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

Introduction

L'échantillonnage adaptatif de réseaux est utilisé pour étudier des variables 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. Il a été introduit en 1990 par Steven Thompson [3] et développé dans le livre [4] qui en a détaillé les idées et les exemples. Il s'est avéré très utile dans des questions de statistique écologique (répartition de certaines espèces animales ou végétales) où sociale (usagers de drogues, sans-logis, rupins, immigrés, etc). Malheureusement la littérature en Français est très limitée : quelques traductions dans Technique d'Enquête, un article de David Lévy aux JMS de 2009 [2] et peut-être quelques utilisations diverses où la méthode statistique est mal expliquée, et peut-être pas très bien comprise.

Le but de cet article est d'introduire ce type de méthodes que j'ai découvert il y a quelques mois, en utilisant une présentation un peu nouvelle et de donner quelques idées associées à de nouveaux résultats.

1. Structure de la population

On s'intéresse à un population finie U d'éléments courants notés $k,l,\ldots$. Cette population est munie d'une relation de voisinage (ou d'analogie chez Bourbaki) qui est réflexive et symétrique et donc représentable par un graphe non orienté. Cette relation induit de façon naturelle sur U des notions de chemin, de composante connexe et une distance à valeurs entières qu'on notera $\delta(k,l)$ (avec la convention que cette distance est infinie pour des éléments appartenant à des composantes connexes différentes). Elle est définie par le plus petit nombre d'arcs d'un chemin joignant k et l.

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 dans [3] $c_k=\mathbf{1}(y_k>5)$ .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 $r°=\{k : \delta(k,r)=1\}$. 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 intérioriser 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 !

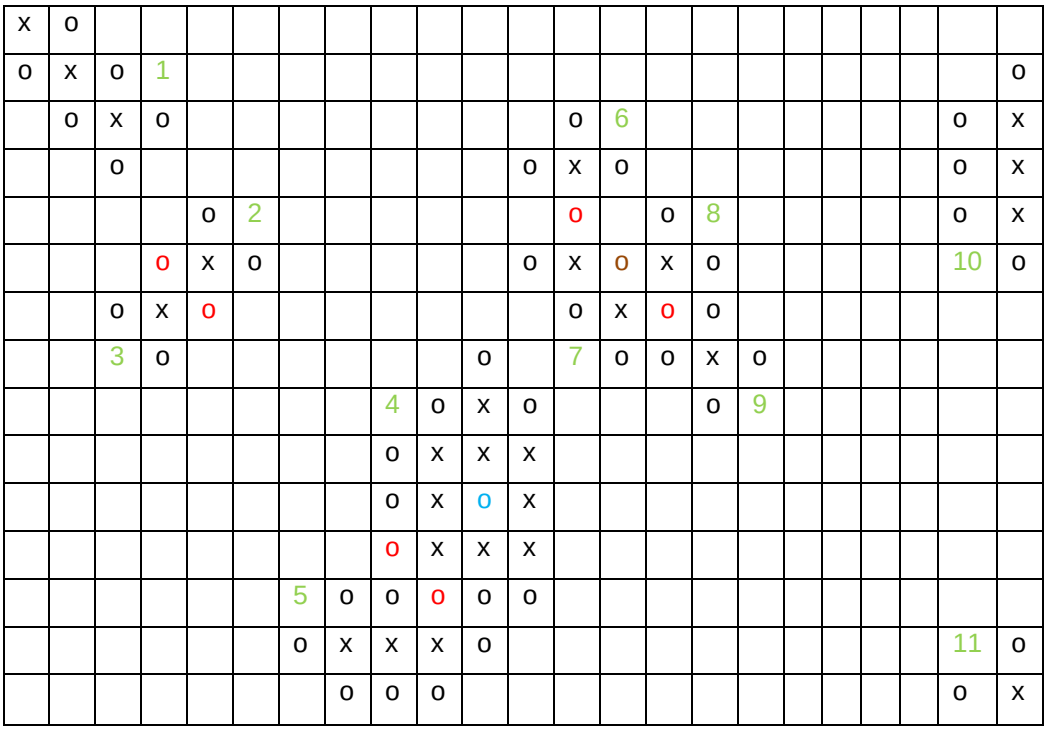

Figure 1 : La population compte 10 grappes d'unités de *C*.

2. Echantillonnage initial et échantillon final

On réalise dans *U* un échantillonnage de type quelconque. Pour faire comme dans [3] ou [4] , on s'intéressera aussi à des tirages indépendants avec remise de taille fixe à probabilités (se sommant à 1) p_k égales ou pas. Si les tirages sont sans remise (le cas habituel de la théorie des sondages) on notera comme d'habitude les probabilités d'inclusion par $\pi_k \; et \; \pi_{kl}$. On pourrait aussi s'intéresser à des tirages séquentiel, mais ne chargeons pas trop. Ces tirages sont régis par une loi de probabilités *p* sur l'ensemble des échantillons possibles et on notera s_0 l'échantillon tiré, dit initial. Pour tout $k \in s_0$ on obtiendra les valeurs de y_k et de 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 .

Il est clair que si $c_k=1$ on intègre dans l'échantillon 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+ son 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 consacrera les deux paragraphes suivant. 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 poissonnien 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.

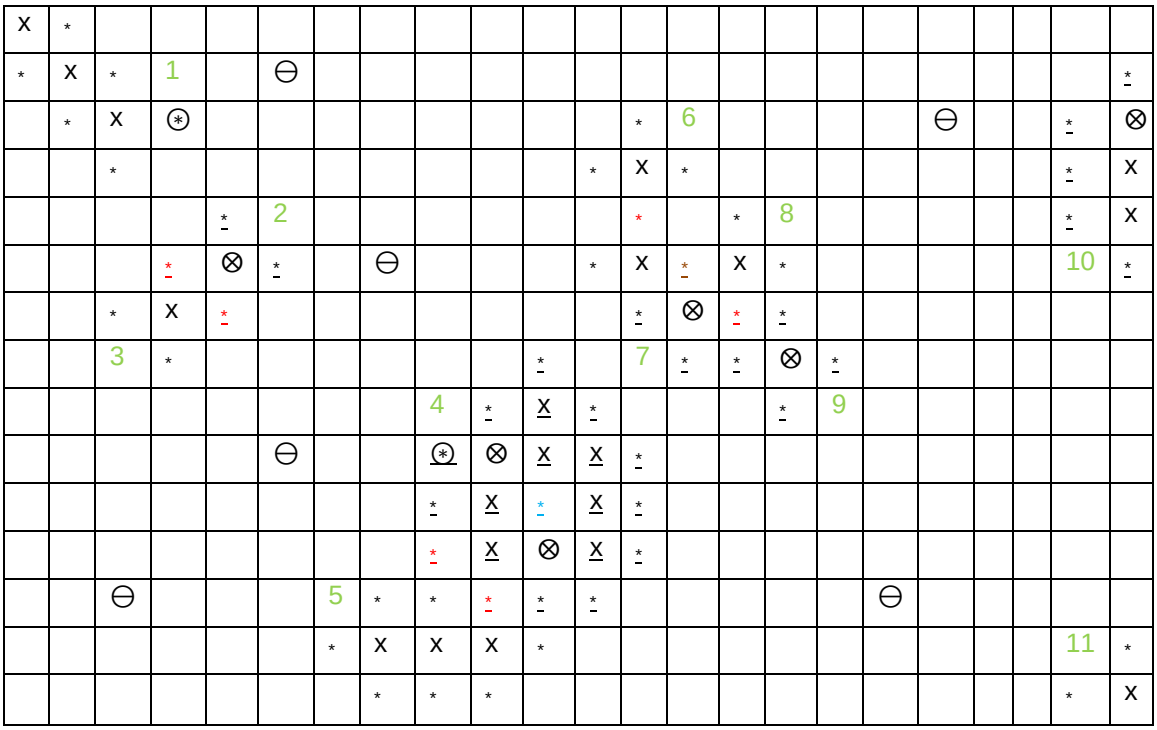

Figure 2 : Les cases avec des ronds sont l'échantillon initial. Ils entourent un X si une unité du domaine privilégié est tiré, un – sinon. Les unités frontières échantillonnées sont soulignées. Elles sont en rouge si elles appartiennent aux bords de plusieurs réseaux.

3. Echantillonnage de grappes

On va traiter d'abord l'échantillonnage de grappes en soi, sans trop se préoccuper de ce qui précède. Autrement dit on s'intéresse au cas où on connait la partition de *U* en grappes *g* et où 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. On ne parle pas pour le moment d'unité-frontière ni de distance. Exemple : échantillonner des ménages à partir d'individus, des aires géographiques à partir d'adresses, etc.

Comme dans le cas du sondage adaptatif, le point est que nous pouvons souvent attraper plusieurs fois la même grappe soit à partir d'individus différents dans le cas sans remise, soit, en plus, à partir de tirages différents dans le cas avec remise en ayant tiré le même individu plusieurs fois. 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….

3.1. Estimateur à la Hansen-Hurvitz

A chaque unité *k* on associe $\bar{Y}_{g(k)}$ moyenne de *y* dans la grappe *g(k).* Remarquons que si la grappe est de taille 1, $Y_{g(k)} = y_k$. Les estimateurs 'naturels' suivant sont sans biais :

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

Sous cette forme 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. La dénomination 'à la HH' vient du fait que, dans les deux cas, la même grappe peut être sélectionnée plusieurs fois si des *k* appartenant à la même grappe sont tirées dans s_0. Dans le cas avec remise, de plus, le même *k* peut être sorti plusieurs fois.

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- $s_G = \{g : g = g(k), pour\ un\ k \in s_0\}$ - on a :

$$\hat{Y}_{HH} = \sum_{g \in s_G} w_g Y_g \qquad (Y_g \text{ total dans } g)$$

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 :

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

3.2 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, bien sur, $\hat{Y}_{HT} = \sum_{g \in s_G} w_g Y_g$ soit aussi $\hat{Y}_{HT} = \sum_{k \in s'} w_{g(k)} \bar{Y}_{g(k)}$ avec $w_g = 1/\pi_g$ .Cette probabilité d'inclusion dépend du plan initial *p*. Il est facile de voir que dans le cas avec remise (SAR) elle vaut $\pi_g = 1 - (1 - \sum_{k \in g} p_k)^n$. Dans le cas du SAS elle vaut $1 - \binom{N - |g|}{n} / \binom{N}{n}$.

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

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

La probabilité π_{gh} vaut $(1 - \sum_{k \in g \cup h} p_k)^n$ dans le cas SAR et $1 - \binom{N - |g| + |h|}{n} / \binom{N}{n}$ dans le cas 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'.

Ces résultats figurent dans [4].

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

Le cas d'un échantillonnage initial <u>poissonnien</u> (POIS) associé à des probabilités d'inclusion π_k est particulièrement simple : $\pi_g = 1 - \prod_{k \in g}(1 - \pi_k)$. Du fait de l'indépendance des tirages individuels, les tirages de grappes sont eux aussi indépendants, autrement dit l'échantillonnage de grappes est encore poissonnien. Le problème du calcul de la variance et de son estimation s'en trouvent réglés sans autre forme de procès.

Pour le <u>poissonnien conditionnel</u> à une taille fixée *n* (POISCOn) et de support *V* contenu dans *U,* associé à des probabilités d'inclusion π_k , on sait calculer de façon récursive les probabilités d'inclusion $\pi_k^{n,V}$ a partir des $\omega_k = \pi_k/(1 - \pi_k)$ (voir [6] [7] ou [8]). Soit $k_1,k_2,..,k_{|g|}$ 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 :

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

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

En ce qui concerne un échantillonnage <u>stratifié</u> avec dans les strates des échantillonnages (indépendants par définition !) des types précédents, on obtient encore assez facilement la probabilité d'inclusion parce que : $\pi_g = 1 - \prod_{ens\ des\ strates}(1 - \Pr\,(la\ strate\ coupe\ g))$ 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 <u>plusieurs degrés</u> 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.

3.3 Rao-blackwellisation : rappels et exemples/exercices

L'échantillonnage à la HT est fonction de la statistique exhaustive minimale (*s',*$(y_k$ *,g(k)) ;k∈s'*). Il est donc son propre Rao-Blackwellisé.
On peut se demander, si on est un peu curieux, 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 : $E\big(\sum_{k\in s'} w_{g(k)}\bar{Y}_{g(k)}\,\big|s'\big) = E\big(\sum_{g\in s_G} w_g Y_g\,\big|s_G\big)$ avec $w_g = \sum_{k\in s_0\cap g} 1/\pi_k$
soit $E\big(\hat{Y}_{HH}\big|s_G\big) = \sum_{g\in s_G} E\big(w_g\big|s_G\big)Y_g$

On verra au §4 (entre autres) comment on peut s'en tirer pour le calculer dans des cas assez fréquents.

En attendant, à titre <u>d'exemple</u>, considérons une population constituée d'une grappe *g* comptant 2 unités et d'une grappe *h* qui en compte 3. L'échantillon s_0 est un SAS de taille 2. On a donc au total 10 échantillons équiprobables. L'échantillon de grappes est *s'= g* une fois, *h* trois fois et $g \cup h$ six fois.
On a donc $\pi_g = .7$ et $\pi_h = .9$. L'estimation à la HT conduit donc à :

Y_g *.10/7*	*avec proba 1/10*
Y_h *.10/9*	*avec proba 3/10*
et Y_g *.10/7 +* Y_h *.10/9*	*avec proba 6/10.*

On vérifie qu'il est sans biais bien qu'il ne donne jamais la vraie valeur !

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

Y_g *.10/4*	*avec proba 1/10*
Y_h *.10/6*	*avec proba 3/10*
et Y_g *.5/4 +* Y_h *.5/6*	*avec proba 6/10 .*

On vérifie qu'il est sans biais bien qu'il ne donne jamais la vraie valeur !
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.
Tant qu'on s'amuse, voici <u>deux authentiques exemples</u> de Rao-Blackwellisation.
La première est le plus petit exemple qu'on puisse construire avec un SAS. On garde la même population, mais la taille du SAS est maintenant de 3, et les probas d'inclusion valent 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 :

Y_h *.15/9*	*avec proba 1/10*
Y_g *.10/6 +* Y_h *.5/9*	*avec proba 3/10*
et Y_g *.5/6 +* Y_h *.10/9*	*avec proba 6/10.*

C'est sans biais, of course, bien la vraie valeur soit impossible, mais on commence à être habitué.
L'échantillon *s'* est *h* avec la proba 1/10 et $g \cup h$ avec la proba 9/10 . Le Rao-Blackwellisé vaut donc :

Y_h *.15/9*	*avec proba 1/10*
et Y_g *.10/9 +* Y_h *.25/27*	*avec proba 9/10 .*

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

Avec proba inégales, on prend deux grappes de 2 , *g={1,4}* et *h={2,3}* ;pour faire simple $\pi_k = 0.2\ k$ avec π_{14}=0.2 π_{23}=0.2 π_{24}=0.2 et π_{34}=0.4. On a $\pi_g = \pi_h = 0.8$.

Estimateur HT : Y_g *.5/4 ou* Y_h *.5/4* avec proba *0.2 ,* (Y_g *.+* Y_h *).5/4 avec proba 0.6* .Sans biais.

HH : 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.

Remarque : Ces trois exemples sont libres de droits et peuvent servir d'exercices pour des cours de sondage pas trop nuls. Ca serait bien pourtant sympa d'en citer l'origine, contrairement à ce qui s'est fait dans certains recueils d'exercices bien connus dont les droits d'auteurs sont protégés.

4 Echantillonnage adaptatif : Rao-Blackwellisation

Le but est maintenant de prendre en compte l'information apportée par les unités-frontière dans l'échantillonnage adaptatif. Remarquons que nous ne connaissons ce caractère frontalier que pour ce qui concerne 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 voisinent n'ont pas été tirés) où 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 $\hat{Y}_{HT}$ .David Lévy [2] n'utilise que l'estimateur à la HT de grappes sans évoquer une possible Rao-Blackwellisation. Il est vrai que celle-ci, dans la cas qu'il étudie, reviendrait à ajouter un estimateur égal à 0. Dans [3] et [4], le traitement de la Rao-Blackwellisation est sommaire. 0n ne donne aucune méthode explicite de pondération même dans les cas élémentaires. On va tacher d'aller un peu plus loin.

4.1 Rao-Blackwellisation

Un peu de rappels. 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*∈*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 en 3-3, 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*∩*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°* de taille *n"* , l'échantillon de réseaux avec leurs bords et s_r l'ensemble des v réseaux constituant s_c . La taille du réseau *r* sera notée |r| et si *k*∈s_c *r(k)* désigne le réseau auquel *k* appartient.

La RB consiste à utiliser le plan conditionnel $p(s_0 \mid s)$. Son support est l'ensemble $\boldsymbol{S}$ des s_0 dont l'extension 'adaptative' redonne s. On voit facilement que $\boldsymbol{S}$ 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 :

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

Le plan $p(s_0 \mid s)$ vérifie donc $p(s_0 \mid s) = p(s_0)/p(\boldsymbol{S})$ si $s_0 \in \boldsymbol{S}$ et 0 sinon. On le notera p^* , E^* l'espérance associée, π^* les diverses probabilités d'inclusion associées. I^*_k désigne l'indicatrice d'appartenance pour ce plan de sorte que $E^*(I^*_k) = \pi^*$. On utilisera donc, si le calcul est possible, soit l'estimateur $\hat{Y}_{HTRB} = E^*(\hat{Y}_{HT})$ soit $\hat{Y}_{HHRB} = E^*(\hat{Y}_{HH})$ où les estimateurs initiaux sont définis par les formules du §3.1 et 3.2 avec $s' = s_0 \cup s_c$ .De façon plus adaptée, nous pouvons écrire que l'estimateur initial s'écrit aussi :

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

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 .

Dés que p est <u>sans remise</u> on a donc :

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

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.

4.2 Cas du plan avec remise

<u>Cas HT</u> : C'est le cas le plus facile car c'est déjà le RB de l'estimateur par grappes. La partie $s_{ex}+s_c$ reste inchangée. Pour s° on doit regarder de plus prés le plan conditionnel. C'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 $m = n - n_{ex} - \nu$ 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, <u>sous le plan p</u> , 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 :$(1 - (1 - p'_k)^n)/(1 - (1 - p_k)^n$. Bon, c'était presque ce qu'il y a de plus simple.

<u>Cas HH</u> : Ca ressemble un peu : le plan conditionnel est un SAR(n) avec la même condition que ci-dessus. La difficulté supplémentaire est que nous avons à calculer l'<u>espérance</u> 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 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.

4.3 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 (**) .

<u>Cas HT</u> : La partie $s_{ex}+s_c$ reste inchangée (sa proba d'inclusion conditionnelle vaut 1 comme toujours dans un cas HT). Pour la partie s°, le poids si k est sélectionné 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* !

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

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

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.

4.4 **Cas du plan aléatoire simple de taille *n***

Le plan conditionnel à s est un plan simple de taille *n*, 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 sur s''=$s°$+s_c conditionné par (*) qui s'écrit de façon équivalente :

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

Cette condition est assez difficile à manier. Toutefois, 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 $q_\sigma = \binom{n'' - \sum_\sigma |r|}{n} \Big/ \binom{n''}{n}$ 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 : 0n 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. Si on ne veut pas appliquer les formules il suffit de compter. 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 !).
Remarque : La même qu'à la fin du 3-3.

4.5 Cas du plan poissonnien conditionnel de taille fixe

On va se contenter de quelques indications. 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°	*r2*	*r3*
1 044	3 434	5 579
2 264	4 749	6 930

Les espérances utiles sont $E*n_{r2}=1,183$ et $E*n_{r3}=1,509$.

Un petit dernier pour la route : on part de :

s°	*r2*	*r3*
1 0.8	3 0.6	5 0.4
2 0.6	4 0.4	6 0.2

On trouve :

s°	*r2*	*r3*
1 654	3 716	5 775
2 245	4 360	6 250

Les espérances utiles sont $E*n_{r2}=1,076$ et $E*n_{r3}=1,025$.
Remarque : La même qu'à la fin du 4-4.

4.6 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, 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).

5 Sous Rao-Blackwellisation

Une idée astucieuse est utilisée dans [5]. Elle consiste à prendre l'espérance des estimateurs conditionnés par une statistique exhaustive non minimale à savoir : $d'=(s\,, y_k, c_k\,; k \in s \;\; et \; \{n_r \; r \in s_r\})$. 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 peut permettre de limiter les volumes de calculs si la structure de *s* s'y prête. Elle consiste à utiliser les composantes connexes de *s''* et à 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 au 4-5 soient surmontables. Nous n'entrerons pas ici dans les détails.

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 $\{n_r\ r \in s_r\}$), qui permet de calculer plus facilement une espérance conditionnelle $E\left(\hat{Y}\middle|d, Inf(s)\right)$ 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.

6 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 comme nous l'avons vu assez facile à manier. C'est la condition (*) qui pose des problèmes de calcul et d'algorithmique. On peut s'en affranchir en tirant de façon répétée des échantillons de ce type et en faisant la moyenne des 'pseudo-estimations' ainsi obtenues.

Par ailleurs la variance de l'estimateur initial est connue et estimable. Or $Var(\hat{T}) = Var(\hat{T}_{RB}) + E(\hat{T} - \hat{T}_{RB})^2$ où $\hat{T}$ est n'importe lequel des estimateurs initiaux envisagés. Si on dispose de *n** 'pseudo-estimations' , leur moyenne a pour variance $Var(\hat{T}_{RB}) + E(\hat{T} - \hat{T}_{RB})^2/n^*$ et 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 $\sum_{i=1}^{n^*}\left(\hat{T}_i^* - \bar{\hat{T}}\right)^2/(n^* - 1)$ 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. Conditionnellement aux effectifs on sait se débrouiller comme au § 5.

7 Diverses choses et conclusion

Ce texte est une introduction à la statistique de l'échantillonnage adaptatif de réseaux. Certaines techniques y ont été présentées de façon très incomplète et rapide. Elles méritent de plus ample développements. Par exemple les échantillonnages poissoniens conditionnels n'ont pas encore livré tous leurs secrets !
D'autres sont possibles :

- Les questions de calage et d'usage d'informations auxiliaires. Il semble a priori que cela ne soulève pas beaucoup de nouveaux problèmes cependant. Voir éventuellement [5] .
- Les questions liées à la non-réponse. Sans doute quelque chose d'un peu plus épineux.
- 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.

Bibliographie (provisoire)

[1] Dryver A.L., Thompson S.K., « Improving Unbiased estimators in adaptive cluster sampling», *Journal of the Royal Statistical Society: Series B (Statistical Methodology),*vol 67, Issue 1, pp 157–166, february 2005.
[2] Levy D, « Une méthode pour adapter l'échantillonnage au cours de la collecte afin de mieux répondre à un objectif », *JMS Insee,* 2009.
[3] Thompson S.K., «Adaptive Cluster Sampling», *Journal of the American Statistical Associat*ion, Vol. 85, n° 412), pp. 1050-1059, december 1990.
[4] Thompson S.K., Seber G.A.F, « Adaptive Sampling» , Wiley 1996.
[5] Dryver, A.L. and Chao, CT, « Ratio estimators in adaptive cluster sampling», *Environmetrics*, 18, pp 607-620, 2007.
[6] Deville, J.C. et Qualité, L. , « Échantillonnage multi-dimensionnel (de plusieurs échantillons à la fois) à entropie maximum : définition, propriétés, algorithmes et programmes », *JMS Insee,* 2005.
[7] Deville, J.C., « Sondages approfondis », *Notes de cours ENSAE,* 2005 à 2012.
[8] Tillé, Y., « Sampling algorithms », *Springer,* 2006

↺⇆⇄↞↬↢↣↰↱↲↳⇇⇉↞↠↶ ↷↺↻⊸⇜⇝ ∅∫∫∫ ⨍ ⊠∉∈∋∈∋? tℂℭ𝒫𝒮𝕞ℙwℜ ⊂⊖ ⊗⊕ ‡⊙