PLANS DE SONDAGE A DISPERSION MINIMALE

Jean-Claude DEVILLE(*), Mohamed El HADJ TIRARI (*)

(*) Ensai/Crest,Laboratoire de Statistique d'Enquête

Introduction

On s'intéresse à l'échantillonnage dans une population finie *U* de taille *N*. Ses individus sont notés conventionnellement par un index *k* variant de *1* à *N*. Un plan de sondage est une loi de probabilité sur l'ensemble ***S*** des parties de *U* autrement dit une famille de nombres *p(s)* associés à chaque élément de ***S***- ou échantillon- vérifiant $p(s) \geq 0$ et $\sum_{s \in \mathbf{S}} p(s) = 1$. Le support *supp(p)* est l'ensemble des s tels que *p(s)>0* , c'est-à-dire l'ensemble des échantillons que l'on accepte a priori de sélectionner. Dans ce travail, nous adopterons la contrainte habituelle, motivée par le coût du recueil des données, qui consiste à restreindre le support aux échantillons ayant une taille *n* fixée avec *n<N*. Formellement on impose donc $\mathrm{supp}(s) \subset \mathbf{S_n}$ ensemble des échantillons de taille *n*. On impose par ailleurs de respecter des probabilités d'inclusion $\pi_k = \sum_{s:k \in \mathbf{s}} p(s)$, données pour chaque unité *k* de *U.*

Il est bien connu que la somme des probabilités d'inclusion d'un plan de taille fixée vaut *n*, autrement dit le *N*-vecteur des π_k appartient au simplexe $\sum_{k \in U} \pi_k = n$ de $\boldsymbol{R}^N$, bord exclu si on se limite aux probabilités différentes de *0* ou *1*.

Ce problème est aussi vieux que la théorie des sondages et est l'objet de trop nombreux développement dans la littérature pour qu'on puisse envisager de donner des références exhaustives. On pourra consulter l'ancienne synthèse de Brewer et Hanif[1] et surtout le 'nouveau testament' de Tillé [16]. La façon dont nous approchons la question consiste à formuler un critère de choix parmi tous les plans vérifiant les contraintes de support et de probabilités d'inclusion. Intuitivement, en l'absence d'autres informations se traduisant par des contraintes, on cherche à n'accorder aucun privilège à quelque échantillon que ce soit. En particulier, si cela est possible, c'est-à-dire quand toutes les π_k sont égales (et donc égales à *n/N*), on rendra tous les échantillons équiprobables, et donc on choisira le sondage aléatoire simple. Si les probabilités d'inclusion sont inégales, ce choix est évidemment impossible car l'échantillon comportant les *n* unités les plus probables devra recevoir une

probabilité supérieure à celui qui contient les *n* plus 'petites'. On peut même souhaiter voir ce dernier recevoir une probabilité nulle de façon à éviter complètement l'éventualité d'un échantillon trop mal balancé.

Un critère de choix naturel consistera à minimiser la dispersion des nombres *p(s)* compte tenu des contraintes, en un sens que nous allons préciser. Cette approche a déjà été utilisée par Joe[9] puis Chen [4,5] ainsi que par Deville[7] ou Berger[2] essentiellement dans le but de justifier un critère du type de l'entropie. Les deux premiers auteurs situent leur démarche dans le cadre d'un modèle où les probabilités d'inclusion (vues comme des lois marginales) sont connues ou estimées et où on désire estimer la probabilité des combinaisons de *n* unités : typiquement la probabilité de gagner au loto avec un pari sur six nombres parmi quarante-huit. Sans les examiner en détail, Joe évoque des critères liés aux entropies de Renyi[14,15] analogues aux distances de Cressie-Read [6] utilisées parfois par les économètres.

Dans la première partie de cet article nous examinerons une formulation assez générale de la notion de dispersion et du problème de sa minimisation sous contraintes. Les cas typiques sont les critères d'entropie ou de variance. L'optimisation conduit pour le premier critère à une solution où tous les *p(s)* admissibles sont strictement positifs alors que c'est généralement le contraire pour le second. Dans la section 2, on caractérisera les critères de dispersion selon cette propriété et on montrera que l'algorithme de Metropolis ou ses améliorations [3, 13] peuvent permettre un échantillonnage effectif selon les schémas optimaux. Dans la section 3 nous nous intéresserons plus spécifiquement au critère de variance et nous établirons ses liens avec le schéma de Midzuno-Lahiri [11,12 par exemple]. La section 4 exhibera un algorithme efficace de calcul des *p(s)* et la section 5 une méthode générale d'échantillonnage pour le plan de variance minimum qui étend de façon surprenante la méthode de Midzuno. Les sections 6 et 7 étendent ces résultats au cas d'un critère général de dispersion, en particulier en ce qui concerne l'utilisation du schéma de Midzuno. Enfin la section 8 est consacrée à des illustrations des résultats établis auparavant.

1. Plans à dispersion minimale

La notion de dispersion peut être axiomatisée de façons très diverses. Nous utiliserons ici un type de critère assez simple mais relativement naturel et répandu dans la pratique.

On se base sur une fonction φ strictement convexe (pour tout couple $x<y$ on a $\varphi(\frac{x+y}{2}) < \frac{\varphi(x)+\varphi(y)}{2}$) définie sur *[0,1]* qu'on supposera, pour éviter des complications inutiles, continument dérivable, la dérivée en *0* (resp *1*) pouvant valoir $-\infty$ (resp $+\infty$) . Le critère

$$\Phi(p) = \sum_{s\in \mathbf{S}_n} \varphi(p(s)) \qquad (1)$$

est une fonctionnelle sur l'ensemble des plans de taille *n*. Sa valeur minimum est obtenue pour l'échantillonnage aléatoire simple (SAS), c'est-à-dire pour la répartition uniforme p_u sur $\boldsymbol{S_n}$. C'est une mesure de dispersion au sens où, si *p* n'est pas uniforme, les lois $p_t=p_u + t(p-p_u)$ sont, intuitivement, de plus en plus dispersées quand $|t|$ croît et où $\Phi(p_t)$ est une fonction strictement convexe qui prend son minimum pour $t=0$. Ce critère reste le même à une constante additive près si on ajoute à φ une fonction affine arbitraire $\alpha p + \beta$. On peut donc choisir α et β de façon à ce que pour tout *s* on ait $\phi(p_u(s)) = 0$ et $\phi'(p_u(s)) = 0$. Avec cette normalisation $\Phi(p)$ peut se voir comme une 'distance' au Sondage Aléatoire Simple.

Les contraintes sur les probabilités d'inclusion se formalisent sous forme linéaire. Par abus de notation, notons aussi par *p* le $\boldsymbol{S}_n$ –vecteur des *p(s)* ; soit $\boldsymbol{U}$ la $N\text{x}\boldsymbol{S}_n$ matrice des $\boldsymbol{1}_k(s)$ (1 si $k \in s$ et *0* sinon), et π le *N*-vecteur des probabilités d'inclusion. On écrira donc les contraintes sous la forme compacte :

$$\boldsymbol{U}p=\pi \ . \quad (2)$$

Notons enfin $\boldsymbol{1}_N$ et $\boldsymbol{1}_{\boldsymbol{S}}$ les vecteurs composés uniquement de *1* de tailles respectives *N* et *card*($\boldsymbol{S}_n$) . On a :

$$\boldsymbol{1'}_N \boldsymbol{U} = n\, \boldsymbol{1'}_{\boldsymbol{S}} \quad \text{et } \boldsymbol{U}\, \boldsymbol{1}_{\boldsymbol{S}} = \binom{N}{n} \boldsymbol{1}_N$$

Il en résulte, en particulier, qu'on doit avoir :

$$\sum_U \pi_k = \mathbf{1'}_N \pi = \mathbf{1'}_N \mathbf{U}\ p = n\, \mathbf{1'}_S\ p = n \ .$$

Les contraintes sur les probabilités d'inclusion impliquent que la somme des *p(s)* vaut 1 . Il est donc inutile d'ajouter cette contrainte pour la résolution du problème de minimisation de (1) sous les contraintes (2) et de non-négativité des *p(s)* .

On appellera cela le problème (P). Voici quelques unes de ses propriétés :

Propriété 1 : Le problème (P) possède toujours une solution unique.

En effet *p* varie dans un domaine compact et convexe et on minimise une fonction strictement convexe sous des contraintes linéaires.

La propriété suivante caractérise les solutions de (P) selon que tous les échantillons de $\boldsymbol{S}_n$ reçoivent une probabilité strictement positive (toutes les solutions sont 'intérieures') ou qu'il existe des vecteurs π menant à des solutions 'au bord', c'est à dire comportant des échantillons à probabilité nulle.

Propriété 2 : Une condition nécessaire et suffisante pour qu'il n'y ait jamais de solution 'au bord' est que $\varphi'(0) = -\infty$.

La démonstration fait l'objet de l'annexe 1.

Corollaire : Si *n>1*, cette condition implique que toutes les probabilités d'ordre deux $\pi_{kl} = \sum_{s:k\ et\ l\in s} p(s)$

sont strictement positives et qu'il existe un estimateur sans biais de la variance d'échantillonnage.

Le cas opposé évite des configurations désagréables comme la possibilité de tirer des échantillons contenant trop de petites unités.

Propriété 3 : Si $\varphi'(0) > -\infty$ il existe des systèmes de probabilités d'inclusion strictement positives tels que *p(s)=0* pour certains échantillons de $\boldsymbol{S}_n$.

Ce n'est autre que la contraposée de la proposition 2.

Si φ est prolongeable à [0, +∞] le critère peut se généraliser en une distance entre lois de probabilités par $d(p,q) = \sum_{s\in \mathbf{S}_n} \varphi(p(s)p_u / q(s))\ q(s)$.

La fonction φ la plus usuelle est l'entropie négative $\varphi_{ent}(p) = p\ log(p)$ (avec la convention *0 log(0)=0*) . Il est bien connu [5,7, 9…] que l'entropie d'une distribution de probabilité est maximum pour la distribution uniforme et que, sous les contraintes que nous utilisons, sa maximisation conduit à l'échantillonnage de Poisson conditionnel (ou réjectif).

D'autres critères sont aussi assez naturels, en particulier la variance pour laquelle on utilise $\varphi_{var}(p)=(p-p_u)^2$. On peut, par normalisation à une fonction affine près, utiliser simplement la fonction convexe $\varphi_{var}(p)=p^2$. Mais on peut aussi utiliser les fonctions de type 'puissance' $\varphi_a(p)= p^a$ si *a>1* ou

$\varphi_a(p)= - p^a$ *0 <a <1* .

La minimisation de certaines distances bien connues se ramène à cette famille comme la distance de Hellinger $d_H(p,q) = \sum_s (p(s)^{1/2} - q(s)^{1/2})^2$ qui conduit à utiliser $\varphi_{-1/2}$. Comme on le verra, la maximisation d'une des entropies (informations) de Renyi $H_\alpha(p) = \frac{1}{1-\alpha}\log(\sum_s p(s)^\alpha)$ se ramène aussi à cette famille.

Exemple 1 : Si Φ est l'entropie $-\sum p(s)\ log(p(s))$ tous les échantillons de taille *n* ont une probabilité strictement positive. Il en va de même pour le critère associé à la distance de Hellinger.

Exemple 2 : Si Φ est la variance $\sum p(s)^2$ on obtient facilement une solution 'au bord' . Par exemple, pour *N=4* et *n=2* , prenons $\pi_1 = \pi_2 = 0.2$, $\pi_3 = \pi_4 = 0.8$. On obtiens *p(1 ,2) =0 , p(1 , 3) = p(2 , 3) = p(1 ,4) = p(2 , 4) =0.1* et *p(3 ,4) = 0.6* .

Si on minimise l'entropie négative ces trois valeurs sont respectivement *0.0141 , 0.0930* et *0.61141* .

2. Propriétés des plans à dispersion minimale et échantillonnage de type Metropolis

La solution de (P) peut s'obtenir à l'aide de la technique de Kühn et Tucker [10]. En introduisant un vecteur de multiplicateur de Lagrange $\lambda = \{\lambda_k ; k \in U\}$ pour les contraintes d'égalité et des $\mu_s > 0$ pour les contraintes de positivité actives –soit *p(s)>0* -, ou égaux à zéro sinon, la solution vérifie $\varphi'(p) = U'\lambda + \mu$, en notant $\varphi(p)$ le $\mathbf{S}_n$ vecteur des $\varphi(p(s))$ et μ le *N*-vecteur des μ_s . Comme φ est strictement convexe et continûment dérivable, φ' est croissante strictement et continue. Elle admet une fonction réciproque ψ qui croît continûment de 0 à 1 quand son argument croît de $\varphi'(0)$ à $\varphi'(1)$. On a en particulier :

Propriété 4 : Si $\varphi'(0) = -\infty$, ψ est définie de $-\infty$ à *1* , toutes les contraintes sont inactives et $p(s) = \psi(\sum_s \lambda_k)$.

Si $\varphi'(0) > -\infty$, on adoptera désormais la normalisation $\varphi'(0) = 0$. La solution de (P) admet donc la forme générale $p(s) = \psi(\sum_{k \in s} \lambda_k)$ si *p(s)>0* et donc si $\sum_{k \in s} \lambda_k > 0$. Pour les *s* de probabilité nulle on aura $p(s) = \psi(\sum_{k \in s} \lambda_k + \mu_s)$ avec $\sum_{k \in s} \lambda_k < 0$, ce qui identifie $\mu_s = -\sum_{k \in s} \lambda_k$.

La difficulté numérique réside dans l'identification des contraintes actives. Les λ_k étant supposées connues, nous avons donc le résultat suivant :

Propriété 5 : Si $\varphi'(0)>-\infty$ la solution de (P) est caractérisée par un vecteur de *N* réels λ_k définis de façon unique en fonction des probabilités d'inclusion tels que :

$$p(s)=\psi(\sum_{k\in s}\lambda_k)>0 \text{ si } \sum_{k\in s}\lambda_k>0.$$

$$p(s)=0 \qquad \text{si } \sum_{k\in s}\lambda_k<0 .$$

Propriété 6 : Les λ_k sont ordonnés comme les π_k : $\pi_k<\pi_l$ si et seulement si $\lambda_k<\lambda_l$.

En effet $\pi_k=\sum_{s:k\in s}\psi(\lambda_k+\sum_{l\in s-k}\lambda_l)$ de sorte que $\pi_k-\pi_l=\sum_{s\in Skl}\psi(\lambda_k+\sum_{j\in s}\lambda_j)-\sum_{s\in Skl}\psi(\lambda_l+\sum_{j\in s}\lambda_j)$, où S_{kl} est l'ensemble des échantillons de taille *n-1* ne contenant ni *k* ni *l*. Comme ψ est strictement croissante $\lambda_k<\lambda_l$ implique que $\pi_k<\pi_l$. La réciproque résulte du caractère bijectif de la correspondance entre π et λ .

La forme simple des *p(s)* permet de calculer facilement le ratio de deux d'entre elles. On peut donc envisager d'échantillonner dans ce plan en utilisant l'algorithme de Metropolis [13] ou une de ses variantes [3] . Cette méthode consiste à réaliser sur l'ensemble des échantillons possibles une marche aléatoire markovienne réversible dont la loi ergodique est celle des *p(s)* . On peut procéder typiquement de la façon suivante.

Arrivé à l'échantillon s_t , un candidat s_{t+1} est obtenu, par exemple, en remplaçant une unité tirée au hasard dans s_t par une nouvelle unité tirée elle aussi à probabilités égales. Si $p(s_{t+1})>p(s_t)$ on va en s_{t+1} . Sinon, $p(s_{t+1})/p(s_t)=r$ est une probabilité ; on va en s_{t+1} avec la probabilité *r* et on reste à s_t avec la probabilité *1-r* . Malheureusement cette technique demande une quantité énorme d'itérations (de l'ordre de plusieurs fois le nombre d'échantillons possibles) et son utilisation s'avère peu efficace. Dans le cas du critère d'entropie on connaît maintenant des méthodes rapides et efficaces même quand les tailles des populations et des échantillons sont relativement grandes ([5],[7],[16]). Dans le cas général on propose au paragraphe 6 une méthode relativement satisfaisante.

Remarque : La maximisation des entropies de Renyi se ramène au problème (P) comme celle de tout critère de la forme $H(p)=F(\sum_{s\in\mathbf{S}_n}\varphi(p(s)))$ où *F* est une fonction régulière et assurant la

convexité du critère. Le calcul conduit en effet aux équations $\varphi'(p)\, F'(\sum_{s\in \mathbf{S}_n} \varphi(p(s))) = U'\lambda + \mu$ qui sont identiques à une constante près à celles de la résolution de (P).

3. Le plan à variance minimale et l'algorithme de Midzuno

Dans cette section on examine le cas où on minimise la variance des *p(s)*, ce qui équivaut à minimiser $\sum p(s)^2$. On peut prendre alors simplement $\psi(t)=t$. Examinons d'abord le cas où toutes les probabilités sont strictement positives. Les multiplicateurs vérifient donc, comme $\mu=0$ par hypothèse, $0<p(s)=\lambda' s= \sum_{k\in s} \lambda_k$, soit, matriciellement, $p=U'\lambda$. L'identification de λ résulte de $Up= UU'\lambda=\pi$ .Or

$UU'=(S_1 - S_2)\mathbf{I}_N + S_2 \mathbf{1}_N \mathbf{1}_N'$ où $S_1=\begin{pmatrix} N-1 \\ n-1 \end{pmatrix}$ et $S_2=\begin{pmatrix} N-2 \\ n-2 \end{pmatrix}$.

On trouve donc facilement que $\lambda=C^{te}(\pi -(n-1)/(N-1)\mathbf{1}_N)$.

Propriété 7 : Le plan à variance minimale charge tous les échantillons de taille *n* si et seulement si $\sum_{s\min} \pi_k > n(n-1)/(N-1)$ où *smin* est l'échantillon contenant les *n* plus petites unités. La probabilité de *s* est alors proportionnelle à $\sum_s (\pi_k - (n-1)/(N-1))$.

Ce plan est très proche de celui qu'implémente le schéma de Midzuno-Lahiri. Celui ci fonctionne, rappelons le, de la façon suivante : x_k est une variable positive ou nulle et on pose $p_k= x_k/X$ où *X* est le total des x_k .On tire une première unité avec la loi des p_k et on complète l'échantillon par sondage simple de *n-1* unités parmi les *N-1* restantes. On voit facilement que $p(s) = (\sum_s p_k) / \begin{pmatrix} N-1 \\ n-1 \end{pmatrix}$ et que les probabilités d'inclusion valent $\pi_k = \frac{n-1}{N-1} + p_k(1 - \frac{n-1}{N-1}) \geq \frac{n-1}{N-1}$. La propriété la plus amusante de ce plan est de rendre sans biais l'estimateur par ratio $X\, \bar{y} / \bar{x}$ du total d'une variable *y* quelconque.

Même lorsqu'il charge tout $\mathbf{S}_n$ le plan à variance minimum est un peu plus général que le schéma de Midzuno car ce dernier demande que chaque π_k soit supérieur à *(n-1)/(N-1)* .

4. Plan à variance minimale : résolution

Dans le cas de la variance, le problème (P) s'écrit :

Min p'p sous les contraintes $Up=\pi$ et $p \geq 0$.

C'est un problème banal de programmation quadratique pour lequel on connaît, depuis longtemps, des algorithmes. Ils demandent, cependant, une grosse capacité de mémoire et de calcul et ils s'avèrent inefficaces dès que *N* dépasse 15 à 20 et *n* 7 à 10. Il se trouve qu'on peut trouver une résolution bien plus rapide et économique, et, de plus ayant une interprétation statistique intéressante. Elle se base sur le résultat suivant :

Propriété 8 : Soit p_0 la solution de *Min p'p* sous les contraintes $U_0 p=\pi$ où U_0 est la restriction de *U* à un ensemble de colonnes $\mathbf{S}_0 \subset \mathbf{S}_n$. Soit $\mathbf{S}_{0+}$ l'ensemble des *s* vérifiant $p(s)>0$. Le plan p_* optimal obtenu en ajoutant les contraintes $p \geq 0$ vérifie $p_*(s) > 0$ sur un ensemble $\mathbf{S}_{*+}$ contenu $\mathbf{S}_{0+}$.

La démonstration est donnée en annexe 2.

Il en résulte l'algorithme suivant :

-Soit p_1 la solution sans contraintes de positivité $p_1 = U'(UU')^{-1}\pi = U'\lambda$, $\mathbf{S}_1$ l'ensemble des *s* tels que $p_1(s)>0$ et U_1 la restriction de *U* à cet ensemble. D'après la propriété 7 le support de p_* est contenu dans $\mathbf{S}_1$.

-On cherche p_2 à support dans $\mathbf{S}_1$ optimal sans contrainte de positivité, soit :

$$\begin{aligned} p_2 &= U_1'(U_1U_1')^{-1}\pi = U_1'\lambda_2 \quad \text{si } s \in \mathbf{S_1} \\ &= 0 \qquad\qquad\qquad\qquad\quad \text{si } s \in \mathbf{S}_n - \mathbf{S_1} \end{aligned}$$

-Si $supp(p_2)=supp(p_1)$ alors $p_* = p_1$ et fin.

-Sinon $supp(p_2)$ est inclus strictement dans $supp(p_1)$ et permet de remplacer U_1 par une matrice dont le nombre de colonnes est inférieur et de retourner à l'étape 2 jusqu'à convergence.

Cet algorithme semble voisin de celui utilisé par Joe [9].

5. Echantillonnage pour le plan à variance minimale

Les conditions de Midzuno ($\pi_k \geq (n-1)/(N-1)$) correspondent au cas où le support du plan de variance minimale est $\boldsymbol{S}_n$ tout entier et où tous les λ_k sont non négatifs. Dans le cas général on a une structure du plan qui vient de la propriété 4:

Propriété 9 : Le plan de variance minimale est caractérisé par un *N*-vecteur λ_k tel que :

$$p(s)=0 \qquad \text{si} \quad \sum_{k\in s}\lambda_k \leq 0$$

$$p(s)=\sum_{k\in s}\lambda_k \qquad \text{si} \quad \sum_{k\in s}\lambda_k > 0.$$

On peut alors utiliser une extension du schéma de Midzuno. Soit $w_k=sup(\lambda_k, 0) \geq 0$. Le schéma de tirage est le suivant :

Etape 1 : on tire un échantillon *s* selon le schéma de Midzuno en utilisant la variable w_k .

Etape 2 : soit $r = sup(0, \sum_s \lambda_k / \sum_s w_k)$. On accepte *s* avec la probabilité *r* ou on retourne à l'étape 1 avec la probabilité *1-r* .

Remarques : Comme $w_k \geq \lambda_k$, *r* est toujours compris entre 0 et 1. Sa valeur est *0* si $\sum_s \lambda_k \leq 0$ c'est à dire si *p(s)=0*. Elle est de 1 si tous les λ_k de *s* sont positifs. Dans ces deux cas il est donc inutile de générer un aléatoire uniforme pour le comparer à *r* de sorte que cette nécessité survient assez rarement. En revanche la nécessité de revenir à l'étape 1 peut être assez fréquente si les probabilités π_k sont très dispersées.

6. Cas d'un critère de dispersion plus général : plan de sondage

On cherche ici à étendre aux plans minimisant le critère (1) ce qui fonctionne dans le cas de la minimisation de la variance. Le premier axe concerne le calcul de λ .

On peut commencer par la normalisation des fonctions ψ. Les φ peuvent se normaliser par $\varphi(0)=0$, $\varphi'(0)=0$ ou $-\infty$ selon le cas et $\varphi'(1)=1$ ou $+\infty$ selon le cas. Comme φ' est croissante et sera de plus supposée continument dérivable, ψ est croissante de 0 à 1, continument dérivable sur son intervalle de définition qui est, selon le cas $[0,1]$, $[0,+\infty[$, $]-\infty,1]$ ou $]-\infty, +\infty[$. Dans les deux premiers cas on prolonge ψ par $\psi(\lambda)=0$ si $\lambda<0$ ce qui peut introduire une discontinuité sans importance en *0*

pour ψ' . Dans tous les cas on est amené à rechercher par des méthodes itératives la solution λ à $Up=U\,\psi(U'\lambda)$ avec ψ opérant coordonnée par coordonnée. On peut évidemment utiliser la méthode de Newton, surtout si ψ est suffisamment régulière (ayant une dérivée seconde assez peu variable) ou si les probabilités π_k ne sont pas très dispersées. La difficulté consiste à trouver une valeur initiale des itérations suffisamment proche de la solution exacte. La plus naturelle est de partir du SAS p_u avec donc $\lambda_0=\varphi'(\lambda_u)$. On peut imaginer tenir compte du fait que la fonction $\pi \to \lambda$ est monotone, mais cela est automatiquement réalisée par la première itération de la méthode de Newton qui donne $\lambda_1=\lambda_0 +(U\psi'(\lambda_0)U')^{-1}(\pi-n/N\mathbf{1}_N)= \varphi'(\lambda_u)+ \varphi''(\lambda_u)(\pi-n/N\mathbf{1}_N)$.

Cette technique fonctionne bien dans le cas de la variance et dans celui de l'entropie. Dans d'autres cas, en particulier pour $\varphi_{1/2}$ ou φ_3 , $\varphi''(\lambda_u)$ est trop grosse et il faut raccourcir le pas d'itération quand N , n ou la dispersion des π_k augmente. Nous n'avons pu dégager que quelques règles empiriques qui permettent une convergence de l'algorithme avec un peu de tâtonnement. Pour $\psi(\lambda)=\lambda^2$ par exemple, la règle consistant à diminuer le pas de la méthode de Newton d'un facteur égal à $100-niter$ (nombre d'itérations) fonctionne assez bien.

Cependant la principale difficulté de ces méthodes est en amont. Elle réside dans la nécessité d'énumérer tous les échantillons possibles, ce que nous ne sommes arrivé à faire commodément que pour N ne dépassant pas 30. A cet égard, le critère d'entropie fonctionne de façon tout à fait particulière car on sait que $\lambda_k \approx \log(\pi_k/(1-\pi_k))$ et que, de ce fait, la matrice inverse qui intervient dans la méthode de Newton peut être prise égale à l'identité. Dans ce cas on peut traiter sans problème des dimensions de l'ordre de quelques centaines.

Remarque sur le calcul des probabilités d'inclusion d'ordre 2 : Puisque le vecteur p peut être calculé, la matrice des probabilités s'inclusion d'ordre 2 est donnée par $U\,diag(p)\,U'$ et ce calcul ne pose aucune difficulté nouvelle puisque l'ensemble des échantillons a été énuméré. Ceci étant, le critère d'entropie a l'avantage de permettre un calcul récursif de ces probabilités d'inclusion [7] dont les limitations en volume de stockage sont beaucoup moindres. Ceci permet donc pouvoir utiliser de plus grandes valeurs de n et N .

7. Cas d'un critère de dispersion plus général : échantillonnage

L'autre axe de généralisation concerne l'échantillonnage quand λ est connu. L'idée consiste à utiliser l'algorithme de Midzuno pour présélectionner un échantillon selon un plan proche de celui qui nous intéresse puis à accepter ou non cet échantillon de façon rapide et simple, par analogie avec le cas de la minimisation de la variance. Pour ce faire nous aurons besoin d'une fonction convexe $w \geq \psi$ et donc positive ou nulle. Comme pour tout échantillon s on a $w_k = 1/n \sum_s w(n\lambda_k) \leq w(\sum_s \lambda_k) \leq \psi(\sum_s \lambda_k) = p(s)$, on peut tirer un échantillon par la méthode de Midzuno avec des probabilités proportionnelles aux w_k puis accepter l'échantillon avec la probabilité $n\ \psi(\sum_s \lambda_k)/\sum_s w(n\lambda_k)$. La méthode envisagée dans le cas de la variance revient à prendre $w(\lambda)=sup(\lambda,0)$.

Dans le cas où ψ est convexe, il est clair que qu'on ne peut faire mieux que prendre $w=\psi$. Dans une situation plus générale les fonctions admissibles sont les fonctions convexes minimales parmi celles qui majorent ψ. Si, par exemple, ψ est concave sur [*0,1*] , on tombe sur la famille *sup(0, aλ+b)* où *aλ+b* est l'équation d'une droite tangente au graphe de ψ sur [*0,1*]. Ceci posé il est nécessaire d'avoir des critères de choix plus fins dans cette famille. Le coût de l'échantillonnage est directement lié à la proportion de rejets dans la deuxième phase. Pour le minimiser il semble utile de choisir *w* proche de ψ surtout au voisinage de la valeur moyenne des λ_s qu'on peut calculer facilement. Néanmoins, si ψ est 'fortement convexe' (comme $\psi(\lambda)=\lambda^2$ simplement) la proportion d'échecs devient vite très forte si les π_k sont un peu dispersés. On obtient cependant des résultats plus rapides et plus fiables qu'avec un méthode de type Metropolis.

8. Illustration et commentaires

On a expérimenté assez systématiquement cinq critères de dispersion, qui sont, par « ordre de convexité croissante » :

- la distance de Hellinger *($\psi(u)=u^{-2}$, $\varphi(p)=-sqrt(p)$)*,
- l'entropie *($\psi=exp$, $\varphi(p)=p\ log(p)$*) ,
- le « carré » ($\psi(u)=u^2$, $\varphi(p)=p^{3/2}$) ,
- la variance ($\psi(u)=max(0,u)$, $\varphi(p)=p^2$)
- la « racine » ($\psi(u)=sqrt(u)$ si $u>0$, 0 sinon et $\varphi(p)=p^3$).

On a exploré des tailles de populations allant jusqu'à 25 et tailles d'échantillon allant jusqu' à 12. Au delà, l'énumération des échantillons s'est avérée trop lourde avec l'algorithme utilisé. L'annexe 3 décrit les résultats en détail pour un cas simple. Les conclusions sont cependant tout à fait générales et conformes à ce que l'on peut attendre des développements qui précédent. La distance de Hellinger et le critère d'entropie octroient des probabilités strictement positives à tous les échantillons de $\boldsymbol{S}_n$. La distance de Hellinger privilégie les échantillons extrêmes, petits ou gros. Les trois autres critères font apparaître, sur cet exemple choisi ad hoc, des probabilités nulles pour les échantillons les plus petits (le critère de classement est la probabilité obtenue pour le critère d'entropie). Sur cet exemple, il y a une jolie croissance de cet ensemble avec la convexité du critère surtout si on sait que le quatorzième échantillon à, en fait, une probabilité de $1.6\ 10^{-6}$ pour le « carré ». Ce comportement reste vrai dans ses grandes lignes pour tous les exemples que nous avons explorés. Si les probabilités d'inclusion π_k sont moins dispersées ont peut obtenir un ensemble de *p(s)* nuls de plus en plus petit, voire vide, et

de plus en plus facilement quand le critère est plus faiblement convexe. Pour les « gros » échantillons, l'augmentation de la convexité se traduit par une diminution des *p(s)* .Les conséquences sur les probabilités d'inclusion d'ordre 2 sont claires et faciles à comprendre. Le fait de ne plus disposer d'une estimation de variance sans biais est un inconvénient sans doute mineur.

Numériquement les critères d'entropie et de variance ne posent pas de problème. Les trois autres demandent une adaptation par tâtonnement à chaque cas particulier de tailles de population et d'échantillon. Pour ce qui concerne l'échantillonnage, la méthode proposée au paragraphe 7 marche parfaitement pour l'entropie, très bien pour la variance et Hellinger, de façon plus chaotique pour les deux autres critères.

En résumé, l'entropie est un critère presque parfait, mais la variance permet commodément d'éviter des échantillons extrêmes mal équilibrés.

Bibliographie

[1] Brewer, K.R.W. and Hanif, M., Sampling with Unequal Probabilities, *Springer-Verlag*, New-York,1983

[2] Berger, Y.G., Rate of Convergence to Normal Distribution for the Horvitz-Thompson Estimator,*Journal of Statistical Planning and Inference*, vol 67 , pp 209-226,1998.

[3] Bernd A. Berg. Markov Chain Monte Carlo Simulations and Their Statistical Analysis. Singapore, World Scientific 2004.

[4] Chen, S.X., Weighted polynomial models and weighted sampling schemes for finite population, *Annals of Statistics,* vol 26, pp 1894-1915 , 1998,

[5] Chen, S.X. , Dempster, A.P. and Liu, J.S.,Weighted finite population sampling to maximize entropy, *Biometrika*, vol 81 , pp457-469, 1994

[6] Cressie,N., Read, T.R.C., "Multinomial goodness-of-fit tests", *Journal of the Royal Statistical Society, Serie B,* vol 46, pp 440-464, 1984

[7] Deville,J-C. ,Qualité, L., « Echantillonnage multidimensionnel à entropie maximum : définitions, propriétés, algorithmes et programmes », *Journée de Méthodologie Statistiques 2005*.

[8] W.K. Hastings. "Monte Carlo Sampling Methods Using Markov Chains and Their Applications", *Biometrika*, vol 57, pp 97-109, 1970

[9] Joe, H., A Winning Strategy for Lotto Games , *The Canadian Journal of Statistics*, vol 18 , pp 233-244 , 1990

[10] Kuhn, H. W.; Tucker, A. W. "Nonlinear programming". Proceedings of 2nd Berkeley Symposium: pp 481-492, Berkeley: University of California Press, 1951.

[11] Lahiri, D.B. (1951). A method of sample selection providing unbiased ratio estimates, *Bulletin of the International Statistical Institute*, vol 33, pp 133-140 , 1951.

[12] Midzuno, H. (1952). On the sampling system with probability proportionate to sum of Sizes, *Annals of the Institute of Statistical Mathematics*, vol 3, pp 99-107.

[13] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, and E. Teller. "Equations of State Calculations by Fast Computing Machines*", Journal of Chemical Physics*, vol 21pp , 1087-1092, 1953.

[14] Rényi, A., "On measures of information and entropy". *Proceedings of the 4th Berkeley Symposium on Mathematics, Statistics and Probability 1960*: pp 547-561 , 1961.

[15] Rényi, A., "Calcul des probabilités suivi de Introduction à la théorie de l'information",*Dunod,*1966, réimpression *Gabay* 1992.

[16] Tillé Y., Sampling algorithms ,*Springer-Verlag,* New York , 2006.

ANNEXES

Annexe 1 : démonstration des propriétés 2 et 3

Propriété 2 : Supposons que $\varphi'(0)=-\infty$ et qu'une solution de $Up=\pi$, $p\geq 0$ soit réalisée 'en coin', c'est à dire avec $p(s)=0$ pour un ensemble non-vide S_0 d'échantillons. Soit une solution intérieure $p_1=p+\delta$ c'est à dire vérifiant les contrainte d'égalités et $p_1(s)>0$ pour tout s , et donc $\delta(s)>0$ pour s dans S_0.

Soit $\Phi(t)= \Sigma_S \varphi (p(s)+t\delta(s)) = \Sigma_{S0} \varphi(t\delta(s)) + t \Sigma_{S+} \varphi'(p(s)+t^*\delta(s))$ avec t^* entre 0 et 1 d'après le théorème des accroissements finis. La quantité $(\Phi(t)- \Phi(0)) /t$ tend vers moins l'infini quand t tend vers zéro, car c'est la somme d'un terme qui tend vers moins l'infini et d'un terme qui reste borné. $\Phi(t)- \Phi(0)$ est donc négatif sur un voisinage de 0 ce qui prouve que le minimum du critère ne peut pas être en coin.

Réciproquement : Si $\varphi'(0)$ est fini, que $n>2$ et $N-n>2$, on peut trouver un vecteur de π_k strictement positives tel que l'optimum soit 'en coin'.

On commence par regarder comment les choses se passent si on admet des probabilités d'inclusion nulles. Plus précisément, prenons $\pi_1 = \pi_2 =0$ et $\pi_k =n/(N-2)$ pour $k= 3$ à N. La solution du problème *Min* $\Sigma_S \varphi (p(s))$ avec $Up=\pi$, $p\geq 0$ conduit évidemment à $p(s)=0$ si s contient 1 ou 2 et $p(s)=\begin{pmatrix} N-2 \\ n \end{pmatrix}^{-1}$ si s ne contient ni 1 ni 2. On va montrer que si $\pi_1 = \pi_2 =\varepsilon$ suffisamment petit, l'optimum nécessite que $p(s)=0$ quand s contient 1 et 2 . Pour d'évidentes raisons de symétrie celui ci est de la forme :

-Si s contient 1 et 2 $p(s)=p_2$ soit $\begin{pmatrix} N-2 \\ n-2 \end{pmatrix}$ échantillons.

-Si s contient soit 1 soit 2 $p(s)=p_1$ soit $2.\begin{pmatrix} N-2 \\ n-1 \end{pmatrix}$ échantillons.

-Si *s* ne contient ni *1* ni *2* $p(s)=p_0$ soit $\binom{N-2}{n}$ échantillons.

Au facteur $\binom{N}{n}$ près, le critère à minimiser s'écrit, avec $f_1=n/N$ et $f_2=n(n-1)/N(N-1)$:

$$\Phi = f_2\phi(p_2) + 2f_1\phi(p_1) + (1-2f_1-f_2)\,\phi(p_0)$$

sous les contraintes :

$\varepsilon = f_2 p_2 + f_1 p_1$ et $1 = f_2 p_2 + 2f_1 p_1 + (1-2f_1-f_2)\,p_0$, soit $p_0 = (1- f_2 p_2 - 2f_1 p_1)/(1-2f_1-f_2)$.

Φ est différentiable en $\pi_1 = \pi_2 = 0 = p_1 = p_2$ et on a $d\Phi = (\phi'(0)-\phi'(n/(N-2)))(f_2\,dp_2 + 2f_1\,dp_1)$.

Comme $\phi'(0)<\phi'(n/(N-2))$, que dp_1 et dp_2 sont positives et que $f_2 < 2f_1$, l'optimum évolue au voisinage de *0* (ε petit) avec $dp_1>0$ et $dp_2=0$. Pour ε suffisamment petit on aura donc un optimum en coin.

Remarque : cet exemple montre qu'on a alors $\pi_{12}=0$ et donc qu'on ne peut pas assurer la positivité des probabilités d'inclusion d'ordre deux pour tous les couples *{k, l}*.

<u>Annexe 2 : démonstration du lemme de la propriété 8.</u>

Pour *t* appartenant à [*0,1*] soit $p_t = p_0 + t\,(p_* - p_0)$. On a, pour tout *t* , $Up_t = \pi$. Par ailleurs, comme p_0 est la projection orthogonale de l'origine sur cette variété affine on a $(p_* - p_0)'\,p_0 = 0$ et $p_t'\,p_t = p_0'\,p_0 + t^2\,(p_* - p_0)'(p_* - p_0)$. Soit $\boldsymbol{S}_{-+} = (\boldsymbol{S}_0 - \boldsymbol{S}_{0+}) \cap supp(p_*)$ l'ensemble des coordonnées *s* telles que $p_0(s) \le 0$ et $p_*(s) > 0$.

Prouver le lemme équivaut à montrer que $\boldsymbol{S}_{-+}$ est vide. Si ce n'est pas vrai, soit θ la plus petite valeur de *t* telle que $p_t(s) \ge 0$ pour tout *s* de $\boldsymbol{S}_{-+}$. Toutes les coordonnées de p_θ sont positives ou nulles et ce plan vérifie donc les contraintes. Or , clairement, $\theta < 1$. Il en résulte que $p_\theta'\,p_\theta < p_*'\,p_*$ ce qui contredit le fait que p_θ réalise l'optimum.

Annexe 3 : Illustration

Pour rester lisible mais néanmoins présenter les principales caractéristiques des plans à dispersion minimale, nous avons choisi le cas $N=7$et $n=3$. Le vecteur des probabilités d'inclusion est π=[.05 .1 .35 .45 .55 .7 .8] , soit quelque chose d'assez dispersé avec deux petits éléments. Au total le support de p comporte au plus 35 éléments, ce qui permet une lecture 'à la main' si on peut oser cette champignaquerie. Les conclusions reste néanmoins qualitativement valide empiriquement jusqu'à N de l'ordre de 20 à 30 et n de l'ordre de 12.

On a utilisé dans cette exercice quatre fonctions, par « ordre de convexité croissante » : la distance de Hellinger $(\psi(u)=u^{-2}, \varphi(p)=-sqrt(p))$, l'entropie $(\psi=exp, \varphi(p)=p\ log(p))$, le « carré » ($\psi(u)=u^2, \varphi(p)=p^{3/2}$) , la variance ($\psi(u)=max(0,u), \varphi(p)=p^2$) et la « racine » ($\psi(u)=sqrt(u)$ si $u>0$, 0 sinon et $\varphi(p)=p^3$).

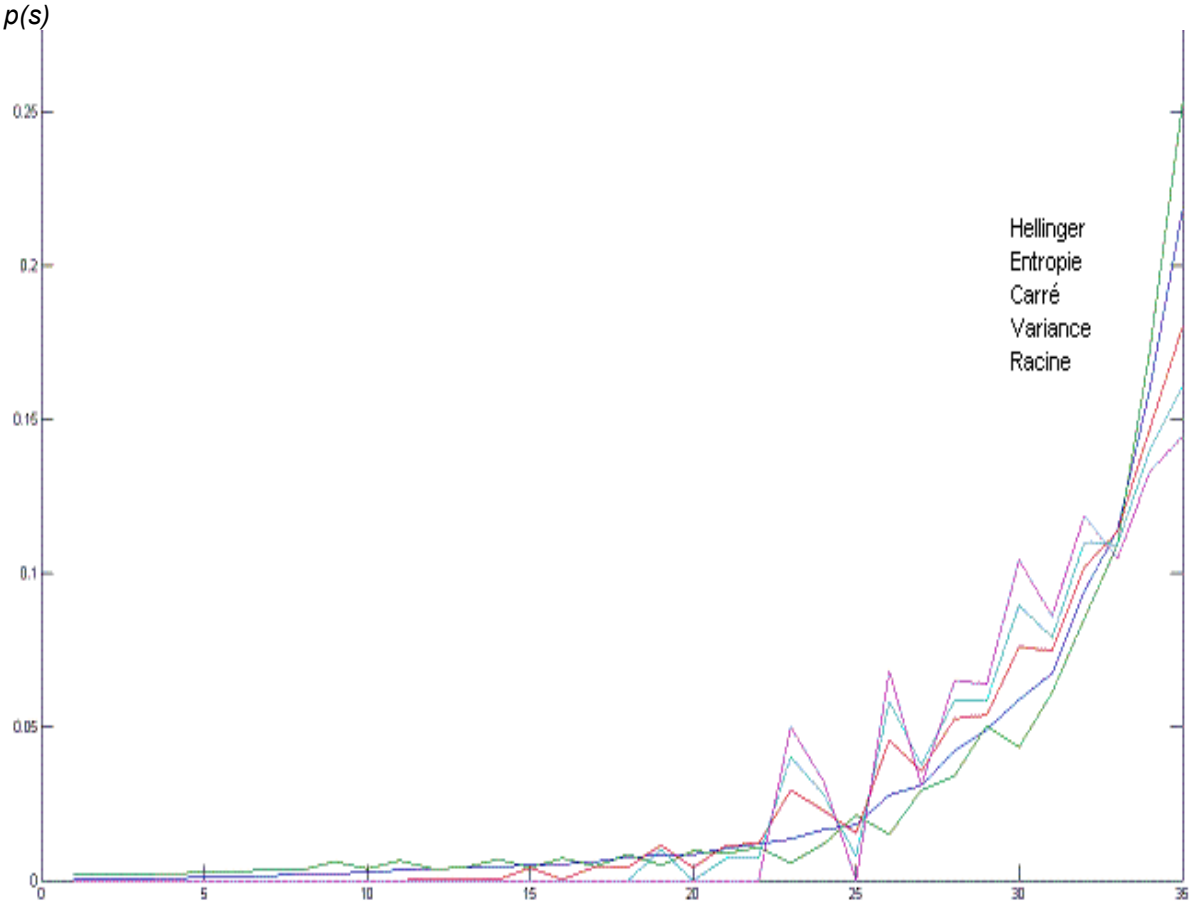
On obtient :

		p(s)			s		
Hellinger	Entropie	Carré	Linéaire	Racine			
0,00173	0,00019	0	0	0	1	2	3
0,00182	0,00027	0	0	0	1	2	4
0,00189	0,00037	0	0	0	1	2	5
0,00207	0,00063	0	0	0	1	2	6
0,00235	0,00101	0	0	0	1	2	7
0,00301	0,00112	0	0	0	1	3	4
0,00315	0,00153	0	0	0	1	3	5
0,00338	0,00215	0	0	0	1	4	5
0,00584	0,00230	0	0	0	2	3	4
0,00355	0,00260	0	0	0	1	3	6
0,00624	0,00314	0	0	0	2	3	5
0,00382	0,00364	0,00050	0	0	1	4	6
0,00421	0,00416	0,00039	0	0	1	3	7
0,00689	0,00441	0	0	0	2	4	5
0,00403	0,00498	0,00405	0	0	1	5	6
0,00737	0,00533	0,00036	0	0	2	3	6
0,00456	0,00583	0,00433	0	0	1	4	7
0,00821	0,00748	0,00425	0	0	2	4	6
0,00484	0,00797	0,01148	0	0	1	5	7
0,00947	0,00854	0,00392	0	0	2	3	7
0,00888	0,01022	0,01135	0,00714	0	2	5	6
0,01072	0,01197	0,01182	0,00714	0	2	4	7
0,00559	0,01353	0,02926	0,04000	0,05000	1	6	7
0,01173	0,01637	0,02252	0,02786	0,03199	2	5	7
0,02129	0,01810	0,01545	0,00762	0,00000	3	4	5
0,01479	0,02776	0,04578	0,05786	0,06801	2	6	7
0,02933	0,03070	0,03542	0,03762	0,03071	3	4	6
0,03409	0,04197	0,05268	0,05833	0,06498	3	5	6
0,05021	0,04915	0,05370	0,05833	0,06385	3	4	7
0,04341	0,05885	0,07596	0,08929	0,10431	4	5	6
0,06131	0,06720	0,07456	0,07905	0,08577	3	5	7
0,08535	0,09422	0,10185	0,11000	0,11839	4	5	7
0,10921	0,11398	0,11355	0,10905	0,10469	3	6	7
0,17214	0,15982	0,14673	0,14000	0,13274	4	6	7
0,25351	0,21851	0,18011	0,16071	0,14456	5	6	7

NB : les échantillons sont ordonnés par probabilité croissante pour le critère d'entropie. Les trois dernières colonnes sont les numéros d'ordre des éléments de l'échantillon.

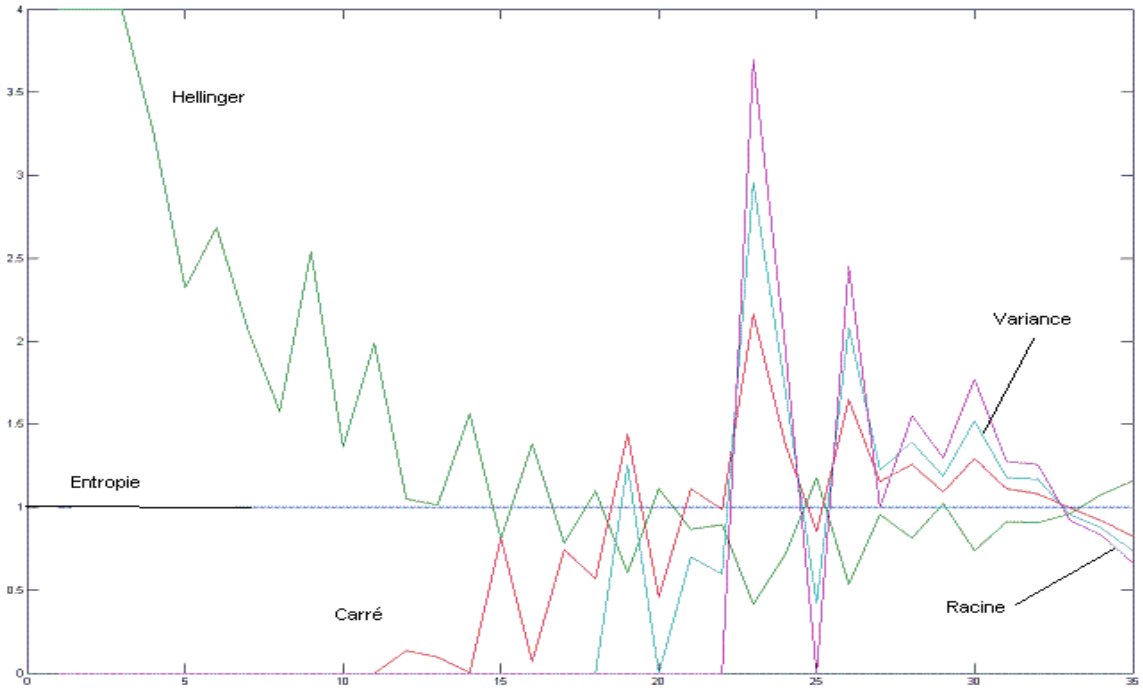
Graphiquement :

p(s)



Numéro d'ordre de l'échantillon

Ratio $p(s)/p_{ent}(s)$:



Numéro d'ordre de l'échantillon

Probabilités d'inclusion d'ordre 2 :

Hellinger

$$\begin{bmatrix} 0.0099 & & & & & \\ 0.0156 & 0.0306 & & & & \\ 0.0166 & 0.0335 & 0.1097 & & & \\ 0.0173 & 0.0356 & 0.1261 & 0.1603 & & \\ 0.0191 & 0.0413 & 0.1835 & 0.2569 & 0.3439 & \\ 0.0216 & 0.0491 & 0.2344 & 0.3230 & 0.4167 & 0.5552 \end{bmatrix}$$

Entropie

$$\begin{bmatrix} 0.0025 & & & & & \\ 0.0096 & 0.0195 & & & & \\ 0.0130 & 0.0264 & 0.1014 & & & \\ 0.0170 & 0.0345 & 0.1319 & 0.1777 & & \\ 0.0254 & 0.0514 & 0.1946 & 0.2605 & 0.3345 & \\ 0.0325 & 0.0656 & 0.2430 & 0.3210 & 0.4043 & 0.5336 \end{bmatrix}$$

Carré

$$\begin{bmatrix} 0 & & & & & \\ 0.0004 & 0.0043 & & & & \\ 0.0048 & 0.0161 & 0.1045 & & & \\ 0.0155 & 0.0339 & 0.1427 & 0.1933 & & \\ 0.0338 & 0.0617 & 0.2020 & 0.2629 & 0.3242 & \\ 0.0455 & 0.0840 & 0.2461 & 0.3184 & 0.3905 & 0.5154 \end{bmatrix}$$

Variance

$$\begin{bmatrix} 0 & & & & & \\ 0 & 0 & & & & \\ 0 & 0.0071 & 0.1036 & & & \\ 0.0100 & 0.0350 & 0.1450 & 0.2069 & & \\ 0.0400 & 0.0650 & 0.2050 & 0.2669 & 0.3155 & \\ 0.0500 & 0.0929 & 0.2464 & 0.3155 & 0.3876 & 0.5076 \end{bmatrix}$$

Racine

$$\begin{bmatrix} 0 & & & & & \\ 0 & 0 & & & & \\ 0 & 0 & 0.0946 & & & \\ 0 & 0.0320 & 0.1508 & 0.2227 & & \\ 0.0500 & 0.0680 & 0.2004 & 0.2678 & 0.3139 & \\ 0.0500 & 0.1000 & 0.2543 & 0.3150 & 0.3807 & 0.5000 \end{bmatrix}$$