

Données d'enquête

Estimation de variance et introduction à l'estimation sur petits domaines

Guillaume Chauvet

Atelier Statistique de la Sfds

Novembre 2011

Objectifs du cours

Fournir aux participants :

- ▶ un rappel des principales méthodes d'échantillonnage utilisées,
- ▶ une présentation des méthodes d'estimation de variance utilisées dans les enquêtes,
- ▶ une introduction à des méthodes d'estimation pour des populations spécifiques (domaines, petits domaines),
- ▶ des exemples pratiques d'utilisation.

Panorama du cours

1. Rappels sur l'échantillonnage
2. Estimation de variance
3. Estimation sur domaine

Partie 1

Estimation de variance

En résumé

Les estimations basées sur des données d'enquête sont issues d'un processus d'échantillonnage et de redressement complexes. Le calcul de précision de ces estimations doit prendre en compte l'ensemble des étapes de ce processus.

Pour une statistique non linéaire, il existe principalement deux possibilités pour obtenir une estimation de variance : la technique de linéarisation et les méthodes de rééchantillonnage.

Les méthodes de rééchantillonnage sont moins générales que la linéarisation, mais elles peuvent permettre à l'utilisateur de s'affranchir de la connaissance du plan de sondage pour réaliser une estimation de variance.

Rappels sur l'échantillonnage en population finie

Notations

Estimation de Horvitz-Thompson

Calcul de précision

Estimation d'un paramètre complexe

Estimateur par substitution

Précision d'un estimateur

Technique de linéarisation

Application : Enquête Logement 2006

Méthodes de rééchantillonnage

Le Jackknife

Le Bootstrap

Rééchantillonnage pour données d'enquête

Plans de sondage classiques

Estimation de variance par rééchantillonnage

Application : Enquête HHANES

Rappels sur l'échantillonnage en population finie

Notations

Notations

On se place dans le cadre d'une population finie U de N *individus* ou *unités statistiques*, supposées identifiables par un label. On notera simplement

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

On s'intéresse à une *variable d'intérêt* y (éventuellement vectorielle), qui prend la valeur y_k sur l'individu k de U . La variable y est vue ici comme non aléatoire : la population U étant fixée, la valeur prise par y sur chaque individu est parfaitement définie et déterministe.

On souhaite disposer d'indicateurs pour la population U (totaux, moyennes, fractiles, indices, ...), à l'aide de données collectées sur un échantillon S .

Exemples

Exemple 1 : Les enquêtes-ménages de l'Insee visent à décrire les conditions de vie des ménages (emploi, logement, patrimoine, ...). Les ménages enquêtés sont sélectionnés dans un échantillon de zones appelé l'*Echantillon-Maître* (communes ou groupes de communes dans le rural, pâtés de maison dans l'urbain).

Exemple 2 : Les enquêtes-entreprises sont réalisées à l'aide d'une base de sondage (répertoire SIRENE) et de sources externes. La définition de l'individu statistique est un problème à part entière (Rivière, 1998).

Paramètre d'intérêt

On s'intéresse à un *paramètre d'intérêt* de la forme

$$\theta(y_k, k \in U) \equiv \theta.$$

Un **estimateur** de ce paramètre sera de la forme

$$\begin{aligned}\hat{\theta}(y_k, k \in S) &\equiv \hat{\theta}(S) \\ &\equiv \hat{\theta},\end{aligned}$$

où S désigne l'échantillon aléatoire finalement observé.

Paramètre d'intérêt

Total et moyenne

On peut s'intéresser au total

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

d'une variable quantitative sur la population, ou encore à sa valeur moyenne

$$\mu_y = \frac{1}{N} \sum_{k \in U} y_k.$$

Exemple :

Chiffre d'affaires total des entreprises d'un secteur d'activité, pourcentage d'étudiants fumeurs, ...

Paramètre d'intérêt

Estimation sur domaine (voir Partie 2)

Un cas particulier important est celui de l'estimation sur une sous-population U_d (appelée *domaine*) d'un total

$$t_{yd} = \sum_{k \in U_d} y_k$$

ou d'une moyenne

$$\mu_{yd} = \frac{1}{N_d} \sum_{k \in U_d} y_k$$

avec N_d la taille du domaine.

Il peut s'agir d'un domaine au sens géographique (habitants d'une région), socio-démographique (individus de moins de 20 ans), temporel (individus présents à une date donnée), ...

Paramètre d'intérêt

Estimation par substitution

Savoir estimer un total permet de traiter le cas de très nombreux paramètres qui peuvent s'exprimer comme des fonctions de totaux. C'est le cas d'un ratio, d'un coefficient de corrélation, d'une variance, d'un coefficient de régression, ...

Ces paramètres sont estimés par substitution, en estimant chaque total inconnu par son estimateur.

Exemple 1 :

$$R = \frac{t_y}{t_x} \quad \text{estimé par} \quad \hat{R} = \frac{\hat{t}_y}{\hat{t}_x}.$$

Plan de sondage

On suppose ici que la sélection de l'échantillon aléatoire S se fait à l'aide d'un plan de sondage p sur U , c'est à dire à l'aide d'une loi de probabilité sur les parties de U :

$$\forall s \subset U \quad p(s) \geq 0 \text{ et } \sum_{s \subset U} p(s) = 1.$$

On utilisera la notation S pour l'échantillon aléatoire et s pour un échantillon particulier. On distingue également

- ▶ l'estimateur $\hat{\theta}(y_k, k \in S) \equiv \hat{\theta}(S)$,
- ▶ l'estimation $\hat{\theta}(y_k, k \in s) \equiv \hat{\theta}(s)$.

La taille de l'échantillon S , a priori aléatoire, est notée $n(S)$.

Exemple

Soit la population $U = \{1, 2, 3, 4\}$, et $p(\cdot)$ le plan de sondage défini par :

$$\begin{array}{llll} p(\{1, 2\}) & = & 0.2 & p(\{1, 4\}) & = & 0.1 & p(\{3, 4\}) & = & 0.3 \\ p(\{1, 2, 3\}) & = & 0.3 & p(\{2, 3, 4\}) & = & 0.1 & & & \end{array}$$

A la différence des lois de probabilités classiques (normale, exponentielle, binomiale, ...) l'aléatoire ne porte pas sur la variable mais sur le sous-ensemble d'individus observés.

Mesures de précision

L'espérance d'un estimateur $\hat{\theta}$ est donnée par la valeur moyenne des estimations :

$$E_p \left[\hat{\theta} \right] = \sum_{s \subset U} p(s) \hat{\theta}(s).$$

Le biais d'un estimateur $\hat{\theta}$ correspond à l'erreur moyenne :

$$\begin{aligned} B_p \left[\hat{\theta} \right] &= E_p \left[\hat{\theta} - \theta \right] \\ &= \sum_{s \subset U} p(s) \left[\hat{\theta}(s) - \theta \right]. \end{aligned}$$

Mesures de précision

On s'intéressera également :

- ▶ à la Variance

$$\begin{aligned}V_p [\hat{\theta}] &= E_p \left[\left(\hat{\theta} - E_p \hat{\theta} \right)^2 \right] \\ &= \sum_{s \subset U} p(s) \left[\hat{\theta}(s) - E_p \hat{\theta} \right]^2,\end{aligned}$$

- ▶ à l'Erreur Quadratique Moyenne (EQM)

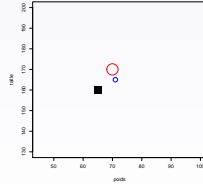
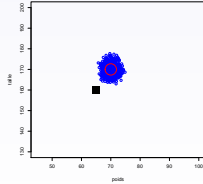
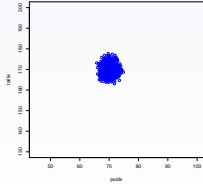
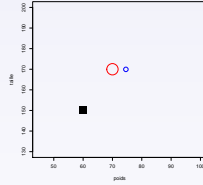
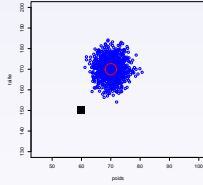
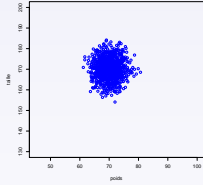
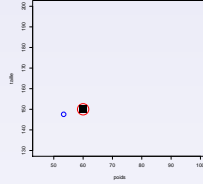
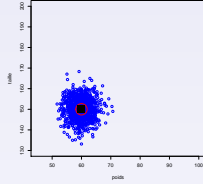
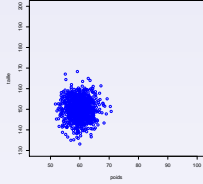
$$\begin{aligned}EQM [\hat{\theta}] &= E_p \left[\left(\hat{\theta} - \theta \right)^2 \right] \\ &= B_p \left[\hat{\theta} \right]^2 + V_p \left[\hat{\theta} \right].\end{aligned}$$

Quelques simulations

Pour illustrer la notion de biais et de variance, on considère l'exemple d'une population (fictive) de $N = 1\,000$ individus âgés de 15 à 20 ans.

Dans cette population, un échantillon de taille $n = 50$ est sélectionné et enquêté. Pour chaque individu enquêté, on obtient son poids (en kg), sa taille (en cm) et son âge.

On s'intéresse à l'estimation du poids moyen et de la taille moyenne (point noir). Chaque échantillon permet d'obtenir une estimation (point bleu) de ces paramètres. La moyenne des estimations est représentée par le point rouge.



Probabilités d'inclusion d'ordre 1

On note π_k la *probabilité d'inclusion* de l'unité k , c'est à dire la probabilité que l'unité k soit retenue dans l'échantillon :

$$\pi_k = \mathbb{P}(k \in S) = \sum_{s/k \in s} p(s)$$

La somme des probabilités d'inclusion donne la taille moyenne de l'échantillon sélectionné :

$$\sum_{k \in U} \pi_k = E_p [n(S)] .$$

En pratique, les probabilités d'inclusion π_k sont fixées avant le tirage à l'aide d'une **information auxiliaire**. On utilise ensuite un plan de sondage qui respecte ces probabilités d'inclusion.

Probabilités d'inclusion d'ordre 2

On note π_{kl} la probabilité que deux unités distinctes k et l soient sélectionnées conjointement dans l'échantillon :

$$\pi_{kl} = \mathbb{P}(k, l \in S) = \sum_{s/k, l \in s} p(s)$$

Ces probabilités doubles interviennent notamment dans la variance des estimateurs. Il est souvent difficile de les calculer exactement, sauf pour des plans de sondage particuliers.

Estimation de Horvitz-Thompson

Objectif

Nous nous intéressons dans un premier temps à l'estimation du total

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

de la variable y , et éventuellement à l'estimation d'une moyenne

$$\mu_y = \frac{1}{N} \sum_{k \in U} y_k.$$

La π -estimation

La connaissance des probabilités π_k permet une estimation sans biais d'un total sous le plan de sondage, i.e. sous le mécanisme aléatoire associé au plan de sondage. Ainsi, le total t_y est estimé sans biais par

$$\hat{t}_{y\pi} = \sum_{k \in S} \frac{y_k}{\pi_k} = \sum_{k \in U} \frac{y_k}{\pi_k} I_k \quad (1)$$

si tous les π_k sont > 0 . On parle d'estimateur de Horvitz-Thompson ou encore de π -estimateur.

La π -estimation

L'estimateur de Horvitz-Thompson est un estimateur pondéré :

$$\hat{t}_{y\pi} = \sum_{k \in S} d_k y_k,$$

où les *poids de sondage* $d_k = 1/\pi_k$ ne dépendent pas de la variable d'intérêt.

⇒ les mêmes poids peuvent être utilisés pour toutes les variables.

Principe : un individu k de l'échantillon représente $d_k = \frac{1}{\pi_k}$ individus de la population.

Biais de couverture

Si certaines probabilités d'inclusion sont nulles, le π -estimateur est biaisé :

$$\begin{aligned} E \left[\hat{t}_{y\pi} \right] &= \sum_{\substack{k \in U \\ \pi_k > 0}} y_k \\ &= t_y - \sum_{\substack{k \in U \\ \pi_k = 0}} y_k. \end{aligned}$$

Ce problème peut notamment se poser :

- ▶ en cas de défaut de couverture de la base de sondage (liste des individus pas à jour, ou individus impossibles à joindre),
- ▶ quand on choisit de laisser de côté une partie de la population (cut-off sampling, parfois utilisé dans les enquêtes-entreprise).



Enquête "Sans-Domicile 2001" (De Peretti et al., 2006)

Sans-domicile : personne qui dort dans un lieu non prévu pour l'habitation ou prise en charge par un organisme fournissant un hébergement gratuit ou à faible participation.

Méthode d'échantillonnage indirect : sélection d'un échantillon de jours \times services d'aide (hébergement, restauration).

Champ de l'enquête : sans-domicile ayant fréquenté, au moins une fois dans la semaine d'enquête, soit un service d'hébergement, soit une distribution de repas chauds.

Exclut les personnes :

- ▶ qui dorment dans la rue pour une période de temps courte et ne font pas appel à un centre ou à une distribution de repas,
- ▶ qui ne font pas (ou ne peuvent pas faire) appel au circuit d'assistance.

Variance

La variance du π -estimateur est donnée par

$$V_p [\hat{t}_{y\pi}] = \sum_{k,l \in U} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} \Delta_{kl}. \quad (2)$$

Cette variance peut être estimée sans biais par

$$v_{HT} [\hat{t}_{y\pi}] = \sum_{k,l \in S} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} \frac{\Delta_{kl}}{\pi_{kl}} \quad (3)$$

si tous les π_{kl} sont strictement positifs. On parle de l'estimateur de variance de Horvitz-Thompson.

Variance

Pour un plan de taille fixe n (i.e., seuls les échantillons de taille n ont une probabilité non nulle d'être sélectionnés), la variance du π -estimateur peut se réécrire sous la forme

$$V_p [\hat{t}_{y\pi}] = -\frac{1}{2} \sum_{k \neq l \in U} \left[\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l} \right]^2 \Delta_{kl}. \quad (4)$$

Cette variance peut être estimée sans biais par

$$v_{YG} [\hat{t}_{y\pi}] = -\frac{1}{2} \sum_{k \neq l \in S} \left[\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l} \right]^2 \frac{\Delta_{kl}}{\pi_{kl}} \quad (5)$$

si tous les π_{kl} sont strictement positifs. On parle de l'estimateur de variance de Yates-Grundy.

Biais de l'estimateur de variance

Proposition

Pour un plan de sondage quelconque, on a :

$$E_p [v_{HT} [\hat{t}_{y\pi}]] = V_p [\hat{t}_{y\pi}] + \sum_{\substack{k,l \in U \\ \pi_{kl}=0}} y_k y_l.$$

Pour un plan de sondage de taille fixe, on a :

$$E_p [v_{YG} [\hat{t}_{y\pi}]] = V_p [\hat{t}_{y\pi}] - \frac{1}{2} \sum_{\substack{k,l \in U \\ \pi_{kl}=0}} \pi_k \pi_l \left[\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l} \right]^2.$$

Si y est à valeurs positives, les deux estimateurs de variance sont respectivement biaisés positivement et négativement si certaines probabilités π_{kl} sont nulles.

Choix des probabilités d'inclusion

D'après la formule de Yates-Grundy, la variance est nulle si les probabilités d'inclusion sont proportionnelles à la variable d'intérêt.

En pratique, on peut définir ces probabilités d'inclusion proportionnellement à une mesure de taille :

- ▶ une enquête comporte généralement de nombreuses variables d'intérêt,
- ▶ ces variables sont inconnues au stade de l'échantillonnage.

Interprétation : si les individus peuvent être de tailles très différentes, on utilise les probabilités d'inclusion pour lisser les rapports y_k/π_k .

Probabilités proportionnelles à la taille

La taille moyenne d'échantillon sélectionné est donnée par

$$E_p [n(S)] = \sum_{k \in U} \pi_k.$$

Si n désigne la taille d'échantillon souhaitée, les probabilités d'inclusion proportionnelles à une variable auxiliaire positive x sont données par

$$\pi_k = n \frac{x_k}{\sum_{l \in U} x_l}.$$

Si certaines unités sont particulièrement grosses, on peut obtenir des probabilités d'inclusion supérieures à 1 \Rightarrow on sélectionne d'office les unités correspondantes, et on recalcule les probabilités d'inclusion des autres unités.

Intervalle de confiance

Intervalle de confiance

On suppose que $\hat{t}_{y\pi}$ estime sans biais t_y . Alors un intervalle de confiance pour t_y de niveau approximatif $1 - \alpha$ est donné par :

$$IC_{1-\alpha} [t_y] = \left[\hat{t}_{y\pi} \pm z_{1-\frac{\alpha}{2}} \sqrt{V_p [\hat{t}_{y\pi}]} \right],$$

avec $z_{1-\frac{\alpha}{2}}$ le quantile d'ordre $1 - \frac{\alpha}{2}$ d'une loi normale centrée réduite $\mathcal{N}(0, 1)$.

Rappel :

- ▶ $\alpha = 0.05 \Rightarrow z_{0.975} = 1.96$
- ▶ $\alpha = 0.10 \Rightarrow z_{0.95} = 1.64$

Interprétation (pour $\alpha = 0.05$) : le vrai total t_y est contenu dans l'intervalle de confiance pour (approximativement) 95% des échantillons.

Intervalle de confiance

Comme la vraie variance $V_p [\hat{t}_{y\pi}]$ est généralement inconnue, on la remplace par un estimateur noté $v [\hat{t}_{y\pi}]$.

On obtient l'intervalle de confiance estimé :

$$\widehat{IC}_{1-\alpha} [t_y] = \left[\hat{t}_{y\pi} \pm z_{1-\frac{\alpha}{2}} \sqrt{v [\hat{t}_{y\pi}]} \right].$$

L'intervalle de confiance est (approximativement) valide :

- ▶ si l'estimateur $\hat{t}_{y\pi}$ suit approximativement une loi $\mathcal{N} [t_y, V_p (\hat{t}_{y\pi})]$,
- ▶ si l'estimateur de variance $v (\hat{t}_{y\pi})$ est faiblement consistant.

Coefficient de variation

La précision de l'estimation du total peut également être donnée sous la forme du coefficient de variation

$$CV_p [\hat{t}_{y\pi}] = \frac{\sqrt{V_p(\hat{t}_{y\pi})}}{\hat{t}_{y\pi}} \quad \text{estimé par} \quad \hat{C}V [\hat{t}_{y\pi}] = \frac{\sqrt{v(\hat{t}_{y\pi})}}{\hat{t}_{y\pi}}.$$

Il s'agit d'une grandeur sans dimension, plus facile à comparer et à interpréter que la variance.

Coefficient de variation

Avec un niveau de confiance de 0.95, l'intervalle de confiance du total est donné par

$$\begin{aligned}\widehat{IC}_{0.95} [t_y] &= \left[\hat{t}_{y\pi} \pm 1.96 \sqrt{v [\hat{t}_{y\pi}]} \right] \\ &= \hat{t}_{y\pi} \left[1 \pm 1.96 \frac{\sqrt{v [\hat{t}_{y\pi}]}}{\hat{t}_{y\pi}} \right] \\ &\simeq \hat{t}_{y\pi} \left[1 \pm 2 \widehat{CV} [\hat{t}_{y\pi}] \right].\end{aligned}$$

Interprétation : un CV de $x\%$ correspond à un total connu à plus ou moins $2 x\%$, avec un niveau de confiance de 0.95.

Estimation d'un paramètre complexe

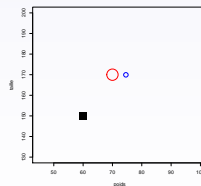
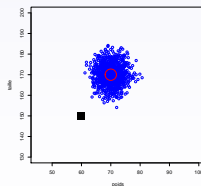
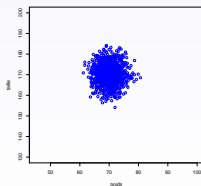
Précision d'un estimateur

Mesures de précision

Soit $\hat{\theta}$ l'estimateur d'un paramètre θ . La précision de cet estimateur peut être mesurée par :

- ▶ son biais $B_p [\hat{\theta}] = E_p [\hat{\theta} - \theta]$
- ▶ sa variance $V_p [\hat{\theta}] = E_p \left[(\hat{\theta} - E_p \hat{\theta})^2 \right]$
- ▶ son Erreur Quadratique Moyenne

$$EQM [\hat{\theta}] = B_p [\hat{\theta}]^2 + V_p [\hat{\theta}].$$



Biais d'un estimateur

On recherche un estimateur efficace pour le paramètre θ , mais pas nécessairement non biaisé :

- ▶ pour un paramètre θ donné, il peut être difficile voire impossible de trouver un estimateur sans biais,
- ▶ un estimateur faiblement biaisé peut avoir une variance beaucoup plus faible qu'un estimateur non biaisé.

Cependant, le biais a des conséquences notamment en termes de taux de couverture réel des intervalles de confiance. On utilisera en pratique des estimateurs dont le biais est d'un ordre de grandeur plus faible que la variance.

Estimateur par substitution

Si les probabilités d'inclusion d'ordre 1 et 2 sont connues, il est possible d'estimer sans biais un total, et d'obtenir une mesure de précision de cette estimation.

En pratique, on peut s'intéresser à des paramètres plus complexes :

- ▶ Estimation d'un ratio (ex : taux de chômage),
- ▶ Estimation d'un coefficient de régression, ou d'un coefficient de corrélation,
- ▶ Estimation d'un fractile ou d'un indice (ex : mesure des inégalités avec l'indice de Gini ou l'indice de Theil).

On suppose que le paramètre à estimer est de la forme

$$\theta = f(t_{\mathbf{y}})$$

où $\mathbf{y}_k = (y_{1k}, \dots, y_{qk})^T$ désigne un q -vecteur de variables d'intérêt,
et $f : \mathbf{R}^q \rightarrow \mathbf{R}$.

Définition

On dit que la fonction f (et par extension, le paramètre θ) est homogène d'ordre m si

$$\forall \alpha > 0 \quad \forall \mathbf{x} \in \mathbf{R}^q \quad f(\alpha \mathbf{x}) = \alpha^m f(\mathbf{x})$$

Exemples

1. $\theta = t_{y1}$ $f(x) = x$
paramètre homogène d'ordre 1
2. $\theta = \frac{t_{y1}}{t_{y2}}$ $f(x_1, x_2) = \frac{x_1}{x_2}$
paramètre homogène d'ordre 0
3. $\theta = t_{y1} \times t_{y2}$ $f(x_1, x_2) = x_1 \times x_2$
paramètre homogène d'ordre 2

Remarques

Si f est homogène d'ordre m , alors

$$\begin{aligned}\theta &= f(N \mu_{\mathbf{y}}) \\ &= N^m f(\mu_{\mathbf{y}})\end{aligned}$$

où la moyenne $\mu_{\mathbf{y}} = t_{\mathbf{y}}/N$ est homogène d'ordre 0 (ou sans dimension). Le paramètre θ est donc "de l'ordre de N^m ".

Si f est homogène d'ordre m et deux fois différentiable, alors les différentielles f' et f'' sont homogènes d'ordre $m - 1$ et $m - 2$, respectivement.

Estimation du paramètre

Il semble naturel d'estimer θ en remplaçant le total t_y inconnu par son π -estimateur. On obtient l'**estimateur par substitution** :

$$\hat{\theta}_\pi = f(\hat{t}_{y\pi}).$$

Questions :

1. Cet estimateur est-il (approximativement) sans biais?
2. Quelle est la variance (approchée) de cet estimateur?

Technique de linéarisation

Principe

Soit $\theta = f(t_{\mathbf{y}})$, avec f une fonction homogène d'ordre m . En utilisant un développement de Taylor à l'ordre 1, on obtient :

$$\begin{aligned}\hat{\theta}_{\pi} - \theta &= f(\hat{t}_{\mathbf{y}\pi}) - f(t_{\mathbf{y}}) \\ &\simeq [f'(t_{\mathbf{y}})]^T [\hat{t}_{\mathbf{y}\pi} - t_{\mathbf{y}}] \\ &= \hat{t}_{u\pi} - t_u,\end{aligned}\tag{6}$$

en notant

$$\begin{aligned}u_k &= [f'(t_{\mathbf{y}})]^T [\mathbf{y}_k] \\ &= \sum_{i=1}^q \frac{\partial f}{\partial x_i}(t_{\mathbf{y}}) y_{ik}.\end{aligned}$$

Sous l'approximation (6), on obtient donc

$$E_p \left[\hat{\theta}_{\pi} - \theta \right] \simeq 0.$$

Principe (suite)

Le résultat précédent implique que l'estimateur par substitution $\hat{\theta}$ est approximativement non biaisé. On peut obtenir un résultat plus précis à l'aide d'un développement de Taylor à l'ordre 2. On a :

$$\begin{aligned} N^{-m} \left[\hat{\theta}_{\pi} - \theta \right] &= f(\hat{\mu}_{\mathbf{y}\pi}) - f(\mu_{\mathbf{y}}) \\ &\simeq X_1 + X_2, \end{aligned}$$

avec

$$\begin{aligned} X_1 &= [f'(\mu_{\mathbf{y}})]^T [\hat{\mu}_{\mathbf{y}\pi} - \mu_{\mathbf{y}}] = O_p(1/\sqrt{n}) \\ X_2 &= [\hat{\mu}_{\mathbf{y}\pi} - \mu_{\mathbf{y}}]^T f''(\mu_{\mathbf{y}}) [\hat{\mu}_{\mathbf{y}\pi} - \mu_{\mathbf{y}}] = O_p(1/n). \end{aligned}$$

Le terme X_1 est donc majoritaire dans le développement de Taylor. Sa variance est utilisée pour obtenir une approximation de la variance de $\hat{\theta}_{\pi}$.

Proposition

Soit $\theta = f(t_y)$, où f est une fonction homogène d'ordre m , deux fois différentiable et dont les dérivées secondes sont continues au point μ_y . Alors l'approximation par linéarisation de la variance de $\hat{\theta}_\pi$ est donnée par :

$$V_p [\hat{t}_{u\pi}],$$

où

$$u_k \equiv u_k(\theta) = [f'(t_y)]^T [\mathbf{y}_k]$$

est appelée la variable linéarisée du paramètre θ .

Pour obtenir la variance (approchée) de $\hat{\theta}_\pi$, il suffit donc :

- ▶ de disposer d'une formule (ou d'un estimateur) de variance associée au plan de sondage utilisé,
- ▶ de calculer la variable linéarisée du paramètre θ .

Quelques règles de calcul

Linéarisée d'une somme : $u_k(\theta_1 + \theta_2) = u_k(\theta_1) + u_k(\theta_2)$.

Linéarisée d'un produit : $u_k(\theta_1 \times \theta_2) = \theta_1 \times u_k(\theta_2) + \theta_2 \times u_k(\theta_1)$.

Linéarisée d'une fonction : Si θ est un paramètre scalaire, et g est une fonction dérivable $\mathbf{R} \rightarrow \mathbf{R}$, alors

$$u_k [g(\theta)] = g'(\theta) u_k(\theta).$$

Corollaire de la règle précédente :

$$u_k [\ln(\theta)] = \frac{u_k(\theta)}{\theta}.$$

Exemple du ratio

Paramètre $R = \frac{t_y}{t_x}$, homogène d'ordre $m = 0$.

Calcul de la variable linéarisée :

$$f(x_1, x_2) = \frac{x_1}{x_2} \quad f'(x_1, x_2) = \left(\frac{1}{x_2}, -\frac{x_1}{(x_2)^2} \right)$$

$$u_k(R) = \frac{1}{t_x} y_k - \frac{t_y}{(t_x)^2} x_k = \frac{1}{t_x} (y_k - R x_k)$$

Variable linéarisée estimée :

$$\hat{u}_k(R) = \frac{1}{\hat{t}_{x\pi}} (y_k - \hat{R} x_k).$$

La variable linéarisée peut également être calculée à l'aide de règles de calcul, proches de celles de la dérivation, et rappelées par Deville (1999).

Linéarisation de Taylor

La notion de linéarisation peut être étendue sous une condition de différentiabilité plus faible. Si le paramètre d'intérêt est différentiable au sens de Gâteaux, Deville (1999) introduit la linéarisation selon la fonction d'influence : la variable linéarisée est alors définie comme la fonction d'influence du paramètre (Hampel et al., 1986, Goga et al., 2009).

Une approche basée sur les équations estimantes est également proposée par Kovacevic et Binder (1997).

Estimation de variance

Estimation de variance

En pratique, la formule de variance donnée par linéarisation n'est pas utilisable car :

- ▶ seuls les individus $k \in S$ sont observés,
- ▶ la variable linéarisée u_k dépend de paramètres inconnus.

L'estimateur de variance par linéarisation s'obtient en 4 étapes :

1. Calcul de la variable linéarisée u_k ,
2. Formule de variance pour $\hat{t}_{u\pi}$, le π -estimateur du total de u_k ,
3. Estimateur de variance (intermédiaire) pour $\hat{t}_{u\pi}$: certains paramètres sont inconnus,
4. Estimateur de variance final : les paramètres inconnus sont estimés.

Exemple du ratio(suite)

On suppose que l'échantillon a été sélectionné selon un SRS(n).

$$\text{Etape 1 : } V_p \left[\hat{R}_\pi \right] \simeq V_p \left[\hat{t}_{u\pi} \right] \text{ avec } u_k = \frac{1}{t_x} (y_k - R x_k)$$

$$\begin{aligned} \text{Etape 2 : } V_p \left[\hat{t}_{u\pi} \right] &= N^2 \frac{1-f}{n} S_u^2 \\ &= \frac{N^2}{t_x^2} \frac{1-f}{n} \left[S_y^2 - 2R S_{xy} + R^2 S_x^2 \right] \\ &= \frac{1-f}{n} \frac{1}{\mu_x^2} \left[S_y^2 - 2R S_{xy} + R^2 S_x^2 \right] \end{aligned}$$

Exemple du ratio(suite)

On suppose que l'échantillon a été sélectionné selon un SRS(n).

$$\begin{aligned}\text{Etape 3 : } \tilde{v} [\hat{t}_{u\pi}] &= N^2 \frac{1-f}{n} s_u^2 \\ &= \frac{1-f}{n \mu_x^2} [s_y^2 - 2R s_{xy} + R^2 s_x^2]\end{aligned}$$

$$\text{Etape 4 : } v [\hat{R}_\pi] = \frac{1-f}{n \bar{x}^2} [s_y^2 - 2\hat{R}_\pi s_{xy} + \hat{R}_\pi^2 s_x^2].$$

Exemple

Population U de taille 10, dans laquelle un échantillon de taille $n = 4$ est sélectionné selon un SAS. Estimation du ratio t_y/t_x .

k	x_k	y_k	
1	5	1	
2	1	3	
3	4	2	
4	8	10	
	$\bar{x} = 4.5$ $s_x^2 = 8.3$	$\bar{y} = 4$ $s_y^2 = 16.7$	

$$\hat{t}_{x\pi} = N\bar{x} = 45$$

$$\hat{t}_{y\pi} = N\bar{y} = 40$$

$$v[\hat{t}_{x\pi}] = N^2 \frac{1-f}{n} s_x^2 = 125$$

$$v[\hat{t}_{y\pi}] = N^2 \frac{1-f}{n} s_y^2 = 250$$

Exemple

Population U de taille 10, dans laquelle un échantillon de taille $n = 4$ est sélectionné selon un SAS. Estimation du ratio t_y/t_x .

k	x_k	y_k	$\hat{u}_k = \frac{1}{\hat{t}_{x\pi}}(y_k - \hat{R}x_k)$
1	5	1	-0.08
2	1	3	0.05
3	4	2	-0.03
4	8	10	0.06
	$\bar{x} = 4.5$ $s_x^2 = 8.3$	$\bar{y} = 4$ $s_y^2 = 16.7$	$\bar{\hat{u}} = 0$ $s_{\hat{u}}^2 = 4.4 \cdot 10^{-3}$

$$\hat{t}_{x\pi} = N\bar{x} = 45$$

$$\hat{t}_{y\pi} = N\bar{y} = 40$$

$$\hat{R} = \frac{\bar{y}}{\bar{x}} = 0.89$$

$$v[\hat{t}_{x\pi}] = N^2 \frac{1-f}{n} s_x^2 = 125$$

$$v[\hat{t}_{y\pi}] = N^2 \frac{1-f}{n} s_y^2 = 250$$

$$v[\hat{R}] = N^2 \frac{1-f}{n} s_{\hat{u}}^2 = 0.07$$

La linéarisation : résumé

Avantages :

- ▶ S'adapte à un plan de sondage (presque) quelconque
- ▶ Utilisable avec un logiciel standard d'estimation de variance (tel que POULPE) → mettre en entrée la variable linéarisée
- ▶ Calcul rapide

Inconvénients :

- ▶ Nécessite le calcul d'une variable linéarisée pour chaque statistique
- ▶ Nécessite une connaissance précise du plan de sondage
- ▶ Difficile à utiliser si la stratégie d'estimation est complexe

Application

Enquête Logement 2006

En résumé

L'Enquête Logement 2006 est une enquête auprès des ménages, qui a donné lieu à une extension régionale et à plusieurs extensions locales au niveau de la région Bretagne notamment. Un complément d'échantillon a également été sélectionné dans des bases externes.

Un plan de sondage et une technique d'estimation complexes ont été nécessaires pour la mise en commun et l'exploitation des différents sous-échantillons, la prise en compte de la non-réponse et le redressement des estimateurs.

Un outil SAS de calcul de précision basé sur les formules d'estimation de variance a été mis au point pour les partenaires de l'Enquête et les chargés d'étude de la Direction Régionale.

Présentation de l'enquête

L'Enquête Logement est une des plus grosses enquêtes réalisées par l'Insee auprès des ménages. Elle a lieu environ tous les quatre ans (dernières éditions en 2002 et 2006).

Le champ est celui des logements résidences principales en 2006, accessibles à l'aide du RP99 et de la Base de Sondage de Logements Neufs (BSLN).

Objectifs de l'enquête :

- ▶ connaître le parc de logements (ancienneté de la construction, nombre de maisons individuelles/appartements, nombre de propriétaires/locataires,...),
- ▶ décrire les conditions de vie des ménages (mobilités et causes de mobilité, confort du logement, emprunts,...).

Sélection de l'Enquête Logement

L'échantillon est sélectionné en quatre temps :

- ▶ Sélection de l'échantillon national dans l'Echantillon Maître de 99 (RP99, BSLN),
- ▶ Sélection d'une extension régionale dans l'EMEX, pour les régions concernées,
- ▶ Sélection d'extensions d'échantillon au niveau local, pour les régions concernées,
- ▶ Sélection d'échantillons complémentaires dans des bases externes.

L'EM 99

L'échantillon maître de 1999 (EM99) est une réserve de logements destinée à servir de base de sondage pour les enquêtes auprès des ménages.

Il est obtenu par un tirage stratifié (selon le degré d'urbanisation) et à plusieurs degrés. On sélectionne des communes dans le rural, des districts (pâtés de maisons) dans l'urbain, ... (Ardilly, 2006).

Les échantillons destinés aux enquêtes ménages seront ensuite tirés dans les zones sélectionnées. Dans ces zones, une liste à jour de logements est fournie par le RP99 et la BSLN.

L'EMEX

Pour les extensions régionales, il existe un échantillon-maître spécifique : l'EMEX (Bourdalle et al., 2000), constitué selon des principes voisins de l'EM :

- ▶ tirage stratifié (selon le degré d'urbanisation) et à plusieurs degrés,
- ▶ même système de rotation des logements situés dans l'EMEX,
- ▶ disjonction par rapport à l'EM.

A quoi servent les extensions? La précision est liée à la taille d'échantillon : plus le domaine d'estimation est petit, plus la précision se détériore. La sélection d'un échantillon dans l'EMEX vise à assurer de meilleures estimations au niveau régional.

Extensions locales d'échantillon

Un complément d'échantillon peut être sélectionné autour de zones particulières pour lesquelles on souhaite produire des estimations fiables.

En Bretagne, c'est le cas des 6 principales aires urbaines (Brest, Lorient, Quimper, Rennes, Saint-Brieuc et Vannes).

Cet échantillon est sélectionné en excluant les logements précédemment échantillonnés dans l'EM ou l'EMEX au titre de l'Enquête Logement.

Bases externes

Enfin, pour surreprésenter des sous-populations particulières, des échantillons ont été sélectionnés dans des fichiers externes :

- ▶ Base des adresses situées dans les Zones Urbaines Sensibles (ZUS),
- ▶ Base d'allocataires de prestations.

Cette sélection n'est pas disjointe de celle des autres sous-échantillons.

Au niveau de la Bretagne, seul l'échantillon ZUS a été utilisé (fusion des autres échantillons très problématique).

Schéma récapitulatif

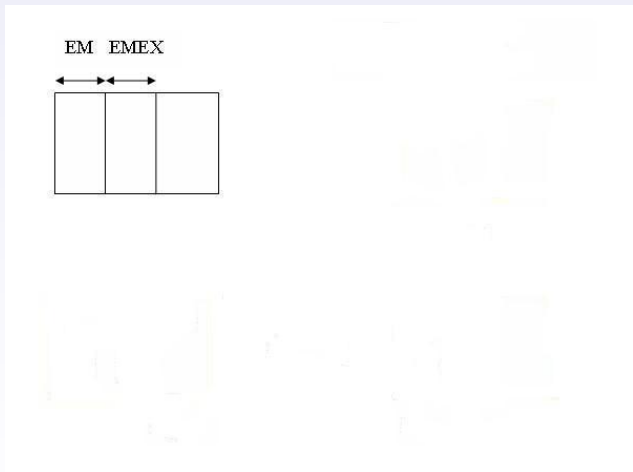


Schéma récapitulatif

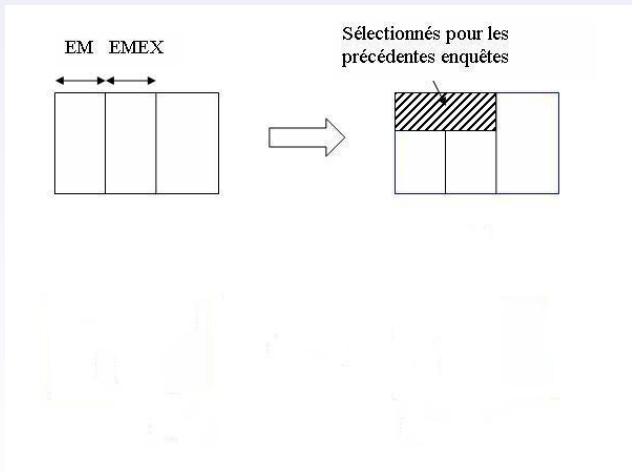


Schéma récapitulatif

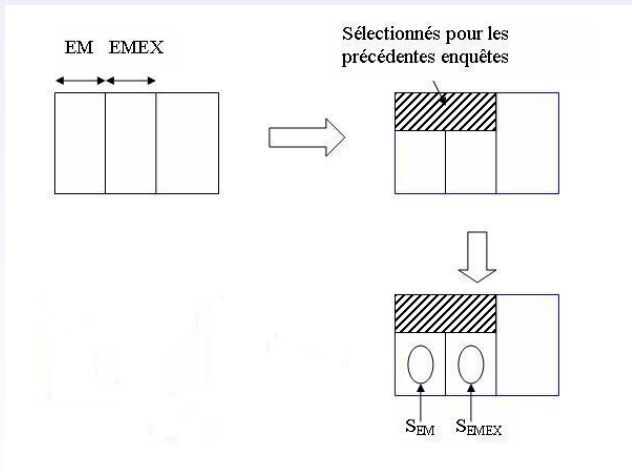


Schéma récapitulatif

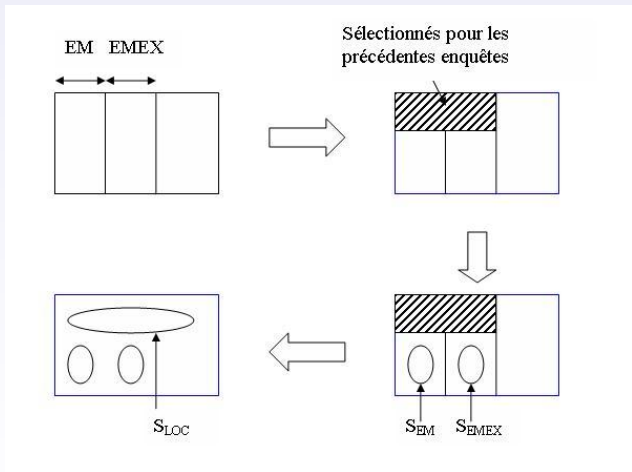


Schéma récapitulatif

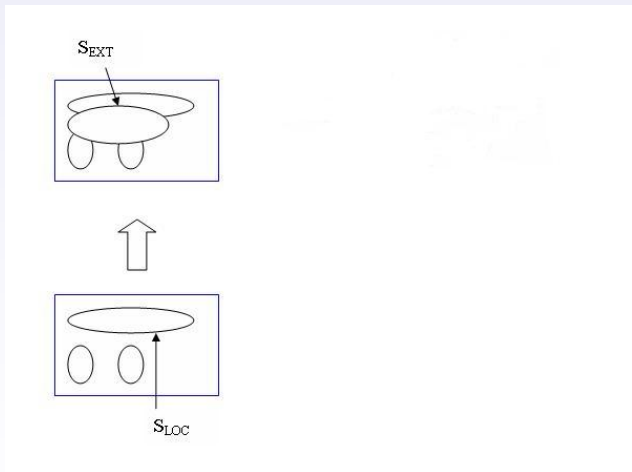
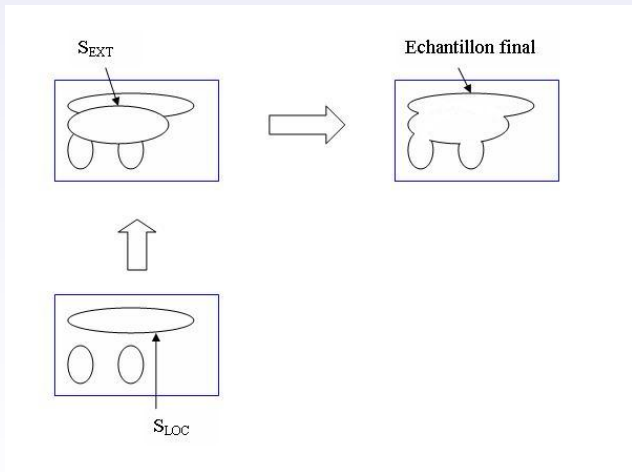


Schéma récapitulatif



Mise en commun des sous-échantillons

La difficulté consiste à produire un estimateur sans biais en gérant les trois sous-échantillons et leur intersection. Il y a essentiellement deux problèmes :

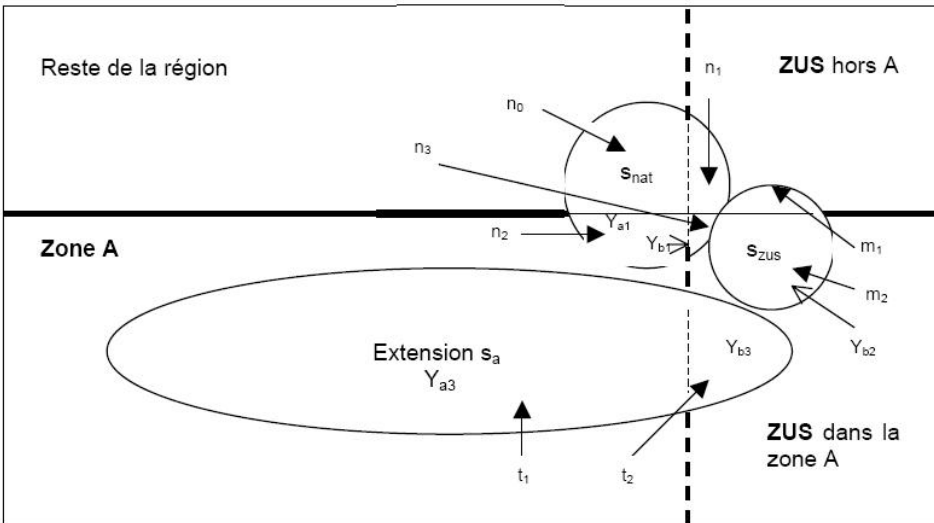
- ▶ Des sous-échantillons même disjoints peuvent représenter une même sous-population. On a donc un risque de biais dû à des doubles ou triples comptes.
- ▶ Certains logements sont sélectionnés dans deux échantillons différents.

Les sous-échantillons sont mis en commun à l'aide de la technique d'estimation composite. On note

$$\hat{t}_{y,d} = \sum_{k \in S} d_k y_k$$

l'estimateur obtenu, avec S la réunion des sous-échantillons et d_k le poids du logement k .

Schéma récapitulatif (Le Guennec, 2009)



Estimation de précision

Les étapes de l'enquête

Les différentes étapes de traitement de l'enquête (Le Guennec, 2009) sont :

1. Sélection des sous-échantillons,
2. Mise en commun des sous-échantillons,
3. Redressement de la non-réponse partielle (imputation),
4. Redressement de la non-réponse totale (méthode des groupes homogènes de réponse),
5. Calage sur une information externe.

L'estimation de variance réalisée ne prend pas en compte la variance d'imputation.

Calcul de variance (1)

On note

$$\hat{t}_{y,w} = \sum_{k \in S} w_k y_k$$

l'estimateur obtenu avec les poids calés w_k . On a :

$$V[\hat{t}_{y,w}] \simeq V[\hat{t}_{e,d}]$$

avec e_k le résidu de régression de y sur les variables de calage.

Cette variance se décompose en

$$V[\hat{t}_{e,d}] = V_p[\hat{t}_{e,d}] + V_{nr}[\hat{t}_{e,d}],$$

où la variance due à l'échantillonnage $V_p[\cdot]$ et la variance due à la non-réponse $V_{nr}[\cdot]$ sont estimées séparément.

Calcul de variance (2)

L'estimateur $\hat{t}_{e,d}$ peut se décomposer sur les trois sous-échantillons tirés nationalement (N), localement (L) ou dans les ZUS (Z) :

$$\hat{t}_{e,d} = \hat{t}_{\tilde{e},d}^N + \hat{t}_{\tilde{e},d}^L + \hat{t}_{\tilde{e},d}^Z$$

avec $\tilde{e}_k = f(e_k)$ la variable synthétique associée à la technique d'estimation composite.

En utilisant l'indépendance (réelle ou approchée) des trois sous-échantillons :

$$V_p [\hat{t}_{e,d}] \simeq V_p [\hat{t}_{\tilde{e},d}^N] + V_p [\hat{t}_{\tilde{e},d}^L] + V_p [\hat{t}_{\tilde{e},d}^Z].$$

Un exemple de calcul de précision

On souhaite estimer, sur l'ensemble des résidences principales de l'aire urbaine de Rennes :

- ▶ La structure des logements selon le nombre de chambres,
- ▶ La surface moyenne par habitant.

On est dans le cas d'une estimation sur un **domaine** D , c'est à dire sur une sous-population de U . Cette estimation ne pose pas de problème particulier, en remarquant que

$$t_{yD} = \sum_{k \in D} y_k = \sum_{k \in U} y_k 1_{k \in D}$$

où $1_{k \in D}$ vaut 1 si le logement k est dans le domaine, et 0 sinon.

On va donc estimer :

- ▶ des effectifs (nombre de logements ne comptant aucune chambre, comptant 1 chambre, ...),
- ▶ un ratio (surface totale rapportée au nombre d'habitants).

Dans le premier cas, on estime un total (variable indicatrice pour un effectif).

Dans le second cas, on estime un ratio de totaux : l'estimation de variance se fait par linéarisation (Deville, 1999).

Résultats obtenus

Paramètre	Estim.	Var.	CV (%)	BI (95%)	BS (95%)	DEFF	DCAL	NR (%)
Surf. moy.	38,27	0,23	1,25	37,34	39,21	0,48	0,42	21,79
% Log.								
0 cha.	0,08	$4,3 \cdot 10^{-5}$	7,75	0,07	0,10	0,46	0,99	17,34
1 cha.	0,18	$1,1 \cdot 10^{-4}$	5,69	0,16	0,20	0,59	0,90	21,80
2 cha.	0,26	$2,1 \cdot 10^{-4}$	5,68	0,23	0,29	0,91	0,96	19,15
3 cha.	0,26	$3,0 \cdot 10^{-4}$	6,64	0,23	0,29	1,26	0,98	18,06
4 cha.	0,18	$1,7 \cdot 10^{-4}$	7,43	0,15	0,20	0,97	0,80	18,82
5 cha.	0,04	$8,5 \cdot 10^{-5}$	23,6	0,02	0,06	1,86	0,97	19,96
6 cha.	$3 \cdot 10^{-3}$	$1,7 \cdot 10^{-6}$	48,5	10^{-4}	$5,2 \cdot 10^{-3}$	0,52	1,01	17,36
+ 6 cha.	$4 \cdot 10^{-4}$	$3,3 \cdot 10^{-7}$	152	-10^{-3}	$1,5 \cdot 10^{-3}$	0,72	1,01	17,21

Méthodes de rééchantillonnage

Introduction

Contexte

Cette section (y compris les notations) s'appuie largement sur Shao et Tu (1994) : "The Jackknife and the Bootstrap".

On souhaite étudier les propriétés d'une population (finie ou infinie), en utilisant les données relevées sur un échantillon i.i.d. X_1, \dots, X_n . Cet échantillon est généré selon une distribution inconnue F , d'espérance m et de variance σ^2 supposées finies.

On notera également :

$$\begin{aligned} T_n &\equiv T_n(X_1, \dots, X_n) \\ &= T_n(X_1, \dots, X_n, F) \end{aligned}$$

un estimateur (ou statistique) quelconque.

Exemples

1) La moyenne simple :

$$T_n = \bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i,$$

estimateur naturel de l'espérance m de la loi F .

2) Une fonction de moyennes :

$$T_n = f(\bar{X}_n),$$

où on impose généralement des conditions de régularité sur la fonction $f(\cdot)$ (continuité, différentiabilité, ...).

Cas particuliers : ratio, coefficient de corrélation ou de régression, ...

Exemples (2)

3) La distribution empirique :

$$F_n(x) = \frac{1}{n} \sum_{i=1}^n 1(X_i \leq x)$$

qui estime la loi inconnue F , avec $1(\cdot)$ variable indicatrice.

4) Un quantile empirique :

$$T_n = F_n^{-1}(t),$$

avec $t \in [0, 1]$ et

$$F_n^{-1}(t) = \text{Inf}\{x; F_n(x) \geq t\}.$$

Quelques rappels

Moyenne

Nous considérons tout d'abord le cas de données X_1, \dots, X_n unidimensionnelles, et de l'estimation de l'espérance m de la loi F .

L'estimateur de m est donné par

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i,$$

et sa variance est donnée par

$$V(\bar{X}_n) = \frac{\sigma^2}{n}.$$

Moyenne : estimation de variance

Cette variance peut être estimée sans biais par

$$v(\bar{X}_n) = \frac{s_X^2}{n},$$

avec

$$s_X^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$$

la dispersion (corrigée) dans l'échantillon.

Moyenne : intervalle de confiance

La loi limite de l'estimateur \bar{X}_n est donnée par le théorème central-limite :

$$\frac{\bar{X}_n - m}{s_X/\sqrt{n}} \rightarrow_{\mathcal{L}} \mathcal{N}(0, 1),$$

On obtient l'intervalle de confiance de niveau $1 - 2\alpha$ pour le paramètre m :

$$\left[\bar{X}_n \pm u_\alpha \frac{s_X}{\sqrt{n}} \right]$$

avec u_α le fractile d'ordre α d'une loi $\mathcal{N}(0, 1)$.

$$\text{Exemples : } u_{0.025} = 1.96 \quad u_{0.05} = 1.64$$

Fonction de moyennes

Nous considérons maintenant le cas de p -vecteurs X_1, \dots, X_n générés selon une loi commune F de dimension p , et d'espérance m .

On souhaite estimer le paramètre $\theta = f(m)$, où la fonction $f(\cdot)$ est supposée continûment différentiable.

Ce paramètre peut être estimé par substitution (ou plug-in) par

$$\hat{\theta} = f(\bar{X}_n) = T_n.$$

Principe :

$$\begin{aligned}\hat{\theta} - \theta &= f(\bar{X}_n) - f(m) \\ &\simeq [f'(m)]^T (\bar{X}_n - m)\end{aligned}$$

Fonction de moyennes : estimation de variance

On en déduit :

$$V(\hat{\theta}) \simeq V\left([f'(m)]^T \bar{X}_n\right).$$

Cette variance peut être estimée asymptotiquement sans biais par

$$v(\hat{\theta}) = \frac{s_U^2}{n},$$

avec

$$s_U^2 = \frac{1}{n-1} \sum_{i=1}^n (U_i - \bar{U}_n)^2$$

et $U_i = [f'(\bar{X}_n)]^T X_i$ la variable linéarisée.

⇒ estimateur de variance par linéarisation.

Fonction de moyennes : intervalle de confiance

La loi limite de l'estimateur $\hat{\theta}$ est donnée par :

$$\frac{\hat{\theta} - \theta}{s_U / \sqrt{n}} \rightarrow_{\mathcal{L}} \mathcal{N}(0, 1),$$

On obtient l'intervalle de confiance de niveau $1 - 2\alpha$ pour le paramètre θ :

$$\left[\hat{\theta} \pm u_\alpha \frac{s_U}{\sqrt{n}} \right]$$

avec u_α le fractile d'ordre α d'une loi $\mathcal{N}(0, 1)$.

Méthodes de rééchantillonnage

Utilisations des méthodes de rééchantillonnage

Très variées :

- ▶ Estimation du biais (statistique),
- ▶ Estimation de variance,
- ▶ Production d'intervalles de confiance,
- ▶ Approximation de la loi d'un estimateur.

Le Jackknife

Le Jackknife

Le Jackknife a été introduit à l'origine par Quenouille (1949a,b) pour estimer le biais d'une statistique, puis a été proposé pour l'estimation de variance par Tukey (1958).

Principe du Jackknife : on recalcule la statistique estimée en supprimant chaque unité tour à tour. La variabilité des statistiques Jackknifées est utilisée afin d'estimer la variance.

Cette technique est encore appelée le delete-1 Jackknife.

Cas d'une moyenne

Nous considérons tout d'abord le cas de données X_1, \dots, X_n unidimensionnelles, et de l'estimation de l'espérance m de la loi F .

L'estimateur de m est donné par

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$$

et l'estimateur basé sur l'échantillon privé d'un individu j est donné par

$$\bar{X}_{n,-j} = \frac{1}{n-1} \sum_{\substack{i=1 \\ i \neq j}}^n X_i.$$

Cas d'une moyenne

L'estimateur Jackknife du biais de \bar{X}_n est donné par

$$b_{JACK} [\bar{X}_n] = \frac{n-1}{n} \sum_{j=1}^n (\bar{X}_{n,-j} - \bar{X}_n).$$

On montre que cet estimateur est égal à 0
 \Rightarrow cohérent avec le caractère non biaisé de \bar{X}_n .

L'estimateur Jackknife de la variance de \bar{X}_n est donné par

$$\begin{aligned} v_{JACK} [\bar{X}_n] &= \frac{n-1}{n} \sum_{j=1}^n (\bar{X}_{n,-j} - \bar{X}_n)^2 \\ &= \frac{s_X^2}{n}. \end{aligned}$$

\Rightarrow restitue l'estimateur sans biais de la variance

Exemple

Echantillon de taille $n = 4$ sélectionné avec remise.

i	X_i				
1	1				
2	3				
3	2				
4	10				

Exemple

Echantillon de taille $n = 4$ sélectionné avec remise.

i	X_i				
1	1				
2	3				
3	2				
4	10				
	$\bar{X}_n = 4$				

$$v[\bar{X}_n] = \frac{s_X^2}{n} = 4.17$$

Exemple

Echantillon de taille $n = 4$ sélectionné avec remise.

i	X_i	S_{-1}			
1	1	-			
2	3	3			
3	2	2			
4	10	10			
	$\bar{X}_n = 4$				

$$v[\bar{X}_n] = \frac{s_X^2}{n} = 4.17$$

Exemple

Echantillon de taille $n = 4$ sélectionné avec remise.

i	X_i	S_{-1}			
1	1	-			
2	3	3			
3	2	2			
4	10	10			
	$\bar{X}_n = 4$	$\bar{X}_{n,-1} = 5$			

$$v[\bar{X}_n] = \frac{s_X^2}{n} = 4.17$$

Exemple

Echantillon de taille $n = 4$ sélectionné avec remise.

i	X_i	S_{-1}	S_{-2}	S_{-3}	S_{-4}
1	1	-	1	1	1
2	3	3	-	3	3
3	2	2	2	-	2
4	10	10	10	10	-
	$\bar{X}_n = 4$	$\bar{X}_{n,-1} = 5$			

$$v[\bar{X}_n] = \frac{s_X^2}{n} = 4.17$$

Exemple

Echantillon de taille $n = 4$ sélectionné avec remise.

i	X_i	S_{-1}	S_{-2}	S_{-3}	S_{-4}
1	1	-	1	1	1
2	3	3	-	3	3
3	2	2	2	-	2
4	10	10	10	10	-
	$\bar{X}_n = 4$	$\bar{X}_{n,-1} = 5$	$\bar{X}_{n,-2} = 4.33$	$\bar{X}_{n,-3} = 4.67$	$\bar{X}_{n,-4} = 2$

$$v[\bar{X}_n] = \frac{s_X^2}{n} = 4.17$$

Exemple

Echantillon de taille $n = 4$ sélectionné avec remise.

i	X_i	S_{-1}	S_{-2}	S_{-3}	S_{-4}
1	1	-	1	1	1
2	3	3	-	3	3
3	2	2	2	-	2
4	10	10	10	10	-
	$\bar{X}_n = 4$	$\bar{X}_{n,-1} = 5$	$\bar{X}_{n,-2} = 4.33$	$\bar{X}_{n,-3} = 4.67$	$\bar{X}_{n,-4} = 2$

$$v[\bar{X}_n] = \frac{s_X^2}{n} = 4.17$$

$$v_{JACK}[\bar{X}_n] = \frac{n-1}{n} \sum_{i=1}^n (\bar{X}_{n,-i} - \bar{X}_n)^2 = 4.17$$

Cas d'une fonction de moyennes

Nous considérons maintenant le cas de p -vecteurs X_1, \dots, X_n générés selon une loi commune F de dimension p , et d'espérance m .

On souhaite estimer le paramètre $\theta = f(m)$, où la fonction $f(\cdot)$ est supposée continûment différentiable.

Ce paramètre est estimé par substitution par

$$\hat{\theta} = f(\bar{X}_n) = T_n.$$

L'estimateur basé sur l'échantillon privé de l'individu j est donné par

$$\hat{\theta}_{-j} = f(\bar{X}_{n,-j}).$$

Cas d'une fonction de moyennes (2)

L'estimateur Jackknife du biais de $\hat{\theta}$ est donné par

$$b_{JACK} [\hat{\theta}] = \frac{n-1}{n} \sum_{i=1}^n \left(\hat{\theta}_{-i} - \frac{1}{n} \sum_{j=1}^n \hat{\theta}_{-j} \right).$$

L'estimateur Jackknife de la variance de $\hat{\theta}$ est donné par

$$v_{JACK} [\hat{\theta}] = \frac{n-1}{n} \sum_{i=1}^n \left(\hat{\theta}_{-i} - \frac{1}{n} \sum_{j=1}^n \hat{\theta}_{-j} \right)^2.$$

Pour une fonction $f(\cdot)$ continûment différentiable au point m , le Jackknife donne une estimation de variance consistante pour $\hat{\theta}$.

Cas d'une fonction de moyennes (3)

L'estimateur Jackknife de la variance de $\hat{\theta}$ peut se réécrire

$$\begin{aligned} v_{JACK} [\hat{\theta}] &= \frac{n-1}{n} \sum_{i=1}^n \left(\hat{\theta}_{-i} - \frac{1}{n} \sum_{j=1}^n \hat{\theta}_{-j} \right)^2 \\ &= \frac{n-1}{n} \sum_{i=1}^n \left([\hat{\theta}_{-i} - \hat{\theta}] - \frac{1}{n} \sum_{j=1}^n [\hat{\theta}_{-j} - \hat{\theta}] \right)^2. \end{aligned} \quad (7)$$

Principe :

- ▶ $\hat{\theta}_{-j} - \hat{\theta} \simeq [f'(\bar{X}_n)]^T (\bar{X}_{n,-j} - \bar{X}_n)$,
- ▶ en injectant cette approximation dans (7), on retrouve l'estimateur de variance par linéarisation.

Linéarisation Jackknife

L'estimateur Jackknife de la variance peut se réécrire

$$v_{JACK} [\hat{\theta}] = \frac{1}{n(n-1)} \sum_{i=1}^n \left(\tilde{\theta}_i - \frac{1}{n} \sum_{j=1}^n \tilde{\theta}_j \right)^2 = \frac{s_{\tilde{\theta}}^2}{n},$$

où

$$\tilde{\theta}_i = (n-1)(\hat{\theta} - \hat{\theta}_{-i})$$

donne une approximation numérique de la variable linéarisée (Davison et Hinkley, 1997, p. 50).

Le Jackknife peut donc être vu comme une méthode numérique permettant d'éviter le calcul des variables linéarisées.

Shao et Tu (1995) montrent que l'estimation de variance est également consistante sous des conditions plus faibles de différentiabilité.

Exemple

On génère une population artificielle de taille $N = 1\,000$ contenant deux variables X et Y . La variable X est générée selon une loi gamma de paramètres 2 et 5. La variable Y est générée de façon à ce que $\rho(X, Y) \simeq 0.7$.

On prélève un échantillon de taille $n = 100$ dans U , par sondage aléatoire simple avec remise. L'objectif est d'estimer le ratio

$$R = \frac{m_Y}{m_X}.$$

On compare pour le paramètre R la linéarisée de Taylor et la linéarisée Jackknife.

Exemple

i				
12				
14				
16				
71				
82				
122				
126				
128				

Exemple

i	X_i	Y_i		
12	3.28	7.37		
14	14.75	18.02		
16	16.66	13.52		
71	15.95	11.86		
82	7.31	21.80		
122	11.53	21.05		
126	12.25	4.28		
128	24.04	31.34		

Exemple

i	X_i	Y_i	Linéarisée U_i	
12	3.28	7.37	0.395	
14	14.75	18.02	0.373	
16	16.66	13.52	-0.212	
71	15.95	11.86	-0.303	
82	7.31	21.80	1.380	
122	11.53	21.05	0.939	
126	12.25	4.28	-0.682	
128	24.04	31.34	0.793	

Exemple

i	X_i	Y_i	Linéarisée U_i	Lin. Jackknife $\tilde{\theta}_i$
12	3.28	7.37	0.395	0.392
14	14.75	18.02	0.373	0.375
16	16.66	13.52	-0.212	-0.213
71	15.95	11.86	-0.303	-0.304
82	7.31	21.80	1.380	1.376
122	11.53	21.05	0.939	0.940
126	12.25	4.28	-0.682	-0.683
128	24.04	31.34	0.793	0.803

Intervalle de confiance

La consistance de l'estimateur de variance Jackknife implique que :

$$\frac{\hat{\theta} - \theta}{\sqrt{v_{JACK}[\hat{\theta}]}} \rightarrow_{\mathcal{L}} \mathcal{N}(0, 1),$$

On peut donc utiliser cet estimateur de variance pour produire un intervalle de confiance de niveau $1 - 2\alpha$ pour le paramètre θ :

$$\left[\hat{\theta} \pm u_{\alpha} \sqrt{v_{JACK}[\hat{\theta}]} \right].$$

Avantage : variable linéarisée remplacée par une approximation numérique.

Delete-d jackknife

Le delete-1 Jackknife implique un total de n suppressions, ce qui peut être prohibitif si la taille d'échantillon est grande. Les suppressions peuvent être également réalisées par blocs de d unités à la fois (Shao et Wu, 1985). On parle alors de delete-d Jackknife.

Cette méthode a également été étudiée par Shao et Wu (1989) afin de produire une estimation consistante de variance pour des paramètres non lisses tels que les fractiles. En particulier, les nombres d et $n - d$ d'individus supprimés et d'individus conservés doivent tendre vers l'infini avec n .

Le Bootstrap

Le Bootstrap

Le Bootstrap a été introduit par Efron (1979) dans le cadre d'une population infinie. L'idée de base consiste à reproduire le mécanisme d'échantillonnage d'origine.

Le Bootstrap permet d'obtenir une approximation de la distribution d'un estimateur. Il permet d'obtenir une estimation de variance consistante, y compris pour des paramètres non lisses tels que les fractiles.

L'adaptation au cas d'une population finie est assez problématique, et fait l'objet d'une littérature abondante, voir en particulier Shao et Tu (1995), Davison et Hinkley (1997), Davison et Sardy (2007).

Principe

L'idée de base du Bootstrap est un principe de substitution (ou plug-in). Soit X_1, \dots, X_n un échantillon iid de distribution commune $F(\cdot)$. Soit F_n la fonction de répartition empirique définie par

$$F_n(x) = \frac{1}{n} \sum_{i=1}^n 1(X_i \leq x).$$

Un paramètre $\theta(F)$ est estimé par plug-in par $\theta(F_n)$

Exemples

1) L'espérance de la loi F

$$m = \int x dF(x) = \theta_1(F)$$

est estimée par

$$\theta_1(F_n) = \int x dF_n(x) = \bar{X}_n.$$

2) Une fonction de moyennes :

$$f(m) = f\left(\int x dF(x)\right) = \theta_2(F)$$

est estimée par

$$\theta_2(F_n) = f\left(\int x dF_n(x)\right) = f(\bar{X}_n).$$

Exemples (2)

3) Un quantile

$$F^{-1}(t) = \theta_3(F)$$

est estimé par

$$\theta_3(F_n) = F_n^{-1}(t).$$

Notons que le principe de substitution s'applique même à des paramètres fortement non-linéaires.

Approximation de Monte-Carlo

L'estimateur $\theta(F_n)$ peut être difficile à déterminer en pratique :

$$\begin{aligned}\theta_1 = F^{-1}(t) &\Rightarrow \hat{\theta}_1 = F_n^{-1}(t) \\ \theta_2 = V [F_n^{-1}(t)] &\Rightarrow \hat{\theta}_2 = ?\end{aligned}$$

On peut utiliser à la place une approximation de Monte-Carlo. On note :

- ▶ S^* un échantillon de taille n tiré selon la loi F_n (i.e., par sondage aléatoire simple avec remise dans S),
- ▶ $F_n^*(\cdot)$ la fonction de répartition empirique calculée sur le rééchantillon S^* .

Approximation de Monte-Carlo

On répète cette procédure B fois indépendamment, pour obtenir les rééchantillons S_1^*, \dots, S_B^* , et $\theta(F_n)$ est approché par

$$\theta^B(F_n) = \frac{1}{B} \sum_{b=1}^B \theta(F_n^{*b}),$$

avec F_n^{*b} la fonction de répartition empirique calculée sur le rééchantillon S_b^* .

Moyenne : estimation du biais

Le biais de \bar{X}_n est donné par

$$\begin{aligned} B[\bar{X}_n] &= E[\bar{X}_n - m] \\ &= E\left[\int x dF_n(x) - \int x dF(x)\right] \\ &= 0. \end{aligned}$$

L'estimateur Bootstrap du biais de \bar{X}_n est donné par

$$\begin{aligned} b_{BOOT}[\bar{X}_n] &= E^*[\bar{X}_n^* - \bar{X}_n] \\ &= E^*\left[\int x dF_n^*(x) - \int x dF_n(x)\right]. \end{aligned}$$

avec $E^*(\cdot)$ l'espérance sous le mécanisme de rééchantillonnage. Cet estimateur est égal à 0.

\Rightarrow cohérent avec le caractère non biaisé de \bar{X}_n .

Moyenne : estimation de la variance

La variance de \bar{X}_n est donnée par

$$\begin{aligned}V[\bar{X}_n] &= E[\bar{X}_n - m]^2 \\ &= E\left[\left(\int x d(F_n - F)(x)\right)^2\right] = \sigma^2/n.\end{aligned}$$

L'estimateur Bootstrap de la variance de \bar{X}_n est donné par

$$\begin{aligned}v_{BOOT}[\bar{X}_n] &= E^*[\bar{X}_n^* - \bar{X}_n]^2 \\ &= E^*\left[\left(\int x d(F_n^* - F_n)(x)\right)^2\right] = \left(\frac{n-1}{n}\right) \frac{s_y^2}{n}.\end{aligned}$$

Rééchantillonner avec une taille $m = n - 1$ permet de supprimer le biais (Efron, 1982).

Exemple

Population U de taille 10, dans laquelle un échantillon de taille $n = 4$ est sélectionné selon un SAS avec remise.

i	X_i						
1	1						
2	3						
3	2						
4	10						

Exemple

Population U de taille 10, dans laquelle un échantillon de taille $n = 4$ est sélectionné selon un SAS avec remise.

i	X_i						
1	1						
2	3						
3	2						
4	10						
	$\bar{X} = 4$						

$$v[\bar{X}] = \frac{s_X^2}{n} = 4.17$$

Exemple

Population U de taille 10, dans laquelle un échantillon de taille $n = 4$ est sélectionné selon un SAS avec remise.

i	X_i	W_1					
1	1	0					
2	3	3					
3	2	1					
4	10	0					
	$\bar{X} = 4$						

$$v[\bar{X}] = \frac{s_X^2}{n} = 4.17$$

Exemple

Population U de taille 10, dans laquelle un échantillon de taille $n = 4$ est sélectionné selon un SAS avec remise.

i	X_i	W_1					
1	1	0					
2	3	3					
3	2	1					
4	10	0					
	$\bar{X} = 4$	$\bar{X}_1^* = 2.75$					

$$v[\bar{X}] = \frac{s_X^2}{n} = 4.17$$

Exemple

Population U de taille 10, dans laquelle un échantillon de taille $n = 4$ est sélectionné selon un SAS avec remise.

i	X_i	W_1	W_2	W_3	W_4	W_5	W_6
1	1	0	0	1	0	1	1
2	3	3	0	1	0	1	3
3	2	1	1	0	2	1	0
4	10	0	3	2	2	1	0
	$\bar{X} = 4$	$\bar{X}_1^* = 2.75$					

$$v[\bar{X}] = \frac{s_X^2}{n} = 4.17$$

Exemple

Population U de taille 10, dans laquelle un échantillon de taille $n = 4$ est sélectionné selon un SAS avec remise.

i	X_i	W_1	W_2	W_3	W_4	W_5	W_6
1	1	0	0	1	0	1	1
2	3	3	0	1	0	1	3
3	2	1	1	0	2	1	0
4	10	0	3	2	2	1	0
	$\bar{X} = 4$	$\bar{X}_1^* = 2.75$	$\bar{X}_2^* = 8$	$\bar{X}_3^* = 6$	$\bar{X}_4^* = 6$	$\bar{X}_5^* = 4$	$\bar{X}_6^* = 2.5$

$$v[\bar{X}] = \frac{s_{\bar{X}}^2}{n} = 4.17$$

Exemple

Population U de taille 10, dans laquelle un échantillon de taille $n = 4$ est sélectionné selon un SAS avec remise.

i	X_i	W_1	W_2	W_3	W_4	W_5	W_6
1	1	0	0	1	0	1	1
2	3	3	0	1	0	1	3
3	2	1	1	0	2	1	0
4	10	0	3	2	2	1	0
	$\bar{X} = 4$	$\bar{X}_1^* = 2.75$	$\bar{X}_2^* = 8$	$\bar{X}_3^* = 6$	$\bar{X}_4^* = 6$	$\bar{X}_5^* = 4$	$\bar{X}_6^* = 2.5$

$$v[\bar{X}] = \frac{s_X^2}{n} = 4.17$$

$$v_{BOOT}^B[\bar{X}] = 3.87$$

Fonction de moyennes : estimation du biais

Nous considérons maintenant le cas de p -vecteurs X_1, \dots, X_n générés selon une loi commune F de dimension p , et d'espérance m . On souhaite estimer le paramètre $\theta = f(m)$, où la fonction $f(\cdot)$ est supposée continûment différentiable.

L'estimateur basé sur le rééchantillon S^* est donné par $\hat{\theta}^* = f(\bar{X}_n^*)$.

Estimateur Bootstrap du biais de $\hat{\theta}$:

$$b_{BOOT} [\hat{\theta}] = E^* [f(\bar{X}_n^*) - f(\bar{X}_n)].$$

Approximation de Monte-Carlo :

$$b_{BOOT}^B [\hat{\theta}] = \frac{1}{B} \sum_{b=1}^B [f(\bar{X}_n^{*b}) - f(\bar{X}_n)].$$

Fonction de moyennes : estimation de la variance

Estimateur Bootstrap de la variance de $\hat{\theta}$:

$$v_{BOOT} [\hat{\theta}] = E^* [f(\bar{X}_n^*) - E^*[f(\bar{X}_n^*)]]^2$$

Approximation de Monte-Carlo :

$$v_{BOOT}^B [\hat{\theta}] = \frac{1}{B-1} \sum_{b=1}^B \left[f(\bar{X}_n^{*b}) - \frac{1}{B} \sum_{c=1}^B f(\bar{X}_n^{*c}) \right]^2 .$$

Si la fonction $f(\cdot)$ est continument différentiable au point μ_y , le Bootstrap donne une estimation de variance consistante pour $\hat{\theta}$.

Principe :

- ▶ $\hat{\theta}^* - \hat{\theta} = f(\bar{X}_n^*) - f(\bar{X}_n) \simeq [f'(\bar{X}_n)]^T [\bar{X}_n^* - \bar{X}_n]$,
- ▶ consistance de l'estimateur de variance par linéarisation.

Généralisation

Shao et Tu (1995) montrent que l'estimation de variance est également consistante sous des conditions de différentiabilité plus faible pour le paramètre θ .

Contrairement au delete-1 Jackknife, on peut également obtenir un résultat de consistance pour l'estimation d'un quantile.

Intervalle de confiance

Un des intérêts du Bootstrap est que l'on estime non seulement la variance de $\hat{\theta}$, mais aussi sa distribution. Les statistiques Bootstrappées peuvent être utilisées pour produire un intervalle de confiance.

La méthode des percentiles est une solution simple pour obtenir un intervalle de confiance :

- ▶ On tire B rééchantillons $S_1^*, \dots, S_B^* \Rightarrow \hat{\theta}_1^*, \dots, \hat{\theta}_B^*$.
- ▶ On les trie par ordre croissant $\Rightarrow \hat{\theta}_{(1)}^*, \dots, \hat{\theta}_{(B)}^*$.
- ▶ On supprime les $\alpha\%$ les plus faibles et les $\alpha\%$ les plus grandes pour obtenir l'intervalle de confiance de niveau $(1 - 2\alpha)$:

$$\left[\hat{\theta}_{(L)}^*, \hat{\theta}_{(U)}^* \right] .$$

Intervalle de confiance

La méthode des percentiles est simple à mettre en oeuvre, et peu gourmande en temps de calcul. Elle est utilisable aussi bien pour des paramètres lisses (fonctions de moyennes) que non lisses (fractiles).

Une méthode plus fine appelée le t-Bootstrap consiste à estimer les fractiles de la statistique pivotale

$$\frac{\hat{\theta} - \theta}{\sqrt{v[\hat{\theta}]}} \equiv \frac{\theta(F_n) - \theta(F)}{\sqrt{v[\theta(F_n)]}}$$

en utilisant l'équivalent Bootstrap de cette statistique :

$$\frac{\hat{\theta}^* - \hat{\theta}}{\sqrt{v[\hat{\theta}^*]}} \equiv \frac{\theta(F_n^*) - \theta(F_n)}{\sqrt{v[\theta(F_n^*)]}}$$

Intervalle de confiance

On sélectionne B rééchantillons Bootstrap, ce qui permet d'obtenir les statistiques pivotales bootstrappées

$$\frac{\hat{\theta}_1^* - \hat{\theta}}{\sqrt{v[\hat{\theta}_1^*]}} \quad \dots \quad \frac{\hat{\theta}_B^* - \hat{\theta}}{\sqrt{v[\hat{\theta}_B^*]}}$$

En les ordonnant, on en déduit une estimation des fractiles u_α et $u_{1-\alpha}$ de $\frac{\hat{\theta} - \theta}{\sqrt{v[\hat{\theta}]}}$.

On en déduit l'intervalle de confiance de niveau $(1 - 2\alpha)$:

$$\left[\hat{\theta} - \hat{u}_{1-\alpha} \sqrt{v[\hat{\theta}]}, \hat{\theta} - \hat{u}_\alpha \sqrt{v[\hat{\theta}]} \right].$$

Intervalle de confiance

L'inconvénient de cette méthode est que pour chaque statistique Bootstrap

$$\frac{\hat{\theta}^* - \hat{\theta}}{\sqrt{v[\hat{\theta}^*]}},$$

on doit disposer d'un estimateur de variance pour le rééchantillon :

- ▶ double Bootstrap pour estimer la variance,
- ▶ estimateur Jackknife de variance,
- ▶ estimateur de variance linéarisé.

Cette méthode donne de bons résultats quand le paramètre est une fonction lisse de moyennes. En revanche, elle est inconsistante dans le cas d'un fractile.

Rééchantillonnage pour données d'enquête

L'adaptation des méthodes de rééchantillonnage à des données d'enquête fait l'objet d'une vaste littérature, mais se heurte à de nombreuses difficultés techniques.

Dans cette section, nous donnons tout d'abord quelques rappels sur des plans de sondage classiques en population finie.

Nous revenons ensuite sur la technique de substitution permettant d'estimer un paramètre complexe, avant d'évoquer des adaptations des méthodes de rééchantillonnage au cas d'une population finie.

Plans de sondage classiques

Notations

On considère une population finie U de taille N , dans laquelle un échantillon S de taille n est sélectionné à l'aide d'un plan de sondage $p(\cdot)$.

On note y_k la valeur prise par une variable y (éventuellement vectorielle) sur l'individu k de U , et $t_y = \sum_{k \in U} y_k$ le total de cette variable.

Les probabilités d'inclusion π_k des unités dans l'échantillon sont supposées connues et contrôlées. Le total t_y peut être estimé sans biais sous le plan de sondage par le π -estimateur

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

Plans de sondage classiques

Dans cette partie, nous nous intéressons au cas de l'estimation du total t_y , et à l'estimation de la variance associée au seul mécanisme d'échantillonnage. Estimation de variance en présence de données imputées : voir Haziza (2009).

Nous revenons sur quelques plans de sondage classiques :

- ▶ sondage aléatoire simple,
- ▶ sondage à probabilités inégales,
- ▶ sondage stratifié,
- ▶ sondage à plusieurs degrés.

Sondage aléatoire simple

Le sondage aléatoire simple **avec remise**

On sélectionne dans la population n unités, avec remise et à probabilités égales. Le total t_y peut être estimé sans biais par l'estimateur d'Hansen-Hurvitz (1943) :

$$\begin{aligned}\hat{t}_{yHH} &= \sum_{k \in S} \frac{y_k}{n p_k} \text{ avec } p_k = \frac{1}{N} \\ &= N \bar{y}\end{aligned}$$

avec

$$\bar{y} = \frac{1}{n} \sum_{k \in S} y_k$$

la moyenne simple de la variable y calculée sur l'échantillon S .

Le sondage aléatoire simple avec remise

La variance de l'estimateur de Hansen-Hurvitz est donnée par

$$V [\hat{t}_{yHH}] = N^2 \frac{\sigma_y^2}{n} \quad \text{avec} \quad \sigma_y^2 = \frac{1}{N} \sum_{k \in U} (y_k - \mu_y)^2$$

et peut être estimée sans biais par

$$v [\hat{t}_{yHH}] = N^2 \frac{s_y^2}{n} \quad \text{avec} \quad s_y^2 = \frac{1}{n-1} \sum_{k \in S} (y_k - \bar{y})^2.$$

Les méthodes de Jackknife et de Bootstrap classiques s'appliquent facilement au cas d'un sondage aléatoire simple avec remise.

Le sondage aléatoire simple **sans remise**

Sondage Aléatoire Simple (SAS) de taille n : plan de taille fixe, où tous les échantillons contenant n unités **distinctes** ont la même probabilité d'être sélectionnés.

$$\pi_k = \frac{n}{N}.$$

Le π -estimateur peut se réécrire

$$\hat{t}_{y\pi} = \frac{N}{n} \sum_{k \in S} y_k = N\bar{y},$$

avec $\bar{y} = \frac{1}{n} \sum_{k \in S} y_k$ la moyenne simple de la variable y calculée sur l'échantillon S .

Le sondage aléatoire simple **sans remise**

La variance du π -estimateur est donnée par

$$V [\hat{t}_{y\pi}] = N^2(1 - f) \frac{S_y^2}{n}$$

avec $f = n/N$ le taux de sondage, et $S_y^2 = \frac{1}{N-1} \sum_{k \in U} (y_k - \mu_y)^2$ la dispersion corrigée de y dans la population.

Cette variance est estimée sans biais par

$$v [\hat{t}_{y\pi}] = N^2(1 - f) \frac{s_y^2}{n},$$

avec $s_y^2 = \frac{1}{n-1} \sum_{k \in S} (y_k - \bar{y})^2$ la dispersion corrigée de y dans l'échantillon.

Le sondage aléatoire simple **sans remise**

Le facteur $1 - f$ appelé **correction de population finie** représente le gain de variance obtenu comparé à un tirage avec remise de même taille.

Une des difficultés des méthodes de rééchantillonnage est de prendre en compte cette correction.

La variance du sondage aléatoire simple peut également être utilisée comme majorant pour un tirage systématique à probabilités égales, couramment utilisé en pratique.

Exemple : tirage de ménages dans les communes de l'échantillon-maître (Ardilly, 2006).

Le tirage à probabilités inégales

Le tirage à probabilités inégales

En l'absence d'information, on n'a pas de raison de privilégier la sélection d'un individu et on utilise des probabilités égales de tirage.

En présence d'une **information auxiliaire**, on peut individualiser les probabilités de sélection et utiliser des probabilités inégales de tirage.

Exemple : tirage à probabilités proportionnelles à la taille

$$\pi_k = n \frac{x_k}{\sum_{l \in U} x_l}$$

pour la sélection d'unités primaires de l'Echantillon-Maître 99 (Bourdalle et al, 2000).

Le tirage à probabilités inégales

Parmi les nombreux algorithmes de tirage à probabilités inégales (cf Tillé, 2006) :

- ▶ le tirage systématique (Madow and Madow, 1944),
- ▶ le tirage réjectif (Hajek, 1964),
- ▶ la méthode du pivot (Deville et Tillé, 1998).

Ces méthodes permettent de respecter exactement les probabilités d'inclusion souhaitées, et conduisent à une taille fixe d'échantillon.

La méthode du pivot est un cas particulier de la méthode du Cube (Deville et Tillé, 2004) permettant de sélectionner des échantillons équilibrés (non développé ici).

Variance du π -estimateur

Pour un tirage à probabilités inégales de taille fixe, la variance du π -estimateur est également donnée par la formule de Yates-Grundy (1953)

$$V[\hat{t}_{y\pi}] = -\frac{1}{2} \sum_{k \neq l \in U} \left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l} \right)^2 (\pi_{kl} - \pi_k \pi_l).$$

Cette variance peut être estimée sans biais par

$$v_{YG}[\hat{t}_{y\pi}] = -\frac{1}{2} \sum_{k \neq l \in S} \left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l} \right)^2 \frac{\pi_{kl} - \pi_k \pi_l}{\pi_{kl}}$$

si toutes les probabilités d'inclusion π_{kl} d'ordre 2 sont non nulles.

Estimation de variance

Les estimateurs précédents sont biaisés si certaines probabilités π_{kl} sont nulles (ex : tirage systématique).

De plus, ces estimateurs de variance sont difficiles à utiliser en pratique car :

- ▶ les probabilités π_{kl} peuvent être difficiles à calculer,
- ▶ problème du stockage des π_{kl} ,
- ▶ ces estimateurs font appel à une somme double et peuvent être numériquement instables (Matei et Tillé, 2005).

Le tirage à probabilités inégales

Un estimateur de variance simplifié a été proposé par Hajek (1964) pour le tirage réjectif :

$$v_{HAJ} [\hat{t}_{y\pi}] = \sum_{k \in S} (1 - \pi_k) \left(\frac{y_k}{\pi_k} - \hat{R} \right)^2$$

où

$$\hat{R} = \frac{\sum_{k \in S} \frac{y_k}{\pi_k} (1 - \pi_k)}{\sum_{k \in S} (1 - \pi_k)}.$$

D'autres estimateurs de variance sont possibles (Matei et Tillé, 2005).

Le tirage à probabilités inégales

Ce type d'estimateur de variance peut également être utilisé pour un tirage à probabilités inégales à forte entropie (Berger 1996, 1998, 2005).

On peut également l'utiliser comme estimateur de variance conservatif pour un plan à plus faible entropie (ex : tirage systématique).

Un estimateur de ce type est utilisé dans le logiciel POULPE permettant d'estimer la variance pour une enquête complexe (Caron, 1998, Petit, 1998).

Le sondage stratifié

Le sondage stratifié

L'échantillonnage est rarement réalisé directement dans la population. Les méthodes vues précédemment servent de base à la construction de plans de sondage plus précis.

Une possibilité consiste à découper la population U en sous-populations U_1, \dots, U_H appelées strates, dans lesquelles des échantillons sont sélectionnés indépendamment, par exemple selon un SAS.

Exemple (Caron, 2002): cas des enquêtes entreprises
≡ plan stratifié en croisant l'activité principale de l'entreprise (APE)
et la taille en tranches d'effectifs salariés
+ sondage aléatoire simple dans chaque strate.

Le sondage aléatoire simple stratifié

Principe : la population U est partitionnée en H strates U_1, \dots, U_H de tailles respectives N_1, \dots, N_H . On sélectionne indépendamment des échantillons S_h dans chaque strate, selon des SAS(n_h).

$$\pi_k = \frac{n_h}{N_h} \text{ pour } k \in U_h$$

Le π -estimateur peut se réécrire

$$\hat{t}_{y\pi} = \sum_{h=1}^H \frac{N_h}{n_h} \sum_{k \in S_h} y_k = \sum_{h=1}^H N_h \bar{y}_h.$$

Le sondage aléatoire simple stratifié

La variance du π -estimateur découle de l'indépendance des tirages :

$$V [\hat{t}_{y\pi}] = \sum_{h=1}^H N_h^2 (1 - f_h) \frac{S_{yh}^2}{n_h}$$

que l'on estime sans biais par

$$v [\hat{t}_{y\pi}] = \sum_{h=1}^H N_h^2 (1 - f_h) \frac{s_{yh}^2}{n_h}.$$

Allocations classiques : allocation proportionnelle, allocation de Neyman.

Tirage à probabilités inégales dans les strates

Par indépendance des tirages réalisés dans les strates, on obtient là encore un estimateur de variance pour $\hat{t}_{y\pi}$ en sommant les estimateurs de variance associés à chacune des strates.

Il est également possible de coupler un plan de sondage stratifié avec une stratégie d'échantillonnage à probabilités inégales dans chaque strate.

Exemple: Dans la sélection de l'Echantillon-Maître 99 :

stratification des Unités Primaires sur le rural/urbain
+ tirage des UP à probabilités proportionnelles à la taille dans les strates les moins urbaines.

Tirage à probabilités inégales dans les strates

Dans le cas d'un plan de sondage stratifié, avec tirage à probabilités inégales et à forte entropie dans chaque strate, on peut par exemple utiliser l'estimateur de Hajek :

$$v_{HAJ} [\hat{t}_{y\pi}] = \sum_{h=1}^H \sum_{k \in S_h} (1 - \pi_k) \left(\frac{y_k}{\pi_k} - \hat{R}_h \right)^2$$

où

$$\hat{R}_h = \frac{\sum_{k \in S_h} \frac{y_k}{\pi_k} (1 - \pi_k)}{\sum_{k \in S_h} (1 - \pi_k)}.$$

Estimation de variance

Augmenter le nombre de strates diminue la dispersion intra-strates, et améliore donc la précision du π -estimateur.

Cas particulier important : tirage fortement stratifié, avec tirage de taille $n_h = 2$ dans chaque strate (MacCarthy, 1969).

Le nombre H de strates est alors important. Un biais même minime d'un estimateur de variance dans chaque strate peut conduire à un biais important pour la variance du π -estimateur.

Le sondage multi-degrés

Le tirage multi-degrés

Les individus sont regroupés au sein de grosses unités (on peut avoir plusieurs niveaux de regroupement). On échantillonne ces grosses unités, puis (éventuellement) les individus à l'intérieur.

Exemple : plan à 3 degrés pour une enquête auprès des ménages

1. un échantillon de communes (UP),
2. puis un échantillon de quartiers dans les communes sélectionnées (US),
3. puis finalement un échantillon de ménages dans les quartiers sélectionnés (UT).

Moins coûteux qu'un échantillonnage direct si la population est dispersée géographiquement, mais moins précis.

Le tirage multi-degrés

On se limite ici au cas de l'échantillonnage à deux degrés, mais la discussion peut être étendue à un nombre de degrés supérieur.

La population U est partitionnée en M Unités Primaires (UP)

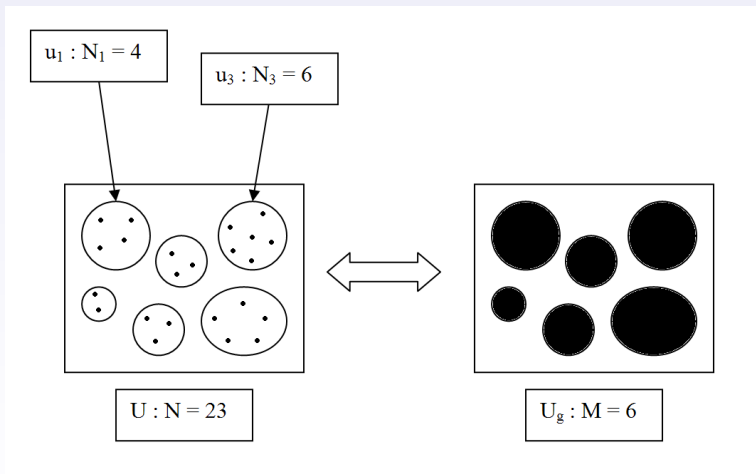
$$u_1, \dots, u_i, \dots, u_M.$$

L'ensemble des unités primaires est noté

$$U_I = \{u_1, \dots, u_i, \dots, u_M\}.$$

Soit M le nombre d'UP, N_i le nombre d'individus dans l'UP u_i , et $N = \sum_{u_i \in U_I} N_i$ la taille globale de la population U .

Notation pour le tirage à deux degrés



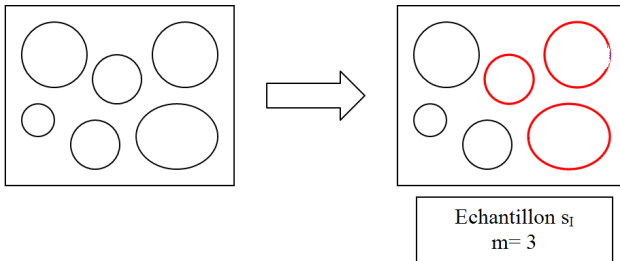
Le tirage multi-degrés

1. Un échantillon S_I de m UP est tiré dans U_I selon un plan de sondage $p_I(\cdot)$
 \Rightarrow probabilité d'inclusion π_{Ii} pour l'UP u_i
2. Pour chaque UP u_i sélectionnée, un échantillon S_i de n_i Unités Secondaires (US) est tiré dans u_i selon un plan $p_i(\cdot)$
 \Rightarrow probabilité d'inclusion $\pi_{k|i}$ pour l'US $k \in u_i$

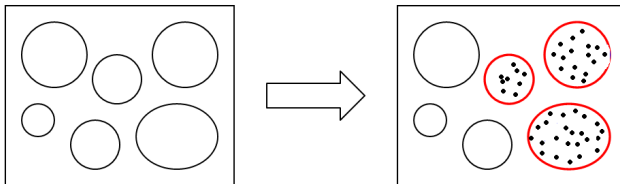
Les plans $p_i(\cdot)$ sont indépendants : le sous-échantillonnage est réalisé indépendamment d'une UP à l'autre.

La réunion des sous-échantillons S_i donne l'échantillon final S .

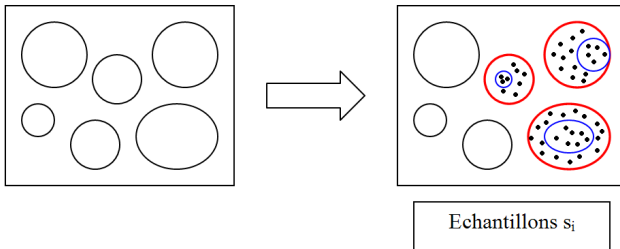
Principe du tirage à deux degrés



Principe du tirage à deux degrés



Principe du tirage à deux degrés



Le tirage multi-degrés

Le total t_y est estimé sans biais par

$$\hat{t}_{y\pi} = \sum_{u_i \in S_I} \frac{1}{\pi_{Ii}} \sum_{k \in S_i} \frac{y_k}{\pi_{k|i}} = \sum_{u_i \in S_I} \frac{\hat{t}_{y_i\pi}}{\pi_{Ii}}.$$

Sa variance se décompose en deux termes :

$$V[\hat{t}_{y\pi}] = V_{UP} + V_{US},$$

où :

- ▶ la première composante V_{UP} représente la variance due au 1er degré de tirage.
- ▶ la seconde composante V_{US} représente la variance due au 2nd degré de tirage.

Le tirage multi-degrés

Le premier terme V_{UP} est généralement prépondérant, particulièrement si le taux de sondage m/M est faible.

Les deux termes de variance diminuent si le nombre d'UP sélectionné augmente.

En pratique, la variabilité due au 1er degré de tirage peut être réduite en stratifiant les UP et/ou en les sélectionnant avec des probabilités d'inclusion proportionnelles à la taille.

Tirage sans remise des UP

La variance peut être estimée sans biais par

$$v [\hat{t}_{y\pi}] = \hat{V}_1 + \hat{V}_2.$$

Ce n'est pas un estimateur de variance sans biais terme à terme :

- ▶ \hat{V}_1 n'est pas un estimateur sans biais de V_{UP} ,
- ▶ \hat{V}_2 n'est pas un estimateur sans biais de V_{US} .

Cet estimateur de variance peut être justifié par une approche renversée (Shao et Steel, 1999, Haziza et Beaumont, 2005). L'approche renversée permet également de justifier des estimateurs simplifiés de variance (voir plus loin).

Cas d'un SAS à chaque degré

$$\begin{aligned} V \left[\hat{Y}_\pi \right] &= M^2 \left(1 - \frac{m}{M} \right) \frac{S_I^2}{m} + \frac{M}{m} \sum_{i \in U_I} N_i^2 \left(1 - \frac{n_i}{N_i} \right) \frac{s_i^2}{n_i} \\ &= V_{UP} + V_{US} \end{aligned}$$

$$\begin{aligned} v \left[\hat{Y}_\pi \right] &= M^2 \left(1 - \frac{m}{M} \right) \frac{s_I^2}{m} + \frac{M}{m} \sum_{i \in s_I} N_i^2 \left(1 - \frac{n_i}{N_i} \right) \frac{s_i^2}{n_i} \\ &= \hat{V}_1 + \hat{V}_2 \end{aligned}$$

Tirage sans remise des UP

Une des difficultés de l'estimation de variance par réplication pour un tirage sans remise des UP est de reproduire chacun de ces termes "artificiels".

De façon générale, un tirage à D degrés implique un estimateur de variance sans biais composé de D termes, mais ces termes n'isolent pas la variance due à chaque degré de tirage.

Tirage avec remise des UP

On se limite au cas d'une seule strate avec tirage à probabilités inégales avec remise des UP. La variance peut être estimée sans biais par (Cochran, 1977, p. 306) :

$$v [\hat{t}_{y\pi}] = \frac{m}{m-1} \sum_{u_i \in S_I} \left[\frac{\hat{t}_{y_i\pi}}{\pi_{Ii}} - \frac{\hat{t}_{y\pi}}{m} \right]^2.$$

Cet estimateur utilise un seul terme, qui représente la variance due aux deux degrés de tirage.

Se généralise à un nombre quelconque de degrés de tirage, avec des plans quelconques à partir du degré 2 : il suffit de disposer d'estimateurs sans biais dans les UP.

Tirage avec remise des UP

Cet estimateur de variance est plus simple à "attraper" avec les méthodes de réplication. L'ensemble de la variabilité du plan de sondage est capturé par la variabilité des $\frac{\hat{t}_{y_i} \pi}{\pi_{I_i}}$.

Cet estimateur est parfois utilisé comme estimateur conservatif de la variance dans le cas d'un tirage sans remise des unités primaires.

Si la fraction de sondage m/M du 1er degré n'est pas négligeable, la différence peut être importante.

Estimation de variance par rééchantillonnage

Principe

La linéarisation est une technique qui permet de tenir compte de la forme de la statistique dans le calcul de variance, mais qui n'intègre pas la prise en compte du plan de sondage. Elle nécessite :

- ▶ une connaissance précise du plan de sondage,
- ▶ le calcul d'une variable linéarisée pour chaque paramètre d'intérêt.

Elle peut être également difficile à mettre en oeuvre si la stratégie d'estimation est complexe (et notamment en situation de non-réponse).

Principe

Les méthodes de réplication consistent à remplacer une formule explicite de variance par de la puissance de calcul. La variance est estimée à l'aide de tirages répétés dans l'échantillon d'origine S .

Parmi les méthodes de réplication, on trouve les demi-échantillons équilibrés, ainsi que des adaptations du Jackknife et du Bootstrap au cas d'une population finie.

Une présentation détaillée des méthodes de rééchantillonnage en sondage est donnée dans Shao et Tu (1995), Davison et Sardy (2006).

Les demi-échantillons équilibrés

Cette méthode a été développée à l'origine par Mac Carthy (1969) pour le cas d'un sondage stratifié avec tirage **avec remise** de deux unités dans chaque strate.

Un demi-échantillon est obtenu en prélevant dans chaque strate un individu de l'échantillon de départ. Cette méthode permet une estimation asymptotiquement sans biais de la variance, y compris pour des fractiles.

L'extension de cette méthode a des plans de sondage plus complexes a été étudiée notamment par Rao et Shao (1996) ; voir également Davison et Sardy (2007) pour plus de références.

Le Jackknife

Sondage aléatoire simple sans remise

Comme la linéarisation, le Jackknife est une technique permettant de prendre en compte la forme de la statistique dans le calcul de variance, mais pas le plan de sondage.

Pour obtenir une estimation consistante de variance avec un SAS sans remise, l'estimateur de variance doit être remplacé par

$$\begin{aligned}v_{JACK} \left[\hat{\theta} \right] &= \frac{1-f}{n} s_{\hat{\theta}}^2 \\ &= (1-f) \left[\frac{n-1}{n} \sum_{k \in S} \left(\hat{\theta}_{-k} - \frac{1}{n} \sum_{l \in S} \hat{\theta}_{-l} \right)^2 \right].\end{aligned}$$

On peut le voir comme l'estimateur de variance par linéarisation, où la variable linéarisée est remplacée par son approximation Jackknife.

Tirage stratifié

L'estimation de variance Jackknife se généralise facilement au cas d'un sondage aléatoire simple stratifié. Si le paramètre d'intérêt est de la forme $\theta = \sum_{h=1}^H \theta_h$, l'estimateur de variance Jackknife est donné par

$$\begin{aligned} v_{JACK} \left[\hat{\theta} \right] &= \sum_{h=1}^H v_{JACK} \left[\hat{\theta}_h \right] \\ &= \sum_{h=1}^H \frac{1-f_h}{n_h} S_{\hat{\theta}_h}^2 \\ &= \sum_{h=1}^H (1-f_h) \frac{n_h-1}{n_h} \sum_{k \in S_h} \left(\hat{\theta}_{h,-k} - \frac{1}{n_h} \sum_{l \in S_h} \hat{\theta}_{h,-l} \right)^2 \end{aligned}$$

où l'estimateur $\hat{\theta}_{h,-k}$ est recalculé en supprimant l'unité k de l'échantillon S_h , et en multipliant les autres unités de S_h par un facteur $n_h/(n_h-1)$.

Tirage à probabilités inégales

Dans le cas d'un tirage à probabilités inégales, Campbell (1980) propose d'utiliser l'estimateur de variance de Horvitz-Thompson ou de Yates-Grundy, en injectant l'approximation Jackknife de la variable linéarisée (voir aussi Berger et Skinner, 2005).

Berger (2007) a proposé d'utiliser l'estimateur de variance de Hajek dans le cas d'un tirage à forte entropie, i.e. d'utiliser l'estimateur de variance

$$v_{HAJ} [\hat{t}_{y\pi}] = \sum_{k \in S} (1 - \pi_k) \left(\frac{y_k}{\pi_k} - \hat{R} \right)^2 \quad \text{où} \quad \hat{R} = \frac{\sum_{k \in S} \frac{y_k}{\pi_k} (1 - \pi_k)}{\sum_{k \in S} (1 - \pi_k)},$$

en remplaçant la variable y_k par une linéarisée de type Jackknife $\tilde{\theta}_k$.

Tirage multi-degrés

Le Jackknife peut être étendu au cas d'un tirage multidegrés où les unités primaires sont sélectionnées **avec remise** : on supprime successivement chaque unité primaire.

L'estimateur Jackknife de variance s'obtient en remplaçant dans l'estimateur de variance

$$v [\hat{t}_{y\pi}] = \frac{m}{m-1} \sum_{u_i \in S_I} \left[\frac{\hat{t}_{y_i\pi}}{\pi_{Ii}} - \frac{\hat{t}_{y\pi}}{m} \right]^2$$

la variable d'intérêt y par la linéarisée Jackknife.

Krewski et Rao (1981) montrent la consistance de cet estimateur de variance, voir également Rao et Wu (1985), Kovar et al. (1988).

Extensions

Le delete-1 Jackknife implique un total de n suppressions, ce qui peut être prohibitif si la taille d'échantillon est grande. Les suppressions peuvent être également réalisées par blocs de d unités à la fois (Shao et Wu, 1985). On parle alors de delete- d Jackknife.

Cette méthode a également été proposée par Shao et Wu (1989) afin de produire une estimation consistante de variance pour le Jackknife pour des paramètres non lisses tels que les quantiles.

Le Jackknife : résumé

L'estimation de variance Jackknife est fortement comparable à l'estimation de variance par linéarisation (avec ses avantages et ses inconvénients).

Le Jackknife permet de prendre en compte la forme du paramètre dans l'estimation de variance, mais la connaissance du plan de sondage reste nécessaire.

Le Jackknife permet d'éviter le calcul explicite de la variable linéarisée, mais sa mise en oeuvre peut être complexe si le nombre de suppressions est grand.

Le Bootstrap

Sondage aléatoire simple sans remise

L'extension au cas d'un sondage simple sans remise est problématique. Deux approches ont principalement été proposées dans la littérature :

- ▶ Approche 1 : Reproduire le mécanisme d'échantillonnage d'origine, dans l'esprit du Bootstrap d'Efron.
- ▶ Approche 2 : Reproduire un estimateur de variance (approximativement) sans biais dans le cas de l'estimation d'une moyenne (ou d'un total).

Approche 1 : la méthode de Gross

Gross (1980) suggère d'utiliser l'échantillon S afin de recréer une pseudopopulation U^* , dans laquelle on rééchantillonne selon un SAS sans remise :

1. Obtenir U^* en dupliquant N/n fois chaque unité k de S ,
2. Tirer dans U^* un rééchantillon S^* selon un SAS sans remise de taille n , pour obtenir un estimateur $\hat{\theta}^*$,
3. Répéter l'étape 2 indépendamment B fois, et estimer $V[\theta]$ par

$$v_{BOOT}^B = \frac{1}{B-1} \sum_{b=1}^B \left[f(\bar{y}_b^*) - \frac{1}{B} \sum_{c=1}^B f(\bar{y}_c^*) \right]^2.$$

Procédure connue sous le nom de Bootstrap sans remise (BWO), ou Bootstrap populationnel.

Approche 1 : la méthode de Gross

Dans le cas où la variable y est unidimensionnelle, avec $\theta = \mu_y$ et $\hat{\theta} = \bar{y}$, on a :

$$\begin{aligned}v_{BOOT} [\hat{\theta}] &= \frac{N(n-1)}{n(N-1)} \frac{1-f}{n} s_y^2 \\ &= \frac{N(n-1)}{n(N-1)} v[\bar{y}].\end{aligned}$$

L'estimateur de variance Bootstrap est donc biaisé. Le biais est négligeable si la taille d'échantillon est grande, mais peut être appréciable si n est borné (cas d'un sondage aléatoire simple stratifié avec une stratification très fine).

Exemple

Population U de taille 10, dans laquelle un échantillon de taille $n = 5$ est sélectionné selon un SAS sans remise.

Valeurs de la variable y échantillonnées :

$$S \equiv \{1, 3, 2, 10, 8\}.$$

Pseudo-population obtenue :

$$U^* \equiv \{1, 1, 3, 3, 2, 2, 10, 10, 8, 8\}.$$

On rééchantillonne dans la population U^* selon un SAS sans remise de taille 5. Une variable de poids Bootstrap donne le nombre de fois où l'unité est sélectionnée dans le rééchantillon.

Exemple

k	y_k						
1	1						
2	3						
3	2						
4	10						
5	8						

Exemple

k	y_k						
1	1						
2	3						
3	2						
4	10						
5	8						
	$\bar{y} = 4.8$						

$$v[\bar{y}] = (1 - f) \frac{s_y^2}{n} = 1.57$$

Exemple

k	y_k	W_1					
1	1	2					
2	3	0					
3	2	1					
4	10	1					
5	8	1					
	$\bar{y} = 4.8$						

$$v[\bar{y}] = (1 - f) \frac{s_y^2}{n} = 1.57$$

Exemple

k	y_k	W_1					
1	1	2					
2	3	0					
3	2	1					
4	10	1					
5	8	1					
	$\bar{y} = 4.8$	$\bar{y}_1^* = 4.4$					

$$v[\bar{y}] = (1 - f) \frac{s_y^2}{n} = 1.57$$

Exemple

k	y_k	W_1	W_2	W_3	W_4	W_5	W_6
1	1	2	2	0	0	1	1
2	3	0	2	2	2	1	0
3	2	1	1	2	0	2	1
4	10	1	0	1	2	0	1
5	8	1	0	0	1	1	2
	$\bar{y} = 4.8$	$\bar{y}_1^* = 4.4$					

$$v[\bar{y}] = (1 - f) \frac{s_y^2}{n} = 1.57$$

Exemple

k	y_k	W_1	W_2	W_3	W_4	W_5	W_6
1	1	2	2	0	0	1	1
2	3	0	2	2	2	1	0
3	2	1	1	2	0	2	1
4	10	1	0	1	2	0	1
5	8	1	0	0	1	1	2
	$\bar{y} = 4.8$	$\bar{y}_1^* = 4.4$	$\bar{y}_2^* = 2$	$\bar{y}_3^* = 4$	$\bar{y}_4^* = 6.8$	$\bar{y}_5^* = 3.2$	$\bar{y}_6^* = 5.8$

$$v[\bar{y}] = (1 - f) \frac{s_y^2}{n} = 1.57$$

Exemple

k	y_k	W_1	W_2	W_3	W_4	W_5	W_6
1	1	2	2	0	0	1	1
2	3	0	2	2	2	1	0
3	2	1	1	2	0	2	1
4	10	1	0	1	2	0	1
5	8	1	0	0	1	1	2
	$\bar{y} = 4.8$	$\bar{y}_1^* = 4.4$	$\bar{y}_2^* = 2$	$\bar{y}_3^* = 4$	$\bar{y}_4^* = 6.8$	$\bar{y}_5^* = 3.2$	$\bar{y}_6^* = 5.8$

$$v[\bar{y}] = (1 - f) \frac{s_y^2}{n} = 1.57$$

$$B = 50 \Rightarrow v_{BOOT}^B[\bar{y}] = 1.35$$

Extensions de la méthode de Gross

L'algorithme de Bootstrap suppose que le nombre de duplications N/n est entier. Dans le cas contraire, une adaptation est nécessaire.

Bickel et Freedman (1985) et Chao et Lo (1984) proposent une randomisation entre deux pseudo-populations. Booth et al. (1994) proposent d'arrondir N/n pour la duplication, et de compléter la pseudo-population U^* par sondage aléatoire simple dans S .

Sitter (1992) propose une randomisation sur le nombre de duplications et la taille de rééchantillon afin de retrouver (en moyenne) le bon estimateur de variance pour \bar{y} .

Approche 2 : le Bootstrap avec remise

Mac Carthy et Snowden (1985) proposent de rééchantillonner avec remise dans S , pour obtenir un rééchantillon S^* de taille m . Procédure connue sous le nom de Bootstrap avec remise (BWR).

Mac Carthy et Snowden suggèrent le choix

$$m = \frac{n - 1}{1 - f},$$

qui conduit à une estimation sans biais de variance pour \bar{y} . Si ce choix est impossible, une randomisation sur la taille de rééchantillon est nécessaire.

Approche 2 : le Rescaling Bootstrap

Rao et Wu (1985) proposent également de rééchantillonner avec remise dans S , pour obtenir un rééchantillon S^* de taille m . Dans chaque rééchantillon S^* , les poids Bootstrap sont ajustés afin de restituer l'estimateur sans biais de variance pour \bar{y} , voir également Rao et al. (1992).

Rao et Wu argumentent de la consistance de l'estimateur de variance, dans le cas où le paramètre est une fonction lisse de moyennes. Ils suggèrent une valeur optimale pour la taille de rééchantillon m , mais cette valeur peut être non entière.

Approche 2 : le Mirror-Match Bootstrap

Cette méthode proposée par Sitter (1992) combine les deux approches. Un rééchantillon S^* est obtenu en sous-échantillonnant k fois selon un SAS de taille n^* , et en agglomérant les sous-échantillons obtenus.

Le choix $k = \frac{n(1-f)}{n^*(1-f^*)}$ où $f^* = n^*/n$ permet de retrouver l'estimateur de variance sans biais de \bar{y} . Sitter suggère le choix $f^* = f$, et de randomiser k et n^* si nécessaire.

Là encore, Sitter argumente de la consistance de l'estimateur de variance, dans le cas où le paramètre est une fonction lisse de moyennes.

Quelle approche choisir?

La littérature est assez contradictoire à ce sujet. Certains jeux de simulation (Rao et Wu, 1984, Sitter, 1993) plaident en faveur des méthodes ad-hoc (approche 2). D'autres études (Presnell et Booth, 1994, Davison et Hinkley, 1997, Davison et Sardy, 2007) suggèrent le contraire.

La méthode proposée par Gross est intuitivement plus proche du principe de Bootstrap proposé par Efron. Toutes les méthodes de Bootstrap proposées présentent des difficultés pratiques. La généralisation à un plan de sondage complexe est laborieuse, en dehors de quelques plans de sondage particuliers.

Tirage stratifié

Les procédures proposées se généralisent facilement au cas d'un sondage aléatoire simple stratifié : la même procédure est appliquée dans chaque strate.

Si la stratification est fine, un biais même faible d'un estimateur de variance dans chaque strate peut conduire à un biais global important. Pour cette raison, Rao et Wu (1984) et Sitter (1993) recommandent l'utilisation des méthodes du Rescaled Bootstrap et du Mirro-Match Bootstrap.

Les résultats obtenus par Davison et Hinkley (1997) montrent au contraire qu'une approche de type Gross donne de bons résultats, même avec un nombre important de strates.

Tirage à probabilités inégales

Le principe de la méthode de Gross peut se généraliser à un tirage à probabilités inégales, en utilisant le principe d'estimation de Horvitz-Thompson : une unité k de l'échantillon représente $1/\pi_k$ unités de la population.

La pseudo-population U^* est obtenue $1/\pi_k$ fois chaque individu k de l'échantillon. Le plan de sondage d'origine est appliqué dans U^* . Une simulation (Chauvet, 2007) montre un bon comportement de cette méthode pour un tirage à forte entropie, mais :

- ▶ problème de la duplication ($1/\pi_k$ rarement entier)
- ▶ validation théorique de la méthode?
- ▶ échec pour un tirage à faible entropie (ex : tirage systématique).

Voir également Antal et Tillé (2009), pour une approche ad-hoc.

Exemple

Population U de taille $N = 10$, dans laquelle un échantillon S de taille $n = 2$ est sélectionné selon un tirage systématique ordonné à probabilités égales (tri de la population selon une variable auxiliaire x).

k	1	2	3	4	5	6	7	8	9	10
-----	---	---	---	---	---	---	---	---	---	----

Exemple

Population U de taille $N = 10$, dans laquelle un échantillon S de taille $n = 2$ est sélectionné selon un tirage systématique ordonné à probabilités égales (tri de la population selon une variable auxiliaire x).

k	1	2	3	4	5	6	7	8	9	10
x_k	1	3	4	9	12	15	20	35	36	39

Exemple

Population U de taille $N = 10$, dans laquelle un échantillon S de taille $n = 2$ est sélectionné selon un tirage systématique ordonné à probabilités égales (tri de la population selon une variable auxiliaire x).

k	1	2	3	4	5	6	7	8	9	10
x_k	1	3	4	9	12	15	20	35	36	39
y_k	2	1	7	5	9	11	8	14	5	20

Exemple

Population U de taille $N = 10$, dans laquelle un échantillon S de taille $n = 2$ est sélectionné selon un tirage systématique ordonné à probabilités égales (tri de la population selon une variable auxiliaire x).

k	1	2	3	4	5	6	7	8	9	10
x_k	1	3	4	9	12	15	20	35	36	39
y_k	2	1	7	5	9	11	8	14	5	20

On tire un individu au hasard parmi les 5 premiers.

Exemple

Population U de taille $N = 10$, dans laquelle un échantillon S de taille $n = 2$ est sélectionné selon un tirage systématique ordonné à probabilités égales (tri de la population selon une variable auxiliaire x).

k	1	2	3	4	5	6	7	8	9	10
x_k	1	3	4	9	12	15	20	35	36	39
y_k	2	1	7	5	9	11	8	14	5	20

On tire un individu au hasard parmi les 5 premiers. Puis on fait un bond de taille 5.

Exemple

Si on applique la méthode de Gross, on obtient la pseudo-population U^* :

k	2	7	2	7	2	7	2	7	2	7
-----	---	---	---	---	---	---	---	---	---	---

Exemple

Si on applique la méthode de Gross, on obtient la pseudo-population U^* :

k	2	7	2	7	2	7	2	7	2	7
x_k	3	20	3	20	3	20	3	20	3	20

Exemple

Si on applique la méthode de Gross, on obtient la pseudo-population U^* :

k	2	7	2	7	2	7	2	7	2	7
x_k	3	20	3	20	3	20	3	20	3	20
y_k	1	8	1	8	1	8	1	8	1	8

On la trie selon la variable x avant de réaliser le tirage systématique de taille 2 :

k	2	2	2	2	2	7	7	7	7	7
x_k	3	3	3	3	3	20	20	20	20	20
y_k	1	1	1	1	1	8	8	8	8	8

⇒ on échantillonne toujours les mêmes valeurs.

Tirage multi-degrés

Le Bootstrap peut être facilement appliqué au cas d'un tirage multidegrés où les unités primaires sont sélectionnées **avec remise** : on rééchantillonne parmi les unités primaires uniquement.

Dans le cas où les unités primaires sont sélectionnées sans remise, on obtient une estimation de variance approximativement sans biais si la fraction du sondage du 1er degré est faible. Le Bootstrap permet en fait de capter le 1er terme de variance donné par l'approche renversée (Haziza, 2009).

Cas d'un tirage multidegrés général encore peu traité (Funaoka et al., 2006).

Application : Enquête HHANES

Présentation de l'enquête (Korn et Graubard, 1999)

L'Enquête nationale sur la santé et la nutrition auprès de la population hispanique (HHANES) a été conduite aux Etats-Unis en 1982-1983. Lors de cette édition, un échantillon de 8 500 individus environ a été sélectionné.

Objectifs de l'enquête :

- ▶ estimer la prévalence de certaines maladies, et les facteurs de risque auprès de la population hispanique,
- ▶ estimer des paramètres de santé dans la population,
- ▶ étudier les liens entre les maladies et les facteurs de risque.

Présentation de l'enquête (Korn et Graubard, 1999)

L'échantillon a été sélectionné selon un plan de sondage à plusieurs degrés.

Le territoire est d'abord découpé en M Unités Primaires (donnés par des comtés ou des regroupements de comtés) regroupées au sein de $H = 8$ strates. On sélectionne $m_h = 2$ UP dans chaque strate avec des probabilités de sélection inégales.

La sélection des UP est assimilée à un tirage avec remise de 2 UP dans chaque strate.

Présentation de l'enquête (Korn et Graubard, 1999)

La sélection de ces UP est suivie de 3 autres degrés de tirage :

- ▶ sélection d'aires (villes ou quartiers) dans les UP,
- ▶ sélection de ménages dans les aires,
- ▶ sélection d'individus dans les ménages.

Le tableau suivant donne, dans chaque strate et pour chaque UP, l'estimation du niveau moyen d'hémoglobine chez les femmes.

Strate h	UP u_{hi}	Taille d'éch. n_{hi}	Niveau moyen estimé \bar{y}_{hi}	Somme des poids $W_{hi} (\times 10^5)$
1	1	404	13.46	4.58
	2	391	13.34	4.23
2	1	87	13.04	1.34
	2	84	13.08	1.31
3	1	181	13.06	1.84
	2	239	13.42	2.61
4	1	148	13.37	2.04
	2	127	13.58	1.92
5	1	215	13.57	2.80
	2	214	13.28	3.13
6	1	238	14.04	2.64
	2	193	12.91	2.39
7	1	231	13.86	2.67
	2	223	13.09	2.69
8	1	142	13.62	1.47
	2	252	14.05	3.36

Estimation du niveau moyen d'hémoglobine

L'estimateur du niveau moyen d'hémoglobine chez les femmes, noté μ_y , est donné par

$$\bar{y} = \frac{\sum_{h=1}^H \sum_{u_{ih} \in S_{Ih}} W_{hi} \bar{y}_{hi}}{\sum_{h=1}^H \sum_{u_{ih} \in S_{Ih}} W_{hi}}.$$

En utilisant la linéarisation, on obtient qu'un estimateur (approximativement) sans biais de variance est donné par

$$v[\bar{y}] = \sum_{h=1}^H \left(\frac{m_h}{\sum_{h=1}^H \sum_{u_{ih} \in S_{Ih}} W_{hi}} \right)^2 \frac{s_{eh}^2}{m_h}$$

avec

$$s_{eh}^2 = \frac{1}{m_h - 1} \sum_{u_{ih} \in S_{Ih}} (W_{hi} \bar{e}_{hi})^2 \quad \text{et} \quad e_k = y_k - \bar{y}.$$

Estimation du niveau moyen d'hémoglobine

L'application numérique donne $\bar{y} = 13.46$ (g/dl) et $v[\bar{y}] = 0.01$, soit un coefficient de variation de $cv[\bar{y}] \simeq 0.8\%$.

L'estimateur de variance Jackknife est donné par

$$v_{JACK}[\bar{y}] = \sum_{h=1}^H \frac{m_h - 1}{m_h} \sum_{u_{ih} \in S_{Ih}} \left(\bar{y}_{h,-i} - \frac{1}{m_h} \sum_{u_{jh} \in S_{Ih}} \bar{y}_{h,-j} \right)^2$$

Strate h	UP u_{hi}	\bar{y}_{hi}	W_{hi}	$W_{h,-1}$	$W_{h,-2}$	$W_{h,-3}$
1	1	13.46	4.58	0	9.16	4.58
	2	13.34	4.23	8.46	0	4.23
2	1	13.04	1.34	1.34	1.34	0
	2	13.08	1.31	1.31	1.31	2.62
3	1	13.06	1.84	1.84	1.84	1.84
	2	13.42	2.61	2.61	2.61	2.61
4	1	13.37	2.04	2.04	2.04	2.04
	2	13.58	1.92	1.92	1.92	1.92
5	1	13.57	2.80	2.80	2.80	2.80
	2	13.28	3.13	3.13	3.13	3.13
6	1	14.04	2.64	2.64	2.64	2.64
	2	12.91	2.39	2.39	2.39	2.39
7	1	13.86	2.67	2.67	2.67	2.67
	2	13.09	2.69	2.69	2.69	2.69
8	1	13.62	1.47	1.47	1.47	1.47
	2	14.05	3.36	3.36	3.36	3.36
		\bar{y} .	13.46	13.44	13.47	13.46

Estimation du niveau moyen d'hémoglobine

On obtient finalement $v_{JACK}[\bar{y}] \simeq 0.01$ (quasiment la même valeur qu'avec la linéarisation).

On peut également obtenir une estimation de variance par Bootstrap en tirant des rééchantillons S_b^* , $b = 1, \dots, B$ avec remise, de taille 2 dans chaque strate, mais la variance est sous-estimée avec un facteur $m_h/(m_h - 1) = 2$.

On peut également appliquer une correction en imposant que chaque rééchantillon contienne exactement une UP de chaque strate : on obtient la méthode des demi-échantillons équilibrés et l'estimateur de variance

$$v_{HS}[\bar{y}] = \frac{1}{B} \sum_{b=1}^B \left(\bar{y}_{hb}^* - \frac{1}{B} \sum_{c=1}^B \bar{y}_{hc}^* \right)^2.$$

Strate h	UP u_{hi}	\bar{y}_{hi}	W_{hi}	$W_{h,1}$	$W_{h,2}$	$W_{h,3}$
1	1	13.46	4.58	0	9.16	9.16
	2	13.34	4.23	8.46	0	0
2	1	13.04	1.34	0	0	0
	2	13.08	1.31	2.62	2.62	2.62
3	1	13.06	1.84	3.68	0	0
	2	13.42	2.61	0	5.22	5.22
4	1	13.37	2.04	0	0	4.08
	2	13.58	1.92	3.84	3.84	0
5	1	13.57	2.80	0	5.60	5.60
	2	13.28	3.13	6.26	0	0
6	1	14.04	2.64	5.28	0	0
	2	12.91	2.39	0	4.78	4.78
7	1	13.86	2.67	5.34	0	5.34
	2	13.09	2.69	0	5.38	0
8	1	13.62	1.47	0	0	0
	2	14.05	3.36	6.72	6.72	6.72
		\bar{y} .	13.46	13.58	13.44	13.52

Bibliographie

Sur le rééchantillonnage en Statistique classique

- ▶ Davison, A.C., and Hinkley, D.V. (1997). *Bootstrap Methods and their application*. Cambridge University Press.
- ▶ Efron, B. (1979). *Bootstrap methods : another look at the Jackknife*. Annals of Statistics, 7, p. 1-26.
- ▶ Hampel, F.R., and Ronchetti, E.M., and Rousseeuw, P.J., and Stahel, W.A. (1986). *Robust Statistics: The Approach Based on Influence Functions*. New York: Wiley.
- ▶ Quenouille, M. (1949). *Approximation tests of correlation in time series*. JRSS B, 11, p. 18-84.
- ▶ Shao, J., and Tu, D. (1995). *The jackknife and Bootstrap*. New-York, Springer.
- ▶ Shao, J., and Wu, C.F.J. (1989). *A general theory for jackknife variance estimation*. Annals of Statistics, 17, p. 1176-1197.
- ▶ Tukey, J. (1958). *Bias and confidence in not quite large samples*. Annals of Mathematical Statistics, 29, p. 614.

Sur le Bootstrap en Sondage

- ▶ Bickel, P.J., and Freedman, D.A. (1981). *Asymptotic normality and the bootstrap in stratified sampling*. Annals of Statistics, 12, 470-482.
- ▶ Booth, J.G., and Butler, R.W., and Hall, P. (1994). *Bootstrap Methods for Finite Populations*. JASA, 89, p. 1282-1289.
- ▶ Chao, H., and Lo, K.Y. (1985). *A Bootstrap Method for Finite Populations*. Sankhya, 47, p. 399-405.
- ▶ Chauvet, G. (2007). *Méthodes de Bootstrap en population finie*. PhD Thesis, université de Rennes 2.
- ▶ Kovar J.G., and Rao J.N.K., and Wu C.F.J (1988). *Bootstrap and other methods to measure errors in survey estimates*. Canadian Journal of Statistics, 16, p. 25-45.
- ▶ Mc Carthy, P.J., and Snowden, C.B. (1985). *The Bootstrap and finite population sampling*. Public Health Service Publication 1369.
- ▶ Gross, S.T. (1980). *Median estimation in sample surveys*. Proceedings of the Survey Research Methods Section, American Statistical Association, p. 181-184.
- ▶ Rao, J.N.K., and Wu, C.F.J. (1988). *Resampling inference with complex survey data*. JASA, 83, p. 231-241.
- ▶ Rao, J.N.K., and Wu, C.F.J., and Yue, K. (1992). *Some recent work on resampling methods for complex surveys*. Survey Methodology, 18, p. 209-217.
- ▶ Sitter, R.R. (1992). *A resampling procedure for complex survey data*. JASA, 87, p. 755-765.

Sur le rééchantillonnage en Sondage

- ▶ Berger, Y.G. (2007). *A Jackknife Variance Estimator for Unistage Stratified Samples with Unequal Probabilities*, *Biometrika*, 94, p. 953-964.
- ▶ Berger Y. G., and Skinner C. J. (2005). *A jackknife variance estimator for unequal probability sampling*. *JRSS B*, 67, p. 79-89.
- ▶ Campbell C. (1980). *A different view of finite population estimation*. *Proceedings of the Survey Research Methods Section, American Statistical Association*, p. 319-324.
- ▶ Davison, A.C., and Sardy, S. (2006). *Méthodes de rééchantillonnage pour l'estimation de variance en sondages*. *Journal de la SFDS*, 147, p. 3-32.
- ▶ Krewski D., and Rao, J.N.K. (1981). *Inference from stratified samples: properties of the linearization, jackknife and balanced repeated replication methods*. *Annals of Statistics*, 9, p. 1010-1019.
- ▶ Mc Carthy, P.J. (1969). *Pseudo-replication: Half-samples*. *International Statistical Review*, 37, p. 239-264.
- ▶ Rao, J.N.K., and Shao, J. (1996). *On balanced half-sample variance estimation in stratified sampling*. *JASA*.
- ▶ Rao, J.N.K., and Wu, C.F.J. (1985). *Inference from stratified samples: second-order analysis of three methods for nonlinear statistics*. *JASA*, 80, p. 620-630.

Sur l'estimation de variance en Sondage

- ▶ Berger, Y.G. (1996), *Asymptotic Variance for Sequential Sampling without Replacement with Unequal Probabilities*, Survey Methodology, 22.
- ▶ Berger, Y.G. (1998), *Variance Estimation Using List Sequential Scheme for Unequal Probability Sampling*, Journal of Official Statistics, 14.
- ▶ Caron, N. (1998). *Le logiciel POULPE : aspects méthodologiques*, Actes des Journées des Méthodologie Statistique, Insee.
- ▶ Deville, J.-C. (1999), *Variance estimation for complex statistics and estimators : linearization and residual techniques*, Survey Methodology, 25.
- ▶ Haziza, D., and Beaumont, J-F. (2005). *Estimation de la variance dans le cas d'échantillonnage à deux phases*. Actes du colloque francophone sur les Sondages, Québec.
- ▶ Kovacevic, M. S., and Binder, D.A. (1997). *Variance estimation for measures of income inequality and polarization*. Journal of Official Statistics, 13, p. 41-58.
- ▶ Matei, A, Tillé, Y. (2005), *Evaluation of variance approximations and estimators in unequal probability sampling with maximum entropy*, Journal of Official Statistics, 21.
- ▶ Petit, J-N. (1998), *Le logiciel POULPE : modélisation informatique*, Actes des Journées des Méthodologie Statistique, Insee.
- ▶ Shao, J., Steel, P. (1999). *Variance Estimation for Survey Data with Composite Imputation and Nonnegligible Sampling Fractions*. JASA 94, p. 254-265.

Sur les Sondages

- ▶ Ardilly, P. (2006). *Les Techniques de Sondage*. Technip.
- ▶ Cochran, W.G. (1977). *Sampling Techniques*. Wiley, New-York.
- ▶ Deville, J-C., and Tillé, Y. (1998). *Unequal probability sampling without replacement through a splitting method*. *Biometrika*, 85, p. 89-101.
- ▶ Deville, J-C., and Tillé, Y. (2004). *Efficient balanced sampling: the cube method*. *Biometrika*, 128, p. 569-591.
- ▶ Goga, C., Deville, J-C., and Ruiz-Gazen, A. (2009). *Use of functionals in linearization and composite estimation with application to two-sample survey data*. *Biometrika*, 96, P. 691-710.
- ▶ Hájek, J. (1964). *Asymptotic theory of rejective sampling with varying probabilities from a finite population*, *Annals of Mathematical Statistics*, 35, p. 1491-1523.
- ▶ Haziza, D. (2009). *Imputation and inference in the presence of missing data*. *Handbook of Statistics, Volume 29, Sample Surveys : Theory Methods and Inference*, Editors : C.R. Rao and D. Pfeffermann.
- ▶ Madow, L.H, and Madow, W.G. (1944). *On the theory of systematic sampling, II*. *Annals of Mathematical Statistics*, 20, p. 333-354.