\sigma)^{-1} R(\zeta \wedge I,\sigma), \text{ où } I = a \circ R(\zeta,\sigma),$$

est une bijection. Nous avons $R(\zeta \wedge I,\sigma) = R(\zeta \wedge I, R(\zeta,\sigma))$ et

$$\begin{aligned} R(\zeta \wedge I, R(\zeta,\sigma)) &= R(\zeta \wedge \left[a \circ R(\zeta,\sigma)\right], R(\zeta,\sigma)) \\ &= R(\zeta \circ R(\zeta,\sigma)^{-1} \wedge a, id)\, R(\zeta,\sigma). \end{aligned}$$

On peut montrer que chaque permutation $\sigma \in S_U$ a une décomposition unique $\sigma = \mu\alpha$ avec $\alpha \in S_\zeta$ et $R(\zeta, \mu) = R(\zeta, id)$. Alors

$$R(\zeta, \sigma) = R(\zeta, \mu\alpha) = R(\zeta \circ \alpha^{-1}, \mu)\alpha = R(\zeta, \mu)\alpha = R(\zeta, id)\alpha$$

et

$$\tilde{\zeta} = \zeta \circ R(\zeta, \sigma)^{-1} = (\zeta \circ \alpha^{-1}) \circ R(\zeta, id)^{-1} = \zeta \circ R(\zeta, id)^{-1},$$

d'où il suit que $\tilde{\zeta} = \zeta \circ R(\zeta, \sigma)^{-1}$, la stratification standard, ne dépend plus de σ □ et que

$$\psi_{\zeta,a}(\sigma) = \sigma\left[\alpha^{-1} R(\zeta, id)^{-1} R(\tilde{\zeta} \wedge a, id) R(\zeta, id)\alpha\right]$$

Si on pose

$$\kappa = \kappa(\zeta, a) = R(\zeta, id)^{-1} R(\tilde{\zeta} \wedge a, id) R(\zeta, id),$$

alors la renumérotation est donnée par

$$\sigma = \mu\alpha \mapsto \psi_\kappa(\sigma) = \sigma\left[\alpha^{-1}\kappa\alpha\right] = \mu(\kappa\alpha).$$

Finalement, on montre par un calcul direct que κ est le produit de permutations cycliques sur les strates, et donc $\kappa \in S_\zeta$.

On note que la permutation $R(\tilde{\zeta} \wedge a, id)$ est une permutation cyclique sur les strates standards, et que les permutations $R(\zeta, id)$ et $R(\zeta, id)^{-1}$ permettent de se remettre dans les strates effectives. Dans notre exemple, $\zeta = \tilde{\zeta}$, $R(\zeta, id) = id$ et $\kappa = R(\zeta \wedge a, id) = 341278956$, cf. la table ci-dessous.

k	ζ	a	id	$R(\zeta \wedge a, id)$
1	1	1	1	3
2	1	1	2	4
3	1	0	3	1
4	1	0	4	2
5	2	1	5	7
6	2	1	6	8
7	2	1	7	9
8	2	0	8	5
9	2	0	9	6

4. Coordination positive

On s'intéresse dans cette dernière section à la coordination positive d'un échantillon aléatoire simple avec un échantillon stratifié. Nous considérons une population U de taille N et un vecteur de nombres aléatoires $(\omega_k, k \in U)$ indépendants et uniformément distribués sur $(0,1)$. Le premier échantillon tiré est un échantillon aléatoire simple S_1 de taille n_1. Pour le deuxième tirage, nous avons une stratification de $U = \bigcup_{h \in H} U_h$ en H strates U_h de tailles N_h. On tire dans chaque strate un échantillon aléatoire simple S_{2h} de taille n_{2h}. La taille totale de l'échantillon au temps $t = 2$ est $n_2 = \sum_{h \in H} n_{2h}$. Pour avoir une coordination positive entre S_1 et S_2 on utilise le même vecteur de nombres aléatoires ω pour le tirage de S_1 et S_2 avec les algorithmes des sections 1 et 3.1. On veut calculer l'espérance de la taille du recouvrement

$E|S_1 \cap S_2|$. Ceci revient à calculer les probabilités jointes d'inclusion $\pi_{k,12}$ d'une unité $k \in U$ dans les deux échantillons S_1 et S_2, puisque $E|S_1 \cap S_2| = \sum_{k \in U} \pi_{k,12}$. Une borne supérieure pour $\pi_{k,12}$ est donnée par $\min(\pi_{k1}, \pi_{k2})$. Dans la situation que nous considérons, $\pi_{k1} = n_1 / N = f$ et, pour $k \in U_h$, $\pi_{k2} = n_{2h} / N_h = f_h$. Alors, pour une coordination positive optimale on aurait

$$E_{opt}|S_1 \cap S_2| = \sum_h N_h \min(f, f_h) = \sum_h \min\left(n_1 \frac{N_h}{N}, n_{2h}\right).$$

Nous allons montrer que pour le tirage considéré,

$$E\ |S_1 \cap S_2| = \sum_h E(\min(n_{1h}, n_{2h}))$$

où n_{1h} est de loi hypergéométrique $\mathcal{H}(N, N_h, n_1)$, ce qui est aussi ce qu'on obtiendrait avec la méthode des substitutions de Kish et Scott (1971). Comme $E(n_{1h}) = n_1 (N_h / N)$ nous pouvons aussi écrire l'espérance de la taille du recouvrement optimal comme $E_{opt}|S_1 \cap S_2| = \sum_h \min(E(n_{1h}), n_{2h})$.

Avec nos notations, nous avons une permutation aléatoire σ de loi uniforme sur S_U, une stratification $\zeta : U \to H$ et les vecteurs d'allocation $a_1 = (1_{n_1}, 0_{N-n_1})$ et $a_2 = ((1_{n_{2h}}, 0_{N_h - n_{2h}}),$ $h \in H)$. Le premier échantillon est donné par le TAS $I_1 = a_1 \circ \sigma$ et le deuxième échantillon est donné par le TASST $I_2 = a_2 \circ R(\zeta, \sigma)$. Comme dans la section 1, on peut calculer $\pi_{k,12}$ comme une somme sur l'ensemble des permutations de U :

$$\pi_{k,12} = E(I_{k1} I_{k2}) = \sum_{\sigma \in S_U} (a_1 \circ \sigma)(k)\ (a_2 \circ R(\zeta, \sigma))(k)\ P(\sigma).$$

Soit $S_\zeta = \{\alpha \in S_U ; \zeta \circ \alpha = \zeta\}$ le sous-groupe des permutations de U qui laissent la stratification ζ invariante et $M = \{\mu \in S_U ; R(\zeta, \mu) = R(\zeta, id)\}$. On peut montrer que chaque permutation $\sigma \in S_U$ a une décomposition unique $\sigma = \mu\alpha$ avec $\alpha \in S_\zeta$ et $\mu \in M$. Nous avons alors

$$R(\zeta, \sigma) = R(\zeta, \mu\alpha) = R(\zeta \circ \alpha^{-1}, \mu)\alpha = R(\zeta, \mu)\alpha = R(\zeta, id)\alpha.$$

Il suit que

$$\pi_{k,12} = \frac{1}{N!} \sum_{\mu \in M} \sum_{\alpha \in S_\zeta} (a_1 \circ \mu)\ (\alpha(k))\ (a_2 \circ R(\zeta, id))\ (\alpha(k)).$$

La sommation sur S_ζ se fait essentiellement comme pour le calcul des probabilités d'inclusion pour un TAS, cf. la section 1. On trouve que, pour $k \in U_h$ et $x, y \in \mathbb{R}^U$,

$$\sum_{\alpha \in S_\zeta} x(\alpha(k))\, y(\alpha(k)) = (N_h - 1)! \prod_{i \neq h} x_l y_l.$$

Nous avons donc, pour $k \in U_h$,

$$\pi_{k,12} = \frac{1}{N_h}\binom{N}{N_1 \cdots N_H}^{-1} \sum_{\mu \in M} \sum_{l \in U_h} a_1(\mu(l))\,(a_2 \circ R(\zeta, id))(l).$$

Il reste à calculer la somme $\sum_{\mu \in M} \sum_{l \in U_h} a_{\mu(l)}\, b_l$, où $a = a_1$ et $b = a_2 \circ R(\zeta, id)$. On peut montrer que

$$\sum_{\mu \in M} \sum_{l \in U_h} a(\mu(l))\, b(l) = \frac{N_h!(N - N_h)!}{\prod_{h \in H} N_h!} \sum_{1 \le j_1 < j_2 < \cdots < j_{N_h} \le N} \sum_{g=1}^{N_h} a(j_g)\, b(i_g),$$

où $U_h = (1 \le i_1 < i_2 < \cdots < i_{N_h} \le N)$. Alors, pour $k \in U_h$,

$$\pi_{k,12} = \frac{1}{N_h}\binom{N}{N_h}^{-1} \sum_{1 \le j_1 < j_2 < \cdots < j_{N_h} \le N} \sum_{g=1}^{N_h} a(j_g)\, b(i_g).$$

Dans le cas qui nous intéresse, nous avons $a = a_1 = (1_{n_1}, 0_{N-n_1})$ et $b = a_2 \circ R(\zeta, id)$ avec $a_2 = ((1_{n_{2h}}, 0_{N_h - n_{2h}}), h \in H)$. Nous avons ainsi

$$a(\mu(U_h)) = (a(j_1), \ldots, a(j_{N_h})) = (1_{n_{1h}}, 0_{N_h - n_{1h}}),$$

où n_{1h} est la taille de l'intersection de $\{1, \ldots, n_1\}$ avec $\mu(U_h) = (j_1 < \cdots < j_{N_h})$. On note qu'il y a $\binom{n_1}{n_{1H}}\binom{N - n_1}{N_h - n_{1h}}$ sous-ensembles de U de taille N_h ayant une intersection de taille n_{1h} avec $\{1, \ldots, n_1\}$ et que $\max(0, N_h - (N - n_1)) \le n_{1h} \le \min(n_1, N_h)$ alors $b(U_h) = (1_{n_{2h}}, 0_{N_h - n_{2h}})$.

Il suit que

$$\sum_{g=1}^{N_h} a(j_g)\, b(i_g) = \min(n_{1h}, n_{2h})$$

et

$$\sum_{1 \le j_1 < j_2 < \cdots < j_{N_h} \le N} \sum_{g=1}^{N_h} a(j_g)\, b(i_g) = \sum_{n_{1h}} \binom{n_1}{n_{1h}}\binom{N - n_1}{N_h - n_{1h}} \min(n_{1h}, n_{2h}).$$

Nous trouvons ainsi que, pour $k \in U_h$,

$$\pi_{k,12} = \frac{1}{N_h} \sum_{n_{1h}} \frac{\binom{n_1}{n_{1h}}\binom{N - n_1}{N_h - n_{1h}}}{\binom{N}{N_h}} \min(n_{1h}, n_{2h})$$

Ainsi, la probabilité jointe d'inclusion d'une unité $k \in U_h$ est donnée par

$$\pi_{k,12} = \frac{1}{N_h} E(\min(n_{1h}, n_{2h})),$$

où n_{1h} est de loi hypergéométrique $\mathcal{H}(N, n_1, N_h) = \mathcal{H}(N, N_h, n_1)$ et l'espérance de la taille du chevauchement est donnée par

$$E|S_1 \cap S_2| = \sum_{k \in U} \pi_{k,12} = \sum_h E(\min(n_{1h}, n_{2h})).$$

Annexe

A Rangs

Soit U une population de taille N. Pour un $\omega \in (0,1)^U$, sans doublons, nous considérons la statistique d'ordre :

$$\tilde{\omega} = (\omega_{(1)} < \cdots < \omega_{(k)} < \cdots < \omega_{(N)}$$

c'est-à-dire la suite ordonnée des éléments de ω,et le rang r_k de ω_k dans la suite ω :

$$r_k = \sum_{l \in U} (\omega_k \geq \omega_t),$$

où $(\omega_k \geq \omega_l) = 1$ si $\omega_k \geq \omega_l$ et $(\omega_k \geq \omega_l) = 0$ autrement. Ainsi le rang de ω_k dans la suite ω est le nombre d'éléments de la suite inférieurs ou égaux à ω_k. En particulier, $\omega_{(1)} = \min(\omega)$ et $\omega_{(N)} = \max(\omega).$

Comme ω est sans doublons, à chaque élément $\tilde{\omega}_i$ de la suite $\tilde{\omega}$ correspond un et un seul élément ω_j de la suite ω. Puisque le rang de $\tilde{\omega}_i$ dans $\tilde{\omega}$ est égal au rang de ω_j dans ω et que le rang de $\tilde{\omega}_i$ dans $\tilde{\omega}$ est égal à i, nous pouvons conclure que $i = r_j$, et donc que $\tilde{\omega}_i = \omega_j$, c'est-à-dire $\tilde{\omega}_{r_j} = \omega_j$.

Un vecteur $\omega \in (0,1)^U$ peut aussi être vu comme une application

$$\omega : U \to (0,1)\,,\ k \mapsto \omega(k) = \omega_k$$

Un vecteur $\omega \in (0,1)^U$ sans doublons, c'est-à-dire tel que $\omega_k \neq \omega_l$ si $k \neq l$, correspond ainsi à une application injective. Nous définissons donc $\Omega = \left\{\omega \in (0,1)^U \ ;\ \omega \text{ est injective}\right\}$. Pour un $\omega \in \Omega$, les composantes du vecteur des rangs sont toutes différentes. Vu comme une application de U sur U, le vecteur des rangs est donc une permutation, un élément de S_U,le groupe symétrique sur U. Nous pouvons donc définir les rangs comme une application

$$R : \Omega \to S_U\,, \quad \omega \mapsto R(\omega)$$

où $R(\omega)\,(k) = \sum_{l \in U} (\omega_k \geq \omega_l).$

On obtient la statistique d'ordre $\tilde{\omega}$ par permutation des composantes de ω. Il existe donc une permutation $\sigma \in S_U$ telle que $\tilde{\omega} = \omega \circ \sigma.$ Il suit de la définition de R que $R(\tilde{\omega}) = id.$ On montre que $R(\omega \circ \sigma) = R(\omega)\sigma.$ En effet, $R(\omega \circ \sigma)\,(k) = \sum_{l \in U} (\omega_{\sigma(k)} \geq \omega_{\sigma(l)})$ et $R(\omega)\,(\sigma(k)) = \sum_{l \in U} (\omega_{\sigma(k)} \geq \omega_l).$ Puisque σ est une permutation, nous avons bien le résultat voulu. Alors de $id = R(\tilde{\omega}) = R(\omega \circ \sigma)$ on déduit $id = R(\omega)\sigma,$ d'où $\sigma = R(\omega)^{-1}.$ Nous avons donc $\tilde{\omega} = \omega \circ R(\omega)^{-1}.$ ˜. Cette dernière relation est équivalente à $\omega_{r_j} = \omega_j$ car $\tilde{\omega} \circ R(\omega) = \omega.$

Soit $c \in \mathbb{N}^U$. Nous considérons la transformation de Ω définie par

$$\omega \mapsto (\omega \text{ ordonné par } c \text{ et } \omega) =: \tilde{\omega}_c.$$

Nous définissons les rangs de ω par rapport au vecteur c comme une application

$$R : \mathbb{N}^U \times \Omega \to S_U, \quad (c, \omega) \mapsto R(c, \omega),$$

où

$$R(c,\omega)\,(k) = \sum_{l \in U} \left[(c_k > c_l) + (c_k = c_l)\,(\omega_k \geq \omega_l) \right].$$

On note que l'on obtient $\tilde{\omega}_c$ par une permutation des composantes de ω. Il existe donc une permutation $\sigma \in S_U$ telle que $\tilde{\omega}_c = \omega \circ \sigma$. Par définition de σ nous avons $c \circ \sigma = \tilde{c}$. En effet, si on applique à c la permutation σ qui ordonne ω par c et ω, on obtient le vecteur c ordonné par c, c'est-à-dire la statistique d'ordre de c. Finalement, il suit de la définition de R que $R(\tilde{c}, \tilde{\omega}_c) = id$. Comme pour les rangs classiques, on montre que $R(c \circ \sigma, \omega \circ \sigma) = R(c, \omega)\sigma$, ce qui nous permet de conclure que la permutation σ qui ordonne ω par c et ω est donnée par $\sigma = R(c, \omega)^{-1}$. Nous avons ainsi montré que $\tilde{\omega}_c = \omega \circ R(c, \omega)^{-1}$.

Références

[1] Cotton, F. and Hesse, C. (1992). Tirages coordonnés d'échantillons. Document de travail de la Direction des Statistiques Economiques E9206. Rapport technique, INSEE, Paris.

[2] De Ree, J. (1999). Co-ordination of business samples using measured response burden. Invited paper, 52nd Session of the ISI Helsinki, Book 2, 289-292.

[3] Kish, L. and Scott, A. (1971). Retaining units after changing strata and probabilities. Journal of the American Statistical Association, 66, 461-470.

[4] Rivière, P. (2001). Random permutations of random vectors as a way to co-ordinate samples. Technical report, University of Southampton, UK.