Une nouvelle approche du tirage systématique en population infinie

Matthieu Wilhelm[1] & Yves Tillé[1]

[1] *Université de Neuchâtel, Pierre à Mazel 7, 2000 Neuchâtel*

Résumé. On développe une famille de plans de sondage dont le tirage systématique est un cas limite. Cette famille de plans de sondage dépend d'un paramètre permettant de contrôler l'étalement. La définition de cette famille de plans de sondage peut être étendue lorsque la population considérée n'est pas finie, par exemple lorsque l'intervalle [0,1] est considéré. On est en mesure de calculer les probabilités d'inclusions d'ordre 2 et ainsi d'estimer la variance de l'estimateur du total d'une variable observée.

Mots-clé. Echantillonnage systématique, processus ponctuel

1 Echantillonnage continu et processus ponctuel

L'échantillonnage en population infinie requiert quelques outils mathématiques plus subtils que l'échantillonnage en population finie. Si on considère un domaine compact $\Omega \subset \mathbb{R}$, un échantillon s de n individus tirés dans Ω est en fait un $n-$tuple particulier de Ω. Toutefois, on ne peut pas directement l'identifier à Ω^n, pour des raisons techniques sur lesquelles nous ne nous arrêterons pas (voir Deville, 1989). Toutefois, on se limitera à considérer que pour chaque unité $x_i,\ i = 1, \ldots, n$ de l'échantillon, il existe une densité f_i sur Ω qui est en fait la densité marginale de l'unité x_i. On peut alors définir l'intensité (Cordy, 2003), $\pi : \Omega \to \mathbb{R}$ comme:

$$\pi(x) = \sum_{i=1}^{n} f_i(x).$$

On remarque que π est une mesure mais pas une densité de probabilité puisque

$$\int_{\Omega} \pi(x)\ dx = n.$$

Cette intensité joue un rôle similaire à la probabilité d'inclusion d'une unité en population finie. En fait, on remarque qu'un enchantillon dans le continu est la réalisation d'un processus ponctuel. Cela justifie l'usage du terme *intensité* pour désigner la fonction π, qui est en fait l'intensité du processus ponctuel qui caractérise un échantillon. L'intensité peut aussi être vue comme la dérivée de Radon-Nikodym de la mesure de comptage du processus ponctuel qui caractérise l'échantillon.

On peut aussi définir *la densité produit de second ordre* (on utilise ici la traduction de la terminologie utilisée dans Møller & Waagepetersen (2003)) comme la dérivée de Radon-Nikodym de la mesure du deuxième moment factoriel du processus ponctuel. De la même manière que l'intensité joue le rôle de la probabilité d'inclusion, la densité produit de second ordre joue le rôle de la probabilité d'inclusion d'ordre deux.

On peut alors simplement définir l'analogue continu de l'estimateur de Horvitz-Thompson à l'aide des notions d'intensité et de densité produit de second ordre. Ce développement est détaillé dans Cordy (2003).

2 Le tirage systématique

Le tirage systématique est un concept qui est naturel et qu'il est facile de définir. En effet, si l'on demande à un enfant de disposer n points sur ligne droite, il y a une forte probabilité qu'il dispose ces points de manière systématique. Cela traduit le fait qu'un échantillon systématique maximise l'étalement des points. On peut définir une mesure de l'étalement comme la variance de l'écart entre deux points successifs. Si le tirage est systématique, l'écart entre deux points est déterministe et donc sa variance est nulle. Cette définition peut-être étendue au cas multi-dimensionel lorsque l'on remplace le concept d'écart entre deux points par le concept d'aire des polygones de Voronoï.

Le tirage systématique peut-être défini à la fois lorsque la population est finie et lorsque la population considérée est infinie et incomptable, à l'instar d'un intervalle de l'ensemble des nombres réels. On peut définir le tirage systématique en continu comme suit:

Définition 1 (*Tirage systématique sur* $[0,1]$)
On choisit un nombre x_1 uniformément dans l'intervalle $]0, \frac{1}{n}[$. Ensuite, on sélectionne les unité suivantes:

$$x_k = x_1 + k \cdot \frac{N}{n},\ k = 0, \ldots, n-1.$$

On remarque que ce tirage donne lieu à une intensité uniforme pour toute la population. Il est alors possible de définir le tirage systématique d'une densité quelconque. En effet, on définit un tirage systématique selon une loi de distribution F comme l'image par F^{-1} d'un tirage systématique sur $[0,1]$.

3 Une approximation d'un plan de sondage systématique dans le continu

On considère à présent $\Omega = [0,1]$. Ce qui suit peut être facilement étendu au cas où $\Omega \subset \mathbb{R}$ est un compact quelconque. Nous proposons un plan de sondage de taille fixe n, $p_r : [0,1]^n \to [0,1]$, dépendant d'un paramètre r. Lorsque r tend vers l'infini, on obtient le plan de sondage systématique. Une telle convergence est moins triviale qu'il

n'y paraît. En effet, puisque l'objet considéré est un processus ponctuel, il s'agit de définir proprement ce que l'on entend par convergence.

Définition 2 (plan de sondage p_r)
Soit $f : \Omega \to [0,1]$ une densité de probabilité définie sur $[0,1]$. On considère

$$x_1, \ldots, x_{n \cdot r} \overset{\text{i.i.d}}{\sim} f.$$

Soit $i_1 \sim$ uniforme $\{1, \ldots, r\}$ et $i_k = i_1 + (k-1) \cdot r,\ k = 2, \ldots, n$. Finalement, l'échantillon est composé des unités $x_{i_1}, \ldots, x_{i_n}$.

On peut voir sur les figure 1 et 2, des tirages avec différentes valeurs de r. La figure 2 illustre comment notre plan de sondage p_r peut aussi approcher le tirage systématique selon une distribution quelconque f. Dans le cas présent, on a utilisé une loi Beta(α, β) de paramètres $\alpha = 2, \beta = 5$. Ainsi que le montre la figure 1, l'étalement augmente

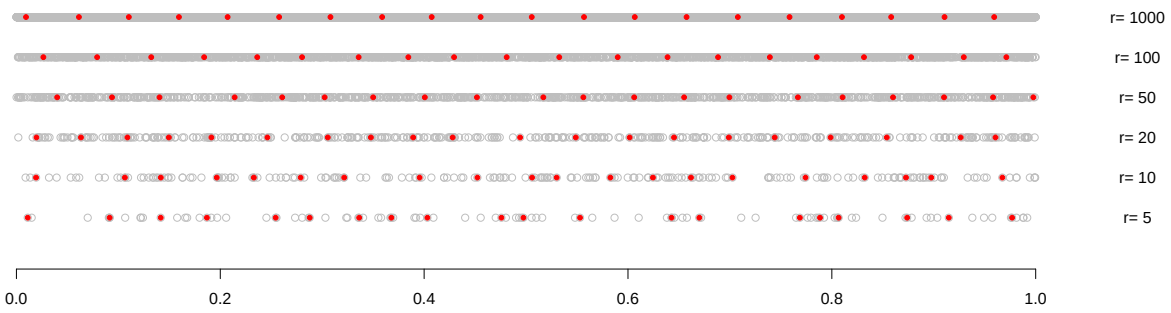

Figure 1: Echantillons de taille $n = 20$ sur l'intervalle $[0,1]$ pour différentes valeurs de r et pout f densité uniforme.

jusqu'à devenir quasiment systématique. L'avantage que présente ce plan de sondage est que la densité produit de second ordre est différente de 0, sur tout l'intervalle $[0,1]$. De la même manière que dans le cas discret, l'estimateur de variance de l'estimateur Horvitz-Thompson est non-biaisé si la densité produit de second ordre est strictement positive pour tout couple de l'intervalle. Toutefois, plus r devient grand, plus la variance de l'estimateur de variance de l'estimateur de Horvitz-Thompson va être grande. Le choix du paramètre r peut donc être vu comme un compromis entre précision de l'estimation de variance et précision de l'estimation elle-même, si la variable observée est fortement auto-corréllée.

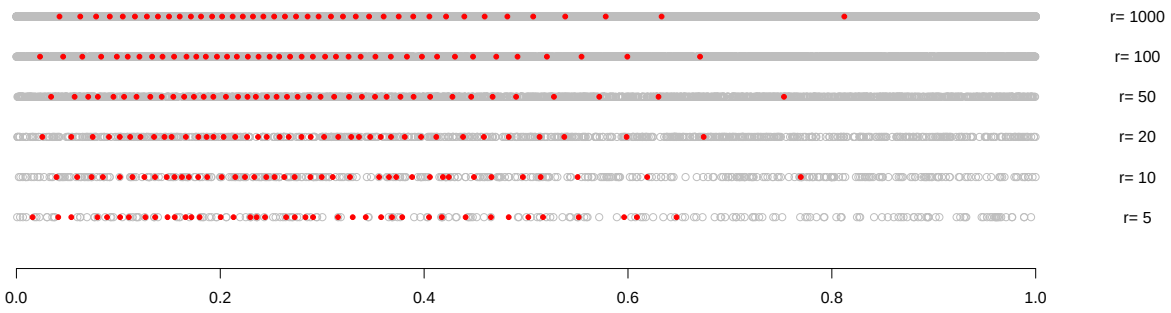

Figure 2: Echantillons de taille $n = 20$ sur l'intervalle $[0, 1]$ pour différentes valeurs de r et pout f suivant une loi Beta(α, β) de paramètres $\alpha = 2, \beta = 5$.

Bibliographie

[1] C. B. Cordy (1993), An extension of the Horvitz-Thompson theorem to point sampling from a continuous universe, *Statistics & Probability Letters*, 18, 353-362.
[2] Deville J-C (1989), Une théorie simplifiée des sondages, *in Les ménages: mélanges en l'honneur de Jacques Desabie*, INSEE, Paris, 191-214.
[3] Møller J. & Waagepetersen R.P. (2003), *Statistical Inference and Simulation of Spatial Point Processes*, Chapman & Hall/CRC, Boca Raton.