ECHANTILLONNAGE MULTIDIMENSIONNEL (= DE PLUSIEURS ECHANTILLONS A LA FOIS) A ENTROPIE MAXIMUM : DEFINITION, PROPRIETES, ALGORITHMES ET PROGRAMMES .

Jean-Claude DEVILLE (CREST/ENSAI) et Lionel QUALITE (CREST/ENSAI puis Université de Neuchâtel)

On désire tirer dans une même population U de taille N, Q échantillons disjoints en même temps (et non pas successivement !) ayant des probabilités d'inclusion π^i_k ($i=1$ à Q, $k=1$ à N) fixées et de tailles fixées n_i .

Exemple:
-Contrôle de qualité du recensement dans deux ateliers, l'un s'occupant de l'analyse ménage-famille, l'autre de la qualité des libellés pour la codification automatique. On doit de tirer des districts entiers, les premiers avec des probabilités proportionnelles au nombre de gros ménages, le second avec des probabilités proportionnelles au nombres d'actifs.

-Imputation d'une variable quantitative en cas de non-réponse.

Echantillonnage à entropie maximum

On s'intéresse ici au cas d'un seul échantillon tire avec des probabilités d'inclusion fixée $0 < \pi_k < 1$ et un support fixé $\boldsymbol{S}$. De plus, on veut que les variables indicatrices I_k (=1 si k dans s, 0 sinon) soient les plus indépendantes possibles, ou, ce qui revient sensiblement au même, que les $p(s)$ aient une dispersion minimale. Sans faire trop de philosophie, un bon critère est de maximiser l'entropie $\sum_{s \in \mathbf{S}} -p(s)\log(p(s))$ du plan de sondage.

Les N contraintes sont associées à des multiplicateurs de Lagrange λ_k La solution du problème d'optimisation donne $p(s)=C(\boldsymbol{S})\ exp(\lambda.s)$ où apparaît le produit scalaire du N-vecteur des λ_k avec le vecteur s des 'coordonnées de s ': 1 si k est dans s, 0 sinon. La constante $C(\boldsymbol{S})$ normalise la somme des probabilités à un.

-Supposons que ***S*** soit l'ensemble de toutes les parties de *U*. Posons $exp\lambda_k = \omega_k = \pi_k^*(1-\pi_k^*)$. On voit que:

$$p(s)=C(\mathbf{S})\prod_s \omega_k=\prod_s \pi_k^*\prod_{U-s}(1-\pi_k^*)$$

ce qui signifie que nous avons obtenu un échantillonnage de Poisson et que $\pi_k^* = \pi_k$.

-Si ***S*** est une famille arbitraire de parties de *U* , on a donc un échantillonnage de Poisson conditionnel à ***S,*** et les π_k sont les probabilités conditionnelles, 'sachant ' que *s* est dans ***S*** . Les π_k^* sont maintenant différents des π

Echantillonnage de Poisson conditionnel à une taille fixe

Le calcul des probabilités conditionnelles est étonnamment simple!

$For\ n = 0 \quad \pi_k^0 = 0$

$For\ \ n = 1 \quad \pi_k^1 = \omega_k \ / \sum_U \omega_l$

$$\pi_k^{n+1} = \frac{C(n)}{C(n+1)} \omega_k (1 - \pi_k^n)$$

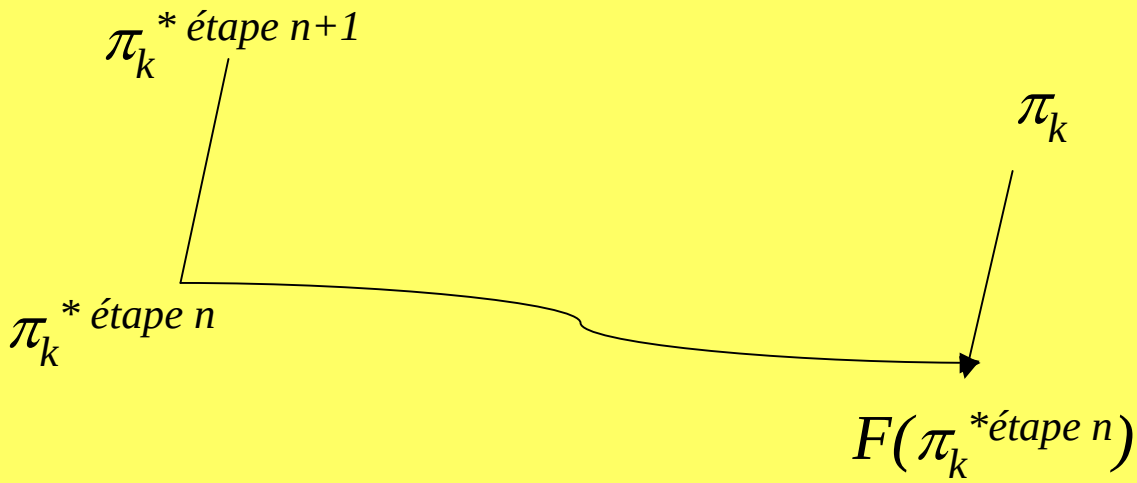
Pour les probabilités du second ordre, c'est à peine plus compliqué . La différence vient de ce que la récurrence fonctionne de *n* à *n+2* :

$$\pi_{kl}^{n+2} = \frac{C(n)}{C(n+2)} \omega_k \omega_l (1 - \pi_k^n - \pi_k^l + \pi_{kl}^n)$$

avec:

$$\pi_{kl}^0 = \pi_{kl}^1 = 0$$

Il reste à calculer les π_k^* à partir des π_k , c'est à dire à inverser la fonction construite ci-dessus . L'algorithme peut s'expliquer graphiquement comme ceci:



Ceci conduit à deux algorithmes de tirage:
-unité par unité (modèle de l'urne)- pas terrible
-séquentiel (plus commode , plus 'souple')

Echantillonnage multidimensionnel

Q échantillons dont on désire contrôler les tailles et les tailles des intersections. Pour deux échantillons, par exemple, on désire avoir des probabilités d'inclusion fixées dans chaque échantillon π^1_k et π^2_k , des tailles fixées $n_{1 \text{ ou } 2}$, et, de plus, contrôler la taille n_3 de l'intersection.

On se ramène à tirer trois échantillons <u>disjoints</u>, de taille n_1-n_3, n_2-n_3 et n_3 avec des probabilités d'inclusion que nous savons calculer (patience, ça va venir).

Une loi de probabilité sur l'ensemble des suites ordonnées de Q parties de U , vérifiant donc ,

$$\sum_{s1,\ldots,s_Q} p(s_1,\ldots,s_Q) = 1 \qquad \text{et} \qquad p(s_1,\ldots,s_Q) \geq 0 \qquad .$$

On définit sans aucune des difficultés techniques qui font le charme pervers de la Théorie des Probabilités, des lois marginales comme

$$p_1(s_1) = \sum_{s2\ldots,s_Q} p(s_1,\ldots,s_Q)$$

à une où plusieurs dimensions et des lois conditionnelles comme

$$p_{1|2}(s_1|s_2) = p_{12}(s_1,s_2) / p_2(s_2)$$

.

Echantillonnage à entropie maximum: vecteurs de probabilités d'inclusion π^i pour $i = 1$ à Q et support $\boldsymbol{S}$

Maximiser :

$$-\sum_{(s1,..,sQ)\in \mathbf{S}} p(s_1,\ldots,s_Q)\, log(p(s_1,\ldots,s_Q))$$

Sous les QxN contraintes:

$$\sum_{si \supset k} p_i(s_i) = \pi^i_k = \sum \mathbf{1}(k \in s_i)\ p_i(s_i)$$

Avec Q N-vecteurs de multiplicateurs λ^i , on trouve:

$$p(s_1,...,s_Q) = C(\mathbf{S})\ \exp(\textstyle\sum_{i=1}^{Q} \lambda^i . s_i)$$

-Supposons que ***S*** soit l'ensemble des suites de Q parties de U.
Posons $exp\lambda^i{}_k = \omega^i{}_k = \pi^i{}_k{}^*(1-\pi^i{}_k{}^*)$. On voit que :

$$p(s_1,...,s_Q) = C(\mathbf{S})\prod_{i=1}^{Q}(\prod_{si}\omega_k^i) = \prod_{i=1}^{Q}\ (\prod_{si}\pi_k^{i\,*}\prod_{U-si}(1-\pi_k^{i\,*}))$$

ce qui signifie que nous avons obtenu Q échantillonnages de Poisson <u>*indépendants*</u> et que $\pi^i{}_k{}^* = \pi^i{}_k$.
On a donc en particulier $p(s_1,...,s_Q) = \prod_{i=1}^{Q}(p_i(s_i))$. Par suite, les plans 'intersection' comme $p(s_1 \cap s_2)$ sont aussi poissonnien pour les probabilités d'inclusion $\pi_k{}^1\pi_k{}^2$.

-Si ***S*** est une famille arbitraire de suites de Q parties de U , on a une famille d'échantillonnages de Poisson indépendants conditionnellement à ***S*** **.** Les $\pi^i{}_k$ sont les probabilités conditionnelles, 'sachant ' que $(s_1,\ldots,s_Q)$ est dans ***S*** . Les $\pi^i{}_k{}^*$ sont maintenant différents des $\pi^i{}_k$.

S est contraint par des tailles fixes $card(s_i)=n_i$ et par le fait que les échantillons sont disjoints.
La somme en k des π^i_k doit donc être égale à n_i .

<u>Calcul des probabilités conditionnelles</u> (étonnamment simple… sur le plan des principes!). Notons $\pi_k^{i;\underline{n}}$ ces probabilités d'inclusion conditionnelles, avec $\underline{n}=(n_1,\ldots,n_Q)$ et $\underline{i}=(0,\ldots,1,..,0)$ le vecteur avec *1* à la $i^{ème}$ position et des zéros ailleurs. A partir de la relation de définition des probabilités d'inclusion, on obtient :

$$\pi_k^{i;\bar{n}}=C_{\bar{n}}\sum_{Dik}\exp\left(\sum_{j=1}^{Q}\lambda^i.s_j\right)=C_{\bar{n}}\omega_k^i\sum_{Eik}\exp\left(\sum_{j=1}^{Q}\lambda^j.s_j\right)=(C_{\bar{n}}/C_{\bar{n}-\bar{i}})\omega_k^i(1-\sum\nolimits_j\pi_k^{j;\bar{n}-\bar{i}})$$

où D_{ik} est l'ensemble des $\underline{n}$-échantillons tels que s_i contienne k ,
et E_{ik} l'ensemble des $\underline{n}$-$\underline{i}$-échantillons dont aucun ne contient k.

De ce fait, on obtient un programme de calcul relativement simple, le plus délicat étant l'initialisation d'une récurrence 'à double coque' et la gestion des multi-indices .

Calcul des π_k^{i*} du tirage multipoissonnien sous jacent (= les $\lambda_k{}^i$)
à partir des π_k^i , c'est à dire à inverser la fonction construite ci-dessus. On normalise la somme des π_k^{i*} à n_i .Comme dans le cas d'un seul échantillon de taille fixe, la matrice des derivées partielles est très proche de l'identité, et la méthode de Newton se simplifie.
De fait, on utilise le programme unidimensionnel ci-dessus en lui donnant comme entrée l'empilement des π_k^i . Rien de neuf donc !

Et les probabilités d'ordre deux, direz vous ?
Même analogie que pour les probas d'ordre un. La forme exponentielle des probabilités du tirage poissonnien conduit à une récurrence de même nature, et toujours de deux en deux :

$$\pi_{kl}^{ij;\bar{n}} = C_{\bar{n}} \sum_{Dik} \exp\left(\sum_{j=1}^{Q} \lambda^i . s_j\right) = C_{\bar{n}} \omega_k^i \omega_l^j \sum_{Eik} \exp\left(\sum_{h=1}^{Q} \lambda^h . s_h\right) = \left(C_{\bar{n}} / C_{\bar{n}-\bar{i}-\bar{j}}\right) \omega_k^i \omega_l^j \left(1 - \sum_h \pi_{kl}^{h;\bar{n}-\bar{i}-\bar{j}}\right)$$

où D_{ik} est l'ensemble des $\underline{n}$-échantillons tels que s_i contienne k et l ,
et E_{ik} l'ensemble des $\underline{n}$-$\underline{i}$-$\underline{j}$-échantillons dont aucun ne contienne k $\underline{\text{ou}}$ l .

Algorithmes de tirage:

L'algorithme séquentiel est le plus naturel et le plus simple à écrire:

Unité 1: elle est attribuée à l'échantillon i avec la proba π_1^i ***(et à aucun d'eux avec la proba*** $1-\Sigma_i \pi_1^i$***) ;***

- Calculer $\pi_k^{i,\ U-1\ ,\underline{n}-\underline{i}}$ ***ou les*** $\pi_k^{i,\ U-1\ ,\underline{n}}$ ***selon le résultat du tirage.***

…

- Unité k+1 : soit $\underline{n}_k = (n_k^i)$ ***le nombre d'unités selectionnées dans l'échantillon i après k 'étapes' (=unités). On calcule les proba*** $\pi_k^{i,\ U-\{1,\ldots k\}\ ,\ \underline{n}-\underline{n}_k}$ ***. L'unité k+1 est attribuée à l'échantillon i avec ces probabilités (où n'est pas attribuée avec la proba complémentaire de leur somme).***

L'algorithme 'draw by draw' n'a pas d'intérêt.

Applications et compléments

a) **Tirage de deux échantillons avec une intersection contrôlée**. Il suffit de tirer trois échantillons. Si les deux échantillons de base sont Poissoniens, le paramètre pour l'intersection sera $\lambda^1 + \lambda^2$. Ceci dit on peut même spécifier les probas d'inclusion dans l'intersection (sauf 0 sinon ça gueule). Ceci dit, pour limiter la portée du miracle, on remarquera que les échantillons marginaux ne sont pas poissonniens. Il en va de même de l'échantillon intersection

b) **Imputation d'une variable qualitative**
Les sommes $\Sigma_k \pi_k^i = n_i$ sont supposées être des entiers et on veut Imputer n_i unités à la modalité i .

c) Quand est-ce que la disjonction des échantillons provoque une covariance négative?

C’EST FINI!

Merci pour votre patience!