# Tirages coordonnés et permutations

Paul-André Salamin
Office fédéral de la statistique, Service de méthodes statistiques

JMS2005 Paris

# 1 Le problème du sondage coordonné

Le but du sondage coordonné est de gérer le recouvrement de plusieurs échantillons tirés successivement dans une population $U$

Suite d'échantillons $(S_t, t \in T)$

Plans marginaux $P(S_t)$, probabilités d'inclusion marginales $\pi_{kt}$, $\pi_{kl,t}$

Coordination : probabilités jointes d'inclusion $\pi_{k,ut}$, $u \leq t$

Probabilités jointes d'inclusion

$$\pi_{k,12} = P\left(\{(S_1, S_2)\,; S_1 \ni k, S_2 \ni k\}\right)$$

Tirages indépendants : $\pi_{k,12} = \pi_{k1}\pi_{k2}$

Taille du recouvrement

$$E\,|S_1 \cap S_2| = \sum_{k \in U} \pi_{k,12}$$

Bornes pour $\pi_{k,12}$

$$\underbrace{\max\left(0, \pi_{k1} + \pi_{k2} - 1\right)}_{\text{coordination négative optimale}} \leq \pi_{k,12} \leq \underbrace{\min\left(\pi_{k1}, \pi_{k2}\right)}_{\text{coordination positive optimale}}$$

Tirage de $S_t$ qui

- respecte le plan pour $S_t$ : $P(S_t)$
- optimise les recouvrements avec les échantillons précédants : $\pi_{k,ut}$, $u \leq t$

Algorithme pour le sondage coordonné

- validité de l'algorithme
- qualité de l'algorithme

Autres aspects

- populations en évolution
- rotation d'un panel
- contrôle du temps hors échantillons

# 2 Tirage aléatoire simple (TAS)

Algorithme TAS ($|U| = N$, $|S| = n$)

1. Générer des nombres aléatoires $\omega_k \overset{i.i.d}{\sim} Unif(0,1)$ pour $k \in U$

2. Ordonner les unités de la population par nombres aléatoires $\omega_k$ croissants

3. Sélectionner dans l'échantillon $S \subset U$ les $n$ premières unités

Sélectionner des unités qui ont reçu de petits nombres aléatoires $\Leftrightarrow$
Sélectionner des unités qui ont des rangs bas dans la suite $\omega = (\omega_k, k \in U)$

Les nombres aléatoires *i.i.d.* $\omega_k$, $k \in U$, sont générés à partir d'une loi ayant une fonction de distribution continue $\Rightarrow$
Les *rangs* des unités $k \in U$ dans la suite $\omega$ sont de distribution uniforme sur $\mathcal{S}_U$
$\mathcal{S}_U =$ groupe des permutations des éléments de la population $U$

Algorithme TAS

1. Choisir une permutation aléatoire $\sigma \in \mathcal{S}_U$ telle que $P(\sigma) = 1/N!$

2. Sélectionner l'échantillon $S = \sigma^{-1}\{1, \ldots, n\}$.

$$N = 5, n = 3, \quad \sigma = \begin{pmatrix} 12345 \\ 52413 \end{pmatrix} \Rightarrow S = \{2, 4, 5\}$$

Remplacer $S \subseteq U$ par $I = (I_k, k \in U) \in \{0,1\}^U$
où $I_k = 1$ si $k \in S$ et $I_k = 0$ st $k \notin S$

$$S = \{2,4,5\} \subset U = \{1,2,3,4,5\} \leftrightarrow I = (01011)$$

Algorithme TAS

1. Choisir une permutation aléatoire $\sigma \in \mathcal{S}_U$ telle que $P(\sigma) = 1/N!$

2. Sélectionner l'échantillon $I = a \circ \sigma$ où $a = (1_n, 0_{N-n})$

$$I = a \circ \sigma = \binom{12345}{11100}\binom{12345}{52413} = \binom{12345}{01011} \leftrightarrow S = \{2,4,5\}$$

Calcul des probabilités d'inclusion: $I = a \circ \sigma$, $a = (1_n, 0_{N-n})$, $P(\sigma) = 1/N!$.

$$\pi_k = \sum_{\sigma \in \mathcal{S}_U} a(\sigma(k)) P(\sigma) = \frac{1}{N} \sum_{i \in U} a_i = \frac{n}{N}$$

$$\pi_{kl} = \sum_{\sigma \in \mathcal{S}_U} a(\sigma(k)) a(\sigma(l)) P(\sigma) = \frac{1}{N(N-1)} \sum_{i \neq j \in U} a_i a_j = \frac{n(n-1)}{N(N-1)}$$

# 3 Tirage aléatoire simple stratifié (TASST)

Population $U$ de taille $N$. Stratification de $U = \bigcup_{h \in H} U_h$ en $H$ strates $U_h$ de tailles $N_h$.

| $k$ | $\zeta$ |
|---|---|
| **1** | **1** |
| 2 | 2 |
| 3 | 2 |
| **4** | **1** |
| **5** | **1** |
| 6 | 2 |

Stratification $\zeta : U \longrightarrow H$.

Population de taille $N = 6$ avec la stratification $\zeta = (122112)$
Deux strates $U_1 = \{1, 4, 5\}$ et $U_2 = \{2, 3, 6\}$ de tailles $N_1 = N_2 = 3$
Echantillons de tailles $n_1 = n_2 = 2$
Permutation aléatoire pour la sélection de l'échantillon$R(\omega) = \sigma = 354621$
Strate $U_1 = \{1, 4, 5\}$ : rangs $362 \longrightarrow$ échantillon $S_1 = \{1, 5\}$
Strate $U_2 = \{2, 3, 6\}$ : rangs $541 \longrightarrow$ échantillon $S_2 = \{3, 6\}$

| $k$ | $\zeta$ | $\sigma$ | $I$ |
|---|---|---|---|
| **1** | **1** | **3** | **1** |
| 2 | 2 | 5 | 0 |
| 3 | 2 | 4 | 1 |
| **4** | **1** | **6** | **0** |
| **5** | **1** | **2** | **1** |
| 6 | 2 | 1 | 1 |

Algorithme TASST

1. Choisir une permutation aléatoire $\sigma \in \mathcal{S}_U$ telle que $P(\sigma) = 1/N!$
2. Sélectionner l'échantillon $I = a \circ R(\zeta, \sigma)$ où $a = \left(\left(1_{n_h}, 0_{N_h - n_h}\right), h \in H\right)$

$R(\zeta, \sigma) = R_\zeta(\sigma) =$ rangs de la permutation $\sigma$ par rapport à la stratification $\zeta$

$$\begin{array}{rccc} R_\zeta : & \mathcal{S}_U & \to & R_\zeta(id)\,\mathcal{S}_\zeta \\ & \sigma = \mu\alpha & \mapsto & R_\zeta(\alpha) \end{array}$$

$\mathcal{S}_\zeta = \{\sigma \in \mathcal{S}_U; \zeta \circ \sigma = \zeta\}$ le sous-groupe de $\mathcal{S}_U$ qui laisse la stratification $\zeta$ invariante.

$R(\zeta, \sigma)$ = rangs de la permutation $\sigma$ par rapport à la stratification $\zeta$

| $k$ | $\zeta$ | $\sigma$ | $R(\zeta, \sigma)$ | $I$ |
|---|---|---|---|---|
| **1** | **1** | **3** | **2** | **1** |
| 2 | 2 | 5 | 6 | 0 |
| 3 | 2 | 4 | 5 | 1 |
| **4** | **1** | **6** | **3** | **0** |
| **5** | **1** | **2** | **1** | **1** |
| 6 | 2 | 1 | 4 | 1 |

Exemple : $N_1 = N_2 = 3$, $n_1 = n_2 = 2$

$$I = a \circ R(\zeta, \sigma) = \begin{pmatrix} 123456 \\ 110110 \end{pmatrix} \begin{pmatrix} 123456 \\ 265314 \end{pmatrix} = \begin{pmatrix} 123456 \\ 101011 \end{pmatrix} \leftrightarrow S_1 = \{1, 5\}, S_2 = \{3, 6\}$$

Par exemple dans S-Plus

```
I=a[inverse(order(zeta,sigma))]
```

Propriétés des rangs par rapport à une stratification

- $R(\zeta, \sigma)$ est la permutation qui ordonne $\sigma$ par $\zeta$ et $\sigma$
- Pour toute permutation $\tau$, $R(\zeta \circ \tau, \sigma\tau) = R(\zeta, \sigma)\,\tau$

| $\zeta$ | 1212 | 1212 | 1212 | 1212 |
|---|---|---|---|---|
| | $M_\zeta$ | $M_\zeta \cdot 1432$ | $M_\zeta \cdot 3214$ | $M_\zeta \cdot 3412$ |
| $1324 \cdot \mathcal{S}_\zeta$ | 1324 | 1423 | 2314 | 2413 |
| $\mathcal{S}_\zeta$ | 1234 | 1432 | 3214 | 3412 |
| $1243 \cdot \mathcal{S}_\zeta$ | 1243 | 1342 | 4213 | 4312 |
| $2134 \cdot \mathcal{S}_\zeta$ | 2134 | 2431 | 3124 | 3421 |
| $2143 \cdot \mathcal{S}_\zeta$ | 2143 | 2341 | 4123 | 4321 |
| $3142 \cdot \mathcal{S}_\zeta$ | 3142 | 3241 | 4132 | 4231 |
| $R_\zeta \downarrow$ | $\downarrow$ | $\downarrow$ | $\downarrow$ | $\downarrow$ |
| $1324 \cdot \mathcal{S}_\zeta$ | 1324 | 1423 | 2314 | 2413 |

**Justification de l'algorithme**

Décomposition $\sigma = \mu\alpha$, où $\alpha \in \mathcal{S}_\zeta$

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

On peut choisir $\mu$ tel que $R(\zeta, \mu) = R(\zeta, id)$.
Alors $R(\zeta, \sigma) = R(\zeta, id)\, \alpha$ et

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

# 4 Validité d'un algorithme de coordination négative

Coordination négative pour une suite de TASST

Algorithme de Cotton et Hesse (algorithme CH)

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

Coordination négative pour des TASST, algorithme CH

1. Générer $\omega_1 = (\omega_k, k \in U)$, $\omega_k \overset{i.i.d}{\sim} Unif\,(0,1)$, $k \in U$
2. Sélection de l'échantillon $S_t$ de taille $n_t$
   (a) Ordonner les unités $k \in U$ par $\omega_t$ croissant au sein des strates
   (b) Sélectionner dans $S_t$ les $n_{ht}$ premières unités de $U_h$, $h \in H$
3. Réattribution des $\omega_t$ qui respecte les rangs de $\omega_t$ dans $S_{ht}$ et $U_{ht} \setminus S_{ht}$ et
   (a) qui associe aux unités de $S_{ht}$ les $n_{ht}$ plus grands nombres aléatoires
   (b) qui associe aux unités de $U_{ht} \setminus S_{ht}$ les $N_{ht} - n_{ht}$ plus petits nombres aléatoires

Coordination négative pour des TASST, algorithme CH

$N_1 = 4$, $N_2 = 5$, $n_1 = 2$, $n_2 = 3$

| $k$ | $\zeta_1$ | $\sigma_1$ | $I_1$ | $\sigma_2$ |
|---|---|---|---|---|
| 1 | 1 | 2 | 1 | 6 |
| 2 | 1 | 8 | 0 | 4 |
| **3** | **2** | **1** | **1** | **5** |
| **4** | **2** | **9** | **0** | **3** |
| **5** | **2** | **5** | **1** | **9** |
| **6** | **2** | **7** | **0** | **1** |
| **7** | **2** | **3** | **1** | **7** |
| 8 | 1 | 4 | 1 | 8 |
| 9 | 1 | 6 | 0 | 2 |

Coordination négative pour des TASST, algorithme CH

1. Choisir une permutation aléatoire $\sigma_1 \in \mathcal{S}_U$ telle que $P(\sigma_1) = 1/N!$
2. Sélectionner l'échantillon $I_t = a_t \circ R(\zeta_t, \sigma_t)$ où $a_t = \left(\left(\mathbf{1}_{n_{th}}, \mathbf{0}_{N-n_{th}}\right), h \in H\right)$
3. Transformation $\sigma_{t+1} = \sigma_t R(\zeta_t, \sigma_t)^{-1} R(\zeta_t \wedge I_t, \sigma_t)$

Si $\sigma_t = \mu_t \alpha_t$, où $\alpha_t \in \mathcal{S}_\zeta$ et $R(\zeta_t, \mu_t) = R(\zeta_t, id)$, alors

$$\sigma_{t+1} = \mu_t \left( \prod_{h \in H} \kappa_h^{-n_h} \right) \alpha_t$$

où $\kappa_h = \left(i_1 i_2 \dots i_{N_h}\right) \in \mathcal{S}_\zeta$ est la permutation cyclique fondamentale sur la strate $U_h = \left\{ i_1 < i_2 < \dots < i_{N_h} \right\}$ de la stratification $\zeta_t$.

La transformation $\sigma_t \mapsto \sigma_{t+1}$ est bijective, l'algorithme CH effectue bien des TASST.

# 5 Comparaison de trois algorithmes

Coordination négative pour une suite de TAS, algorithme EDS
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.

1. Générer $\omega = (\omega_k, k \in U)$, $\omega_k \overset{i.i.d}{\sim} Unif\,(0,1)$, $k \in U$
2. Sélection de l'échantillon $S_t$ de taille $n_t$
   (a) Ordonner les unités $k \in U$ par charge cumulée croissante et par nombre aléatoire $\omega$ croissant
   (b) Sélectionner dans l'échantillon $S_t$ les $n_t$ premières unités
3. Actualiser la charge cumulée $b_t(k) = \begin{cases} b_{t-1}(k) + 1 & \text{si } k \in S_t \\ b_{t-1}(k) & \text{si } k \in U \setminus S_t \end{cases}$

Coordination négative pour une suite de TAS, algorithme CH

1. Générer $\omega_1 = (\omega_k, k \in U)$, $\omega_k \overset{i.i.d}{\sim} Unif\,(0,1)$, $k \in U$
2. Sélection de l'échantillon $S_t$ de taille $n_t$
   (a) Ordonner les unités $k \in U$ par $\omega_t$ croissant
   (b) Sélectionner dans l'échantillon $S_t$ les $n_t$ premières unités
3. Réattribution des $\omega_t$ qui respecte les rangs de $\omega_t$ dans $S_t$ et $U \setminus S_t$ et
   (a) qui associe aux unités de $S_t$ les $n_t$ plus grands nombres aléatoires
   (b) qui associe aux unités de $S_t^c$ les $N - n_t$ plus petits nombres aléatoires

Coordination négative pour une suite de TAS, algorithme RIV

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

1. Générer $\omega_1 = (\omega_k, k \in U)$, $\omega_k \overset{i.i.d}{\sim} Unif\,(0,1)$, $k \in U$
2. Sélection de l'échantillon $S_t$ de taille $n_t$
   (a) Ordonner les unités $k \in U$ par $\omega_t$ croissant
   (b) Sélectionner dans l'échantillon $S_t$ les $n_t$ premières unités
3. Rénumérotation
   (a) Actualiser la charge cumulée $b_t = b_{t-1} + c_t I_t$
   (b) Rénumérotation de $\omega_t$ basé sur la charge cumulée
   $\omega_{t+1} = \xi\,[b_t]\,(\omega_t) = \omega_t \circ R(\omega_t)^{-1} R(b_t, \omega_t)$

Coordination négative pour une suite de TAS,
Forme standard des algorithmes EDS, CH et RIV

1. Générer $\sigma_1 \in \mathcal{S}_U$, $P(\sigma_1) = 1/N!$
2. Sélection de l'échantillon $I_t = a_t \circ \sigma_t$ où $a_t = \left(\mathbf{1}_{n_t}, \mathbf{0}_{N-n_t}\right)$
3. Actualisation de la charge cumulée $b_t = b_{t-1} + c_t I_t$ ($c_t = \mathbf{1}$ pour EDS)
4. Rénumérotation $\sigma_{t+1} = \begin{cases} R(b_t, \sigma_1) & \text{EDS} \\ R(I_t, \sigma_t) & \text{CH} \\ R(b_t, \sigma_t) & \text{RIV} \end{cases}$

EDS $\Leftrightarrow$ CH

# 6 Coordination positive

$S_1$ : TAS, $S_2$ : TASST

Tirages avec les *mêmes* nombres aléatoires.

Taille du recouvrement

$$E\,|S_1 \cap S_2| = \sum_{k \in U} \pi_{k,12}$$

Probabilités jointes d'inclusion

$$\pi_{k,12} = P\left(\{(S_1, S_2)\,; S_1 \ni k, S_2 \ni k\}\right) = \sum_{S_1 \ni k, S_2 \ni k} P\left(S_1, S_2\right)$$

où $P\left(S_1, S_2\right) = P\left(S_2 \mid S_1\right) P\left(S_1\right)$

Permutation aléatoire $\sigma$, $P(\sigma) = 1/N!$, stratification $\zeta$,
vecteurs d'allocation $a_1 = \left(1_{n_1}, 0_{N-n_1}\right)$ et $a_2 = \left(\left(1_{n_{2h}}, 0_{N_h - n_{2h}}\right), h \in H\right)$.

$$\pi_{k,12} = \sum_{\sigma \in \mathcal{S}_U} (a_1 \circ \sigma)(k)\,(a_2 \circ R(\zeta, \sigma))(k)\,P(\sigma)$$

"You see what I have done?" he asked the ceiling, ... "I have transformed the problem from an intractably difficult and possibly quite insoluble conundrum into a mere linguistic puzzle. Albeit," he muttered, after a long moment of silent pondering, "an intractably difficult and possibly insoluble one"

Douglas Adams "Dirk Gently's holistic detective agency"

$\mathcal{S}_\zeta = \{\alpha \in \mathcal{S}_U; \zeta \circ \alpha = \zeta\}$
$M_\zeta = \{\mu \in \mathcal{S}_U; R(\zeta, \mu) = R(\zeta, id)\}$
Chaque permutation $\sigma \in \mathcal{S}_U$ a une décomposition unique $\sigma = \mu\alpha$
avec $\alpha \in \mathcal{S}_\zeta$ et $\mu \in M_\zeta$

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

alors

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

Sommation sur $\mathcal{S}_\zeta$ : essentiellement comme pour le calcul des probabilités d'inclusion pour un TAS.

Pour $k \in U_h$,

$$\sum_{\alpha \in \mathcal{S}_\zeta} x\left(\alpha\left(k\right)\right) y\left(\alpha\left(k\right)\right) = \left(N_h - 1\right)! \prod_{i \neq h} N_i! \sum_{l \in U_h} x_l y_l$$

Pour $k \in U_h$,

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

Pour $k \in U_h$,

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

où $n_{1h} = |S_1 \cap U_h|$ est de loi hypergéométrique $\mathcal{H}\left(N, N_h, n_1\right)$ et $n_{2h} = |S_2 \cap U_h|$.
Taille du recouvrement

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

$$E\left(\min\left(n_{1h}, n_{2h}\right)\right) = \frac{1}{\binom{N}{n_1}} \sum_m \binom{N_h}{m} \binom{N - N_h}{n_1 - m} \min\left(m, n_{2h}\right)$$

où $\max\left(0, n_1 - (N - N_h)\right) \leq m \leq \min\left(N_h, n_1\right)$.

Bornes pour $\pi_{k,12}$

$$\underbrace{\max\left(0, \pi_{k1} + \pi_{k2} - 1\right)}_{\text{coordination négative optimale}} \leq \pi_{k,12} \leq \underbrace{\min\left(\pi_{k1}, \pi_{k2}\right)}_{\text{coordination positive optimale}}$$

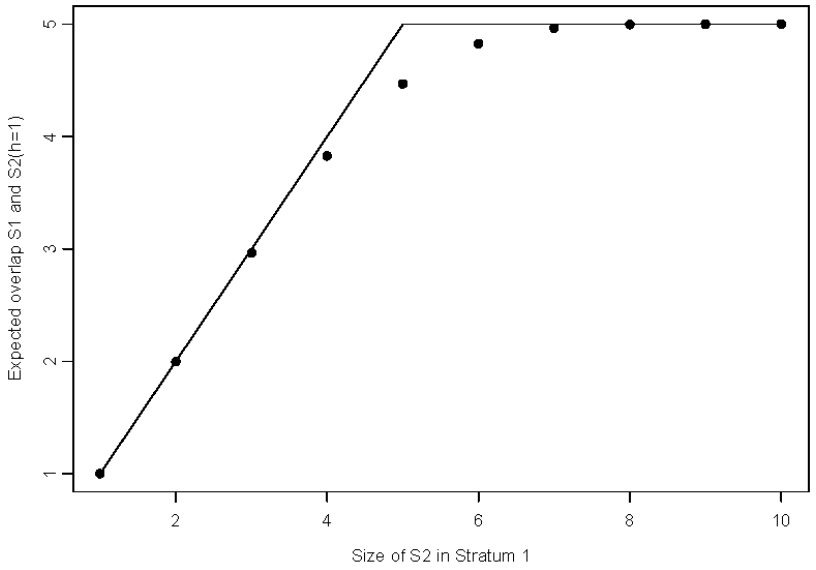
Coordination positive optimale

$$E_{opt}\left|S_1 \cap S_2\right| = \sum_h \min\left(n_1 \frac{N_h}{N}, n_{2h}\right) = \sum_h \min\left(E\left(n_{1h}\right), n_{2h}\right)$$

Exemple

| $h$ | $N_h$ | $E(n_{1h})$ | $n_{2h}$ | $E(n_{1h}) \wedge n_{2h}$ | $E(n_{1h} \wedge n_{2h})$ |
|---|---|---|---|---|---|
| 1 | 10 | 5 | 8 | 5 | 4.9956 |
| 2 | 30 | 15 | 12 | 12 | 11.9956 |
| | 40 | 20 | 20 | 17 | 16.9912 |

| $h$ | $N_h$ | $E(n_{1h})$ | $n_{2h}$ | $E(n_{1h}) \wedge n_{2h}$ | $E(n_{1h} \wedge n_{2h})$ |
|---|---|---|---|---|---|
| 1 | 10 | 5 | 5 | 5 | 4.4683 |
| 2 | 30 | 15 | 15 | 15 | 14.4683 |
| | 40 | 20 | 20 | 20 | 18.9366 |


Expected overlap S1 and S2(h=1)
1
2
3
4
5
2
4
6
8
10
Size of S2 in Stratum 1

# 7 Conclusion

Techniques pour étudier les propriétés de certains algorithmes de coordination (nombres aléatoires permanents, TASST). Trois exemples. Beaucoup d'autres applications possibles :

- Autres algorithmes (OFS, Y. Tillé)
- Tirages de Bernoulli par strate
- Méthode des substitutions de Kish et Scott
- Probabilités jointes d'inclusion dans des cas plus complexes
- Populations en évolution
- Rotation de panels
- Temps hors échantillons