# ECHANTILLONNAGE EQUILIBRE EXACT POISSONNIEN

Jean-Claude Deville

17 bis rue Gabriel Péri 92250-La Garenne-Colombes
deville@ensai.fr

**Résumé.** Un échantillonnage équilibré est exact si tous les sommets du polytope des probabilités d'inclusion permises, *K*, sont aussi des sommets du cube unité de dimension *N* (taille de la population. Autrement dit, l'algorithme du cube n'a pas besoin de phase d'atterrissage.
S'il existe au moins un échantillon vérifiant l'équilibre, les trois assertions suivantes sont des CNS d'exactitude :

a- Toutes les sous matrices carrées de plein rang de *A* ont au signe près le même déterminant.
b- La matrice des contraintes peut se mettre sous une forme totalement unimodulaire.
c- Le sous groupe de $\mathbb{Z}^N$ généré par les arêtes de *K* est identique au groupe des entiers de *kerA*.

Il en résulte que les contraintes peuvent s'écrire à l'aide d'une matrice dont les éléments appartiennent à {-1,0,+1} et que pour tout $X$ de $IM(S) = \{As; s\ sommet\ de\ C\}$, l'échantillonnage contraint par *As=X* est exact. Les points extrêmes du polytope *conv*$(IM(S))$ de $\mathbb{R}^p$ correspondent à des problèmes dégénérés où *K* est réduit à un sommet unique de *C*.
Pour un $\pi$ intérieur à *C* on peut alors s'intéresser à l'échantillonnage poissonnien conditionnel à *As=X* pour tout *X* de *IM*(*S*). On obtient des algorithmes qui calculent les probabilités d'inclusion conditionnelles, un plan poissonnien dont une conditionnelle est donnée et un procédé d'échantillonnage dans la loi conditionnelle. On est donc capable de réaliser un échantillonnage poissonnien conditionnel équilibré exactement sur *X* et respectant des probabilités d'inclusion données.

**Mots-clés.** Echantillonnage équilibré, plan poissonnien et poissonnien conditionnel, matrice unimodulaire, sous groupes de $\mathbb{Z}^N$ , algorithme merveilleux .

## 1 Position du problème

Une stratégie est représentative au sens de Hajek ([4]) si un couple plan-estimateur appliqué à $p$ variables auxiliaires restitue exactement leurs totaux $X = \sum_U x_k$. L'échantillonnage est équilibré si ce but est obtenu en utilisant l'estimateur (linéaire) de Horvitz-Thompson dont les poids sont les inverses des probabilités d'inclusion $\pi_k$ fixées pour tout *k* de la population *U*. Autrement dit on a pour tout échantillon *s* possible $X = \sum_s x_k/\pi_k = As$ en notant *A* la *pxN* matrice dont les colonnes sont les $a_k = x_k/\pi_k$ et *s* le *N*-vecteurs qui code l'appartenance à l'échantillon, qui peut être vu comme un sommet du *N*-cube unité $= [0,1]^N$ . Les contraintes que doivent vérifier l'échantillon s'écrivent donc *As=X=Aπ* ( $\pi$ vecteur des probabilités d'inclusion). Un problème d'échantillonnage équilibré a donc pour paramètres un vecteur $\pi$ situé dans l'intérieur de *C* (pour des raisons évidentes on peut éliminer des probabilités d'inclusion égales à 0 ou 1) et une *pxN* matrice *A*.

Malheureusement, on sait bien que ce problème n'a que rarement une solution exacte. La méthode du cube (Deville et Tillé [2] et [3] ) donne une solution presque exacte (en un certain sens la plus exacte possible) grâce à une 'phase d'atterrissage ' où on relâche légèrement la contrainte. Dans ce texte on présente des conditions nécessaires et suffisantes d'exactitude. Elle font apparaitre une structure beaucoup plus riche qu'on pouvait l'imaginer : si le problème est exact pour une valeur de

$X=A\pi$ , il l'est aussi pour toute valeur de $X \in IM(S) = \{As\,;\, s\ sommet\ de\ C\}$ .

Des lors, si on se donne un plan poissonnien sur *U*, on peut s'intéresser à la famille des plans conditionnels à *X* qui sont des plans équilibrés pour les probabilités d'inclusion conditionnelles $\pi^X$ . On les calculera grâce à un algorithme qui généralise ce qu'on sait faire (Deville ([0]) ou Tillé ([9])) dans le cas où la contrainte est la taille de l'échantillon. Inversement, étant donnée une conditionnelle $\pi^X$, on calculera le $\pi$ d'un plan poissonnien admettant $\pi^X$ comme conditionnelle et vérifiant la contrainte $A\pi^X = X$ .

## 2 Eléments de géométrie et conditions nécessaires et suffisantes d'exactitude

Soit $K = C \cap (\pi + \ker(A))$ l'ensemble des probabilités d'inclusion admissibles. C'est un polytope convexe de dimension *M=N-p* puisque $\pi$ est à l'intérieur de *C*. L'exactitude du problème se traduit par le fait que les sommets de *K* sont aussi des sommets de C (les échantillons équilibrés).Par suite tous les *s-s'* joignant deux sommets de *K* , en particulier les arêtes, sont des éléments de $\mathbb{Z}^N$ (dont les coordonnées valent *0,+1* ou *-1* ) ; ces vecteurs engendrent donc un sous groupe $G_K$ de $\mathbb{Z}^N$ contenu dans ker(A) et donc dans le groupe $G_A$ des points de coordonnées entières de ker(A). On a le résultat suivant :

Théorème (je dis rarement ça !) : Les propositions suivantes sont équivalentes :

(i) Le problème (A,π) est exact.

(ii) Toutes les *pxp* sous-matrices carrées de A de plein rang ont le même déterminant en valeur absolue.

(iii) $G_A=G_K$ .

(iv) Il existe une *pxp* matrice inversible *W* telle que $WA = (I_p \quad B)$ où $I_p$ est la matrice identité d'ordre *p* et *B une p x(N-p)* matrice totalement unimodulaire.

Rappelons qu'une matrice est totalement unimodulaire si toutes ses sous-matrices carrées ont un déterminant qui vaut +1,0 ou -1.

Conséquence : Comme les *1x1* sous-matrices sont les éléments de B ceux-ci valent +1,0 ou -1. Une condition d'équilibrage est donc que les contraintes puissent s'écrire sous cette forme. Mais ce n'est pas suffisant comme le montre l'exemple minimal suivant :

$$A = \begin{pmatrix} 1 & 0 & 1 & 1 \\ 0 & 1 & -1 & 1 \end{pmatrix} \quad ker(A) \text{ est généré par } \begin{pmatrix} 1 & 0 & 1/2 & 1/2 \\ 0 & 1 & 1/2 & -1/2 \end{pmatrix}.$$

Commentaire : la chose profonde est évidemment le point (iii). Techniquement, il implique de façon assez simple les trois autres. En revanche, sauf méprise de ma part, la démonstration de (i)=>(iii) est plutôt ardue et longuette. Intuitivement, la maille du réseau de $G_K$ est trop petite pour laisser échapper des points de coordonnées entières.

Remarque : Le 'théorème 3.15' donné en annexe, trouvé je ne sais trop où sur internet, exhibe une très intéressante CNS d'unimodularité à son point (vii). Traduit dans notre problématique, il dit qu'une matrice *n'est pas* TUM si et seulement si on peut faire apparaître un bloc $\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$ sans faire disparaitre une sous matrice identité par des manipulations *A* se limitant à :

- Soustraire ou additionner une ligne à une ligne donnée
- Changer le signe d'une colonne.

Il me semble que ça pourrait servir à quelque chose, mais je ne sais pas encore à quoi.

## 3 Structure de l'espace image

La remarque précédente laisse penser que le groupe $A(\mathbb{Z}^N)$, qui est égal à $\mathbb{Z}^p$ dans le cas TUM doit avoir des 'trous' quand ce n'est plus le cas. Néanmoins je ne connais pas, à ce jour, de caractérisation de l'exactitude basée sur la structure de l'espace image.

Pour des raisons statistiques nous avons limité le problème au cas où $\pi$ est intérieur à $C$ ; autrement dit il n'a aucune coordonnée égale à 0 ou1, ce qui revient à dire que ker($A$) ne contient aucune $p$-face ($p$>0) du cube. Néanmoins on ne peut pas éviter d'examiner ce qui se passe quand on a des coordonnées de $\pi$ égales à 0 ou 1, ne serait ce que parce que dans l'algorithme du cube le but est d'atteindre de tels points! Dans l'algorithme c'est simple : on élimine la dimension pour poursuivre, et c'est la raison même pour la quelle on avait exclus cette éventualité dans la définition du problème.

On a besoin de la construction suivante : on normalise la matrice $A$ mise sous la forme $(I_p \quad B)$ du théorème précédent au point (iv). De plus chaque $b_k$ dont la première coordonnée non nulle est négative est remplacée par $-b_k$ ce qui revient à changer le système de coordonnées en $s \rightarrow 1-s$ pour les coordonnées correspondantes. Soit $U^\circ$ l'ensemble des colonnes distinctes de cette matrice et $A^\circ$ la matrice $(I_p \quad B^\circ)$ formée par cet ensemble de colonnes. Pour chaque $a$ de $U^\circ$ on aura $N_a$>0 colonnes de $A$ égales à $a$ . Les valeurs possibles des contraintes sont donc de la forme $\sum_{U^\circ} n_a \; a \;\; avec \; 0 \leq n_a \leq N_a$ ; *0* et $\sum_{U^\circ} n_a$ sont évidemment des points extrêmes de $IM(S)$ correspondant aux échantillons $s=\emptyset$ et $s=U$. Mais ce ne sont pas les seuls. On établit les résultats suivant :

- $A(C)$ est un polytope convexe de dimension $p$. Ses sommets sont aussi les points extrêmes de $IM(S)$. C'est l'image du parallélotrope $0 \leq x_a \leq N_a$ de $\mathbb{R}^{N^\circ}$ par $A^\circ$.
- le point $1/2 \sum_{U^\circ} N_a \; a$ est centre de symétrie de $A(C)$ et tout symétrique d'un point extrême est aussi extrême.
- Pour tout $X$ intérieur à $A(C)$ le polytope $A\pi=X$ est de dimension $M=N-p$ et son intérieur un $N$-vecteur de probabilités de ]0,1[ .
- Pour des $X$ sur le bord de $A(C)$ , la dimension de $A\pi=X$ peut varier de 0 (point extrême) à $M$-1 et son intérieur un $N$-vecteur de probabilités comportant de $N$ à $N-N_a$ 0 ou 1 selon le cas.
- $X_1$ et $X_2$ sont connexes s'il existe $a$ tel que $X_1 = X_2 \pm a$. Tout point $X$ intérieur à $A(C)$ est connecté à un point extrême de plusieurs façons. Si ce sommet est 0, ceci est obtenu comme une addition de vecteurs a.

## 4 Echantillonnage poissonnien conditionnel début

Soit $\pi \in ]0,1[^N$ un vecteur de probabilités d'inclusion définissant un échantillonnage poissonnien par $p(s) = \prod_s \pi_k \prod_{U-s}(1-\pi_k) \; = p(\emptyset) \; \prod_s \omega_k \;\; avec \; \omega_k = \frac{\pi_k}{1-\pi_k}$ . Pour tout $X \in IM(S)$ l'échantillonnage poissonnien conditionnel à $X$ est défini par $p_X(s) = p(X)^{-1} \; \prod_s \omega_k$ si $As = X$ et 0 sinon. Si $X$ est extrême, $p_X(s_X)=1$ pour l'unique échantillon réalisant la contrainte. Avec $A$ normalisé comme au 3 $X$ est un $p$-vecteur d'entiers $n=(n_1,\ldots,n_p)$ . Soit $t=(t_1,\ldots,t_p)$ et $t^n$ le monôme $\prod_{i=1}^{p} {t_i}^{n_i}$ .

**a-** On obtient les probabilités $p(X)=p(n)$ grâce à l'égalité suivante :

$$P_U(t) = \prod\nolimits_{k \in U} (1 + \; \omega_k t^{a_k}) = \; P_U(\mathbf{1})^{-1} \sum\nolimits_{IM(S)} p(n) \; t^n$$

Le calcul des coefficient du polynôme consiste à remplir un tableau à $(N+1)^p$ entrées de façon récursive . Il nécessite surtout un espace de stockage relativement grand mais le volume de calcul

est assez restreint et la programmation très simple (environ dix lignes de matlab).

Remarque : le fait que certains *a* aient des coordonnées négatives ne doit effrayer personne !

**b-** Probabilités d'inclusion conditionnelle.

Il y a au moins deux méthodes. La première est assez brutale : *X* ou *n* étant donné, pour tout *k* on calcule le polynôme $P_{U-k}(t) = \prod_{l\in U-k}(1 + \omega_l t^{a_l})$ et donc, en notant $[n](P)$ le coefficient de $t^n$ dans le polynôme *P* on a $1\text{-}\pi_k^n = [n](P_{U-k})/[n](P_U)$ . Pour les probabilités d'inclusion d'ordre deux on peut faire la même chose avec des polynômes de type $P_{U\text{-}k,l}$ .

La seconde se calque sur ce qu'on fait dans le cas où la seule contrainte est la taille fixe. Il suffit de constater que $\pi_k^n / p(n) = \omega_k(1 - \pi_k^{n-a_k})/p(n-a_k)$ . La récurrence débute en 0 où toutes les probabilités d'inclusion sont nulles et se poursuit en augmentant les tailles de *n* de façon convenable pour que tous les $\pi_k^{n-a_k}$ soient calculés quand on arrive à l'étape *n* . Techniquement, c'est beaucoup plus délicat à programmer mais un peu moins lourd en calculs et peut être plus stable numériquement.

Remarque : la première méthode est tout à fait applicable pour la taille fixe , mais bien plus lourde que les récurrences 'standard' qui figurent dans Deville ([0]) où Tillé ([9]).

## 5 Echantillon poissonnien conditionnel suite

**a-** Posons $\lambda_k = \log(\omega_k)$, logit de $\pi_k$. On a $p_n(s) = C_n \exp(\lambda' s)\ 1(As = n)$ avec $C_n$ constante de normalisation. Pour tout $\mu$ de $\mathbb{R}^p$ soit $\lambda_\mu = \lambda + A'\mu$ . On a $\exp(\lambda'_\mu\, s) = \exp(\lambda' s)\,.\exp(\mu' As) = p_n(s)$ à une constante qui ne dépend que de *n* près. Les poissonniens ayant pour vecteurs de logit les $\lambda_\mu$ ont donc tous les mêmes loi conditionnelles pour tout *n* .

Inversement, étant donné *x* intérieur à *A*(*C*) , il existe un $\pi$ unique de la famille (donc un $\mu$ unique de $\mathbb{R}^p$) tel que $A\pi = x$ . En particulier, étant donné un problème exact $A\pi^{\,n} = n$, il existe un poissonnien $\pi$ vérifiant $A\pi = x$ admettant $\pi^n$ comme conditionnelle à *n* . On a évidemment $\pi = \sum_{m\in IM(S)} p(m)\,\pi^m$ de sorte que $\pi$ est plus près du centre de *K* (projection orthogonale du centre de *C* sur *K*) que $\pi^n$ .

Cette dernière remarque permet d'accélérer l'algorithme qui permet de calculer effectivement $\pi$ et qui est une généralisation évidente de celui qui est donné dans Deville ([0]) où Tillé ([9]).

**b-** Reste à voir comment on réalise l'échantillonnage. On dispose donc de probabilités $\pi^n$ vérifiant $A\pi^{\,n} = n$, et on calcule le $\pi$ du poissonnien qui admet $\pi^{\,n}$ comme conditionnelle à *n*. Maintenant encore, il y a deux méthodes, la brutale et la subtile.

La méthode brutale décalque ce qu'on fait dans le cas de la taille fixe. On sélectionne $k=1$ avec la proba $\pi_k^{U,n}$ . Si *k* est sélectionné, on calcule les conditionnelles $\pi_k^{U-1,n-a_1}$ pour $k>1$ et pose $s_1=1$ ; sinon on pose $s_1=0$ et on calcule les conditionnelles $\pi_k^{U-1,n}$ pour $k>1$. Ceci permet de fixer le sort de l'individu 2.

Supposons que nous soyons arrivés à l'étape $k°$. Posons $(k°) = n - \sum_{l=1}^{l=k°} s_l a_l$ . On calcule les conditionnelles $\pi_k^{U-\{1,..,k°\},n(k°)}$ pour $k>k°$ . Ceci permet de fixer le sort de l'individu $k°+1$.

L'inconvénient (c'est le côté brutal de l'histoire), c'est d'avoir à calculer toutes les probabilités conditionnelles pour ne se servir que d'une seule. L'avantage (théorique et parfois pratique), c'est que le suivant de $k°$ a le droit d'être n'importe qui.

La méthode subtile. Pour tout $a \in U°$ notons $U_a$ la 'strate' des *k* ayant $a_k = a$ (de taille $N_a$). Sans contrainte, les échantillons $s_a = s \cap U_a$ sont poisonniens et indépendants. La loi de la taille $n_a$ s'obtient sans problème (polynôme $P_{U_a}(t) = \prod_{k\in U_a}(1 + \omega_k t) = P_{U_a}(1)^{-1} \prod_0^{N_a} p(n_a)\, t^{n_a}$ ) .

Conditionnellement à $n_a$ l'échantillonnage dans $U_a$ est un poissonien conditionnel 'ordinaire'. Il suffit donc d'échantillonner les $n_a$ sous la contrainte $\sum_{U_0} n_a\, a = n$ . Or cela peut se faire grâce à une variante de l' 'algorithme merveilleux' décrit dans Deville ([1]).

## 6 Conclusion

Un échantillonnage équilibré exact n'est possible que pour des contraintes vérifiant des conditions très particulières mais assez souvent vérifiées dans la pratique. L'équilibrage sur les marges d'un tableau de contingence en est l'exemple typique. L'échantillonnage poissonnien conditionnel est alors possible et représente une alternative plus satisfaisante que la méthode du cube.

## Bibliographie


[0] Deville, J.-C.,(2000 à 2012), *Cours de sondage approfondi*, notes de cours ENSAE.

[1] Deville, J.-C.,(2014) *Atelier sur le sondage adaptatif*, 8ème colloque francophone sur les sondages.

[2] Deville, J.-C., Tillé, Y.,(2004). Efficient balanced sampling: The cube method, *Biometrika*, vol 91,pp 893-912.

[3] Deville, J.-C., Tillé, Y., (2005), Variance approximation under balanced sampling, *Journal of Statistical Planning and Inference*, vol 128, pp 411-425.

[4] Hajek,J, (1981) ,Sampling from a finite population, *Dekker*, New York.

[5] Kiaer, A.W (1895). The Representative Method of Statistical Surveys*, English translation, 1976, Oslo: Statistik Sentralbyro.*

[6a] Kruskal, W.,Mosteller, F. (1979 a):„Representative sampling". In: *Non-scientific literature. International Statistical Review 47*, 13-24.

[6b] Kruskal, William/Mosteller, Frederick (1979 b):„Representative sampling, II. Scientific literature,excluding statistics". In: *International Statistical Review 47*, 111-127.

[6c] Kruskal, William/Mosteller, Frederick (1979 c):„Representative sampling, III: The current statistical literature". In: *International Statistical Review 47*, 245-265.

[6d] Kruskal, William/Mosteller, Frederick (1980): „Representative sampling, VI: the history of the concept in statistics, 1895-1939". In: *International Statistical Review 48*, 169-195.

[7] Neyman, J. (1934). On the different aspects of the representative method: the method of stratified sampling and the method of purposive selection, *Journal of the Royal Statistical Society*, 97.

[8] Royall, R.M. and Herson, J.(1973). Robust estimation in finite populations {I}, *Journal of the American Statistical Association*, vol 68,pp880-889.

[9] Tillé, Y. (2006). *Sampling Algorithms*. New York: Springer

[10] Valliant, R., Dorfman, A.H., Royall, R.M (2000). Finite Population Sampling and Inference: A Prediction Approach , *Wiley*,New York.

[11] Ziegler,G,M (2000) : Lectures on polytopes, *Springer*.

[12] G.M. Ziegler. Lectures on 0-1-polytopes. In G. Kalai and G.M. Ziegler, editors, Polytopes – Combinatorics and Computation, volume 29 of DMV Seminars, pages 1–41. Birkhauser-verlag, Basel, 2000.

**Annexe:**
Un énoncé synthétique qui a son intérêt même si c'est de l'angliche.

**Theorem 3.15:** Let *A* be a matrix with entries 0, +1 or -1. Then the following are equivalent:

(i) A is totally unimodular, i.e. each square submatrix of *A* has determinant 0, +1, or -1.

(ii) **[Hoffman & Kruskal]** For each integral vector *b* the polyhedron $\{x: x \geq 0\ ,\ Ax \leq b\}$ has only integral vertices.

(iii) **[Hoffman & Kruskal]** For all integral vectors *a,b,c,d* the polyhedron $\{x: c \leq x \leq d\ , a \leq Ax \leq b\}$ has only integral vertices.

(iv) **[Ghouila-Houri]** Each collection of columns of *A* can be split into two parts so that the sum of the columns in one part minus the sum of the columns in the other part is a vector with entries only 0, +1, or -1.

(v) **[Camion]** Each nonsingular submatrix of *A* has a row with an odd number of nonzero components.

(vi) **[Camion]** The sum of the entries in any square submatrix with even row and column sums is divisible by four.

(vii) **[R.E.Gomory]** No square submatrix of *A* has determinant +2 or -2.