Plans de sondage déterminantaux.

V. LOONIS [*] X. MARY [†]

30 mars 2015

Résumé

Dans cet article, on importe dans la théorie des sondages certains concepts et résultats des processus ponctuels. On définit et on étudie ainsi les propriétés théoriques d'une classe de plans de sondage : les plans de sondage déterminantaux. Pour ces plans, on dispose de l'ensemble de la loi de probabilité et d'un algorithme de sélection. Les calculs de précision sont rendus possibles par la connaissance des probabilités d'inclusion simple et double et l'existence de théorèmes asymptotiques et à distance finie. La recherche de plans optimaux selon un critère telle que la variance des estimations est rendue possible.

L'objectif de la théorie des sondages est de parvenir à la connaissance d'un paramètre d'intérêt θ à partir d'information partielle. Ce paramètre est fonction des y_k, valeurs prises par une variable d'intérêt y sur l'ensemble des individus k composant U de taille N. On ne dispose pas de l'ensemble des y_k mais seulement de ceux correspondant aux individus inclus dans un échantillon s. L'agrégation par une fonction $\hat{\theta}$, appelée estimateur de θ, des valeurs $\{y_k,\ k \in s\}$ conduit à une valeur $\hat{\theta}(s)$. Le passage de la connaissance de $\hat{\theta}$ à celle de θ est qualifiée d'inférence statistique. Les propriétés de l'inférence peuvent être formalisées dès lors que le processus conduisant à la sélection de s est aléatoire. On note $\mathbb{S}$ une variable aléatoire à valeur dans l'ensemble des échantillons possible. La probabilité d'inclusion d'un individu est la probabilité que l'individu k soit sélectionné : $\pi_k = pr(k \in \mathbb{S})$. La probabilité d'inclusion double que les individus k et l soient sélectionnés est $\pi_{kl} = pr((k,l) \in \mathbb{S})$. A chaque individu de U est associé un poids $w_k(\mathbb{S})$ dépendant éventuellement de l'échantillon aléatoire $\mathbb{S}$.

Un paramètre classique à estimer est le total sur U d'une variable y : $t_y = \sum_{k \in U} y_k$. Un estimateur du total t_y est dit linéaire homogène s'il s'´écrit sous la forme :

$$\hat{t}_{yw} = \sum_{k \in \mathbb{S}} w_k(\mathbb{S}) y_k \tag{1}$$

Quand les poids ne dépendent pas de $\mathbb{S}$, l'erreur quadratique moyenne de $\hat{t}_{yw}$ est :

$$EQM(\hat{t}_{yw}) \quad = \quad \overbrace{y^T \delta_w \Delta \delta_w y}^{Variance} + \overbrace{[e^T(\delta_w \delta_\pi - I_N)y]}^{Biais}{}^2 \tag{2}$$

[*] INSEE Division des Méthodes et Référentiels Géographiques (DMRG)

[†] Université Paris-Ouest Nanterre-La Défense : laboratoire MODAL'X

où y est le vecteur $(y_1, \cdots, y_N)^T$ de dimension N , δ_w et δ_π sont des matrices diagonales dont le k^{ieme} terme est respectivement w_k et π_k, et Δ est la matrice de terme général :

$$\begin{cases} \Delta_{kl} = \pi_{kl} - \pi_k \pi_l, \text{ si k} \neq \text{l} \\ \Delta_{kk} = \pi_k(1 - \pi_k), \text{ autrement} \end{cases} \tag{3}$$

Le plus connu des estimateurs linéaires d'un total dont les poids ne dépendent pas de $\mathbb{S}$ est l'estimateur d'Horvitz-Thompson [18] (aussi appelé estimateur de Narain-Horvitz-Thompson en référence aux travaux antérieurs de Narain [37]). Il est défini par $w_k = \pi_k^{-1}$ si π_k est strictement positive et w_k arbitraire sinon. On le note $\hat{t}_{yHT}$. Son intérêt provient notamment du fait que, quand π_k est strictement positive pour tout k, il est le seul à être sans biais parmi les estimateurs linéaires dont les poids ne dépendent pas de $\mathbb{S}$. On montre par ailleurs que si π_{kl} est également strictement positive pour tout couple (k, l) alors $EQM(\hat{t}_{yHT}) = var(\hat{t}_{yHT})$ est estimée sans biais par :

$$\hat{var}(\hat{t}_{yHT}) = \sum_{k \in \mathbb{S}} \sum_{l \in \mathbb{S}} \frac{y_k y_l}{\pi_k \pi_l \pi_{kl}} \Delta_{kl}. \tag{4}$$

Les propriétés statistiques de l'estimateur du total dépendent ainsi des probabilités d'inclusion simple et double. Une difficulté est que quand les π_k sont quelconques, les π_{kl} ne sont connus au mieux qu'asymptotiquement pour certains plans parmi les plus utilisés dans la pratique (sytématique, équilibré, sampford...). L'estimation de la variance passe alors par des approximations dont la validité est asymptotique. Dans cet article nous introduisons une famille de plans de sondages, les plans de sondage déterminantaux, dont les probabilités d'inclusion sont connues par une formule explicite à n'importe quel ordre. Cette formulation dépend en outre d'un paramètre dont la dimension est adaptable. La présence de ce paramètre autorise la recherche, à l'intérieur de la famille déterminantale, de plans optimaux pour un critère donné.

1 Plan de sondage déterminantal

1.1 Définition

La définition d'un plan de sondage sur une population U finie de taille N correspond exactement à celle d'un processus ponctuel simple sur $\mathscr{X}$, quand $\mathscr{X}$ est également un ensemble fini.

Définition 1.1. *[33] Un plan de sondage $\mathcal{P}$ non ordonné sans remise est une loi de probabilité sur $\mathcal{S}$=2^U ensemble des parties de U. On note alors $pr(s)$ pour tout $s \in \mathcal{S}$ la probabilité de l'échantillon s. On notera $\mathbb{S}$ une variable aléatoire à valeurs dans $\mathcal{S}$ de loi $\mathcal{P}$ telle $pr(\mathbb{S} = s) = pr(s)$.*

Définition 1.2. *[5] Un processus ponctuel simple $\mathcal{P}$ sur $\mathscr{X}$ est une loi de probabilité sur l'ensemble $2^{\mathscr{X}}$ des parties de $\mathscr{X}$.*

Il est donc possible d'importer dans une discipline certains résultats ou concepts de l'autre. Un processus ponctuel simple ayant fait l'objet de nombreux travaux récents est le processus déterminantal ([5], [20], [21], [26], [29]). La restriction au cas hermitien et

la transcription dans la terminologie de la théorie des sondages conduit à la définition du plan de sondage déterminantal.

Définition 1.3. *Un plan de sondage $\mathcal{P}$ est déterminantal s'il existe une matrice hermitienne K, appelée noyau, indéxée par les individus de U, telle que pour tout $s \in \mathcal{S}$:*

$$P(s \subseteq \mathbb{S}) = \det(K_{|s})$$

où $K_{|s}$ est la sous matrice de K indéxée par les individus de s. Un tel plan est noté $\mathcal{P}_K$.

Macchi [27] et Shoshnikov [29] montrent qu'un tel plan existe si et seulement si K est contractante, c'est à dire que ses valeurs propres sont à valeur dans $[0; 1]$. Ce résultat fondamental permet de voir les plans de sondages déterminantaux comme une famille paramètrique de plans de sondages, paramétrée par les matrices contractantes. Ces matrices admettent de nombreuses paramétrisations différentes (décomposition spectrale, réduction de forme quadratique, décomposition lagrangienne). Nous utiliserons en pratique la paramétrisation suivante : $K = \mu V \overline{V}^T$, avec V matrice (N, p) et $0 \leq \mu \leq \lambda_{max}$ où λ_{max} est la plus grande valeur propre de $\overline{V}^T V$. Il existe des processus déterminantaux associés à des matrices non-hermitiennes, mais pour lesquels l'absence de paramétrisation simple rend leur utilisation plus difficile en sondages.

Exemple 1.1. *Un premier exemple de matrice contractante est celui des matrices de projections (orthogonales), les valeurs propres appartenant à $\{0, 1\}$. A toute matrice P de projection est donc associé un plan de sondage déterminantal $\mathcal{P}_P$. Ces plans de sondage jouissent de propriétés particulières, qui seront étudiées précisément dans la suite de l'article.*

Exemple 1.2. *Un second exemple est celui des matrices diagonales dont les coefficients diagonaux $\pi_k \in [0, 1]$. De la définition, on déduit directement*

$$pr(s \subseteq \mathbb{S}) = \prod_{k \in s} \pi_k \tag{5}$$

et en utilisant le principe d'inclusion-exclusion

$$pr(\mathbb{S} = s) = \prod_{k \in s} \pi_k \prod_{k \notin s} (1 - \pi_k). \tag{6}$$

On retrouve l'équation qui caractérise le plan de sondage Poissonnien, qui est donc un plan de sondage déterminantal.

Exemple 1.3. *De manière plus générale, si Γ est une matrice diagonale dont les termes diagonaux appartiennent à $[0, 1]$ et P une matrice orthogonale alors $K = P\Gamma\overline{P}^T$ est contractante.*

Exemple 1.4. *Soit L une matrice définie-positive, alors $K = L(I_N + L)^{-1}$ est une matrice contractante. Macchi [27] établit la propriété suivante : le plan de sondage déterminantal associé à K vérifie*

$$pr(\mathbb{S} = s) = \frac{\det(L_{|s})}{\det(I_N + L)}.$$

On appelle un tel plan de sondage un L-plan de sondage.

Un autre processus ponctuel récemment étudié est le processus permanental. Nous ne le considérons pas dans cet article pour les deux raisons suivantes :

1. il ne s'agit pas d'un processus simple et correspondrait donc à un plan avec remise. Si de tels plans sont utiles dans certains domaines de la statistique (bootstrap non-paramétrique par exemple), l'application de la Rao-Blackwellisation à ces plans avec remise montre qu'ils n'ont que peu d'intérêt pour l'estimation d'un total ;
2. contrairement aux plans déterminantaux, les plans permanentaux ne peuvent pas être de taille fixe.

1.2 Probabilités d'inclusion

De la définition d'un plan de sondage déterminantal on déduit l'expresssion des probabilités d'inclusion d'un plan de sondage déterminantal.

$$\pi_k = P(k \in \mathbb{S}) = K_{kk}; \tag{7}$$

$$\pi_{kl} = P(\{k,l\} \in \mathbb{S}) = K_{kk}K_{ll} - \mid K_{kl} \mid^2 \ \ (k \neq l); \tag{8}$$

$$\Delta_{kl} = \left\{ \begin{array}{l} \pi_{kl} - \pi_k\pi_l = - \mid K_{kl} \mid^2 (k \neq l) \\ \pi_k(1-\pi_k) = K_{kk}(1-K_{kk})(k=l) \end{array} \right. ; \tag{9}$$

ou sous forme matricielle

$$\Delta = \overline{(I-K)} * K = (I-K) * \overline{K} \tag{10}$$

où $*$ désigne le produit de Schur-Hadamard.

Exemple 1.5. *Pour tout $0 < \alpha < 1$, il existe des plans de sondage dont les probabilités d'inclusion simple et double sont*

$$\begin{array}{rcl} \pi_k & = & \alpha \\ \pi_{kl} & = & \alpha^2(1 - \exp^{-2\beta(|x_k - x_l|_1)}) \end{array} ; \tag{11}$$

pour tout β assez grand, où x est une information auxiliaire de dimension Q disponible sur tous les éléments de U.

Le noyau de Laplace : $f^{\alpha,\beta}(x,y) = \alpha \exp^{-\beta|x-y|_1}$, est défini positif sur $\mathbb{R}^Q$ si α et β) sont positifs. Par plongement, il est en est de même de la matrice $L^{\alpha,\beta}$ de coefficients $L^{\alpha,\beta}_{kl} = f^{\alpha,\beta}(x_k, x_l)$. On choisit alors β de manière à ce que les valeurs propres de $L^{\alpha,\beta}$ soient plus petites que 1. La quantité $N\alpha$ s'interprête comme l'espérance de la taille de l'échantillon.

Proposition 1.1. *D'après* (9) *un plan de sondage déterminantal vérifie les conditions appelées de Sen-Yates-Grundy : $\pi_{kl} - \pi_k\pi_l \leq 0 (k \neq l)$.*

Ces conditions assurent la positivité de l'estimateur défini en 4 quand le plan est de taille fixe. Plus généralement, un plan de sondage déterminantal est un processus ponctuel à associations négatives ([26]). En particulier, il vérifie

$$P(A \cup B \subseteq \mathbb{S}) \leq P(A \subseteq \mathbb{S})P(B \subseteq \mathbb{S}) \tag{12}$$

Les processus déterminantaux vérifient en fait une propriété plus forte que l'association négative, à savoir la la propriété forte de Rayleigh ([6], [28]). Cette propriété concerne (la localisation des zéros de) la fonction génératrice du processus et s'avère avoir des conséquences très intéressantes pour l'étude des variables aléatoires dépendantes.

1.3 Taille des échantillons et plans déterminantaux de taille fixe

Les propriétés de la taille des échantillons aléatoires issus d'un plan de sondage déterminantal sont données par le théorème 7 de Hough et alii [20]. Pour tout ensemble A, on note $\sharp A$ son cardinal.

Théorème 1. *Soit $\mathbb{S}$ une variable aléatoire à valeur dans $\mathcal{S}$ dont la loi de probabilité est un plan de sondage déterminantal de noyau K. Alors la variable aléatoire $\sharp\mathbb{S}$ a la même loi qu'une somme de N variables indépendantes de Bernouilli de paramètre λ_i, où les λ_i sont les valeurs propres de K.*

Les corollaires suivants s'en déduisent directement.

Corollaire 1.1. *Soit $\mathcal{P}_\mathcal{K}$ un plan de sondage déterminantal associé à la matrice K. Alors :*

1. $E(\sharp\mathbb{S}) = tr(K)$ *et* $var(\sharp\mathbb{S}) = tr(K - K^2)$,
2. $\sum\limits_{k,l \in U} \Delta_{kl} = \sum\limits_{k \in U} \lambda_k(1 - \lambda_k)$
3. $pr(\mathbb{S} = \emptyset) = 0 \iff 1$ *est valeur propre de* K,
4. $\mathcal{P}_\mathcal{K}$ *est de taille fixe si et seulement si* K *est une matrice de projection orthogonale.*

Le point 2 du corollaire généralise le résultat connu pour les plans de sondage de taille fixe : $\sum_{l \in U} \Delta_{kl} = 0$ pour tout k et donc $\sum_{l,k \in U} \Delta_{kl} = 0$. Il provient de l'application à $\sum_{l,k \in U} \Delta_{kl} = e^T \Delta e = e^T \left[(I - K) * \overline{K}\right] e$, où e est le vecteur de taille N tel que $e^T = (1 \ \cdots 1)$, de la formule d'algèbre linéaire reliant produit de Hadamard et produit matriciel classique :

$$\overline{x}^T (A * B) y = Tr(\delta_x A \delta_y B). \tag{13}$$

où x et y sont des vecteurs de taille N, δ_x et δ_y des matrices diagonales de diagonale x et y et A et B des matrices (N, N).

Corollaire 1.2. *Soit $L \neq 0$ une matrice definie positive, alors le L-plan de sondage associé vérifie $pr(\mathbb{S} = \emptyset) > 0$. En particulier il n'est pas de taille fixe.*

Démonstration Soit $\mathcal{P}$ un L-plan de sondage. C'est donc un plan de sondage déterminantal de matrice $K = L(I + L)^{-1}$. Soit $X \in \mathbb{C}^N, KX = X$. Alors $(I + L)KX = LX = (I + L)X$ et $X = 0$. Donc 1 n'est pas valeur propre de K et $pr(\mathbb{S} = \emptyset) > 0$. Comme L est définie positive, $K \neq 0$ et ce n'est donc pas une projection. □

1.4 Autres propriétés des plans de sondage déterminantaux

Nous donnons dans cette section quelques résultats probabilistes généraux sur les plans de sondages déterminantaux et leur interprétation en théorie des sondages.

Proposition 1.2. *Soit $\mathcal{P}_K$ un plan de sondage determinantal de matrice K. Alors le plan de sondage complémentaire $\mathcal{P}_K^c$, qui consiste à choisir comme échantillon le complémentaire $\mathbb{S}^c$ de $\mathbb{S}$ ($P(\mathbb{S}^c = s) = P(\mathbb{S} = s^c)$), est un plan de sondage déterminantal de matrice $I - K$.*

Proposition 1.3 (Domaine)**.** *Soit $\mathcal{P}_K$ un plan de sondage determinantal sur la population $U = \{1, \cdots, N\}$ de matrice K, et $A \subseteq U$ une sous-population (un domaine). Alors la restriction du plan de sondage $\mathcal{P}_K$ à la sous-population A notée $\mathcal{P}_{K_{|A}}$ est un plan de sondage determinantal sur A de matrice $K_{|A}$, matrice de K restreinte aux individus de A.*

Proposition 1.4 (Plan de sondage Déterminantal stratifié)**.** *Soit $\{U_1, \cdots U_H\}$ une partition U en H strates, le plan de sondage déterminantal $\mathcal{P}_K$ de matrice K est stratifié si et seulement si il admet une écriture diagonale par blocs relativement à ces strates, c'est à dire $k \in U_h, l \in U_{h'}, h \neq h' \Rightarrow K_{kl} = 0$.*

En utilisant le principe d'inclusion-exclusion, on obtient que les probabilités de disjonction sont également données par un déterminant (Lyons 2003 [26], Théorème 5.1 equation 5.2 pour les plans de taille fixe et équation 8.1 pour les plans de taille aléatoire) :

Théorème 2 (Disjonction)**.** *Soit $\mathcal{P}_K$ un plan de sondage déterminantal de matrice K. Alors*

$$P(s \in \mathbb{S}, s' \notin \mathbb{S}) = det(B) \tag{14}$$

avec B matrice carrée de taille $\sharp s + \sharp s'$ indexée par $s \dot{\cup} s'$ et définie par $B_{kl} = K_{kl}$ si $k \in s$ et $B_{kl} = (I_N - K)_{kl}$ si $k \in s'$.

Cette formule permet en particulier de calculer l'entropie de plans de sondages déterminantaux.

Proposition 1.5. *Soit $\mathcal{P}_K$ un plan de sondage determinantal de matrice K, alors :*

$$pr(s) = \det(B(s)) \tag{15}$$

$$H(\mathcal{P}) = -\sum_{s \in \mathbb{S}} pr(s) \log(pr(s)) = -\sum_{s \in \mathbb{S}} \det(B(s)) \log(\det(B(s))) \tag{16}$$

où $B(s)$ est la matrice carrée de taille N définie par $B_{kl} = K_{kl}$ si $k \in s$ et $B_{kl} = (I_N - K)_{kl}$ si $k \notin s$.

Enfin, un plan de sondage déterminantal de taille fixe conditionné à contenir certains individus et en exclure certains autres est encore determinantal (sur les individus restant), la matrice de projection associée étant connue ([26], Proposition 6.3 et équation 6.5). Ce résultat s'étend aux plans de sondages déterminantaux de taille aléatoire, mais la formule n'est plus explicite ([26]).

1.5 Algorithme

La sélection d'unités dans U selon un plan de sondage déterminantal est rendue possible par l'algorithme 18 de Hough et alii [20]. On trouve une formulation très proche de cet algorithme dans Scardicchio et alii [31] et un autre basé sur la procédure de Gram-Schmidt dans Lavancier et alii [25]. Pour des raisons de lecture, on note $A(i,j)$ le coefficient en ligne i et colonne j d'une matrice A.

Le premier algorithme décrit la sélection d'un échantillon pour un plan de sondage de taille fixe n. Soit K_n la matrice de projection de rang n sur $H_n \subseteq \mathbb{C}^N$.

Algorithme 1.1. *pour i de n à 1*

1. *On sélectionne un individu parmi N avec les probabilités $\frac{K_i(k,k)}{i}$, soit k_i cet individu,*
2. *On note F_i la matrice représentative d'une base orthonormée de H_i, on pose*

$$Z_i = \frac{\overline{F}_i^T e_{k_i} e_{k_i}^T F_i}{K_i(k_i,k_i)}$$

   où e_{k_i} est le vecteur dont les N coordonnées sont nulles sauf la $k_i - ème$ qui vaut 1.
3. *On pose $K_{i-1} = F_i(I_i - Z_i)\overline{F}_i^T$, et H_{i-1} l'espace de projection associé.*

L'échantillon constitué des individus k_n à k_1 est une réalisation d'un plan de sondage déterminantal de matrice K_n.

Quand K n'est pas une matrice de projection, on se ramène au cas de la matrice de projection grâce au Théorème 7 de Hough et alii [20].

Théorème 3. *Soit $\mathcal{P}_K$ un plan de sondage déterminantal associé à la matrice K dont la décompostion spectrale est :*

$$K = \sum_{i=1}^{i=N} \lambda_i \phi_i \overline{\phi}_i^T$$

où λ_i est la $i - ème$ valeur propre associée au vecteur propre ϕ_i et soit $\mathcal{P}_{K_I}$ le plan associé à la matrice de projection

$$K_I = \sum_{i=1} B_i \phi_i \overline{\phi}_i^T$$

où les $B_i,\ 1 \leq i \leq N$, sont des variables de Bernouilli indépendantes de paramètre λ_i alors

$$\mathcal{P}_K \sim \mathcal{P}_{K_B}.$$

De manière concrête, on génére le vecteur b réalisation de N Bernouilli indépendantes de paramètres λ_i. Dans un deuxième temps, on construit la matrice de projection K_b associée à ce vecteur. On sélectionne alors un échantillon selon l'algorithme précédent. Le thérorème de Hough et alii assure que cet échantillon est une réalisation de $\mathcal{P}_K$.

1.6 Plans à probabilités d'inclusion fixées

Il est courant de contraindre les probabilités d'inclusion simple π_k à être égales à une quantité donnée Π_k, où Π un vecteur de taille N tel que $0 \leq \Pi_k \leq 1 (k \in 1, \ldots, N)$. Il existe un plan de sondage déterminantal très simple vérifiant ces probabilités : le plan de sondage Poissonnien donné par l'équation 6. Mais il n'est pas de taille fixe. Le théorème suivant prouve qu'il existe des plans de sondages déterminantaux de taille fixe et de probabilités d'inclusion simple fixées, sous l'hypothèse que $\sum_{k=1}^{N} \Pi_k$ soit un entier.

Théorème 4. *Soit Π un vecteur de taille N tel que $0 \leq \Pi_k \leq 1 (k \in 1, \ldots, N)$ et $\sum_{k=1}^{N} \Pi_k = n \in \mathbb{N}$. Alors il existe un plan de sondage déterminantal de taille fixe n et de probabilités d'inclusions simples $\pi_k = \Pi_k (k \in 1, \ldots, N)$.*

Démonstration Il s'agit d'une application du Théorème de Schur-Horn [22]. Posons $\sum_{k=1}^{N} \Pi_k = n$, $\lambda_k = 1$ pour $1 \leq k \leq n$ et $\lambda_k = 0$ pour $n+1 \leq k \leq N$. Alors pour tout $1 \leq l \leq N$, on a $\sum_{i=1}^{l} \Pi_k \leq l \wedge n \leq \sum_{k=1}^{l} \lambda_k$ et $\sum_{k=1}^{N} \Pi_k = n = \sum_{k=1}^{N} \lambda_k$. D'après le Théorème de Schur-Horn, il existe une matrice Hermitienne K dont les valeurs propres sont les $\{\lambda_k, 1 \leq k \leq N\}$ et de diagonale $\{\Pi_k, 1 \leq k \leq N\}$. La matrice K est donc une projection de rang n. □

Réciproquement, si $\mathcal{P}_K$ est un tel plan de sondage de matrice associée K, alors $\sum \Pi_k = tr(K) = rg(K) \in \mathbb{N}$. De plus, $K_{k,k} = (K^2)_{k,k} = \sum_{l=1}^{N} K_{kl}^2$ d'où $K_{k,k} - K_{k,k}^2 \geq 0$, soit $0 \leq \Pi_k \leq 1$.

Kadison [24] propose un algorithme basé sur les rotations pour construire une telle projection. Plus généralement, [14] propose des algorithmes pour construire des matrices hermitiennes de diagonale et spectre fixés.

Les plans à probabilités constantes c sont des cas particuliers des plans à probabilités d'inclusion simple fixée. On appelle de tels plans de sondage des plans équipondérés. D'après le théorème 4, il existe donc un plan de sondage déterminantal de taille fixe dont les probabilités d'inclusion simples sont égales à c, lorsque cN est entier. Le résultat qui suit propose une construction directe de tels plans de sondages, basée sur les racines $N-$ièmes de l'unité.

Soit N entier fixé. Le polynôme $z^N - 1$ admet N racines et $\phi(N)$ racines primitives (telles que N soit le plus petit entier d, $z^d = 1$), où ϕ est l'indicatrice d'Euler. Notons z une racine primitive N-ième de l'unité. Alors le groupe $\{z, \ldots, z^N = 1\}$ est cyclique d'ordre N.

Posons $c = n/N$. On définit pour tout $k = 0, \ldots, n-1$ les vecteur :

$$U_k = \frac{\sqrt{c}}{\sqrt{n}} ((z^k)^1, \ldots, (z^k)^N)^T$$

et $V_k = {}^t\overline{U_k}$.

Lemme 1. *La famille $V_1, \ldots, V_n$ est orthonormale.*

Démonstration Soit $k \in \{0, \ldots, n-1\}$. Alors

$$U_k V_k = \sum_{j=1}^{N} U_k(j) V_k(j) = n^{-1} c \sum_{j=1}^{N} |z|^2 k = n^{-1} cN = 1.$$

Soit $k > l,\ k, l \in \{0, \ldots, n-1\}$. Alors

$$U_k V_l = n^{-1} c \sum_{j=1}^{N} z^{jk} \overline{z}^{jl} = n^{-1} c \sum_{j=1}^{N} (z^{k-l})^j = 0.$$

En effet, pour toute racine q de l'unité différente de 1 (d'où l'interêt de considérer des racines primitives), $\sum_{j=1}^{N} q^j = \sum_{j=0}^{N-1} q^j = (1-q^N)(1-q)^{-1} = 0$. Comme la famille est orthonormale, la matrice P est une projection. □

La matrice $P = \sum_{k=1} n V_k U_k$ ainsi construite est donc une projection de rang n. Sa diagonale vérifie $P_{j,j} = \sum_{k=1}^{n} V_k(j) U_k(j) = n^{-1} c \sum_{k=1}^{n} |z|^2 k = c$ pour tout $j = 1, \ldots, N$. On a ainsi construit $\phi(N)$ projections de diagonale constante $\frac{n}{N}$, pour tout $n \leq N$.

Le sondage aléatoire simple faisant partie des plans à probabilités d'inclusion simple fixée, on peut se demander s'il appartient à la classe des plans de sondages détérminantaux. On montre en annexe que le sondage aléatoire simple n'est pas déterminantal en général.

Lemme 2. *Si $n \notin \{0, 1, N, N-1\}$, le plan de sondage aléatoire simple n'est pas déterminantal.*

On peut cependant se demander s'il existe un plan de sondage déterminantal ayant les mêmes probabilités d'inclusion simple et double que le sondage aléatoire simple de taille n, c'est à dire vérifiant $\pi_k = \frac{n}{N}$ et $\pi_{kl} = \frac{n(n-1)}{N(N-1)}$.

Définition 1.4. *On appelle plan de sondage determinantal (N, n)-simple un plan de sondage déterminantal $\mathcal{P}$ sur une population de taille N vérifiant $\pi_k = \frac{n}{N}$ et $\pi_{kl} = \frac{n(n-1)}{N(N-1)}$.*

Si $\mathcal{P}$ est un plan de sondage déterminantal (N, n)-simple, alors le plan de sondage $\mathcal{P}^c$ consistant à prendre le complémentaire de l'échantillon $\mathbb{S}$ est déterminantal $(N, N-n)$-simple.

Lemme 3. *Soient $n \leq N$ deux entiers et $\mathcal{P}$ un plan de sondage déterminantal (N, n)-simple. Alors $\mathcal{P}$ est de taille fixe n.*

Démonstration Soit $n_{\mathbb{S}}$ le cardinal de l'échantillon aléatoire. Sa variance dépend uniquement des probabilités d'inclusions simples et double, qui sont celles du plan de sondage aléatoire simple, qui est de taille fixe. Donc $var(n_{\mathbb{S}}) = 0$, et $\mathcal{P}$ est de taille fixe. □

Lemme 4. *Soient $n \leq N$ deux entiers et $\mathcal{P}$ un plan de sondage déterminantal $(N,n)-$simple. Alors la matrice associée K vérifie*

$$K_{k,k} = \frac{n}{N}, \qquad \forall\, 1 \leq k \leq N \tag{17}$$

$$|K_{kl}|^2 = \frac{n(N-n)}{N^2(N-1)}, \qquad \forall\, 1 \leq k \neq |l \leq N \tag{18}$$

Théorème 5. *Soient $0 < n < N$ deux entiers. Alors il existe un plan de sondage déterminantal (N,n)-simple $\mathcal{P}$ seulement si $N \leq \min\{n^2, (N-n)^2\}$.*

Démonstration Soit K la matrice d'un tel plan de sondage. C'est une projection de rang n. Nous allons utiliser la sous-multiplicativité du rang pour le produit de Hadamard, noté $*$. Soit J la matrice définie par $J_{kl} = 1\ \forall 1 \leq k,l \leq N$. On a

$$K * \overline{K} = \alpha I + \beta J \tag{19}$$

avec $\alpha = -\frac{n(n-1)}{N(N-1)}$ et $\beta = \frac{n(N-n)}{N^2(N-1)}$. En effet, $(K*\overline{K})_{k,k} = \frac{n^2}{N^2}$ pour tout $1 \leq k \leq N$ et $(K * \overline{K})_{kl} = |K_{kl}|^2 = \frac{n(N-n)}{N^2(N-1)}$ pour $k \neq l$ d'après le Lemme précédent. Comme les valeurs propres de J sont 0 et N, et que $\beta\alpha \notin \{0, -N\}$, $K * \overline{K}$ est inversible et donc de rang N. Mais $rg(K * \overline{K}) \leq rg(K)rg(\overline{K}) = rg^2(K)$ par sous-multiplicativité et finalement, $N \leq n^2$. De la même façon, en étudiant la matrice de projection $I - K$ (associée à $\mathcal{P}^c$ de rang $N-n$) et le produit de Hadamard $(I-K) * \overline{(I-K)}$ on obtient $N \leq (N-n)^2$. □

La preuve du théorème n'est pas nouvelle, elle apparaît dans le contexte des frames équiangulaires ([34].)

Les matrices considérées dans ce théorème sont à valeurs complexes. Si l'on considère seulement des matrices réelles, on trouve des conditions beaucoup plus restrictives :

Théorème 6 (théroème 4.1 [11])**.** *Soit $1 < n < N-1$. Quand $N \neq 2n$ une condition nécessaire pour l'existence d'un plan de sondage (N,n)-simple, associé à une matrice K réelle, est que les deux quantités*

$$\alpha = \sqrt{\frac{n(N-1)}{N-n}}; \beta = \sqrt{\frac{(N-n)(N-1)}{n}}$$

sont des entiers impairs.
Quand $N = 2n$, il est nécessaire que n soit un entier impair et que $N-1$ soit la somme de deux carrés.

Les articles [32] et [11] permettent de déduire l'ensemble des plans de sondages déterminantaux (N,n)-simples à noyau réel pour respectivement $N \leq 100$ ([32] Table I) et $n \leq 50$ ([11] Table III). Les tables II et III de [32] donnent certains plans de sondages (N,n)-simples à noyau complexe. La table 1 reprend certaines de ces informations pour $n < 9$. Dans cette table l'indication $\mathbb{C}$ indique qu'il n'existe pas de plan de sondage (N,n)-simple à noyau réel mais qu'il en existe un à noyau complexe.

TABLE 1 – Existence de plans de sondage (N,n)-simples, selon le type de noyau (réel ou complexe) pour $n<9$.

n	3	3	4	4	5	5	6	6	6	7	7	7	8	8	8
N	6	7	7	13	10	11	11	16	31	14	15	28	15	29	57
	$\mathbb{R}$	$\mathbb{C}$	$\mathbb{C}$	$\mathbb{C}$	$\mathbb{R}$	$\mathbb{C}$	$\mathbb{C}$	$\mathbb{R}$	$\mathbb{C}$	$\mathbb{R}$	$\mathbb{C}$	$\mathbb{R}$	$\mathbb{C}$	$\mathbb{C}$	$\mathbb{C}$

Exemple 1.6 (plan (6,3)-simple)**.** *Soit la matrice*

$$K=\frac{1}{2}\begin{pmatrix} 1 & \frac{1}{\sqrt{5}} & \frac{1}{\sqrt{5}} & \frac{1}{\sqrt{5}} & \frac{1}{\sqrt{5}} & \frac{1}{\sqrt{5}} \\ \frac{1}{\sqrt{5}} & 1 & -\frac{1}{\sqrt{5}} & -\frac{1}{\sqrt{5}} & \frac{1}{\sqrt{5}} & \frac{1}{\sqrt{5}} \\ \frac{1}{\sqrt{5}} & -\frac{1}{\sqrt{5}} & 1 & \frac{1}{\sqrt{5}} & -\frac{1}{\sqrt{5}} & \frac{1}{\sqrt{5}} \\ \frac{1}{\sqrt{5}} & -\frac{1}{\sqrt{5}} & \frac{1}{\sqrt{5}} & 1 & \frac{1}{\sqrt{5}} & -\frac{1}{\sqrt{5}} \\ \frac{1}{\sqrt{5}} & \frac{1}{\sqrt{5}} & -\frac{1}{\sqrt{5}} & \frac{1}{\sqrt{5}} & 1 & -\frac{1}{\sqrt{5}} \\ \frac{1}{\sqrt{5}} & \frac{1}{\sqrt{5}} & \frac{1}{\sqrt{5}} & -\frac{1}{\sqrt{5}} & -\frac{1}{\sqrt{5}} & 1 \end{pmatrix}.$$

C'est une matrice de projection, et le plan de sondage déterminantal associé est $(6,3)$*-simple. On vérifie aisément qu'il n'est pas aléatoire simple car les échantillons* $\{1,2,3\}$ *et* $\{4,5,6\}$ *n'ont pas les mêmes probabilités (respectivement* $\frac{1}{8}(1-\frac{3}{5}-\frac{2}{5\sqrt{5}})$ *et* $\frac{1}{8}(1-\frac{3}{5}+\frac{2}{5\sqrt{5}})$*).*

On constate que certains plans aléatoires simples non déterminantaux ont un analogue déterminantal présentant les mêmes probabilités d'inclusion simple et double. Dans le cas général, on définit pour tout plan de sondage $\mathcal{P}$ et toute matrice ζ quelconque la matrice $K^{\Delta,\zeta}=\zeta * K^{\Delta}$ avec

$$K^{\Delta}_{kl}=\begin{cases} \sqrt{\pi_{kl}-\pi_k\pi_l}\ (k\neq l) \\ \pi_k\ (k=l) \end{cases}$$

Il existe alors un plan de sondage déterminantal de même probabilités d'inclusion simple et double que $\mathcal{P}$ si et seulement si il existe une matrice $\zeta_{\mathcal{P}}$ telle que

$$\zeta_{kl}=\begin{cases} \overline{\zeta_{lk}}=\frac{1}{\zeta_{kl}}\ (k\neq l) \\ 1\ (k=l) \end{cases}$$

et $K^{\Delta,\zeta_{\mathcal{P}}}$ soit contractante.

Dans l'exemple précédent, le passage d'un plan de sondage aléatoire simple de taille $n=3$ dans une population de taille 6 à un plan $(6,3)$-simple est possible grâce à la matrice :

$$\zeta=\begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 & 1 & 1 \\ 1 & -1 & 1 & 1 & -1 & 1 \\ 1 & -1 & 1 & 1 & 1 & -1 \\ 1 & 1 & -1 & 1 & 1 & -1 \\ 1 & 1 & 1 & -1 & -1 & 1 \end{pmatrix}.$$

2 Estimation d'un total

2.1 Erreur Quadratique Moyenne

Dans le cas d'un plan de sondage déterminantal, les formules (9) et (2) permettent d'écrire l'erreur quadratique moyenne de l'estimateur linéaire homogène de poids w, indépendant de $\mathbb{S}$, du total t_y d'une variable d'intérêt y :

Proposition 2.1. *Soit $\mathcal{P}_K$ un plan de sondage déterminantal de noyau K et $\hat{t}_{yw}$ un estimateur linéaire homogène de t_y, alors l'erreur quadratique moyenne de $\hat{t}_{yw}$ est telle que :*

$$EQM(\hat{t}_{yw}) = y^T\delta_w((I-K)*\overline{K})\delta_w y + [e^T(\delta_w(I*K)-I_N)y]^2 \quad (20)$$

Dans le cas où la variable d'intérêt est positive, l'EQM est telle que :

$$= \langle Z(I-K)Z, ZKZ\rangle + [\langle I, ZKZ\rangle - t_y]^2 \quad (21)$$
$$= \langle\langle I-K, K\rangle\rangle + [\langle I, ZKZ\rangle - t_y]^2 \quad (22)$$

où Z est la matrice diagonale de diagonale $\sqrt{w_i y_i}$ et $\langle A, B\rangle = trace(\overline{A}^T B) = \sum_{k,l}\overline{a_{k,l}}b_{k,l}$ le produit scalaire canonique sur l'espace des matrices carrées de taille N. Ce produit scalaire apparaît suite à l'application de la formule (13). On notera $|A|$ la norme associée (norme de Frobenius). Enfin on pose $\langle\langle A, B\rangle\rangle = \langle ZAZ, ZBZ\rangle$ ($||A|| = |ZAZ|$) ($\langle\langle ., .\rangle\rangle$ est alors un produit scalaire).

Dans le cas de l'estimateur d'Horvitz-Thompson, on obtient :

$$EQM(\hat{t}_{yHT}) = y^T\delta_{\pi^{-1}}((I-K)*\overline{K})\delta_{\pi^{-1}}y \quad (23)$$
$$= \langle Z(I-K)Z, ZKZ\rangle \quad (24)$$
$$= \langle\langle I-K, K\rangle\rangle \quad (25)$$

Les formulations précédentes permettent de savoir si, pour un vecteur y strictement positif donné, il est possible d'estimer t_y avec un plan de sondage déterminantal à diagonale fixée de telle sorte que l'EQM de l'estimateur soit 0. Il faut tout d'abord que l'estimateur soit non biaisé. On suppose donc que les poids vérifient $w_k = \frac{1}{\pi_k},\ k = 1, \ldots, N$ (estimateur D'Horvitz-Thompson).

Proposition 2.2. *Soit y une vecteur strictement positif. Un total t_y sera estimé avec une EQM=0 si et seulement si il existe une matrice de projection Kde diagonale Π telle que K commute avec Z^2.*

Démonstration Pour toutes matrices semi-définies positives A et B, on a $tr(AB) \geq 0$, avec égalité si et seulement si $AB = 0$. Soit K une matrice de C, telle que $var(\hat{t}_{yHT}) = 0$, on a alors $tr(Z(I-K)ZZKZ) = 0$, avec $Z(I-K)Z$ et ZKZ semi-définies positives. Donc $Z(I-K)Z^2KZ = 0$ soit, en multipliant par Z^{-1} à gauche et à droite $Z^2K = KZ^2K$, et en prenant l'adjoint, $Z^2K = KZ^2K = KZ^2$. Finalement, K commute avec Z^2. On en déduit alors que $Z^2K = Z^2K^2$ et en multipliant par Z^{-2} à gauche, $K^2 = K$, soit K est une projection. Réciproquement, si K est une projection commutant avec Z^2 on a bien $Z(I-K)Z^2KZ = Z^3(I-K)KZ = 0$. □

(Dans le cas où le vecteur y peut prendre des valeurs nulles, on peut adapter la preuve en utilisant l'inverse généralisée de groupe $Z^\sharp$ de la matrice Z, et l'on obtient comme condition nécéssaire et suffisante $ZZ^\sharp KZZ^\sharp$ est une projection commutant avec Z^2.)

L'existence d'une telle matrice de projection est fortement contrainte. Soit $\alpha_1, \ldots, \alpha_q$ les valeurs distinctes de $\frac{y_k}{\Pi_k},\ k = 1, \ldots, N$, et A_j l'ensemble des indices k tels que $\frac{y_k}{\Pi_k} = \alpha_j$. Alors

Corollaire 6.1. *Le problème d'estimation avec une erreur quadratique moyenne nulle admet une solution exacte parmi les plan de sondages déterminantaux si et seulement si* $\forall j = 1, \ldots, q,\ \sum_{k \in A_j} \Pi_k \in \mathbb{N}^*$. *Les plans de sondages déterminantaux sont alors les plans de sondages stratifiés par les strates* $A_j,\ j = 1, \ldots, q$, *respectant les probabilités d'inclusions simples et de taille fixe sur chaque strate.*

La formulation de l'EQM d'un plan de sondage déterminantal sous forme de produit scalaire peut être étendu à tout plan de sondage en utilisant les notations de la section 1 :

Proposition 2.3. *Soit* $\mathcal{P}$ *un plan de sondage quelconque et* $\hat{t}_{yw}$ *un estimateur linéaire homogène de* t_y, *alors l'erreur quadratique moyenne de* $\hat{t}_{yw}$ *est telle que :*

$$\begin{aligned} & EQM(\hat{t}_{yw}) \\ & \quad = \quad y^T \delta_w((I - K^{\Delta,\zeta}) * \overline{K^{\Delta,\zeta}})\delta_w y + [e^T(\delta_w(I * K^{\Delta,\zeta}) - I_N)y]^2 \end{aligned}$$

et dans le cas d'une variable positive,

$$\begin{aligned} & = \quad \langle Z(I - K^{\Delta,\zeta})Z, ZK^{\Delta,\zeta}Z\rangle + [\langle I, ZKZ\rangle - t_y]^2 \\ & = \qquad \langle\langle I - K^{\Delta,\zeta}, K^{\Delta,\zeta}\rangle\rangle + [\langle I, ZKZ\rangle - t_y]^2 \end{aligned}$$

pour toute matrice ζ *verifiant*

$$\zeta_{kl} = \begin{cases} \overline{\zeta_{lk}} = \frac{1}{\zeta_{kl}}\ (k \neq l) \\ 1\ (k = l) \end{cases}$$

Dans le cas général, la question de l'existence un plan de sondage à probabilités d'inclusion fixées tel que $EQM(\hat{t}_{yHT}) = 0$ a été récemment étudiée par Deville [**?**].

Proposition 2.4. *Si le vecteur* y *est positif et pour des probabilités d'inclusion fixées* $\pi_k = \Pi_K > 0$, *la variance de l'estimateur d'Horvitz-Thompson du total de* y *est maximale pour le plan poissonnien parmi les plans de sondages déterminantaux de diagonale fixée.*

Démonstration Soit $\mathcal{P}_K$, un plan de sondage déterminantal de noyau K tel que $\pi_k = \Pi_k > 0$, la variance de l'estimateur d'Horvitz-Thompson se déduite de la formule 2 :

$$var(\hat{t}_{yHT}) = y^T(I_N * K)^{-1}\Delta(I_N * K)^{-1}y$$

On décompose Δ en $\Delta^{poiss} + (\Delta - \Delta^{poiss})$ où Δ^{poiss} est la matrice diagonale de terme général $\pi_k(1-\pi_k)$. On déduit que :

$$\begin{aligned} var(\hat{t}_{yHT}) &= y^T(I_N * K)^{-1}\Delta^{poiss}(I_N * K)^{-1}y - \sum_{k\in U}\sum_{l\in U, l\neq U}\frac{y_k y_l}{\pi_k \pi_l} \mid K_{kl} \mid^2 \\ &\leq y^T(I_N * K)^{-1}\Delta^{poiss}(I_N * K)^{-1} = var^{poiss}(\hat{t}_{yHT}) \end{aligned}$$

□

De manière plus générale, le plan poissonnien est à variance maximale parmi les plans vérifiant les conditions Sen-Yates-Grundy, plans dont font partie les plans déterminantaux.

2.2 Propriétés statistiques de l'estimateur du total

Lorsqu'on ne considère pas de modèle de superpopulation, le cadre usuel est celui d'une suite de populations finies emboîtées U_N, et une suite de données y_k, tel que décrit dans Isaki and Fuller [23]. Dans ce cadre, les principaux résultats portent sur la convergence en moyenne quadratique de l'estimateur d'Horvitz-Thompson. Nous rappelons ici le Théorème de Cardot et alii [9] appliqué aux plans de sondage vérifiant les conditions de Sen-Yates-Grundy par Chauvet [13].

Théorème 7. *Soit $\{\mathcal{P}_N, N \in \mathbb{N}\}$ une suite de plans de sondage de taille (aléatoire) d'espérance n_N sur U_N vérifiant les conditions de Sen-Yates-Grundy, et $(y_k, k \in \mathbb{N})$ une suite positive. Si*

1. $\lim_{N\to\infty} \frac{n_N}{N} \to f \in]0,1[$;
2. $\min_{k\in U_N} \pi_k > \lambda_1 > 0$;
3. $\frac{1}{N}\sum_{k\in U}(\frac{y_k}{\pi_k})^2 = O(1)$;

alors l'estimateur de la moyenne $\frac{\hat{t}^N_{yHT}}{N}$ est convergent en moyenne quadratique.

L'inégalité de Bienaymé-Tchebychev permet d'en déduire classiquement la convergence en probabilité.

Le théorème de Schur sur les matrices semi-définies positives permet d'améliorer ce résultat dans le cadre des plans de sondage déterminantaux, pour des variables y de signe quelconque.

Théorème 8. *Soit $\{\mathcal{P}_N, N \in \mathbb{N}\}$ une suite de plans de sondage déterminantaux sur $U_N = \{1, \cdots, N\}$ de matrices associées $\{K_N, n \in \mathbb{N}\}$ de termes diagonaux non-nuls. Si*

$$\frac{1}{N^2}\sum_{k\in U}\frac{y_k^2}{K_N(k,k)} \to 0$$

alors l'estimateur de la moyenne $\frac{\hat{t}^N_{yHT}}{N}$ est convergent en moyenne quadratique.

Démonstration Soit $\mathcal{P}$ un plan de sondage déterminantal de matrice K, π le vecteur des termes diagonaux. D'après les résultats de la section précédente, $V(\hat{t}_{yHT}) = y^T\delta_{\pi^{-1}}((I-K)*\overline{K})\delta_{\pi^{-1}}y$. Les matrices I, K, $(I-K)$ et $\overline{K}$ étant semi-définies positives, le Théorème de Schur affirme que $((I-K)*\overline{K})$, $I*\overline{K}$ et $K*\overline{K}$ sont semi-définies positives. Or $((I-K)*\overline{K}) = I*\overline{K} - K*\overline{K}$, donc $((I-K)*\overline{K}) \leq I*\overline{K}$ pour l'ordre partiel des matrices semi-définies positives. On a donc

$$\begin{aligned} y^T\delta_{\pi^{-1}}((I-K)*\overline{K})\delta_{\pi^{-1}}y &\leq y^T\delta_{\pi^{-1}}(I*\overline{K})\delta_{\pi^{-1}}y \\ &\leq y^T\delta_{\pi^{-1}}\delta_{\pi}\delta_{\pi^{-1}}y \\ &\leq y^T\delta_{\pi^{-1}}y \end{aligned}$$

Avec les notations du théorème, on obtient alors

$$V\left(\frac{\hat{t}^N_{yHT}}{N}\right) \leq \frac{1}{N^2}\sum_{k\in U}(\frac{y_k^2}{K_N(k,k)} \longrightarrow 0.$$

□

Dans le cadre d'une suite de plans de sondage déterminantaux équipondérés de taille d'espérance n_N et d'une variable $y_k \in \{0,1\}$, la condition suffisante est simplement $n_N \to \infty$.

En supposant que les $w_k y_k$ prennent leur valeurs dans un ensemble fini, et en supposant de plus que le plan de sondage est *assez* aléatoire, on obtient un théorème centrale limite en appliquant les résultats de Soshnikov ([29],[30]). Dans ces articles on trouve plusieurs théorèmes sur la la normalité asymptotique de fonctionnelles de processus déterminantaux. Le théorème portant sur les fonctions étagées est directement applicable à l'étude des plans de sondages déterminantaux et aux estimateurs linéaires homogènes dont les poids ne dépendent pas de l'échantillon :

Théorème 9 (Soshnikov 2000 et 2002)**.** *Soit $\{\mathcal{P}_N, N \in \mathbb{N}\}$ une suite de plans de sondage déterminantaux sur $U_N = \{1,\cdots,N\}$ de matrices associées $\{K_N, n \in \mathbb{N}\}$. Soit $\{w_k, k \in \mathbb{N}\}$ une suite de réels telle que les variables $w_k y_k, k \in \mathbb{N}$ prennent leurs valeurs dans un ensemble fini de nombres réels $\{\alpha_1,\cdots,\alpha_r\}$. On définit pour tout N l'estimateur linéaire homogène*

$$\hat{t}^N_{yw} = \sum_{k\in\mathbb{S}_N} w_k y_k = \sum_{k\in\mathbb{S}_N}\sum_{1\leq i\leq r}\alpha_i 1_{\{w_k y_k=\alpha_i\}} = \sum_{1\leq i\leq r}\alpha_i.\sharp\{k\in U_N | w_k y_k = \alpha_i\}.$$

Si la variance $Var(\hat{t}^N_{yw})$ tend vers $+\infty$ et si, pour tout $1 \leq i \leq r$,

$$Var\left(\sharp\{k\in\mathbb{S}_N | w_k y_k = \alpha_i\}\right) = O\left(Var\left(\hat{t}^N_{yw}\right)\right)$$

alors

$$\frac{\hat{t}^N_{yw} - E(\hat{t}^N_{yw})}{\sqrt{Var(\hat{t}^N_{yw})}} \overset{loi}{\to} \mathcal{N}(0,1). \tag{26}$$

L'hypothèse selon laquelle les $w_k y_k$ prennent leurs valeurs dans un ensemble fini est classique (cf Hartley et Rao 68, Ericson 69 et Scott 71). Elle est de plus aisèment acceptable en pratique puisque l'on ne considérera que des populations finies, et que le nombre r peut-être arbitrairement grand.

Plus récemment, la normalité asymptotique a été étendue à des classes de processus plus généraux (incluant les processus déterminantaux) : les processus à associations négatives ou positives [38], et les processus vérifiant la propriété forte de Rayleigh [**?**]. D'un tout autre point de vue, les travaux de Berger [4] permettent d'obtenir la normalité asymptotique sous des conditions d'entropie asymptotiquement maximale.

Pour les processus vérifiant la propriété forte de Rayleigh, Pemantle et Peres [28] ont récemment démontré des inégalités de concentration et de déviation dont l'application à la théorie des sondages permet d'obtenir des résultats suivants.

Théorème 10. *Soit $\mathcal{P}_K$ un plan de sondage déterminantal de noyau K de trace μ. Soit y une variable d'intérêt et $\hat{t}_{yw}$ un estimateur linéaire du total t_y de y dont les poids ne dépendent pas de l'échantillon aléatoire $\mathbb{S}$. On note $C = \max_{k\in U} |w_k y_k|$. Alors, pour tout a positif,*

$$pr(\hat{t}_{yw} - E(\hat{t}_{yw}) > a) \leq 3\exp\left(-\frac{a^2}{16\left(aC + 2\mu C^2\right)}\right) \tag{27}$$

$$pr(|\hat{t}_{yw} - E(\hat{t}_{yw})| > a) \leq 5\exp\left(-\frac{a^2}{16^2\left(aC + 2\mu C^2\right)}\right). \tag{28}$$

Si $\mathcal{P}_K$ est de taille fixe $\mu = n$, alors :

$$pr(\hat{t}_{yw} - E(\hat{t}_{yw}) > a) \leq \exp\left(-\frac{a^2}{8nC^2}\right) \tag{29}$$

$$pr(|\hat{t}_{yw} - E(\hat{t}_{yw})| > a) \leq 2\exp\left(-\frac{a^2}{8nC^2}\right). \tag{30}$$

Démonstration La fonction qui à s associe $\sum_{k\in U} w_k y_k 1_{\{k\in s\}}$ est C-Lipschitzienne pour la distance de Hamming ce qui permet d'appliquer les théorèmes 3.1 et 3.2 de Pemantle et Peres (2013). □

Une appliction des ces inégalités à l'estimation d'une proportion $p = N^{-1}t_y$ avec $y_k \in \{0, 1\}$, avec un plan déterminantal équipondéré de taille $n = 20000$ dans une population de taille $N = 6.10^6$, montre qu'avec une probabilité plus grande que 0.95 la proportion est estimé avec erreur maximale de 3.8 points.

Ces résultats permettent également, dans le cadre des plans de sondage déterminantaux, d'obtenir une convergence en probabilité de $\hat{t}_{yHT}$ vers t_y sous une condition plus faible que celle du Théorème 7, et assez proche de celle du Théorème 8.

Corollaire 10.1. *Soit $\{\mathcal{P}_N, N \in \mathbb{N}\}$ une suite de plans de sondages déterminantaux sur $U_N = \{1, \cdots, N\}$ de matrices associées $\{K_N, n \in \mathbb{N}\}$. Si*

$$\frac{1}{N}\max_{k\in U_N}\frac{y_k}{K_N(k,k)} \to 0,$$

alors

$$\frac{\left(\hat{t}_{yHT} - t_y\right)}{N} \xrightarrow{pr} 0.$$

Démonstration Posons $C_N = \max_{k \in U_N} \frac{y_k}{K_N(k,k)}$, on a

$$pr(|\hat{t}_{yHT} - t_y| > Na) \leq 5 \exp\left(-\frac{N^2 a^2}{16^2 \left(NaC_N + 2\mu C_N^2\right)}\right). \tag{31}$$

et le terme de droite tend vers 0 d'après l'hypothèse. □

Soit $\{\mathcal{P}_N, N \in \mathbb{N}\}$ une suite de plans de sondage déterminantaux équipondérés de taille $n_N = \sqrt{N}$ et $y_k \in \{0, 1\}$ telle que $\sum_{k \in U} y_k \to \infty$, alors

$$\frac{1}{N} \max_{k \in U_N} \frac{y_k}{K_N(k,k)} \leq \frac{1}{\sqrt{N}} \to 0 \tag{32}$$

Cependant, la proportion $\frac{n_N}{N} = \frac{1}{\sqrt{N}} \to 0$ et

$$\frac{1}{N} \sum_{k \in U} (\frac{y_k}{\pi_k})^2 = \sum_{k \in U} y_k \to \infty$$

Les hypothèses du Théorème 7 ne sont pas vérifiées.

2.3 Estimation du nombre d'individus dans un domaine

Si l'on considère le problème d'estimation du nombre d'individus dans un domaine par un plan de sondage déterminantal à probabilités d'inclusion simple constantes, on connaît également les lois à distance finies. Soit donc la population $U = \{1, 2, \cdots, N\}$, un domaine $D \subseteq U$ et la variable d'intéret $y_k = 1_{k \in D}$. On s'intéresse à l'estimation de $\theta = \sharp D$ par un plan de sondage déterminantal équipondéré de matrice K fixée, de diagonale constante $K_{k,k} = \pi_k = c$. On a $\theta = \sum_{k=1}^{N} y_k$, et on considère donc naturellement l'estimateur d'Horvitz-Thomson $\hat{\theta} = c^{-1} \sum_{k \in \mathbb{S}} y_k = c^{-1} \sharp(D \cap \mathbb{S})$. Cette écriture permet de spécifier entièrement la loi de l'estimateur grâce au Théorème 1 :

Théorème 11. *La variable aléatoire $c\hat{\theta}$ suit une loi Poisson-binomiale, somme de $\sharp D$ variables aléatoires de Bernoulli indépendantes de paramètre λ_i, valeurs propres de la matrice $K_{|D}$.*

Démonstration La restriction du processus déterminantal $\mathbb{S}$ au domaine D est un processus déterminantal de matrice $K_{|D}$. D'après le Théorème 1, le nombre de points de ce processus sur une loi Poisson-binomiale, somme de $\sharp D$ variables aléatoires de Bernoulli indépendantes de paramètre λ_i valeurs propres de la matrice $K_{|D}$. □

Le théorème de Lindeberg-Feller pour les systèmes trianguaires ("triangular arrays") appliqué à des sommes de Bernouillis indépendantes (mais de paramètres non

nécéssairement égaux) permet alors de retrouver le théorème centrale limite précédent. Des théorèmes de type Berry-Essen permettent également d'obtenir, en plus de la normalité asymptotique, une vitesse de convergence. Enfin, l'inégalité de Bernstein permet d'obtenir une inégalité de deviation.

3 Plans de sondages déterminantaux optimaux

3.1 Stratégie représentative

Hàjek [19] définit une stratégie d'échantillonnage et d'estimation comme un couple $(\mathcal{P}, w)$ où $\mathcal{P}$ est un plan de sondage et w un système de pondérations. La stratégie est qualifiée de *représentative*, pour un ensemble de Q variables auxiliaires, si l'estimateur linéaire homogène associé conduit à une estimation parfaite du total de chaque variable : $EQM(\hat{t}_{x^q w}) = 0 \ \forall 1 \leq q \leq Q$.

En pratique, les deux composantes d'une stratégie ne sont pas fixées simultanément. Pour un jeu de probabilités d'inclusion fixé, Deville et Tillé [15] définissent un algorithme de tiragé équilibré conduisant à une stratégie approximativement représentative quand les pondérations sont astreintes à être celles de l'estimateur d'Horvitz-Thompson : $EQM(\hat{t}_{x^q HT}) \approx 0 \ \forall 1 \leq q \leq Q, \ \pi_k = \Pi_k$.

Pour un plan de sondage donné, Deville et Särndal [16] cherchent un jeu de pondérations w, dépendant de $\mathbb{S}$ conduisant à une stratégie représentative appelée calage. Pour atténuer, l'effet du sur-apprentissage du calage, Guggemos et Tillé [17] proposent un calage pénalisé plus robuste lors de l'estimation du total d'une variable d'intérêt y.

Avec un plan de sondage déterminantal, il est possible de coordonner la recherche des deux composantes d'une stratégie représentative en fixant en tant que solution d'un problème général du type :

$$\underset{(K,w)\in\Theta}{\mathcal{M}in} g(X, K, w) + \lambda pen(K, w) \tag{33}$$

où g est telle que $g(X, K, w) = 0$ si et seulement si $(\mathcal{P}_k, w)$ est représentatif pour X. pen est un critère de pénalisation qui permet d'éviter le surapprentissage. De manière classique, il permet également de transformer un problème d'optimisation sous contraintes en un problème non contraint. λ est un paramètre d'ajustement. Les paramètres inconnus sont K et w.

1. Exemple de fonctions g
   - Une fonction simple permettant de caractériser la représentativité est :

$$g(X, K, w) = \sum_{q=1}^{Q} EQM(\hat{t}_{x^q w})$$

   - Un arbitrage biais-variance et une importance différente accordée aux variables auxiliaires x_q peut être prise en compte au travers de la fonction suivante :

$$g(X, K, w) = \sum_{q=1}^{q=Q} \alpha_q \frac{\sqrt{V(\hat{t}_{x^q w})}}{t_{x^q}} + \beta_q \frac{|biais(\hat{t}_{x^q w})|}{t_{x^q}} \tag{34}$$

Les coefficients α_q et β_q sont fixés par le sondeur.

2. Exemple de Θ
   - Pour une contrainte d'échantillon de taille fixe n, on prendra pour Θ l'ensemble des matrices de projection de rang n et $w \in \mathbb{R}_+^N.$; si on souhaite seulement fixé la taille moyenne de l'échantillon on se limitera aux matrices contractantes K de trace fixée.
   - Si on s'intéresse uniquement à l'estimateur d'Horvitz-Thompson, on prendra $\Theta = \{(K,w)|spec(K) \in [0,1]^N, w_k = \frac{1}{K_{kk}}, \forall 1 \leq k \leq N\}$
   - Soit $\Pi \in [0,1]^N$ un vecteur fixé. Alors $\Theta = \{(K,w)|spec(K) \in [0,1]^N, K_{kk} = \Pi_k, \forall 1 \leq k \leq N\}$ est l'ensemble des plans déterminantaux de probabilités d'inclusion fixées.
3. Exemple de fonctions pen
   - $pen(K,w) = tr(K)^d$ permet de pénaliser les plans de grande taille ;
   - $pen(K,w) = \sum_k K_{kk}(w_k - \frac{1}{K_{kk}}))^2$, relaxation convexe de la contrainte $\Theta = \{(K,w)|w_k = \frac{1}{K_{kk}}, \forall 1 \leq k \leq N\}$, permet de limiter l'effet de surapprentissage lors de la recherche de la représentativité. Par analogie avec la méthode classique de calage *a posteriori*, d'autres distances ou pseudo-distances usuelles dans les peuvent être utilisées [15].
   - Soit C une matrice de contiguïté telle $c_{kl} = 1$ si k et l sont contigus et 0 autrement alors la fonction $pen(K,w) = \sum_k \sum_{l>k} c_{kl}(K_{kk}K_{ll} - |K_{kl}|^2)$ pénalise les plans de sondage en fonction du nombre moyen de voisins dans l'échantillon aléatoire.

Ces problèmes de minimisation seront en général bien posés car l'ensemble des matrices contractantes est un convexe compact. Cependant, le problème d'optimisation n'est pas convexe car la fonction objectif ne l'est pas en général, ce qui peut poser des difficultés algorithmiques.

3.2 Un cas particulier : minimsation de l'Erreur Quadratique Moyenne à probabilités d'inclusion fixés

Soit $Y = (y_1, \ldots, y_N)$ un vecteur strictement positif fixé, et Π un vecteur de probabilités d'inclusion simples strictement positives fixé. Le problème de minimisation de l'Erreur Quadratique Moyenne de l'estimateur linéaire homogène, pour les plans de sondages déterminantaux, revient à minimiser ce critère parmi les matrices de contraction de diagonale fixée Π. Il s'écrit de manière analytique :

$$(P) = \underset{0 \leq K \leq I, diag(K) = \Pi}{\arg\min} y' \delta_w ((I-K) * \overline{K}) \delta_w y + [e'(\delta_w (I * K) - I_N) y]^2$$

et puisque le biais $e'(\delta_w(I*K) - I_N)y$ est constant,

$$(P) = \underset{0 \leq K \leq I, diag(K) = \Pi}{\arg\min} y' \delta_w ((I-K) * \overline{K}) \delta_w y$$

Ce problème s'interprète de façon géométrique. Soit S_N^+ le cône des matrices semidéfinies positives. L'ensemble des matrices contractantes est $S_n^+ \cap (I - S_n^+)$ (et

on peut donc considérer $\frac{1}{2}I$ comme le "centre" de cet ensemble). Soit également D_Π l'ensemble des matrices de diagonale Π. Alors le domaine sur lequel on optimise est l'ensemble

$$C = S_n^+ \cap (I - S_n^+) \cap D_\Pi = \{0 \leq K \leq I, diag(K) = \Pi\}.$$

C est un ensemble convexe fermé comme intersection de convexes fermés, et comme il est borné et non vide (d'après le Théorème 4), c'est un compact et le problème d'optimisation à donc toujours (au moins) une solution (Théorème de Weierstrass pour les fonctions continues dans $\mathbb{R}^p$). Le domaine C est plus précisement un spectrahèdre projeté, c'est à dire la projection de l'intersection du cône des matrices semi-définies positives avec un espace affine. En effet, l'inégalité $0 \leq X \leq I$ est équivalente à $\left(\begin{pmatrix} X & 0 \\ 0 & I-X \end{pmatrix}\right) \geq 0$. Le problème d'optimisation est donc un problème d'optimisation semi-définie (non-linéaire). Ce type d'optimisation s'est fortement développé (dans le cadre linéaire) à partir des années 1990 ([1], [35] par exemple), mais présente déjà dans ce cadre linéaire de grandes difficultés.

En interprétant ce produit scalaire simplement en fonction de la norme associée $||.||$, on obtient :

Proposition 3.1. *Le problème (P) est équivalent aux problèmes suivants :*

1. $\arg\max_{X \in C} \langle\langle I - X, -X \rangle\rangle$*, soit "le plus grand angle (aigu) possible en $X \in C$ du triangle $0XI$" ;*
2. $\arg\max_{X \in C} ||I - X||^2 + ||X||^2$*, soit "le plus long chemin de 0 à I passant par C" ;*
3. $\arg\max_{X \in C} ||X||^2$ *soit "le plus grand vecteur de C" ;*
4. $\arg\max_{X \in C} ||X - \frac{1}{2}I||^2$ *;*
5. $\arg\max_{X \in C} ||I - X||^2$.

Démonstration il suffit de développer le produit scalaire et de remarquer que $\langle\langle I, X \rangle\rangle$ est constant dans C. □

On observe que la fonction objectif du problème de minimisation est concave en X. Elle atteint donc son miminum en un point extremal du convexe C. Malheureusement, ces points extremaux n'admettent pas de caractérisation simple en général (en particulier, ce ne sont pas des projections). Pour illustrer ce point, considérons un cas particulier simple.

Soit σ le spectrahèdre des matrices semi-définies positives de diagonale $\frac{1}{N}$. A une homothétie près, il s'agit de l'elliptope $\mathcal{E}$ des matrices de correlation, étudiée dans de nombreux articles [**?**]. Soit X une matrice de σ. Alors sa trace vaut 1, et comme ses valeurs propres sont toutes positives, elles sont également toutes inférieures ou égales à 1. Le spectrahedre σ et le spectrahedre projeté C considéré précedemment coincident.

De manière suprenante, le problème d'optimisation (P) est très simple, alors que la structure de σ est complexe. En effet, on a (pour des matrices à coefficients réels) :

Théorème 12. *1. Pour tout entier k tel que $\binom{k+1}{2} \leq N$, il existe une matrice de rang k extrémale dans σ ;*

2. *Les sommets de σ (points extremaux où le cône normal est de rang N) sont les matrices de projection de σ, qui sont toutes de rang 1 ;*
3. *Le problème d'optimisation linéaire $\max_{X \in \sigma} \langle A, X \rangle$ est NP-complet et la solution n'est pas nécéssairement un sommet.*

Cependant, on a la proposition suivante :

Proposition 3.2. *Le problème (P) sur σ admet exactement l'ensemble des matrices de projection de σ comme ensemble solution.*

L'ensemble des solutions est donc l'ensemble des sommets de σ, sous-ensemble strict de l'ensemble des points extremaux.

Démonstration Soit X une matrice de C. On a $|X_{k,l}|^2 \leq \frac{1}{N^2}$. On a donc $\varphi(X) = \sum_{k \in U} (w_k y_k)^2 - \sum_{k,l \in U} w_k y_k w_l y_l |X_{kl}|^2 \geq \sum_{k \in U} (w_k y_k)^2 - \frac{1}{N^2} \sum_{k,l \in U} w_k y_k w_l y_l$, avec égalité uniquement si tous les coefficients hors diagonale sont de module $\frac{1}{N}$.
Soit P une projection de C. Comme sa trace vaut 1, elle est de rang 1 et s'écrit $P = b\bar{b}^T$. On a $P_{k,k} = |b_k|^2 = \frac{1}{N}$ pour tout k, et $|P_{k,l}| = |b_k \overline{b_l}| = |b_k|.|\overline{b_l}| = \frac{1}{N}$. P est donc solution de (P).
Réciproquement, soit X solution de (P). Alors ses coefficients hors diagonale sont de module $\frac{1}{N}$, et les probabilités d'inclusion double $\pi_{k,l}$ sont donc nulles pour tout k, l. Le plan de sondage est donc de taille fixe (égale à 1), et X est une projection. □

Plus généralement, cette preuve s'adapte directement au cas des matrices dont la somme des coefficients diagonaux vaut 1. La conclusion est également la même pour un vecteur y constant, quelque soit le vecteur Π.

Une question ouverte est alors de savoir si l'une de ces conclusions reste valide dans le cas de spectrahèdres plus complexes. L'ensemble des solutions est-il inclus dans l'ensemble des projections ? des sommets ? Contient-il toujours une projection ? un sommet ? Si c'était le cas, cette conclusion s'obtiendrait également pour le problème (P) à diagonale libre. En pratique, les simulations numériques des problèmes d'optimisations précédents (pour des Π de somme entière) ont toujours conduit des projections.

4 Application

4.1 Principe

Nous paramétrons les matrices contractantes de rang $p \geq n$ sous la forme $K = \lambda_{max}^{-1} V V^T$ où n est la taille moyenne de l'échantillon, V est une matrice de taille (N, p) et λ_{max} la plus grande valeur propre de VV^T. Le nombre de paramètres complexes, de l'ordre de $2Np$, peut être très grand si la taille de la population est importante. Cette situation associée aux difficultés attendues dans la recherche d'un plan

optimal conduit à ce que nous nous limitons à des populations de petite taille N et, dans un premier temps, à des matrices réelles.

La sélection d'échantillon dans des populations de petite taille se rencontre en pratique. L'INSEE sélectionne les logements à enquêter pour ses enquêtes auprès des ménages selon le principe d'un échantillon maître à deux degrés, que nous présentons ici de manière simplifiée. Au premier degré sont sélectionnés des unités primaires (UP) constitués sur la base de regroupements d'entités géographiques contigües. Elles sont selectionnées pour une période donnée. Dans l'échantillon d'UP et sur la période considéreé sont ensuite sélectionnés pour chaque enquête des unités secondaires (US) correspondant aux logements, et dont tous les occupants sont interviewés.

L'échantillon d'UP est stratifié par région et degré d'urbanisation. Dans la région Rhône-Alples, la strate des agglomérations de 20000 à 100000 habitants contient ainsi 19 UP. Le calibrage de l'ensemble conduit à vouloir sélectionner 9 UP. Les probabilités d'inclusion sont calculées de manière à être proportionnelle au nombre de logements. On note Π_k^1 la probabilité que l'UP k soit sélectionnée. Les UP primaires sont selectionnées selon la méthode du cube de Deville Tillé et telle que programmée par Chauvet. Les variables d'équilibrage sont : le nombre d'habitants, les nombres de logements HLM, individuels (maison), occupés par leur propriétaire. On note X l'ensemble de ces variables d'équilibrage. Afin de disposer d'un échantillon de taille fixe l'équilibrage s'effectue également sur la variable donnant les probabilités d'inclusion.

Les résultats de cette stratégie sont inclus, pour comparaison, dans ceux obtenus par différentes méthodes suivantes

1. Sondage Aléatoire Simple,
2. Plan déterminantal de taille fixe et de probabilités d'inclusion constantes,
3. Méthode de Sampford de probabilités Π_k^1,
4. Plan déterminantal de taille fixe et de probabilités d'inclusion $\pi_k = \Pi_k^1$,
5. Méthode de Deville Tillé de probabilités constantes et équilibré sur X,
6. Plan déterminantal représentatif des totaux de X d'après le critère (34) avec $\alpha_q = 1$ sous contraintes de probabilités constantes et de poids correspondant à ceux d'Horvitz-Thompson,
7. Méthode de Deville Tillé de probabilités Π_k^1 et équilibré sur X,
8. Plan déterminantal représentatif des totaux de X d'après le critère (34) avec $\alpha_q = 1$ sous contraintes $\pi_k = \Pi_k^1$ et de poids correspondant à ceux d'Horvitz-Thompson,
9. Plan déterminantal représentatif des totaux de X d'après le critère (34) avec $\alpha_q = 1$ sans contraintes sur les probabilités d'inclusion et de poids correspondant à ceux d'Horvitz-Thompson.

4.2 Résultats

Le tableau 2 donne pour chacune des 9 stratégies la valeur, à l'optimum le cas échéant, de la fonction :

$$g(X) = \sum_{q=1}^{q=Q} \alpha_q \frac{\sqrt{V(\hat{t}_{x^q HT})}}{t_{x^q}} \tag{35}$$

Pour le sondage aléatoire simple, et les plans déterminataux, la précision a été obtenue directement la formule (2) de l'EQM. Pour les plans de Sampford et selon la méthode de Deville-Tillé, la précision a été obtenue par simulation.

TABLE 2 – Précision moyenne des estimations selon la stratégie

Méthode		π_k	Equilibrage	$g(X)$
1	SAS	$\frac{n}{N}$		0.6013106
2	Déterminant	$\frac{n}{N}$	π	0.5839211
3	Deville-Tillé	$\frac{n}{N}$	X,π	0.2667057
4	Déterminant	$\frac{n}{N}$	X	0.2523981
5	Sampford	Π_k^1		0.2241620
6	Déterminant	Π_k^1	π	0.2240632
7	Deville-Tillé	Π_k^1	X,π	0.1156012
8	Déterminant	Π_k^1	X	0.1130944
9	Déterminant	opt	X	0.1071314

On constate que dans tous les cas le plan déterminantal a de meilleures performances que son équivalent non déterminantal, même si parfois le gain est très faible. L'intérêt de la stratégie déterminantal est qu'elle autorise le calcul des probabilités d'inclusion optimales. A la différence de l'équilibrage à la Deville-Tillé, l'approche déterminantal fournit par ailleurs, de manière exacte, les probabilités d'inclusion doubles.

Quand les probabilités d'inclusion simples sont constantes et que la population des unités primaires est indicée selon le nombre croissant de résidences principales, les probabilités d'inclusion doubles sont d'autant plus faibles que l'on s'approche de la diagonale (graphique 1.) Pour deux UP ayant des tailles dissemblables, le plan déterminantal équilibré conduit à les sélectionner de manière quasi indépendante.

Quand les probabilités d'inclusion simples sont proportionnelles au nombre de résidences principales, le constat précédent reste globalement valable. Il est accentué pour les unités primaires 7,8 et 9 qui ont des probabilités d'inclusion double nulles.

FIGURE 1 – Stratégie 4 : Disques proportionnels à π_k sur la diagonale (noir), à $\frac{\pi_{kl}}{\pi_k \pi_l}$ autrement : bleu $0.95 \leq \frac{\pi_{kl}}{\pi_k \pi_l} \leq 1$, rouge $0 \leq \frac{\pi_{kl}}{\pi_k \pi_l} \leq 0.95$. UP triées par taille.

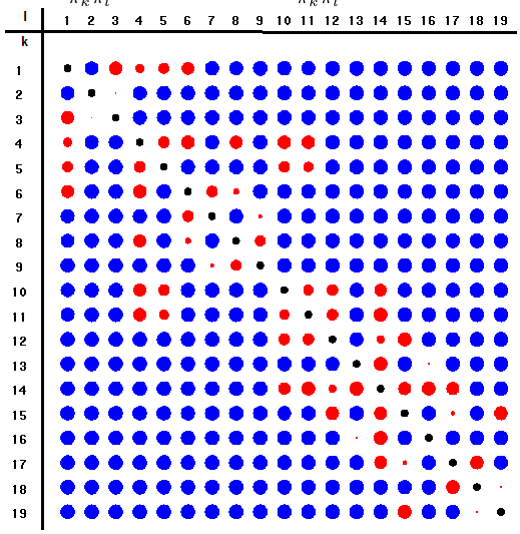
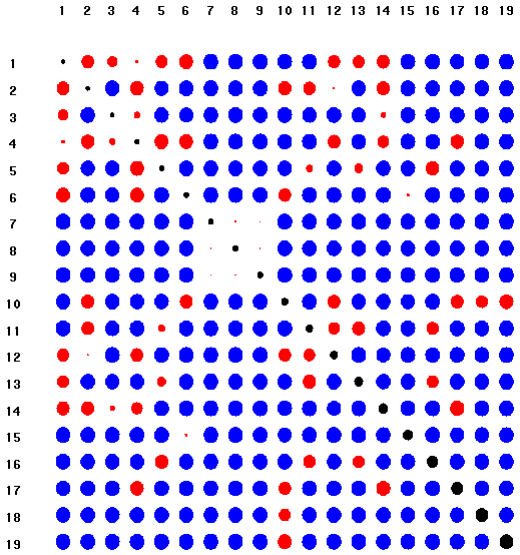


FIGURE 2 – Stratégie 8 : Disques proportionnels à π_k sur la diagonale (noir), à $\frac{\pi_{kl}}{\pi_k \pi_l}$ autrement : bleu $0.95 \leq \frac{\pi_{kl}}{\pi_k \pi_l} \leq 1$, rouge $0 \leq \frac{\pi_{kl}}{\pi_k \pi_l} \leq 0.95$. UP triées par taille.



Appendix 1

Démonstration Soit $n \notin \{0, 1, N, N-1\}$, et supposons que le plan de sondage aléatoire simple soit déterminantal. Alors sa restriction à un domaine l'est aussi d'après

la Proposition 1.3, et d'après le Théorème 1, la loi du nombre de point ν tombant dans un domaine D donné est une loi de Poisson-binomiale somme de $\sharp D$ loi binomiales. Prenons comme domaine un domaine à deux éléments, et notons p_1 et p_2 les coefficients des lois binomiales associées. Alors $\mathbb{P}(\nu = 0) = \frac{\left(\begin{array}{c} N-2 \\ n \end{array} \right)}{\left(\begin{array}{c} N \\ n \end{array} \right)}$, $\mathbb{P}(\nu = 1) = 2\frac{\left(\begin{array}{c} N-2 \\ n-1 \end{array} \right)}{\left(\begin{array}{c} N \\ n \end{array} \right)}$ et $\mathbb{P}(\nu = 2) = \frac{\left(\begin{array}{c} N-2 \\ n-2 \end{array} \right)}{\left(\begin{array}{c} N \\ n \end{array} \right)}$. L'espérance du nombre de point est alors égale à $2\frac{\left(\begin{array}{c} N-1 \\ n-1 \end{array} \right)}{\left(\begin{array}{c} N \\ n \end{array} \right)}$. On vérifie alors que le système

$$\left\{ \begin{array}{rcl} (1-p_1)(1-p_2) & = & \frac{\left(\begin{array}{c} N-2 \\ n \end{array} \right)}{\left(\begin{array}{c} N \\ n \end{array} \right)} \\ p_1(1-p_2)+(1-p_1)p_2 & = & 2\frac{\left(\begin{array}{c} N-2 \\ n-1 \end{array} \right)}{\left(\begin{array}{c} N \\ n \end{array} \right)} \\ p_1p_2 & = & \frac{\left(\begin{array}{c} N-2 \\ n-2 \end{array} \right)}{\left(\begin{array}{c} N \\ n \end{array} \right)} \end{array} \right.$$

n'admet pas de solution dans $[0,1]^2$. En effet, p_1 et p_2 sont nécessairement les solutions de l'équation du second degré $\left(\begin{array}{c} N \\ n \end{array} \right) X^2 - 2\left(\begin{array}{c} N-1 \\ n-1 \end{array} \right) X + \left(\begin{array}{c} N-2 \\ n-2 \end{array} \right) = 0$, soit $n(N-1)X^2 - 2N(N-1)X + n(n-1) = 0$, et l'une des solutions est plus grande que $\frac{N}{n} > 1$. □

Références

[1] , G. Blekherman, P. A. Parrilo, R. R. Thomas, Semidefinite Optimization and Convex Algebraic Geometry SIAM 2013

[2] Basu, D. and Ghosh, J. K. (1967) : Sufficient statistics in sampling from a finite universe, Bull. Int. Stat. Inst. 42, BK. 2, 850–859.

[3] Basu, D. (1969) : Role of sufficiency and likelihood principles in sample survey theory, Sankhya A 31, 441–454.

[4] Y.G. Berger, Rate of convergence for asymptotic variance of the Horvitz-Thompson estimator. J. Statist. Plann. Inference 74 (1998), no. 1, 149–168.

[5] A. Borodin, Determinantal point processes. The Oxford handbook of random matrix theory, 231–249, Oxford Univ. Press, Oxford, 2011.

[6] J. B. Borcea, P. Brändèn and T. M. Liggett, Negative dependance and the geometry of polynomials. J. Amer. Math. Soc. 22 (2009), no. 2, 521–567.

[7] P. Brändèn ; J. Jonasson, Negative dependence in sampling. Scand. J. Stat. 39 (2012), no. 4, 830–838.

[8] P.G. Casazza, M. Fickus, J. Kovacevic, M.T. Leon, J.C. Tremain, A Physical Interpretation for Finite Tight Frames

[9] H. Cardot, M. Chaouch, C. Goga, C. Labruère, Properties of design-based functional principal components analysis. J. Statist. Plann. Inference 140 (2010), no. 1, 75–91.

[10] Chen, X-H, Dempster, A. P. and Liu, J. S. (1994). Weighted finite population sampling to maximize entropy. Biometrika 81(3) 457–469

[11] P.G. Casazza, D. Redmond, J.C. Tremain, Real Equiangular Frames, CISS Meeting, Princeton, NJ (2008).

[12] G. Chauvet, D. Bonnery, J-C Deville., Optimal inclusion probabilities for balanced sampling, Journal of Statistical Planning and Inference, 2011.

[13] G. Chauvet, A note on the consistency of the Narain-Horvitz-Thompson-Estimator arXiv :1412.2887v1

[14] I.S. Dhillon, R.W. Heath Jr, M.A. Sustik, J.A. Tropp, Generalized finite algorithms for constructing Hermitian matrices with prescribed diagonal and spectrum, SIAM J. Matrix Anal. Appl. 27 (2005) 61–71.

[15] J-C. Deville, Y. Tillé, Efficient balanced sampling : The cube method,

[16] J-C. Deville, C-E. Sarndal, Calibration Estimator in Survey Sampling, Journal of the American Statiscal Association, 1992.

[17] F. Guggemos, Y. Tillé, Penalized Calibration in survey sampling, Journal of Statistical Planning and Inference, 2010.

[18] D. G. Horvitz, D. J. Thompson, A generalization of sampling without replacement from a finite universe, Journal of the American Statistical Association, 1952.

[19] J. Hajek, (1981). Sampling from a finite population. Marcek Dekker.

[20] J. B. Ben Hough, M. Krishnapur, Y ; Peres, B. Virag, Determinantal processes and independence. Probability Surveys, 3 (2006) 206–229.

[21] J. B. Ben Hough, M. Krishnapur, Y ; Peres, B. Virag, Zeros of Gaussian analytic functions and determinantal point processes. University Lecture Series, 51. American Mathematical Society, Providence, RI, 2009.

[22] A. Horn, Doubly stochastic matrices and the diagonal of a rotation matrix, American Journal of Mathematics 76 (1954), 620–630.

[23] Isaki, C.T. and Fuller, W.A. (1982). Survey design under the regression superpopulation model. J. Am. Statist. Ass. 77, 89–96.

[24] R.V. Kadison, The Pythagorean theorem. I. The finite case. Proc. Natl. Acad. Sci. USA 99 (2002), no. 7, 4178–4184.

[25] Lavancier F., Moller J., Rubak E. Determinantal point process models and statistical inference. To appear in Journal of the Royal Statistical Society, series B.

[26] R. Lyons. Determinantal probability measures. Publ. Math. Inst. Hautes Etudes Sci. 98 (2003), 167–212.

[27] O. Macchi, The coincidence approach to stochastic point processes. Adv. Appl. Probab., 7 (1975), 83–122.

[28] R. Pemantle, Y. Peres, Concentration of Lipschitz functionals of determinantal and other strong Rayleigh measures. Combin. Probab. Comput. 23 (2014), no. 1, 140–160.

[29] A. Soshnikov, Determinantal random point fields. Russian Math. Surveys, 55, (2000), 923–975.

[30] A. Soshnikov, Gaussian limit for determinantal random point fields. Ann. Probab. 30 (2002), no. 1, 171–187.

[31] A. Scardicchio, C.E Zachary, S. Torquato (2009). Statistical properties of determinantal point processes in high dimenseional Euclidian spaces. Physical Review, 79.

[32] M.A. Sustik, J.A. Tropp, I.S. Dhillon, R.W. Heath Jr. On the existence of equiangular tight frames. Linear Algebra Appl. 426 (2007), no. 2–3, 619–635.

[33] Y. Tillé. Théorie des sondages. Dunod.

[34] J.A. Tropp, Complex equiangular tight frames, in : Proceedings of SPIE 2005 (Wavelets XI), San Diego, August 2005.

[35] L. Vandenberghe, S. Boyd, Semidefinite Programming, SIAM Review 38, March 1996, pp. 49–95.

[36] Fan, C. T., Muller, Mervin E., and Rezucha, Ivan , Development of Sampling Plans by Using Sequential (Item by Item) Selection Techniques and Digital Computers, Journal of the American Statistical Association, 57 , 387–402.

[37] R. Narain, On sampling without replacement with varying probabilities. Journal of the Indian Society of Agricultural Statistics, 3 (1951) 169–175.

[38] M. Yuan, C. Su, T. Hu, A central limit theorem for random fields of negatively associated processes, J. Theoret. Probab. 16, (2003) 309–323.