

Probabilités d'inclusion optimales pour un échantillonnage équilibré

Guillaume Chauvet

Travail joint avec Daniel Bonnéry et Jean-Claude Deville
(Ensaï)

Journée Méthodes Avancées pour l'Analyse de Sondages
Complexes
Toulouse, 22/10/2009



En résumé

Une information auxiliaire connue à l'étape du plan de sondage permet de réduire la variance de l'estimateur de Horvitz-Thompson à l'aide de méthodes d'échantillonnage équilibré.

En résumé

Une information auxiliaire connue à l'étape du plan de sondage permet de réduire la variance de l'estimateur de Horvitz-Thompson à l'aide de méthodes d'**échantillonnage équilibré**.

Une manière complémentaire de réduire sa variabilité est de définir des probabilités d'inclusion qui minimisent la variance résiduelle, au moins approximativement.

En résumé

Une information auxiliaire connue à l'étape du plan de sondage permet de réduire la variance de l'estimateur de Horvitz-Thompson à l'aide de méthodes d'**échantillonnage équilibré**.

Une manière complémentaire de réduire sa variabilité est de définir des probabilités d'inclusion qui minimisent la variance résiduelle, au moins approximativement.

Nous proposons ici une méthode permettant de calculer de telles probabilités d'inclusion. Cette méthode nécessite généralement de faire appel à des estimations issues d'une autre enquête.

- 1 Introduction
- 2 L'échantillonnage équilibré
- 3 Calcul de probabilités d'inclusion optimales
- 4 Etude par simulations
- 5 Perspectives

Introduction

Objectif

On considère une population finie d'individus

$$U = \{1, \dots, k, \dots, N\},$$

où un individu est supposé identifiable par son label k . On note y_k la valeur prise par une variable d'intérêt y sur l'individu k de U .

Objectif

On considère une population finie d'individus

$$U = \{1, \dots, k, \dots, N\},$$

où un individu est supposé identifiable par son label k . On note y_k la valeur prise par une variable d'intérêt y sur l'individu k de U .

L'objectif est ici l'estimation du total

$$t_y = \sum_{k \in U} y_k$$

de la variable y .

Plan de sondage

L'échantillon aléatoire S est sélectionné au moyen d'un plan de sondage $p(\cdot)$ sans remise et défini sur U . La taille n d'échantillon voulue est supposée fixée.

Plan de sondage

L'échantillon aléatoire S est sélectionné au moyen d'un plan de sondage $p(\cdot)$ sans remise et défini sur U . La taille n d'échantillon voulue est supposée fixée.

Les probabilités $\pi_k = \mathbb{P}(k \in S)$ et $\pi_{kl} = \mathbb{P}(k, l \in S)$ sont appelées probabilités d'inclusion d'ordre 1 et 2.

Plan de sondage

L'échantillon aléatoire S est sélectionné au moyen d'un plan de sondage $p(\cdot)$ sans remise et défini sur U . La taille n d'échantillon voulue est supposée fixée.

Les probabilités $\pi_k = \mathbb{P}(k \in S)$ et $\pi_{kl} = \mathbb{P}(k, l \in S)$ sont appelées probabilités d'inclusion d'ordre 1 et 2.

Le total t_y est estimé sans biais sous le plan de sondage par le π -estimateur :

$$\hat{t}_{y\pi} = \sum_{k \in S} \frac{y_k}{\pi_k}.$$

Echantillonnage équilibré

Un échantillon est dit **équilibré** sur un jeu de variables auxiliaires \mathbf{x} si les équations

$$\hat{t}_{\mathbf{x}\pi} = t_{\mathbf{x}}$$

sont respectées.

Echantillonnage équilibré

Un échantillon est dit **équilibré** sur un jeu de variables auxiliaires \mathbf{x} si les équations

$$\hat{t}_{\mathbf{x}\pi} = t_{\mathbf{x}}$$

sont respectées.

Par extension, un plan de sondage est dit **équilibré** sur les variables \mathbf{x} si le support du plan de sondage est réduit aux échantillons équilibrés sur \mathbf{x} .

Echantillonnage équilibré

Un échantillon est dit **équilibré** sur un jeu de variables auxiliaires \mathbf{x} si les équations

$$\hat{t}_{\mathbf{x}\pi} = t_{\mathbf{x}}$$

sont respectées.

Par extension, un plan de sondage est dit **équilibré** sur les variables \mathbf{x} si le support du plan de sondage est réduit aux échantillons équilibrés sur \mathbf{x} .

Question : Quelle est la stratégie optimale d'échantillonnage, i.e. menant à un π -estimateur de variance minimale ?

Stratégie optimale d'échantillonnage

Nedyalkova et Tillé (2008) donnent une réponse sous le modèle de superpopulation

$$y_k = \mathbf{x}'_k \beta + \epsilon_k,$$

$$\text{avec } E_m(\epsilon_k) = 0 \quad V_m(\epsilon_k) = v_k^2 \sigma^2 \quad \text{Cov}_m(\epsilon_k, \epsilon_l) = 0.$$

Stratégie optimale d'échantillonnage

Nedyalkova et Tillé (2008) donnent une réponse sous le modèle de superpopulation

$$y_k = \mathbf{x}'_k \beta + \epsilon_k,$$

$$\text{avec } E_m(\epsilon_k) = 0 \quad V_m(\epsilon_k) = v_k^2 \sigma^2 \quad \text{Cov}_m(\epsilon_k, \epsilon_l) = 0.$$

Théorème

Si les v_k sont supposés connus, la stratégie optimale d'échantillonnage consiste à sélectionner un échantillon

- *équilibré sur les variables \mathbf{x}_k ,*
- *avec des probabilités d'inclusion proportionnelles aux v_k .*

Stratégie optimale d'échantillonnage

On peut réécrire le π -estimateur sous la forme

$$\hat{t}_{y\pi} = \beta' \hat{t}_{\mathbf{x}\pi} + \hat{t}_{\epsilon\pi}.$$

Stratégie optimale d'échantillonnage

On peut réécrire le π -estimateur sous la forme

$$\hat{t}_{y\pi} = \beta' \hat{t}_{\mathbf{x}\pi} + \hat{t}_{\epsilon\pi}.$$

Principe :

- Le **respect** des probabilités d'inclusion permet d'obtenir une estimation sans biais,

Stratégie optimale d'échantillonnage

On peut réécrire le π -estimateur sous la forme

$$\hat{t}_{y\pi} = \beta' \hat{t}_{\mathbf{x}\pi} + \hat{t}_{\epsilon\pi}.$$

Principe :

- Le **respect** des probabilités d'inclusion permet d'obtenir une estimation sans biais,
- La **restriction du support** du plan de sondage aux échantillons équilibrés permet d'annuler la variabilité du 1er terme,

Stratégie optimale d'échantillonnage

On peut réécrire le π -estimateur sous la forme

$$\hat{t}_{y\pi} = \beta' \hat{t}_{\mathbf{x}\pi} + \hat{t}_{\epsilon\pi}.$$

Principe :

- Le **respect** des probabilités d'inclusion permet d'obtenir une estimation sans biais,
- La **restriction du support** du plan de sondage aux échantillons équilibrés permet d'annuler la variabilité du 1er terme,
- Le **choix** des probabilités d'inclusion permet de minimiser la variabilité du 2nd terme.

But de ce travail

On recherche une stratégie optimale d'échantillonnage sans l'hypothèse d'un modèle de superpopulation, mais sous un modèle de travail

$$y_k = \mathbf{x}'_k B + E_k.$$

But de ce travail

On recherche une stratégie optimale d'échantillonnage sans l'hypothèse d'un modèle de superpopulation, mais sous un modèle de travail

$$y_k = \mathbf{x}'_k B + E_k.$$

Pour un plan de sondage équilibré sur les variables \mathbf{x} , on recherche les probabilités d'inclusion permettant de minimiser la variance du π -estimateur.

L'échantillonnage équilibré

Définition

On suppose dans cette partie que les probabilités d'inclusion π_k sont fixées.

Définition

On suppose dans cette partie que les probabilités d'inclusion π_k sont fixées.

Les variables x_k sont supposées disponibles au moment de l'échantillonnage pour chaque individu k de la population.

Définition

On suppose dans cette partie que les probabilités d'inclusion π_k sont fixées.

Les variables x_k sont supposées disponibles au moment de l'échantillonnage pour chaque individu k de la population.

La plupart des plans de sondage utilisant de l'information auxiliaire peuvent être décrits comme des stratégies d'échantillonnage équilibré.

Exemples

Plan à probabilités inégales de taille fixe

Equilibrage sur la variable $\mathbf{x}_k = x_k = \pi_k$.

Exemples

Plan à probabilités inégales de taille fixe

Equilibrage sur la variable $\mathbf{x}_k = x_k = \pi_k$.

Sondage aléatoire simple stratifié

Equilibrage sur les variables

$$\mathbf{x}_k = [1(k \in U_1), \dots, 1(k \in U_H)].$$

Exemples

Plan à probabilités inégales de taille fixe

Equilibrage sur la variable $\mathbf{x}_k = x_k = \pi_k$.

Sondage aléatoire simple stratifié

Equilibrage sur les variables

$$\mathbf{x}_k = [1(k \in U_1), \dots, 1(k \in U_H)].$$

Plan stratifié de taille fixe dans chaque strate

Equilibrage sur les variables

$$\mathbf{x}_k = [\pi_k 1(k \in U_1), \dots, \pi_k 1(k \in U_H)]$$

Représentation du Cube

Deville et Tillé (2004) proposent un algorithme général pour la sélection d'échantillons équilibrés sur un nombre quelconque de variables, avec des probabilités d'inclusion $\pi = (\pi_1, \dots, \pi_N)'$ quelconques : la méthode du Cube.

Représentation du Cube

Deville et Tillé (2004) proposent un algorithme général pour la sélection d'échantillons équilibrés sur un nombre quelconque de variables, avec des probabilités d'inclusion $\pi = (\pi_1, \dots, \pi_N)'$ quelconques : la méthode du Cube.

Un échantillon s est vu comme un sommet $(s_1, \dots, s_N) \in \{0, 1\}^N$ du N -cube $C = [0, 1]^N$.

Représentation du Cube

Deville et Tillé (2004) proposent un algorithme général pour la sélection d'échantillons équilibrés sur un nombre quelconque de variables, avec des probabilités d'inclusion $\pi = (\pi_1, \dots, \pi_N)'$ quelconques : la méthode du Cube.

Un échantillon s est vu comme un sommet $(s_1, \dots, s_N) \in \{0, 1\}^N$ du N -cube $C = [0, 1]^N$.

Les équations d'équilibrage définissent l'espace des contraintes :

$$\pi + Ker(A) \text{ où } A = (\mathbf{x}_k / \pi_k)_{k \in U}$$

Représentation du Cube

Deville et Tillé (2004) proposent un algorithme général pour la sélection d'échantillons équilibrés sur un nombre quelconque de variables, avec des probabilités d'inclusion $\pi = (\pi_1, \dots, \pi_N)'$ quelconques : la méthode du Cube.

Un échantillon s est vu comme un sommet $(s_1, \dots, s_N) \in \{0, 1\}^N$ du N -cube $C = [0, 1]^N$.

Les équations d'équilibrage définissent l'espace des contraintes :

$$\pi + Ker(A) \text{ où } A = (\mathbf{x}_k / \pi_k)_{k \in U}$$

L'algorithme consiste à arrondir aléatoirement des composantes du vecteur π par une marche aléatoire dans l'espace des contraintes.



Etape de base : la martingale équilibrante

On initialise avec $\pi^{(0)} = \pi$.

Etape de base : la martingale équilibrante

On initialise avec $\pi^{(0)} = \pi$. A l'étape t , $\pi^{(t)} = \pi^{(t-1)} + \delta^{(t)}$, avec

$$\delta^{(t)} = \begin{cases} +\lambda_1(t) u(t) & \text{avec la probabilité } \lambda_2(t)/(\lambda_1(t) + \lambda_2(t)) \\ -\lambda_2(t) u(t) & \text{avec la probabilité } \lambda_1(t)/(\lambda_1(t) + \lambda_2(t)) \end{cases}$$

Etape de base : la martingale équilibrante

On initialise avec $\pi^{(0)} = \pi$. A l'étape t , $\pi^{(t)} = \pi^{(t-1)} + \delta^{(t)}$, avec

$$\delta^{(t)} = \begin{cases} +\lambda_1(t) u(t) & \text{avec la probabilité } \lambda_2(t)/(\lambda_1(t) + \lambda_2(t)) \\ -\lambda_2(t) u(t) & \text{avec la probabilité } \lambda_1(t)/(\lambda_1(t) + \lambda_2(t)) \end{cases},$$

où

Etape de base : la martingale équilibrante

On initialise avec $\pi^{(0)} = \pi$. A l'étape t , $\pi^{(t)} = \pi^{(t-1)} + \delta^{(t)}$, avec

$$\delta^{(t)} = \begin{cases} +\lambda_1(t) u(t) & \text{avec la probabilité } \lambda_2(t)/(\lambda_1(t) + \lambda_2(t)) \\ -\lambda_2(t) u(t) & \text{avec la probabilité } \lambda_1(t)/(\lambda_1(t) + \lambda_2(t)) \end{cases},$$

où

- $\lambda_1(t), \lambda_2(t) > 0$
→ choisis afin qu'au moins une unité soit sélectionnée ou définitivement rejetée.

Étape de base : la martingale équilibrante

On initialise avec $\pi^{(0)} = \pi$. A l'étape t , $\pi^{(t)} = \pi^{(t-1)} + \delta^{(t)}$, avec

$$\delta^{(t)} = \begin{cases} +\lambda_1(t) u(t) & \text{avec la probabilité } \lambda_2(t)/(\lambda_1(t) + \lambda_2(t)) \\ -\lambda_2(t) u(t) & \text{avec la probabilité } \lambda_1(t)/(\lambda_1(t) + \lambda_2(t)) \end{cases},$$

où

- $\lambda_1(t), \lambda_2(t) > 0$
→ choisis afin qu'au moins une unité soit sélectionnée ou définitivement rejetée.
- $u(t) \in \text{Ker}(A)$ est un vecteur (non aléatoire)
→ assure que les équations d'équilibrage sont exactement respectées.

Etape de base : la martingale équilibrante

On initialise avec $\pi^{(0)} = \pi$. A l'étape t , $\pi^{(t)} = \pi^{(t-1)} + \delta^{(t)}$, avec

$$\delta^{(t)} = \begin{cases} +\lambda_1(t) u(t) & \text{avec la probabilité } \lambda_2(t)/(\lambda_1(t) + \lambda_2(t)) \\ -\lambda_2(t) u(t) & \text{avec la probabilité } \lambda_1(t)/(\lambda_1(t) + \lambda_2(t)) \end{cases},$$

où

- $\lambda_1(t), \lambda_2(t) > 0$
→ choisis afin qu'au moins une unité soit sélectionnée ou définitivement rejetée.
- $u(t) \in \text{Ker}(A)$ est un vecteur (non aléatoire)
→ assure que les équations d'équilibrage sont exactement respectées.
- le choix aléatoire assure que les probabilités d'inclusion sont exactement respectées.

Fin de l'échantillonnage

L'algorithme précédent s'arrête quand il n'est plus possible de trouver un vecteur $u(t)$ respectant les contraintes précédentes : c'est la fin de la **phase de vol**.

Fin de l'échantillonnage

L'algorithme précédent s'arrête quand il n'est plus possible de trouver un vecteur $u(t)$ respectant les contraintes précédentes : c'est la fin de la **phase de vol**.

La **phase d'atterrissage** permet de terminer l'échantillonnage pour les unités restantes (au plus q). L'impact sur la variance peut généralement être négligé si le nombre de variables d'équilibrage est faible.

Fin de l'échantillonnage

L'algorithme précédent s'arrête quand il n'est plus possible de trouver un vecteur $u(t)$ respectant les contraintes précédentes : c'est la fin de la **phase de vol**.

La **phase d'atterrissage** permet de terminer l'échantillonnage pour les unités restantes (au plus q). L'impact sur la variance peut généralement être négligé si le nombre de variables d'équilibrage est faible.

Le vecteur $\pi(T)$ obtenu à la dernière étape de l'algorithme donne le résultat de l'échantillonnage.

Approximation de variance

Il est difficile d'obtenir un estimateur sans biais de variance car les probabilités d'inclusion d'ordre 2 sont généralement impossibles à calculer exactement.

Approximation de variance

Il est difficile d'obtenir un estimateur sans biais de variance car les probabilités d'inclusion d'ordre 2 sont généralement impossibles à calculer exactement.

Pour un plan de sondage exactement équilibré et à entropie maximale, Deville et Tillé (2005) proposent l'approximation de variance

$$V_{app}(\hat{t}_{y\pi}) = \frac{N}{N-q} \sum_{k \in U} \frac{1 - \pi_k}{\pi_k} e_k^2(\pi), \quad (1)$$

Approximation de variance

Il est difficile d'obtenir un estimateur sans biais de variance car les probabilités d'inclusion d'ordre 2 sont généralement impossibles à calculer exactement.

Pour un plan de sondage exactement équilibré et à entropie maximale, Deville et Tillé (2005) proposent l'approximation de variance

$$V_{app}(\hat{t}_{y\pi}) = \frac{N}{N-q} \sum_{k \in U} \frac{1 - \pi_k}{\pi_k} e_k^2(\pi), \quad (1)$$

où $e_k(\pi) = y_k - y_k^*(\pi)$ donne l'erreur dans la prédiction de y_k obtenue avec les q variables d'équilibrage \mathbf{x}_k .

Calcul de probabilités d'inclusion optimales

Principe

Nous supposons que les variables d'équilibrage sont fixées avant le calcul des probabilités d'inclusion.

Principe

Nous supposons que les variables d'équilibrage sont fixées avant le calcul des probabilités d'inclusion. Poursuivant une idée suggérée par Tillé et Favre (2005), nous proposons de calculer les probabilités π_k qui minimisent l'approximation de variance V_{app} , sous la contrainte de taille

$$\sum_{k \in U} \pi_k = n.$$

Principe

Nous supposons que les variables d'équilibrage sont fixées avant le calcul des probabilités d'inclusion. Poursuivant une idée suggérée par Tillé et Favre (2005), nous proposons de calculer les probabilités π_k qui minimisent l'approximation de variance V_{app} , sous la contrainte de taille

$$\sum_{k \in U} \pi_k = n. \quad (2)$$

On montre que ces probabilités d'inclusion vérifient

$$\pi_l = n \frac{|e_l(\pi)|}{\sum_{m \in U} |e_m(\pi)|}.$$

Principe

Nous supposons que les variables d'équilibrage sont fixées avant le calcul des probabilités d'inclusion. Poursuivant une idée suggérée par Tillé et Favre (2005), nous proposons de calculer les probabilités π_k qui minimisent l'approximation de variance V_{app} , sous la contrainte de taille

$$\sum_{k \in U} \pi_k = n. \quad (2)$$

On montre que ces probabilités d'inclusion vérifient

$$\pi_l = n \frac{|e_l(\pi)|}{\sum_{m \in U} |e_m(\pi)|}.$$

Interprétation : les probabilités d'inclusion doivent être proportionnelles à l'erreur de prédiction.

Résolution pratique

Problème : la résolution de ce système nécessite la connaissance de la variable y_k pour tous les individus de U .

Résolution pratique

Problème : la résolution de ce système nécessite la connaissance de la variable y_k pour tous les individus de U .

Nous proposons d'ajouter la contrainte suivante : on suppose définie une partition de U en H domaines U_1, \dots, U_H , et on impose

$$\pi_k = \alpha_h \text{ pour tout } k \in U_h. \quad (3)$$

Résolution pratique

Problème : la résolution de ce système nécessite la connaissance de la variable y_k pour tous les individus de U .

Nous proposons d'ajouter la contrainte suivante : on suppose définie une partition de U en H domaines U_1, \dots, U_H , et on impose

$$\pi_k = \alpha_h \text{ pour tout } k \in U_h. \quad (3)$$

On minimise alors V_{app} sous les contraintes de taille globale, et de probabilités égales dans les domaines.

Résolution pratique

La résolution de ce second problème d'optimisation conduit à

$$\alpha_h = n \frac{\sigma_h(\alpha)}{\sum_{j=1}^H N_j \sigma_j(\alpha)},$$

avec

$$\sigma_h(\alpha) = \frac{1}{N_h} \sum_{k \in U_h} e_k^2(\alpha).$$

Résolution pratique

La résolution de ce second problème d'optimisation conduit à

$$\alpha_h = n \frac{\sigma_h(\alpha)}{\sum_{j=1}^H N_j \sigma_j(\alpha)},$$

avec

$$\sigma_h(\alpha) = \frac{1}{N_h} \sum_{k \in U_h} e_k^2(\alpha).$$

Ce système peut être résolu de façon itérative selon une méthode de point fixe. Cela nécessite la connaissance des totaux

$$\sum_{k \in U_h} (\mathbf{x}_k \mathbf{x}'_k, \mathbf{x}_k y_k, y_k^2).$$

Résolution pratique

La résolution de ce second problème d'optimisation conduit à

$$\alpha_h = n \frac{\sigma_h(\alpha)}{\sum_{j=1}^H N_j \sigma_j(\alpha)},$$

avec

$$\sigma_h(\alpha) = \frac{1}{N_h} \sum_{k \in U_h} e_k^2(\alpha).$$

Ce système peut être résolu de façon itérative selon une méthode de point fixe. Cela nécessite la connaissance des totaux

$$\sum_{k \in U_h} (\mathbf{x}_k \mathbf{x}'_k, \mathbf{x}_k y_k, y_k^2).$$

Ces totaux sont supposés connus, ou de façon plus réaliste estimables à l'aide d'une autre enquête.

Algorithme et résultat principal

Algorithme

Algorithme et résultat principal

Algorithme

- 1 Initialiser avec un vecteur quelconque α^0 .

Algorithme et résultat principal

Algorithme

- 1 Initialiser avec un vecteur quelconque α^0 .
- 2 A l'étape t , calculer α^t tel que

$$\alpha_h^t = n \frac{\sigma_h(\alpha^{t-1})}{\sum_{j=1}^H N_j \sigma_j(\alpha^{t-1})} \text{ pour tout domaine } h.$$

Algorithme et résultat principal

Algorithme

- 1 Initialiser avec un vecteur quelconque α^0 .
- 2 A l'étape t , calculer α^t tel que

$$\alpha_h^t = n \frac{\sigma_h(\alpha^{t-1})}{\sum_{j=1}^H N_j \sigma_j(\alpha^{t-1})} \text{ pour tout domaine } h.$$

- 3 Arrêter à l'étape T quand $\text{Max}|\alpha_h^T - \alpha_h^{T-1}| \leq \epsilon$, fixé.

Algorithme et résultat principal

Algorithme

- 1 Initialiser avec un vecteur quelconque α^0 .
- 2 A l'étape t , calculer α^t tel que

$$\alpha_h^t = n \frac{\sigma_h(\alpha^{t-1})}{\sum_{j=1}^H N_j \sigma_j(\alpha^{t-1})} \text{ pour tout domaine } h.$$

- 3 Arrêter à l'étape T quand $\text{Max}|\alpha_h^T - \alpha_h^{T-1}| \leq \epsilon$, fixé.

Propriété

A toute étape t de l'algorithme précédent, $V(\alpha^t) \leq V(\alpha^{t-1})$.



Etude par simulations

Cadre

On génère une population de $N = 1\,000$ individus, partitionnés en 4 domaines U_1, \dots, U_4 de même taille.

Cadre

On génère une population de $N = 1\,000$ individus, partitionnés en 4 domaines U_1, \dots, U_4 de même taille.

Les variables d'équilibrage $\mathbf{x} = (x_1, x_2)$ sont générées de la façon suivante :

U_1	U_2
$\mathbf{x}_k = (1, 1)$	$\mathbf{x}_k = (1, 2)$
U_3	U_4
$\mathbf{x}_k = (2, 1)$	$\mathbf{x}_k = (2, 2)$

Cadre

On génère ensuite trois variables d'intérêt y_1, y_2, y_3 selon le modèle

$$y_{ik} = \phi_{ih} + \eta_{hk},$$

où les η_{hk} sont générés à l'aide d'une loi normale centrée, de façon à obtenir un R^2 de 0.6 environ (respectivement 0.3).

Cadre

On génère ensuite trois variables d'intérêt y_1, y_2, y_3 selon le modèle

$$y_{ik} = \phi_{ih} + \eta_{hk},$$

où les η_{hk} sont générés à l'aide d'une loi normale centrée, de façon à obtenir un R^2 de 0.6 environ (respectivement 0.3).

On utilise :

- $\phi_1 = (0.5, 0.5, 1.5, 1.5) \Rightarrow y_1$ liée à x_1 ,

Cadre

On génère ensuite trois variables d'intérêt y_1, y_2, y_3 selon le modèle

$$y_{ik} = \phi_{ih} + \eta_{hk},$$

où les η_{hk} sont générés à l'aide d'une loi normale centrée, de façon à obtenir un R^2 de 0.6 environ (respectivement 0.3).

On utilise :

- $\phi_1 = (0.5, 0.5, 1.5, 1.5) \Rightarrow y_1$ liée à x_1 ,
- $\phi_2 = (0.5, 1.5, 0.5, 1.5) \Rightarrow y_2$ liée à x_2 ,

Cadre

On génère ensuite trois variables d'intérêt y_1, y_2, y_3 selon le modèle

$$y_{ik} = \phi_{ih} + \eta_{hk},$$

où les η_{hk} sont générés à l'aide d'une loi normale centrée, de façon à obtenir un R^2 de 0.6 environ (respectivement 0.3).

On utilise :

- $\phi_1 = (0.5, 0.5, 1.5, 1.5) \Rightarrow y_1$ liée à x_1 ,
- $\phi_2 = (0.5, 1.5, 0.5, 1.5) \Rightarrow y_2$ liée à x_2 ,
- $\phi_3 = (0.25, 0.75, 1.25, 2.00) \Rightarrow y_3$ liée à l'interaction de x_1 et x_2 .

Simulation 1

On suppose tout d'abord connus les totaux

$$\sum_{k \in U_h} (\mathbf{x}_k \mathbf{x}'_k, \mathbf{x}_k y_k, y_k^2).$$

Simulation 1

On suppose tout d'abord connus les totaux

$$\sum_{k \in U_h} (\mathbf{x}_k \mathbf{x}'_k, \mathbf{x}_k y_k, y_k^2).$$

On compare deux stratégies : échantillonnage équilibré de taille $n = 100$ à probabilités égales (EGAL), ou selon les probabilités obtenues à l'aide de la méthode de point fixe (OPTI).

Simulation 1

On suppose tout d'abord connus les totaux

$$\sum_{k \in U_h} (\mathbf{x}_k \mathbf{x}'_k, \mathbf{x}_k y_k, y_k^2).$$

On compare deux stratégies : échantillonnage équilibré de taille $n = 100$ à probabilités égales (EGAL), ou selon les probabilités obtenues à l'aide de la méthode de point fixe (OPTI).

On compare les deux stratégies en termes d'efficacité relative

$$RE = \frac{EQM(\hat{t}_{y\pi}^{OPTI})}{EQM(\hat{t}_{y\pi}^{EGAL})},$$

où l'EQM est calculée à l'aide de 10 000 simulations indépendantes.

Résultats obtenus

	y_1	y_2	y_3
$\sigma^{(1)}$	0.90	0.88	0.86
$\sigma^{(2)}$	0.91	0.88	0.90

Simulation 2

On suppose maintenant que les totaux inconnus

$$\sum_{k \in U_h} (\mathbf{x}_k \mathbf{x}'_k, \mathbf{x}_k y_k, y_k^2)$$

sont estimés à l'aide d'un échantillon S_0 de taille $n_0 = 50$
(respectivement $n_0 = 100$) indépendant.

Simulation 2

On suppose maintenant que les totaux inconnus

$$\sum_{k \in U_h} (\mathbf{x}_k \mathbf{x}'_k, \mathbf{x}_k y_k, y_k^2)$$

sont estimés à l'aide d'un échantillon S_0 de taille $n_0 = 50$ (respectivement $n_0 = 100$) indépendant.

Les deux mêmes stratégies sont comparées. Notons que les probabilités d'inclusion obtenues par la méthode de point fixe dépendent de l'échantillon S_0 , et sont donc recalculées pour chaque simulation.

Résultats obtenus

	$\sigma^{(1)}$			$\sigma^{(2)}$		
	y_1	y_2	y_3	y_1	y_2	y_3
$n_0 = 50$	0.92	0.93	0.90	0.92	0.94	0.94
$n_0 = 100$	0.88	0.89	0.91	0.91	0.92	0.90

Perspectives

Choix du partitionnement de U

Une bonne partition de U devrait conduire à des domaines U_h dans lesquels les résidus sont approximativement constants ... mais les résidus sont inconnus au stade de l'échantillonnage.

Choix du partitionnement de U

Une bonne partition de U devrait conduire à des domaines U_h dans lesquels les résidus sont approximativement constants ... mais les résidus sont inconnus au stade de l'échantillonnage.

Les domaines U_h doivent être suffisamment grands pour que les totaux qui interviennent dans l'algorithme puissent être estimés de façon fiable.

Choix du partitionnement de U

Une bonne partition de U devrait conduire à des domaines U_h dans lesquels les résidus sont approximativement constants ... mais les résidus sont inconnus au stade de l'échantillonnage.

Les domaines U_h doivent être suffisamment grands pour que les totaux qui interviennent dans l'algorithme puissent être estimés de façon fiable.

Il est possible d'imposer une condition de taille fixe dans les domaines U_h .

Travail futur

Travail futur

Cas d'un vecteur général de variables d'équilibrage.

Travail futur

Cas d'un vecteur général de variables d'équilibrage.

Cas où l'approximation de variance Deville-Tillé ne s'applique pas (échantillonnage équilibré avec tri informatif).

Travail futur

Cas d'un vecteur général de variables d'équilibrage.

Cas où l'approximation de variance Deville-Tillé ne s'applique pas (échantillonnage équilibré avec tri informatif).

Application à une fonctionnelle non linéaire.

Bibliographie

Deville, J.-C., and Tillé, Y. (2004). *Efficient balanced sampling : the cube method*, Biometrika, 91, pages 893-912.

Deville, J.-C., and Tillé, Y. (2005). *Variance approximation under balanced sampling*, Journal of Statistical Planning and Inference, 128, pages 569-591.

Nedyalkova, D., and Tillé, Y. (2008). *Optimal sampling and estimation strategies under the linear model*, Biometrika, 95, pages 521-537.

Tillé, Y., and Favre, A.-C. (2005). *Optimal allocation in balanced sampling*. Statistics and Probability Letters, 74, pages 31-37.