

ECHANTILLONNAGE DE RESEAUX, UNE RELECTURE DE S.K.THOMPSON AVEC UNE NOUVELLE PRESENTATION ET QUELQUES NOUVEAUTES

Jean-Claude DEVILLE ()*

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

L'échantillonnage adaptatif de réseaux :

-variable dont les plus fortes valeurs et la diversité se trouvent surtout dans des populations assez rares et groupées au sens d'une certaine notion de voisinage.

-Steven Thompson (jasa 1990)

-Statistique écologique -forestry- :répartition de certaines espèces animales ou végétales

-sociale : usagers de drogues, porteurs du VIH, sans-logis, , immigrés, rupins etc).

-Littérature en Français très limitée

-Une présentation un peu nouvelle et quelques idées associées à de nouveaux résultats.

Structure de la population

- U d'éléments courants notés k, l, \dots .

-Relation de voisinage qui est réflexive et symétrique et donc représentable par un graphe non orienté.

-Chemin : suite d'arcs joignant k et l .

-Composante connexe : tout couple admet un chemin qui les joint

-Distance $\delta(k, l)$: le plus petit nombre d'arcs d'un chemin joignant k et l
(convention : cette distance est infinie pour des éléments appartenant à des composantes connexes différentes).

-On s'intéresse l'estimation du total $Y = \sum_U y_k$ d'une variable numérique (éventuellement vectorielle) y .

-Celle-ci est surtout 'localisée' dans un domaine C caractérisé par une variable binaire $c_k=1$. Par exemple $c_k=1(y_k>5)$.

-Pour la distance δ , C se décompose en composantes connexes notées r et appelées généralement les réseaux (*network*).

-Le bord de r est l'ensemble $r^\circ = \{k : \delta(k,r)=1\}$.

-Les réseaux sont disjoints mais un k donné de $U - C$ peut appartenir aux bords de plusieurs réseaux.

-Enfin on appellera 'grappes' les éléments de la partition de U qui comporte les réseaux et les singletons (ensembles réduits à un élément) de $U - C$.

Un coup d'œil à la figure 1 aidera à mieux interioriser cette structure. Une autre façon consiste à gagner une partie de démineur : le domaine C est l'ensemble des cases où sont les bombes, les bords sont les cases où il y a des chiffres !

X	O																					
O	X	O	1																		O	
	O	X	O								O	6								O	X	
		O								O	X	O								O	X	
				O	2						O		O	8						O	X	
			O	X	O					O	X	O	X	O						10	O	
		O	X	O							O	X	O	O								
		3	O							O		7	O	O	X	O						
								4	O	X	O				O	9						
								O	X	X	X											
								O	X	O	X											
								O	X	X	X											
						5	O	O	O	O	O											
						O	X	X	X	O											11	O
							O	O	O												O	X

Figure 1 : La population compte 10 grappes d'unités de C.

1	2	1	1									1	2	2	1
1	3	3	3	2	1	1						1	1	1	1
	1	1	1	2	1	1						1	2	2	1
1	3	4	3	2	2	2	1		1	1	1			1	1
1	1	1	1		1	1	2	1	2	1	1			1	1
1	2	2	1		1	1	2	1	3	2	2			1	1
1	1					1	2	2	2	1	1			1	1
1	1				1	2	1	2	3	2	2			1	1
2	3	1	2	1	2	1	5	1	4	1	2			1	1
1	3	1	2	1	3	3	1	1	4	1	2				
1	4	2	3	1	2	1	3	2	2	1	1				
1	3	1	2		2	2	2				1	2	2	1	
	2	1	2		1	1	1			1	3	1	1	1	
	1	1	1		1	1	1	1	1	2	1	1	3	1	
			1	2	2	1		1	1	2	3	4	3	1	
			1	1	1	1		1	1	1	1	1	1	1	

Echantillonnage initial et échantillon final

On réalise dans U un échantillonnage de type quelconque :

- n tirages indépendants avec remise à probabilités (de somme 1) p_k égales ou pas.
- Si les tirages sont sans remise on notera comme d'habitude les probabilités d'inclusion par π_k et π_{kl} .
- Tirages régis par une loi de probabilités p sur l'ensemble des échantillons possibles
- On note s_0 l'échantillon tiré, dit initial. Pour tout $k \in s_0$ on obtient y_k et c_k .
- **Si $c_k=1$ on augmente l'échantillon de tous les voisins de k . Ceux-ci peuvent être dans C , et alors on itère la procédure. Si $c_k=0$ on la stoppe après avoir mesuré y_k .**

Si $c_k=1$ on attrape donc tout le réseau de k ainsi que le bord de ce réseau.

On note s l'ensemble des k échantillonnés initialement et des groupes (réseau+ bord) ainsi sélectionnés, autrement dit, s est la réunion de s_0 , des réseaux ayant une intersection non vide avec s_0 et de leur bords.

Certaines de ces unités peuvent être sélectionnées de plusieurs façons différentes ; en particulier les bords de réseaux peuvent contenir des éléments de s_0 . Un coup d'œil à la figure 2 aidera à mieux intérioriser cette mécanique.



La question est maintenant d'utiliser cette information pour estimer le total des y .

On peut évidemment utiliser l'échantillon initial et son estimateur standard \hat{Y}_0 . On se demande alors à quoi bon tout ce cirque.

Une information est relativement facile à utiliser : celle des grappes. On y consacra les points qui suivent tout de suite.

On verra ensuite comment utiliser de façon complète l'information des groupes grâce à la Rao-Blackwellisation.

Pour cela on devra utiliser de façon plus précise la technique d'échantillonnage de l'échantillon initial s_0 . Thompson et ses successeurs ne considèrent que le sondage simple. Nous donnerons de nouveaux résultats relatifs au sondage poissonien et au poissonien réjectif de taille fixe.

Puis, nous regarderons quelques méthodes utiles dans les cas où la Rao-Blackwellisation mène à des calculs trop lourds pour les possibilités habituelles des bécanes du marché aujourd'hui : sous rao-blackwellistion et simulation.

Echantillonnage de grappes : des choses dont on ne parle pas beaucoup !

On ne se préoccupe pas de ce qui précède !

On suppose connue (et non pas trouvée sur la terrain) une partition de U en grappes g

L'échantillonnage initial $p(s_0)$ dans U permet d'attraper pour tout k de s_0 la grappe entière $g(k)$ à laquelle il appartient.

Exemple : échantillonner des ménages à partir d'individus, des aires géographiques à partir d'adresses, etc.

Nous pouvons attraper plusieurs fois la même grappe

Se pose alors la question du choix d'un estimateur améliorant \hat{Y}_0 en utilisant toutes les valeurs de l'échantillon de grappes.

Remarque : Nous nous intéressons à l'estimation de totaux et, en vertu du principe de substitution, de fonctions de totaux. L'estimation de paramètres plus complexes mérite peut-être attention....

Estimateur à la Hansen-Hurvitz

A chaque k on associe $\bar{Y}_{g(k)}$ moyenne de la grappe $g(k)$. $\bar{Y}_{g(k)} = y_k$ si elle est de taille 1.

Les estimateurs 'naturels' suivant sont sans biais :

$$\text{Cas sans remise : } \hat{Y}_{HH} = \sum_{k \in s_0} \bar{Y}_{g(k)} / \pi_k \quad \text{cas avec remise : } \hat{Y}_{HH} = \frac{1}{n} \sum_{i=1}^n \bar{Y}_{g(k_i)} / p_{k_i}$$

La variance et l'estimation de variance sont données par les formules habituelles associées aux plans de sondage initial, c'est-à-dire Horvitz-Thompson (HT) dans le cas sans remise et Hansen-Hurvitz (HH) dans le cas avec .

Par rapport à l'estimateur n'utilisant que s_0 , on a remplacé les y_k par les $\bar{Y}_{g(k)}$.

Dans les deux cas on peut aussi voir le sondage comme portant sur la population G des grappes. En notant s_G l'échantillon de grappes- $s_G = \{g: g = g(k), \text{ pour un } k \in s_0\}$ - on a :

$$\hat{Y}_{HH} = \sum_{g \in s_G} w_g Y_g \quad (Y_g \text{ total de la grappe } g)$$

avec $w_g = \sum_{k \in s_0 \cap g} 1/\pi_k$ pour le cas sans remise et $w_g = \sum_{k \in s_0 \cap g} 1/np_k$ dans le cas avec remise.

Dans le premier cas l'estimateur n'est autre que celui que donne la méthode de partage des poids.

Dans les deux cas on peut encore voir le sondage comme fournissant un échantillon s' (sans remise) réunissant tous les $g(k)$ pour $k \in s_0$. Les estimateurs peuvent s'écrire :

$$\hat{Y}_{HH} = \sum_{k \in s'} w_{g(k)} \bar{Y}_{g(k)} \quad (\text{avec } \bar{Y}_g = Y_g/|g| \text{ moyenne de } y \text{ sur } g).$$

Estimateur à la Horvitz-Thompson

L'estimation à la HT consiste à partir de l'échantillon de grappes s_G et à utiliser l'estimateur de HT après avoir calculé, *si c'est possible*, la probabilité d'inclusion π_g de la grappe g . L'estimateur s'écrit alors :

$$\hat{Y}_{HT} = \sum_{g \in s_G} w_g Y_g \quad \text{soit aussi} \quad \hat{Y}_{HT} = \sum_{k \in s'} w_{g(k)} \bar{Y}_{g(k)} \quad \text{avec} \quad w_g = 1/\pi_g$$

Cette probabilité d'inclusion dépend du plan initial p :

$$\text{-cas SAR : } \pi_g = 1 - (1 - \sum_{k \in g} p_k)^n \quad \text{-cas SAS : } 1 - \binom{N - |g|}{n} / \binom{N}{n}$$

Les probabilités d'inclusion d'ordre deux se calculent par la même méthode :

$$p((g \notin s_G) \text{ et } (h \notin s_G)) = 1 - \pi_g - \pi_h + \pi_{gh} .$$

$$\pi_{gh} = (1 - \sum_{k \in g \cup h} p_k)^n \quad \text{pour le SAR et} \quad 1 - \binom{N - |g| + |h|}{n} / \binom{N}{n} \quad \text{pour le SAS.}$$

L'expression de la variance des estimateurs concernés s'en déduit de même que celle de l'estimateur sans biais de la variance 'naturel'.

Deux cas important pour la pratique semblent n'avoir jamais été envisagés.

-Echantillonnage poissonnien (POIS) : $\pi_g = 1 - \prod_{k \in g} (1 - \pi_k)$.

-Le tirages des grappes *est aussi* poissonnien.

- Le problème du calcul de la variance et de son estimation s'en trouvent réglés sans autre forme de procès.

-Dans un POISCON de support V contenu dans U associé à des probabilités d'inclusion π_k du POISS sous-jacent, on sait calculer de façon récursive les probabilités d'inclusion $\pi_k^{n,V}$ à partir des $\omega_k = \pi_k / (1 - \pi_k)$ (voir [6] [7] ou [8]). Soit $k_1, k_2, \dots, k_{|g|}$ l'ensemble des éléments de g numérotés de façon arbitraire, et $V_i = V_{i-1} - k_i$ pour $i=2$ à $|g|$ avec la convention $V_1=U$. La probabilité d'inclusion de la grappe g vaut : $\pi_g = 1 - \prod_{i=1}^{|g|} (1 - \pi_{k_i}^{n,V_i})$, ce qui se calcule sans trop de mal et avec précision. Pour obtenir les probabilités d'ordre deux, on suit aveuglément et de façon touchante la même méthode que dans le cas simple. On a :

$$p((g \notin s_G) \text{ et } (h \notin s_G)) = 1 - \pi_g - \pi_h + \pi_{gh} \text{ avec } \pi_{gh} = \prod_{i=1}^{|g|+|h|} (1 - \pi_{k_i}^{n,V_i}),$$

où $i=1$ à $|g| + |h|$ numérote les éléments de $g \cup h$, et où les V_i sont construits selon la même convention, soit $V_i = V_{i-1} - k_i$ pour $i=2$ à $|g| + |h|$ avec $V_1=U$.

-Echantillonnage stratifié : échantillonnages dans les strates indépendants par définition! On obtient donc encore assez facilement la probabilité d'inclusion parce que :

$$\pi_g = 1 - \prod_{\text{ens des strates}} (1 - \text{Pr}(\text{la strate coupe } g))$$

et la dernière probabilité de la formule est obtenue par l'un des procédés ci-dessus. Les probabilités d'inclusion d'ordre deux peuvent s'obtenir de façon analogue, bien que ce soit un peu plus sportif.

Le cas d'un échantillonnage initial à plusieurs degrés est plus délicat. En effet, le calcul d'une probabilité d'intersection avec une grappe nécessite de connaître celle de chaque unité primaire, y compris celles qui ne sont pas échantillonnées. Bien que formellement possible, le calcul devient vraiment très compliqué et risque d'être numériquement instable.

Rao-blackwellisation : rappels et exemples/exercices

L'échantillonnage à la HT est fonction de la statistique exhaustive minimale $(s', (y_k, g(k)) ; k \in s')$. Il est donc son propre Rao-Blackwellisé.

Quel est le Rao-Blackwellisé de l'estimateur de type HH. Est-ce, par hasard, l'estimateur de type HT ? On va voir avec les 'petits' exemples qui suivent que c'est rarement le cas. Cette chose s'écrit :

$$E\left(\sum_{k \in s'} w_{g(k)} \bar{Y}_{g(k)} \mid s'\right) = E\left(\sum_{g \in s_G} w_g Y_g \mid s_G\right) \text{ avec } w_g = \sum_{k \in s_0 \cap g} 1/\pi_k$$

$$\text{soit } E(\hat{Y}_{HH} \mid s_G) = \sum_{g \in s_G} E(w_g \mid s_G) Y_g .$$

On va voir tout de suite comment on peut s'en tirer pour le calculer dans des cas assez fréquents. En attendant :

Exemples : p

1 -Population =une grappe g ayant 2 unités et une grappe h en ayant 3.

- s_0 = SAS de taille 2 soit 10 échantillons équiprobables.

-Echantillon de grappes : s' = g une fois, h trois fois et $g \cup h$ six fois.

- On a donc $\pi_g = .7$ et $\pi_h = .3$.

L'estimation à la HT conduit donc à :

$$\begin{array}{ll} Y_g . 10/7 & \text{avec proba } 1/10 \\ Y_h . 10/9 & \text{avec proba } 3/10 \\ \text{et } Y_g . 10/7 + Y_h . 10/9 & \text{avec proba } 6/10. \end{array}$$

Quant à HH il donne (comme tous les π_k valent 2/5) :

$$\begin{array}{ll} Y_g . 10/4 & \text{avec proba } 1/10 \\ Y_h . 10/6 & \text{avec proba } 3/10 \\ \text{et } Y_g . 5/4 + Y_h . 5/6 & \text{avec proba } 6/10 . \end{array}$$

De plus, dans ce cas, l'estimateur à la HH est son propre Rao-Blackwellisé, mais n'est pas l'estimateur à la HT, bien que tous deux soient fonction de la statistique exhaustive minimale. Cela prouve que cette dernière n'est pas complète.

2-Tant qu'on s'amuse, voici deux authentiques exemples de Rao-Blackwellisation.

Même population, mais SAS de taille 3, (probas d'inclusion = 3/5).

Il y a encore 10 échantillons, un égal à h , 3 contenant g et un élément de h , et 6 contenant un élément de g et deux éléments de h . L'estimateur HH donne donc :

$$\begin{array}{ll} Y_h . 15/9 & \text{avec proba } 1/10 \\ Y_g . 10/6 + Y_h . 5/9 & \text{avec proba } 3/10 \\ \text{et } Y_g . 5/6 + Y_h . 10/9 & \text{avec proba } 6/10. \end{array}$$

L'échantillon s' est h avec la proba 1/10 et $g \cup h$ avec la proba 9/10 .

$$\begin{array}{ll} \text{Le Rao-Blackwellisé vaut donc : } & Y_h . 15/9 \quad \text{avec proba } 1/10 \\ & \text{et } Y_g . 10/9 + Y_h . 25/27 \quad \text{avec proba } 9/10 . \end{array}$$

C'est toujours sans biais, of course, mais la variance a baissé.

3-Avec proba inégales :

-On prend deux grappes de 2 , $g=\{1,4\}$ et $h=\{2,3\}$;

- $\pi_k = 0.2$ k avec $\pi_{14}=0.2$ $\pi_{23}=0.2$ $\pi_{24}=0.2$ et $\pi_{34}=0.4$. On a $\pi_g = \pi_h = 0.8$.

Estimateur HT : $Y_g .5/4$ avec proba 0.2
 $Y_h .5/4$ avec proba 0.2
 $(Y_g .+ Y_h).5/4$ avec proba 0.6 .

Sans biais.

Hansen-Huvitz : Y_g 25/8 proba 2/10
 Y_h 25/12 2/10
 $Y_g.5/8 + Y_h .5/4$ 2/10
et $Y_g .5/8 + Y_h .5/6$ 4/10.

HHRao-B : Y_g 25/8 proba 2/10
 $Y_h .25/12$ 2/10)
et $Y_g.5/8 + Y_h 35/36$ 6/10.

C'est sans biais, exotique et de variance réduite.

Echantillonnage adaptatif : généralités sur l'estimation et Rao-Blackwellisation

But : prendre en compte l'information apportée par les unités-frontière dans l'échantillonnage adaptatif, caractère que nous ne connaissons que pour les réseaux capturés dans l'échantillon. Cette information est donc conditionnelle à s .

En particulier, certaines unités de s_0 peuvent être des unités frontières sans qu'on le sache (les réseaux qu'elles voient n'ont pas été tirés) ou être voisines de réseaux tirés (ce dont on s'aperçoit lors de l'extension adaptative de s_0 à s) sans qu'on sache si elles sont voisines d'autres réseaux, non tirés. C'est fondamentalement ce qui empêche de pouvoir calculer une probabilité d'inclusion dans s de ces unités et donc de leur donner un poids de façon naturelle.

Dans une certaine pratique de l'échantillonnage adaptatif on ignore purement et simplement les unités-frontière, et on se contente d'utiliser les données comme un échantillonnage de grappes. On utilise alors l'estimateur \hat{Y}_{HH} ou \hat{Y}_{HT} .

Dans la littérature le traitement de la Rao-Blackwellisation est plutôt sommaire et pédestre. On va tacher d'aller un peu plus loin.

Rao-Blackwellisation

Rappel : Le Rao-Blackwellisé (RB pour faire court) d'un estimateur \hat{T} pour une statistique exhaustive S est $\hat{T}_{RB} = E(\hat{T}|S)$. Si \hat{T} est sans biais, il est également sans biais et de variance inférieure ou égale à celle de \hat{T} . En effet $Var(\hat{T}) = Var(\hat{T}_{RB}) + E(\hat{T} - \hat{T}_{RB})^2$. Si la statistique S est complète, alors on obtient un estimateur sans biais de variance minimale, ce qui est d'ailleurs quasiment tautologique avec la définition de la complétude.

Dans le cas de la statistique de population finie, la statistique exhaustive est $d=(s, \{z_k, k \in s\})$ où s est l'ensemble des unités échantillonnées (sans tenir compte des répétitions éventuelles) et z_k la variable d'intérêt. Dans notre présentation de l'échantillonnage adaptatif $z_k=(y_k, c_k)$, variable dont on cherche le total et identificateur du domaine d'intérêt. Comme on l'a vu, cette statistique n'est pas complète et les estimateurs de type HH et HT donnent en général des RB-isés différents.

L'échantillon s se décompose en trois parties :

$s_c = s \cap C$ caractérisé par $c_k = 1$,

s° caractérisé par $\delta(k, s_c) = 1$

et s_{ex} , composé des n_{ex} éléments (distincts) restant de s_0 , c'est-à-dire vérifiant $\delta(k, s_c) > 1$.

Soit $s'' = s_c + s^\circ$ de taille n'' , l'échantillon de réseaux avec leurs bords et s_r l'ensemble des ν réseaux constituant s_c .

La taille du réseau r sera notée $|r|$ et si $k \in s_c$ $r(k)$ désigne le réseau auquel k appartient.

La RB consiste à utiliser le plan conditionnel $p(s_0 | s)$. Son support est l'ensemble \mathbf{S} des s_0 dont l'extension 'adaptative' redonne s . Il est caractérisé par les conditions suivantes où, par un léger abus de notation, s_0 désigne aussi l'ensemble des unités distinctes de s_0 :

- ($^\circ$) $s_0 \subset s$
- ($*$) $\forall r \in s_r : s_0 \cap r \neq \emptyset$
- ($**$) $s_0 \cap s_{ex} = s_{ex}$.

On a donc $p(s_0 | s) = p(s_0) / p(\mathbf{S})$ si $s_0 \in \mathbf{S}$ et 0 sinon.

On le notera p^* , E^* l'espérance associée, π^* les diverses probabilités d'inclusion associées. I_k^* note l'indicatrice d'appartenance pour ce plan de sorte que $E^*(I_k^*) = \pi_k^*$.

On utilisera soit l'estimateur $\hat{Y}_{HTRB} = E^*(\hat{Y}_{HT})$ soit $\hat{Y}_{HHRB} = E^*(\hat{Y}_{HH})$.

Les estimateurs initiaux sont définis par les formules de l'estimateur du sondage de grappes déjà citées avec $s' = s_0 \cup s_c$.

L'estimateur initial s'écrit aussi :

$$\hat{Y}_{HT \text{ ou } HH} = \sum_{s_{ex}} w_k y_k + \sum_{s_c} w_{r(k)} \bar{Y}_{r(k)} + \sum_{s^\circ \cap s_0} w_k y_k$$

Avec :-Pour le cas HT $w_k = 1/\pi_k$ sur $s_{ex} + s^\circ$ et $1/\pi_{r(k)}$ sur s_c

-Pour le cas HH les poids de s_0 sur $s_{ex} + s^\circ$ et $w_{r(k)} = \sum_{l \in s_0 \cap r(k)} 1/\pi_l$ sur s_c .

Dans ce qui suit on va appliquer cette méthode aux plans SAS, POIS, et, dans une certaine mesure, POISCon. On commencera par le plan SAR qui a droit à un traitement spécifique.

a- Cas du plan avec remise

Cas HT : Cas plutôt facile car c'est déjà le RB de l'estimateur par grappes. La partie $s_{ex}+s_c$ reste inchangée.

Pour s° on remarque que le plan conditionnel est un SAR(n) sur s avec la contrainte d'une observation au moins sur chaque élément de s_{ex} et de s_r (conditions (*) et (**)) ci-dessus). Reste donc $m=n -n_{ex}-v$ tirages indépendants libres de tomber où ils veulent dans s avec probabilités $p'_k = p_k / \sum_s p_l$. La probabilité d'une observation au moins en k vaut donc $1 - (1 - p'_k)^n$. Comme le poids, sous le plan p , d'une observation tombant dans s° serait de $1/(1 - (1 - p_k)^n)$, le poids de chaque unité de s° dans $\hat{Y}_{HTRB,SAR}$ sera égal à $:(1 - (1 - p'_k)^n)/(1 - (1 - p_k)^n)$.

Cas HH : Le plan conditionnel est bien sur le même soit un SAR(n). La difficulté supplémentaire est que nous avons à calculer l'espérance sous ce plan du nombre d'observations tombant en chaque k de s .

La loi de proba des n_k (k dans $s-s_c$) et n_r est donc une multinomiale conditionnée par $\forall k \in s_{ex} n_k \geq 1$ et $\forall r \in s_r n_r \geq 1$.

C'est un calcul que je sais faire (par des méthodes récursives) mais trop compliqué pour que je l'explique ici.

Remarque préliminaire pour les p sans remise ; on a :

$$\hat{Y}_{HTRB} = \sum_{s_{ex}} y_k / \pi_k + \sum_{s_c} y_k / \pi_{r(k)} + \sum_{s^\circ} y_k \pi_k^* / \pi_k.$$

b- Cas du plan poissonnien

Tout est assez simple. Le plan conditionnel à s est tout simplement un POISS gardant les mêmes probabilités d'inclusion π_k mais de support s et vérifiant toujours les conditions (*) et (**).

Cas HT : La partie $s_{ex}+s_c$ reste inchangée.

Pour s° , le poids si k est dans s_0 est $1 / \pi_k$ mais la proba d'inclusion conditionnelle vaut aussi π_k de sorte que le poids dans $\hat{Y}_{HTRB,POISS}$ vaut tout simplement 1 !

Cas HH : La partie s_{ex} reste inchangée. Par contre le poids d'un réseau devient

$$E \left(\sum_{k \in s_0 \cap r} 1 / \pi_k \mid n_r \geq 1 \right) = |r| / \pi_r$$

de sorte que la partie s_c demeure elle aussi inchangée et que $\hat{Y}_{HHRB,POISS} = \hat{Y}_{HTRB,POISS}$, ce qui peut éviter certaines angoisses métaphysique sur la choix de l'estimateur.

c- Cas du plan aléatoire simple de taille n_0

Le plan conditionnel à s est un plan simple de taille n_0 , de support s et vérifiant les conditions (*) et (**). La seconde implique que s_{ex} est sélectionné avec la probabilité 1. Reste donc à trouver les probabilités d'inclusion d'un plan simple de taille $n=n_0 - n_{ex}$ sur $s''=s^\circ+s_c$ ('population' de taille n'') conditionné par (*) qui s'écrit de façon équivalente :

$$(*) \quad \forall r \in s_r : n_r \geq 1$$

Si on est capable de calculer la probabilité $q_n(*)$ de cet événement pour tout n , on voit que la probabilité d'inclusion π_k^* d'un élément de s° vaut $\pi_k^* = \binom{n''-1}{n-1} q_{n-1}(*) / \binom{n''}{n} q_n(*)$. Quand à $q_n(*)$, dans le cas simple où $\nu = 1$, c'est la même chose que la probabilité d'inclusion du réseau, si $\nu = 2$, c'est la proba d'ordre 2, etc. De façon générale soit σ une partie de s_r et $q_\sigma = \binom{n'' - \sum_\sigma |r|}{n} / \binom{n''}{n}$ la probabilité que s'' aie une intersection vide avec $\sum_\sigma r$.

On a alors le résultat à partir de la formule sommatoire 'inclusion-exclusion' :

$$q_n(*) = 1 + \sum_{\{\sigma\}} q_\sigma (-1)^{|\sigma|} \text{ soit } 2^\nu \text{ termes.}$$

Bien que possible le calcul devient vite très lourd, s'il reste même numériquement faisable vu le caractère explosif de la combinatoire sous-jacente.

Cas HT : La partie $s_{ex}+s_c$ reste inchangée. Les poids des éléments de s° sont égaux à $N\pi_k^*/n$.

Le calcul est donc théoriquement possible assez facilement.

Cas HH : Les poids des éléments du réseau r sont donnés par $E^*(n_r^*) N/n$. Le calcul de cette expression est analogue à ce qui vaut pour les éléments de s° . En effet $E^*(n_r^*) = |r| \pi_k^*$ où π_k^* est la probabilité d'inclusion commune des éléments de r . On se rend compte assez facilement qu'elle vaut :

$$\pi_k^* = \binom{n''-1}{n-1} q_{n-1}(*-r) / \binom{n''}{n} q_n(*)$$

où $q_{n-1}(*-r)$ est la probabilité pour que le SAS($n''-1, n-1$) ait une intersection non vide avec chaque réseau de s_c-r .

Deux petits exemples : On veut tirer un échantillon de taille 3 dans un ensemble à 6 éléments, 1 dans s° , 2 dans un premier réseau et 3 dans un second.

Il y a 15 échantillons possibles, la probabilité d'inclusion dans s° vaut $2/5$ ($< 1/2$) , $3/5$ dans le petit réseau ($E^*(n_r^*) = 1.2$) et $7/15$ dans le gros ($(E^*(n_r^*) = 1.4)$. On vérifie que la somme des probas d'inclusion fait bien 3 comme de juste.

Si on recommence avec 2 éléments dans chaque ensemble, il y a 12 échantillons, la probabilité d'inclusion dans s° vaut $1/3$ et $7/12$ sur les quatre unités des réseaux (ce qui est assez exotique !).

d- Cas du plan poissonnien conditionnel de taille fixe

Le même type de méthodes que celles utilisées pour le SAS sont possibles. Cependant, chaque coefficient binomial doit être remplacé par une somme ayant le même nombre de termes que ce coefficient. Ce terme indice une partie s^{**} à n ou $n-1$ éléments, selon le cas et vaut , avec $\omega_k = \pi_k / (1 - \pi_k)$ (π_k probabilité d'inclusion du poissonnien non contraint sous-jacent)) $\prod_{k \in s^{**}} \omega_k$.

Disons simplement que le traitement de l'estimateur HT est possible si le nombre de réseaux (de taille supérieure à 2 naturellement) ne dépasse pas trois ou peut être quatre. Au delà la combinatoire me semble devenir assez démentielle. Néanmoins cela donne une piste pour le calcul, et il est peut-être possible de trouver des astuces qui le simplifient.

Encore un petit exemple : Dans le deuxième exemple du 4-4 (2+2+2), on colle maintenant des probabilités inégales (du poissonnien sans aucune contrainte, pour ne pas trop compliquer le bazar) données par le tableau suivant :

s°		$r2$		$r3$	
1	0.1	3	0.4	5	0.6
2	0.4	4	0.6	6	0.9

Les ω_k sont donnés au facteur 1/18 près par :

s°		$r2$		$r3$	
1	2	3	12	5	27
2	12	4	27	6	162

Après quelques calculs tout à fait automatisables (en Matlab par exemple), on trouve les probas conditionnelles : (en dix-millièmes)

s°	r_2	r_3
1 044	3 434	5 579
2 264	4 749	6 930

Les espérances utiles sont $E^*n_{r_2}=1,183$ et $E^*n_{r_3}=1,509$.

Un petit dernier pour la route : on part de :

s°		r_2		r_3	
1	0.8	3	0.6	5	0.4
2	0.6	4	0.4	6	0.2

On trouve :

s°		r_2		r_3	
1	654	3	716	5	775
2	245	4	360	6	250

Les espérances utiles sont $E^*n_{r_2}=1,076$ et $E^*n_{r_3}=1,025$.

Conclusion pour cette partie

L'utilisation de la RB semble nécessaire quand s° est gros par rapport à s_c , ce qui dépend de la structure de la population (nombre de voisins par unité par exemple), du plan p initial...et de l'échantillon s auquel on arrive.

Cela plaide pour l'utilisation d'un plan initial poissonnien pour lequel tous les calculs sont simples.

Un autre avantage est de pouvoir utiliser des probabilités variables ce qui s'avère souvent utile dans les applications de la méthode d'échantillonnage adaptative.

L'inconvénient unique est qu'on ne contrôle pas exactement la taille de l'échantillon initial. Ceci dit, la taille de l'échantillon final est toujours aléatoire et sa variabilité doit être généralement supérieure à celle du poissonnien (qui est relativement faible).

Et maintenant quelques idées pour éviter tout ça !

Sous Rao-Blackwellisation

Une idée astucieuse est utilisée dans Dryver-Chao (cf biblio) : prendre l'espérance des estimateurs conditionnés par une statistique exhaustive non minimale à savoir : $d'=(s, y_k, c_k ; k \in s \text{ et } \{n_r, r \in s_r\})$.

Sous ce conditionnement les calculs deviennent simples. Par exemple l'échantillonnage conditionnel dans s° , si p est un poisson conditionnel de taille n (donc en particulier un SAS) est un poisson conditionnel de taille n_{s° , nombre d'unités de s_0 qui sont tombées dans s° . C'est une importante restriction du nombre des échantillons possibles dans la RB, mais peut-être pas si grave que ça pour la précision de l'estimation.

Une autre idée ... si la structure de s s'y prête ! Utiliser les composantes connexes de s'' et sous-RB-iser conditionnellement à la taille de l'échantillon initial dans chaque composante connexe. La condition est moins restrictive que la précédente mais demande, pour être applicable, que le nombre de réseaux dans chaque composante soit relativement faible pour que les difficultés signalées il y a une minute soient surmontables.

L'idée générale de ce type de méthode est la suivante.

On ajoute à la statistique exhaustive minimale d une information supplémentaire $Inf(s)$ (par exemple $\{n_r \mid r \in s_r\}$), qui permet de calculer plus facilement une espérance conditionnelle $E(\hat{Y} \mid d, Inf(s))$ et qui a un biais conditionnel. L'espérance de ce biais est nulle (puisque \hat{Y} est supposé sans biais), mais sa variance reste incluse dans celle de l'estimateur final.

Estimation par simulation

Bien que naturelle et dans l'air du temps, cette méthode semble ne jamais avoir été évoquée.

Le plan conditionnel à s est connu sans la condition (*) est entièrement connu et ressemble à l'échantillonnage initial.

C'est dans tous les cas quelque chose de simple et manipulable sans la condition (*) qui pose des problèmes de calcul et d'algorithmique. On peut s'en affranchir en tirant de façon répétée des échantillons dans ce plan conditionnel et en ne conservant que ceux qui vérifient la condition (effectif noté n^*). Pour chacun d'eux on peut calculer une 'pseudo-estimation', puisque y est observée sur tout s .

L'estimateur 'simulé' est la moyenne des n^* 'pseudo-estimations' obtenues.

Il est évidemment sans biais dès que l'estimateur initial l'est.

Variance : si où \hat{T} est n'importe lequel des estimateurs initiaux envisagés, on a :

$$Var(\hat{T}) = Var(\hat{T}_{RB}) + E(\hat{T} - \hat{T}_{RB})^2 .$$

L'estimateur simulé pour variance : $Var(\hat{T}_{RB}) + E(\hat{T} - \hat{T}_{RB})^2/n^*$.

On se rapproche donc très vite de $Var(\hat{T}_{RB})$: dès le deuxième pseudo-échantillon obtenu on a fait la moitié du chemin (et même dès le premier si on utilise des probabilités égales car s_0 peut être utilisé dans ce cas) ! Enfin $E(\hat{T} - \hat{T}_{RB})^2$ s'estime sans biais par $\sum_{i=1}^{n^*} (\hat{T}_i^* - \bar{\hat{T}})^2 / (n^* - 1)$ d'où une estimation sans biais de la variance de l'estimateur.

La difficulté peut venir du caractère 'rare' de l'événement (*). Une solution peut venir d'une implémentation du tirage où on commence par tirer les effectifs tombant dans les r et dans s^0 . Dans certains cas, en effet on arrive à simuler cette loi et conditionnellement aux effectifs on sait se débrouiller.

Diverses choses et conclusion

Cet exposé n'est qu'une introduction à la statistique de l'échantillonnage adaptatif de réseaux. Certains trucs y ont été présentés de façon très incomplète et rapide. Ils méritent de plus amples développements. Par exemple les échantillonnages poissoniens conditionnels n'ont pas livré tous leurs secrets !

Autres sujets de méditation sont possibles :

- Les questions de calage et d'usage d'informations auxiliaires.
- Les questions liées à la non-réponse.
- Variance et estimation de variance. Les cas simples sont néanmoins résolus de façon complète.
- Lien entre RB et bootstrap.
- Les échantillonnages initiaux à plusieurs degrés ou en deux phases.
- Déjà pas mal traité dans la littérature, l'échantillonnage où les voisins visités sont sélectionnés de façon aléatoire.
- On n'a pas de hiérarchie bien nette entre les estimateurs (HH ou HT ?)
- Le domaine des sous-RB est-il très ouvert ?
- Approche par le modèle.
-

Bref il s'agit d'un domaine de recherche assez passionnant et dont l'utilité est loin d'être négligeable.

**MERCI POUR VOTRE
PATIENCE !**

I love you madly

comme disait qui déjà...