OPTIMUM DE TYPE NEYMAN POUR L'ECHANTILLONNAGE EQUILIBRE SUR DES MARGES

Daniel Bonnéry
Guillaume Chauvet
Jean-Claude Deville

Journées de Méthodologie Statistique - Paris

25/03/2009

Présentation du problème

Le sondage stratifié

On se donne

- une population $U = \{k \in [\![1, N]\!]\}$
- découpée en H strates $U = \cup_{h=1}^{H} U_h$
- U_h étant de taille N_h
- pour chaque U_h un tirage aléatoire simple de taille n_h parmi N_h
- une variable y définie sur U
- la dispersion de y dans la strate U_h est notée S_h^2

L'allocation de Neyman

La variance de l'estimateur de Horvitz Thompson $\hat{Y}$ du total Y de la variable y est alors :

$$V[\hat{Y}] = \sum_{h=1}^{H} N_h^2 (1 - \frac{n_h}{N_h}) \frac{S_h^2}{n_h}$$

L'allocation de Neyman est le H-uplet $n_1, \ldots, n_H$ qui minimise l'expression de la variance ci-dessus, sous la contrainte : $\sum n_h = n$.

Le problème revient à déterminer des probabilités d'inclusion constantes par strate $\pi_k = \frac{n_h}{N_h}$ optimales pour l'estimation de Y.

Rappels sur le sondage équilibré

Etant donné une variable auxiliaire z définie sur U, un tirage est équilibré sur z lorsque l'estimateur de Horvitz Thompson du total Z de z est exact.

Le sondage stratifié avec sondage aléatoire simple dans chaque strate est un tirage équilibré sur la taille :

- de l'échantillon dans chaque strate ($z = \pi$, $\hat{n}_h = n_h$)
- de la population dans chaque strate ($z = \mathbb{1}_{U_h}$, $\hat{N}_h = N_h$)

La méthode du cube

Il s'agit d'une méthode de tirage permettant de respecter :

- des contraintes d'équilibrage
- des probabilités d'inclusion fixées à l'avance

Deville et Tillé [2005] proposent une formule d'approximation de la variance pour l'estimateur de Horvitz-Thompson du total Y dans le cas d'un tirage équilibré sur les variables auxiliaires $z^1 \dots z^H$ qui peuvent dépendre de π :

$$V[\hat{Y}] = \sum_{k=1}^{N} (\frac{1}{\pi_k} - 1)(y_k - \hat{y}_k(\pi))^2$$

$\hat{y}(\pi)$ désignant le prédicteur de y par la régression sur $z^1 \dots z^H$ pondérée par les poids $(\frac{1}{\pi_k} - 1)$

La méthode du cube

La variance peut être vue comme une fonction de π :

$$V : \pi \mapsto \sum_{k=1}^{N} (\frac{1}{\pi_k} - 1)(y_k - \hat{y}_k(\pi))^2$$

On peut donc, comme dans le cas de l'optimum de Neyman, chercher à déterminer les probabilités d'inclusion qui minimisent cette approximation de la variance.

Equilibrage sur deux variables catégorielles

Variables d'équilibrage

On se donne deux variables catégorielles qui définissent deux partitions de U :

$$U = \cup_{i \in [\![1,I]\!]} U_{i.} = \cup_{j \in [\![1,J]\!]} U_{.j}$$

On réalise un sondage équilibré sur les variables : $z^{i.} = \mathbb{1}_{U_{i.}}$, $z^{.j} = \mathbb{1}_{U_{.j}}$
On cherche les probabilités d'inclusion qui minimisent l'approximation de variance sous les contraintes linéaires $A\pi = B$:

- probabilités constantes dans chaque cellule $U_{ij} = U_{i.} \cap U_{.j}$
- espérance de la taille de l'échantillon égale à n, i.e. $\sum \pi_k = n$

Recherche d'un point fixe

On remarque que

$$\begin{aligned} V(\pi) &= \sum_{k=1}^{N}(\frac{1}{\pi_k}-1)(y_k-\hat{y}_k(\pi))^2 \\ &= v(\pi,\hat{y}(\pi)) \end{aligned}$$

avec

$$v:(\pi,\tilde{y})\mapsto\sum_{k=1}^{N}(\frac{1}{\pi_k}-1)(y_k-\tilde{y}_k)^2$$

Ceci reprend une idée de **Tillé et Favre (2005)**

Etant donné des probabilités de départ π^0, on définit une suite $(\pi^r)_{r\in\mathbb{N}}$: $\pi^{r+1}=argmin_{\tilde{\pi}|A\pi=B}v(\tilde{\pi},\hat{y}(\pi^r))$
On démontre qu'en cas de convergence de la suite (π^r), lorsque les variables d'équilibrage ne dépendent pas de π, on obtient un minimum local contraint de la fonction V en la limite.

Quotas probabilistes

On garde les partitions $U_{i.}$, $U_{.j}$ définies précédemment.

On réalise un sondage équilibré sur les variables :

$$z^{i.} = \pi_k \mathbb{1}_{U_{i.}} \qquad z^{.j} = \pi_k \mathbb{1}_{U_{.j}}$$

On obtient alors un sondage équilibré sur **des tailles d'échantillon** sur chaque $U_{i.}$ et chaque $U_{.j}$: on maîtrise par exemple le nombre d'hommes enquêtés, et le nombre de personnes enquêtées par tranches d'âge, sans contrôler le nombre d'individus enquêtés appartenant à un croisement des populations d'une tranche d'âge et de sexe donnés.

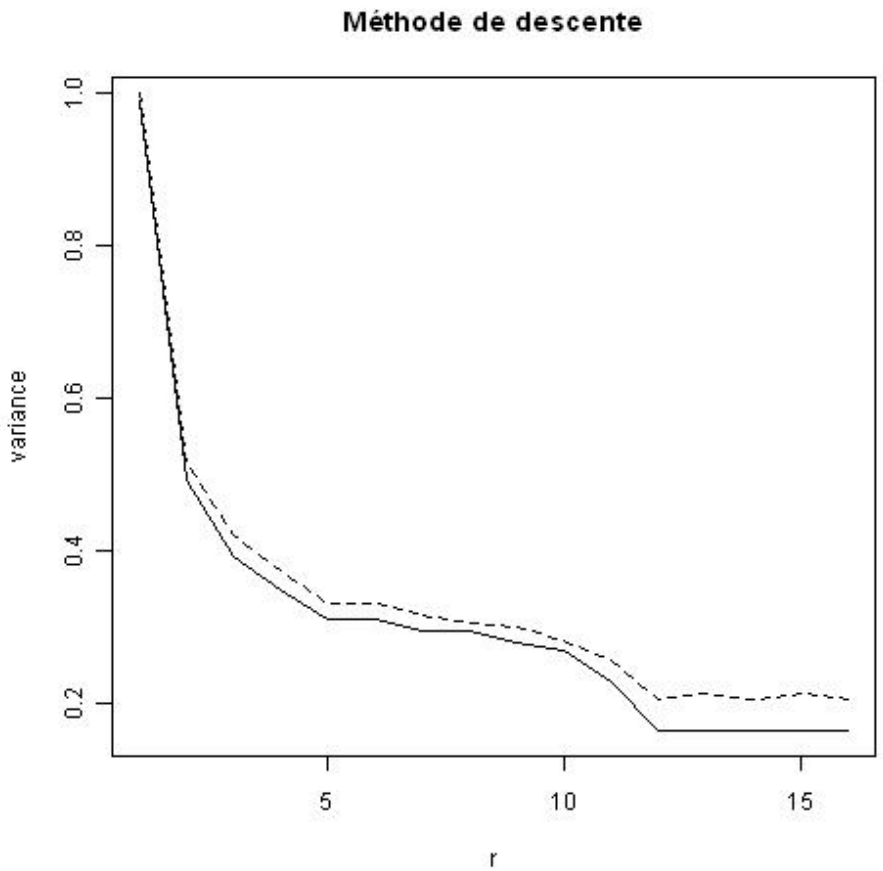
Méthode de descente

Il s'agit de déterminer des probabilités d'inclusion optimales sous des contraintes linéaires :

- probabilités constantes sur $U_{ij} = U_{i.} \cap U_{.j}$
- espérance de la taille de l'échantillon égale à n $(\sum \pi_k = n)$

La méthode précédente ne peut être appliquée. On utilise alors une méthode de descente : il s'agit de calculer la dérivée de $V(\pi)$ et de faire évoluer (π) dans le sens de la pente.

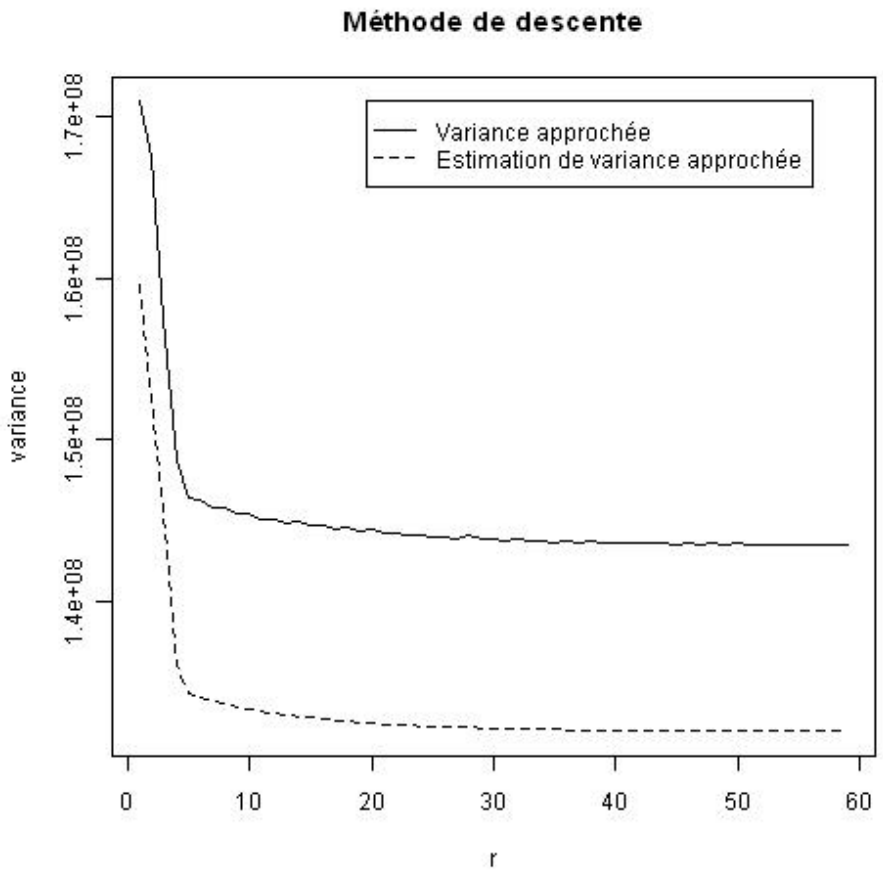
Evolution comparée des variances approchée et empirique



On simule une variable y par un modèle ANOVA hétéroscédastique ou les facteurs sont générés aléatoirement sur une population de 600 individus. On constate les mêmes variations de la variance approchée (trait continu) et de la variance empirique (pointillé) à chaque itération. La variance empirique a été obtenue par simulations de tirages équilibrés.

Cas pratique

La méthode de recherche de probabilités optimales nécessite uniquement la connaissance de $\sum_{k \in U_{ij}} y_k$ et $\sum_{k \in U_{ij}} y_k^2$ et N_{ij} pour tout i, j.
Supposons qu'on dispose de valeurs approchées de y, $z^{i\cdot}$, $z^{\cdot j}$ sur un échantillon s de la population, ce qui permet d'obtenir $\hat{Y}_{ij}$, $\widehat{Y_{ij}^2}$, $\hat{N}_{ij}$ des estimateurs de $Y_{ij} = \sum_{k \in U_{ij}} y_k$, $Y_{ij}^2 = \sum_{k \in U_{ij}} y_k^2$, N_{ij}.
On peut alors lancer le programme d'optimisation à partir des valeurs estimées.



On simule une variable y par un modèle ANOVA hétéroscédastique ou les facteurs sont générés aléatoirement sur une population de 600 individus. On constate que les variations de la variance approchée (en trait continu) suivent celles de son estimation (en pointillés) à partir d'un sondage aléatoire simple de 60 individus parmi 600.

Conclusion

Deux méthodes numériques ont été appliquées pour obtenir un minimum local contraint de la variance approchée de l'estimateur de Horvitz-Thompson d'un sondage équilibré.

Ces méthodes peuvent être généralisées, mais leur application dans un cas général peut nécessiter la connaissance de la variable y.

Bibliographie


Tillé Favre.
Optimal allocation in balanced sampling.
Statistics and probability letters, 2005.


Deville Tillé.
Efficient balanced sampling : the cube method.
Biometrika, 2004.


Deville Tillé.
Variance approximation under balanced sampling.
Journal of statistical planning and Inference, 2005.