Résultats asymptotiques pour la méthode systématique de Deville

Guillaume Chauvet, Jean-Claude Deville

Résultats asymptotiques pour la méthode systématique de Deville

Guillaume Chauvet & Jean-Claude Deville

Ecole Nationale de la Statistique et de l'Analyse de l'Information

Abstract

We are interested in a sampling algorithm known in the literature as Deville's Systematic Sampling. This algorithm enables without-replacement selection of fixed-size samples, with given inclusion probabilities. In this paper, we show that this algorithm may be well-approximated by another sampling algorithm, where units are selected independently. From this comparison, we deduce asymptotic results for the Horvitz-Thompson estimator.

Keywords : Asymptotic Normality - Sampling Algorithm - Unequal Probabilities - Variance Approximation

Résumé

Nous nous intéressons à un algorithme d'échantillonnage connu dans la littérature sous le nom de méthode systématique de Deville. Cet algorithme permet de sélectionner sans remise des échantillons de taille fixe, avec des probabilités d'inclusion fixées. Dans ce travail, nous montrons que cet algorithme peut être bien approché par une autre procédure de tirage où des individus sont sélectionnés indépendamment. Cette comparaison permet d'obtenir des résultats asymptotiques pour l'estimateur de Horvitz-Thompson.

Mots-clés : Algorithme de tirage - Approximation de variance - Normalité asymptotique - Probabilités inégales

1 Notation et tirage systématique de Deville

Nous considérons une population finie d'individus U constituée de N unités représentées par leurs labels $1, \ldots, k, \ldots, N$, avec les caractéristiques associées $y(1), \ldots, y(k), \ldots, y(N)$. Un échantillon s est un sous-ensemble de U sélectionné avec des probabilités d'inclusion $\boldsymbol{\pi} = [\pi(1), \ldots, \pi(k), \ldots, \pi(N)]'$. On suppose sans perte de généralité que les probabilités

d'inclusion vérifient $0 < \pi(k) < 1$ pour chaque unité $k \in U$, avec $n = \sum_{k \in U} \pi(k)$ la taille d'échantillon souhaitée. Soit

$$V(k) = \sum_{l=1}^{k} \pi(l) \text{ pour tout } k \in U, \tag{1}$$

avec $V(0) = 0$. Une unité k sera dite *frontalière* s'il existe un entier i tel que $V(k-1) < i \leq V(k)$. Les unités frontalières sont notées $k_i,\ i = 1, \ldots, n-1$, et pour toute unité frontalière k_i on note $a_i = i - V(k_i - 1)$ et $b_i = V(k_i) - i$. En particulier, on a $0 \leq a_i, b_i \leq 1$ et $\pi(k_i) = a_i + b_i$. Une *microstrate* $U_i,\ i = 1, \ldots, n$, est l'ensemble des unités k de U telles que $k_{i-1} \leq k \leq k_i$, avec par convention $k_0 = 0$ et $k_n = N$. Notons que les microstrates peuvent se chevaucher, puisqu'une unité frontalière peut appartenir à deux microstrates adjacentes. Pour toute microstrate U_i, on note également

$$\boldsymbol{\alpha}_i = [\alpha_i(k_{i-1}), \ldots, \alpha_i(k), \ldots, \alpha_i(k_i)]', \tag{2}$$

avec $\alpha_i(k) = b_{i-1}$ si $k = k_{i-1}$, $\alpha_i(k) = a_i$ si $k = k_i$ et $\alpha_i(k) = \pi(k)$ sinon. On a $\sum_{k \in U_i} \alpha_i(k) = 1$.

Nous nous intéressons à un plan de sondage particulier, appelé dans la littérature *tirage systématique de Deville* (Deville, 1998 ; Tillé, 2006). Le tirage systématique de Deville de paramètre $\boldsymbol{\pi}$ est défini dans l'algorithme 1.

Algorithme 1 Sélection d'un échantillon selon la méthode systématique de Deville

1. A l'étape 1 :
   (a) On génère une variable aléatoire u_1 selon une loi uniforme sur $[0, 1]$.
   (b) Pour toute unité k de U, on pose $I_1(k) = 1$ si $V(k-1) \leq u_1 < V(k)$, et $I_1(k) = 0$ sinon.
2. A l'étape i :
   (a) On génère une variable aléatoire u_i
      i. si l'unité k_{i-1} a été sélectionnée à l'étape $i-1$, selon une loi uniforme sur $[b_{i-1}, 1]$.
      ii. si l'unité k_{i-1} n'a pas été sélectionnée à l'étape $i-1$, selon une loi uniforme sur $[0, b_{i-1}]$ avec une probabilité $\frac{a_{i-1}b_{i-1}}{(1-a_{i-1})(1-b_{i-1})}$, et selon une loi uniforme sur $[0, 1]$ sinon.
   (b) Pour toute unité k de U, on pose $I_i(k) = 1$ si $V(k-1) \leq u_i < V(k)$, et $I_i(k) = 0$ sinon.

L'échantillon s est constitué des unités k telles que $I_i(k) = 1$ pour un entier $i = 1, \ldots, n$. Il est obtenu par des tirages successifs de taille 1 dans chaque microstrate, et le

plan de sondage est donc de taille fixe par construction. Les variables aléatoires u_i sont générées de façon à ce qu'une unité frontalière ne puisse pas être sélectionnée deux fois. Deville (1998) montre que cet algorithme respecte exactement les probabilités d'inclusion $\boldsymbol{\pi}$, et donne une formule explicite pour les probabilités d'inclusion d'ordre deux. Le π-estimateur du total $Y = \sum_{k \in U} y_k$ peut s'écrire sous la forme

$$\hat{Y} = \sum_{k \in s} \frac{y(k)}{\pi(k)} = \sum_{i=1}^{n} \sum_{k \in U_i} \frac{y(k)}{\pi(k)} I_i(k).$$

2 Procédure de sélection conjointe

Dans le cas particulier où $b_i = 0,\ i = 1, \ldots, n-1$, les microstrates sont non chevauchantes et la méthode systématique de Deville est équivalente au tirage stratifié de taille 1 dans chaque microstrate, avec les probabilités d'inclusion $\boldsymbol{\pi}$. Intuitivement, l'algorithme est proche d'un tirage stratifié si les probabilités d'inclusion des unités frontalières sont faibles. Les échantillons sélectionnés à l'aide de cet algorithme bénéficient alors d'un effet de stratification, qui tend à réduire la variance si la variable d'intérêt est positivement corrélée à la variable ordonnant la population.

Algorithme 2 Sélection coordonnée d'un échantillon selon la méthode systématique de Deville et d'un échantillon de type stratifié

---

1. A l'étape 1 :
   (a) On génère une variable aléatoire $u_1 = v_1$ selon une loi uniforme sur $[0, 1]$.
   (b) Pour toute unité k de U, on pose $I_1(k) = J_1(k) = 1$ si $V(k-1) \leq v_1 < V(k)$, et $I_1(k) = J_1(k) = 0$ sinon.
2. A l'étape i :
   (a) On génère une variable aléatoire v_i selon une loi uniforme sur $[0, 1]$.
   (b) Pour toute unité k de U, on pose $J_i(k) = 1$ si $V(k-1) \leq v_i < V(k)$, et $J_i(k) = 0$ sinon.
   (c) On génère une variable aléatoire u_i
      i. si l'unité k_{i-1} a été sélectionnée à l'étape $i-1$, selon une loi uniforme sur $[b_{i-1}, 1]$.
      ii. si l'unité k_{i-1} n'a pas été sélectionnée à l'étape $i-1$, selon une loi uniforme sur $[0, b_{i-1}]$ avec une probabilité $\frac{a_{i-1} b_{i-1}}{(1-a_{i-1})(1-b_{i-1})}$, et en prenant $u_i = v_i$ sinon.
   (d) Pour toute unité k de U, on pose $I_i(k) = 1$ si $V(k-1) \leq u_i < V(k)$, et $I_i(k) = 0$ sinon.

Pour le montrer, une procédure de sélection conjointe de deux échantillons est donnée dans l'algorithme 2. Clairement, le vecteur aléatoire $(u_1, \ldots, u_n)'$ obtenu dans cet algorithme possède la même distribution que le vecteur obtenu dans l'algorithme 1, et conduit donc également à la sélection d'un échantillon selon la méthode systématique de Deville. Le vecteur aléatoire $(v_1, \ldots, v_n)'$ obtenu dans l'algorithme 2 conduit à la sélection d'un échantillon de type stratifié, avec des sélections indépendantes d'une microstrate à l'autre, mais avec la possibilité de sélectionner deux fois dans l'échantillon une unité frontalière. A chacune des étapes de l'algorithme, il existe une très forte probabilité que le même alea soit utilisé pour la sélection des deux échantillons. En conséquence, les estimateurs correspondant à ces deux échantillons seront proches. On note

$$\begin{aligned} \hat{Y}^{(sd)} &= \sum_{i=1}^{n} \sum_{k \in U_i} \frac{y(k)}{\pi(k)} I_i(k) \\ &= \sum_{i=1}^{n} \hat{Y}_i^{(sd)} \end{aligned}$$

et

$$\begin{aligned} \hat{Y}^{(st)} &= \sum_{i=1}^{n} \sum_{k \in U_i} \frac{y(k)}{\pi(k)} J_i(k) \\ &= \sum_{i=1}^{n} \hat{Y}_i^{(st)} \end{aligned}$$

On montre facilement que

$$E\left[\hat{Y}^{(sd)}\right] = E\left[\hat{Y}^{(st)}\right] = Y,$$

où $E(\cdot)$ désigne l'espérance sous le mécanisme d'échantillonnage associé à l'algorithme ci-dessus. D'autre part, en raison de l'indépendance des $v_i,\ i = 1, \ldots, n$, on a

$$\begin{aligned} V\left[\hat{Y}^{(st)}\right] &= \sum_{i=1}^{n} V\left[\hat{Y}_i^{(st)}\right] \\ &\equiv \sum_{i=1}^{n} V_i, \end{aligned} \tag{3}$$

où $V(\cdot)$ désigne la variance sous le mécanisme d'échantillonnage.

3 Résultats obtenus

Nous considérons une suite d'expériences obtenues de la façon suivante. Nous supposons l'existence d'une suite de plans de sondage permettant de sélectionner des échantillon s_N de taille n_N dans les populations U_N, avec $n_N \to \infty$ quand $N \to \infty$. Pour simplifier, l'indice N est omis dans ce qui suit. Nous faisons également les hypothèses suivantes :

C1 : $\max_i \pi_{p_i} = O(n^{-1/2})$,

C2 : $\max_i \left[\frac{a_i \left(\frac{y(k_i)}{\pi(k_i)} - Y_i \right)^2}{V_i} \right] \to 0 \quad \max_i \left[\frac{b_i \left(\frac{y(k_i)}{\pi(k_i)} - Y_{i+1} \right)^2}{V_i} \right] \to 0.$

La condition $C1$ signifie que les probabilités d'inclusion des unités frontalières deviennent négligeables quand la taille d'échantillon devient grande. La condition $C2$ est

une condition de type Noether, et garantit que la variance associée aux unités frontalières est arbitrairement faible, pour les tirages associés à chacune des microstrates. Notons que ces conditions ne portent que sur les unités frontalières ; en particulier, aucune contrainte n'est imposée pour les probabilités d'inclusion des unités non frontalières.

Proposition 1 *Sous les conditions $C1$ et $C2$, on a :*

$$\frac{E\left[\hat{Y}^{(sd)} - \hat{Y}^{(st)}\right]^2}{V\left[\hat{Y}^{(st)}\right]} \to 0.$$

En utilisant une remarque de Hajek (1960), on en déduit que $\hat{Y}^{(sd)}$ et $\hat{Y}^{(st)}$ ont les mêmes propriétés en termes de variance et de lois limites. Comme $\hat{Y}^{(st)}$ est une somme de variables aléatoires indépendantes, sa variance s'obtient facilement à l'aide de la formule (3), et sa normalité asymptotique peut être obtenue sous une condition de type Lindeberg.

Bibliographie

[1] Deville, J.-C. (1998). *Une nouvelle (encore une !) méthode de tirage à probabilités inégales*, Rapport Technique, Insee.
[2] Hajek, J. (1960). *Limiting Distributions in Simple Random Sampling From a Finite Population*. Publications of the Mathematical Institute of the Hungarian Academy of Sciences, 5, 361-374.
[3] Tillé, Y. (2006). *Sampling Algorithms*. New-York, Wiley.