THÈSE

présentée devant

L'UNIVERSITÉ DE RENNES 2
École doctorale - Humanités et Sciences de l'Homme

MÉTHODES DE BOOTSTRAP EN POPULATION FINIE

Option : Mathématiques Appliquées
Discipline : Statistiques

Présentée par Guillaume CHAUVET

Sous la direction de Michel CARBON et Jean-Claude DEVILLE

Soutenue le 14 décembre 2007 devant le jury composé de :

M. Pascal ARDILLY	Administrateur de l'Insee	Rapporteur
M. Patrice BERTAIL	Prof. Université de Paris 10	Rapporteur
M. Michel CARBON	Prof. Université de Rennes 2	Directeur de recherche
M. Jean-Claude DEVILLE	Inspecteur Général de l'Insee, Crest-ENSAI	Directeur de recherche
M. Peter HALL	Prof. Université de Melbourne	Rapporteur
M. Yves TILLÉ	Prof. Université de Neuchâtel	Examinateur

Laboratoire de Statistique d'Enquêtes, CREST-ENSAI

If nobody makes you do it, it counts as fun.
Bill Watterson

Facts are stubborn things, but statistics are more pliable.
Mark Twain

Remerciements

Je remercie très sincérement Jean-Claude Deville de la confiance qu'il m'a accordée en me donnant l'opportunité de faire cette thèse, et ce même si celle-ci se sera singulièrement éloignée de sa thématique de départ. La plus grande partie de ce que je sais sur l'estimation de variance en particulier, et la théorie des Sondages en général, provient des discussions formelles ou informelles que nous avons pu avoir durant ces quatre dernières années. Je lui en suis particulièrement reconnaissant.
Je remercie également Michel Carbon d'avoir bien voulu accepter de m'encadrer officiellement dans ce travail de thèse, et de son aide dans la préparation de cette soutenance.

Les travaux de Patrice Bertail et Peter Hall dans le domaine du Bootstrap, et de Pascal Ardilly dans le domaine de la Statistique d'Enquête, sont pour moi des références dont j'ai largement bénéficié dans la rédaction de cette thèse. Je les remercie d'avoir accepté d'en être les rapporteurs. Merci tout particulièrement à Pascal Ardilly pour sa lecture critique et méticuleuse d'une version préliminaire de ce manuscrit.
Le cours et les livres d'Yves Tillé ont également constitué pour moi de précieux outils de compréhension de la théorie des Sondages. Je le remercie d'avoir accepté de participer à mon jury de thèse.

Mon travail de recherche a largement bénéficié des discussions et collaborations que j'ai pu initier durant cette thèse. Je remercie Anne, Camélia, Christophe, David, Denaïs, Mohammed, Yves et bien d'autres pour le travail passé et à venir. Je remercie également tous les élèves que j'ai pu cotoyer dans le cadre des projets statistiques, projets qui m'ont souvent été d'une aide très précieuse.

Parce que toute réussite individuelle est avant tout collective, merci à toutes les personnes, enseignants, permanents, élèves et anciens de l'Ensai qui m'ont soutenu de leurs encouragements et de leur amitié.

Merci à ma famille, Anaëlle, Estelle, Frédéric, Michèle et Pierre pour leur soutien sans faille. Enfin et surtout, merci à Brigitte et Nora pour leur patience et leur amour, et pour me rappeler le plus souvent possible que les Sondages c'est sympa, mais qu'il fait beau dehors.

Table des matières

Liste des tableaux

Table des figures

Présentation

Cette thèse est consacrée à l'estimation de précision dans le cas d'une population finie, et plus particulièrement à l'utilisation de méthodes de type Bootstrap. Le Bootstrap est un outil très largement utilisé dans le cas d'une analyse statistique en population infinie. Beaucoup d'adaptations ont déjà été proposées dans le cas d'un sondage en population finie. Nous montrons dans ce travail que le principe de substitution qui est à la base du Bootstrap admet un équivalent naturel en population finie, le principe d'estimation de Horvitz-Thompson, et que la méthode proposée à l'origine par Gross (1980) peut se généraliser à une gamme très étendue de plans de sondage.

Par rapport à la technique générale et efficace qu'est la linéarisation, le gain que nous attendons d'une méthode de type Bootstrap est avant tout d'ordre pratique. La linéarisation permet de produire très simplement des estimations de variance et des intervalles de confiance basés sur une approximation normale, pour peu que l'on dispose du fichier d'enquête et d'une vision détaillée des différentes étapes de traitement (échantillonnage, correction de la non-réponse, repondération, ...). Si ces deux éléments sont généralement accessibles au concepteur de l'enquête, ils ne le sont que rarement pour l'utilisateur. L'existence d'une méthode de Bootstrap qui permettrait d'accoler au fichier d'enquête des poids issus du rééchantillonnage donnerait la possibilité de produire a posteriori une estimation de précision, ou un intervalle de confiance de type percentile, pour une gamme très large de statistiques et pour un domaine quelconque.

Le premier chapitre introduit quelques rappels sur l'échantillonnage en population finie. Afin que chaque chapitre puisse se lire de la façon la plus indépendante possible, nous avons choisi de limiter ces rappels à l'essentiel, c'est pourquoi certains passages pourront paraître redondants dans la suite

de ce manuscrit. Nous proposons une présentation synthétique de la notion d'asymptotique en théorie des Sondages, et présentons les principales méthodes d'estimation de précision.

Le chapitre 2 propose un rappel sur les méthodes de type Bootstrap proposées dans la littérature pour un sondage aléatoire simple. Ces méthodes sont nombreuses, et se répartissent schématiquement en deux groupes : celles qui se calent explicitement sur l'estimateur de variance dans le cas linéaire (on parlera de méthodes ad-hoc), et celles qui s'appuient sur un principe de substitution (on parlera de méthodes de type plug-in). Deux algorithmes simplifiés et une nouvelle méthode de Bootstrap sont également proposés.

Le chapitre 3 introduit un nouvel algorithme de Bootstrap, qui généralise la méthode de Booth et al. (1994) au cas des probabilités inégales. Nous montrons à l'aide de la linéarisation par la fonction d'influence que cet algorithme donne une estimation consistante de variance pour les plans de sondage de grande entropie, incluant le tirage poissonien et le tirage réjectif. Une méthode simplifiée de Bootstrap est également proposée afin de réduire le fardeau de rééchantillonnage ; elle n'induit pas de perte significative de précision si les probabilités d'inclusion restent limitées.

Nous donnons dans le chapitre 4 quelques rappels sur l'échantillonnage équilibré, et sa mise en oeuvre via l'algorithme du Cube de Deville and Tillé (2004). Une méthode rapide d'échantillonnage équilibré est proposée. Nous donnons une justification des formules de variance de Deville and Tillé (2005) à l'aide d'une technique d'approximation des probabilités d'inclusion proposée à l'origine par Hájek (1981), et justifions la consistance de la méthode de Bootstrap introduite au chapitre précédent dans le cas d'un échantillonnage équilibré à entropie maximale. Une nouvelle formule analytique d'approximation de précision est également proposée, ainsi qu'une extension de la méthode mirror-match de Sitter (1992) pour un tirage équilibré à entropie maximale avec des probabilités d'inclusion égales.

Le chapitre 5 aborde le cas d'un échantillonnage complexe. Les algorithmes de Bootstrap proposés au chapitre 3 s'étendent de façon immédiate au cas d'un plan stratifié avec un nombre fixe et fini de strates. Nous montrons également que ces algorithmes peuvent être, avec une modification, étendus au cas de l'échantillonnage multidegrés, et que le calage éventuel d'un es-

timateur peut être aisément pris en compte dans l'estimation Bootstrap de précision.

Le chapitre 6 propose une application des méthodes de Bootstrap développées au cas du Nouveau Recensement de la Population. Une autre méthode de Bootstrap est également développée, afin de prendre en compte la méthode de régression géographique pondérée retenue pour produire des estimations basées sur trois années de collecte du Recensement.

Bibliographie

Booth, J., Butler, R., and Hall, P. (1994). Bootstrap methods for finite populations. *Journal of the American Statistical Association*, 89 :1282–1289.

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

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

Gross, S. (1980). Median estimation in sample surveys. *ASA Proceedings of Survey Research*, pages 181–184.

Hájek, J. (1981). *Sampling from a Finite Population*. Marcel Dekker, New York.

Sitter, R. (1992). A resampling procedure for complex survey data. *Journal of the American Statistical Association*, 87 :755–765.

Chapitre 1

Généralités sur la théorie des Sondages et le calcul de précision

Dans ce chapitre, nous rappelons les principaux outils et résultats de la théorie des sondages dite basée sur le plan, et nous introduisons des notations qui nous seront utiles dans la suite. Nous avons choisi de ne présenter dans ce chapitre que le socle minimal nécessaire, à notre sens, à la compréhension de l'ensemble. Afin que les autres chapitres puissent s'appréhender de la façon la plus indépendante possible, une présentation plus détaillée de différents plans de sondage, pour lesquels on recherche une stratégie Bootstrap d'estimation de la précision, sera donnée dans les sections correspondantes.

Pour une présentation détaillée de la théorie des Sondages dite basée sur le plan et des principaux résultats obtenus dans le cadre d'une population finie, on pourra consulter les ouvrages de référence que sont les livres de Särndal et al. (1992), Tillé (2001) et Ardilly (2006). Pour une approche basée sur le modèle, on pourra se référer à Valliant et al. (2000).

1.1 Généralités sur la théorie des Sondages

1.1.1 Population finie et variable d'intérêt

On appelle **population** l'ensemble des éléments que l'on étudie. On notera généralement U cette population : les éléments qui la composent sont appelés unités ou **individus**. Si ces individus sont en nombre fini, que l'on note alors

N, on dit que la population est finie. Notons que même si la population est finie, N lui-même n'est pas nécessairement connu et peut faire l'objet d'une inférence.

On suppose que chaque individu peut être entièrement caractérisé par un identifiant ou un numéro d'ordre. Par abus de langage, on notera alors

$$U = \{u_1, \ldots, u_N\} \equiv \{1, \ldots, N\}$$

en assimilant un individu u_k avec son label k. On dit que les unités sont **identifiables** (Cassel et al., 1976).

On note y une variable (éventuellement vectorielle) qui peut être mesurée sur chacun des individus de U, mais dont la valeur sur chacun de ces individus est inconnue. y est appelée **variable d'intérêt** ; la valeur prise par y sur le $k^{\text{ième}}$ individu est notée y_k. L'objet d'un sondage est d'estimer une fonction de la variable d'intérêt

$$\theta = \theta\left(y_k, k \in U\right)$$

que l'on appellera **paramètre d'intérêt** ou **fonctionnelle**, et d'évaluer la précision de cette estimation (généralement sous forme d'une variance ou d'un intervalle de confiance). Parmi les fonctionnelles, on trouve le total de la variable y sur la population U

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

la moyenne

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

la variance

$$\sigma_y^2 = \frac{1}{N} \sum_{k \in U} \left(y_k - \mu_y\right)^2,$$

et la dispersion (ou variance corrigée)

$$S_y^2 = \frac{1}{N-1} \sum_{k \in U} \left(y_k - \mu_y\right)^2 = \frac{N}{N-1} \sigma_y^2.$$

On peut également s'intéresser à des fonctionnelles plus complexes, telles que le ratio des totaux de deux variables y et z

$$R = \frac{t_y}{t_z},$$

le coefficient de corrélation entre les variables y et z

$$\rho = \frac{\sum_{k\in U}(y_k - \mu_y)(z_k - \mu_z)}{\sqrt{\sum_{k\in U}(y_k - \mu_y)^2 \sum_{k\in U}(z_k - \mu_z)^2}},$$

ou le fractile t_α d'ordre α de la variable y

$$t_\alpha = Inf\{x\ ; F(x) \geq \alpha\} \quad \text{avec} \quad F(x) = \frac{1}{N}\sum_{k\in U} \delta_{y_k \leq x},$$

δ_u désignant la mesure de Dirac au point u.

1.1.2 Plan de sondage

Si l'on enquêtait l'ensemble de la population (on parle alors de **recensement**) sur les variables d'intérêt, il serait théoriquement possible d'obtenir des estimations exactes pour les différentes fonctionnelles. On se heurte cependant à des difficultés pratiques :

→ Il n'est pas envisageable de recourir systématiquement à un recensement de la population. Ce type d'opération est coûteux et nécessite un personnel important (préparation de l'enquête ; collecte de l'information ; saisie, redressement et apurement des questionnaires ; exploitation des données).

→ Le volume d'information recueilli (on parle encore de fardeau de réponse) a des conséquences sur le coût de l'enquête, mais aussi sur les taux de réponse.

→ Pour fournir des statistiques récentes, il est nécessaire d'avoir des échantillons de taille limitée (dans le recensement français de 1999, exploitation au quart pour la plupart des estimations, exploitation au vingtième pour des estimations rapides).

→ Même si on enquêtait (théoriquement) exhaustivement la population, il subsiste toujours des problèmes de non-réponse (de l'ordre de 10 à 20 % dans les enquêtes obligatoires avec relance, jusqu'à 90 % et plus avec les échantillons de volontaires).

On se contente donc généralement d'enquêter une partie des individus de la population, appelée **échantillon**. Nous supposerons ici que l'échantillon est sélectionné au moyen d'un plan de sondage p sans remise, i.e d'une loi de probabilité sur l'ensemble des parties de U. p vérifie donc

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

On notera $E_p(.)$ (respectivement $V_p(.)$) l'espérance (respectivement la variance) sous le plan de sondage p, ou encore plus simplement $E(.)$ et $V(.)$ s'il n'y a pas de confusion possible. Pour un estimateur $\widehat{\theta}(S)$, on a

$$\begin{array}{rcl} E\left(\widehat{\theta}(S)\right) & = & \sum_{s \subset U} p(s)\widehat{\theta}(s) \\ V\left(\widehat{\theta}(S)\right) & = & \sum_{s \subset U} p(s)\left(\widehat{\theta}(s) - E\left(\widehat{\theta}(S)\right)\right)^2 \end{array}$$

Définition 1.1. *On appelle* ***support*** *du plan de sondage p, et on note $\mathcal{S}(p)$, l'ensemble des échantillons ayant une probabilité non nulle d'être sélectionnés :*

$$\mathcal{S}(p) = \{s \subset U; p(s) > 0\}$$

Définition 1.2. *Un plan de sondage p est dit* ***de taille fixe*** *égale à n si son support est inclus dans l'ensemble des échantillons de taille n, autrement dit si seuls les échantillons de taille n ont une probabilité non nulle d'être sélectionnés.*

On note S l'échantillon aléatoire. Sa taille, qui peut être également aléatoire, sera notée n_S. Dans le cas particulier où le plan de sondage est de taille fixe, on notera simplement n la taille de l'échantillon. Dans la mesure du possible, on utilise dans les enquêtes des plans de taille fixe, ce qui évite d'ajouter à la variance (incompressible) liée à l'échantillonnage, un alea supplémentaire dû à l'incertitude sur la taille de l'échantillon. Cependant, certains plans de sondage ne permettent pas d'obtenir une taille fixe d'échantillon (notamment le plan de Poisson et le tirage par grappes).

1.1.3 Estimation de Horvitz-Thompson

Probabilités d'inclusion

En suivant les notations de Särndal et al. (1992), nous notons $I_k = \delta_{k \in S}$ l'indicatrice d'appartenance à l'échantillon pour l'unité k de U.

Pour un plan de sondage p fixé, on appellera probabilité d'inclusion d'ordre 1 de l'unité k la probabilité π_k qu'a cette unité d'être retenue dans l'échantillon. Cette probabilité dépend du plan de sondage :

$$\pi_k = \sum_{s \subset U / k \in s} p(s).$$

On appelle probabilité d'inclusion d'ordre 2 la probabilité que deux unités distinctes k et l soient retenues conjointement dans l'échantillon :

$$\pi_{kl} = \sum_{s \subset U / k,l \in s} p(s).$$

Propriété 1.1. *Soit p un plan de sondage, $(\pi_k)_{k \in U}$ (respectivement $(\pi_{kl})_{k,l \in U}$) les probabilités d'inclusion d'ordre 1 (respectivement d'ordre 2) associées. Alors pour toutes les unités $k, l \in U$, les variables indicatrices I_k et I_l vérifient les propriétés suivantes :*

- $E(I_k) = \pi_k$
- $V(I_k) = \pi_k(1 - \pi_k)$
- $Cov(I_k, I_l) = \pi_{kl} - \pi_k \pi_l$

Démonstration. Voir Tillé (2001), page 31. □

Propriété 1.2. *Soit p un plan de taille fixe égale à n, $(\pi_k)_{k \in U}$ (respectivement $(\pi_{kl})_{k,l \in U}$) les probabilités d'inclusion d'ordre 1 (respectivement d'ordre 2) associées. Alors :*

- $\sum_{k \in U} \pi_k = n$
- $\forall l \in U \ \sum_{k \in U / k \neq l} \pi_{kl} = \pi_l (n - 1)$
- $\forall l \in U \ \sum_{k \in U} (\pi_{kl} - \pi_k \pi_l) = 0$

Démonstration. Cette propriété est une conséquence de la précédente, en utilisant le fait que, comme le tirage est de taille fixe, $\sum_{k \in U} I_k = n$. Voir Tillé (2001), page 32. □

Choix des probabilités d'inclusion et π-estimation

Lorsque l'on réalise une enquête, il est théoriquement possible de calculer la probabilité de sélection d'une partie quelconque de U, éventuellement à un facteur près ; en revanche, il n'est pas envisageable d'utiliser la distribution globale de probabilité. On impose généralement au minimum que le plan de sondage respecte des probabilités d'inclusion d'ordre 1 préalablement fixées (généralement, proportionnellement à une variable auxiliaire). Le respect de ces probabilités d'inclusion est primordial, car le théorême suivant assure que leur connaissance permet d'estimer sans biais une fonctionnelle linéaire.

Théorème 1.3. *Horvitz and Thompson (1952)*

Si pour toute unité k de U on a $\pi_k > 0$, alors

$$\widehat{t_{y\pi}} = \sum_{k \in S} \frac{y_k}{\pi_k}$$

est un estimateur sans biais de t_y.

Démonstration.

$$E\left(\widehat{t_{y\pi}}\right) = E\left(\sum_{k \in U} \frac{y_k}{\pi_k} I_k\right) = \sum_{k \in U} \frac{y_k}{\pi_k} E(I_k) = t_y$$

□

$\widehat{t_{y\pi}}$ est appelé **estimateur de Horvitz-Thompson** (ou **π-estimateur**) du total t_y. On utilise également la dénomination d'estimateur par les valeurs dilatées, car il s'agit d'un estimateur pondéré qui affecte un poids $d_k = 1/\pi_k$ à chaque unité k de l'échantillon. On dit encore que l'unité k de l'échantillon représente $1/\pi_k$ unités de la population dans l'estimation du total.

Nous supposerons dans la suite que la condition

$$\forall k \in U \quad \pi_k > 0$$

est toujours vérifiée. On peut toujours se ramener à ce cas en considérant que les éventuels individus tels que $\pi_k = 0$ constituent une strate à part, non enquêtée (on parle alors d'échantillonnage de type cut-off).

Remarque 1.1. *L'estimateur de Horvitz-Thompson est le seul estimateur linéaire sans biais d'un total de la forme*

$$\sum_{k\in U} w_k 1_{k\in S} y_k$$

*où les w_k soient **indépendants** de l'échantillon. Ici $w_k = 1/\pi_k$ est égal à l'inverse de la probabilité d'inclusion, c'est à dire au poids d'échantillonnage qui est disponible dans la base de sondage.*

1.1.4 Calcul et estimation de variance

Nous commençons ce paragraphe par un théorême, donnant la forme générale de variance pour un π-estimateur de total.

Théorème 1.4. *Horvitz and Thompson (1952)*

Le π-estimateur de total $\widehat{t_{y\pi}}$ a pour variance

$$V(\widehat{t_{y\pi}}) = \sum_{k\in U}\sum_{l\in U} \frac{y_k}{\pi_k}\frac{y_l}{\pi_l}(\pi_{kl} - \pi_k\pi_l)$$

Cette variance peut être estimée sans biais par

$$\widehat{V}_1(\widehat{t_{y\pi}}) = \sum_{k\in S}\left(\frac{y_k}{\pi_k}\right)^2 (1-\pi_k) + \sum_{k\in S}\sum_{l\neq k\in S} \frac{y_k}{\pi_k}\frac{y_l}{\pi_l}\frac{\pi_{kl}-\pi_k\pi_l}{\pi_{kl}}$$

si et seulement si $\pi_{kl} > 0\ \forall k, l \in U$. Cet estimateur est appelé estimateur de variance de Horvitz-Thompson.

Démonstration. La démonstration du premier point est une conséquence de la propriété 1.1. Le second point est une conséquence du résultat suivant : soit $g(.,.)$ une fonction quelconque. Alors la fonctionnelle

$$\sum_{k\in U}\sum_{l\neq k\in U} g(y_k, y_l)$$

est estimée sans biais par

$$\sum_{k\in S}\sum_{l\neq k\in S} \frac{g(y_k, y_l)}{\pi_{kl}}$$

si et seulement si $\pi_{kl} > 0\ \forall k, l \in U$; voir Tillé (2001), théorême 3.9, page 36. □

Dans le cas d'un plan de taille fixe, Sen (1953) et Yates and Grundy (1953) ont montré que la variance admettait une forme particulière, qui conduit à un second estimateur de variance.

Théorème 1.5. *(Sen, 1953; Yates and Grundy, 1953)*

Si le plan est de taille fixe, le π-estimateur de total $\widehat{t_{y\pi}}$ a pour variance

$$V(\widehat{t_{y\pi}}) = \sum_{k\in U} \sum_{l\neq k\in U} \left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l}\right)^2 (\pi_k\pi_l - \pi_{kl})$$

Cette variance peut être estimée sans biais par

$$\widehat{V}_2(\widehat{t_{y\pi}}) = \sum_{k\in S} \sum_{l\neq k\in S} \left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l}\right)^2 \frac{\pi_k\pi_l - \pi_{kl}}{\pi_{kl}}$$

si $\pi_{kl} > 0\ \forall k, l \in U$. Cet estimateur est appelé estimateur de variance de Sen-Yates-Grundy.

Démonstration. Le premier point est une conséquence de la proposition 1.2. Le second point découle là encore du théorême 3.9 de Tillé (2001).

□

Une condition suffisante pour que cet estimateur soit positif est que

$$\pi_{kl} \leq \pi_k\pi_l \qquad \forall k \neq l \in U \tag{1.1}$$

Les conditions 1.1 sont appelées **conditions de Sen-Yates-Grundy**.

1.1.5 Formules simplifiées de variance pour les plans à forte entropie

Les formules précédentes sont applicables de façon générale (et sous réserve d'un plan de taille fixe pour les formules de Sen-Yates-Grundy). En pratique, il est souvent difficile de les utiliser :

→ Ces formules utilisent les probabilités d'inclusion d'ordre 2. Or, en dehors de quelques plans de sondage (sondage aléatoire simple, voir chapitre 2 ; tirages poissonien et réjectif, voir chapitre 3), ces probabilités sont généralement très difficiles à calculer.

→ Même si les probabilités d'inclusion d'ordre 2 sont exactement calculables, ce calcul n'est pas forcément réalisable avec les seules données d'enquête. Par exemple, dans le cas du tirage réjectif, le calcul des probabilités d'inclusion doubles nécessite de connaître les probabilités d'inclusion d'ordre 1 pour l'ensemble des individus de la population.
→ Même si les probabilités d'inclusion d'ordre 2 sont calculables, certaines peuvent être nulles (notamment dans le cas d'un tirage systématique non randomisé), auquel cas les estimateurs de variance de Horvitz-Thompson et de Sen-Yates-Grundy sont biaisés.
→ Même si les probabilités d'inclusion d'ordre 2 sont strictement positives, ces deux estimateurs de variance peuvent se révéler instables (Matei and Tillé, 2005).

Compte-tenu de ces difficultés, il est fréquent d'utiliser des estimateurs simplifiés basés sur une approximation de la variance. Bien que (faiblement) biaisés, ces estimateurs sont généralement stables et présentent une erreur quadratique moyenne plus faible que celle des estimateurs de Horvitz-Thompson et de Sen-Yates-Grundy.

Brewer and Donadio (2003) et Matei and Tillé (2005) donnent une revue très détaillée de différents estimateurs simplifiés de variance. Ces derniers recommandent notamment l'utilisation de l'estimateur de Deville (1993) :

$$\widehat{V}_{dev}(\widehat{t_{y\pi}}) = \sum_{k \in S} \frac{c_k}{\pi_k^2} \left(y_k - \hat{y_k}\right)^2$$

avec

$$\hat{y_k} = \pi_k \frac{\sum_{l \in S} c_l y_l / \pi_l}{\sum_{l \in S} c_l}$$

et

$$c_k = (1 - \pi_k) \frac{n}{n-1}$$

Cet estimateur de variance donne de bons résultats pour les plans proches de l'entropie maximale (voir chapitre 3) dès que la taille d'échantillon est supérieure à 6 (Deville, 1993).

1.2 L'asymptotique en théorie des Sondages

Pour déterminer la précision de l'estimateur $\hat{\theta}$ d'un paramètre θ, on utilise souvent la notion d'intervalle de confiance. On détermine (ou plus exacte-

ment, on estime) un intervalle, généralement centré sur l'estimateur $\hat{\theta}$, et contenant la vraie valeur du paramètre avec un niveau de confiance (on parle encore de **taux de couverture**) fixé. Produire un intervalle de confiance suppose de connaître, même de façon approchée, la loi asymptotique de l'estimateur $\hat{\theta}$.

1.2.1 Le théorême central-limite

Nous commençons par rappeler une forme de base du théorême central-limite dans le cas d'une population infinie.

Théorème 1.6. *Soit $(X_n)_{n\in\mathbb{N}}$ une suite de variables aléatoire i.i.d. On suppose que $\mu = E(X_i)$ et $\sigma^2 = V(X_i)$ existent. Alors $\overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i$ suit asymptotiquement une loi normale.*

Démonstration. Voir par exemple Feller (1966). □

Pour une suite de variables aléatoires indépendantes, il est donc possible de connaître la loi asymptotique de la suite des sommes partielles. Dans le cadre d'un sondage, si les unités sont sélectionnées indépendamment, il est possible d'invoquer un résultat de ce type pour établir la loi asymptotique du π-estimateur. Cependant, la sélection des unités dans l'échantillon se fait dans la plupart des cas de façon dépendante : pour que l'échantillonnage soit efficace, on s'interdit en particulier de sélectionner plusieurs fois un même individu dans l'échantillon (on montre par exemple que le sondage aléatoire simple sans remise est plus efficace que le sondage aléatoire simple avec remise). Dans ce cas, les résultats classiques ne sont pas applicables.

Dans le cadre d'une population finie, un théorême central limite a été démontré pour certains plans de sondage. Indépendamment, Erdös and Renyi (1959) et Hájek (1960) l'ont établi pour le sondage aléatoire simple sans remise. La population U est vue ici comme un élément d'une suite croissante de populations, notée (U_ν) ; la taille de la population N et la taille n de l'échantillon sont également extraites de deux suites croissantes (N_ν) et (n_ν), donnant respectivement la taille de la population U_ν et la taille de l'échantillon sélectionné dans U_ν. L'échantillon S_ν est extrait de U_ν avec un taux de sondage $f_\nu = n_\nu/N_\nu$. On note respectivement $\bar{y}_\nu$ et $\mu_{y\nu}$ la moyenne simple de la variable y_ν sur l'échantillon S_ν et sur la population U_ν. On note $S_{\nu y}$ la dispersion de la variable y_ν sur la population U_ν.

Théorème 1.7. *Hájek (1960)*

On suppose que $n_\nu \to \infty$*, et* $N_\nu - n_\nu \to \infty$ *quand* $\nu \to \infty$*. Alors, dans le cas d'un sondage aléatoire simple*

$$\sqrt{n_\nu}\frac{(\bar{y}_\nu - \mu_{y\nu})}{\sqrt{1-f_\nu}S_{\nu y}} \to_d \mathcal{N}(0,1) \quad \textit{quand } \nu \to \infty$$

si et seulement si la suite $(Y_{\nu j})$ *vérifie la condition de Lindeberg-Hájek*

$$lim_{\nu\to\infty} \sum_{T_\nu(\delta)} \frac{y_{\nu j} - \mu_{y\nu}}{(N_\nu - 1)S_{\nu y}^2} = 0 \quad \textit{quel que soit } \delta > 0$$

où $T_\nu(\delta)$ *désigne l'ensemble des unités de* U *pour lesquelles*

$$|y_{\nu j} - \mu_{y\nu}|/\sqrt{1-f_\nu}S_{\nu y} > \delta\sqrt{n_\nu}.$$

En ce qui concerne les plans de sondage à probabilités inégales, la normalité asymptotique a essentiellement été étudiée pour les tirages réjectif et successif. Le tirage réjectif peut être vu comme un tirage poissonien conditionnel à la taille (voir chapitre 3), mais également comme un tirage avec remise avec des probabilités α_i, conditionné par le fait que toutes les unités sélectionnées soient distinctes (Hájek, 1964). Le tirage successif (Rosen, 1972a) consiste à échantillonner les unités avec remise avec des probabilités α_i : si une unité est sélectionnée deux fois, le doublon est rejeté et le tirage continue jusqu'à obtenir n unités distinctes. Hájek (1964) a démontré la normalité asymptotique du π-estimateur de total dans le cas d'un tirage réjectif, et Rosen (1972a,b) obtient un résultat similaire pour le tirage successif. La normalité asymptotique de l'estimateur de la moyenne dans le cas d'un sondage aléatoire simple stratifié a été discutée par Bickel and Freedman (1984) et Krewski and Rao (1981). Dans le cas d'un tirage multi-degrés, Sen (1980, 1988) établit la normalité asymptotique d'un estimateur de type Horvitz-Thompson pour un échantillonnage successif des unités du premier degré. Isaki and Fuller (1982) proposent également un cadre asymptotique souvent utilisé comme référence, voir aussi Fuller and Isaki (1981). Une présentation plus détaillée de ces résultats est donnée dans Thompson (1997).

1.2.2 Développements d'Edgeworth

Le théorême central-limite établit qu'asymptotiquement, la fonction de répartition du π-estimateur est proche de celle d'une loi normale. C'est un

résultat au premier ordre, car il ne donne pas d'ordre de grandeur de l'erreur obtenue en utilisant cette approximation.

On peut obtenir des approximations à des ordres plus grands en utilisant les développements d'Edgeworth, dont on trouvera une présentation détaillée dans Hall (1992) et Thompson (1997). Soient $X_1, \ldots, X_n$ n variables aléatoires i.i.d., admettant des moments finis jusqu'à l'ordre 4. La moyenne est notée $\theta = \mu$ et la variance σ^2. On suppose que la moyenne standardisée $Z_n = n^{1/2}(\bar{X} - \theta)/\sigma$ admet une fonction de densité f. Alors

$$f(x) = \phi(x) \left\{ 1 + \frac{\kappa_3}{3!} h_3(x) + \frac{\kappa_4}{4!} h_4(x) + \frac{\kappa_3^2}{2(3!)^2} h_6(x) \right\} + o(n^{-1})$$

où $\phi(.)$ désigne la fonction de densité de la loi normale centrée réduite, κ_r le cumulant d'ordre r de la distribution de densité f et les $h_j(.)$ sont des polynômes d'Hermite, définis par

$$h_j(x) = (-1)^j \phi^{(j)}(x)/\phi(x)$$

avec $\phi^{(j)}(.)$ la dérivée $j^{\text{ème}}$ de $\phi(.)$. On a en particulier :

$$\begin{aligned}
h_1(x) &= x \\
h_2(x) &= x^2 - 1 \\
h_3(x) &= x^3 - 3x \\
h_4(x) &= x^4 - 6x^2 + 3 \\
h_5(x) &= x^5 - 10x^3 + 15x \\
h_6(x) &= x^6 - 15x^4 + 45x^2 - 15
\end{aligned}$$

On en déduit un développement analogue pour la fonction de répartition de Z_n :

$$\mathbb{P}(Z_n \leq x) = \Phi(x) - \phi(x) \left\{ \frac{\kappa_3}{3!} h_2(x) + \frac{\kappa_4}{4!} h_3(x) + \frac{\kappa_3^2}{2(3!)^2} h_5(x) \right\} + o(n^{-1})$$

Sous certaines hypothèses, on peut obtenir des expansions à des ordres supérieurs ; on peut également dériver des formules analogues pour des estimateurs plus complexes que la moyenne simple (Hall, 1992).

Dans le cadre d'une population finie, la dépendance entre les tirages des unités rend délicate la construction de tels développements. Pour le sondage

aléatoire simple, ce problème a été étudié par Robinson (1978), Babu and Singh (1985) et plus récemment par Bloznelis and Götze (2001, 2002). Voir également Sugden and Smith (1997) et Sugden et al. (2000). Sous une condition technique, qui est vérifiée en particulier si les valeurs de la variable d'intérêt y sur la population sont des réalisations i.i.d. d'une variable aléatoire ayant une densité et des moments finis jusqu'à l'ordre 5, Robinson (1978) (voir également Thompson (1997)) établit que la fonction de répartition de la moyenne standardisée

$$Z_n = \frac{\bar{y} - \bar{Y}}{\sqrt{(1-f)S_y^2/n}}$$

admet le développement

$$\mathbb{P}(Z_n \leq x) = \Phi(x) + \frac{p_1(x)\phi(x)}{n^{1/2}} + \frac{p_2(x)\phi(x)}{n} + O(n^{-3/2})$$

avec

$$p_1(x) = -\frac{\gamma_1}{6}\frac{1-2f}{\sqrt{1-f}}(x^2-1),$$

$$p_2(x) = \left\{-\gamma_2\frac{1-6f(1-f)}{24(1-f)} + \frac{f}{4}\right\}(x^3-3x) - \frac{\gamma_1^2}{72}\frac{(1-2f)^2}{\sqrt{1-f}}(x^5-10x^3+15x),$$

$$\gamma_1 = \frac{\mu^{(3)}}{\sigma_y^3},$$

$$\gamma_2 = \frac{\mu^{(4)}}{\sigma_y^4} - 3,$$

$$\sigma_y^2 = \frac{N-1}{N}S_y^2 = \frac{1}{N}\sum_{k\in U}(y_k - \bar{Y})^2,$$

$$\mu^{(r)} = \frac{1}{N}\sum_{k\in U}(y_k - \bar{Y})^r.$$

Un développement analogue pour la fonction de répartition de la moyenne studentisée

$$U_n = \frac{\bar{y} - \bar{Y}}{\sqrt{(1-f)s_y^2/n}}$$

est donné par exemple dans Sugden et al. (2000) :

$$\mathbb{P}(U_n \leq x) = \Phi(x) + \frac{q_1(x)\phi(x)}{n^{1/2}} + \frac{q_2(x)\phi(x)}{n} + O(n^{-3/2})$$

avec

$$q_1(x) = \gamma_1 \left\{ \frac{1}{2}\sqrt{1-f} + \frac{1}{3}\,\frac{1-f/2}{\sqrt{1-f}}(x^2-1) \right\},$$

$$\begin{aligned} q_2(x) \;=\; & x\left[\gamma_2 \left\{ \tfrac{2-6f+3f^2}{24(1-f)}(x^2-3) - \tfrac{1}{2}f \right\} - \left\{1 + \tfrac{1}{4}(x^2-3)\right\}\right. \\ & \left. -\; \gamma_1^2 \left\{ 1 - f + \tfrac{2-f}{3}(x^2-3) + \tfrac{(1-f/2)^2}{18(1-f)}(x^4 - 10x^2 + 15) \right\} \right]. \end{aligned}$$

Les conditions précises de validité sont discutées dans Thompson (1997).

1.2.3 Discussion

A l'heure actuelle, la normalité asymptotique n'est strictement établie que pour un nombre limité de plans de sondages, évoqués ci-dessus, et des développements plus fins de type Edgeworth ne sont disponibles que pour le sondage aléatoire simple. D'autre part, dans une situation pratique d'enquête, le sondeur est souvent confronté à des mécanismes aléatoires non maîtrisés (tels que la non-réponse partielle ou totale de certains individus enquêtés) dont l'influence sur le comportement asymptotique des estimateurs est difficile à prendre en compte.

Cela signifie t-il qu'il faut renoncer à définir des intervalles de confiance ? Evidemment non. Même si elle n'est pas totalement validée sur le plan théorique, de nombreuses études empiriques par simulations ont démontré le bien-fondé de l'hypothèse de normalité dans des situations réalistes, voir par exemple Särndal et al. (1992) page 276. Une pratique courante consiste à utiliser le cadre asymptotique de Fuller and Isaki (1981) et Isaki and Fuller (1982), où les tailles de la population et de l'échantillon tendent vers l'infini (pour simplifier les notations, on omettra l'indice ν), en supposant que pour toute variable vectorielle $\mathbf{x}$ les hypothèses suivantes sont vérifiées :

H1 : $N^{-1}n \to f \in \,]0,1[$
H2 : $N^{-1}t_{\mathbf{x}}$ a une limite finie
H3 : $N^{-1}(\widehat{t_{\mathbf{x}\pi}} - t_{\mathbf{x}}) \to 0$ en probabilité
H4 : $n^{1/2}N^{-1}(\widehat{t_{\mathbf{x}\pi}} - t_{\mathbf{x}}) \to \mathcal{N}(0,\Sigma)$ en loi

Comme nous venons de le souligner, il est difficile de disposer de résultats asymptotiques, même au premier ordre, dans un cadre réaliste. Le comportement d'une méthode de Bootstrap au second ordre, pour un plan de sondage de référence tel que le sondage aléatoire simple, est une propriété appréciable mais non suffisante. Il nous semble qu'une bonne méthode d'estimation de précision devrait fournir une estimation consistante de variance pour une gamme étendue de plans de sondage, mettant en jeu des techniques usuelles telles que l'échantillonnage à probabilités inégales, l'échantillonnage équilibré, l'échantillonnage multidegrés, l'imputation ou les méthodes de calage.

1.3 Méthodes de calcul de précision

Les formules présentées en 1.1.4 ne permettent d'estimer effectivement la variance que dans le cas d'une fonctionnelle linéaire, estimée à l'aide des poids de sondage. Or, on peut s'intéresser à des fonctionnelles non linéaires, de type ratios ou corrélations, voire à des fonctionnelles plus complexes de type indices, par exemple dans des études économiques sur la pauvreté. D'autre part, l'estimateur naturel de Horvitz-Thompson est généralement redressé, pour améliorer la précision et/ou tenir compte de la non-réponse.

Nous présentons dans cette section quelques techniques classiques permettant de calculer la précision d'un estimateur complexe. Une présentation synthétique est donnée dans Wolter (2007). Voir également Deville (1987), Kovar et al. (1988), Rao et al. (1992), Shao and Tu (1995) et Davison and Sardy (2007). Dans la suite de ce texte, nous dirons qu'un estimateur de variance est **consistant** s'il restitue la vraie variance asymptotiquement sans biais.

1.3.1 La technique de linéarisation

La linéarisation est probablement la technique la plus générale actuellement disponible pour l'estimation de variance. Elle est utilisable pour une gamme très étendue de statistiques (voir ce qui suit). Elle ne nécessite que de disposer d'une estimation analytique de variance pour un estimateur de total, ce qui est accessible avec de nombreux logiciels (tels le logiciel POULPE utilisé à l'Insee) y compris pour des plans de sondage très complexes. La prise en compte d'un plan de sondage complexe est plus délicate avec les méthodes réplicatives (Jackknife, demi-échantillons équilibrés ou Bootstrap).

La linéarisation de Taylor est applicable lorsque le paramètre θ à estimer peut être exprimé comme une fonction explicite de totaux

$$\theta = f(t_{y_1}, \ldots, t_{y_p})$$

où f est supposée dérivable. Supposons que l'on estime θ à l'aide de son estimateur par substitution

$$\widehat{\theta} = f(\widehat{t_{y_1\pi}}, \ldots, \widehat{t_{y_p\pi}})$$

où chaque total est remplacé par son π-estimateur. La méthode consiste à approcher $\widehat{\theta}$ par un estimateur linéaire $\widehat{\theta_0}$ obtenu à l'aide de la linéarisation de Taylor de la fonction f au point $(t_{y_1}, \ldots, t_{y_p})$. Plus précisément :

$$\widehat{\theta} \simeq \widehat{\theta_0} = \theta + \sum_{i=1}^{p} a_i(\widehat{t_{y_i\pi}} - t_{y_i})$$

avec $a_i = \frac{\partial f}{\partial u_i}|_{(u_1,\ldots,u_p)=(t_{y_1},\ldots,t_{y_p})}$. La variance de $\widehat{\theta}$ est approchée par celle de $\widehat{\theta_0}$.

Propriété 1.8. *Särndal et al. (1992)*

Avec les notations précédentes, la variance approchée de $\widehat{\theta}$ est égale à

$$V_{app}(\widehat{\theta}) = \sum_{k,l \in U} (\pi_{kl} - \pi_k \pi_l) \frac{u_k}{\pi_k} \frac{u_l}{\pi_l}$$

avec $u_k = \sum_{i=1}^{p} a_i y_{ki}$, et un estimateur consistant de variance est donné par

$$\widehat{V}(\widehat{\theta}) = \sum_{k,l \in S} \frac{\pi_{kl} - \pi_k \pi_l}{\pi_{kl}} \frac{\hat{u_k}}{\pi_k} \frac{\hat{u_l}}{\pi_l}$$

avec $\hat{u_k} = \sum_{i=1}^{p} \hat{a}_i y_{ki}$, $\hat{a}_i$ s'obtenant de a_i en remplaçant chaque total par le π-estimateur correspondant.

Démonstration. Voir par exemple Krewski and Rao (1981). □

Il existe des méthodes permettant de mettre à profit la linéarisation de Taylor pour produire des intervalles de confiance de fonctionnelles fortement non linéaires de type médiane (Woodruff, 1952; Kuk, 1988) ; mais cette technique n'est utilisable pour l'estimation de variance que dans le cas d'une fonction explicite de totaux.

Pour résoudre ce problème, Deville (1999) a proposé une approche plus générale basée sur la fonction d'influence ; voir également Goga et al. (2007). Supposons que dans la population U, un échantillon S ait été sélectionné selon un plan de sondage p quelconque, affectant la probabilité d'inclusion π_k à l'individu k de U. On note M la mesure qui place une masse unité sur l'unité k de U. De nombreux paramètres peuvent êre vus comme des fonctionnelles de M, par exemple un total

$$t_y = \int y dM,$$

un ratio

$$t_y/t_x = \int y dM / \int x dM,$$

le fractile t_α d'ordre α

$$t_\alpha = Inf\{x\ ; F(x) \geq \alpha\} \quad \text{avec} \quad F(x) = \frac{1}{N}\sum_{k\in U} \delta_{y_k \leq x}$$

ou un indice de Gini (hors constante)

$$G = \frac{\int y T_y dM}{N \int y dM} \quad \text{avec} \quad T_z = \int \delta_{. \leq z} dM.$$

Définition 1.3. *Deville (1999)*

La fonction d'influence d'une fonctionnelle $\theta(M)$ au point k est définie (sous réserve d'existence) par

$$I\theta_k(M) = \lim_{\epsilon \to 0} \frac{\theta(M + \epsilon\delta_k) - \theta(M)}{\epsilon}$$

Cette définition diffère de la définition classique de la fonction d'influence (Hampel et al., 1985), en raison de la masse totale de la mesure M, souvent

inconnue. Selon le principe de Horvitz-Thompson, la mesure M est estimée par sa "mesure par substitution" $\hat{M}$, qui place une masse $1/\pi_k$ sur chaque unité k de l'échantillon S. La fonctionnelle $\theta(M)$ est estimée par $\theta(\hat{M}) = \hat{\theta}$.

Définition 1.4. *Deville (1999)*

La variable linéarisée de la fonctionnelle $\theta(M)$, notée u_k, est égale à la fonction d'influence de $\theta(M)$ au point k :

$$u_k = I\theta_k(M)$$

Le résultat suivant donne une approximation de variance pour l'estimateur $\theta(\hat{M})$. Nous nous plaçons dans le cadre asymptotique introduit en 1.2.3, et nous supposerons de plus que

H5 : θ est homogène de degré β, i.e. quel que soit $r > 0$: $\theta(rM) = r^\beta \theta(M)$

H6 : $\lim_{N\to\infty} N^{-\beta}\theta(M) < \infty$

H7 : $\theta(M)$ est Fréchet différentiable

Théorème 1.9. *Deville (1999)*

Sous les hypothèses $H1 - H7$, l'estimateur par substitution $\theta(\hat{M})$ est linéarisable et

$$\begin{aligned} n^{1/2}N^{-\beta}\left(\theta(\hat{M}) - \theta(M)\right) &= n^{1/2}N^{-\beta}\int I\theta_k(M)d(\hat{M} - M) + o_p(1) \\ &= n^{1/2}N^{-\beta}\textstyle\sum_{k\in U} u_k(1/\pi_k - 1) + o_p(1) \end{aligned}$$

avec $u_k = I\theta_k(M)$.

On peut donc approcher $V(\hat{\theta})$ par

$$V_{app}(\widehat{\theta}) = \sum_{k,l\in U}(\pi_{kl} - \pi_k\pi_l)\frac{u_k}{\pi_k}\frac{u_l}{\pi_l}$$

et un estimateur consistant de variance est donné par

$$\widehat{V}_{lin}(\widehat{\theta}) = \sum_{k,l\in S}\frac{\pi_{kl} - \pi_k\pi_l}{\pi_{kl}}\frac{\hat{u}_k}{\pi_k}\frac{\hat{u}_l}{\pi_l}$$

où $\hat{u}_k$ s'obtient en remplaçant dans u_k chaque fonctionnelle inconnue par son estimateur par substitution.

En pratique, une statistique sera généralement homogène de degré 1 (cas d'un total), de degré 0 (cas d'un ratio et d'un paramètre sans dimension), et parfois de degré 2.

1.3.2 Le Jackknife

Le Jackknife a été introduit par Quenouille (1949a,b) pour l'estimation du biais d'une statistique. Tukey (1958) l'a proposé comme une méthode générale d'estimation de variance. L'idée consiste à supprimer dans l'échantillon d'origine des groupes d'individus, et à recalculer la statistique d'intérêt sur les individus restants. La dispersion des statistiques Jackknife est alors utilisée comme estimateur de variance. On trouvera une présentation synthétique des méthodes de Jackknife adaptées au cas d'une population finie dans Davison and Sardy (2007), et plus détaillée dans Shao and Tu (1995).

Supposons que l'on dispose d'un échantillon $X_1, \ldots, X_n$ i.i.d.. On s'intéresse à la précision d'un estimateur $\widehat{\theta} = g(\overline{X_n})$ d'un paramètre θ, où $\overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i$. Le Jackknife simple consiste à supprimer à tour de rôle chaque individu de l'échantillon. On obtient ainsi n statistiques Jackknife $\widehat{\theta}_1, \ldots, \widehat{\theta}_n$, avec

$$\widehat{\theta}_i = g(\overline{X_{n,i}}) \text{ et } \overline{X_{n,i}} = \frac{n\overline{X_n} - X_i}{n-1}.$$

Le biais est estimé par

$$(n-1)(\overline{\widehat{\theta}_n} - \widehat{\theta}) \text{ avec } \overline{\widehat{\theta}_n} = \frac{1}{n}\sum_{i=1}^{n} \widehat{\theta}_i$$

et la variance par

$$\frac{n-1}{n}\sum_{i=1}^{n}\left(\widehat{\theta}_i - \overline{\widehat{\theta}_n}\right)^2.$$

Ces estimateurs sont consistants si la fonction g est continument différentiable (Shao and Tu, 1995, théorême 2.1). De façon plus générale, on montre que le Jackknife simple fournit une estimation consistante de variance si la statistique est suffisamment lisse. En revanche, il est inconsistant pour l'estimation de précision de quantiles ou d'estimateurs basés sur les quantiles (Efron, 1982). Pour remédier à ce problème, Wu (1986, 1990) propose une généralisation appelée le delete-d Jackknife. La statistique Jackknife est obtenue en supprimant non plus une observation, mais un échantillon de taille

d. Shao and Wu (1989) montrent que le delete-d Jackkine est consistant pour l'estimation de précision de quantiles si $d \to \infty$ quand $n \to \infty$.

L'utilisation du Jackknife est complexe dans le cadre d'une population finie. Dans le cas d'un sondage aléatoire simple (éventuellement stratifié), l'estimateur Jackknife de variance doit être ajusté car il ne capte pas la correction de population finie. Un estimateur Jackknife de variance, proche de l'estimateur obtenu par linéarisation, est proposé par Berger and Rao (2006) dans le cas d'un plan à probabilités inégales. Dans le cas d'un tirage multidegrés, le Jackknife simple conduit à supprimer tour à tour chaque unité primaire (Rao and Wu, 1985) ; le nombre de recalculs peut être important si la taille d'échantillon est grande. Voir également Rao and Tausi (2004) pour une approche par les fonctions estimantes. Pour obtenir une estimation consistante pour des quantiles, on peut utiliser le delete-d Jackknife (d devant être d'autant plus grand que la statistique est moins lisse), mais le fardeau de calcul est encore plus lourd. Une méthode consiste à scinder la population en blocs de d unités, et de supprimer tour à tour chacun de ces blocs (Krewski and Rao, 1981; Shao and Tu, 1995).

1.3.3 Les demi-échantillons équilibrés

La méthode des demi-échantillons équilibrés (Mac Carthy, 1969) a été à l'origine développée dans le cas d'un plan de sondage stratifié, avec tirage de deux unités primaires avec remise dans chaque strate. Dans ce cas, un demi-échantillon est obtenu en prélevant un individu dans chaque strate dans l'échantillon de départ. On recalcule ainsi 2^H demi-échantillons, où H désigne le nombre de strates.

Pour limiter la charge de calcul, on peut prélever un nombre restreint de demi-échantillons choisis de façon équilibrée, i.e. tels que

$$\sum_{r=1}^{R} \alpha_{hr}\alpha_{kr} = 0 \text{ pour tous } h \neq k = 1 \dots H$$

où R désigne le nombre de demi-échantillons, α_{hr} vaut 1 si la première unité de la strate h est sélectionnée dans le demi-échantillon r et -1 sinon. Autrement dit, les unités de chaque strate se retrouvent le même nombre de fois

dans les R demi-échantillons. L'estimateur de variance est alors

$$R^{-1} \sum_{r=1}^{R} (\widehat{\theta}_r - \widehat{\theta})^2$$

où $\widehat{\theta}_r$ désigne la statistique $\widehat{\theta}$ calculée à partir du demi-échantillon r. Cette méthode donne une estimation de variance consistante, y compris pour des estimateurs basés sur des quantiles (Krewski and Rao, 1981; Shao and Wu, 1992; Shao and Rao, 1994). Les demi-échantillons équilibrés peuvent être construits à l'aide de matrices de Hadamard (Wolter, 1985).

Comme pour le Jackknife, la méthode des demi-échantillons équilibrés ne permet pas de capter la correction de population finie dans le cas d'un sondage aléatoire simple sans remise ; une correction, analogue à celle proposée par Rao and Wu (1988) pour le Bootstrap, a été suggérée par Wu (1991). Des adaptations sont également disponibles dans le cas général où l'on tire plus de deux individus dans chaque strate (Rao and Shao, 1996), mais la mise en oeuvre peut être délicate (Davison and Sardy, 2007).

1.3.4 Le Bootstrap

Le Bootstrap est sans doute la méthode d'estimation de précision par réplication la plus générale. Elle a été initialement proposée par Efron (1979) dans le cadre d'une population infinie. Nous proposons ici un bref aperçu de la méthode de Bootstrap de base. Nous détaillerons dans les chapitres suivants les différentes adaptations du Bootstrap au cas d'un sondage en population finie.

Depuis l'article fondateur d'Efron, de nombreux ouvrages ont été consacrés au Bootstrap ; Hall (1992), Efron and Tibshirani (1993), Shao and Tu (1995) et Davison and Hinkley (1997) sont parmi les principales références. Ces deux derniers ouvrages présentent une revue des principales méthodes de Bootstrap en population finie. Sur ce sujet, voir également Deville (1987), Presnell and Booth (1994), Nigam and Rao (1996) et Lahiri (2003).

Principe de la méthode

Soit un échantillon $X_1, \ldots, X_n$ i.i.d. et distribué selon une loi F inconnue. Supposons que l'on souhaite estimer un paramètre $\theta(F)$. Alors, selon un

principe de plug-in, $\theta(F)$ est estimé par $\theta(\hat{F})$, obtenu en remplaçant F par la fonction de répartition empirique $\hat{F}$ calculée sur l'échantillon. Si $\theta(\hat{F})$ ne peut être calculé directement, on peut l'approcher par simulations : soient $(X_{1b}^*, \ldots, X_{nb}^*)_{b=1\ldots B}$ B échantillons sélectionnés indépendamment selon la loi $\hat{F}$ conditionnellement à l'échantillon de départ $(X_1, \ldots, X_n)$. On obtient ainsi une approximation de Monte-Carlo

$$\frac{1}{B}\sum_{b=1}^{B}\theta(\hat{F}^{*b})$$

pour $\theta(\hat{F})$, où $\hat{F}^{*b}$ désigne la fonction de répartition empirique calculée sur le rééchantillon $(X_{1b}^*, \ldots, X_{nb}^*)$. On peut ainsi fournir une estimation Bootstrap du biais d'un estimateur, ou de sa variance. Notons que par rapport au Jackknife de base, le Bootstrap présente l'avantage d'être consistant, sous de faibles hypothèses, pour l'estimation de précision d'un paramètre non lisse de type fractile.

Intervalle de confiance

Le Bootstrap permet également de donner une estimation de la fonction de répartition dans son ensemble. Intéressons-nous à la fonction de répartition d'une variable aléatoire $R(X_1, \ldots, X_n, F)$, que l'on note

$$H_F(x) = \mathbb{P}(R(X_1, \ldots, X_n, F) \leq x).$$

En suivant le principe de plug-in, $H_F(x)$ est estimé par

$$H_{\hat{F}}(x) = \mathbb{P}(R(X_1^*, \ldots, X_n^*, \hat{F}) \leq x | (X_1, \ldots, X_n))$$

où $(X_1^*, \ldots, X_n^*)$ est un échantillon i.i.d. tiré selon la loi $\hat{F}$. Là encore, $H_{\hat{F}}(x)$ peut être approché via des méthodes de Monte-Carlo par

$$\frac{1}{B}\sum_{i=1}^{B} I(R(X_{1i}^*, \ldots, X_{ni}^*, \hat{F}) \leq x)$$

où les échantillons $(X_{1b}^*, \ldots, X_{nb}^*)$ sont sélectionnés indépendamment selon la loi $\hat{F}$. On peut ainsi produire des intervalles de confiance Bootstrap pour une statistique. Dans la suite, nous nous intéresserons principalement à deux techniques largement utilisées : la méthode des percentiles (Efron, 1981), et

la méthode du t-Bootstrap (Efron, 1982).

La méthode des percentiles utilise directement l'estimation de la fonction de répartition donnée par le Bootstrap. Soit $\theta(F)$ un paramètre estimé par $\theta(\hat{F})$, $\theta(\hat{F}^*)$ son équivalent Bootstrap calculé sur un rééchantillon $(X_1^*, \ldots, X_n^*)$ et

$$G_{boot}(x) = \mathbb{P}(\theta(\hat{F}^*) \leq x | X_1, \ldots, X_n).$$

Alors on utilise comme intervalle de confiance de niveau $1 - 2\alpha$ pour θ, l'intervalle

$$\left[G_{boot}^{-1}(\alpha), G_{boot}^{-1}(1-\alpha)\right].$$

Là encore, $G_{boot}(.)$ peut être approchée à l'aide de simulations : pratiquement, on sélectionne B rééchantillons sur chacun desquels la statistique Bootstrap $\theta(\hat{F}^*)$ est calculée. Ces estimateurs Bootstrap sont ordonnés par ordre croissant, et l'intervalle de niveau $1 - 2\alpha$ est obtenu en supprimant les B α statistiques Bootstrap les plus faibles et les B α statistiques Bootstrap les plus fortes.
Cette méthode est simple et intuitive, mais Shao and Tu (1995) soulignent qu'à moins que la taille d'échantillon ne soit grande, elle peut être nettement plus imprécise que des techniques telles que le t-Bootstrap. Des variantes ont été proposées, telles que le Bootstrap bias-corrected percentile (Efron, 1981; Schenker, 1985), ou le Bootstrap accelerated bias-corrected percentile (Efron, 1987; DiCiccio and Tibshirani, 1987; Konishi, 1991). L'examen de ces variantes dans le cas d'une population finie ne sera pas discuté ici, et nous nous restreindrons à la méthode des percentiles de base qui a tout de même l'avantage de fournir un bon compromis entre précision et fardeau de rééchantillonnage.

La méthode du t-Bootstrap passe par une estimation de la statistique studentisée

$$T(F, \hat{F}) = (\theta(\hat{F}) - \theta)/\sqrt{v(\theta(\hat{F}))},$$

où $v(\theta(\hat{F}))$ désigne un estimateur de variance de $\theta(\hat{F})$. La version Bootstrap de la statistique studentisée est

$$T(\hat{F}, \hat{F}^*) = (\theta(\hat{F}^*) - \theta(\hat{F}))/\sqrt{v^*(\theta(\hat{F}^*))}$$

où $v^*(\theta(\hat{F}^*))$ désigne un estimateur de variance de $\theta(\hat{F}^*)$. On notera que, en vertu du principe de plug-in, l'estimateur de variance $v(\theta(\hat{F}))$ (associé

à $\theta(\hat{F})$) qui apparaît dans la statistique studentisée est remplacé dans la version Bootstrap de cette statistique studentisée par un autre estimateur de variance $v^*(\theta(\hat{F}^*))$ associé à $\theta(\hat{F}^*)$, et non par une approximation de la variance de $\theta(\hat{F}^*)$. Un intervalle de confiance t-Bootstrap de niveau $1-\alpha$ est donné par

$$\left[\theta(\hat{F}) - t_U^* \sqrt{v(\theta(\hat{F}))}, \theta(\hat{F}) - t_L^* \sqrt{v(\theta(\hat{F}))}\right]$$

où t_L^* et t_U^* désignent respectivement les fractiles d'ordre α et $1-\alpha$ de $T(\hat{F}, \hat{F}^*)$. Hall (1986, 1988) établit que le t-Bootstrap est exact au second ordre (c'est à dire qu'il permet d'avoir une approximation de la fonction de répartition de $\theta(\hat{F})$ qui restitue les deux premiers du développement d'Edgeworth de cette fonction de répartition) dans le cas où $\theta(F)$ est une fonction lisse de moyennes.

Le t-Bootstrap n'est en revanche pas exact au second ordre pour des paramètres non lisses de type fractile, et peut dans ce cas se révéler moins précis que la méthode des percentiles. D'autre part, s'il n'est pas possible de produire un estimateur direct $v^*(\theta(\hat{F}^*))$, le t-Bootstrap nécessite d'itérer le Bootstrap, c'est à dire pour chaque rééchantillon $(X_1^*, \ldots, X_n^*)$ de rééchantillonner selon la loi $\hat{F}^*$ afin de produire un estimateur de variance pour le calcul de la statistique Bootstrap studentisée. Cela demande un temps de calcul très important. Une alternative suggérée par Sitter (1992) consiste à utiliser un estimateur Jackknife de variance pour $v^*(\theta(\hat{F}^*))$; on peut également utiliser la technique de linéarisation (Deville, 1999), ce que l'on fera dans certaines simulations pour obtenir des approximations d'intervalles de confiance de type t-Bootstrap.

Cas d'une population finie

L'adaptation du Bootstrap au cas d'une population finie a suscité une littérature abondante, depuis l'article de Gross (1980) ; sans prétendre à l'exhaustivité, nous présenterons des méthodes proposées depuis lors dans les chapitres suivants. Il est important de souligner la parenté entre le principe de plug-in en population infinie, et le principe d'estimation de Horvitz-Thompson en population finie, présenté aux paragraphes 1.1.3 et 1.3.1 ; la mesure M joue le rôle de la loi inconnue F, que l'on estime par la mesure $\hat{M} = \sum_{k \in S} \delta_k / \pi_k$ calculée sur l'échantillon en tenant compte des poids de Horvitz-Thompson.

Comme nous l'avons souligné précedemment, les enquêtes sont généralement entachées de non-réponse (partielle ou totale) qui détériorent l'échantillonnage d'origine. S'il est possible d'obtenir des résultats de validité au second ordre dans le cas idéal où tout l'échantillon sélectionné est effectivement enquêté, ces résultats sont plus délicats à justifier dans une situation pratique où des phénomènes complexes de non-réponse entrent en jeu. En particulier, il ne paraît pas évident de justifier d'une plus grande pertinence du Bootstrap par rapport à l'approximation normale. D'autre part, des techniques telles que la linéarisation constituent des outils très généraux, disponibles sous forme de logiciels et largement utilisés. Dans ce cas, quelle utilité pour des méthodes de Bootstrap en population finie ?

On peut d'abord noter que, même si la validation théorique du Bootstrap en population finie pose des difficultés, de nombreuses simulations montrent le bon comportement d'une méthode de Bootstrap judicieusement choisie, et des taux de couverture théoriques souvent mieux respectés qu'avec l'approximation normale pour des paramètres non lisses de type fractiles. Nous montrerons également dans les chapitres suivants qu'une méthode de Bootstrap basée sur le principe de π-estimation permet d'obtenir une estimation consistante de variance pour une gamme étendue de plans de sondage.

D'autre part, une technique telle que la linéarisation suppose de pouvoir retracer l'ensemble du plan de sondage ; or, la chaîne de traitements d'une enquête sépare généralement le concepteur et l'utilisateur, ce dernier ne disposant que du fichier d'enquête éventuellement muni d'une variable de poids synthétisant les différents traitements (échantillonnage, traitement de la non-réponse, redressement, ...). Le calcul par le concepteur de poids Bootstrap (issus par exemple d'une méthode de type percentile) adjoints au fichier d'enquête fournirait à l'utilisateur un moyen simple de calcul de précision, pour n'importe quel domaine de l'enquête. Enfin, certaines enquêtes mettent en jeu une stratégie (échantillonnage et estimation) complexe, dont la précision est délicate à évaluer avec la linéarisation ; le Bootstrap peut alors fournir une alternative intéressante (voir le chapitre 6), bien qu'il soit dans ce cas davantage fondé sur des principes de bon sens plutôt que sur une démonstration rigoureuse.

Bibliographie

Ardilly, P. (2006). *Les Techniques de Sondage*. Technip, Paris.

Babu, G. and Singh, K. (1985). Edgeworth expansions for sampling without replacement from a finite population. *Journal of Multivariate Analysis*, 17 :261–278.

Berger, Y. and Rao, J. (2006). Adjusted jackknife for imputation under unequal probability sampling without replacement. *Journal of the Royal Statistical Society*, B68 :531–547.

Bickel, P. and Freedman, D. (1984). Asymptotic normality and the bootstrap in stratified sampling. *The Annals of Statistics*, 12 :470–482.

Bloznelis, M. and Götze, F. (2001). An edgeworth expansion for symmetric finite population statistics. *The Annals of Statistics*, 29 :899–917.

Bloznelis, M. and Götze, F. (2002). An edgeworth expansion for symmetric finite population statistics. *The Annals of Probability*, 30 :1238–1265.

Brewer, K. and Donadio, M. (2003). The high entropy variance of the horvitz-thompson estimator. *Survey Methodology*, 29 :189–196.

Cassel, C., Särndal, C., and Wretman, J. (1976). Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika*, 63 :615–620.

Davison, A. and Hinkley, D. (1997). *Bootstrap Methods and Their Application*. Cambridge University Press, Cambridge.

Davison, A. and Sardy, S. (2007). Méthodes de rééchantillonnage pour l'estimation de variance. *soumis au Journal de la Société Française de Statistique*.

Deville, J. (1987). Réplications d'échantillons, demi-échantillons, jackknife, bootstrap. In *Les Sondages*, Paris. Economica.

Deville, J.-C. (1993). Estimation de la variance pour les enquêtes en deux phases. Note interne manuscrite, INSEE, France.

Deville, J.-C. (1999). Variance estimation for complex statistics ans estimators : linearization and residual techniques. *Techniques d'enquête, Survey methodology*, 25 :193–204.

DiCiccio, T. and Tibshirani, R. (1987). Bootstrap confidence intervals and bootstrap approximations. *Journal of the American Statistical Association*, 82 :163–170.

Efron, B. (1979). Bootstrap methods : another look at the jackknife. *Annals of Statistics*, 7 :1–26.

Efron, B. (1981). Nonparametric standard errors and confidence intervals (with discussion). *Canadian Journal of Statistics*, 9 :139–172.

Efron, B. (1982). *The jackknife, the Bootstrap and Other Resampling Plans*, volume 38. ACBMS-N SIAM.

Efron, B. (1987). Better bootstrap confidence intervals (with discussions). *Journal of the American Statistical Association*, 82 :171–200.

Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap*. Chapman and Hall, New-York.

Erdös, P. and Renyi, A. (1959). On the central limit theorem for samples from a finite population. *Publ.Math.Inst.Hung.Acad.Sci.*, 4 :49–57.

Feller, W. (1966). *An introduction to Probability Theory and its applications*. Wiley, New-York.

Fuller, W. and Isaki, C. (1981). *Survey design under superpopulation models*, chapter Currents topics in survey sampling, pages 196–226. Academic Press, New York.

Goga, C., Deville, J., and Ruiz-Gazen, A. (2007). Composite estimation and linearization method for two-sample survey data. *en révision dans Biometrika*.

Gross, S. (1980). Median estimation in sample surveys. *ASA Proceedings of Survey Research*, pages 181–184.

Hájek, J. (1960). Limiting distributions in simple random sampling from a finite population. *Publ.Math.Inst.Hung.Acad.Sci.*, 5 :361–374.

Hájek, J. (1964). Asymptotic theory of rejective sampling with varying probabilities from a finite population. *Annals of Mathematical Statistics*, 35 :1491–1523.

Hall, P. (1986). On the bootstrap and confidence intervals. *Annals of Mathematical Statistics*, 14 :1431–1452.

Hall, P. (1988). Theoretical comparisons of bootstrap confidence intervals. *Annals of Mathematical Statistics*, 16 :927–953.

Hall, P. (1992). *The Bootstrap and Edgeworth Expansion*. Springer-Verlag, New-York.

Hampel, F., Ronchetti, E., Rousseuw, P., and Stahel, W. (1985). *Robust Statistics : The Approach Based on the Influence Function*. Wiley, New-York.

Horvitz, D. and Thompson, D. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47 :663–685.

Isaki, C. and Fuller, W. (1982). Survey design under a regression population model. *Journal of the American Statistical Association*, 77 :89–96.

Konishi, S. (1991). Normalizing transformations and bootstrap confidence intervals. *Annals of Statistics*, 19 :2209–2225.

Kovar, J., Rao, J., and Wu, C. (1988). Bootstrap and other methods to measure errors in survey estimates. *The Canadian Journal of Statistics*, 16 :25–45.

Krewski, D. and Rao, J. (1981). Inference from stratified samples : properties of the linearization, jackknife and balanced repeated replication methods. *Annals of Statistics*, 9 :1010–1019.

Kuk, A. (1988). Estimation of distribution functions and medians under sampling with unequal probabilities. *Biometrika*, 75 :97–103.

Lahiri, P. (2003). On the impact of bootstrap in survey sampling and small-area estimation. *Statistical Science*, 18 :199–210.

Mac Carthy, P. (1969). Pseudo-replication : Half samples. *Review of the International Statistics Institute*, 37 :239–264.

Matei, A. and Tillé, Y. (2005). Evaluation of variance approximations and estimators in maximum entropy sampling with unequal probability and fixed sample size. *Journal of Official Statistics*, 21 :543–570.

Nigam, A. and Rao, J. (1996). On balanced bootstrap for stratified multistage samples. *Statistica Sinica*, 6 :199–214.

Presnell, B. and Booth, J. (1994). Resampling methods for sample surveys. Technical report.

Quenouille, M. (1949a). Approximate tests of correlation in time-series. *Journal of the Royal Statistical Society*, B11 :68–84.

Quenouille, M. (1949b). Notes on bias in estimation. *Biometrika*, 43 :353–360.

Rao, J. and Shao, M. (1996). On balanced half-sample variance estimation in stratified random sampling. *Journal of the American Statistical Association*, 91 :343–348.

Rao, J. and Tausi, M. (2004). Estimatinf function jackknife variance estimators under stratified multistage sampling. *Communications in Statistics Theory and Methods*, 33 :2087–2095.

Rao, J. and Wu, C. (1985). Inference from stratified samples : Second-order analysis of three methods for nonlinear statistics. *Journal of the American Statistical Association*, 80 :620–630.

Rao, J. and Wu, C. (1988). Resampling inference with complex survey data. *Journal of the American Statistical Association*, 83 :231–241.

Rao, J., Wu, C., and Yue, K. (1992). Some recent work on resampling methods for complex surveys. *Survey Methodology*, 18 :209–217.

Robinson, J. (1978). An asymptotic expansion for samples from a finite population. *Annals of Statistics*, 6 :1005–1011.

Rosen, B. (1972a). Asymptotic theory for successive sampling i. *Annals of Mathematical Statistics*, 43 :373–397.

Rosen, B. (1972b). Asymptotic theory for successive sampling ii. *Annals of Mathematical Statistics*, 43 :748–776.

Schenker, N. (1985). Qualms about bootstrap confidence intervals. *Journal of the American Statistical Association*, 80 :360–361.

Sen, A. (1953). On the estimate of the variance in sampling with varying probabilities. *Journal of Indian Society for Agricultural Statistics*, 5 :119–127.

Sen, P. (1980). Limit theorem for an extended coupon's collector problem and for successive sub-sampling with varying probabilities. *Calcutta Statistical Association Bulletin*, 29 :113–132.

Sen, P. (1988). Asymptotics in finite population sampling. *Handbooks of Statistics*, 6 :291–331.

Shao, J. and Tu, D. (1995). *The Jackknife and The Bootstrap*. Springer, New-York.

Shao, M. and Rao, J. (1994). Standard errors for lowincome proportions estimated from stratified multistage samples. *Sankhya*, B55 :393–414.

Shao, M. and Wu, C. (1989). A general theory for jackknife variance estimation. *Annals of Statistics*, 17 :1176–1197.

Shao, M. and Wu, C. (1992). Asymptotic properties of the balanced repeated replication method for sample quantiles. *Annals of Statistics*, 20 :1571–1593.

Sitter, R. (1992). Comparing three bootstrap methods for survey data. *Canadian Journal of Statistics*, 20 :135–154.

Särndal, C.-E., Swensson, B., and Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer Verlag, New York.

Sugden, R. and Smith, T. (1997). Edgeworth approximations to the distribution of the sample mean under simple random sampling. *Statistics and Probability Letters*, 34 :293–299.

Sugden, R., Smith, T., and Jones, R. (2000). Cochran's rule for simple random sampling. *Journal of the Royal Statistical Society*, B62 :787–793.

Thompson, M. (1997). *Theory of Sample Surveys*. Chapman and Hall, London.

Tillé, Y. (2001). *Théorie des sondages : échantillonnage et estimation en populations finies*. Dunod, Paris.

Tukey, J. (1958). Bias and confidence in not quite large samples. *Annals of Mathematical Statistics*, 29 :614.

Valliant, R., Dorfman, A., and Royall, R. (2000). *Finite population sampling and inference : a prediction approach*. Wiley Series in Probability and Statistics, New-York.

Wolter, K. (1985). *Introduction to Variance Estimation*. Springer-Verlag, New York.

Wolter, K. (2007). *Introduction to Variance Estimation*. Springer, New York.

Woodruff, R. (1952). Confidence intervals for medians and other position measures. *Journal of the American Statistical Association*, 47 :635–646.

Wu, C. (1986). Jackknife, bootstrap and other resampling methods in regression analysis (with discussions). *Annals of Statistics*, 14 :1261–1350.

Wu, C. (1990). On the asymptotic properties of the jackknife histogram. *Annals of Statistics*, 18 :1438–1452.

Wu, C. (1991). Balanced repeated replications based on mixed orthogonal arrays. *Biometrika*, 78 :181–188.

Yates, F. and Grundy, P. (1953). Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society*, B15 :235–261.

Chapitre 2

Bootstrap pour le sondage aléatoire simple

Malgré sa relative simplicité, le sondage aléatoire simple est fondamental car il est à la base de plans de sondage plus complexes et très utilisés en pratique, alliant la stratification et le tirage multidegrés. D'où son importance dans la littérature et la littérature Bootstrap en particulier, et les nombreuses méthodes qui ont été proposées pour estimer la précision d'un échantillon aléatoire simple.

L'objectif de ces méthodes peut être résumé en quelques mots : capter la correction de population finie, c'est à dire le gain obtenu en échantillonnant sans remise plutôt qu'avec remise. Cette correction n'est généralement pas captée par le Bootstrap d' Efron (1982), voir par exemple Mac Carthy and Snowden (1985). D'où les nombreux algorithmes proposés, que l'on peut en suivant la classification suggérée par Presnell and Booth (1994) répartir schématiquement en deux groupes : les **méthodes "ad-hoc"**, se calant explicitement sur l'estimateur de variance dans le cas linéaire, et les **méthodes de type plug-in**, faisant intervenir un principe de substitution similaire à celui de la théorie classique. Dans la suite de ce chapitre, nous utiliserons la notation $[x]$ pour désigner l'entier le plus proche de x. La valeur $[x - 1/2]$ correspondra donc à la partie entière de x.

Le chapitre est organisé de la façon suivante. En section 1, nous effectuons quelques rappels sur le sondage aléatoire simple. En section 2, nous rappelons les principales méthodes de Bootstrap proposées pour le sondage aléatoire

simple, et argumentons sur les avantages d'une méthode de type plug-in. En section 3, nous proposons trois méthodes de Bootstrap visant à limiter le fardeau de simulation, et nous testons ces méthodes à l'aide de quelques simulations.

2.1 Rappels sur le sondage aléatoire simple

2.1.1 Définition

Un plan de sondage est dit **simple** si tous les échantillons de même taille ont la même probabilité d'être sélectionnés. Le sondage aléatoire simple (sans remise) de taille n, ici noté p, est un plan de sondage simple et de taille fixe. Ces deux propriétés permettent de déterminer entièrement le plan de sondage :

$$p(s) = \begin{cases} 1/C_N^n \text{ si } \sharp s = n \\ 0 \text{ sinon} \end{cases}$$

Les probabilités d'inclusion peuvent être déterminées exactement à tout ordre. On établit en particulier la propriété suivante :

Propriété 2.1. *Soit p un plan de sondage simple sans remise de taille n. Alors :*

- $\rightarrow$ $\pi_k = \frac{n}{N}$ *pour tout* $k \in U$
- $\rightarrow$ $\pi_{kl} = \frac{n(n-1)}{N(N-1)}$ *pour tous* $k \neq l \in U$

Démonstration. Voir Tillé (2001). □

2.1.2 Estimation et calcul de précision

Compte tenu de la propriété précédente, le π-estimateur du total de y est égal à $N/n \sum_{k \in s} y_k = N\bar{y}$, où $\bar{y}$ désigne la moyenne de la variable y sur l'échantillon S. Par linéarité, μ_y est estimée sans biais par $\bar{y}$: dans un sondage aléatoire simple, la moyenne simple sur l'échantillon estime sans biais la moyenne simple sur la population.

La précision de $\bar{y}$ peut être calculée exactement.

Propriété 2.2. *Soit S un échantillon tiré selon un sondage aléatoire simple, et $\bar{y}$ la moyenne simple calculée sur l'échantillon. Alors :*

$$V(\bar{y}) = \frac{1-f}{n} S_y^2$$

où $S_y^2 = 1/(N-1)\sum_{k\in U}(y_k - \mu_y)^2$ désigne la dispersion de la variable y sur la population U.

Cette variance peut être estimée sans biais par

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

où $s_y^2 = 1/(n-1)\sum_{k\in S}(y_k - \bar{y})^2$ désigne la dispersion de la variable y sur S.

Démonstration. Voir Tillé (2001). □

N est généralement grand en pratique, de sorte que S_y^2 est peu différent de la variance (non corrigée) $\sigma_y^2 = N^{-1}\sum_{k\in U}(y_k - \mu_y)^2$. Le terme $1-f$, appelé correction de population finie, représente le gain obtenu en échantillonnant sans remise plutôt qu'avec remise. Ce gain peut être conséquent : par exemple, dans le cas d'un tirage stratifié où l'on surreprésente certaines strates dans lesquelles la variable d'intérêt est particulièrement dispersée.

2.2 Les méthodes de Bootstrap existantes

De nombreuses adaptations de la méthode d'Efron (1982) ont été proposées dans le cas d'un sondage aléatoire simple sans remise. Le sondage aléatoire simple "direct" est rarement utilisé, mais il sert de base à la construction de plans de sondage plus complexes (plans stratifiés, plans à plusieurs degrés) très employés en pratique ; cela explique l'importance du sondage aléatoire simple dans la littérature sur le Bootstrap en population finie. Certaines des méthodes présentées ci-dessous peuvent se généraliser à des plans à probabilités inégales, nous y reviendrons dans les chapitres suivants. Nous présentons ici les principales méthodes existantes dans le cas d'un sondage aléatoire simple à un degré.

Dans la suite de ce chapitre, nous adopterons les notations suivantes. S désigne un échantillon de taille n sélectionné par sondage aléatoire simple. La moyenne simple de la variable y sur l'échantillon S est notée $\bar{y}$, et l'estimateur sans biais de variance pour $\bar{y}$ est noté $v(\bar{y})$.

2.2.1 Le Bootstrap avec remise (Mac Carthy and Snowden, 1985)

Le Bootstrap avec remise (noté BWR), proposé par Mac Carthy and Snowden (1985), est une transposition directe de la méthode classique de Bootstrap au cas d'un échantillonnage sans remise en population finie. On sélectionne dans S un rééchantillon S^* de taille m selon un sondage aléatoire simple **avec remise**.

On note $\bar{y}^*$ la moyenne simple de la variable y calculée sur le rééchantillon S^*. Mac Carthy and Snowden (1985) établissent que :

$$E(\bar{y}^*|S) = \bar{y}$$
$$V(\bar{y}^*|S) = \left(1 - \frac{1}{n}\right)\frac{s_y^2}{m}$$

Dans le cas où $m = n$, on retrouve la méthode de Bootstrap classique utilisée pour le tirage avec remise. Si de plus $f = 1/n$, autrement dit si $n^2 = N$, cette méthode permet de capter exactement la correction de population finie et la méthode de Bootstrap estime sans biais la variance. Ce cas est rare en pratique.

Dans le cas général, Mac Carthy and Snowden (1985) suggèrent de choisir $m = (n-1)/(1-f)$ afin de capter la correction de population finie. En pratique, $(n-1)/(1-f)$ peut être non entier. Sitter (1992b) suggère de choisir aléatoirement entre les deux entiers les plus proches $[(n-1)/(1-f)]$ et $[(n-1)/(1-f)] + 1$, de façon à ce que l'estimation de variance soit, en moyenne, non biaisée. Une alternative consiste à prendre $m = n - 1$, correction déjà suggérée par Efron (1982). On a alors :

$$V(\bar{y}^*|S) = \frac{s_y^2}{n}$$

Dans ce cas, la méthode ne capte pas la correction de population finie et la variance est sur-estimée. Cependant, le biais dans l'estimation de variance dépend de f et non de n. Quand le taux de sondage est faible, le sondage simple sans remise est très proche du sondage simple avec remise : le BWR peut alors se révéler compétitif par rapport à d'autres méthodes de Bootstrap dont le biais dépend de n (ce qui peut être gênant pour un échantillon de taille réduite).

2.2.2 Le Rescaled Bootstrap (Rao and Wu, 1988)

Rao and Wu (1988) suggèrent d'utiliser le schéma précédent pour sélectionner un rééchantillon s^*. Les valeurs prises par la variable y sur ce rééchantillon sont alors réajustées afin de fournir une estimation sans biais de variance.

Théorème 2.3. *Rao and Wu (1988)*

Soit S^* un rééchantillon sélectionné dans S selon un sondage aléatoire simple avec remise de taille m. Pour $k \in S^*$, soient :

$$\tilde{y_k} = \bar{y} + \left[\frac{m(1-f)}{n-1}\right]^2 (y_k - \bar{y})$$

et

$$\begin{aligned} \bar{\tilde{y}}^* &= 1/m \sum_{k \in S^*} \tilde{y_k} \\ &= \bar{y} + \left[\tfrac{m(1-f)}{n-1}\right]^2 (\bar{y}^* - \bar{y}) \end{aligned}$$

Alors le Rescaled Bootstrap restitue les estimateurs sans biais habituels des moments d'ordre 1 et 2, au sens où :

$$E(\bar{\tilde{y}}^*|S) = \bar{y}$$

$$V(\bar{\tilde{y}}^*|S) = v(\bar{y})$$

Démonstration. Ce théorême résulte des identités

$$E(\bar{y}^* - \bar{y}|S) = 0$$

et

$$V(\bar{y}^* - \bar{y}|S) = V(\bar{y}^*|S) = \frac{1}{m}\frac{n-1}{n}s_y^2$$

□

Notons que le résultat précédent est vrai pour m quelconque. Rao and Wu (1988) montrent que le Rescaled Bootstrap (que l'on notera RB) fournit une estimation consistante de variance pour un estimateur qui peut s'écrire comme une fonction lisse de moyennes $\hat{\theta} = f(\bar{y})$. Ils suggèrent d'utiliser

$$m = \frac{1-f}{(1-2f)^2}\frac{(n-2)^2}{n-1}$$

afin que le Bootstrap restitue l'estimateur sans biais du moment d'ordre 3. Le terme de droite n'est généralement pas entier : il est possible d'introduire une randomisation entre les deux entiers les plus proches, mais cela détériore la stabilité du Bootstrap. Presnell and Booth (1994) et Sitter (1992a,b) soulignent également que le réajustement peut conduire à des valeurs impossibles pour la statistique Bootstrappée. En particulier, une statistique strictement positive peut avoir des répliques Bootstrap strictement négatives avec ce schéma de rééchantillonnage.

2.2.3 Le Mirror-Match Bootstrap (Sitter, 1992b)

Sitter (1992b) propose un autre algorithme, où l'idée consiste à échantillonner de façon répétée dans l'échantillon S, selon un sondage aléatoire simple de taille n', et à réunir les sous-échantillons obtenus jusqu'à obtenir un rééchantillon S^* de taille $n^* = k'n'$, où k' désigne le nombre de sous-échantillonnages réalisés dans S.

Théorème 2.4. *Sitter (1992b)*

Soit n' un entier tel que $1 \leq n' < n$. On pose $f^* = n'/n$ et $k' = \frac{n(1-f^*)}{n'(1-f)}$. k' est supposé entier. On sélectionne dans S, de façon indépendante, k' sous-échantillons $s_1^*, \ldots, s_{k'}^*$ de taille n' par sondage aléatoire simple sans remise.

Soit S^* obtenu en réunissant $s_1^*, \ldots, s_{k'}^*$, et $\bar{y}^*$ la moyenne simple de la variable y sur S^*. Alors le Mirror-Match Bootstrap restitue les estimateurs sans biais habituels des moments d'ordre 1 et 2, au sens où :

$$E(\bar{y}^*|S) = \bar{y}$$

$$V(\bar{y}^*|S) = v(\bar{y})$$

Démonstration. On note $\bar{y}_i^*$ la moyenne de la variable y sur le sous-échantillon s_i^*. On a :

$$\bar{y}^* = \frac{1}{k'} \sum_{i=1}^{k'} \bar{y}_i^*$$

On en déduit que

$$E(\bar{y}^*|S) = E(\bar{y}_i^*|S) = \bar{y}$$

et par indépendance entre les k' sous-échantillonnages

$$\begin{aligned} V(\bar{y}^*|S) &= \tfrac{1}{k'} V(\bar{y}_i^*|S) \\ &= \tfrac{1}{k'} \tfrac{1-n'/n}{n'} s_y^2 \\ &= \tfrac{1-f^*}{n'k'} s_y^2 \\ &= v(\bar{y}) \end{aligned}$$

□

Dans la suite, cette méthode sera notée MMB. Sitter (1992b) suggère d'utiliser $n' = fn$, c'est à dire sous-échantillonner avec le même taux de sondage qu'au départ. Quand il est possible, ce choix est assez intuitif, car il permet véritablement d'appliquer de façon répétée le plan de sondage de départ (au sens du taux de sondage) dans l'échantillon S. D'autre part, ce choix permet d'obtenir un rééchantillon S^* de même taille que S, et de restituer l'estimateur sans biais du moment d'ordre 3. Mais comme le soulignent Presnell and Booth (1994), fn est rarement entier, et il faut alors effectuer une randomisation sur n' et k'. Dans tous les cas, le calage sur l'estimateur sans biais du moment d'ordre 3 est impossible si $n^2 \leq N$; cette situation n'est pas rare en pratique, notamment dans le cas particulier important d'un tirage stratifié de taille 2 dans chaque strate.

Bien que généralement classée dans les méthodes de type Bootstrap, le MMB peut être vu comme une adaptation du delete-d Jackknife (Wu, 1986, 1990). Le sous-échantillonnage dans S avec le taux de sondage de départ f permet de capter la correction de population finie, sans passer par des corrections a posteriori que l'on applique généralement à l'estimateur de variance donné par le Jackknife de base (Shao and Tu, 1995, page 238).

2.2.4 Le Bootstrap sans remise ou BWO (Gross, 1980)

Une autre méthode, proposée à l'origine par Gross (1980), consiste à reproduire les conditions initiales de tirage en constituant à l'aide de l'échantillon S une pseudo-population U^*, image de la population d'origine, dans laquelle le plan de sondage de départ est appliqué de façon répétée (d'où le nom de Bootstrap populationnel que l'on rencontre parfois). Cet algorithme sera noté

BWO.

Gross (1980) suppose que l'inverse du taux de sondage $f^{-1} = N/n$ est entier. Chaque élément de S est dupliqué $1/f$ fois pour créer une pseudo-population U^*. On sélectionne alors dans U^* un rééchantillon S^*, par sondage aléatoire simple de taille n. On note $\bar{y}^*$ la moyenne simple de la variable y sur le rééchantillon S^*.

Théorème 2.5. *Mac Carthy and Snowden (1985)*

Soit S^* un rééchantillon sélectionné dans S selon l'algorithme de Gross. Alors :

$$E(\bar{y}^*|S) = \bar{y}$$
$$V(\bar{y}^*|S) = \frac{N(n-1)}{n(N-1)} v(\bar{y})$$

Démonstration. On a :

$$\begin{aligned} E(\bar{y}^*|S) &= N^{-1} \textstyle\sum_{k\in U^*} y_k \\ &= N^{-1} \tfrac{N}{n} \textstyle\sum_{k\in S} y_k \\ &= \bar{y} \end{aligned}$$

D'autre part, en utilisant la formule de variance d'un estimateur de moyenne dans un sondage aléatoire simple :

$$\begin{aligned} V(\bar{y}^*|S) &= \tfrac{1-f}{n} \tfrac{1}{N-1} \textstyle\sum_{k\in U^*} (y_k - \bar{y})^2 \\ &= \tfrac{1-f}{n} \tfrac{1}{N-1} \tfrac{N}{n} \textstyle\sum_{k\in S} (y_k - \bar{y})^2 \\ &= \tfrac{N(n-1)}{n(N-1)} v(\bar{y}) \end{aligned}$$

□

Le Bootstrap de Gross restitue donc l'estimateur naturel de variance à un facteur $\frac{N(n-1)}{n(N-1)} \doteq \frac{n-1}{n}$ près. Notons que ce biais se retrouve également dans le Bootstrap classique (Efron (1982)).

Intuitivement séduisant, le Bootstrap de Gross présente l'avantage d'obéir, comme le Bootstrap d'Efron, à un véritable principe de plug-in, et n'est donc pas uniquement motivé par un ajustement sur les estimateurs naturels des premiers moments. On définit ici le principe de plug-in comme étant le principe d'estimation de Horvitz-Thompson, présenté au chapitre précédent, où la mesure

$$M = \sum_{k \in U} \delta_k$$

qui place une masse unité sur chaque individu k de U est estimée par sa mesure par substitution

$$\hat{M} = \sum_{k \in S} \frac{\delta_k}{\pi_k}$$

qui place une masse $1/\pi_k$ sur chaque individu k de S (voir également 1.3.1). Un paramètre $\theta(M)$ est donc remplacé par son estimateur par substitution $\theta(\hat{M})$.

La méthode de Gross a été abondamment reprise et discutée dans la littérature ; sans prétendre à l'exhaustivité, nous présentons ci-dessous des variantes qui ont été proposées pour tenir compte du cas (général) où $\frac{N}{n}$ n'est pas entier. Dans ce cas, la constitution de U^* n'est pas immédiate et passe par une étape de randomisation (afin que chaque unité de l'échantillon soit dupliquée en moyenne N/n fois) qui génère une variance parasite. Pour solutionner ce problème, Booth et al. (1994) suggèrent comme principe de plug-in d'estimer un paramètre $\theta = \theta(M)$ par $E\left(\theta(M^*)|S\right)$, où

$$M^* = \sum_{k \in U^*} \delta_k$$

désigne la mesure qui place une masse unité sur chaque unité de U^*.

Variante de Bickel and Freedman (1984) et Chao and Lo (1985)

Bickel and Freedman (1984) et Chao and Lo (1985) proposent, à chaque étape du Bootstrap, de choisir entre dupliquer $[N/n - 1/2]$ fois chaque unité de S avec une probabilité α (on rappelle que $[. - 1/2]$ donne la partie entière), et dupliquer $[N/n - 1/2] + 1$ fois chaque unité, avec une probabilité $1 - \alpha$. On prélève ensuite un échantillon S^* par sondage aléatoire simple de taille n dans la population U^* ainsi constituée. Cette méthode, que l'on notera

BFCL, coincide exactement avec le Bootstrap de Gross quand N/n est entier.

Mac Carthy and Snowden (1985) suggèrent de choisir α de façon à restituer l'estimateur sans biais habituel $v(\bar{y})$, mais un tel choix n'est pas toujours possible, notamment dans le cas où N/n est entier ; voir également Sitter (1992a). Sinon, on peut toujours choisir α tel que

$$E\left(V(\bar{y}^*|U^*)|\, S\right) = \frac{N(n-1)}{n(N-1)} v(\bar{y})$$

Variante de Booth et al. (1994)

Lors de la constitution de U^*, Booth et al. (1994) proposent de dupliquer $[N/n - 1/2]$ fois chaque unité de S, et de compléter les $n \times [N/n - 1/2]$ unités ainsi obtenues en sélectionnant un échantillon de taille $r = N - n \times \lfloor N/n \rfloor$ dans S, pour obtenir finalement une pseudo-population U^*. On prélève ensuite un échantillon S^* par sondage aléatoire simple de taille n dans la population U^* ainsi constituée.

Avec cette méthode, que l'on notera BBH, Presnell and Booth (1994) montrent que l'estimation Bootstrap de $Var(\bar{y})$ est donnée par

$$E\left(V(\bar{y}^*|U^*)|\, S\right) = \left[1 - \frac{r}{N}\left(1 - \frac{r-1}{N-1}\right)\right] \frac{N(n-1)}{n(N-1)} v(\bar{y})$$

Là encore, cette variante coincide exactement avec la méthode de Gross quand N/n est entier, i.e. quand $r = 0$.

Variante de Presnell and Booth (1994)

Presnell and Booth (1994) proposent de constituer la pseudo-population U^* en sélectionnant dans S un échantillon de taille N, par sondage aléatoire simple avec remise. Le rééchantillon S^* est ensuite prélevé dans U^* par sondage aléatoire simple de taille n.

Avec cette méthode, que l'on notera PB, Presnell and Booth (1994) montrent que l'estimation Bootstrap de $Var(\bar{y})$ est donnée par

$$E\left(V(\bar{y}^*|U^*)|\, S\right) = \frac{n-1}{n} v(\bar{y})$$

Cette méthode ne coincide pas avec le Bootstrap de Gross dans le cas où $\frac{N}{n}$ est non entier.

2.2.5 Le Bootstrap pondéré (Bertail and Combris, 1997)

Toute procédure d'estimation de variance par rééchantillonnage peut être vue comme un algorithme consistant à affecter aux individus de l'échantillon d'origine S des poids aléatoires donnés par le résultat du rééchantillonnage. Par exemple, avec la méthode de Gross (1980), le poids affecté à un individu de S est égal au nombre de fois où l'une de ses réplications est sélectionnée par rééchantillonnage dans U^*.

Bertail and Combris (1997) proposent d'appliquer dans le cadre d'une population finie la méthode de Bootstrap pondéré (Lo, 1992; Mason and Newton, 1992; Barbe and Bertail, 1995). L'idée consiste à affecter directement aux individus des poids aléatoires, sans passer par une procédure de rééchantillonnage. Les premiers moments de ces poids sont choisis dans le but de restituer les estimateurs sans biais habituels.

Théorème 2.6. *Bertail and Combris (1997)*

On note W_k le poids aléatoire affecté à l'individu k de S. On suppose que ces poids vérifient les propriétés suivantes :

- $\rightarrow\ E(W_k) = 1$
- $\rightarrow\ V(W_k) = 1 - f$
- $\rightarrow\ k \neq l \Rightarrow Cov(W_k, W_l) = -\frac{1-f}{n-1}$

On note $\bar{y}_W = \frac{1}{n}\sum_{k\in S} W_k y_k$. Alors

$$E(\bar{y}_W|S) = \bar{y}$$

$$V(\bar{y}_W|S) = v(\bar{y})$$

Démonstration. On a :

$$\begin{aligned} E(\bar{y}_W|S) &= \frac{1}{n}\sum_{k\in S} E(W_k) y_k \\ &= \bar{y} \end{aligned}$$

D'autre part

$$\begin{aligned} V(\bar{y}_W|S) &= \tfrac{1}{n^2}\sum_{k\in S} V(W_k)y_k^2 + \tfrac{1}{n^2}\sum_{k\neq l\in S} Cov(W_k,W_l)y_k y_l \\ &= \tfrac{1-f}{n}\left(\tfrac{1}{n}\sum_{k\in S} y_k^2 - \tfrac{1}{n(n-1)}\sum_{k\neq l\in S} y_k y_l\right) \\ &= v(\bar{y}) \end{aligned}$$

□

Toutes les méthodes évoquées précedemment peuvent être vues comme des cas particuliers du Bootstrap pondéré, les poids aléatoires affectés aux unités de S étant donnés par la procédure de rééchantillonnage. Ici, les poids peuvent être générés directement. Bertail and Combris (1997) donnent des méthodes effectives pour générer des variables aléatoires dont les premiers moments sont fixés ; voir également Devroye (1986).

2.2.6 Discussion

Nous nous plaçons dans le même cadre asymptotique que dans le chapitre précédent, où n et N s'accroissent, mais le taux de sondage f reste borné. Les résultats présentés s'étendent de façon immédiate au cas d'un tirage stratifié sur un nombre fini de strates. Une autre asymptotique suppose que le nombre H de strates s'accroît, alors que les tailles d'échantillon n_h et les tailles de strates N_h restent bornées. Ce second cadre, qui présente lui aussi un réel intérêt pratique (une stratégie courante consiste à partitionner la population en autant de strates que possible, avec la seule contrainte d'échantillonner au moins deux individus par strate), ne sera pas discuté ici. Nous donnerons cependant quelques arguments concernant le choix d'une méthode de Bootstrap dans le cas d'une faible taille d'échantillon.

Presnell and Booth (1994) divisent les méthodes de Bootstrap en population finie en deux catégories :

→ d'une part les méthodes visant explicitement à se caler sur les estimateurs sans biais des premiers moments dans le cas linéaire, que nous appellerons **méthodes de type ad-hoc**, pour reprendre la terminologie de Presnell and Booth (1994),

→ d'autre part les méthodes reposant sur un principe de plug-in comme le Bootstrap d'origine proposé par Efron (1982), que nous appellerons **méthodes de type plug-in**

Les méthodes de Bootstrap de type ad-hoc

On peut classer dans cette catégorie les méthodes RB et MMB, ainsi que la variante de la méthode de Gross proposée par Sitter (1992a). Ces méthodes fournissent une estimation de variance exactement sans biais dans le cas linéaire, et consistante pour un estimateur $\hat{\theta} = f(\bar{y})$ pouvant s'écrire comme une fonction suffisamment lisse de moyennes.

Presnell and Booth (1994) argumentent que les méthodes RB et MMB ne sont pas exactes au second ordre (et donc que les taux de couverture ne sont pas mieux respectés qu'avec une approximation normale) et que plusieurs choix sont a priori possibles pour le taux de sondage f^* lors de la phase de rééchantillonnage ; ce point n'est pas détaillé dans Rao and Wu (1988) et Sitter (1992b), ce qui présente un risque pour l'utilisateur. Si la valeur adéquate de f^* n'est pas utilisée, le Bootstrap peut ne pas être exact au premier ordre.

Pour la production d'intervalles de confiance, l'utilisation de techniques de type t-Bootstrap nécessite de pouvoir calculer, pour chaque statistique Bootstrap $\hat{\theta}^*$, une estimation de variance, ce qui passe normalement par un double Bootstrap. En effet, produire une estimation de variance pour $\hat{\theta}^*$ suppose de réappliquer au rééchantillon S^* toute la procédure de Bootstrap. Par double Bootstrap, on entend donc d'itérer la procédure de rééchantillonnage, une première fois sur l'échantillon S, puis sur le rééchantillon S^*. En plus d'un volume de calcul très conséquent, l'absence d'un principe de plug-in rend la réalisation de ce double Bootstrap délicate ; pour pallier à cette difficulté, Sitter (1992b) utilise une estimation de variance de type Jackknife. On peut également envisager d'utiliser un estimateur de variance obtenu par linéarisation.

Les méthodes de Bootstrap de type plug-in

Cette catégorie regroupe essentiellement les méthodes de type BWO. Dans une moindre mesure, et pour un taux de sondage faible, la méthode BWR

peut être vue comme l'approximation d'une méthode BWO.

Les différentes variantes proposées permettent de traiter le cas où N/n est non entier. Bien que le principe de plug-in proposé par Booth et al. (1994) et évoqué plus haut ne soit pas celui initialement retenu par Bickel and Freedman (1984) et Chao and Lo (1985), Presnell and Booth (1994) montrent que pour être exact au moins au premier ordre, ce principe doit également être utilisé pour la méthode BFCL. Davison and Hinkley (1997) soulignent que les différentes variantes de la méthode BWO donnent des résultats assez proches.
L'utilisation d'un principe de plug-in a l'avantage de donner une méthode très simple à utiliser, même pour les statistiques fortement non linéaires de type fractile. Il permet également de tenir compte simplement d'un éventuel calage à l'étape de la repondération : voir Canty and Davison (1999), ainsi que le chapitre 5 de ce document.

Le principe inconvénient des méthodes de type BWO est qu'elles nécessitent l'utilisation d'une sorte de "double Bootstrap", sur la population U^* et le réechantillon S^*, pour capter le terme $E(V(\theta(\widehat{M^*})|U^*)|S)$ utilisé. L'utilisation de ce terme pour estimer $V(\widehat{\theta(M)})$ permet de supprimer la variance parasite liée à la randomisation sur la constitution de U^*. Bien que l'existence d'un principe de plug-in rende l'utilisation de la méthode de t-Bootstrap théoriquement plus accessible qu'avec les méthodes de type ad-hoc, il semble pratiquement impossible d'itérer le double Bootstrap nécessaire pour obtenir une estimation de variance de chaque statistique $\hat{\theta}^*$ en raison d'un temps de calcul prohibitif. Ce point n'apparaît pas dans Presnell and Booth (1994), car le seul estimateur considéré est celui de la moyenne stratifiée, pour lequel on dispose d'un estimateur direct (et sans biais) de variance. Pour des fonctionnelles non linéaires, on peut envisager un estimateur de variance de type Jackknife ou un estimateur de variance linéarisé.

L'utilisation du "double Bootstrap" rend également les méthodes BWO difficiles à utiliser pour des enquêtes complexes, où on rencontre fréquemment plusieurs degrés d'échantillonnage. Dans ce cadre, il serait souhaitable de disposer d'une estimation Bootstrap "directe" de variance (même approchée), telle que la méthode des percentiles.

2.3 Résultats obtenus

Dans cette section, nous proposons trois méthodes de Bootstrap approché, de type BWO. Nous évaluons le comportement de ces méthodes à l'aide de quelques simulations. L'objectif est avant tout de tester des méthodes de Bootstrap ne nécessitant pas un fardeau de rééchantillonnage trop lourd. La réduction de ce fardeau se traduit par une perte d'efficacité, au sens où les méthodes proposées peuvent présenter un biais au niveau de l'estimation de variance. Dans le même souci d'avoir un volume de calcul limité, les intervalles de confiance sont produits uniquement à l'aide de la méthode des percentiles.

Pour une comparaison plus détaillée des méthodes de Bootstrap pour un sondage aléatoire simple, on pourra notamment se reporter à Presnell and Booth (1994) et Davison and Hinkley (1997).

2.3.1 Méthode BWO tronquée

Une généralisation simple, et couramment pratiquée, de la méthode de Gross (1980) dans le cas où N/n est non entier, consiste à dupliquer chaque unité de S exactement $p = [N/n]$ fois (on rappelle que [.] désigne l'entier le plus proche). On obtient ainsi une population U^* de taille $N^* = np$, dans laquelle on rééchantillonne de façon répétée selon le plan de sondage d'origine.

Avec une démonstration analogue à celle de 2.5 :

$$E(\bar{y}^*|S) = \bar{y}$$

$$V(\bar{y}^*|S) = \frac{N^*(n-1)}{n(N^*-1)} \frac{1-f^*}{n} s_y^2$$

Le biais obtenu est négligeable si le taux de sondage est faible.

2.3.2 Méthode BBH simplifiée

La simplification passe par une modification du principe de plug-in : $\theta(M)$ est simplement estimé par $\theta(\hat{M})$. La population U^* est constituée en dupliquant $p = [N/n - 1/2]$ fois chaque unité de S, puis en sélectionnant un échantillon de taille $r = N - np$ dans S. On sélectionne ensuite un rééchantillon S^* de taille n dans U^* par sondage aléatoire simple. A chaque itération, une nouvelle population U^* est constituée et un échantillon S^* est prélevé dans U^* :

on évite donc le double Bootstrap, i.e. les rééchantillonnages multiples dans chaque population U^*.

Soit $\bar{y}^*$ la moyenne simple calculée sur le rééchantillon S^*.

Propriété 2.7.

$$E(\bar{y}^*|S) = \bar{y}$$

$$V(\bar{y}^*|S) = \left(\frac{r}{N}\right)^2 \frac{1-r/N}{r} s_y^2 + \left[1 - \frac{r}{N}\left(1 - \frac{r-1}{N-1}\right)\right] \frac{N(n-1)}{n(N-1)} v(\bar{y})$$

Démonstration. On note a_k le nombre (aléatoire) de réplications de l'individu k de S. Alors :

$$E(\bar{y}^*|U^*, S) = \frac{1}{N} \sum_{k \in S} a_k y_k$$

$$V(\bar{y}^*|U^*, S) = \frac{1-f}{n} \frac{1}{N-1} \sum_{k \in S} a_k (y_k - \bar{y_a})^2$$

avec $\bar{y_a} = N^{-1} \sum_{k \in S} a_k y_k$.

Le terme $E(V(\bar{y}^*|U^*, S)|S)$ est donné dans 2.2.4, et $V(E(\bar{y}^*|U^*, S)|S)$ se calcule à l'aide de la formule de variance d'un estimateur de moyenne pour un sondage aléatoire simple.

□

Cette méthode de Bootstrap donne donc une estimation de variance biaisée, même asymptotiquement. Cependant, ce biais reste limité si le taux de sondage est faible. Notons au passage que ce terme de biais peut encore être diminué, en complétant les p duplications de chaque individu de S par un échantillon sélectionné dans S et équilibré sur des variables auxiliaires disponibles (voir le chapitre 4). L'utilisation d'un sondage aléatoire simple, qui apparaît ici de façon naturelle, revient à équilibrer sur la variable constante.

2.3.3 Méthode BWO calée

Notre méthode calée s'inspire du Bootstrap pondéré de Bertail and Combris (1997). Les différentes variantes de la méthode de Gross présentées ci-dessus peuvent être vues comme des cas particuliers. Cette méthode cherche à capter le bon estimateur de variance dans le cas linéaire, en évitant le double

Bootstrap de Booth et al. (1994) et Presnell and Booth (1994).

FIG. 2.1 – Méthode de Bootstrap unifiée pour le sondage aléatoire simple

---

On souhaite estimer la variance de $\theta(\hat{M})$. La procédure Bootstrap est la suivante :

Etape 1. On génère un vecteur aléatoire de n poids (entiers) $\mathbf{a} = (a_1, \ldots, a_n)$, suivant une distribution $\mathbb{L}$ fixée a priori. Le poids a_k est affecté à l'individu k de s.

Etape 2. On constitue une pseudo-population U^* en dupliquant a_k fois chaque individu k de s.

Etape 3. On prélève un échantillon s^* dans U^* selon un SAS de taille n.

Etape 4. On répète les étapes 1 à 3 un grand nombre de fois (disons B). On prend comme estimateur de la variance de $\hat{\theta}$:

$$\frac{1}{B}\sum_{i=1}^{B}\left(\hat{\theta}(s_i^*) - \frac{1}{B}\sum_{j=1}^{B}\hat{\theta}(s_j^*)\right)^2$$

Notons que cette méthode revient à affecter les individus de l'échantillon de poids aléatoires $\mathbf{W} = (w_1, \ldots, w_n)$ suivant une loi hypergéométrique multivariée $\mathbb{H}(n; \mathbf{a})$, où le vecteur $\mathbf{a}$ est lui-même aléatoire et distribué selon la loi $\mathbb{L}$.

Dans le cas où $\frac{N}{n}$ est entier, on retrouve la méthode de Gross avec la distribution dégénérée $\mathbb{L}$ définie par :

$$\mathbf{a} = (\frac{N}{n}, \ldots, \frac{N}{n}) \text{ avec probabilité 1}$$

On retrouve la généralisation de Bickel and Freedman (1984) et Chao and Lo (1985) avec la distribution $\mathbb{L}$ définie par :

$$\mathbf{a} = \begin{cases} \left(\left[\frac{N}{n}\right], \ldots, \left[\frac{N}{n}\right]\right) & \text{avec probabilité } \alpha \\ \left(\left[\frac{N}{n}\right] + 1, \ldots, \left[\frac{N}{n}\right] + 1\right) & \text{avec probabilité } 1 - \alpha \end{cases}$$

α étant choisi, quand cela est possible de façon à retrouver l'estimateur sans biais de variance dans le cas linéaire, et sinon de façon à retrouver le même estimateur de variance que dans le cas où $\frac{N}{n}$ est entier.

La construction de U^* proposée par Presnell and Booth (1994) revient à choisir $\mathbb{L}$ comme étant une distribution multinomiale $\mathbb{M}\left(N; \frac{1}{n}, \ldots, \frac{1}{n}\right)$.
On retrouve la variante de Booth et al. (1994) avec :

$$\mathbf{a} = \left(\left[\frac{N}{n}\right], \ldots, \left[\frac{N}{n}\right]\right) + \mathbb{H}\left(N - n\left[\frac{N}{n}\right]; 1, \ldots, 1\right)$$

En introduisant des contraintes sur les moments de $\mathbb{L}$, on peut retrouver l'estimateur sans biais de variance dans le cas linéaire. On impose les contraintes suivantes :

$\rightarrow$ Afin de se caler sur la taille de la population d'origine, on impose que U^* soit également de taille N, c'est à dire que $\sum_{i=1}^{n} a_i = N$
$\rightarrow$ Pour des raisons de symétrie, on impose que la loi $\mathbb{L}$ soit échangeable. Si $\mathbb{L}$ désigne une loi discrète dont le support est inclus dans $\mathbb{R}^n$, on dira que la loi $\mathbb{L}$ est échangeable si pour tout vecteur aléatoire X suivant la loi $\mathbb{L}$

$$\mathbb{P}(X = (a_1, \ldots, a_n)) = \mathbb{P}(X = (a_{\sigma 1}, \ldots, a_{\sigma_n}))$$

où $a = (a_1, \ldots, a_n)$ désigne un vecteur quelconque et σ une permutation quelconque de $\{1, \ldots, n\}$.

a - Calage sur l'estimateur de variance d'un total

Un bon algorithme Bootstrap devrait restituer l'estimateur sans biais de la variance dans le cas linéaire ; autrement dit, on souhaite que :

$$V(\bar{y}^*|S) = v(\bar{y})$$

Théorème 2.8. *On suppose que* **a** *est échangeable, avec* $\sum_{k=1}^{n} a_k = N$. *Alors si*

$$V(a_1) = \frac{N^2(1-f)^2}{n^2(n-1)}$$

on a

$$V(\bar{y}^*|S) = v(\bar{y})$$

Démonstration. On a :

$$\begin{array}{rcl} \bar{y}^* - \bar{y} & = & \frac{1}{n}\sum_{k\in s}(w_k - 1)y_k \\ \Rightarrow \quad V(\bar{y}^*|S) & = & \frac{1}{n^2}\left(\sum_{k\in s} V(w_k)y_k^2 + \sum_{k\neq l\in s} Cov(w_k, w_l)y_k y_l\right) \end{array}$$

Les a_k étant échangeables, les w_k le sont également et la condition $\sum_{k=1}^{n} w_k = n$ impose :

$$k \neq l \Rightarrow Cov(w_k, w_l) = -\frac{1}{n-1}V(w_1)$$

On a donc :

$$\begin{array}{rcl} V(\bar{y}^*|S) & = & \frac{1}{n^2}V(w_1)\sum_{k\in s} y_k^2 - \frac{1}{n^2(n-1)}V(w_1)\sum_{k\neq l\in s} y_k y_l \\ & = & \frac{1}{n}\ s_y^2\ V(w_1) \end{array}$$

D'autre part :

$$V(w_1) = E((w_1 - 1)^2) = E(w_1(w_1 - 1) - (w_1 - 1)) = E(w_1(w_1 - 1))$$

D'après Johnson et al. (1997), page 173 :

$$E(w_1(w_1 - 1)|\mathbf{a}) = \frac{n(n-1)}{N(N-1)}a_1(a_1 - 1)$$

Il reste à remarquer que

$$V(a_1) = \frac{N^2(1-f)^2}{n^2(n-1)}$$

est équivalent à

$$E(a_1(a_1 - 1)) = \frac{N(N-1)}{n(n-1)}(1-f)$$

On a donc

$$V(w_1) = 1 - f$$

et

$$V(\bar{y}^*|S) = v(\bar{y})$$

□

Pratiquement, il n'est pas toujours possible de se placer sous les conditions du théorême précédent. Si $n^2 < N$, on montre qu'il est possible de générer les a_k de la façon suivante :

$$\mathbf{a} \sim \mathbb{M}(N; p_1, \ldots, p_n)$$

$$\text{où } \mathbf{p} = (p_1, \ldots, p_n) \sim Dirichlet(\theta_1, \ldots, \theta_1) \text{ avec } \theta_1 = \frac{n - 2 + f}{1 - nf}$$

On peut également montrer que si $n^2 \geq N$, cette méthode ne permet pas de se placer sous les conditions du théorème 2.8.

La condition suivante est un peu moins restrictive : on peut générer des poids $\mathbf{a}$ ayant les propriétés voulues si

$$\frac{N}{n} - \left[\frac{N}{n}\right] \leq \frac{1}{n-1}\left(\frac{N}{n} - 1\right)$$

Cette condition est toujours vérifiée si $n^2 < N$. Dans le cas où $n^2 \geq N$ et si la condition précédente est vérifiée, on procède de la façon suivante : soit $K = \left[\frac{N}{n}\right]$ que l'on suppose différent de $\frac{N}{n}$. On note j l'entier tel que $N = (n-1)K + j$. Alors soient $\mathbf{b}$ et $\mathbf{c}$ deux vecteurs aléatoires, de loi échangeable, tels que

$$\mathbb{P}(b_1 = K, \ldots, b_{n-1} = K, b_n = j) = \frac{1}{n}$$

$$\mathbb{P}(c_1 = K - 1, \ldots, c_{n-1} = K - 1, c_n = j + (n-1)) = \frac{1}{n}$$

Alors soit le vecteur aléatoire

$$\mathbf{a} = \begin{cases} \mathbf{b} & \text{avec probabilité } \alpha \\ \mathbf{c} & \text{avec probabilité } 1 - \alpha \end{cases}$$

où

$$\alpha = \frac{\left(\frac{N}{n} - k + 1\right)^2 - \frac{1}{(n-1)^2}\left(\frac{N}{n} - 1\right)^2}{2\left(\frac{N}{n} - K\right) + 1}.$$

Notons également qu'il existe certains cas particuliers pour lesquels on peut toujours trouver un vecteur $\mathbf{a}$ adéquat :

1. Si $n = 2$, on peut prendre $\mathbf{a}$ de loi :

$$\mathbb{P}(a_1 = N - 1, a_2 = 1) = \mathbb{P}(a_1 = 1, a_2 = N - 1) = \frac{1}{2}$$

2. Si $N^2 = n$, on peut prendre $\mathbf{a}$ échangeable avec :

$$\mathbb{P}(a_1 = n + 1, \ldots, a_{n-1} = n + 1, a_n = 1) = \frac{1}{n}$$

3. Si $\frac{N}{n}$ est entier, on obtient $\mathbf{a}$ en mélangeant le vecteur échangeable $\mathbf{b}$ vérifiant

$$\mathbb{P}(b_1 = 1, \ldots, b_{n-1} = 1, b_n = N - n + 1) = \frac{1}{n}$$

avec le vecteur constant

$$\left(\frac{N}{n}, \ldots, \frac{N}{n}\right)$$

avec probabilité

$$\alpha = \left(\frac{1}{n-1}\right)^2$$

Ces méthodes particulières conduisent cependant à générer des pseudo-populations très variables. Bien que sans biais, l'estimateur de variance peut dans ce cas avoir une variance importante et devenir très instable. Ces cas particuliers ont donc un intérêt assez limité, et il semble que le seul cas où notre méthode puisse se révéler intéressante est celui où $n^2 < N$.

Remarque 2.1.

1. *Sous les hypothèses de la proposition précédente, on a :*

$$V_*(E(\bar{y}_*|\mathbf{a}, s)) = \frac{1 - f}{n - 1}\widehat{V}(\bar{y})$$

C'est un terme d'ajustement, de l'ordre de n fois plus petit que $\widehat{V}(\bar{y})$.

2. *Si* $\frac{N}{n}$ *est entier, on ne retrouve pas la méthode proposée par Gross.*

b - Calage sur l'estimateur du moment d'ordre 3

On notera ici μ_3 le moment centré d'ordre 3, c'est à dire que pour une variable aléatoire X

$$\mu_3(X) = E\left(X - E(X)\right)^3.$$

Théorème 2.9. *On suppose que* **a** *est échangeable, avec* $\sum_{k=1}^{n} a_k = N$. *Alors si*

$$E(a_k(a_k-1)(a_k-2)) = \frac{N(N-1)(N-2)}{n(n-1)(n-2)}(1-f)(1-2f)$$

on a

$$\mu_3(\bar{y}_*|S) = \hat{\mu_3} = \frac{(1-f)(1-2f)}{(n-1)(n-2)}\ \frac{1}{n}\sum_{k\in s}(y_k - \bar{y})^3$$

Démonstration. Notons tout d'abord la formule suivante :

$$\left(\sum_{k\in s} z_k\right)^3 = \sum_{k\in s} z_k^3 + 3\sum_{\substack{k,\,l\,\in\, s\\ l\,\neq\, k}} z_k^2 z_l + \sum_{\substack{k,\,l,\,m\,\in\, s\\ l\,\neq\, k\\ m\,\neq\, k,\,l}} z_k z_l z_m$$

Dans la suite de la preuve, on notera simplement

$$\sum_{l\neq k\in s} z_k^2 z_l \text{ pour } \sum_{\substack{k,\,l\,\in\, s\\ l\,\neq\, k}} z_k^2 z_l$$

et

$$\sum_{m\neq l\neq k\in s} z_k z_l z_m \text{ pour } \sum_{\substack{k,\,l,\,m\,\in\, s\\ l\,\neq\, k\\ m\,\neq\, k,\,l}} z_k z_l z_m$$

On a :

$$\begin{aligned} n^3(\bar{y}^* - \bar{y})^3 &= \sum_{k\in s}(w_k-1)^3 y_k^3 \\ &+ 3\sum_{l\neq k\in s}(w_k-1)^2(w_l-1)y_k^2 y_l \qquad (2.1)\\ &+ \sum_{m\neq l\neq k\in s}(w_k-1)(w_l-1)(w_m-1)y_k y_l y_m \end{aligned}$$

Comme $(w_k-1)^3 = w_k(w_k-1)(w_k-2)+(w_k-1)$, on a $E((w_k-1)^3|S) = E(w_k(w_k-1)(w_k-2)|S)$. En utilisant les formules de Johnson et al. (1997), page 173 :

$$E(w_k(w_k-1)(w_k-2)|\mathbf{a},S) = \frac{n(n-1)(n-2)}{N(N-1)(N-2)}a_k(a_k-1)(a_k-2)$$

On en déduit que :

$$E((w_k-1)^3|S) = (1-f)(1-2f)$$

Les a_k étant échangeables, les w_k le sont également et la condition $\sum_{k\in s}(w_k-1)=0$ impose :

$$k\neq l \Rightarrow E\left((w_k-1)^2(w_l-1)|S\right) = -\frac{1}{n-1}E\left((w_k-1)^3|S\right)$$

$$k\neq l\neq m \Rightarrow E\left((w_k-1)(w_l-1)(w_m-1)|S\right) = \frac{2}{(n-1)(n-2)}E\left((w_k-1)^3|S\right)$$

Donc :

$$\begin{aligned}
E((\bar{y}^*-\bar{y})^3|S) &= \tfrac{(1-f)(1-2f)}{n^3}\left(\textstyle\sum_{k\in s}y_k^3 - \tfrac{3}{n-1}\sum_{l\neq k\in s}y_k^2y_l\right. \\
&+ \left.\tfrac{2}{(n-1)(n-2)}\textstyle\sum_{m\neq l\neq k\in s}y_ky_ly_m\right) \\
&= \tfrac{(1-f)(1-2f)}{(n-1)(n-2)}\left(\tfrac{(n-1)(n-2)}{n^3}\textstyle\sum_{k\in s}y_k^3 - 3\tfrac{n-2}{n^3}\sum_{l\neq k\in s}y_k^2y_l\right. \\
&+ \left.\tfrac{2}{n^3}\textstyle\sum_{m\neq l\neq k\in s}y_ky_ly_m\right) \\
&= \tfrac{(1-f)(1-2f)}{(n-1)(n-2)}\ \tfrac{1}{n}\textstyle\sum_{k\in s}(y_k-\bar{y})^3 \\
&= \hat{\mu_3}
\end{aligned}$$

□

Pour être sous les hypothèses du théorême précédent, on peut générer des poids a_k de la façon suivante :

$$\mathbf{a}\sim\mathbb{M}(N;p_1,\ldots,p_n)$$

$$\mathbf{p} = (\frac{1}{n},\ldots,\frac{1}{n})\wedge_\alpha Dirichlet(\theta_1,\ldots,\theta_1)$$

$$\text{avec } \theta_1 = \frac{n-4+2f}{2-fn} \text{ et } \alpha = 1 - \frac{(n-2)^2-(2-fn)}{2-fn}\frac{1-fn}{(n-1)^2}$$

Il faut pour celà que : $n \geq 4$ et $n^2 < N$. Notons que si $n^2 = N$, on retrouve la variante de Booth et Presnell.

Bien que rien ne démontre théoriquement que cela améliore les performances d'une méthode de Bootstrap, la recherche d'un calage sur l'estimateur sans biais du moment d'ordre 3 apparaît comme un critère de qualité dans une partie de la littérature. En particulier, Rao and Wu (1988) et Sitter (1992b) justifient que leurs méthodes de Bootstrap respectives peuvent, sous certaines conditions, vérifier cette propriété. Si le théorême précédent montre que cette propriété peut également être obtenue avec une approche de type BWO, des simulations montrent que cela conduit en pratique à fortement complexifier l'algorithme de Bootstrap et donc le temps d'exécution pour un gain marginal, voire nul. Nous ne considérerons donc dans les simulations que la méthode plus simple impliquant un calage sur le moment d'ordre 2.

2.3.4 Simulations

Nous utilisons une population artificielle, notée POP, de 1 000 individus. Cette population contient trois variables notées x, y et z. Elles sont générées à l'aide d'une loi exponentielle, de façon à ce que la corrélation entre x et y soit approximativement de 0.2 et que la corrélation entre x et z soit approximativement de 0.4.

Nous réalisons trois séries de simulations, mettant en oeuvre un sondage aléatoire simple de taille égale à 30, 60 et 90 respectivement. Nous testons quatre méthodes de Bootstrap, à savoir :

1. le Bootstrap naïf,
2. la méthode BWO tronquée,
3. la méthode BBH simplifiée,
4. la méthode BWO calée.

Pour la dernière méthode, on se cale uniquement sur le moment d'ordre 2 (i.e., sur l'estimateur de variance sans biais dans le cas linéaire, et non sur l'estimateur sans biais du moment d'ordre 3). Cette méthode n'est testée que pour la première simulation (la condition $n^2 < N$ n'étant pas vérifiée pour

les autres).

On s'intéresse aux paramètres suivants : totaux des variables x et y, ratio du total de x et du total de y, coefficient de corrélation entre les variables x et y, médianes des variables x et y. La précision des estimateurs de totaux est dérivée de façon exacte à l'aide des formules usuelles pour un sondage aléatoire simple. Pour l' estimateur $\hat{\theta}$ d'un autre paramètre, on réalise un grand nombre (noté B) de simulations pour obtenir B échantillons indépendants $S_1, \ldots, S_B$, sur lesquels les estimateurs $\hat{\theta}_1, \ldots, \hat{\theta}_B$ sont calculés. En vertu de la loi des grands nombres,

$$V_{sim}(\hat{\theta}) = \frac{1}{B-1} \sum_{b=1}^{B} \left(\hat{\theta}_b - \hat{\theta}_{\cdot} \right)^2$$

tend quand B augmente vers la vraie valeur $V(\hat{\theta})$ de la variance, avec

$$\hat{\theta}_{\cdot} = \frac{1}{B} \sum_{b=1}^{B} \hat{\theta}_b.$$

On utilisera $V_{sim}(\hat{\theta})$ comme approximation de la vraie variance de $\hat{\theta}$, avec $B = 20\ 000$.

Pour les méthodes de Bootstrap, l'estimation de variance et les intervalles de confiance sont produits à l'aide de 1 000 échantillons indépendants, pour chacun desquels 600 rééchantillons Bootstrap sont sélectionnés. On rappelle que la méthode des percentiles est utilisée pour la production d'intervalles de confiance. Un **indicateur de stabilité** de l'estimation de variance, pour l'estimateur $\theta(\hat{M})$ du paramètre $\theta = \theta(M)$, est donné par le coefficient de variation des estimations Bootstrap de variance

$$\frac{\sqrt{V(\hat{V}(\hat{\theta}))}}{E(\hat{V}(\hat{\theta}))}$$

où $\hat{V}(.)$ désigne l'estimation Bootstrap de variance. Ces indicateur est ici calculé en remplaçant $E(.)$ par la moyenne empirique calculée sur les 1 000 échantillons, et $V(.)$ par la variance empirique correspondante. On donne également la longueur standardisée de l'intervalle de confiance, définie comme

le rapport entre la longueur moyenne de l'intervalle de confiance Bootstrap et la longueur de l'intervalle de confiance donné par l'approximation normale.

Le tableau 2.1 donne pour les différentes méthodes testées l'écart entre l'approximation Bootstrap de précision et la précision donnée par les simulations. Quand elle est applicable, la méthode du BWO calé n'amène pas d'amélioration notable dans l'estimation de précision par rapport à une méthode telle que le BWO tronqué. Les méthodes du BBO tronqué et du BBH approché donnent des résultats assez semblables, et sont généralement moins biaisées pour l'estimation de variance que le Bootstrap naïf (exception faite du coefficient de corrélation).

Les tableaux 2.2, 2.3 et 2.4 donnent, pour plusieurs tailles d'échantillons, les performances des différentes méthodes dans la production d'intervalles de confiance. Toutes les méthodes ont des comportements relativement voisins. Les taux de couverture théoriques sont raisonnablement bien respectés, sauf peut-être pour le coefficient de corrélation.

2.4 Conclusion

Dans ce chapitre, nous donnons un résumé des principales méthodes de Bootstrap développées dans le cadre d'un sondage aléatoire simple. Le bon comportement des méthodes de type BWO a été souligné dans la littérature, voir Presnell and Booth (1994) et Davison and Hinkley (1997) ; l'existence d'un principe de plug-in rend ces méthodes plus facilement applicables, particulièrement pour l'utilisation de techniques avancées de type t-Bootstrap.

L'inconvénient des méthodes de type BWO est un volume de calcul important, qui rend la mise en oeuvre d'une méthode telle que le t-Bootstrap presque impossible pour des fonctionnelles non linéaires. Nous proposons donc plusieurs méthodes de type BWO, simplifiées pour permettre la production rapide d'estimations de variance et d'intervalles de confiance. Des simulations montrent que, si le taux de sondage reste faible, ces méthodes ainsi que le Bootstrap naïf présentent des performances analogues pour la production d'intervalles de confiance. Le BWO tronqué a l'avantage d'être simple d'utilisation, et de fournir une estimation de variance plus fidèle. Nous conseillons donc son utilisation pour un taux de sondage faible. Si le taux de

sondage augmente, nous suggérons d'utiliser plutôt la méthode BBH exacte.

TAB. 2.1 – Ecart relatif de l'estimation de variance pour 4 méthodes de Bootstrap dans le cas d'un SAS

	Ecart relatif (%)			
	Bootstrap naïf	BWO tronqué	BBH approché	BWO calé
		Echantillon de taille 30		
Total de x	-0.4	-3.5	-0.2	-3.1
Total de y	-0.6	-3.9	-0.6	-3.3
Ratio	14.1	10.1	13.6	10.6
Corrélation	-13.7	-16.1	-13.5	-16.1
Médiane de x	26	22	26.7	22.3
Médiane de y	32.3	27.7	32.7	28.3
		Echantillon de taille 60		
Total de x	4.5	-1.4	-1.6	
Total de y	5.1	-1.4	-1.0	
Ratio	8	1.6	1.4	
Corrélation	-8.8	-13.8	-13.9	
Médiane de x	24.9	18.2	18.0	
Médiane de y	28.9	20.0	20.7	
		Echantillon de taille 90		
Total de x	8.6	-1.5	-1.0	
Total de y	8.6	-0.9	-1.2	
Ratio	8.8	-0.9	-0.8	
Corrélation	-1.7	-9.9	-9.7	
Médiane de x	28.7	17.6	18.3	
Médiane de y	24.4	14.0	13.2	

TAB. 2.2 – Taux de couverture, Longueurs standardisées et Stabilité de 3 méthodes de Bootstrap pour un SAS de taille 30

	Taux de couverture								
	2.5 %			5 %			Longueur standardisée		
Méthode	$L^{(a)}$	$U^{(b)}$	$L+U^{(c)}$	L	U	$L+U$	2.5 %	5 %	Stabilité
				Total de la variable x					
Bootstrap naïf	1.7	6.6	8.3	2.7	11.2	13.9	1.15	0.97	0.42
BWO tronqué	1.6	7.2	8.8	2.9	10.7	13.6	1.14	0.96	0.41
BBH approché	1.7	7.1	8.8	2.7	9.9	12.6	1.16	0.97	0.42
BWO calé	1.8	7.1	8.9	3.1	10.9	14.0	1.14	0.96	0.41
				Total de la variable y					
Bootstrap naïf	2.3	6.0	8.3	3.8	10.0	13.8	1.15	0.97	0.44
BWO tronqué	2.5	6.6	9.1	4.1	10.3	14.4	1.13	0.96	0.43
BBH approché	2.1	6.6	8.7	4.1	10.2	14.3	1.15	0.97	0.44
BWO calé	2.1	6.6	8.7	4.2	10	14.2	1.14	0.96	0.43
				Ratio					
Bootstrap naïf	4.5	2.9	7.4	6.6	5.7	12.3	1.20	1.00	0.74
BWO tronqué	4.7	3.0	7.7	6.6	6.1	12.7	1.18	0.98	0.72
BBH approché	4.7	2.6	7.3	6.8	5.0	11.8	1.20	1.00	0.72
BWO calé	4.4	3.0	7.4	6.7	6.0	12.7	1.19	0.99	0.72
				Coefficient de corrélation					
Bootstrap naïf	4.1	5.6	9.7	6.0	10.0	16.0	1.06	0.90	0.40
BWO tronqué	3.9	6.2	10.1	6.2	10.1	16.3	1.05	0.89	0.40
BBH approché	4.5	5.6	10.1	6.7	9.1	15.8	1.07	0.90	0.40
BWO calé	4.0	5.9	9.9	6.9	10.3	17.2	1.05	0.89	0.40
				Médiane de la variable x					
Bootstrap naïf	2.2	4.5	6.7	3.7	6.7	10.4	1.17	0.99	0.84
BWO tronqué	2.4	4.7	7.1	3.6	8.0	11.6	1.16	0.96	0.82
BBH approché	2.2	4.5	6.7	3.7	7.3	11.0	1.17	1.00	0.84
BWO calé	2.5	4.5	7.0	3.8	7.2	12.0	1.16	0.97	0.82
				Médiane de la variable y					
Bootstrap naïf	2.4	3.7	6.1	4.6	7.1	11.7	1.22	1.02	0.97
BWO tronqué	2.3	4.3	6.6	4.6	7.6	12.2	1.20	0.99	0.95
BBH approché	2.4	3.9	6.3	4.4	7.3	11.7	1.22	1.02	0.97
BWO calé	2.6	3.9	6.5	4.4	7.9	12.3	1.20	0.99	0.94

Note de lecture

(a) % des IC pour lesquels le paramètre se situe en-dessous de l'IC (% théorique : 2.5 %)

(b) % des IC pour lesquels le paramètre se situe au-dessus de l'IC (% théorique : 2.5 %)

(c) % des IC pour lesquels le paramètre se situe en-dehors de l'IC (% théorique : 5 %)

TAB. 2.3 – Taux de couverture, Longueurs standardisées et Stabilité de 3 méthodes de Bootstrap pour un SAS de taille 60

	Taux de couverture								
	2.5 %			5 %			Longueur standardisée		
Méthode	L	U	L+U	L	U	L+U	2.5 %	5 %	Stabilité
				Total de la variable x					
Bootstrap naïf	1.5	4.0	5.5	2.6	6.3	8.9	1.20	1.00	0.31
BWO tronqué	1.3	4.1	5.4	3.4	6.7	10.1	1.16	0.98	0.29
BBH approché	1.5	4.0	5.5	3.2	6.9	10.1	1.16	0.98	0.29
				Total de la variable y					
Bootstrap naïf	4.1	2.8	6.9	4.6	6.3	10.9	1.20	1.01	0.32
BWO tronqué	4.1	3.2	7.3	4.4	7.0	11.4	1.17	0.98	0.31
BBH approché	4.3	3.1	7.4	4.5	6.7	11.2	1.17	0.98	0.30
				Ratio					
Bootstrap naïf	2.8	1.9	4.7	5.4	3.9	9.3	1.21	1.00	0.44
BWO tronqué	3.2	2.4	5.6	6.0	4.8	10.8	1.17	0.98	0.42
BBH approché	3.1	2.3	5.4	5.4	4.7	10.1	1.17	0.98	0.41
				Coefficient de corrélation					
Bootstrap naïf	2.2	4.9	7.1	4.9	7.9	12.8	1.10	0.93	0.38
BWO tronqué	2.6	5.3	7.9	5.4	8.4	13.8	1.07	0.91	0.36
BBH approché	2.3	5.3	7.6	5.0	8.6	13.6	1.07	0.91	0.36
				Médiane de la variable x					
Bootstrap naïf	1.6	2.8	4.4	2.4	5.7	8.1	1.21	1.02	0.75
BWO tronqué	1.5	3.6	5.1	2.6	6.2	9.0	1.17	0.99	0.73
BBH approché	1.2	2.9	4.1	2.8	6.9	9.7	1.17	0.99	0.73
				Médiane de la variable y					
Bootstrap naïf	1.7	3.4	5.1	3.5	5.3	8.8	1.25	1.04	0.76
BWO tronqué	2.1	3.5	5.6	3.8	6.3	10.1	1.20	1.01	0.70
BBH approché	2.1	3.5	5.6	3.9	6.2	10.1	1.20	1.01	0.71

Note de lecture : cf. tableau 2.2

TAB. 2.4 – Taux de couverture, Longueurs standardisées et Stabilité de 3 méthodes de Bootstrap pour un SAS de taille 90

	Taux de couverture								
	2.5 %			5 %			Longueur standardisée		
Méthode	L	U	L+U	L	U	L+U	2.5 %	5 %	Stabilité
				Total de la variable x					
Bootstrap naïf	1.4	3.7	5.1	2.7	6.4	9.1	1.23	1.03	0.26
BWO tronqué	1.5	4.2	5.7	3.1	7.6	10.7	1.16	0.98	0.24
BBH approché	1.8	4.4	5.2	3.0	7.3	10.3	1.17	0.98	0.24
				Total de la variable y					
Bootstrap naïf	1.8	4.4	5.2	3.2	6.7	9.9	1.23	1.03	0.27
BWO tronqué	2.0	4.4	6.4	4.6	7.8	12.4	1.17	0.99	0.25
BBH approché	2.3	4.8	7.1	4.5	7.2	11.7	1.17	0.99	0.24
				Ratio					
Bootstrap naïf	2.3	2.1	4.4	4.4	4.9	9.3	1.22	1.02	0.36
BWO tronqué	3.0	2.9	5.9	5.6	5.4	11.0	1.16	0.97	0.33
BBH approché	2.8	2.7	5.5	5.2	5.9	11.1	1.16	0.97	0.33
				Coefficient de corrélation					
Bootstrap naïf	2.9	3.5	6.4	5.0	7.6	12.6	1.15	0.97	0.37
BWO tronqué	3.2	4.0	7.2	6.0	8.2	14.2	1.10	0.93	0.34
BBH approché	3.6	4.4	8.0	6.2	8.2	14.4	1.10	0.93	0.34
				Médiane de la variable x					
Bootstrap naïf	1.7	2.6	4.3	3.9	4.6	8.4	1.26	1.06	0.69
BWO tronqué	2.2	2.7	4.9	4.4	5.9	10.3	1.20	1.01	0.65
BBH approché	1.9	3.0	4.9	5.1	5.7	10.8	1.20	1.01	0.66
				Médiane de la variable y					
Bootstrap naïf	1.6	3.4	5.0	3.1	7.7	10.8	1.25	1.04	0.63
BWO tronqué	1.7	4.4	6.1	4.2	8.0	12.2	1.19	0.99	0.60
BBH approché	1.9	5.0	6.9	4.6	8.6	13.2	1.18	0.99	0.59

Note de lecture : cf. tableau 2.2

Bibliographie

Barbe, P. and Bertail, P. (1995). *The Weighted Bootstrap*. Springer-Verlag, New-York.

Bertail, P. and Combris, P. (1997). Bootstrap généralisé d'un sondage. *Annales d'Economie et de Statistique*, 46 :49–83.

Bickel, P. and Freedman, D. (1984). Asymptotic normality and the bootstrap in stratified sampling. *The Annals of Statistics*, 12 :470–482.

Booth, J., Butler, R., and Hall, P. (1994). Bootstrap methods for finite populations. *Journal of the American Statistical Association*, 89 :1282–1289.

Canty, A. and Davison, A. (1999). Resampling-based variance estimation for labour force surveys. *The Statistician*, 48 :379–391.

Chao, M.-T. and Lo, S.-H. (1985). A bootstrap method for finite population. *Sankhya Series A*, 47 :399–405.

Davison, A. and Hinkley, D. (1997). *Bootstrap Methods and Their Application*. Cambridge University Press, Cambridge.

Devroye, L. (1986). *Non-Uniform Random Variate Generation*. Springer-Verlag, New-York.

Efron, B. (1982). *The jackknife, the Bootstrap and Other Resampling Plans*, volume 38. ACBMS-N SIAM.

Gross, S. (1980). Median estimation in sample surveys. *ASA Proceedings of Survey Research*, pages 181–184.

Johnson, N., Kotz, S., and Balakrishnan, N. (1997). *" Discrete Multivariate Distributions*. Wiley, New-York.

Lo, A. (1992). Bayesian bootstrap clones and a biometry function. *Sankhya Series A*, 53 :320–333.

Mac Carthy, P. and Snowden, C. (1985). The bootstrap and finite population sampling. Technical report.

Mason, D. and Newton, M. (1992). A rank statistic approach to the consistency of a general bootstrap. *Annals of Statistics*, 20 :1611–1624.

Presnell, B. and Booth, J. (1994). Resampling methods for sample surveys. Technical report.

Rao, J. and Wu, C. (1988). Resampling inference with complex survey data. *Journal of the American Statistical Association*, 83 :231–241.

Shao, J. and Tu, D. (1995). *The Jackknife and The Bootstrap*. Springer, New-York.

Sitter, R. (1992a). Comparing three bootstrap methods for survey data. *Canadian Journal of Statistics*, 20 :135–154.

Sitter, R. (1992b). A resampling procedure for complex survey data. *Journal of the American Statistical Association*, 87 :755–765.

Tillé, Y. (2001). *Théorie des sondages : échantillonnage et estimation en populations finies*. Dunod, Paris.

Wu, C. (1986). Jackknife, bootstrap and other resampling methods in regression analysis (with discussions). *Annals of Statistics*, 14 :1261–1350.

Wu, C. (1990). On the asymptotic properties of the jackknife histogram. *Annals of Statistics*, 18 :1438–1452.

Chapitre 3

Bootstrap d'un plan de sondage à probabilités inégales

L'échantillonnage à probabilités inégales est généralement utilisé lorsque l'on dispose sur la base de sondage d'une variable auxiliaire bien corrélée à la variable d'intérêt. Il est alors plus efficace d'échantillonner les unités à probabilités proportionnelles à cette variable auxiliaire, plutôt que d'échantillonner à probabilités égales.

Parmi les très nombreuses méthodes de tirage à probabilités inégales, les plans à entropie maximale tels que le tirage poissonien et le tirage réjectif ont une importance particulière. Sur le plan théorique, car on dispose d'un théorême central-limite pour ces deux plans de sondage. Sur le plan pratique également, car le tirage réjectif présente de nombreuses bonnes propriétés que nous rappellerons dans ce chapitre ; on peut notamment toujours produire une estimation sans biais de variance dans le cas linéaire. Nous proposons ici une extension de l'algorithme de Bootstrap de Booth et al. (1994) au cas des probabilités inégales, et nous montrons que cet algorithme donne une estimation consistante pour un estimateur par substitution dans le cas d'un plan de sondage à entropie maximale, ou proche de l'entropie maximale.

Le chapitre est organisé de la façon suivante. En section 1, nous effectuons quelques rappels sur l'échantillonnage à probabilités inégales et proposons deux algorithmes de Bootstrap. En section 2, nous donnons une condition générale de "bootstrappabilité" pour un plan de sondage quelconque. En section 3, nous rappelons les principales propriétés du plan poissonien, et

montrons qu'il s'agit d'un plan bootstrappable. Nous réalisons le même travail en section 4 pour le tirage réjectif, et en section 5 pour les plans de sondage proches de l'entropie maximale.

3.1 Introduction

3.1.1 Echantillonnage à probabilités inégales

On parle de plan de sondage à probabilités inégales lorsque deux unités au moins de la population ont des probabilités différentes d'être retenues dans l'échantillon. La définition des probabilités d'inclusion se fait généralement à l'aide d'une variable auxiliaire connue sur chaque individu de la population.

Soit p un plan de sondage sur U, et π_k la probabilité de sélection de l'unité $k \in U$. On suppose que π_k est définie proportionnellement à une variable auxiliaire $x > 0$:

$$\pi_k = n\frac{x_k}{t_x}.$$

Comme $\sum_{k\in U} \pi_k$ donne la taille (moyenne) d'échantillon souhaitée, n désigne cette taille moyenne d'échantillon. Notons que si certaines valeurs x_k sont fortes, cette méthode peut conduire à des $\pi_k > 1$. Dans ce cas, les π_k correspondantes sont arrondies à 1, et les probabilités d'inclusion sont recalculées.
Une formulation plus rigoureuse est donnée par Deville and Tillé (1998). Soit

$$h(z) = \sum_{k\in U} min\left(z\frac{x_k}{t_x}, 1\right).$$

Alors les probabilités d'inclusion sont données par

$$\pi_k = min\left(1, h^{-1}(n)\ \frac{x_k}{t_x}\right) \quad \forall k \in U.$$

On suppose ici sans perte de généralité que $\forall k \in U\ 0 < \pi_k < 1$.

L'échantillonnage à probabilités inégales est préférable lorsque l'on dispose d'une information sur chaque individu de la population (généralement, une mesure de taille, mais pas nécessairement) et que cette variable est bien corrélée à la variable d'intérêt. Comme le montre la propriété suivante, on obtient

alors une réduction de la variance.

Propriété 3.1. *Si le plan p est de taille fixe et que les probabilités d'inclusion sont proportionnelles à la variable x, alors :*

$$V(\widehat{t_{x\pi}}) = 0$$

De plus, pour toute variable y, la variance est la même que pour la variable de résidus, pour toute régression de la variable y sur la variable x.

Démonstration. On a :

$$\begin{aligned} \widehat{t_{x\pi}} &= \sum_{k\in S} \frac{x_k}{\pi_k} \\ &= \sum_{k\in S} \frac{x_k}{n\frac{x_k}{\sum_{l\in U} x_l}} \\ &= \frac{t_x}{n} n(S) \end{aligned}$$

où $n(S)$ désigne la taille de l'échantillon. Comme p est de taille fixe, $V(\widehat{t_{x\pi}}) = 0$. Le second résultat est immédiat.

□

Il existe de nombreux algorithmes d'échantillonnage à probabilités inégales permettant de respecter (ou de respecter approximativement) des probabilités d'inclusion d'ordre 1 fixées a priori. On trouvera un inventaire assez détaillé de ces méthodes dans Hanif and Brewer (1980) et Brewer and Hanif (1983), et plus complet dans Tillé (2006).

3.1.2 Algorithme de Bootstrap proposé

Nous définissons dans cette section un algorithme général de Bootstrap pour un échantillonnage à probabilités inégales. Cet algorithme est une généralisation de la méthode de Gross (1980), plus particulièrement de la variante proposée par Booth et al. (1994). Une méthode analogue est évoquée dans Deville (1987). Par la suite, nous examinerons la validité de cette méthode pour différents plans de sondage.

La méthode de Gross (1980), qui s'appuie sur le principe de π-estimation, consistait pour un échantillon de taille n issu d'un sondage aléatoire simple

dans une population de taille N à obtenir une pseudopopulation U^* par N/n duplications de chaque unité de l'échantillon. La variance est alors estimée par tirages répétés dans U^*. Le même principe, transposé au cas des probabilités inégales, est appliqué dans l'algorithme 3.1 : chaque unité k de l'échantillon, sélectionnée avec une probabilité d'inclusion π_k, est dupliquée $1/\pi_k$ fois pour constituer la pseudo-population U^* dans laquelle on effectue des tirages répétés. Une correction de cette méthode est rendue nécessaire par le fait que les inverses des probabilités d'inclusion π_k sont rarement entières : cette correction généralise celle de Booth et al. (1994) pour le sondage aléatoire simple.

Gestion des problèmes d'arrondi

Notons que dans l'algorithme 3.1, un problème se pose si le plan de sondage p utilisé est de taille fixe. En effet, l'étape 1 suppose le tirage d'un complément d'échantillon dans S selon le plan p avec des probabilités α_k, mais le total $\sum_{k\in S}\alpha_k$ n'est généralement pas entier. De la même façon, l'étape 2 suppose le tirage d'un rééchantillon dans U^* selon le plan p, avec des probabilités d'inclusion π_k, mais le total $\sum_{k\in U^*}\pi_k$ n'a pas non plus de raison particulière d'être entier. Ce problème est important, notamment pour l'étape 2, car si la condition de taille fixe n'est pas respectée lors du rééchantillonnage, des simulations montrent que cette procédure peut conduire à une estimation de variance totalement inconsistante.

Ce problème est moins important pour l'étape 1, car les résultats énoncés plus loin au paragraphe 3.2 supposent seulement que, lors de l'étape 1, la population U^* est complétée par un échantillon sélectionné dans S **selon une procédure qui respecte les probabilités d'inclusion** α_k (et donc, non nécessairement de taille fixe). Ce complément d'échantillon peut par exemple être sélectionné selon un tirage poissonien. Alternativement, on peut utiliser le plan de sondage p initial en modifiant légèrement les probabilités d'inclusion α_k afin que leur somme soit entière, ce qui a en pratique peu d'impact sur les résultats.

Comme nous l'avons indiqué plus haut, la situation peut être plus problématique au niveau de l'étape 2. Si les probabilités d'inclusion sont choisies

FIG. 3.1 – Bootstrap général pour un plan à probabilités inégales

Etape 1. Chaque unité k de S est dupliquée $[1/\pi_k - 1/2]$ fois, où $[. - 1/2]$ désigne la partie entière. On complète les unités ainsi obtenues d'un échantillon, sélectionné dans S selon le plan de sondage p avec des probabilités d'inclusion $\alpha_k = 1/\pi_k - [1/\pi_k - 1/2]$. On obtient ainsi une pseudo-population U^*.

Etape 2. On échantillonne dans U^* selon le plan de sondage p (et avec les probabilités π_k d'origine) pour obtenir un échantillon S^*. Soit $\hat{M}^* = \sum_{k \in S^*} \frac{\delta_k}{\pi_k}$ la mesure Bootstrap estimante calculée sur S^*, et $\theta(\hat{M}^*)$ la version Bootstrap de $\theta(M)$.

Etape 3. L'étape 2 est répétée un grand nombre de fois (disons C) pour obtenir $\theta(\hat{M}_1^*), \ldots, \theta(\hat{M}_C^*)$. On note

$$v_{boot} = \frac{1}{C-1} \sum_{i=1}^{C} \left(\theta(\hat{M}_i^*) - \hat{\theta}_\cdot^* \right)^2 \text{ avec } \hat{\theta}_\cdot^* = \frac{1}{C} \sum_{i=1}^{C} \theta(\hat{M}_i^*).$$

Etape 4. Les étapes 1 à 3 sont répétées un grand nombre de fois, disons B, pour obtenir $v_{boot}^1, \ldots, v_{boot}^B$. La variance est estimée par

$$\frac{1}{B} \sum_{i=1}^{B} v_{boot}^i.$$

proportionnellement à une variable de taille, la solution consiste à bootstrapper ce processus. Plus exactement, on redéfinit pour chaque unité de U^* la probabilité

$$\pi_k^* = n \, \frac{x_k}{\sum_{k \in U^*} x_k},$$

avec un recalcul éventuel si certaines de ces probabilités dépassent 1. Le rééchantillon est alors sélectionné dans U^* selon le plan de sondage p en respectant les probabilités π_k^*.

Quelques remarques générales

On peut noter que, dans le cas particulier où p désigne un sondage aléatoire simple, l'algorithme redonne exactement la méthode de Booth et al. (1994). Nous retenons le même principe de plug-in, à savoir qu'un paramètre $\theta(M)$ est estimé par $E\left(\theta(\hat{M}^*)|S\right)$. Ce principe, qui complique l'algorithme puisqu'il implique un double Bootstrap sur la population et l'échantillon, est rendu nécessaire par le caractère non-entier des inverses de probabilités d'inclusion π_k ; il permet d'éliminer la variance parasite générée lors de l'étape 1 de l'algorithme pour constituer U^*.

Cette variance parasite est potentiellement réduite lorsque les probabilités d'inclusion sont faibles, i.e. quand les $1/\pi_k$ sont "à peu près" entières. Dans ce cas, on peut envisager une simplification de la méthode (voir l'algorithme 3.2 au paragraphe suivant) en ne constituant qu'**une seule** pseudo-population U^* dans laquelle les rééchantillons sont tirés de façon répétée. Cette simplification peut s'avérer très utile dans des cas pratiques, où le taux de sondage est effectivement limité : l'utilisation d'un simple Bootstrap sur l'échantillon limite fortement le temps de calcul. D'autre part, cette méthode simplifiée revient à créer pour chaque individu de l'échantillon des poids Bootstrap donnant le résultat du rééchantillonnage, que l'on accole au fichier d'enquête, et qui permettent de produire ultèrieurement des estimations de précision de façon autonome. A contrario, l'utilisation de l'algorithme 3.1 rend les estimations de variance plus difficilement "transportables".

Le résultat principal que nous avons obtenu est que la méthode de Bootstrap est efficace pour un plan de sondage à entropie maximale ; ce résultat, démontré par Bickel and Freedman (1984) et Chao and Lo (1985) pour le sondage aléatoire simple (qui est le plan de taille fixe à entropie maximale pour des probabilités d'inclusion égales), sera démontré dans ce chapitre pour les tirages poissonien et réjectif, et dans le chapitre suivant dans le cas plus général d'un échantillonnage équilibré à entropie maximale. L'argument principal consiste en une approximation des probabilités d'inclusion du second ordre, initialement proposée par Hájek (1981) pour le tirage réjectif, et étendue dans ce travail à un schéma d'échantillonnage plus général.
On peut énoncer un résultat contraposé : la méthode de Bootstrap est inefficace pour un plan à entropie faible, comme le montre l'exemple suivant.

Exemple 3.1. *Soit une population de* 8 *individus sur lesquels la valeur de deux variables* x *et* y *sont supposées connues :*

	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*
x_k	*2*	*3*	*3*	*5*	*9*	*9*	*11*	*15*
y_k	*1*	*3*	*4*	*2*	*6*	*1*	*7*	*10*

On prélève dans U *un échantillon sélectionné selon un tirage systématique de taille* 2*, à probabilités égales, après avoir trié* U *selon les variables croissantes de la variable* x*. Les échantillons possibles sont*

$$\{1,5\},\{2,6\},\{3,7\},\{4,8\}$$

et sont équiprobables. La variance de l'estimateur $\bar{y}$ *de la moyenne* μ_y *est égale à*

$$\begin{aligned} V(\bar{y}) &= \tfrac{1}{4}\sum_s \left(\bar{y}(s)-\mu_y\right)^2 \\ &= 5.3125 \end{aligned}$$

Comme les inverses de probabilités d'inclusion sont entières, on peut se contenter d'utiliser la méthode de Bootstrap en ne créant qu'une seule pseudo-population U^**. Si* $s=\{1,5\}$ *est l'échantillon sélectionné, on obtient*

$$U^* = \{1,5,1,5,1,5,1,5\}$$

en dupliquant 4 *fois chaque unité de* s*. En appliquant le plan de sondage d'origine, la pseudo-population est triée selon la variable* x

$$\Rightarrow U^* = \{1,1,1,1,5,5,5,5\},$$

puis un échantillon de taille 2 *est sélectionné selon un tirage systématique à probabilités égales. Tous les échantillons éligibles sont constitués d'un dupliqué de* 1 *et d'un dupliqué de* 5*, de sorte que*

$$V(\bar{y}^*|S=s_1)=0$$

Par un raisonnement analogue, on en déduit que

$$V(\bar{y}^*|S)=0$$

De façon plus générale, on peut montrer que l'algorithme de Bootstrap n'est pas consistant pour un tirage systématique à probabilités inégales sur un fichier trié de façon informative (c'est à dire selon une variable auxiliaire, par opposition avec un fichier trié selon un ordre aléatoire).

Par la suite, nous adopterons les notations suivantes : S est un échantillon sélectionné selon un plan de sondage p, avec une probabilité π_k pour l'unité k. On note U^* une pseudo-population obtenue à partir de l'échantillon S selon l'étape 1 de l'algorithme, et S^* un rééchantillon sélectionné dans U^* selon l'étape 2. Le π-estimateur du total de la variable y calculé sur S^* est noté $\widehat{t^*_{y\pi}} = \sum_{k \in S^*} \frac{y_k}{\pi_k}$.

3.1.3 Algorithme simplifié

Comme dans le cas de l'algorithme BBH pour le sondage aléatoire simple (voir chapitre 2), l'algorithme 3.1 implique une double randomisation : B pseudopopulations sont générées aléatoirement, et dans chacune d'elles C réechantillons sont prélevés. Ce procédé est assez gourmand en temps de calcul, surtout si on le combine avec une technique d'estimation par intervalle de confiance elle-aussi chronophage, de type t-Bootstrap.

Pour cette raison, il peut être intéressant de disposer d'un algorithme simplifié, possiblement moins précis mais plus rapide. Nous avons montré au chapitre 2 que la méthode BWO tronquée constituait une bonne alternative pour le sondage aléatoire simple quand le taux de sondage est faible ; l'algorithme 3.2 présenté ci-dessous donne une généralisation au cas des probabilités inégales, que nous évaluerons conjointement avec l'algorithme 1 à l'aide de simulations.

Ce schéma présente l'avantage de ne constituer qu'une seule pseudopopulation dans laquelle on rééchantillonne. En contrepartie, les fréquences théoriques d'apparition des unités de l'échantillon dans la pseudopulation ne sont pas exactement respectées, ce qui est peu gênant si les probabilités d'inclusion sont faibles. Nous donnerons dans les paragraphes correspondants une estimation du biais provoqué par la simplification pour les différents plans de sondage traités.

FIG. 3.2 – Bootstrap simplifié pour un plan à probabilités inégales

Etape 1. Chaque unité k de S est dupliquée $[1/\pi_k]$ fois, où $[.]$ désigne l'entier le plus proche. On obtient ainsi une pseudo-population U^*.

Etape 2. On échantillonne dans U^* selon le plan de sondage p (et avec les probabilités π_k d'origine) pour obtenir un échantillon S^*. Soit $\hat{M}^* = \sum_{k \in S^*} \frac{\delta_k}{\pi_k}$ la mesure Bootstrap estimante calculée sur S^*, et $\theta(\hat{M}^*)$ la version Bootstrap de $\theta(M)$.

Etape 3. L'étape 2 est répétée un grand nombre de fois (disons C) pour obtenir $\theta(\hat{M}_1^*), \ldots, \theta(\hat{M}_C^*)$. La variance est estimée par

$$\frac{1}{C} \sum_{i=1}^{C} \left(\theta(\hat{M}_i^*) - \hat{\theta}_\cdot^* \right)^2 \text{ avec } \hat{\theta}_\cdot^* = \frac{1}{C} \sum_{i=1}^{C} \theta(\hat{M}_i^*).$$

Notons que là encore, un problème se pose si le plan p est de taille fixe est que le total $\sum_{k \in U^*} \pi_k$ n'est pas entier. Si les probabilités d'inclusion π_k sont définies proportionnellement à une variable de taille x, ces probabilités peuvent être là encore recalculées sur la population U^*, pour obtenir

$$\pi_k^* = n \, \frac{x_k}{\sum_{k \in U^*} x_k},$$

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

3.2 Un critère général de validité du Bootstrap

La propriété suivante donne une condition suffisante pour que l'algorithme 3.1 donne une estimation de variance asymptotiquement sans biais pour l'estimation d'un total.

Propriété 3.2. *Soit p un plan de sondage quelconque, de taille (moyenne) n, défini à l'aide d'un vecteur de variables auxiliaires $\mathbf{x}$ de dimension p. On suppose qu'il existe une fonctionnelle Q, linéarisable, un entier q et une fonction $f(.) : \mathbb{R}^{p+1} \mapsto \mathbb{R}^q$ tels que, pour une population U et une variable*

$y = (y_k)_{k \in U}$ *quelconques*

$$V(\widehat{t_{y\pi}}) = (1 + o(1))Q(\mathbf{t_{f(y,x)}}) \tag{3.1}$$

avec $\mathbf{t_{f(y,x)}} = \int f(y_k, \mathbf{x_k}) dM(y_k, \mathbf{x_k}) = \sum_{k \in U} f(y_k, \mathbf{x_k})$ *le vecteur des totaux de la variable vectorielle* $f(y_k, \mathbf{x_k})$, $M = \sum_{k \in U} \delta_{y_k, \mathbf{x_k}}$ *et* $o(1) \to 0$ *quand* $n \to \infty$.

Alors l'algorithme 3.1 de Bootstrap donne une estimation de variance consistante pour $\widehat{t_{y\pi}}$.

Démonstration. On utilise les notations de l'algorithme 3.1. En appliquant l'équation 3.1 à la population U^* :

$$V(\widehat{t^*_{y\pi}}|U^*) = (1 + o_p(1))Q(\mathbf{t^*_{f(y,x)}}) \tag{3.2}$$

avec $\mathbf{t^*_{f(y,x)}} = \int f(y_k, \mathbf{x_k}) dM^*(y_k)$.

Le théorême 1.9 appliqué à la fonctionnelle Q, en remplaçant M et $\hat{M}$ par $\hat{M}$ et M^* respectivement implique que

$$E\left(Q\left(\int f(y_k, \mathbf{x_k}) dM^*\right)|S\right) = Q\left(\int f(y_k, \mathbf{x_k}) d\hat{M}\right)(1 + o_p(1))$$

et le même théorême appliqué sous sa forme habituelle implique que

$$E\left(Q\left(\int f(y_k, \mathbf{x_k}) d\hat{M}\right)\right) = Q\left(\int f(y_k, \mathbf{x_k}) dM\right)(1 + o(1))$$

ce qui donne le résultat. □

La propriété suivante établit que ce résultat s'étend de façon plus générale à un estimateur par substitution quelconque.

Propriété 3.3. *Soit p un plan de sondage quelconque, de taille (moyenne) n, défini à l'aide d'un vecteur de variables auxiliaires $\mathbf{x}$ de dimension p. On suppose qu'il existe une fonctionnelle Q, linéarisable, un entier q et une fonction $f(.) : \mathbb{R}^{p+1} \mapsto \mathbb{R}^q$ tels que, pour une population U et une variable $y = (y_k)_{k \in U}$ quelconques*

$$V(\widehat{t_{y\pi}}) = (1 + o(1))Q(\mathbf{t_{f(y,x)}}) \tag{3.3}$$

avec $\mathbf{t_{f(y,x)}} = \int f(y_k, \mathbf{x_k}) dM(y_k, \mathbf{x_k}) = \sum_{k \in U} f(y_k, \mathbf{x_k})$ *le vecteur des totaux de la variable vectorielle* $f(y_k, \mathbf{x_k})$, $M = \sum_{k \in U} \delta_{y_k, \mathbf{x_k}}$ *et* $o(1) \to 0$ *quand* $n \to \infty$.

Soit θ *une fonctionnelle quelconque, vérifiant les hypothèses* $H1 - H7$. *Alors l'algorithme 1 donne une estimation consistante de variance pour l'estimateur par substitution* $\theta(\hat{M})$.

Démonstration. On rappelle que la variable linéarisée de la fonctionnelle $\theta(M)$ est égale à la fonction d'influence de $\theta(M)$ au point k. On note u_k (respectivement, $\hat{u_k}$ et u_k^*) pour $I\theta_k(M)$ (respectivement $I\theta_k(\widehat{M})$ et $I\theta_k(M^*)$). En appliquant le théorême 1.9 au rééchantillon S^* et à la pseudopopulation U^* :

$$\sqrt{n^*}(N^*)^{-\beta}(\theta(\widehat{M^*}) - \theta(M^*)) = \sqrt{n^*}(N^*)^{-\beta} \sum_{k \in U^*} u_k^*(\pi_k^{-1} 1_{k \in S^*} - 1) + o_p(1)$$

où n^* (respectivement N^*) désigne le nombre d'unités du rééchantillon S^* (respectivement de la pseudopopulation U^*).

$Var(\theta(\widehat{M^*})|U^*)$ est donc asymptotiquement équivalente à

$$Var\left(\sum_{k \in U^*} u_k^*(\pi_k^{-1} 1_{k \in S^*} - 1)\,|U^*\right) = Var\left(\sum_{k \in S^*} \frac{u_k^*}{\pi_k} \mid U^*\right).$$

La deuxième partie du théorême 1.9, utilisée en substituant $\hat{M}$ et M^* à M et $\hat{M}$ respectivement, implique que dans l'expression précédente, $u_k^* = I\theta_k(M^*)$ peut être remplacé de façon équivalente par $\widehat{u_k} = I\theta_k(\widehat{M})$. On a d'autre part

$$\begin{aligned} Var\left(\textstyle\sum_{k \in S^*} \frac{u_k^*}{\pi_k} \mid U^*\right) &= (1 + o_p(1)) Q(\mathbf{t}^*_{\mathbf{f}(\hat{\mathbf{u}}, \mathbf{x})}) \\ &= (1 + o_p(1)) Q(\textstyle\int f(\widehat{u_k}, \mathbf{x_k}) dM^*) \end{aligned}$$

et on a vu dans la propriété précédente que

$$E\left(Q\left(\int f(\widehat{u_k}, \mathbf{x_k}) dM^*\right)|S\right) = (1 + o_p(1)) Q\left(\int f(\widehat{u_k}, \mathbf{x_k}) d\hat{M}\right)$$

Par une application directe du théorême 1.9, $Q\left(\int f(\widehat{u_k}, \mathbf{x_k}) d\hat{M}\right)$ estime asymptotiquement sans biais $Q\left(\int f(u_k, \mathbf{x_k}) dM\right)$, ce qui démontre le résultat. □

Ce résultat peut s'énoncer plus simplement : pour un plan de sondage quelconque, l'algorithme de Bootstrap est consistant pour un estimateur par substitution si le π-estimateur admet une approximation de variance qui peut s'écrire comme une fonction de totaux. La définition suivante reprend la terminologie de Deville (1987).

Définition 3.1. *Un plan de sondage vérifiant les hypothèses des deux théorêmes précédents sera dit bootstrappable.*

Comme dans Sitter (1992), nous pourrions utiliser la linéarisation de Taylor pour prouver la consistance de l'algorithme pour un paramètre donné par une fonction lisse de totaux. Notons que l'approche par la fonction d'influence est plus générale, car elle implique également la consistance pour des estimateurs par substitution d'un paramètre non lisse, tel qu'un fractile.

3.3 Le tirage poissonien

3.3.1 Rappels sur le plan poissonien

Le plan poissonien consiste à échantillonner les individus de U indépendamment les uns des autres. Soit π_k la probabilité de sélection de l'unité k dans l'échantillon et I_k l'indicatrice d'appartenance à l'échantillon de l'individu k. Alors les variables aléatoires I_k sont indépendantes, et de distribution

$$\mathbb{P}(I_k = 1) = \pi_k \quad \mathbb{P}(I_k = 0) = 1 - \pi_k$$

Un échantillon poissonien peut être sélectionné de façon séquentielle à l'aide de l'algorithme suivant :

→ Etape 1 : on génère une variable aléatoire u_1 selon une loi uniforme sur $[0,1]$. Si $u_1 < \pi_1$, l'individu 1 est sélectionné dans l'échantillon, sinon il est écarté

→ Etape 2 : on génère une variable aléatoire u_2 selon une loi uniforme sur $[0,1]$, indépendamment de u_1. Si $u_2 < \pi_2$, l'individu 2 est sélectionné dans l'échantillon, sinon il est écarté

...

→ Etape N : on génère une variable aléatoire u_N selon une loi uniforme sur $[0,1]$, indépendamment de $u_1, \ldots, u_{N-1}$. Si $u_N < \pi_N$, l'individu N est sélectionné dans l'échantillon, sinon il est écarté

Dans le cas d'un tirage poissonien, le plan de sondage peut être entièrement défini. On a

$$\forall s \in U \quad p(s) = \prod_{k \in s} \pi_k \prod_{k \in U \setminus s} (1 - \pi_k)$$

Du fait de l'indépendance entre les tirages, les probabilités d'inclusion d'ordre 2 peuvent être calculées exactement :

$$\forall k \neq l \in U \quad \pi_{kl} = \pi_k \pi_l$$

La variance de $\widehat{t_{y\pi}}$ s'obtient à l'aide de la formule de variance de Horvitz-Thompson :

$$V(\widehat{t_{y\pi}}) = \sum_{k \in U} \left(\frac{y_k}{\pi_k} \right)^2 \pi_k (1 - \pi_k)$$

Dans le cas du tirage poissonien, cette variance apparaît donc comme un simple total et peut être estimée sans biais à l'aide du principe de π-estimation :

$$\widehat{V}(\widehat{t_{y\pi}}) = \sum_{k \in S} \left(\frac{y_k}{\pi_k} \right)^2 (1 - \pi_k)$$

Le plan poissonien est le plan d'entropie maximale pour des probabilités d'inclusion d'ordre 1 fixées (Hájek, 1981). Son principal inconvénient est une variance élevée, dûe à l'indépendance entre les tirages des différents individus. On obtient notamment une taille d'échantillon aléatoire :

$$V(n(S)) = V(\sum_{k \in S} \frac{\pi_k}{\pi_k}) = \sum_{k \in U} \pi_k (1 - \pi_k)$$

En raison de cette variance importante, et en dépit de sa simplicité, le plan poissonien est peu utilisé pour un échantillonnage direct. En revanche, il permet de définir des plans de sondage plus complexes appelés plans poissoniens conditionnels, dont nous parlerons dans les sections suivantes. Le plan poissonien est également utilisé pour modéliser la non-réponse totale dans les enquêtes.

3.3.2 Bootstrap pondéré d'un échantillon poissonien (Bertail and Combris, 1997)

La méthode proposée par Bertail and Combris (1997) est une adaptation du Bootstrap pondéré Barbe and Bertail (1995) au cas d'un sondage.
Soit S un échantillon poissonien, et π_k la probabilité de sélection de l'unité k. Le Bootstrap consiste ici à affecter l'individu k de l'échantillon d'un poids aléatoire noté W_k et appellé poids de rééchantillonnage. On notera

$$\widehat{t_{yW}} = \sum_{k \in S} W_k \frac{y_k}{\pi_k}$$

Les moments de ces poids sont choisis de façon à obtenir un estimateur Bootstrap sans biais d'un total, et un estimateur Bootstrap sans biais de la variance du π-estimateur d'un total.

Propriété 3.4. *Bertail and Combris (1997)*

Si les conditions suivantes sont vérifiées :

1. $E(W_k|S) = 1$
2. $V(W_k|S) = 1 - \pi_k$
3. $k \neq l \Rightarrow E(W_k W_l|S) - E(W_k|S)E(W_l|S) = 0$

alors $\widehat{t_{yW}}$ *est un estimateur sans biais de* t_y *et* $V(\widehat{t_{yW}}|S)$ *est un estimateur sans biais de* $V(\widehat{t_{y\pi}})$.

Bertail and Combris (1997) montrent la normalité asymptotique de l'estimateur bootstrappé du total, et donnent des indications sur le choix du coefficient d'asymétrie des poids W_k pour obtenir des résultats au second ordre. Ils montrent que des poids W_k vérifiant les conditions de moments de la proposition précédente peuvent être obtenus à l'aide d'un algorithme de Devroye (1986).

3.3.3 Propriétés de la méthode de Bootstrap proposée

La consistance de l'algorithme 3.1 de Bootstrap pour un estimateur par substitution dans le cas d'un tirage poissonien découle de la caractérisation donnée en section 1. En effet, on a

$$V(\widehat{t_{y\pi}}) = \sum_{k\in U} \left(\frac{y_k}{\pi_k}\right)^2 \pi_k(1-\pi_k).$$

Autrement dit, en utilisant les notations de la section précédente, le plan poissonien est bootstrappable en prenant $x_k = \pi_k$, $f(y_k, x_k) = \left(\frac{y_k}{x_k}\right)^2 x_k(1-x_k)$ et $Q(\mathbf{t_{f(y,x)}}) = \mathbf{t_{f(y,x)}}$.

Si on utilise la version simplifiée du Bootstrap sous la forme de l'algorithme 2, le biais obtenu pour l'estimation de variance de $\widehat{t_{y\pi}}$ est égal à

$$B_{boot}(\widehat{V_{boot}}(\widehat{t_{y\pi}})) = \sum_{k\in U} \pi_k \left(\left[\frac{1}{\pi_k}\right] - \frac{1}{\pi_k}\right) \left(\frac{y_k}{\pi_k}\right)^2 \pi_k(1-\pi_k), \tag{3.4}$$

et reste limité si les probabilités d'inclusion sont faibles. Par exemple, dans le cas où toutes les probabilités d'inclusion sont égales, on a

$$\begin{aligned}
\left|\frac{B_{boot}(\widehat{V_{boot}}(\widehat{t_{y\pi}}))}{V(\widehat{t_{y\pi}})}\right| &= \frac{\sum_{k\in U} \pi_k \left(\left[\frac{1}{\pi_k}\right] - \frac{1}{\pi_k}\right) \left(\frac{y_k}{\pi_k}\right)^2 \pi_k(1-\pi_k)}{\sum_{k\in U} \left(\frac{y_k}{\pi_k}\right)^2 \pi_k(1-\pi_k)} \\
&= \left|\pi\left(\left[\tfrac{1}{\pi}\right] - \tfrac{1}{\pi}\right)\right| \\
&\leq \tfrac{\pi}{2}
\end{aligned}$$

3.3.4 Simulations

Nous utilisons une population artificielle notée POP2. On génère une variable x selon une loi normale de moyenne 20 et de variance 4. Les variables y et z sont également générées à l'aide d'une loi normale, de façon à ce que la corrélation entre x et y (respectivement z) soit approximativement égale à 0.4 (respectivement 0.7). Nous créons deux variables de probabilités d'inclusion, notées π_1 et π_2, proportionnellement à la variable x, et calibrées pour obtenir un échantillon de taille (moyenne) égale à 50 et 100 respectivement.

Nous nous intéressons ici à la précision des π-estimateurs de totaux $\widehat{t_{y\pi}}$ et $\widehat{t_{z\pi}}$, de l'estimateur du ratio $\frac{\widehat{t_{y\pi}}}{\widehat{t_{z\pi}}}$, du coefficient de corrélation estimé entre les variables y et z, des estimateurs de médiane des variables y et z. La variance exacte est approchée à l'aide de 20 000 simulations indépendantes.

L'algorithme 1 de Bootstrap est testé à l'aide de 1 000 échantillons indépendants. Pour chaque échantillon s, $B = 100$ pseudopopulations $U^*_{sb}, b = 1 \dots 100$ sont générées, et dans chaque pseudopopulation U^*_{sb} $C = 30$ rééchantillons $S^*_{sbc}, c = 1 \dots 30$ sont prélevés. Les intervalles de confiance sont produits à l'aide d'une méthode de type percentile. Notons $\theta(\hat{M}^*_{sbc})$ la statistique Bootstrap calculée sur la pseudopopulation b et le rééchantillon c correspondant à l'échantillon s. L'histogramme des $BC = 3\ 000$ estimateurs $\theta(M^*_{sbc})$ est utilisé pour produire l'intervalle de confiance au niveau $1 - 2\alpha$

$$\left(\hat{\theta}^{\alpha}_{L}\ ,\ \hat{\theta}^{\alpha}_{U}\right),$$

où $\hat{\theta}^{\alpha}_{L} = \hat{F}^{-1}(\alpha)$, $\hat{\theta}^{\alpha}_{U} = \hat{F}^{-1}(1-\alpha)$ et

$$\hat{F}(x) = \frac{Card\{\theta(M^*_{sbc}) \leq x\ ;\ b\ =\ 1\ \dots\ B\ ,\ c\ =\ 1\ \dots\ C\}}{BC}.$$

On construit également des intervalles de confiance par une méthode de type t-Bootstrap, de façon analogue à Booth et al. (1994) pour le sondage aléatoire simple. Le principe consiste à estimer la distribution de la statistique $T = \frac{\theta(\hat{M})-\theta(M)}{\sqrt{\hat{V}(\theta(\hat{M}))}}$, où $\hat{V}(\theta(\hat{M}))$ désigne un estimateur de variance consistant de $\theta(\hat{M})$, par celle de la statistique Bootstrap $T^* = \frac{\theta(\hat{M}^*)-\theta(M^*)}{\sqrt{\widehat{V^*}(\theta(\hat{M}^*))}}$, où $\widehat{V^*}(\theta(\hat{M}^*))$ désigne la version Bootstrap de l'estimateur de variance. Cet estimateur est normalement obtenu en réappliquant la procédure de Bootstrap au rééchantillon S^*_{sbc} ; bien que la mise en oeuvre soit théoriquement possible, cela génère un volume de calcul très important. Pour le réduire, nous remplaçons $\widehat{V^*}(\theta(\hat{M}^*))$ par un estimateur de variance de type linéarisation. Il est difficile de produire un tel estimateur de façon automatique (voir Deville (1999)) dans le cas d'une estimation de fractile, aussi le t-Bootstrap ne sera pas utilisé dans le cas de l'estimation de médiane. Les simulations Bootstrap donnent pour l'échantillon s les valeurs $T^*_{sbc}, b = 1 \dots 100, c = 1 \dots 30$. On peut alors obtenir un intervalle de confiance au niveau $1 - 2\alpha$ pour $\theta(M)$, donné par

$$\left(\theta(\hat{M}) - T^*_{sU}\sqrt{\hat{V}(\theta(\hat{M}))}, \theta(\hat{M}) - T^*_{sL}\sqrt{\hat{V}(\theta(\hat{M}))}\right)$$

où T^*_{sL} et T^*_{sU} désignent les fractiles d'ordre α et $1-\alpha$ respectivement de l'histogramme des valeurs Bootstrap $T^*_{sbc}, b = 1 \ldots 100, c = 1 \ldots 30$.

L'algorithme 3.2 de Bootstrap est également testé à l'aide de 1 000 échantillons indépendants. Pour chaque échantillon, on constitue une pseudo-population U^* dans laquelle 1 000 rééchantillons sont prélevés. Les intervalles de confiance sont produits comme précedemment, à l'aide de la méthode des percentiles et de celle du t-Bootstrap. La méthode du t-Bootstrap est ici appliquée de façon exacte, i.e. l'algorithme 3.2 est réappliqué D fois à chaque échantillon Bootstrap afin de disposer d'un estimateur de variance pour chaque statistique Bootstrap T^*. On a utilisé ici $D = 50$.

Le tableau 3.1 donne les écarts relatifs à la vraie variance (donnée par les 20 000 simulations) pour les différentes méthodes d'estimation de précision testées. Pour l'estimation des totaux de y et de z, l'algorithme 3.1 et la linéarisation donnent une estimation de variance sans biais : l'écart à la vraie variance n'est dû qu'au nombre fini de simulations. L'algorithme 3.2 se compare favorablement aux deux autres méthodes. Pour le non linéaire, la linéarisation est moins efficace que les méthodes de Bootstrap, et l'algorithme 3.2 donne dans l'ensemble de meilleurs résultats que l'algorithme 3.1. La simplification de l'algorithme ne semble donc pas engendrer de biais significatif.

Les tableaux 3.2 et 3.3 comparent les taux de couverture effectifs des intervalles de confiance donnés par les deux méthodes de Bootstrap (percentiles et t-Bootstrap) et la linéarisation (approximation normale). On peut noter plusieurs points importants :

- La méthode des percentiles et celle du t-Bootstrap donnent des résultats comparables. On n'observe pas de gain particulier sur le taux de couverture qui justifie la complexité supplémentaire nécessitée par le t-Bootstrap. Ce dernier donne d'ailleurs des résultats médiocres pour les médianes.
- Les algorithmes 3.1 et 3.2 de Bootstrap donnent des résultats assez proches, mais l'algorithme 3.1 tend à mieux respecter les taux de couverture théoriques, notamment avec le t-Bootstrap.
- Le Bootstrap respecte mieux les taux de couverture pour les statistiques non linéaires que la technique de linéarisation.

Compte-tenu des résultats obtenus, nous ne retiendrons dans la suite que la méthode des percentiles pour la production d'intervalles de confiance avec

des techniques de Bootstrap.

TAB. 3.1 – Ecart relatif à la vraie variance pour les algorithmes 3.1 et 3.2 de Bootstrap et la méthode de linéarisation dans le cas d'un tirage poissonien

	Ecart relatif (%)		
Paramètre	Algorithme 1	Algorithme 2	Linéarisation
	Echantillon de taille 50		
Total de y	+0.77	+0.71	+0.38
Total de z	-0.11	+0.81	+0.52
Ratio	+0.42	+0.20	-4.14
Corrélation	-1.38	-0.35	-7.98
Médiane de y	+20.12	+16.85	
Médiane de z	+20.31	+16.79	
	Echantillon de taille 100		
Total de y	+0.04	+0.61	+0.43
Total de z	+0.19	-1.01	-1.12
Ratio	+0.48	+2.89	+1.02
Corrélation	-1.22	+0.67	-2.90
Médiane de y	+12.12	+15.32	
Médiane de z	+13.02	+10.90	

3.4 Le tirage réjectif

3.4.1 Rappels sur le plan réjectif

Nous avons vu précedemment que le plan poissonien était le plan de sondage à entropie maximale, pour des probabilités d'inclusion fixées au préalable. Ce plan présente des propriétés théoriques intéressantes. La normalité asymptotique du π-estimateur découle du théorême central limite dans le cas indépendant mais non identiquement distribué, voir par exemple Feller (1966).

Les probabilités d'inclusion ont une forme particulièrement simple, qui facilite l'estimation de variance. Cependant, l'indépendance dans le tirage des différentes unités engendre une perte d'efficacité. Un compromis consiste à restreindre le support du plan de sondage, en recherchant le plan de taille fixe à entropie maximale pour des probabilités d'inclusion données.

Ce plan de sondage, introduit initialement par Hájek (1964), est appelé le tirage réjectif. Soit $n = \sum_{k \in U} \pi_k$ la taille d'échantillon souhaitée. Hájek (1964) a établi que le plan réjectif pouvait être vu comme un plan poissonien de probabilités d'inclusion $p = (p_1 \ \ldots \ p_k \ \ldots \ p_N)'$, conditionné par l'obtention d'une taille d'échantillon exactement égale à n : pour cette raison, le plan réjectif est également parfois appelé tirage poissonien conditionnel. Ce vecteur $\mathbf{p}$ est unique si l'on ajoute la condition

$$\sum_{k \in U} p_k = n$$

En suivant la terminologie de Hajek, le vecteur correspondant sera appelé **vecteur des probabilités d'inclusion du plan poissonien canoniquement associé au plan réjectif**.

Dupacova (1979) a établi que pour tout jeu de probabilités d'inclusion $\boldsymbol{\pi}$, il existait un jeu de probabilités d'inclusion $\mathbf{p}$ permettant de mettre en oeuvre un plan réjectif de probabilités d'inclusion $\boldsymbol{\pi}$ selon la procédure précédente. Il est important de noter que si les probabilités d'inclusion $\boldsymbol{\pi}$ sont inégales, les deux vecteurs $\mathbf{p}$ et $\boldsymbol{\pi}$ sont toujours différents. Chen et al. (1994) ont proposé un algorithme permettant de passer d'un jeu de probabilités d'inclusion à un autre, permettant la sélection effective d'un échantillon réjectif avec des probabilités d'inclusion exactes. Cette méthode a été améliorée par Deville (2000). D'autres algorithmes de tirage réjectif ont également été proposés (Chen et al., 1994; Deville, 2000).

Le tirage réjectif est un plan de taille fixe présentant de bonnes propriétés. Il donne une variance plus faible que l'échantillonnage avec remise à probabilités inégales de même taille (Qualité, 2006). Il satisfait les conditions de Sen-Yates-Grundy et les probabilités d'inclusion d'ordre deux sont toutes strictement positives, voir par exemple Thompson (1997)), ce qui assure que l'estimateur de variance de Sen-Yates-Grundy est strictement positif et non

biaisé. Les probabilités d'inclusion d'ordre deux peuvent être calculées exactement ; cependant, ce calcul nécessite la connaissance de l'ensemble des probabilités d'inclusion d'ordre 1, qui peuvent ne pas être disponibles à l'étape de l'estimation. L'algorithme de Bootstrap que nous proposons permet d'estimer la précision d'estimateurs par substitution à l'aide des seules données de l'échantillon.

3.4.2 Résultats obtenus pour la méthode de Bootstrap

Si toutes les probabilités d'inclusion sont égales, le tirage réjectif est équivalent au sondage aléatoire simple sans remise. La méthode proposée par Booth et al. (1994) est donc un cas particulier de l'algorithme 3.1.

Dans la suite de cette section, p désigne un plan de sondage réjectif de probabilités d'inclusion $\boldsymbol{\pi}$, et q le plan poissonien canoniquement associé à p, de probabilités d'inclusion $\mathbf{p}$. Dans le cas général où les probabilités d'inclusion sont différentes, et d'inverses non nécessairement entiers, la consistance du Bootstrap repose sur une formule approchée de variance proposée par Hájek (1981).

Théorème 3.5. *Hájek (1981)*

On a :

$$\pi_k = p_k \left[1 - \frac{(\bar{\bar{p}} - p_k)(1 - p_k)}{d} + o(d^{-1}) \right]$$

où $d = \sum_{k \in U} p_k(1 - p_k)$, $\bar{\bar{p}} = d^{-1} \sum_{k \in U} p_k^2$ *et* $d\ o(d^{-1}) \to 0$ *uniformément en tout* $k \in U$.

Théorème 3.6. *Hájek (1981)*

On a :

$$\pi_{kl} = \pi_k \pi_l \left[1 - \frac{(1 - \pi_k)(1 - \pi_l)}{d_0} + o(d_0^{-1}) \right]$$

où $d_0 = \sum_{k \in U} \pi_k(1 - \pi_k)$ *et* $d_0\ o(d_0^{-1}) \to 0$ *uniformément en tout* $k \in U$.

En injectant l'approximation précédente dans la formule de variance donnée par Sen (1953) et Yates and Grundy (1953) pour un plan de taille fixe, Hajek en déduit une formule approché de variance.

Théorème 3.7. *Hájek (1964)*

Soit y une variable quelconque. Alors

$$V(\widehat{t_{y\pi}}) = (1 + o(1)) \sum_{k \in U} \frac{1 - \pi_k}{\pi_k} (y_k - R\pi_k)^2$$

où

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

et $o(1) \to 0$ *quand* $d_0 \to \infty$.

De cette approximation, Hajek déduit un estimateur asymptotiquement sans biais de la variance, égal à

$$\widehat{V}_{HAJ}(\widehat{t_{y\pi}}) = \sum_{k \in S} \frac{1 - \pi_k}{\pi_k^2} (y_k - \hat{R}\pi_k)^2$$

avec

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

Hájek (1981) justifie également que l'on peut obtenir une approximation plus serrée de la variance en approximant les quantités $\pi_k \pi_l - \pi_{kl}$ sous la forme

$$\pi_k \pi_l - \pi_{kl} \simeq c_k c_l = \frac{\pi_k (1 - \lambda_k)\ \pi_l (1 - \lambda_l)}{\sqrt{\sum_{i \in U} \pi_i (1 - \lambda_i)} \sqrt{\sum_{j \in U} \pi_j (1 - \lambda_j)}} \tag{3.5}$$

et en imposant

$$\sum_{l \in U; l \neq k} c_k c_l = \sum_{l \in U; l \neq k} \pi_k \pi_l - \pi_{kl} = \pi_k (1 - \pi_k)$$

Cela conduit à résoudre le système d'équations non linéaires

$$(1 - \lambda_k) \left[1 - \frac{\pi_k (1 - \lambda_k)}{\sum_{l \in U} \pi_l (1 - \lambda_l)} \right] = 1 - \pi_k \quad \forall k \in U$$

Deville and Tillé (2000) montrent que ce problème peut toujours être résolu de façon itérative si

$$\forall k \in U \frac{\pi_k (1 - \pi_k)}{\sum_{l \in U} \pi_l (1 - \pi_l)} \leq \frac{1}{2},$$

condition qui est vérifiée si les probabilités d'inclusion ne sont pas trop dispersées, ce qui est généralement le cas car un tirage à probabilités inégales est généralement précédé d'une phase de stratification. Si cette condition n'est pas vérifiée, Deville and Tillé (2000) préconisent de n'utiliser qu'une itération, comme le suggérait également Hájek (1981). Dans la suite de cette section, nous nous contenterons de l'approximation donnée par le théorême 3.7. Cette approximation permet d'établir la consistance de la méthode de Bootstrap dans le cas d'une fonctionnelle linéaire.

Propriété 3.8. *Soit $\theta(M) = t_y$ le total de la variable y. Pour un plan de sondage réjectif, l'estimateur Bootstrap de variance de $\theta(\widehat{M}) = \widehat{t_{y\pi}}$ est asymptotiquement non biaisé.*

Démonstration. Le théorême 3.7 assure que l'on est sous les hypothèses de la propriété 3.2, ce qui donne le résultat.

□

On en déduit également un résultat analogue pour un estimateur par substitution quelconque.

Propriété 3.9.

Soit $\theta(\widehat{M})$ l'estimateur par substitution d'une statistique $\theta(M)$. Alors sous les hypothèses H1-H7 (voir chapitre 1), l'algorithme 1 est consistant pour l'estimation de variance de $\theta(\widehat{M})$

Dans le cas du Bootstrap simplifié, il est difficile de fournir une mesure exacte du biais, même pour une statistique linéaire. Une approximation de ce biais, pour l'estimation de variance de $\widehat{t_{y\pi}}$, est donnée par

$$\begin{aligned} & B_{boot}^{app}(\widehat{V_{boot}}(\widehat{t_{y\pi}})) \\ = & \sum_{k\in U} \pi_k \left(\left[\frac{1}{\pi_k} \right] - \frac{1}{\pi_k} \right) \pi_k(1-\pi_k) \left(\frac{y_k}{\pi_k} - \frac{\sum_{l\in U} y_l(1-\pi_l)}{\sum_{l\in U} \pi_l(1-\pi_l)} \right)^2 \end{aligned} \tag{3.6}$$

et reste là encore limité si les probabilités d'inclusion sont faibles.

3.4.3 Simulations

On utilise à nouveau la population POP2 présentée à la section précédente. On utilise à nouveau les variables π_1 et π_2, calibrées pour obtenir un échantillon de taille égale à 50 et 100 respectivement. On définit également une

variable π_3, proportionnelle à x, calibrée pour obtenir un échantillon de taille 20. Nous nous intéressons ici à la précision des π-estimateurs de totaux $\widehat{t_{y\pi}}$ et $\widehat{t_{z\pi}}$, de l'estimateur du ratio $\frac{\widehat{t_{y\pi}}}{\widehat{t_{z\pi}}}$, du coefficient de corrélation estimé entre les variables y et z, des estimateurs de médiane des variables y et z. La variance exacte est approchée à l'aide de 20 000 simulations indépendantes.

L'algorithme 3.1 de Bootstrap est testé à l'aide de 1 000 échantillons indépendants. Pour chaque échantillon s, $B = 100$ pseudopopulations $U^*_{sb}, b = 1 \dots 100$ sont générées, et dans chaque pseudopopulation U^*_{sb} $C = 30$ rééchantillons $S^*_{sbc}, c = 1 \dots 30$ sont prélevés. Compte tenu des résultats de la section précédente, seule la méthode des percentiles est utilisée pour produire des intervalles de confiance.

L'algorithme 3.2 de Bootstrap est également testé à l'aide de 1 000 échantillons indépendants. Pour chaque échantillon, on constitue une pseudo-population U^* dans laquelle 1 000 rééchantillons sont prélevés. Les intervalles de confiance sont produits comme précedemment, à l'aide de la méthode des percentiles.

Le tableau 3.4 compare les estimations de variance obtenues avec les méthodes de Bootstrap et la linéarisation. Dans le cas linéaire, la linéarisation (qui correspond alors à une estimation directe) est préférable. La variance est théoriquement estimée sans biais, et l'écart à la vraie variance est uniquement dû au nombre fini de simulations. Dans le cas non linéaire, les méthodes de Bootstrap donnent de meilleurs résultats : l'algorithme 1 est préférable pour les statistiques faiblement non linéaires (ratio et coefficient de corrélation), mais l'algorithme 3.2 donne généralement de meilleurs résultats pour l'estimation d'une médiane.

Le tableau 3.5 compare les taux de couverture effectifs des différentes méthodes. Les taux réels sont mieux respectés avec l'algorithme 3.1. Dans le cas linéaire, les taux de couverture sont mieux respectés avec l'approximation normale qu'avec l'algorithme 3.2, mais la tendance s'inverse pour les statistiques non linéaires.

3.5 Les plans à probabilités inégales proches de l'entropie maximale

Bien que le tirage réjectif ait été défini puis étudié par Hajek dans les années 60, la mise au point d'algorithmes effectifs de tirages pour un vecteur de probabilités d'inclusion quelconque est un résultat récent (Chen et al., 1994; Deville, 2000). Il existe de nombreux autres algorithmes de tirage à probabilités inégales, dont certains sont peu différents du tirage réjectif : on parlera ici de plans proches de l'entropie maximale.

3.5.1 Rappels

Soit $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_N)^t$ un vecteur de probabilités d'inclusion quelconques, q un plan de sondage quelconque de probabilités d'inclusion $\boldsymbol{\pi}$ et p_{rej} le plan réjectif ayant les mêmes probabilités d'inclusion. Berger (1998b) établit que

$$\left| \frac{V_q(\widehat{t_{y\pi}})}{V_{app}(\widehat{t_{y\pi}})} - 1 \right| \leq |o(1)| + |1 + o(1)|$$

où $V_{app}(\widehat{t_{y\pi}})$ désigne l'approximation de variance de Hajek donnée au théorême 3.7, $V_q(.)$ la variance sous le plan de sondage q et

$$D(q) = H(p_{rej}) - H(q),$$

$$H(q) = -\sum_{s \subset U} q(s) \ln(q(s)),$$

où $H(q)$ désigne l'entropie du plan de sondage q, et $D(q)$ est une quantité positive appelée la divergence du plan de sondage q par rapport au plan à entropie maximale p_{rej}.

3.5.2 Bootstrappabilité

Le résultat de Berger implique que si la divergence $D(q) \to 0$, l'approximation de variance de Hajek est également valide pour le plan q, ce qui, compte

tenu des résultats de la section 2, implique également que le plan q est bootstrappable. Ce problème a été étudié par Kervella et al. (2007).

Berger (1998a,d,b,c) a démontré que plusieurs plans de sondage avaient une divergence asymptotiquement nulle. C'est le cas notamment pour la méthode de Rao-Sampford (Rao, 1965; Sampford, 1967), le tirage successif (Hájek, 1964) et la méthode de Chao (Chao, 1982). Berger argumente que l'approximation de variance est également valide pour un plan de taille fixe randomisé : ce point fera l'objet de travaux ultérieurs.

3.5.3 Simulations

Nous présentons ici quelques simulations sur la même population artificielle, dans le cas d'un tirage systématique randomisé. Le tableau 3.6 compare, pour des échantillonnages à probabilités proportionnelles à x de différentes tailles, les performances de l'algorithme de Bootstrap simplifié avec l'estimation de variance obtenue par linéarisation, et les intervalles de confiance obtenus avec une approximation normale. Les conclusions sont sensiblement les mêmes que pour un tirage réjectif. La variance est mieux estimée avec la linéarisation dans le cas d'un total, mais l'algorithme 2 est préférable pour le non linéaire. L'algorithme 2 donne de très bons résultats en termes de taux de couverture dans le cas de l'estimation d'une médiane.

Conclusion

Nous présentons dans ce chapitre une extension de la méthode de Bootstrap de Booth et al. (1994) au cas des probabilités inégales. Nous montrons que cette méthode donne une estimation consistante de variance pour un estimateur par substitution, dans le cas d'un plan de sondage à entropie maximale (tirage poissonien et tirage réjectif) ou à entropie forte. Nous proposons également une version simplifiée de la méthode, qui donne de bons résultats si les probabilités d'inclusion restent limitées. L'algorithme simplifié a l'avantage d'être plus rapide, et de permettre d'accoler au fichier d'enquête des variables de poids Bootstrap permettant de produire a posterirori des estimations de précision très simplement.

Les deux méthodes de Bootstrap sont comparées à la linéarisation à l'aide

de simulations. Nous montrons qu'elles permettent de respecter les taux de couverture attendus de façon plus satisfaisante que la linéarisation, dans le cas de statistiques non linéaires.

TAB. 3.2 – Taux de couverture obtenus avec les deux algorithmes de Bootstrap et la linéarisation (approximation normale) pour un échantillon de taille moyenne égale à 50 sélectionné par tirage poissonien

	Algorithme 1						Algorithme 2						Linéarisation					
	2.5 %			5 %			2.5 %			5 %			2.5 %			5 %		
Statistique	L	U	L+U	L	U	L+U	L	U	L+U	L	U	L+U	L	U	L+U	L	U	L+U
	Méthode des percentiles												App.normale					
Total de y	2.1	2.9	5.0	4.5	5.4	9.9	2.3	3.9	6.2	4.1	6.8	10.9	1.8	4.3	6.1	3.9	7.3	11.2
Total de z	2.5	3.0	5.5	4.1	5.6	9.7	2.1	3.8	5.9	4.1	6.7	10.8	1.8	4.1	5.9	3.5	7.4	10.9
Ratio	2.1	2.6	4.7	4.3	5.6	9.9	2.5	2.6	5.1	5.1	4.9	10.0	3.2	2.2	5.4	5.6	4.1	9.7
Corrélation	3.4	2.2	5.6	6.9	4.8	11.7	2.0	2.9	4.9	4.1	5.4	9.5	2.2	4.3	6.5	4.7	7.5	12.2
Médiane de y	2.9	2.3	5.2	5.3	4.3	9.6	2.3	1.9	4.2	5.3	4.6	9.9						
Médiane de z	3.4	2.1	5.5	5.8	4.6	10.4	3.3	1.8	5.1	6.7	4.3	11.0						
	Méthode du t-Bootstrap																	
Total de y	2.3	1.6	3.9	5.5	3.8	9.3	2.7	1.9	4.6	4.7	4.5	9.2						
Total de z	3.0	1.9	4.9	5.9	3.8	9.7	2.5	2.4	4.9	4.8	4.4	9.2						
Ratio	1.3	3.4	4.7	2.2	6.5	8.7	2.2	2.2	4.4	4.2	4.3	8.5						
Corrélation	4.5	0.8	5.3	7.3	3.2	10.5	1.5	1.3	2.8	2.6	4.0	6.6						
Médiane de y							3.1	2.5	5.6	6.2	5.4	11.6						
Médiane de z							4.2	3.9	8.1	7.4	7.8	15.2						

Note de lecture : cf. tableau 2.2

TAB. 3.3 – Taux de couverture obtenus avec les deux algorithmes de Bootstrap et la linéarisation (approximation normale) pour un échantillon de taille moyenne égale à 100 sélectionné par tirage poissonien

	Algorithme 1						Algorithme 2						Linéarisation					
	2.5 %			5 %			2.5 %			5 %			2.5 %			5 %		
Statistique	L	U	L+U	L	U	L+U	L	U	L+U	L	U	L+U	L	U	L+U	L	U	L+U
	Méthode des percentiles												App.normale					
Total de y	2.4	3.7	6.1	4.6	6.8	11.4	1.4	2.7	4.1	3.3	5.4	8.7	1.2	2.8	4.0	3.4	5.6	9.0
Total de z	2.2	3.3	5.5	4.2	6.4	10.6	1.4	2.9	4.3	3.7	5.8	9.5	1.4	3.2	4.6	3.5	6.0	9.5
Ratio	2.8	3.1	5.9	4.8	5.3	10.1	2.8	3.4	6.2	5.6	6.5	12.1	3.6	2.8	6.4	6.3	6.0	12.3
Corrélation	3.3	2.4	5.7	6.0	5.3	11.3	2.7	2.9	5.6	4.1	5.8	9.9	2.3	4.3	6.6	4.5	7.5	12.0
Médiane de y	1.8	3.4	5.2	4.2	6.2	10.4	2.1	1.7	3.8	4.6	4.5	9.1						
Médiane de z	1.8	3.4	5.2	4.1	5.6	9.7	2.1	2.7	4.8	4.4	5.8	10.2						
	Méthode du t-Bootstrap																	
Total de y	3.1	2.7	5.8	5.8	5.7	11.5	1.5	2.0	3.5	4.4	4.1	8.5						
Total de z	2.9	2.8	5.7	5.2	5.2	10.4	1.9	1.7	3.6	4.3	4.3	8.6						
Ratio	2.1	3.4	5.5	4.0	5.9	9.9	2.7	3.2	5.9	5.6	6.1	11.7						
Corrélation	3.9	1.2	5.1	7.0	3.6	10.6	2.1	2.3	4.4	3.1	4.4	7.5						
Médiane de y							2.8	2.6	5.4	5.9	6.9	12.8						
Médiane de z							2.7	5.0	7.7	4.4	9.5	13.9						

Note de lecture : cf. tableau 2.2

TAB. 3.4 – Ecart relatif à la vraie variance pour les algorithmes 1 et 2 de Bootstrap et la technique de linéarisation dans le cas d'un tirage réjectif

	Ecart relatif (%)		
Paramètre	Echantillon de taille 20	Echantillon de taille 50	Echantillon de taille 100
		Algorithme 1	
Total de y	-4.80	-2.95	-3.13
Total de z	-6.18	-2.94	+0.91
Ratio	+2.70	+0.54	-1.03
Corrélation	-4.18	-3.20	-0.90
Médiane de y	+21.30	+16.79	+12.53
Médiane de z	+22.71	+18.09	+12.29
		Algorithme 2	
Total de y	-7.94	-1.87	-0.99
Total de z	-6.48	-2.41	-0.24
Ratio	-5.29	-1.26	+0.62
Corrélation	-6.56	-4.11	-2.99
Médiane de y	+12.83	+14.61	+13.55
Médiane de z	+20.33	+13.12	+12.80
		Linéarisation	
Total de y	-2.68	+0.02	-0.33
Total de z	-1.35	-0.51	+0.38
Ratio	-6.06	-1.61	-0.22
Corrélation	-14.14	-6.08	-3.97
Médiane de y			
Médiane de z			

TAB. 3.5 – Taux de couverture des deux algorithmes de Bootstrap pour un tirage réjectif (méthode des percentiles)

	Algorithme 1						Algorithme 2						Linéarisation					
	2.5 %			5 %			2.5 %			5 %			2.5 %			5 %		
Statistique	L	U	L+U	L	U	L+U	L	U	L+U	L	U	L+U	L	U	L+U	L	U	L+U
									Echantillon de taille 20									
Total de y	0.6	4.6	5.2	1.6	8.5	10.1	3.2	3.5	6.7	7.1	6.6	13.7	3.1	3.0	6.1	6.4	6.1	12.5
Total de z	0.0	6.5	6.5	0.1	11.5	11.6	3.8	3.7	7.5	6.6	6.1	12.7	3.3	3.4	6.7	6.1	5.8	11.9
Ratio	2.8	2.8	5.6	4.6	5.8	10.4	5.1	2.6	7.7	7.5	5.9	13.4	5.5	1.6	7.1	8.7	3.9	12.6
Corrélation	3.8	3.1	6.9	6.0	7.0	13.0	4.6	3.3	7.9	7.3	6.2	13.5	4.4	6.4	10.8	7.8	9.4	17.2
Médiane de y	2.4	2.3	4.7	5.1	4.0	9.1	3.5	2.6	6.1	6.7	5.1	11.8						
Médiane de z	3.6	3.6	7.2	6.5	5.7	12.2	3.4	3.9	7.3	7.4	5.5	12.9						
									Echantillon de taille 50									
Total de y	1.6	4.1	5.7	2.6	7.5	10.1	3.9	3.4	7.3	5.9	6.3	12.2	3.5	3.4	6.9	5.7	5.4	11.1
Total de z	0.2	4.4	4.6	0.3	9.7	10.0	2.6	3.3	5.9	4.8	5.8	10.6	2.6	3.1	5.7	4.6	5.4	10.0
Ratio	2.7	2.7	5.4	5.3	5.7	11.0	3.2	2.4	5.6	5.7	4.3	10.0	4.0	1.8	5.8	6.2	3.6	9.8
Corrélation	2.9	3.5	6.4	5.6	5.1	10.7	2.9	3.8	6.7	6.2	7.0	13.2	1.9	5.4	7.3	4.9	8.4	13.3
Médiane de y	2.0	2.9	4.9	5.9	5.3	11.2	2.5	4.2	6.7	5.4	7.1	12.5						
Médiane de z	3.0	1.8	4.8	4.9	3.6	8.5	2.6	2.4	5.0	4.2	5.3	9.5						
									Echantillon de taille 100									
Total de y	1.3	2.8	4.1	3.0	7.5	10.5	2.1	3.8	5.9	3.5	6.2	9.7	1.9	3.2	5.1	3.5	5.9	9.4
Total de z	1.1	4.7	5.8	1.5	9.1	10.6	2.0	2.4	4.4	5.4	5.7	11.1	2.4	2.4	4.8	5.1	5.4	10.5
Ratio	2.8	2.2	5.0	5.3	4.9	10.2	3.2	2.1	5.3	6.6	5.0	11.6	3.6	1.8	5.4	7.1	4.2	11.3
Corrélation	2.9	3.1	6.0	5.8	5.2	11.0	3.0	2.7	5.7	5.9	5.2	11.1	2.8	3.5	6.3	5.9	6.0	11.9
Médiane de y	3.6	2.8	6.4	5.8	5.2	11.0	1.9	3.0	4.9	3.5	5.8	9.3						
Médiane de z	3.3	2.7	6.0	5.8	4.9	10.7	3.1	2.5	5.6	4.8	6.1	10.9						

Note de lecture : cf. tableau 2.2

TAB. 3.6 – Ecart relatif (%) à la variance exacte et taux de couverture obtenus avec l'algorithme simplifié de Bootstrap (méthode des percentiles) et la linéarisation (approximation normale) pour un échantillon sélectionné par tirage systématique randomisé

	Algorithme 2							Linéarisation						
		2.5 %			5 %				2.5 %			5 %		
Statistique	Ecart relatif	L	U	L+U	L	U	L+U	Ecart relatif	L	U	L+U	L	U	L+U
	Echantillon de taille 50													
Total de y	+0.28	1.8	3.6	5.4	4.9	6.6	11.5	-0.47	1.9	3.2	5.1	5.3	6.1	11.4
Total de z	+10.46	2.2	2.6	4.8	4.0	5.0	9.0	+2.44	2.5	2.2	4.7	4.6	4.9	9.5
Ratio	-0.00	2.7	2.3	5.0	5.5	4.9	10.4	-0.45	3.3	1.5	4.8	6.6	3.8	10.4
Corrélation	-3.56	3.3	2.4	5.7	6.9	4.9	11.8	-4.99	2.4	3.4	5.8	4.2	6.4	10.6
Médiane de y	+0.84	2.6	2.6	5.2	5.2	5.2	10.4							
Médiane de z	+0.25	3.5	2.3	5.8	5.6	4.1	9.7							
	Echantillon de taille 100													
Total de y	+1.55	2.3	2.0	4.3	4.2	4.9	9.1	+0.01	2.5	2.1	4.6	5.0	4.7	9.7
Total de z	+9.84	2.8	2.2	5.0	4.7	4.5	9.2	-0.59	3.4	2.1	5.5	6.0	4.3	10.3
Ratio	-0.65	1.5	2.7	4.2	5.1	5.4	10.5	-0.39	2.3	1.5	3.8	5.5	4.6	10.1
Corrélation	-1.02	3.8	3.3	7.1	6.1	5.4	11.5	-3.55	2.8	4.5	7.3	5.8	6.6	12.4
Médiane de y	+15.13	3.2	1.8	5.0	5.4	4.8	10.2							
Médiane de z	+12.10	3.2	2.3	5.5	5.4	4.9	10.1							

Note de lecture : cf. tableau 2.2

Bibliographie

Barbe, P. and Bertail, P. (1995). *The Weighted Bootstrap*. Springer-Verlag, New-York.

Berger, Y. (1998a). *Comportements asymptotiques des plans de sondage à probabilités inégales pour un modèle de population fixe*. PhD thesis, Université Libre de Bruxelles.

Berger, Y. (1998b). Rate of convergence for asymptotic variance for the Horvitz-Thompson estimator. *Journal of Statistical Planning and Inference*, 74 :149–168.

Berger, Y. (1998c). Rate of convergence to normal distribution for the horvitz-thompson estimator. *Journal of Statistical Planning and Inference*, 67 :209–226.

Berger, Y. (1998d). Variance estimation using list sequential scheme for unequal probability sampling. *Journal of Official Statistics*, 14 :315–323.

Bertail, P. and Combris, P. (1997). Bootstrap généralisé d'un sondage. *Annales d'Economie et de Statistique*, 46 :49–83.

Bickel, P. and Freedman, D. (1984). Asymptotic normality and the bootstrap in stratified sampling. *The Annals of Statistics*, 12 :470–482.

Booth, J., Butler, R., and Hall, P. (1994). Bootstrap methods for finite populations. *Journal of the American Statistical Association*, 89 :1282–1289.

Brewer, K. and Hanif, M. (1983). *Sampling with Unequal Probabilities*. Springer-Verlag, New-York.

Chao, M. T. (1982). A general purpose unequal probability sampling plan. *Biometrika*, 69 :653–656.

Chao, M.-T. and Lo, S.-H. (1985). A bootstrap method for finite population. *Sankhya Series A*, 47 :399–405.

Chen, X.-H., Dempster, A., and Liu, J. (1994). Weighted finite population sampling to maximize entropy. *Biometrika*, 81 :457–469.

Deville, J. (1987). Réplications d'échantillons, demi-échantillons, jackknife, bootstrap. In *Les Sondages*, Paris. Economica.

Deville, J.-C. (1999). Variance estimation for complex statistics ans estimators : linearization and residual techniques. *Techniques d'enquête, Survey methodology*, 25 :193–204.

Deville, J.-C. (2000). Note sur l'algorithme de Chen, Dempster et Liu. Technical report, CREST-ENSAI.

Deville, J.-C. and Tillé, Y. (1998). Unequal probability sampling without replacement through a splitting method. *Biometrika*, 85 :89–101.

Deville, J.-C. and Tillé, Y. (2000). Selection of several unequal probability samples from the same population. *Journal of Statistical Planning and Inference*, 86 :215–227.

Devroye, L. (1986). *Non-Uniform Random Variate Generation*. Springer-Verlag, New-York.

Dupacova, J. (1979). A note on rejective sampling. In *Contribution to Statistics (J. Hajek memorial volume)*. Academia Prague.

Feller, W. (1966). *An introduction to Probability Theory and its applications*. Wiley, New-York.

Gross, S. (1980). Median estimation in sample surveys. *ASA Proceedings of Survey Research*, pages 181–184.

Hájek, J. (1964). Asymptotic theory of rejective sampling with varying probabilities from a finite population. *Annals of Mathematical Statistics*, 35 :1491–1523.

Hájek, J. (1981). *Sampling from a Finite Population*. Marcel Dekker, New York.

Hanif, M. and Brewer, K. R. W. (1980). Sampling with unequal probabilities without replacement : A review. *International Statistical Review*, 48 :317–335.

Kervella, A., L'Hour, E., Raillard, N., and Volant, S. (2007). Evaluation empirique du bootstrap pour des plans de sondage à probabilités inégales. Rapport de projet statistique, ENSAI.

Qualité, L. (2006). A comparison of conditional poisson sampling versus unequal probability sampling with replacement. Technical report, Neuchâtel University.

Rao, J. (1965). On two simple schemes of unequal probability sampling without replacement. *Journal of the Indian Statistical Association*, 3 :173–180.

Sampford, M. (1967). On sampling without replacement with unequal probabilities of selection. *Biometrika*, 54 :494–513.

Sen, A. (1953). On the estimate of the variance in sampling with varying probabilities. *Journal of Indian Society for Agricultural Statistics*, 5 :119–127.

Sitter, R. (1992). A resampling procedure for complex survey data. *Journal of the American Statistical Association*, 87 :755–765.

Thompson, M. (1997). *Theory of Sample Surveys*. Chapman and Hall, London.

Tillé, Y. (2006). *Sampling algorithms*. Springer, New-York.

Yates, F. and Grundy, P. (1953). Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society*, B15 :235–261.

Chapitre 4

Bootstrap d'un plan de sondage équilibré

Un plan de sondage est dit équilibré s'il permet d'inférer de façon exacte pour les totaux de variables auxiliaires connues sur l'ensemble des individus de la population à l'étape du plan de sondage. Si l'échantillonnage se fait à probabilités égales, l'équilibrage assure que des structures connues sur la population seront également respectées dans l'échantillon, assurant ainsi une "représentativité" de l'échantillon sur les variables de contrôle. L'équilibrage permet donc de mettre à profit une information auxiliaire connue avant tirage afin d'améliorer le résultat de l'échantillonnage.

Un algorithme général d'échantillonnage équilibré, permettant de sélectionner des échantillons à probabilités inégales et équilibrés sur un nombre quelconque de variables, a été proposé par Deville and Tillé (2004). Deville and Tillé (2005) ont également proposé des formules approchées de calcul de précision pour un tirage équilibré à grande entropie ; ils montrent que la variance n'est plus donnée que par les résidus d'une régression de la variable d'intérêt sur les variables d'équilibrage. L'objectif de ce chapitre est double :

→ proposer un algorithme rapide d'échantillonnage équilibré, permettant la sélection d'échantillons sur de grandes bases de sondage et ouvrant la voie à un calcul de précision par rééchantillonnage dans un temps raisonnable,
→ proposer une justification de la formule d'approximation de variance de Deville and Tillé (2005), via une approximation des probabilités d'inclusion

doubles, afin d'établir la consistance de notre méthode de Bootstrap pour un tirage équilibré à entropie maximale.

Le chapitre est organisé de la façon suivante. En section 1, nous donnons quelques rappels sur l'échantillonnage équilibré et présentons l'algorithme général de tirage. En section 2, nous proposons un algorithme plus rapide d'échantillonnage équilibré et introduisons la notion d'échantillonnage équilibré stratifié. En section 3, nous établissons la consistance de la méthode de Bootstrap dans le cas d'un tirage équilibré sur une variable. Le cas d'un équilibrage multidimensionnel est discuté en section 4. En section 5, nous proposons une généralisation de la méthode Mirror-Match de Sitter (1992) dans le cas d'un tirage équilibré à entropie maximale et à probabilités égales.

4.1 L'échantillonnage équilibré

4.1.1 Définition

On suppose que les vecteurs de valeurs

$$\mathbf{x_k} = (x_{k1}, \ldots, x_{kp})'$$

prises par p variables auxiliaires sont connus pour toutes les unités de la population U. On suppose sans perte de généralité que ces p vecteurs sont linéairement indépendants.

L'estimateur de Horvitz-Thompson du total de la variable y, donné par $\widehat{t_{y\pi}} = \sum_{k\in S} y_k/\pi_k$, est un estimateur sans biais de t_y. En particulier, le π-estimateur du total $t_{x_j} = \sum_{k\in U} x_{kj}$ de la $j^{\text{ème}}$ variable auxiliaire est donné par $\widehat{t_{x_j\pi}} = \sum_{k\in S} x_{kj}/\pi_k$ et

$$E\left(\sum_{k\in S}\frac{x_{ki}}{\pi_k}\right) = \sum_{k\in U} x_{ki} = t_{x_i} \tag{4.1}$$

$\widehat{t_{x_j\pi}}$ restitue donc en moyenne la valeur attendue t_{x_j}. On dira qu'un échantillon est équilibré sur la variable x_j si l'équation 4.1 n'est pas vérifiée **en moyenne, mais exactement**.

Définition 4.1. *Avec les notations précédentes, l'échantillon S sélectionné avec des probabilités π_k est dit équilibré sur la variable x_j si*

$$\sum_{k\in S} \frac{x_{kj}}{\pi_k} = t_{x_j}. \tag{4.2}$$

Par extension, le plan de sondage p est dit équilibré sur les variables $x_1, \ldots, x_p$ si son support est restreint aux échantillons équilibrés sur chacune des variables $x_1, \ldots, x_p$. Autrement dit, p est équilibré sur les variables $x_1, \ldots, x_p$ si pour tout échantillon s tel que $p(s) > 0$:

$$\sum_{k\in s} \frac{x_{ki}}{\pi_k} = t_{x_i} \quad \forall i = 1 \ldots p \tag{4.3}$$

Le système d'équations 4.3 est appelé système des équations d'équilibrages.

Exemple 4.1. *L'équilibrage sur la variable probabilité d'inclusion assure un échantillonnage de taille fixe. En effet, 4.2 se réduit alors à*

$$\begin{aligned} \widehat{t_{x\pi}} &= \textstyle\sum_{k\in S} \frac{\pi_k}{\pi_k} = n(S) \\ &= \textstyle\sum_{k\in U} \pi_k = E(n(S)) \end{aligned}$$

et l'équilibrage assure que la taille d'échantillon effectivement obtenue est exactement la taille souhaitée.

En particulier, le tirage réjectif (plan de taille fixe) est, pour des probabilités d'inclusion fixées, le plan à entropie maximale équilibré sur la variable probabilité d'inclusion.

Exemple 4.2. *L'équilibrage sur la variable constante égale à 1 assure que la taille de la population est parfaitement estimée. En effet, 4.2 se réduit alors à*

$$\begin{aligned} \widehat{t_{x\pi}} &= \textstyle\sum_{k\in S} \frac{1}{\pi_k} = \widehat{N_\pi} \\ &= \textstyle\sum_{k\in U} 1 = N \end{aligned}$$

Les deux exemples précédents illustrent l'apport de l'échantillonnage équilibré par rapport aux méthodes d'échantillonnage à probabilités inégales présentées au chapitre précédent. La variable constante et la variable probabilité

d'inclusion sont deux variables "gratuites" et toujours disponibles dans une base de sondage. Equilibrer sur ces deux variables permet d'obtenir un échantillon de taille fixe (propriété déjà assurée par de nombreux algorithmes), mais également d'inférer exactement pour l'effectif total de la population.

Exemple 4.3. *Dans le Nouveau Recensement de la population réalisé en France (cf chapitre 6), les petites communes (moins de 10 000 habitants) sont, au sein de chaque région, partitionnées aléatoirement en 5 échantillons appelés groupes de rotation.*
*Ces échantillons sont sélectionnés à probabilités égales (*1/5*), et en équilibrant sur les variables :* `sexe` *et* `tranche d'âge`*. L'équilibrage assure que la structure de chacun des 5 échantillons sur les modalités des variables* `sexe` *et* `tranche d'âge` *est la même que dans l'ensemble de la population.*

4.1.2 Mise en oeuvre : la méthode du Cube

La recherche d'un algorithme d'échantillonnage équilibré est un problème déjà ancien. Yates (1949) et Thionet (1953) proposent des méthodes d'équilibrage de type réjectif (où on sélectionne préalablement des échantillons, en conservant celui qui assure un bon équilibrage pour les totaux des variables de contrôle) ; voir également Ardilly (1991). Deville et al. (1988) ont également proposé un algorithme, mais en se restreignant au cas d'un échantillonnage équilibré à probabilités égales.

La méthode du CUBE, qui apporte une solution générale au problème d'équilibrage (i.e., équilibrage à probabilités inégales sur un nombre quelconque de variables) a été développée à l'Ensai par Deville and Tillé (2004) ; voir également Tillé (2001). La méthode a été d'abord consacrée à la sélection d'unités primaires dans un sondage à deux degrés, car le temps d'exécution était proportionnel au carré de la taille de la population. Cette méthode a ensuite été appliquée à plusieurs problèmes statistiques importants. Par exemple, les groupes de rotation de communes et d'adresses du Nouveau Recensement français ont été sélectionnés à l'aide de la méthode du Cube (Dumais and Isnard, 2000; Bertrand et al., 2004).

La méthode du Cube est basée sur une représentation géométrique d'un plan de sondage. Les échantillons s peuvent être vus comme les sommets

d'un hypercube de $\mathbb{R}^N$ donné par $C = [0,1]^N$. Le système 4.3 des équations d'équilibrage définit un sous-espace de $\mathbb{R}^N$ égal à $K = \boldsymbol{\pi} + \mathrm{Ker}(\mathbf{A})$, où $\mathbf{A}$ désigne la matrice des contraintes égale à

$$A = \begin{pmatrix} \mathbf{x}_1 & \dots & \mathbf{x}_N \end{pmatrix} = \begin{pmatrix} x_{11}/\pi_1 & & x_{1p}/\pi_1 \\ \vdots & \dots & \dots \\ x_{N1}/\pi_N & & x_{Np}/\pi_N \end{pmatrix}^t .$$

L'idée de base consiste à sélectionner aléatoirement un sommet de $K \cap C$, à l'aide d'une suite de déplacements aléatoires dans $K \cap C$.

Il n'est généralement pas possible de sélectionner un échantillon exactement équilibré. Par exemple, dans une population de 100 personnes comptant 47 hommes et 53 femmes, sélectionner un échantillon de taille 10 exactement équilibré sur la variable sexe signifierait sélectionner un échantillon de 4.7 hommes et 5.3 femmes, ce qui est impossible. Pour cette raison, la méthode du Cube est constituée de deux phases appelées phase de vol et phase d'atterrissage :

→ Pendant la phase de vol, les contraintes sont toujours exactement respectées. L'objectif est d'arrondir aléatoirement presque toutes les probabilités d'inclusion à 0 ou 1. Cette phase est décrite dans l'algorithme 4.1.
→ La phase d'atterrissage consiste à solutionner au mieux le fait que le système 4.3 ne peut être exactement respecté.

Si T est la dernière étape de l'algorithme 4.1 et que $\boldsymbol{\pi}^* = \boldsymbol{\pi}(T)$, alors Deville and Tillé (2004) ont montré que,

1. $E(\boldsymbol{\pi}^*) = \boldsymbol{\pi}$,
2. $\mathbf{A}\boldsymbol{\pi}^* = \mathbf{A}\boldsymbol{\pi}$,
3. Si $q = \mathrm{card}\{k | 0 < \pi_k^* < 1\}$, alors $q \leq p$, où p désigne le nombre de variables auxiliaires.

Le vecteur $\boldsymbol{\pi}^*$ peut être un échantillon, mais dans la plupart des cas il y a au plus q éléments non entiers dans $\boldsymbol{\pi}^*$. Si $q > 0$, le problème d'arrondi est réglé par la phase d'atterrissage. Notons que q est toujours inférieur au nombre de variables d'équilibrage, et que ce dernier est généralement très faible devant la taille de la population : l'atterrissage ne porte donc que sur un nombre

FIG. 4.1 – Procédure générale d'équilibrage, phase de vol

Initialiser à $\pi(0) = \pi$. Ensuite, au temps $t = 0, ..., T$, répéter les trois étapes suivantes

Etape 1. Générer un vecteur quelconque $\mathbf{u}(t) = \{u_k(t)\} \neq 0$, aléatoire ou non, tel que $\mathbf{u}(t)$ soit dans le noyau de la matrice $\mathbf{A}$, et $u_k(t) = 0$ si $\pi_k(t)$ est un entier.

Etape 2. Calculer $\lambda_1^*(t)$ et $\lambda_2^*(t)$, les plus grandes valeurs de $\lambda_1(t)$ et $\lambda_2(t)$ telles que $0 \leq \boldsymbol{\pi}(t) + \lambda_1(t)\mathbf{u}(t) \leq 1$, et $0 \leq \boldsymbol{\pi}(t) - \lambda_2(t)\mathbf{u}(t) \leq 1$. Noter que $\lambda_1(t) > 0$ et $\lambda_2(t) > 0$.

Etape 3. Sélectionner

$$\boldsymbol{\pi}(t+1) = \begin{cases} \boldsymbol{\pi}(t) + \lambda_1^*(t)\mathbf{u}(t) & \text{avec probabilité } q(t) \\ \boldsymbol{\pi}(t) - \lambda_2^*(t)\mathbf{u}(t) & \text{avec probabilité } 1 - q(t), \end{cases} \tag{4.4}$$

où $q(t) = \lambda_2^*(t)/\{\lambda_1^*(t) + \lambda_2^*(t)\}$.

Cette procédure générale est répétée jusqu'à ce qu'il ne soit plus possible de réaliser l'étape 1.

d'unités très limité.

La première solution pour l'atterrissage consiste à relâcher une contrainte et à relancer à nouveau la phase de vol, jusqu'à ce qu'il ne soit plus possible de "bouger" à nouveau à l'intérieur de l'hyperplan des contraintes. Les contraintes sont donc relâchées successivement. La seconde solution utilise un programme linéaire pour obtenir le meilleur plan de sondage équilibré approché, voir Deville and Tillé (2004). Le temps d'exécution est donné essentiellement par la phase de vol. La nouvelle implémentation que nous proposons dans la section suivante ne concerne que la phase de vol, et la phase d'atterrissage reste inchangée.

4.1.3 Calcul de précision analytique

Dans le cas du π-estimateur d'un total, la variance peut être théoriquement calculée en fonction des probabilités d'inclusion d'ordre 2 :

$$\begin{aligned} V(\widehat{t_{y\pi}}) &= \sum_{k\in U}\sum_{l\in U}\frac{y_k}{\pi_k}\frac{y_l}{\pi_l}\Delta_{kl} \\ &= \tilde{y}'\Delta\tilde{y} \end{aligned}$$

avec $\tilde{y} = (y_1/\pi_1 \dots y_i/\pi_i \dots y_N/\pi_N)'$, $\Delta = (\Delta_{kl})_{kl}$ et

$$\Delta_{kl} = \begin{cases} \pi_{kl} - \pi_k\pi_l & \text{si} \quad k \neq l \\ \pi_k(1-\pi_k) & \text{si} \quad k = l \end{cases}$$

En pratique, ces probabilités d'inclusion doubles et l'opérateur de variance-covariance Δ ne peuvent être calculés exactement. Pour résoudre ce problème, Deville and Tillé (2005) proposent une approximation de variance dans le cas où le tirage équilibré est à entropie maximale.

Le tirage poissonien conditionnel

Soit p le plan de sondage respectant les probabilités d'inclusion $\boldsymbol{\pi} = (\pi_k)_{k\in U}$ et le système d'équations 4.3, et d'entropie maximale parmi les plans de sondage vérifiant ces deux propriétés. Hájek (1981) établit que p peut être vu sous forme d'un plan poissonien q de probabilités d'inclusion $(p_k)_{k\in U}$, conditionné par les équations d'équilibrage.

La mise en oeuvre du tirage poissonien conditionnel se fait de la façon suivante : soit S un échantillon sélectionné selon le plan de sondage q. Si S vérifie

$$\sum_{k\in S}\frac{\mathbf{x_k}}{\pi_k} = \sum_{k\in U}\mathbf{x_k} \tag{4.5}$$

alors l'échantillon est conservé. Sinon, on recommence jusqu'à obtenir un échantillon vérifiant l'équation 4.5. L'échantillon finalement retenu sera noté $\tilde{S}$. Le tirage réjectif est un cas particulier de tirage poissonien conditionnel, obtenu avec $\mathbf{x_k}$ scalaire égal à la probabilité d'inclusion.

Notons au passage que

$$E_q\left(\sum_{k\in\tilde{S}}\frac{\mathbf{x_k}}{\pi_k}\right)=\sum_{k\in U}\frac{p_k}{\pi_k}\mathbf{x_k}, \tag{4.6}$$

où $E_q(.)$ désigne l'espérance sous le plan de Poisson. Cela signifie que l'équation d'équilibrage ne représente pas une situation vraie en moyenne, et qu'il peut être nécessaire de rejeter un grand nombre d'échantillons avant d'obtenir un échantillon équilibré. Cependant, le théorême 1 de Deville and Tillé (2005) montre qu'il existe un unique plan poissonien q de probabilités d'inclusion $(p_k)_{k\in U}$, tel que p soit le conditionnel du plan q, et vérifiant

$$\sum_{k\in U}\frac{p_k}{\pi_k}\mathbf{x_k}=\sum_{k\in U}\mathbf{x_k}. \tag{4.7}$$

L'utilisation de ce plan poissonien permet donc théoriquement d'accélérer la vitesse de tirage de l'échantillon équilibré. Le plan poissonien vérifiant les propriétés précédentes sera appelé **plan poissonien canoniquement associé à** p, pour reprendre la terminologie de Hajek. Sauf indication contraire, nous supposerons dans la suite que le poissonien q est canoniquement associé au plan p.

En postulant une distribution normale multivariée pour le vecteur $(\widehat{t_{y\pi}}\ ,\ \widehat{t_{\mathbf{x}\pi}})$, Deville and Tillé (2005) obtiennent la formule approchée de variance

$$V_{app}(\widehat{t_{y\pi}})=\sum_{k\in U}\frac{b_k}{\pi_k^2}(y_k-y_k^*)^2 \tag{4.8}$$

où

$$y_k^*=\mathbf{x_k'}\left(\sum_{l\in U}b_l\frac{\mathbf{x_l}\mathbf{x_l'}}{\pi_l^2}\right)^{-1}\sum_{l\in U}b_l\frac{\mathbf{x_l}y_l}{\pi_l^2}, \tag{4.9}$$

et le poids $b_k=p_k(1-p_k)$ dépend des probabilités d'inclusion du plan poissonien correspondant q. La variable y^* correspond à une prédiction de la variable y à l'aide des variables d'équilibrage. La variance n'est donc plus donnée que par les résidus de la régression pondérée de la variable d'intérêt y sur les variables d'équilibrages $\mathbf{x}$. Par exemple, équilibrer sur la variable probabilité d'inclusion $\boldsymbol{\pi}$ et la variable constante égale à 1 revient à ajouter

une constante dans le modèle expliquant la variable y par la variable $\boldsymbol{\pi}$.

L'approximation 4.8 peut encore s'écrire sous la forme

$$V_{app}(\widehat{t_{y\pi}}) = \tilde{y}'\Delta_{app}\tilde{y}$$

avec $\tilde{y} = (y_1/\pi_1, \ldots, y_N/\pi_N)'$, $\Delta_{app} = (\Delta_{app,kl})_{kl}$ et

$$\Delta_{app,kl} = \begin{cases} b_k \frac{\mathbf{x_k'}}{\pi_k} \left(\sum_{i \in U} b_i \frac{\mathbf{x_i x_i'}}{\pi_i^2} \right)^{-1} \frac{\mathbf{x_k}}{\pi_k} b_k & \text{si} \quad k \neq l \\ b_k - b_k \frac{\mathbf{x_k'}}{\pi_k} \left(\sum_{i \in U} b_i \frac{\mathbf{x_i x_i'}}{\pi_i^2} \right)^{-1} \frac{\mathbf{x_k}}{\pi_k} b_k & \text{si} \quad k = l \end{cases} . \tag{4.10}$$

La difficulté réside dans le fait que le facteur b_k dépend des probabilités d'inclusion inconnues du plan poissonien, et doit être également approché. Plusieurs approximations sont proposées par Deville and Tillé (2005) :

1. La première consiste à considérer que pour n suffisamment grand, $p_k \simeq \pi_k$, ce qui conduit à
$$b_{1k} = \pi_k(1 - \pi_k). \tag{4.11}$$
On note Δ_1 l'opérateur de variance correspondant. Nous donnons dans les sections 3 et 4 un résultat justifiant l'approximation de p_k par π_k.
2. La seconde consiste à partir de l'approximation précédente, en appliquant une correction pour la perte de degrés de liberté
$$b_{2k} = \pi_k(1 - \pi_k)\frac{N}{N - p} \tag{4.12}$$
donnant un opérateur de variance approché Δ_2.
3. La troisième consiste à s'ajuster sur les éléments diagonaux de Δ, toujours connus, ce qui donne l'approximation
$$b_{3k} = \pi_k(1 - \pi_k)\frac{trace\Delta}{trace\Delta_1} \tag{4.13}$$
et l'opérateur de variance approché Δ_3.
4. Le dernière approximation consiste à caler les éléments diagonaux de Δ_{app} sur ceux de Δ, toujours connus. Cela revient à résoudre le système d'équations non linéaires
$$\pi_k(1 - \pi_k) = b_k - b_k \frac{\mathbf{x_k'}}{\pi_k} \left(\sum_{i \in U} b_i \frac{\mathbf{x_i x_i'}}{\pi_i^2} \right)^{-1} \frac{\mathbf{x_k}}{\pi_k} b_k \quad \forall k \in U \tag{4.14}$$

On peut résoudre ce système par une méthode itérative, mais une solution n'existe pas toujours. Une solution consiste à n'utiliser qu'une itération pour obtenir

$$b_k = \pi_k(1-\pi_k)\left(1 - \frac{\mathbf{x_k'}}{\pi_k}\left(\sum_{i\in U}\pi_i(1-\pi_i)\frac{\mathbf{x_i x_i'}}{\pi_i^2}\right)^{-1}\frac{\mathbf{x_k}}{\pi_k}\pi_k(1-\pi_k)\right)^{-1}. \tag{4.15}$$

Deville and Tillé (2005) déduisent de cette approximation de variance la forme générale d'un estimateur de variance :

$$\widehat{V}(\widehat{t_{y\pi}}) = \sum_{k\in S}\frac{c_k}{\pi_k^2}(y_k - \widehat{y_k^*})^2 \tag{4.16}$$

où

$$\widehat{y_k^*} = \mathbf{x_k'}\left(\sum_{l\in S} c_l\frac{\mathbf{x_l x_l'}}{\pi_l^2}\right)^{-1}\sum_{l\in S} c_l\frac{\mathbf{x_l}y_l}{\pi_l^2}, \tag{4.17}$$

Deville and Tillé (2005) proposent plusieurs approximations pour les coefficients c_k, et montrent à l'aide de simulations le bon comportement de ces approximations. Barat et al. (2005) ont élaboré une macro SAS de calcul de précision d'un échantillon équilibré basé sur les formules précédentes.

4.2 Un algorithme rapide d'échantillonnage équilibré

La méthode du Cube est en fait une famille d'algorithmes dont l'implémentation peut prendre différentes formes. Chauvet and Tillé (2006) ont proposé une implémentation très rapide. L'originalité consiste à appliquer l'étape de base à un sous-ensemble d'unités et non à la population entière restante. Ce sous-ensemble évolue à chaque étape de l'algorithme, et le temps d'exécution ne dépend plus du carré de la taille de la population. Cette amélioration permet l'utilisation de la méthode du Cube sur de très grandes bases de sondage, avec un nombre important de variables d'équilibrage. Elle permet également d'envisager le recours à des méthodes d'estimation de précision par rééchantillonnage avec des temps de calcul raisonnables.

4.2.1 Présentation

Le but de cette nouvelle implémentation est donc d'obtenir une réduction du temps d'exécution. Dans l'algorithme général, la recherche d'un vecteur $\mathbf{u}$ de Ker$\mathbf{A}$ est très coûteuse. L'idée de base est d'utiliser une sous-matrice $\mathbf{B}$ contenant uniquement $p+1$ colonnes de $\mathbf{A}$. Notons que le nombre de variables p est plus petit que la taille de la population N, et que rang $\mathbf{B} \leq p$. La dimension du noyau de $\mathbf{B}$ est donc supérieure ou égale à 1.

Un vecteur $\mathbf{v}$ de Ker$\mathbf{B}$ peut alors être utilisé pour construire un vecteur $\mathbf{u}$ de Ker$\mathbf{A}$ en complétant $\mathbf{v}$ avec des zéros pour les colonnes de $\mathbf{B}$ qui ne sont pas dans $\mathbf{A}$. Avec cette idée, tous les calculs peuvent être faits uniquement sur $\mathbf{B}$. Cette méthode est décrite dans l'Algorithme 4.2 dont on trouvera une description à la fin du chapitre.

Si $\widetilde{T}$ est la dernière étape de l'algorithme et que $\widetilde{\boldsymbol{\pi}} = \boldsymbol{\pi}(\widetilde{T})$, alors on a

1. $E(\widetilde{\boldsymbol{\pi}}) = \boldsymbol{\pi}$,
2. $\mathbf{A}\widetilde{\boldsymbol{\pi}} = \mathbf{A}\boldsymbol{\pi}$,
3. si $\tilde{q} = \text{card}\{k|0 < \tilde{\pi}_k < 1\}$, alors $\tilde{q} \leq p$, où on rappelle que p désigne le nombre de variables auxiliaires.

Dans le cas où certaines contraintes peuvent être exactement satisfaites, on peut poursuivre la phase de vol. Supposons que $\mathbf{C}$ désigne la matrice contenant les colonnes de $\mathbf{A}$ qui correspondent à des valeurs non entières de $\widetilde{\boldsymbol{\pi}}$, et que $\boldsymbol{\phi}$ est le vecteur des valeurs non entières de $\widetilde{\boldsymbol{\pi}}$. Si $\mathbf{C}$ n'est pas de plein rang, une ou plusieurs étapes de l'algorithme général peuvent encore être appliquées à $\mathbf{C}$ et $\boldsymbol{\phi}$. Un retour à l'Algorithme 4.1 est donc nécessaire pour les dernières étapes.

4.2.2 Cas de l'échantillonnage à probabilités inégales

Quand $p = 1$ et que la seule variable d'équilibrage est $x_k = \pi_k$, alors tirer un échantillon équilibré revient à échantillonner avec des probabilités inégales. Dans ce cas, $\mathbf{A} = (1 \dots 1)$. A chaque étape, la matrice $\mathbf{B}$ vaut $(1, 1)$, et $\mathbf{u} = (-1, 1)'$. L'algorithme 4.2 peut être grandement simplifié comme présenté dans l'algorithme 4.3, ce qui donne une méthode très simple de sélection d'un échantillon avec des probabilités inégales. L'algorithme 4.3 présenté à la fin du chapitre est en fait une implémentation de la méthode du pivot proposée

initialement par Deville and Tillé (1998) dans le cadre de la méthode de scission.

Dans le cas où on opère un tri aléatoire préalable sur les données, l'algorithme précédent peut être vu comme une variante de la méthode de Rao-Hartley-Cochran (Rao et al., 1962) permettant de respecter exactement des probabilités d'inclusion fixées.

4.2.3 Echantillonnage équilibré stratifié

Notations

Nous conservons les mêmes notations que dans le paragraphe précédent. Nous supposons ici que U est partitionnée en H strates $U_1, \ldots, U_H$. Rappelons que le plan de sondage est dit équilibré sur la variable x si

$$\sum_{k \in S} \frac{x_k}{\pi_k} = \sum_{k \in U} x_k. \tag{4.18}$$

On dit que le plan de sondage est équilibré par strates sur la variable x si

$$\sum_{k \in S_h} \frac{x_k}{\pi_k} = \sum_{k \in U_h} x_k, \text{ pour tout } h = 1, \ldots, H, \tag{4.19}$$

où S_h désigne l'intersection de l'échantillon S et de la strate U_h. Notons que si un plan de sondage est équilibré par strates, il est également globalement équilibré sur la même population.

Le technique d'échantillonnage équilibré stratifié a été utilisée par le Nouveau Recensement français pour la constitution des groupes de rotation de petites communes. Dans chaque région de France, ces groupes de rotation sont constitués en sélectionnant des échantillons globalement équilibrés sur des variables socio-démographiques, et équilibrés par département sur le nombre de ménages, afin d'assurer une bonne représentation des communes de chaque département dans chacun des groupes de rotation.

Inconvénients d'un équilibrage direct par strates

L'échantillonnage équilibré stratifié peut être réalisé en tirant un échantillon directement dans la population entière. En effet, (4.19) est équivalent à

$$\sum_{k\in U} \frac{S_k(x_k\ 1_{k\in U_h})}{\pi_k} = \sum_{k\in U} x_k\ 1_{k\in U_h} \text{ pour tout } h = 1\ldots H.$$

Il suffit donc de sélectionner un échantillon dans U, équilibré sur les variables données par le produit des variables d'équilibrage $x_1, \ldots, x_p$ et des variables indicatrices :

$$1_{k\in U_h} = \begin{cases} 1 & \text{si } k \in U_h \\ 0 & \text{sinon,} \end{cases}$$

ce qui signifie équilibrer que $H \times p$ variables. Cette méthode a plusieurs inconvénients :

→ Si $H \times p$ est trop grand, il est impossible de réaliser une phase d'atterrissage en cherchant à minimiser un critère de distance à l'état d'équilibre, car le nombre d'échantillons possibles est trop important. La seule option d'atterrissage possible consiste à abandonner progressivement les contraintes.
→ Toutes les strates n'ont pas la même qualité d'équilibrage. En abandonnant successivement les contraintes, les strates correspondant aux variables abandonnées au début sont moins bien équilibrées.
→ On peut ne pas obtenir une taille fixe d'échantillon dans chaque strate, ce qui génère une variance supplémentaire pour les estimations intra-strates.

Nous proposons un algorithme de tirage, inspirée d'une remarque sur le traitement des grandes bases de sondage avec l'ancienne macro CUBE (Rousseau and Tardieu, 2004). Cet algorithme a été programmé sous forme d'une macro SAS (Chauvet and Tillé, 2007). Le principe est le suivant :

→ On essaye tout d'abord d'équilibrer par strate : on réalise une phase de vol indépendamment dans chaque strate afin d'équilibrer sur les variables choisies.
→ Quand il n'est plus possible d'équilibrer par strate, on recherche un équilibrage global : on rassemble toutes les unités restantes, n'ayant pas été sélectionnées ou rejetées durant les phases de vol sur les strates, puis on effectue une dernière phase de vol sur ces unités.

→ On applique alors la phase d'atterrissage à toutes les unités qui n'ont été ni sélectionnées ni rejetées.

Nous donnons ci-dessous un exemple d'utilisation de l'échantillonnage équilibré stratifié pour la sélection d'un échantillon dans une base d'adresses. Cette technique nous sera également utile pour mettre en oeuvre de façon effective une généralisation de la méthode Mirror-Match (Sitter, 1992) au cas d'un échantillonnage équilibré à probabilités égales (voir la section 5).

Un exemple numérique

Nous utilisons une population artificielle de 26 471 unités, correspondant à une commune divisée en 36 strates (variable ZONE). A l'aide d'une macro d'échantillonnage équilibré stratifié, nous sélectionnons un échantillon avec des probabilités d'inclusion égales à $\frac{1}{5}$ en équilibrant sur les variables de la table 4.1. Nous souhaitons donc obtenir un échantillon qui soit

→ globalement équilibré sur l'ensemble de la grande commune,
→ approximativement équilibré dans chaque strate,
→ de taille fixe dans chaque strate.

Nous obtenons un échantillon de 5 296 unités en quelques secondes. La table 4.2 compare les tailles d'échantillon obtenues dans chaque strate avec les tailles d'échantillon théoriques. En les arrondissant, la condition de taille fixe est parfaitement réalisée. Les estimations sur l'ensemble de la commune sont présentées dans la table 4.3. L'équilibrage global est parfaitement réalisé.

Tab. 4.1 – Liste des variables socio-démographiques

NLOG	Nombre de logements
NLOGCO	Nombre de logements en adresse collective
H0019	Nombre d'hommes de moins de 20 ans
H2039	Nombre d'hommes de 20 à 39 ans
H4059	Nombre d'hommes de 40 à 59 ans
H6074	Nombre d'hommes de 60 à 74 ans
H7599	Nombre de femmes de 75 ans et plus
F0019	Nombre de femmes de moins de 20 ans
F2039	Nombre de femmes de 20 à 39 ans
F4059	Nombre de femmes de 40 à 59 ans
F6074	Nombre de femmes de 60 à 74 ans
F7599	Nombre de femmes de 75 ans et plus
ACTIFS	Nombre d'actifs
INACTIFS	Nombre d'inactifs
NATFN	Nombre de français de naissance
NATFA	Nombre de français par acquisition
NATHE	Nombre d'étrangers (hors Union Européenne)
NATUE	Nombre d'étrangers de l'Union Européenne

En ce qui concerne les strates, l'équilibrage est également très bien respecté. Nous pourrions réaliser un échantillonnage similaire en tirant un échantillon directement dans l'ensemble de la population. Nous devrions alors utiliser les variables d'équilibrage suivantes :

→ la probabilité d'inclusion (pour obtenir un échantillon de taille fixe) et les 18 variables socio-démographiques pour obtenir un équilibrage global. En prenant en compte les colinéarités, cela représente 17 variables d'équilibrage.

→ Une variable indiquant l'appartenance à chaque strate, pour obtenir une taille fixe d'échantillon par strate. Cela représente 35 variables d'équilibrage.

→ Des variables égales au produit des variables socio-demographiques (18) et des indicatrices d'appartenance à une strate (36) pour obtenir un équilibrage stratifié. En prenant en compte les colinéarités, cela représente $16 \times 35 = 560$ variables d'équilibrage.

TAB. 4.2 – Comparaison entre les tailles d'échantillon effectives et théoriques obtenues dans chaque strate

Strate	1	2	3	4	5	6	7	8	9
Taille voulue	155.8	137.8	148.6	141	148.8	142.4	145.6	149.6	141.8
Taille obtenue	155	138	148	141	148	142	146	150	142

Strate	10	11	12	13	14	15	16	17	18
Taille voulue	147.6	146.4	144.8	153.4	140	141.8	136.8	144	147.2
Taille obtenue	148	147	145	153	140	142	137	144	148

Strate	19	20	21	22	23	24	25	26	27
Taille voulue	150.6	147.2	150.4	153.4	145.4	153.4	151.2	146	144
Taille obtenue	150	147	150	154	146	153	151	146	144

Strate	28	29	30	31	32	33	34	35	36
Taille voulue	155.8	145.8	147.4	151.4	143.8	153.8	142.4	152	146.8
Taille obtenue	156	145	148	152	144	154	143	152	147

Nous aurions donc eu besoin de 612 variables d'équilibrage. L'échantillonnage aurait été beaucoup plus lent, et l'équilibrage de mauvaise qualité dans certaines strates.

4.3 Bootstrap d'un échantillon équilibré sur une variable

Nous supposons dans ce paragraphe que l'on utilise une seule variable d'équilibrage x_k. Nous utilisons une technique développée par Hájek (1981) pour l'approximation de précision d'un tirage réjectif. L'hypothèse clé dont nous avons besoin est que, dans le cas d'un tirage poissonien, la fonction de densité du π-estimateur admet un développement d'Edgeworth à deux termes.
Dans le cas d'un tirage poissonien, les hypothèses asymptotiques H2-H4 (voir chapitre 1, paragraphe 2.3) sont satisfaites, ce qui implique l'existence d'un théorême central limite ; un développement d'Edgeworth peut également être obtenu sous quelques hypothèses techniques, voir Zhidong and Lincheng (1986). Nous supposons ici que ce développement peut être appliqué à la fonc-

TAB. 4.3 – Différence relative entre le vrai total et l'estimateur de Horvitz-Thompson du total pour les variables d'équilibrage

Variable	Estimateur de Horvitz-Thompson du total	Vrai total	Différence relative (%)
NLOG	251 275	251 199	+0.03%
NLOGCO	251 275	251 198	+0.03%
H0019	46 790	46 759	+0.07%
H2039	74 735	74 729	+0.01%
H4059	45 800	45 763	+0.08%
H6074	20 875	20 870	+0.02%
H7599	10 315	10 316	-0.01%
F0019	45 850	45 852	-0.00%
F2039	84 045	84 022	+0.03%
F4059	51 475	51 455	+0.04%
F6074	28 770	28 739	+0.11%
F7599	21 465	21 484	-0.09%
ACTIFS	360 065	359 984	+0.02%
INACTIFS	70 055	70 005	+0.07%
NATFN	323 305	323 219	+0.03%
NATFA	17 460	17 450	+0.06%
NATHE	8 995	8 990	+0.06%
NATUE	80 360	80 330	+0.04%

tion de densité du π-estimateur de total, et que le manque de continuité de ce π-estimateur est négligeable quand la taille n d'échantillon devient grande.

Nous supposerons également que la variable d'équilibrage est bornée :

$$\exists M_x \qquad \forall k \qquad |x_k| \leq M_x, \tag{4.20}$$

ainsi que la variable donnant les probabilités d'inclusion :

$$\exists \epsilon_1, \epsilon_2 \qquad \forall k \qquad \epsilon_1 \leq \pi_k \leq \epsilon_2. \tag{4.21}$$

Ces deux hypothèses assurent que $d_0 = \sum_{k \in U} \left(\frac{x_k}{\pi_k}\right)^2 \pi_k(1-\pi_k)$ est asymptotiquement équivalent à $n = \sum_{k \in U} \pi_k$, ce qui permet d'alléger les notations en écrivant les ordres de grandeur en fonction de n.

4.3.1 Approximation des probabilités d'inclusion

Nous établissons ici un résultat faisant le lien entre les probabilités d'inclusion d'un tirage équilibré à entropie maximale et les proabbilités d'inclusion du tirage poissonien canoniquement associé. Ce théorême constitue essentiellement un résultat intrmédiaire qui nous servira plus loin pour établir une expression approchée des probabilités d'inclusion doubles d'un tirage équilibré en fonction des probabilités d'inclusion d'ordre 1.

Théorème 4.1. *Soit p le plan à entropie maximale, de probabilités d'inclusion $\boldsymbol{\pi}$, équilibré sur la variable x. Soit q le plan poissonien canoniquement associé à p. Alors :*

$$\pi_k = p_k \left(1 - \frac{1}{2} \left(\frac{x_k}{\pi_k} \right)^2 \frac{(1-p_k)(1-2p_k)}{d} + \frac{x_k}{\pi_k} \frac{1-p_k}{d} \frac{c_1}{2} + o(n^{-1}) \right) \quad (4.22)$$

avec

$$c_1 = d^{-1} \sum_{l \in U} \left(\frac{x_l}{\pi_l} \right)^3 p_l(1-p_l)(1-2p_l) \qquad d = \sum_{l \in U} \left(\frac{x_l}{\pi_l} \right)^2 p_l(1-p_l)$$

et $n\ o(n^{-1}) \to 0$ uniformément en tout $k \in U$.

Démonstration. Soit $E_q(.)$ (respectivement $V_q(.)$) l'espérance (respectivement la variance) sous le plan de Poisson q. Soit I_k l'indicatrice d'appartenance de l'unité k à l'échantillon poissonien $\tilde{S}$, et

$$\begin{aligned} \widehat{t_{x\pi}} &= \textstyle\sum_{k \in U} \frac{x_k}{\pi_k} I_k \\ &= \textstyle\sum_{k \in U} \frac{\tilde{x_k}}{p_k} I_k \quad \text{avec} \quad \tilde{x_k} = \frac{p_k}{\pi_k} x_k \\ &= \widehat{t_{\tilde{x}p}} \end{aligned}$$

Pour alléger l'écriture, on notera simplement

$$y = \widehat{t_{\tilde{x}p}} \text{ et } z = \frac{y - t_{\tilde{x}}}{\sqrt{V_q(\widehat{t_{\tilde{x}p}})}}.$$

On a de façon immédiate :

$$\begin{aligned} E_q(y) &= t_{\tilde{x}} \\ &= \textstyle\sum_{k\in U} \frac{p_k}{\pi_k} x_k \\ &= t_x \qquad\qquad \text{car le plan } q \text{ est canonique.} \end{aligned}$$

$$V_q(y) = \sum_{k\in U} \left(\frac{\tilde{x_k}}{p_k}\right)^2 p_k(1-p_k) = \sum_{k\in U} \left(\frac{x_k}{\pi_k}\right)^2 p_k(1-p_k) = d$$

L'identité

$$\begin{aligned} \left(\textstyle\sum_{k\in U} a_k\right)^3 &= \textstyle\sum_{k\in U}\ a_k^3 \\ &+ \textstyle 3\sum_{k\in U}\sum_{l\neq k\in U}\ a_k^2 a_l \\ &+ \textstyle\sum_{k\in U}\sum_{l\neq k\in U}\sum_{m\neq k,l\in U}\ a_k a_l a_m \end{aligned}$$

nous donne

$$\begin{aligned} (y-t_x)^3 &= \left(\textstyle\sum_{k\in U} \frac{x_k}{\pi_k}(I_k - p_k)\right)^3 \\ &= \textstyle\sum_{k\in U} \left(\frac{x_k}{\pi_k}\right)^3 (I_k-p_k)^3 \\ &+ \textstyle 3\sum_{k\in U}\sum_{l\neq k\in U}\ \left(\frac{x_k}{\pi_k}\right)^2 \left(\frac{x_l}{\pi_l}\right)(I_k-p_k)^2(I_l-p_l) \\ &+ \textstyle\sum_{k\in U}\sum_{l\neq k\in U}\sum_{m\neq k,l\in U}\ \frac{x_k}{\pi_k}\frac{x_l}{\pi_l}\frac{x_m}{\pi_m}(I_k-p_k)(I_l-p_l)(I_m-p_m) \\ \Rightarrow\quad E_q\,(y-t_x)^3 &= \textstyle\sum_{k\in U} \left(\frac{x_k}{\pi_k}\right)^3 E_q(I_k-p_k)^3 \\ &= \textstyle\sum_{k\in U} \left(\frac{x_k}{\pi_k}\right)^3 p_k(1-p_k)(1-2p_k) \end{aligned}$$

que l'on note $\mu_3(y)$.

De la même façon, l'identité

$$\begin{aligned}\left(\sum_{k\in U} a_k\right)^4 &= \sum_{k\in U} a_k^4 \\ &+ 4\sum_{k\in U}\sum_{l\neq k\in U} a_k^3 a_l \\ &+ 3\sum_{k\in U}\sum_{l\neq k\in U} a_k^2 a_l^2 \\ &+ 6\sum_{k\in U}\sum_{l\neq k\in U}\sum_{m\neq k,l\in U} a_k^2 a_l a_m \\ &+ \sum_{k\in U}\sum_{l\neq k\in U}\sum_{m\neq k,l\in U}\sum_{o\neq k,l,m\in U} a_k a_l a_m a_o\end{aligned}$$

nous donne avec un peu de calcul

$$\begin{aligned}E_q\,(y-t_x)^4 &= E_q\Big[\sum_{k\in U}\left(\tfrac{x_k}{\pi_k}\right)^4 (I_k-p_k)^4 \\ &\qquad + 3\sum_{k\neq l\in U}\left(\tfrac{x_k}{\pi_k}\right)^2\left(\tfrac{x_l}{\pi_l}\right)^2 (I_k-p_k)^2(I_l-p_l)^2\Big] \\ &= \sum_{k\in U}\left(\tfrac{x_k}{\pi_k}\right)^4 p_k(1-p_k)(1-6p_k(1-p_k)) + 3d^2\end{aligned}$$

que l'on note μ_4.

Soit f la densité de la variable aléatoire z. Alors, de façon similaire à Hájek (1981) :

$$\begin{aligned}\pi_k &= \mathbb{P}(I_k=1|x=0) \\ &= p_k\tfrac{f^{I_k=1}(0)}{f(0)}\end{aligned} \tag{4.23}$$

où $f^{I_k=1}(.)$ désigne la densité de z conditionnellement à $I_k=1$. En utilisant le développement d'Edgeworth formel d'une fonction de densité (Hall, 1992), on obtient

$$
\begin{aligned}
f(0) &= \frac{1}{\sqrt{2\pi d}} \left(1 + \frac{1}{8} d^{-2} \sum_{k \in U} \left(\frac{x_k}{\pi_k} \right)^4 p_k(1-p_k)(1-6p_k(1-p_k)) \right. \\
&\quad - \frac{5}{24} d^{-3} \left(\sum_{k \in U} \left(\frac{x_k}{\pi_k} \right)^3 p_k(1-p_k)(1-2p_k) \right)^2 \\
&\quad \left. + o(n^{-1}) \right) \\
&= \frac{1}{\sqrt{2\pi d}} \left(1 + c_0 d^{-1} + o(n^{-1}) \right)
\end{aligned}
$$

où

$$
\begin{aligned}
c_0 &= \frac{1}{8} d^{-1} \sum_{k \in U} \left(\frac{x_k}{\pi_k} \right)^4 p_k(1-p_k)(1-6p_k(1-p_k)) \\
&\quad - \frac{5}{24} d^{-2} \left(\sum_{k \in U} \left(\frac{x_k}{\pi_k} \right)^3 p_k(1-p_k)(1-2p_k) \right)^2
\end{aligned}
$$

En conditionnant par rapport à $I_k = 1$, on obtient également

$$
\begin{aligned}
E_q^{I_k=1}(y) &= \sum_{l\neq k\in U} \frac{p_l}{\pi_l} x_l + \frac{x_k}{p_k} \\
&= \sum_{l\in U} \frac{p_l}{\pi_l} x_l + \frac{x_k}{\pi_k}(1-p_k), \\
V_q^{I_k=1}(y) &= \sum_{l\neq k\in U} \left(\frac{x_l}{\pi_l}\right)^2 p_l(1-p_l) \\
&= d - \left(\frac{x_k}{\pi_k}\right)^2 p_k(1-p_k), \\
\mu_3^{I_k=1}(y) &= \mu_3 - \left(\frac{x_k}{\pi_k}\right)^3 p_k(1-p_k)(1-2p_k), \\
\mu_4^{I_k=1}(y) &= \mu_4 - \left(\frac{x_k}{\pi_k}\right)^4 p_k(1-p_k)(1-6p_k(1-p_k)) \\
&+ 3\left(d - \left(\frac{x_k}{\pi_k}\right)^2 p_k(1-p_k).\right)^2
\end{aligned}
$$

Et en réappliquant le développement d'Edgeworth de la fonction de densité :

$$
\begin{aligned}
f^{I_k=1}(0) &= \frac{1}{\sqrt{2\pi\left(d-(x_k/\pi_k)^2 p_k(1-p_k)\right)}} \times \exp\left(-\frac{1}{2}\frac{(x_k/\pi_k)^2(1-p_k)^2}{d-(x_k/\pi_k)^2 p_k(1-p_k)}\right) \\
&\times \left(1 + \frac{1}{6}(\delta^3 - 3\delta)\frac{\sum_{l\neq k\in U}(x_l/\pi_l)^3 p_l(1-p_l)(1-2p_l)}{\left(d-(x_k/\pi_k)^2 p_k(1-p_k)\right)^{3/2}}\right. \\
&\quad + \frac{1}{4!}(\delta^4 - 6\delta^2 + 3)\frac{\sum_{l\neq k\in U}(x_l/\pi_l)^4 p_l(1-p_l)(1-6p_l(1-p_l))}{\left(d-(x_k/\pi_k)^2 p_k(1-p_k)\right)^2} \\
&\quad + \frac{10}{6!}(\delta^6 - 15\delta^4 + 45\delta^2 - 15)\frac{\left(\sum_{l\neq k\in U}(x_l/\pi_l)^3 p_l(1-p_l)(1-2p_l)\right)^2}{\left(d-(x_k/\pi_k)^2 p_k(1-p_k)\right)^2} \\
&\quad \left. + o(n^{-1})\right)
\end{aligned}
$$

avec

$$
\delta = \frac{t_x - t_{\tilde{x}} - \frac{x_k}{\pi_k}(1-p_k)}{\sqrt{d-(x_k/\pi_k)^2 p_k(1-p_k)}} = -\frac{(x_k/\pi_k)(1-p_k)}{\sqrt{d-(x_k/\pi_k)^2 p_k(1-p_k)}}.
$$

En simplifiant, on obtient

$$\begin{aligned} f^{I_k=1}(0) &= \frac{1}{\sqrt{2\pi d}}\Bigg(1+\tfrac{1}{2}d^{-1}\left(x_k/\pi_k\right)^2 p_k(1-p_k) \\ &\qquad -\tfrac{1}{2}d^{-1}\left(x_k/\pi_k\right)^2(1-p_k)^2+\tfrac{1}{2}c_1 d^{-1}\left(x_k/\pi_k\right)(1-p_k) \\ &\qquad +c_0 d^{-1}+o(n^{-1})\Bigg) \end{aligned}$$

avec

$$c_1 = d^{-1}\sum_{l\in U}\left(x_l/\pi_l\right)^3 p_l(1-p_l)(1-2p_l).$$

En injectant les deux expressions obtenues pour $f(0)$ et $f^{I_k=1}(0)$ dans 4.23, on obtient :

$$\begin{aligned} \pi_k &= p_k\Bigg[1+\tfrac{1}{2}d^{-1}\left(x_k/\pi_k\right)^2 p_k(1-p_k)-\tfrac{1}{2}d^{-1}\left(x_k/\pi_k\right)^2(1-p_k)^2 \\ &\qquad +\tfrac{1}{2}c_1 d^{-1}\left(x_k/\pi_k\right)(1-p_k)+c_o d^{-1}+o(n^{-1})\Bigg] \\ &\qquad \times\Bigg[1-c_0 d^{-1}+o(n^{-1})\Bigg] \\ &= p_k\Bigg[1-\tfrac{1}{2}d^{-1}\left(x_k/\pi_k\right)^2(1-p_k)(1-2p_k) \\ &\qquad +\tfrac{1}{2}c_1 d^{-1}\left(x_k/\pi_k\right)(1-p_k)+o(n^{-1})\Bigg] \end{aligned}$$

□

Remarque 4.4.

1. *Si $x_k = \pi_k$, on retrouve le théorême 3.5 dû à Hájek.*

2. *On peut obtenir une formule analogue, même si le plan q n'est pas canoniquement associé à p. On a alors :*

$$\begin{aligned}\pi_k &= p_k\Big(1-\tfrac{1}{2}\left(\tfrac{x_k}{\pi_k}\right)^2\tfrac{(1-p_k)(1-2p_k)}{d}\\ &\quad+\tfrac{x_k}{\pi_k}\tfrac{1-p_k}{d}\left(\epsilon+\tfrac{c_1}{2}\right)+o(n^{-1})\Big)\end{aligned}$$

avec $\epsilon = t_x - t_{\tilde{x}}$.

Le théorème que nous établissons ci-dessous généralise l'approximation de Hajek pour les probabilités d'inclusion d'ordre 2 d'un tirage réjectif au cas d'un tirage équilibré à entropie maximale.

Théorème 4.2. *Soit p le plan à entropie maximale, de probabilités d'inclusion $\boldsymbol{\pi}$, équilibré sur la variable x, et q le plan poissonien canoniquement associé à p. Alors avec les notations précédentes :*

$$\pi_{kl} = \pi_k\pi_l\left(1-\frac{x_k}{\pi_k}\frac{x_l}{\pi_l}\frac{(1-\pi_k)(1-\pi_l)}{d_0}+o(n^{-1})\right)$$

avec

$$d_0 = \sum_{l\in U}\left(\frac{x_l}{\pi_l}\right)^2\pi_l(1-\pi_l)$$

et $n\ o(n^{-1}) \to 0$ uniformément pour tous $k, l \in U$.

Démonstration. On garde les mêmes notations que dans la preuve précédente. Un développement analogue à celui de Hájek (1981) donne

$$\pi_{kl} = \mathbb{P}(I_k = 1, I_l = 1)\frac{f^{I_k=I_l=1}(0)}{f(0)} \tag{4.24}$$

où $f^{I_k=I_l=1}(.)$ désigne la densité de z conditionnellement à $I_k = I_l = 1$.En utilisant le développement d'Edgeworth de la densité f, on obtient :

$$
\begin{aligned}
f^{I_k=I_l=1}(0) &= \frac{1}{\sqrt{2\pi\left(d-(x_k/\pi_k)^2p_k(1-p_k)-(x_l/\pi_l)^2p_l(1-p_l)\right)}} \\
&\times \exp\left(-\frac{1}{2}\frac{\left((x_k/\pi_k)(1-p_k)+(x_l/\pi_l)(1-p_l)\right)^2}{d-(x_k/\pi_k)^2p_k(1-p_k)-(x_l/\pi_l)^2p_l(1-p_l)}\right) \\
&\times \left(1+\frac{1}{6}(\eta^3-3\eta)\frac{\sum_{m\neq k,l\in U}(x_m/\pi_m)^3p_m(1-p_m)(1-2p_m)}{\left(d-(x_k/\pi_k)^2p_k(1-p_k)-(x_l/\pi_l)^2p_l(1-p_l)\right)^{3/2}}\right. \\
&\quad + \frac{1}{4!}(\eta^4-6\eta^2+3)\frac{\sum_{m\neq k,l\in U}(x_m/\pi_m)^4p_m(1-p_m)(1-6p_m(1-p_m))}{\left(d-(x_k/\pi_k)^2p_k(1-p_k)-(x_l/\pi_l)^2p_l(1-p_l)\right)^2} \\
&\quad + \frac{10}{6!}(\eta^6-15\eta^4+45\eta^2-15)\frac{\left(\sum_{m\neq k,l\in U}(x_m/\pi_m)^3p_m(1-p_m)(1-2p_m)\right)^2}{\left(d-(x_k/\pi_k)^2p_k(1-p_k)-(x_l/\pi_l)^2p_l(1-p_l)\right)^2} \\
&\quad \left. + o(n^{-1})\right)
\end{aligned}
$$

avec

$$
\eta = -\frac{\frac{x_k}{\pi_k}(1-p_k)+\frac{x_l}{\pi_l}(1-p_l)}{\sqrt{d-(x_k/\pi_k)^2p_k(1-p_k)-(x_l/\pi_l)^2p_l(1-p_l)}}.
$$

En simplifiant, on obtient

$$
\begin{aligned}
f^{I_k=I_l=1}(0) &= \frac{1}{\sqrt{2\pi d}}\left(1+\frac{1}{2}d^{-1}(x_k/\pi_k)^2p_k(1-p_k)+\frac{1}{2}d^{-1}(x_l/\pi_l)^2p_l(1-p_l)\right. \\
&\quad -\frac{1}{2}d^{-1}(x_k/\pi_k)^2(1-p_k)^2-\frac{1}{2}d^{-1}(x_l/\pi_l)^2(1-p_l)^2 \\
&\quad +\frac{1}{2}c_1d^{-1}(x_k/\pi_k)(1-p_k)+\frac{1}{2}c_1d^{-1}(x_l/\pi_l)(1-p_l) \\
&\quad -d^{-1}(x_k/\pi_k)(x_l/\pi_l)(1-p_k)(1-p_l) \\
&\quad \left.+c_0d^{-1}+o(n^{-1})\right).
\end{aligned}
$$

En utilisant le développement de $f(0)$ donné dans la preuve précédente, on a :

$$\begin{aligned}
\frac{f^{I_k=I_l=1}(0)}{f(0)} &= \Bigg[1 - \tfrac{1}{2}d^{-1}\left(x_k/\pi_k\right)^2(1-p_k)(1-2p_k) \\
&\quad - \tfrac{1}{2}d^{-1}\left(x_l/\pi_l\right)^2(1-p_l)(1-2p_l) \\
&\quad - d^{-1}\left(x_k/\pi_k\right)\left(x_l/\pi_l\right)(1-p_k)(1-p_l) \\
&\quad + \tfrac{1}{2}c_1^{-1}\Big(\left(x_k/\pi_k\right)(1-p_k) + \left(x_l/\pi_l\right)(1-p_l)\Big) \\
&\quad + o(n^{-1})\Bigg]
\end{aligned} \tag{4.25}$$

D'autre part, en utilisant le théorème 4.1, on obtient

$$\begin{aligned}
\mathbb{P}(I_k=1, I_l=1) &= p_k p_l \\
&= \pi_k\pi_l\Bigg[1 + \tfrac{1}{2}d^{-1}\left(x_k/\pi_k\right)^2(1-p_k)(1-2p_k) \\
&\quad + \tfrac{1}{2}d^{-1}\left(x_l/\pi_l\right)^2(1-p_l)(1-2p_l) \\
&\quad - \tfrac{1}{2}c_1 d^{-1}\left(\left(x_k/\pi_k\right)(1-p_k) + \left(x_l/\pi_l\right)(1-p_l)\right) \\
&\quad + o(n^{-1})\Bigg]
\end{aligned} \tag{4.26}$$

En injectant 4.25 et 4.26 dans 4.24, on obtient

$$\pi_{kl} = \pi_k\pi_l\left(1 - \frac{x_k}{\pi_k}\frac{x_l}{\pi_l}\frac{(1-p_k)(1-p_l)}{d} + o(n^{-1})\right)$$

Le théorème 4.1 implique que l'on peut remplacer dans l'expression précédente p_k et p_l par π_k et π_l respectivement, et $d = \sum_{l\in U}\left(\frac{x_l}{\pi_l}\right)^2 p_l(1-p_l)$ par

$d_0 = \sum_{l \in U} \left(\frac{x_l}{\pi_l}\right)^2 \pi_l(1-\pi_l)$, d'où le résultat.

□

Remarque 4.5.

1. *Si $x_k = \pi_k$, on retrouve là encore le théorème 3.6 dû à Hájek.*
2. *On peut obtenir une formule analogue, même si le plan q n'est pas canoniquement associé à p. On a alors :*

$$\begin{aligned}\pi_{kl} &= \pi_k\pi_l\Bigg(1 - \epsilon d^{-1}\left(\tfrac{x_k}{\pi_k}(1-\pi_k) + \tfrac{x_l}{\pi_l}(1-\pi_l)\right) \\ &\quad - (1-\epsilon)d^{-1}\tfrac{x_k}{\pi_k}\tfrac{x_l}{\pi_l}(1-\pi_k)(1-\pi_l) \\ &\quad + o(n^{-1})\Bigg)\end{aligned}$$

avec $\epsilon = t_x - t_{\tilde{x}}$.

Remarque 4.6.

Les hypothèses 4.20 et 4.3 impliquent que l'on peut encore écrire l'approximation des probabilités d'ordre 2 sous la forme

$$\pi_{kl} = \pi_k\pi_l\left(1 - \frac{x_k}{\pi_k}\frac{x_l}{\pi_l}\frac{(1-\pi_k)(1-\pi_l)}{d_0}\right)(1 + o(n^{-1})) \tag{4.27}$$

où n $o(n^{-1} \to 0$ uniformément pour tous $k, l \in U$

4.3.2 Approximation de variance et bootstrappabilité

Dans le cas d'un tirage équilibré sur une seule variable x_k, Ardilly and Tillé (2003) donnent une formule exacte de variance :

$$V(\widehat{t_{y\pi}}) = \frac{1}{2}\sum_{k \neq l \in U}\left(\frac{y_k}{x_k} - \frac{y_l}{x_l}\right)^2 x_k x_l \frac{\pi_k\pi_l - \pi_{kl}}{\pi_k\pi_l}. \tag{4.28}$$

En injectant dans la formule précédente l'approximation des probabilités d'inclusion d'ordre 2 donnée par 4.27, on obtient l'approximation de variance

$$\begin{aligned} V_{app,1}(\widehat{t_{y\pi}}) &= \tfrac{1}{2}\sum_{k\neq l\in U}\left(\tfrac{y_k}{x_k}-\tfrac{y_l}{x_l}\right)^2 x_k x_l \tfrac{x_k}{\pi_k}\tfrac{x_l}{\pi_l}\tfrac{(1-\pi_k)(1-\pi_l)}{d_0} \\ &= \sum_{k\in U}\left(\tfrac{1}{\pi_k}-1\right)(y_k - Rx_k)^2 \end{aligned} \tag{4.29}$$

après un peu de calcul, avec

$$R = \frac{\sum_{l\in U}\frac{y_l}{\pi_l}\frac{x_l}{\pi_l}\pi_l(1-\pi_l)}{\sum_{l\in U}\pi_l(1-\pi_l)}.$$

En utilisant les résultats précédents et la remarque 4.6, on en déduit le théorème suivant :

Théorème 4.3.

Soit p le plan de probabilités d'inclusion $\boldsymbol{\pi} = (\pi_k)_{k\in U}$, équilibré sur une variable x quelconque, et à entropie maximale parmi les plans vérifiant ces deux conditions. Soit y une variable quelconque. Alors la variance est donnée asymptotiquement par

$$V(\widehat{t_{y\pi}}) = (1+o(1))\sum_{k\in U}\left(\frac{1}{\pi_k}-1\right)(y_k - Rx_k)^2$$

avec les notations précédentes.

On retrouve ainsi l'approximation de variance donnée par le théorème 3.7 de Hájek dans le cas particulier du tirage réjectif ($x_k = \pi_k$), et l'approximation 4.11 de Deville et Tillé dans le cas d'une variable x quelconque.

Ce résultat est fondamental car il assure que la variance d'un estimateur de total peut, dans le cas d'un tirage équilibré à entropie maximale, être approché asymptotiquement sans biais par une fonctionnelle qui s'écrit comme une fonction de totaux. En vertu des résultats du paragraphe 3.2., cela assure que la méthode de Bootstrap proposée peut être utilisée dans ce cas pour estimer la variance. On peut résumer ces résultats à l'aide de la propriété suivante.

Propriété 4.4. ***Le plan équilibré à entropie maximale est bootstrappable.***

Compte-tenu des résultats de la section 2 du chapitre 3, on en déduit également que le plan à entropie maximale équilibré sur une variable est bootstrappable.

Hájek remarque que, dans le cas du tirage réjectif, l'approximation

$$\pi_{kl} - \pi_k \pi_l \simeq -\frac{\pi_k(1-\pi_k)\pi_l(1-\pi_l)}{\sum_{m\in U}\pi_m(1-\pi_m)} = -c'_k c'_l \tag{4.30}$$

ne permet pas d'obtenir

$$\sum_{l;l\neq k\in U} c'_k c'_l = \pi_k(1-\pi_k) \qquad \forall k \in U.$$

Cela signifie qu'en injectant l'approximation 4.30 des probabilités d'inclusion doubles dans la formule de variance de Horvitz-Thompson, on obtient une variance non nulle pour l'estimateur du total de la variable probabilité d'inclusion (alors que cet estimateur est constant égal à n car le tirage réjectif est de taille fixe). Hájek propose donc d'utiliser une approximation plus serrée, donnée par l'équation 3.5, que nous avons évoquée au chapitre précédent. En nous inspirant de son raisonnement, on peut partir de l'approximation

$$\pi_{kl} - \pi_k \pi_l \simeq -\frac{x_k}{\pi_k}\frac{x_l}{\pi_l}\frac{\pi_k(1-\pi_k)\pi_l(1-\pi_l)}{d_0}$$

et chercher une approximation plus serrée de la forme

$$\pi_{kl} - \pi_k \pi_l \simeq -\frac{x_k}{\pi_k}\frac{x_l}{\pi_l}\frac{b_k b_l}{d_b} \tag{4.31}$$

avec $d_b = \sum_{m\in U}\left(\frac{x_m}{\pi_m}\right)^2 b_m$, et permettant d'annuler la formule de variance de Horvitz-Thompson pour l'estimation du total de la variable x. En injectant l'approximation 4.31 dans la formule de variance de Horvitz-Thompson, on obtient l'approximation de variance

$$V_{app,2}(\widehat{t_{y\pi}}) = \sum_{k\in U}\left(\frac{y_k}{\pi_k}\right)^2 \pi_k(1-\pi_k) - \sum_{k\neq l\in U}\left(\frac{y_k}{\pi_k}\right)\left(\frac{y_l}{\pi_l}\right)\left(\frac{x_k}{\pi_k}\right)\left(\frac{x_l}{\pi_l}\right)\frac{b_k b_l}{d_b}. \tag{4.32}$$

Pour que cette approximation s'annule dans le cas de l'estimation du total t_x pour une variable x quelconque, il est nécessaire que

$$\begin{array}{lll} & V_{app,2}(\widehat{t_{x\pi}}) & = 0 \\ \Leftrightarrow & \sum_{k\in U}\left(\frac{x_k}{\pi_k}\right)^2\pi_k(1-\pi_k) - \sum_{k\neq l\in U}\left(\frac{x_k}{\pi_k}\right)^2\left(\frac{x_l}{\pi_l}\right)^2\frac{b_k b_l}{d_b} & = 0 \\ \Leftrightarrow & \sum_{k\in U}\left(\frac{x_k}{\pi_k}\right)^2\pi_k(1-\pi_k) - \sum_{k\in U} b_k\left(\frac{x_k}{\pi_k}\right)^2 + \sum_{k\in U}\frac{b_k^2}{d_b}\left(\frac{x_k}{\pi_k}\right)^4 & = 0 \\ \Leftrightarrow & \forall k \quad \pi_k(1-\pi_k) - b_k + \frac{b_k^2}{d_b}\left(\frac{x_k}{\pi_k}\right)^2 & = 0 \end{array}$$

$$\Leftrightarrow \quad \forall k \quad b_k\left[1-\left(\frac{x_k}{\pi_k}\right)\left(\sum_{l\in U}\left(\frac{x_l}{\pi_l}\right)^2 b_l\right)^{-1}\left(\frac{x_k}{\pi_k}\right)b_k\right] = \pi_k(1-\pi_k)$$

ce qui correspond à l'approximation 4.14 proposée par Deville and Tillé (2005). Si cette équation itérative n'admet pas de solution, on peut n'utiliser qu'une itération ce qui conduit à l'approximation de variance donnée par 4.15.

4.4 Bootstrap d'un échantillon équilibré : cas général

4.4.1 Approximation des probabilités d'inclusion

Nous nous plaçons ici dans le cadre général d'un échantillonnage équilibré sur p variables $\mathbf{x} = (x_1, \ldots, x_p)'$. Une approximation de variance analogue à celle donnée par Deville and Tillé (2005) peut être formellement développée suivant une technique analogue à celle utilisée dans la section précédente, et basée sur un développement d'Edgeworth de la fonction de densité multivariée du π-estimateur $\widehat{t_{\mathbf{x}\pi}}$ du total de x, voir par exemple Skovgaard (1986).

On obtient :

$$\begin{array}{rlll} & \pi_{kl} & \simeq & \pi_k\pi_l\left[1-\frac{\mathbf{x'}_k}{\pi_k}(1-\pi_k)D^{-1}(1-\pi_l)\frac{\mathbf{x}_l}{\pi_l}\right] \\ \Rightarrow & \pi_k\pi_l - \pi_{kl} & \simeq & \frac{\mathbf{x'}_k}{\pi_k}\pi_k(1-\pi_k)D^{-1}\pi_l(1-\pi_l)\frac{\mathbf{x}_l}{\pi_l} \end{array} \tag{4.33}$$

avec

$$D = \sum_{k\in U}\frac{\mathbf{x}_k\mathbf{x'}_k}{\pi_k^2}\pi_k(1-\pi_k).$$

4.4.2 Approximation de variance et bootstrappabilité

Cette approximation des probabilités d'inclusion doubles nous permet de produire des expressions approchée de la variance. On utilise la régression de la variable y sur les variables auxiliaires $\mathbf{x}$

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

en utilisant, pour des raisons techniques, une structure de variance covariance

$$W = \begin{pmatrix} \frac{1-\pi_1}{\pi_1} & & 0 \\ & \ddots & \\ 0 & & \frac{1-\pi_N}{\pi_N} \end{pmatrix}.$$

On obtient :

$$\alpha = D^{-1} \sum_{k \in U} b_k \frac{\mathbf{x}_k}{\pi_k} \frac{y_k}{\pi_k} \tag{4.34}$$

et

$$\begin{aligned} \epsilon_k &= y_k - \widehat{y_k} \\ &= y_k - \mathbf{x}_k' \alpha \qquad (4.35) \\ &= y_k - \mathbf{x'}_k D^{-1} \textstyle\sum_{k \in U} b_k \frac{\mathbf{x}_k}{\pi_k} \frac{y_k}{\pi_k}. \end{aligned}$$

où $\widehat{y_k}$ désigne le prédicteur de y_k. Notons que

$$\sum_{l \in U} \frac{\mathbf{x}_l (y_l - \mathbf{x}_l' \alpha)}{\pi_l^2} \pi_l (1 - \pi_l) = 0. \tag{4.36}$$

Notons également que l'équilibrage sur $\mathbf{x}$ assure que :

$$V(\widehat{t_{y\pi}}) = V(\widehat{t_{\epsilon\pi}}).$$

Pour obtenir une approximation de variance, une première possibilité consiste à injecter l'approximation 4.33 des probabilités d'inclusion doubles dans la formule de variance de Horvitz-Thompson. On obtient alors :

$$
\begin{aligned}
V(\widehat{t_{y\pi}}) &= V(\widehat{t_{\epsilon\pi}}) \\
&= \sum_{k\in U}\left(\frac{\epsilon_k}{\pi_k}\right)^2\pi_k(1-\pi_k)+\sum_{k\in U}\sum_{l\neq k\in U}\frac{\epsilon_k}{\pi_k}\frac{\epsilon_l}{\pi_l}(\pi_{kl}-\pi_k\pi_l) \\
&\simeq \sum_{k\in U}\left(\frac{\epsilon_k}{\pi_k}\right)^2\pi_k(1-\pi_k) \\
&\quad -\sum_{k\in U}\sum_{l\neq k\in U}\frac{\epsilon_k}{\pi_k}\frac{\epsilon_l}{\pi_l}\left(\frac{\mathbf{x'}_k}{\pi_k}\pi_k(1-\pi_k)D^{-1}\pi_l(1-\pi_l)\frac{\mathbf{x}_l}{\pi_l}\right) \\
&= \sum_{k\in U}\left(\frac{\epsilon_k}{\pi_k}\right)^2\left(\pi_k(1-\pi_k)+\pi_k^2(1-\pi_k)^2\frac{\mathbf{x'}_k}{\pi_k}D^{-1}\frac{\mathbf{x}_k}{\pi_k}\right) \\
&- \sum_{k\in U}\sum_{l\in U}\frac{\epsilon_k}{\pi_k}\frac{\epsilon_l}{\pi_l}\left(\frac{\mathbf{x'}_k}{\pi_k}\pi_k(1-\pi_k)D^{-1}\pi_l(1-\pi_l)\frac{\mathbf{x}_l}{\pi_l}\right) \\
&= \sum_{k\in U}\left(\frac{\epsilon_k}{\pi_k}\right)^2\pi_k(1-\pi_k)\left(1+\pi_k(1-\pi_k)\frac{\mathbf{x'}_k}{\pi_k}D^{-1}\frac{\mathbf{x}_k}{\pi_k}\right) \\
&- \left(\sum_{k\in U}\frac{(y_k-\alpha'\mathbf{x}_k)\mathbf{x'}_k}{\pi_k^2}\pi_k(1-\pi_k)\right)D^{-1}\left(\sum_{l\in U}\frac{\mathbf{x}_l(y_l-\mathbf{x'}_l\alpha)}{\pi_l^2}\pi_l(1-\pi_l)\right) \\
&= \sum_{k\in U}\left(\frac{\epsilon_k}{\pi_k}\right)^2\pi_k(1-\pi_k)\left(1+\pi_k(1-\pi_k)\frac{\mathbf{x'}_k}{\pi_k}D^{-1}\frac{\mathbf{x}_k}{\pi_k}\right) \\
&\quad \text{en appliquant 4.36}
\end{aligned}
$$

On obtient donc l'approximation de variance :

$$
V_{app,1}(\widehat{t_{y\pi}}) = \sum_{k\in U}\left(\frac{y_k-\hat{y}_k}{\pi_k}\right)^2\pi_k(1-\pi_k)\left(1+\pi_k(1-\pi_k)\frac{\mathbf{X'}_k}{\pi_k}D^{-1}\frac{\mathbf{X}_k}{\pi_k}\right) \quad (4.37)
$$

Une autre possibilité consiste à injecter l'approximation 4.33 dans la formule de variance de Sen-Yates-Grundy, si le tirage est de taille fixe (autrement dit, si la probabilité d'inclusion fait partie des variables d'équilibrage). On obtient alors :

$$
\begin{aligned}
V(\widehat{t_{y\pi}}) &= V(\widehat{t_{\epsilon\pi}}) \\
&= \tfrac{1}{2}\sum_{k\in U}\sum_{l\in U}\left(\tfrac{\epsilon_k}{\pi_k}-\tfrac{\epsilon_l}{\pi_l}\right)^2(\pi_k\pi_l-\pi_{kl}) \\
&\simeq \tfrac{1}{2}\sum_{k\in U}\sum_{l\in U}\left(\tfrac{\epsilon_k}{\pi_k}-\tfrac{\epsilon_l}{\pi_l}\right)^2\left(\tfrac{\mathbf{x'}_k}{\pi_k}\pi_k(1-\pi_k)D^{-1}\pi_l(1-\pi_l)\tfrac{\mathbf{x}_l}{\pi_l}\right) \\
&= \sum_{k\in U}\left(\tfrac{\epsilon_k}{\pi_k}\right)^2\left(\tfrac{\mathbf{x'}_k}{\pi_k}\pi_k(1-\pi_k)D^{-1}\sum_{l\in U}\pi_l(1-\pi_l)\tfrac{\mathbf{x}_l}{\pi_l}\right) \\
&- \sum_{k\in U}\sum_{l\in U}\tfrac{\epsilon_k}{\pi_k}\tfrac{\epsilon_l}{\pi_l}\left(\tfrac{\mathbf{x'}_k}{\pi_k}\pi_k(1-\pi_k)D^{-1}\pi_l(1-\pi_l)\tfrac{\mathbf{x}_l}{\pi_l}\right) \\
&= \sum_{k\in U}\left(\tfrac{\epsilon_k}{\pi_k}\right)^2\left(\tfrac{\mathbf{x'}_k}{\pi_k}\pi_k(1-\pi_k)D^{-1}\sum_{l\in U}\pi_l(1-\pi_l)\tfrac{\mathbf{x}_l}{\pi_l}\right) \\
&- \left(\sum_{k\in U}\tfrac{(y_k-\alpha'\mathbf{x}_k)\mathbf{x'}_k}{\pi_k^2}\pi_k(1-\pi_k)\right)D^{-1}\left(\sum_{l\in U}\tfrac{\mathbf{x}_l(y_l-\mathbf{x'}_l\alpha)}{\pi_l^2}\pi_l(1-\pi_l)\right) \\
&= \sum_{k\in U}\left(\tfrac{\epsilon_k}{\pi_k}\right)^2\left(\tfrac{\mathbf{x'}_k}{\pi_k}\pi_k(1-\pi_k)D^{-1}\sum_{l\in U}\pi_l(1-\pi_l)\tfrac{\mathbf{x}_l}{\pi_l}\right) \\
&\quad \text{en appliquant à nouveau 4.36}
\end{aligned}
$$

On obtient donc l'approximation de variance :

$$V_{app,2}(\widehat{t_{y\pi}}) = \sum_{k\in U}\left(\frac{y_k-\hat{y}_k}{\pi_k}\right)^2\pi_k(1-\pi_k)\left(\frac{\mathbf{x'}_k}{\pi_k}D^{-1}\sum_{l\in U}\pi_l(1-\pi_l)\frac{\mathbf{x}_l}{\pi_l}\right) \quad (4.38)$$

Notons que $\mathbf{x'}_k D^{-1}\sum_{l\in U}\pi_l(1-\pi_l)\frac{\mathbf{x}_l}{\pi_l}$ correspond au prédicteur de π_k ; comme la probabilité d'inclusion fait partie des variables d'équilibrage, on a

$$\mathbf{x'}_k D^{-1}\sum_{l\in U}\pi_l(1-\pi_l)\frac{\mathbf{x}_l}{\pi_l} = \pi_k$$

et l'approximation de variance se réduit donc à :

$$V_{app,2}(\widehat{t_{y\pi}}) = \sum_{k\in U}\left(\frac{y_k-\hat{y}_k}{\pi_k}\right)^2\pi_k(1-\pi_k) \quad (4.39)$$

où on retrouve la formule 4.11 de Deville and Tillé (2005).

Enfin, on peut obtenir une approximation de variance plus serrée en cherchant une approximation des probabilités d'inclusion doubles sous la forme

$$\pi_k \pi_l - \pi_{kl} \simeq \frac{\mathbf{x'}_k}{\pi_k} b_k D_b^{-1} b_l \frac{\mathbf{x}_l}{\pi_l}, \tag{4.40}$$

avec

$$D_b = \sum_{k \in U} b_k \frac{\mathbf{x}_k \mathbf{x'}_k}{\pi_k},$$

de façon à annuler la formule de variance de Horvitz-Thompson. Cette même méthode avait été utilisée au paragraphe 3.2 dans le cas d'une seule variable d'équilibrage. Dans le cas de p variables d'équilibrage, un raisonnement analogue conduit à la condition :

$$\forall k \quad b_k = \pi_k(1 - \pi_k) + b_k \frac{\mathbf{x'}_k}{\pi_k} D_b^{-1} b_k \frac{\mathbf{x}_k}{\pi_k}.$$

On retrouve ainsi l'approximation 4.14. Si cette équation n'admet pas de solution, on peut là encore utiliser 4.15.

Bien que l'approximation des probabilités d'ordre deux n'ait pas été développée rigoureusement dans le cas multivarié, les formules d'approximation de variance, qui vérifient la condition de bootstrappabilité donnée au chapitre 3, section 2, sont validées empiriquement par les simulations de Deville and Tillé (2005). Nous donnons ci-dessous quelques simulations complémentaires pour vérifier le comportement de la méthode de Bootstrap.

Notons que les approximations de variance obtenues s'expriment comme des fonctions de totaux, ce qui répond à la condition de bootstrappabilité établie au chapitre 3, section 2. Une validation empirique de la méthode de Bootstrap au cas d'un échantillon équilibré a été réalisée par Dugachard et al. (2006).

4.4.3 Simulations

Nous utilisons la population POP2 présentée au chapitre 3, section 3. Le plan de sondage consiste à tirer, après un tri aléatoire préalable de la base de sondage, un échantillon équilibré sur la variable constante égale à 1 et sur la

probabilité d'inclusion. Les variables π_1, π_2 et π_3 précedemment définies sont utilisées tour à tour. Nous nous intéressons à la précision des π-estimateurs de totaux $\widehat{t_{y\pi}}$ et $\widehat{t_{z\pi}}$, de l'estimateur du ratio $\frac{\widehat{t_{y\pi}}}{\widehat{t_{z\pi}}}$, du coefficient de corrélation estimé entre les variables y et z, des estimateurs de médiane des variables y et z. La variance exacte est approchée à l'aide de 20 000 simulations indépendantes.

L'algorithme 3.1 de Bootstrap est testé à l'aide de 1 000 échantillons indépendants. Pour chaque échantillon s, $B = 100$ pseudopopulations $U^*_{sb}, b = 1 \dots 100$ sont générées, et dans chaque pseudopopulation U^*_{sb} $C = 30$ rééchantillons $S^*_{sbc}, c = 1 \dots 30$ sont prélevés. Les intervalles de confiance sont déterminés à l'aide de la méthode des percentiles et d'une méthode de type t-Bootstrap ; pour cette dernière, on produit un estimateur de variance pour chaque rééchantillon Bootstrap en utilisant la technique de linéarisation (Deville, 1999) et l'estimation de variance fournie par Deville and Tillé (2005), voir le paragraphe 1.3.

Le tableau 4.5 compare l'approximation de variance donnée par le Bootstrap avec la vraie variance, évaluée à l'aide de 20 000 simulations. Pour les statistiques linéaires et faiblement non linéaires, le biais de l'approximation Bootstrap reste limité (de l'ordre de 5%). Ce biais est plus important pour l'estimation de précision dans le cas de la médiane ; de façon générale, il tend à diminuer quand la taille d'échantillon augmente (sauf pour le total de la variable z).

Le tableau 4.6 donne les taux de couverture effectifs obtenus avec l'algorithme 3.1. Pour l'estimation des totaux et du ratio, la méthode des percentiles respecte mieux les taux de couverture attendus ; en revanche, le t-Bootstrap donne de meilleurs résultats pour le coefficient de corrélation. La méthode des percentiles a l'avantage de permettre de produire facilement un intervalle de confiance pour des statistiques fortement non linéaires de type fractiles : les taux de couverture sont respectés de façon satisfaisante dans le cas de l'estimation d'une médiane.

4.5 Une généralisation de la méthode mirror-match

Dans le cas d'un sondage aléatoire simple, la méthode mirror-match (Sitter, 1992) consistait à estimer la précision par des rééchantillonnage répétés dans l'échantillon S d'origine, avec le même taux de sondage qu'initialement. Nous proposons ici une généralisation de cette méthode au cas d'un échantillonnage équilibré sélectionné à probabilités égales. Nous nous limitons ici au cas où l'inverse du taux de sondage N/n est entier. On peut envisager dans le cas général une randomisation sur la taille et le nombre de rééchantillonnages analogue à celle de Sitter (1992), visant à se caler sur un estimateur (approximativement) sans biais de la variance dans le cas linéaire.

4.5.1 Présentation

Le principe est analogue à celui de la méthode mirror-match d'origine. Soit S un échantillon, sélectionné à probabilités égales $\pi_k = n/N$, selon un plan à entropie maximale équilibré sur les variables $\mathbf{x}$. On sélectionne dans S, de façon indépendante, $K = N/n$ sous-échantillons $S_1^*, \ldots, S_K^*$, selon le même plan à entropie maximale équilibré sur les variables $\mathbf{x}$, avec des probabilités égales à $\pi_k = n/N$. Le rééchantillon S^* est obtenu en réunissant $S_1^*, \ldots, S_K^*$.

Pour justifier de la consistance de cette méthode de Bootstrap, nous utilisons la formule approchée 4.8 de Deville and Tillé (2005) obtenue avec l'approximation $b_{1k} = \pi_k(1 - \pi_k)$. Pour un échantillonnage équilibré sur les variables $\mathbf{x}$ avec des probabilités égales, la formule 4.8 se réduit alors à

$$V_{app}(\widehat{t_{y\pi}}) = N^2 \frac{1-f}{n} \frac{1}{N} \sum_{k \in U} (y_k - y_k^*)^2 \tag{4.41}$$

avec

$$y_k^* = \mathbf{x_k'} \left(\sum_{l \in U} \mathbf{x_l} \mathbf{x_l'} \right)^{-1} \sum_{l \in U} \mathbf{x_l} y_l,$$

et conduit à l'estimateur de variance

$$\widehat{V}(\widehat{t_{y\pi}}) = N^2 \frac{1-f}{n} \frac{1}{n} \sum_{k \in S} (y_k - \hat{y}_k^*)^2 \tag{4.42}$$

avec

$$\hat{y}_k^* = \mathbf{x}_\mathbf{k}' \left(\sum_{l \in S} \mathbf{x}_\mathbf{l} \mathbf{x}_\mathbf{l}' \right)^{-1} \sum_{l \in S} \mathbf{x}_\mathbf{l} y_l.$$

En notant que

$$\begin{aligned} \widehat{t_{y\pi}}(S^*) &= \sum_{k \in S^*} \frac{y_k}{\pi_k} \\ &= \sum_{i=1}^K \widehat{t_{y\pi}}(S_i^*), \end{aligned}$$

en utilisant 4.41 sur l'échantillon S et par indépendance entre les K rééchantillonnages, on obtient :

$$\begin{aligned} V_{app}\left(\widehat{t_{y\pi}}(S^*)\right) &= K\ V_{app}\left(\widehat{t_{y\pi}}(S_i^*)\right) \\ &= K\ n^2 \tfrac{1-f}{nf} \tfrac{1}{n} \sum_{k \in S} \left(y_k - \widetilde{y_k^*}\right)^2 \\ &= N^2 \tfrac{1-f}{n} \tfrac{1}{n} \sum_{k \in S} \left(y_k - \widetilde{y_k^*}\right)^2 \end{aligned}$$

avec

$$\widetilde{y_k^*} = \mathbf{x}_\mathbf{k}' \left(\sum_{l \in S} \mathbf{x}_\mathbf{l} \mathbf{x}_\mathbf{l}' \right)^{-1} \sum_{l \in S} \mathbf{x}_\mathbf{l} y_l = \hat{y}_k^*.$$

On retrouve donc l'estimateur de variance approximativement sans biais donné par la formule 4.42. On rappelle que le sondage aléatoire simple est un cas particulier d'échantillonnage équilibré à entropie maximale, obtenu en équilibrant sur la seule probabilité d'inclusion ; la méthode mirror-match de Sitter (1992) est donc un cas particulier de notre algorithme.

4.5.2 Lien avec l'échantillonnage équilibré stratifié

Dans le cas où l'échantillonnage est effectué à probabilités égales, et que l'équilibrage est exact, l'algorithme précédent fournit donc une méthode d'estimation de variance. Mais nous avons vu qu'en pratique, il est rare qu'un échantillon puisse être exactement équilibré : la méthode du CUBE utilise deux phases de tirage, l'une où les contraintes d'équilibrage sont exactement respectées (phase de vol), l'autre où ces contraintes sont relâchées pour terminer l'échantillonnage (phase d'atterrissage). La méthode du CUBE n'est donc pas une méthode exacte, pour la simple raison que la plupart du temps,

l'équilibrage exact n'existe pas.

En assimilant la variance d'un échantillon équilibré à la variance correspondant à la phase de vol, cet algorithme et de façon plus générale les formules d'approximation de variance proposées par Deville et Tillé peuvent être utilisées. Cette approximation est licite si la phase d'atterrissage n'occasionne qu'une variance limitée (ce qui est par exemple le cas si le nombre de contraintes d'équilibrage reste faible). Cependant, la méthode de Bootstrap présentée ci-dessus met théoriquement en jeu, pour un rééchantillon S^*, K échantillonnages et donc K phases d'atterrissage. La variance totale associée à l'ensemble de ces phases d'atterrissage peut ne plus être négligeable, et occasionner un biais dans l'estimation de variance.

Pour résoudre ce problème, nous préconisons d'utiliser l'échantillonnage équilibré stratifié. On réalise K phases de vol indépendantes dans l'échantillon S. On réunit ensuite les unités restantes (au plus K p), sur lesquelles on refait tourner une phase de vol, suivie de la phase d'atterrissage. Le rééchantillon S^* est donné par le résultat des $K+1$ phases de vol et de l'atterrissage. Cette technique a été mise en oeuvre dans le cas des échantillons du Nouveau Recensement, voir le chapitre 6, section 2.

Conclusion

Dans ce chapitre, nous présentons la méthode du Cube, permettant de sélectionner des échantillons équilibrés assurant une inférence exacte pour les totaux des variables de contrôle. Nous proposons un algorithme rapide d'échantillonnage équilibré, obtenu en limitant les unités mises en balance à chaque étape du tirage. La notion d'échantillonnage équilibré stratifié est également introduite.

La production d'un algorithme rapide d'échantillonnage équilibré ouvre la voie à un calcul de précision effectif par des méthodes de rééchantillonnage. Nous donnons une justification aux formules d'approximation de précision de Deville and Tillé (2005), et les utilisons pour justifier de la consistance de notre méthode de Bootstrap dans le cas d'un tirage équilibré à entropie maximale. Une généralisation de la méthode mirror-match de Sitter (1992) est également proposée dans le cas d'un échantillonnage équilibré à entropie

maximale à probabilités égales.

FIG. 4.2 – Algorithme rapide pour la phase de vol

1 : Initialisation

1. Les unités dont les probabilités d'inclusion sont égales à 0 ou 1 sont retirées de la population avant d'appliquer l'algorithme, de sorte que toutes les unités restantes vérifient $0 < \pi_k < 1$.
2. Les probabilités d'inclusion sont chargées dans le vecteur $\boldsymbol{\pi}$.
3. Le vecteur $\boldsymbol{\psi}$ est constitué des $p+1$ premiers éléments de $\boldsymbol{\pi}$.
4. On crée le vecteur de rangs $\mathbf{r} = (1, 2, \ldots, p, p+1)'$.
5. La matrice $\mathbf{B}$ est constituée des $p+1$ premières colonnes de $\mathbf{A}$.
6. On initialise $k = p+2$.

2 : Boucle de base

1. On choisit un vecteur $\mathbf{u}$ dans le noyau de $\mathbf{B}$,
2. Seul $\boldsymbol{\psi}$ est modifié (et pas le vecteur $\boldsymbol{\pi}$) à l'aide de la technique de base. Calculer λ_1^* et λ_2^*, les plus grandes valeurs telles que $0 \leq \boldsymbol{\psi} + \lambda_1 \mathbf{u} \leq 1$, et $0 \leq \boldsymbol{\psi} - \lambda_2 \mathbf{u} \leq 1$. Noter que $\lambda_1^* > 0$ et $\lambda_2^* > 0$.
3. Sélectionner
$$\boldsymbol{\psi} = \begin{cases} \boldsymbol{\psi} + \lambda_1^* \mathbf{u} \text{ avec probabilité } q \\ \boldsymbol{\psi} - \lambda_2^* \mathbf{u} \text{ avec probabilité } 1-q, \end{cases}$$
où $q = \lambda_2^*/(\lambda_1^* + \lambda_2^*)$.
4. (*Les unités correspondant à des $\psi(i)$ entiers sont retirées de* $\mathbf{B}$, *et sont remplacées par les probabilités d'inclusion de nouvelles unités. L'algorithme s'arrête à la fin du fichier.*)
Pour $i = 1, \ldots, p+1$,
Si $\boldsymbol{\psi}(i) = 0$ ou $\boldsymbol{\psi}(i) = 1$ alors

Si $k \leq N$ alors	$\boldsymbol{\pi}(\mathbf{r}(i)) = \boldsymbol{\psi}(i)$ $\mathbf{r}(i) = k$; $\boldsymbol{\psi}(i) = \boldsymbol{\pi}(k)$ Pour $j = 1, \ldots, p, \mathbf{B}(i,j) = \mathbf{A}(k,j)$ $k = k+1$
sinon aller à l'étape 3	

5. Aller à l'étape 2.1

3 : Fin de la première partie de la phase de vol
Pour $i = 1, \ldots, p+1, \boldsymbol{\pi}(\mathbf{r}(i)) = \boldsymbol{\psi}(i)$.

FIG. 4.3 – Méthode du pivot pour des probabilités d'inclusion inégales

1. Trier (éventuellement) la table de façon aléatoire ;
2. Définition a, b, u real ; i, j, k integer ;
3. $a = \pi_1; b = \pi_2; i = 1; j = 2;$
4. Pour $k = 1, \ldots, N : s_k = 0;$
5. $k = 3;$
6. Tant que $k \leq N$:
   u = variable aléatoire uniforme dans [0,1] ;
   si $a + b > 1$ alors | si $u < \frac{1-b}{2-a-b} : b = a + b - 1; a = 1;$ | sinon : $a = a + b - 1; b = 1;$
   sinon | si $u < \frac{b}{a+b} : b = a + b; a = 0;$ | alors : $a = a + b; b = 0;$
   si a est un entier et $k \leq N$ alors $s_i = a; a = \pi_k; i = k; k = k + 1$;
   si b est un entier et $k \leq N$ alors $s_j = b; b = \pi_k; j = k; k = k + 1$;
7. $s_i = a; s_j = b.$

TAB. 4.4 – Indicateurs de la qualité de l'équilibrage stratifié

Strate	1	2	3	4	5	6	7	8	9	10	11	12
Ecart relatif maximum (valeur absolue)	3.2%	2.4%	3.8%	5.7%	1.7%	3.1%	4.0%	2.2%	3.9%	3.2%	3.5%	6.3%
Ecart relatif moyen (valeur absolue)	0.7%	0.8%	1.1%	1.0%	0.7%	1.0%	0.9%	0.7%	1.2%	0.9%	0.9%	0.8%
Strate	13	14	15	16	17	18	19	20	21	22	23	24
Ecart relatif maximum (valeur absolue)	3.1%	4.9%	2.6%	3.2%	2.2%	5.1%	2.6%	2.1%	3.9%	3.2%	2.1%	4.3%
Ecart relatif moyen (valeur absolue)	1.3%	1.4%	0.8%	0.7%	0.5%	1.2%	1.3%	0.5%	1.1%	1.0%	0.7%	1.2%
Strate	25	26	27	28	29	30	31	32	33	34	35	36
Ecart relatif maximum (valeur absolue)	6.8%	1.5%	1.7%	1.8%	4.5%	3.9%	4.0%	1.5%	1.5%	1.8%	2.8%	2.7%
Ecart relatif moyen (valeur absolue)	1.0%	0.5%	0.5%	0.6%	1.4%	1.4%	0.9%	0.6%	0.4%	0.6%	0.6%	0.9%

TAB. 4.5 – Ecart relatif à la vraie variance pour l'algorithmes 1 de Bootstrap dans le cas d'un tirage équilibré

	Ecart relatif (%)		
Paramètre	Echantillon de taille 20	Echantillon de taille 50	Echantillon de taille 100
Total de y	-1.03	-1.94	+0.03
Total de z	+5.74	+5.79	+8.98
Ratio	-3.29	-4.29	-1.86
Corrélation	-5.01	-4.97	-4.10
Médiane de y	+16.73	+18.82	+11.86
Médiane de z	+15.68	+13.25	+11.44

TAB. 4.6 – Taux de couverture obtenus avec l'algorithme 3.1 de Bootstrap pour un tirage équilibré

	Percentiles						t-Bootstrap					
	2.5 %			5 %			2.5 %			5 %		
Statistique	L	U	L+U	L	U	L+U	L	U	L+U	L	U	L+U
						Echantillon de taille 20						
Total de y	1.9	2.1	4.0	4.3	5.1	9.4	1.7	1.8	3.5	3.4	5.0	8.4
Total de z	2.2	2.3	4.5	4.6	4.8	9.4	1.5	1.9	3.4	3.5	3.5	7.0
Ratio	2.8	2.3	5.1	5.3	5.0	10.3	1.1	3.3	4.6	3.9	5.7	9.6
Corrélation	3.2	3.3	6.5	6.8	6.0	12.8	3.6	1.9	5.5	6.3	3.8	10.1
Médiane de y	2.8	3.0	5.8	4.8	5.7	10.5						
Médiane de z	1.9	2.7	4.6	4.0	4.6	8.6						
						Echantillon de taille 50						
Total de y	2.0	2.8	4.8	5.5	5.0	10.5	1.4	2.6	4.0	4.3	4.7	9.0
Total de z	2.4	2.3	4.7	4.5	4.4	8.9	1.4	1.8	3.2	3.3	3.5	6.8
Ratio	2.9	2.3	5.2	5.5	5.3	10.8	1.2	3.7	4.9	3.3	7.4	10.7
Corrélation	4.5	3.9	8.4	8.3	6.3	14.6	4.8	2.3	7.1	7.8	4.4	12.2
Médiane de y	1.5	2.6	4.1	4.1	4.9	9.0						
Médiane de z	3.2	2.8	6.0	5.8	5.5	11.3						
						Echantillon de taille 100						
Total de y	2.4	3.3	5.7	4.8	5.2	10.0	2.4	3.1	5.5	4.6	4.8	9.4
Total de z	2.0	2.7	4.7	5.4	4.9	10.3	1.0	2.2	3.2	4.1	4.5	8.6
Ratio	3.8	3.4	7.2	6.9	5.7	12.6	2.8	4.1	6.9	5.5	6.7	12.2
Corrélation	3.5	2.8	6.3	6.4	10.9	11.3	3.7	2.0	5.7	7.1	3.9	11.0
Médiane de y	3.2	3.9	7.1	6.0	6.9	12.9						
Médiane de z	2.5	2.5	5.0	4.8	5.4	10.2						

Note de lecture : cf. tableau 2.2

Bibliographie

Ardilly, P. (1991). Echantillonnage représentatif optimum à probabilités inégales. *Annales d'Economie et de Statistique 23*, 91–113.

Ardilly, P. and Y. Tillé (2003). *Exercices corrigés de méthodes de sondage*. Paris : Ellipses.

Barat, C., C. Berthet, and D. Cou/Íeff/'e (2005). Calcul de précision d'un échantillon équilibré. Rapport de projet statistique, ENSAI.

Bertrand, P., B. Christian, G. Chauvet, and J.-M. Grosbras (2004). Plans de sondage pour le recensement rénové de la population. In *Séries INSEE Méthodes : Actes des Journées de Méthodologie Statistique*, Paris. Paris : INSEE, to appear.

Chauvet, G. and Y. Tillé (2006). A fast algorithm of balanced sampling. *Computational Statistics 21*, 53–61.

Chauvet, G. and Y. Tillé (2007). Application of the fast sas macros for balancing samples to the selection of addresses. *à paraître dans Case Studies in Business, Industry and Government Statistics*.

Deville, J.-C. (1999). Variance estimation for complex statistics ans estimators : linearization and residual techniques. *Techniques d'enquête, Survey methodology 25*, 193–204.

Deville, J.-C., J.-M. Grosbras, and N. Roth (1988). Efficient sampling algorithms and balanced sample. In *COMPSTAT, Proceeding in computational statistics*, pp. 255–266.

Deville, J.-C. and Y. Tillé (1998). Unequal probability sampling without replacement through a splitting method. *Biometrika 85*, 89–101.

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

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

Dugachard, M., M. Henry, L. Prunier, and T. Palugan (2006). Calcul de précision d'un échantillon équilibré. Rapport de projet statistique, ENSAI.

Dumais, J. and M. Isnard (2000). Le sondage de logements dans les grandes communes dans le cadre du recensement rénové de la population. In *Séries INSEE Méthodes : Actes des Journées de Méthodologie Statistique*, Volume 100. pp. 37-76, Paris : INSEE.

Hájek, J. (1981). *Sampling from a Finite Population*. New York : Marcel Dekker.

Hall, P. (1992). *The Bootstrap and Edgeworth Expansion*. New-York : Springer-Verlag.

Rao, J., H. Hartley, and W. Cochran (1962). On a simple procedure of unequal probability sampling without replacement. *Journal of Royal of Statistical Society Serie B 24*, 482–491.

Rousseau, S. and F. Tardieu (2004). La macro sas cube d'échantillonnage équilibré - documentation de l'utilisateur. Technical report, Insee.

Sitter, R. (1992). A resampling procedure for complex survey data. *Journal of the American Statistical Association 87*, 755–765.

Skovgaard, I. (1986). On multivariate edgeworth expansions. *International Statistical Review 54*, 169–186.

Thionet, P. (1953). *La théorie des sondages*. Paris : INSEE, Imprimerie nationale.

Tillé, Y. (2001). *Théorie des sondages : échantillonnage et estimation en populations finies*. Paris : Dunod.

Yates, F. (1949). *Sampling Methods for Censuses and Surveys*. London : Griffin.

Zhidong, B. and Z. Lincheng (1986). Edgeworth expansion of distribution function of independant random variables. *Scienta Sinica A29*, 1–22.

Chapitre 5

Bootstrap d'un plan de sondage complexe

Dans la plupart des enquêtes complexes, la stratégie d'échantillonnage et d'estimation ne repose pas sur un tirage "direct" avec utilisation des poids de sondage. Les unités sont préalablement partitionnées en sous-populations homogènes (on parle de stratification), éventuellement par paquets (ou unités primaires) en vue de regrouper les individus échantillonnés. D'autre part, le connaissance d'information auxiliaire après la phase d'échantillonnage permet de réaliser un redressement, afin de produire des estimations cohérentes avec les données auxiliaires mais aussi plus précises. Nous examinons dans ce qui suit comment le Bootstrap doit être adapté afin de tenir compte de ces étapes de l'enquête.

Le chapitre s'organise de la façon suivante. En section 1, nous effectuons quelques rappels sur le plan de sondage stratifié, et discutons de la mise en oeuvre d'une méthode de Bootstrap. En section 2, nous rappelons le principe de l'échantillonnage multi-degrés, et proposons une méthode permettant d'adapter l'algorithme de Bootstrap proposé précedemment au cas de plusieurs degrés d'échantillonnage. En section 3, nous rappelons le principe du redressement d'un estimateur, et montrons que l'algorithme de Bootstrap proposé donne une estimation de variance consistante pour un estimateur redressé.

5.1 Le tirage stratifié

Si on dispose d'une information auxiliaire bien corrélée à la variable d'intérêt, on peut s'en servir en partitionnant la population en sous-populations aussi homogènes que possible (vis à vis de la variable d'intérêt). La variance sera d'autant plus réduite que les strates sont homogènes et que l'on tire un échantillon de taille plus importante dans les strates les plus hétérogènes.

5.1.1 Principe

Soit une variable à H modalités supposée connue avant l'échantillonnage sur chaque individu de la population. On l'utilise pour partitionner U en H sous-populations appelées **strates** et notées $U_1, \ldots, U_H$, de tailles respectives $N_1, \ldots, N_H$. On a donc

$$N = N_1 + \ldots + N_H.$$

On suppose que l'on prélève **indépendamment** dans chaque strate U_h un échantillon S_h de taille n_h, selon un plan de sondage quelconque. On note $f_h = n_h/N_h$ le taux de sondage dans la strate U_h.

Soit t_y le total sur la population U d'une variable y, et t_{yh} le total de la même variable sur la strate U_h. Le π-estimateur de t_y s'écrit

$$\begin{aligned} \widehat{t_{y\pi}} &= \sum_{k\in S} \frac{y_k}{\pi_k} \\ &= \sum_{h=1}^{H} \sum_{k\in S_h} \frac{y_k}{\pi_k} \\ &= \sum_{h=1}^{H} \widehat{t_{y_h\pi}} \end{aligned}$$

où $\widehat{t_{y_h\pi}}$ désigne le π-estimateur du total de y sur U_h. La variance se calcule simplement en raison de l'indépendance entre les tirages :

$$\begin{aligned} V(\widehat{t_{y\pi}}) &= V(\sum_{h=1}^{H} \widehat{t_{y_h\pi}}) \\ &= \sum_{h=1}^{H} V(\widehat{t_{y_h\pi}}) \end{aligned}$$

En particulier, l'estimation de variance passe par une estimation individuelle au sein de chaque strate.

5.1.2 Bootstrap d'un échantillon stratifié

L'utilisation d'une méthode de Bootstrap de type BWO ne pose a priori pas de difficulté pour un échantillon stratifié : le Bootstrap doit être appliqué indépendamment sur chacune des strates. On a montré précedemment que pour des plans de sondage de grande entropie (sondage aléatoire simple, tirage poissonien, tirage réjectif) le Bootstrap de type BWO donnait une estimation de variance sans biais asymptotiquement, i.e. pour une grande taille d'échantillon. Les résultats obtenus se généralisent de façon directe au cas d'un tirage stratifié avec un nombre fini de strates.

Cependant, un autre point de vue asymptotique peut être envisagé. La stratification est avant tout, à taille d'échantillon fixée, un arbitrage entre le nombre de strates et la taille d'échantillon tirée dans chacune. Une pratique assez courante consiste à partitionner la population en autant de strates que possible, avec pour seule contrainte de tirer au moins deux individus dans chaque strate (afin de permettre une estimation sans biais de variance). On suppose dans ce cas que le nombre de strates $H \to \infty$, quand les quantités N_h et n_h restent bornées. Dans ce cas, le biais obtenu avec la méthode de Bootstrap pour l'estimation de variance dans chaque strate U_h est de l'ordre de $1/n_h$ et n'est donc pas négligeable, ce qui peut conduire à une estimation de variance inconsistante. Ce problème a notamment été étudié par Sitter (1992) et Chen and Sitter (1993) pour la méthode mirror-match.

Des résultats rassurants sont obtenus par Presnell and Booth (1994) et Davison and Hinkley (1997). Les simulations réalisées sur une population artificielle montrent un bon comportement du Bootstrap de type BWO (Davison and Hinkley (1997), page 99) :

> *Overall the population Bootstrap and modified sample size methods do best in this limited comparison, and coverage is not improved by using the more complicated mirror-match method.*

5.2 Le tirage multi-degrés

Contrairement à la plupart des techniques d'échantillonnage, le tirage à plusieurs degrés est utilisé non pour améliorer la précision de l'enquête, mais pour réduire les coûts. La concentration des unités échantillonnées permet

d'éviter la constitution d'une base de sondage exhaustive, et limite les déplacements éventuels des enquêteurs. De nombreuses enquêtes à grande échelle utilisent un plan stratifié à plusieurs degrés. C'est le cas à l'Insee pour le tirage de l'échantillon maître qui sert de base de sondage aux enquêtes sur les ménages. Afin de réduire la variance associée à l'échantillonnage du premier degré, les unités primaires peuvent être sélectionnées à probabilités proportionnelles à leur taille, ce qui complique l'estimation de variance.

La section est organisée de la façon suivante. Nous définissons les notations dans le paragraphe 1. Le paragraphe 2 rappelle les principales méthodes de Bootstrap existantes pour un tirage à deux degrés. Le paragraphe 3 introduit une méthode générale de Bootstrap pour un échantillonnage à deux degrés, et une méthode simplifiée est proposée dans le paragraphe 4. Ces deux méthodes sont évaluées dans le paragraphe 5 à l'aide de simulations.

5.2.1 Notations

Afin d'alléger les notations, nous nous plaçons dans le cas d'un échantillonnage à deux degrés avec une seule strate de tirage. Les résultats présentés s'étendent de façon immédiate au cas d'un nombre fini de strates. Le cas d'un tirage stratifié dont le nombre de strates $H \to \infty$ ne sera pas considéré ici. On suppose que le plan à deux degrés vérifie les hypothèses classiques d'invariance (à chaque fois que l'unité primaire u_i est sélectionnée dans l'échantillon S_I du premier degré, le même plan p_i est utilisé au second degré dans S_i) et d'indépendance (l'échantillonnage du second degré dans chaque UP se fait indépendamment du tirage de second degré dans les autres UP), voir Särndal et al. (1992, pages 134-135).

La population U_{GR} est constituée de M unités primaires (UP). Chaque UP u_i contient N_i unités secondaires (US). On note π_{Ii} la probabilité d'inclusion de u_i pour le plan de sondage p_I utilisé au 1er degré, c'est-à-dire la probabilité de sélectionner u_i dans l'échantillon S_I d'UP. Soit $m = \sum_{u_i \in U_{GR}} \pi_{Ii}$ la taille (fixe) de S_I. On note également π_{Iij} la probabilité de sélectionner conjointement u_i et u_j dans S_I et $\Delta_{Iij} = \pi_{Iij} - \pi_{Ii}\pi_{Ij}$. Si l'UP u_i est sélectionnée dans S_I, on note S_i l'échantillon d'US sélectionné dans u_i selon un plan de sondage p_i, $\pi_{k|i}$ la probabilité de sélectionner l'US k appartenant à u_i dans S_i, $\pi_{kl|i}$ la probabilité de sélectionner les US k et l appartenant à u_i dans S_I, et $\Delta_{kl|i} = \pi_{kl|i} - \pi_{k|i}\pi_{l|i}$. La taille de l'échantillon S_i est notée $n_i = \sum_{k \in u_i} \pi_{k|i}$.

On notera $U = \bigcup_{i=1}^{M} u_i$ l'ensemble des US. L'échantillon final S d'US est donné par la réunion des S_i.

Le total $t_y(U) = \sum_{k \in U} y_k$ de la variable y sur U peut être estimé sans biais par son π-estimateur

$$\widehat{t_{y\pi}}(S) = \sum_{k \in S} \frac{y_k}{\pi_k} = \sum_{u_i \in S_I} \sum_{k \in S_i} \frac{y_k}{\pi_{Ii}\pi_{k|i}}.$$

Le totak $t_y(u_i)$ de la variable y sur u_i est estimé sans biais par $\widehat{t_{y\pi}}(S_i) = \sum_{k \in S_i} \frac{y_k}{\pi_{k|i}}$. On a donc encore la relation :

$$\widehat{t_{y\pi}}(S) = \sum_{u_i \in S_I} \frac{\widehat{t_{y\pi}}(S_i)}{\pi_{Ii}}.$$

La variance de $\widehat{t_{y\pi}}(S)$ (voir par exemple Tillé (2001)) est égale à

$$\begin{aligned} V(\widehat{t_{y\pi}}(S)) &= V\left(\sum_{u_i \in S_I} \frac{t_y(u_i)}{\pi_{Ii}}\right) + \sum_{u_i \in U_{GR}} \frac{V(\widehat{t_{y\pi}}(S_i))}{\pi_{Ii}} \\ &= \underbrace{\sum_{u_i \in U_{GR}} \sum_{u_j \in U_{GR}} \frac{t_y(u_i)}{\pi_{Ii}} \frac{t_y(u_j)}{\pi_{Ij}} \Delta_{Iij}}_{V_{UP}} + \underbrace{\sum_{u_i \in U_{GR}} \frac{V(\widehat{t_{y\pi}}(S_i))}{\pi_{Ii}}}_{V_{US}} \end{aligned} \tag{5.1}$$

avec $V(\widehat{t_{y\pi}}(S_i)) = \sum_{k \in u_i} \sum_{l \in u_i} \frac{y_k}{\pi_{k|i}} \frac{y_l}{\pi_{l|i}} \Delta_{kl|i}$. La variance se décompose en deux termes V_{UP} et V_{US}, respectivement associés aux premier et second degré de tirage. Si les probabilités d'inclusion π_{Iij} et $\pi_{kl|i}$ sont toutes strictement positives, un estimateur sans biais de variance est donné par

$$\begin{aligned} \widehat{V}(\widehat{t_{y\pi}}(S)) &= \widehat{V}\left(\sum_{u_i \in S_I} \frac{t_y(u_i)}{\pi_{Ii}}\right) + \sum_{u_i \in U_{GR}} \frac{\widehat{V}(\widehat{t_{y\pi}}(S_i))}{\pi_{Ii}} \\ &= \underbrace{\sum_{u_i \in S_I} \sum_{u_j \in S_I} \frac{\widehat{t_{y\pi}}(S_i)}{\pi_{Ii}} \frac{\widehat{t_{y\pi}}(S_j)}{\pi_{Ij}} \Delta_{Iij}}_{\widehat{V_A}} + \underbrace{\sum_{u_i \in S_I} \frac{\widehat{V}(\widehat{t_{y\pi}}(S_i))}{\pi_{Ii}}}_{\widehat{V_B}} \end{aligned}$$

avec $\widehat{V}(\widehat{t_{y\pi}}(S_i)) = \sum_{k \in S_i} \sum_{l \in S_i} \frac{y_k}{\pi_{k|i}} \frac{y_l}{\pi_{l|i}} \frac{\Delta_{kl|i}}{\pi_{kl|i}}$. Il est important de noter que $\widehat{V}(\widehat{t_{y\pi}}(S_i))$ n'est pas un estimateur sans biais terme à terme de $V(\widehat{t_{y\pi}}(S_i))$.

En particulier, V_{US} et $\widehat{V}_B$ ne sont pas du même ordre de grandeur.

Pour l'estimateur $\widehat{V}(\widehat{t_{y\pi}}(S))$, la difficulté réside dans le calcul des probabilités d'inclusion doubles, aussi bien au premier degré qu'au second. Hájek (1964) offre une solution dans le cas d'un tirage de taille fixe à entropie maximale, appelé tirage réjectif. Il donne une approximation de variance, qui conduit à un estimateur asymptotiquement sans biais. Ce résultat est étendu par Berger (1998) aux plans de sondage proches de l'entropie maximale tels que l'échantillonnage de Rao-Sampford (Rao, 1965; Sampford, 1967) et le tirage successif (Hájek, 1964). Dans le cas d'un échantillonnage à plusieurs degrés, on peut utiliser (Tillé, 2001, page 177) :

$$\begin{aligned}\widehat{V}_{HAJ}(\widehat{t_{y\pi}}(S)) &= \widehat{V}_{HAJ}\left(\textstyle\sum_{u_i\in S_I}\frac{t_y(u_i)}{\pi_{Ii}}\right) + \textstyle\sum_{u_i\in S_I}\frac{\widehat{V}_{HAJ}(\widehat{t_{y\pi}}(S_i))}{\pi_{Ii}} \\ &= \underbrace{\sum_{u_i\in S_I}\frac{c_{Ii}}{\pi_{Ii}^2}\left(\widehat{t_{y\pi}}(S_i)-\widetilde{t_{yi}}\right)^2}_{\widehat{V_A^{HAJ}}} + \underbrace{\sum_{u_i\in S_I}\frac{\widehat{V}_{HAJ}(\widehat{t_{y\pi}}(S_i))}{\pi_{Ii}}}_{\widehat{V_B^{HAJ}}}\end{aligned} \tag{5.2}$$

où $\widetilde{t_{yi}} = \pi_{Ii}\frac{\sum_{u_j\in S_I} c_{Ij}\frac{\widehat{t_{y\pi}}(S_j)}{\pi_{Ij}}}{\sum_{u_j\in S_I} c_{Ij}}$, $c_{Ij}=1-\pi_{Ij}$ et

$$\widehat{V}_{HAJ}(\widehat{t_{y\pi}}(S_i)) = \sum_{k\in S_i}\frac{c_{k|i}}{\pi_{k|i}^2}\left(y_k-\tilde{y_k}\right)^2$$

avec $\tilde{y_k} = \pi_{k|i}\frac{\sum_{l\in S_i} c_{l|i}\frac{y_l}{\pi_{l|i}}}{\sum_{l\in S_i} c_{l|i}}$ et $c_{l|i}=1-\pi_{l|i}$. $\widehat{V_A}$ et $\widehat{V_B}$ peuvent être remplacés respectivement par $\widehat{V_A^{HAJ}}$ et $\widehat{V_B^{HAJ}}$ asymptotiquement sans biais, et l'estimateur de Hajek présenté en 5.2 donne une estimation de variance asymptotiquement sans biais de $V(\widehat{t_{y\pi}}(S))$ si l'échantillonnage implique un plan à forte entropie à chaque degré. L'asymptotique est ici celle proposée par Hajek (1964), en supposant que $d_I = \sum_{u_i\in U_{GR}}\pi_{Ii}(1-\pi_{Ii})\to\infty$ et $d_{k|i} = \sum_{k\in u_I}\pi_{k|i}(1-\pi_{k|i})\to\infty$ pour chaque UP u_i.

L'estimateur proposé en 5.2 peut être modifié en prenant $c_{Ii} = \frac{m}{m-1}(1-\pi_{Ii})$ et $c_{k|i} = \frac{n_i}{n_i-1}(1-\pi_{k|i})$, ce qui permet de restituer l'estimateur sans biais habituel dans le cas d'un sondage aléatoire simple à chaque degré.

5.2.2 Méthodes de Bootstrap existantes

La littérature sur le Bootstrap pour un plan à plusieurs degrés concerne presque exclusivement, à notre connaissance, le cas d'un sondage aléatoire simple à chaque degré. Nous nous restreignons donc à ce type de plan de sondage dans la suite de ce paragraphe. Une présentation plus détaillée est donnée dans Shao and Tu (1995).

Le Bootstrap naïf consiste à sélectionner dans l'échantillon d'UP S_I d'origine, un rééchantillon de taille $m^* = m$ d'UP, par sondage aléatoire simple avec remise. Comme cette approche n'est pas consistante, Mac Carthy and Snowden (1985) suggèrent une correction de ce Bootstrap naïf avec un rééchantillon de taille

$$m^* = \frac{m-1}{1-\frac{m}{M}},$$

ce qui n'est pas toujours possible en pratique. En prenant $m^* = m - 1$, on obtient une estimation sans biais de variance dans le cas linéaire si l'échantillonnage du premier degré se fait avec remise. Dans le cas sans remise, on obtient donc une bonne estimation de variance si le taux de sondage est faible. En suivant la terminologie de Mac Carthy and Snowden (1985), cette méthode sera appelée dans ce chapitre le Bootstrap avec remise (BWR).

Rao and Wu (1988) et Sitter (1992) ont également proposés une extension du Rescaled Bootstrap et du Mirror-Match Bootstrap, respectivement, dans le cas d'un tirage à deux degrés avec sondage aléatoire à chaque degré. L'idée consiste à appliquer la méthode de Bootstrap de base à chaque degré, en se calant a posteriori sur l'estimateur de variance sans biais dans le cas linéaire pour le Rescaled Bootstrap, ou en randomisant sur le nombre de sous-échantillons k' et leur taille n' pour le Mirror-Match Bootstrap (voir chapitre 2) afin de restituer cet estimateur de variance sans biais.

5.2.3 Une méthode générale de Bootstrap

Nous avons défini au chapitre 3 l'algorithme 3.1 de Bootstrap, consistant pour le calcul de précision d'un estimateur par substitution pour un tirage à probabilités inégales à un seul degré. L'algorithme 5.1, présenté ci-dessous, adapte ce Bootstrap au cas de l'échantillonnage à deux degrés. L'idée consiste à construire à partir de l'échantillon une population constituée de pseudo

unités primaires dans laquelle on reproduit le tirage du 1er degré, mais pour laquelle le tirage du 2nd degré est modulé afin de reproduire l'estimateur habituel sans biais de variance dans le cas linéaire. Le Bootstrap proposé couvre les cas particuliers importants du tirage avec sondage aléatoire simple à chaque degré et celui du tirage autopondéré.

FIG. 5.1 – Bootstrap général pour le tirage à deux degrés

Etape 1. Soit $u_i \in S_I$. Chaque unité k de S_i est dupliquée $\left[1/\pi_{k|i}\right]$ fois, où $[.]$ désigne l'entier le plus proche, pour créer une pseudo UP que l'on note u_i^*.

Etape 2. Chaque couple (S_i, u_i^*) est dupliqué $[1/\pi_{Ii} - 1/2]$ fois, où $[. - 1/2]$ désigne la partie entière. Soit $\alpha_{Ii} = 1/\pi_{Ii} - [1/\pi_{Ii} - 1/2]$. On complète les couples déjà dupliqués par un échantillon sélectionné dans $\{(S_i, u_i^*)\ ;\ u_i \in S_I\}$ selon le plan p_I, avec les probabilités d'inclusion α_{Ii}, pour $u_i \in S_I$. On obtient ainsi une population U_{GR}^* de pseudo UP.

Etape 3. On tire l'échantillon S_I^* dans U_{GR}^* selon le plan p_I, avec les probabilités d'inclusion π_{Ii}.

Etape 4. Soit $(S_i, u_i^*) \in S_I^*$. On tire un échantillon S_I^{**} de pseudo US dans u_i^* selon le plan de sondage du 2nd degré d'origine. On prend

$\rightarrow\ S_i^* = S_i^{**}$ avec une probabilité π_{Ii},
$\rightarrow\ S_i^* = S_i$ avec une probabilité $1 - \pi_{Ii}$.

La même procédure est appliquée pour chaque couple $(S_i, u_i^*) \in S_I^*$. Le rééchantillon S_* est donné par la réunion des S_i^*.

Etape 5. Les étapes 3 et 4 sont répétées C fois, pour obtenir les rééchantillons $S_1^*, \ldots, S_C^*$. Soit $v = \frac{1}{C-1} \sum_{c=1}^{C} \left(\widehat{\theta}(S_*^c) - \widehat{\theta}_*^m\right)^2$, où $\widehat{\theta}(S_*^c)$ donne la valeur de l'estimateur sur le rééchantillon S_*^c et $\widehat{\theta}_*^m = \frac{1}{C} \sum_{c=1}^{C} \widehat{\theta}(S_*^c)$.

Etape 6. Les étapes 2 à 5 sont répétées B fois, pour obtenir $v_1, \ldots, v_B$. $V\left(\widehat{\theta}\right)$ est estimée par $\widehat{V}_{boot}\left(\widehat{\theta}\right) = \frac{1}{B} \sum_{b=1}^{B} v_b$.

L'algorithme 5.1 est valide pour des probabilités d'inclusion quelconques au

premier degré. Cette méthode présente des similarités avec le Bernoulli Bootstrap de Funaoka et al. (2006). La propriété suivante établit la consistance de la méthode dans le cas linéaire. On suppose ici que les inverses de probabilités d'inclusion du 2nd degré sont approximativement entières, ce qui induit dans le cas de l'estimation de variance pour l'estimateur du total de la variable y un biais conditionnel de l'ordre de

$$\sum_{u_i \in S_I} \frac{1}{\pi_{Ii}} \sum_{k \in S_i} \left(\left[\frac{1}{\pi_{k|i}} \right] - \frac{1}{\pi_{k|i}} \right) \frac{1 - \pi_{k|i}}{\pi_{k|i}^2} \left(y_k - \tilde{y_k} \right)^2$$

(où [.] désigne l'entier le plus proche) qui peut être négligé si les probabilités d'inclusion $\pi_{k|i}$ sont faibles.

Propriété 5.1. *Soit $\theta = t_y$ et $\widehat{\theta} = \widehat{t_{y\pi}}$. Alors l'algorithme 3 donne une estimation de variance consistante pour $\widehat{t_{y\pi}}$.*

Démonstration. Nous démontrons la consistance de l'algorithme 5.1 dans le cas linéaire. Avec une démonstration analogue à celle de la propriété 3.3 pour le tirage à un degré, la consistance peut ensuite être établie sous de faibles hypothèses pour l'estimateur par substitution $\widehat{\theta}$ d'une statistique θ à l'aide de la technique de linéarisation selon la fonction d'influence.

Nous raisonnons conditionnellement à S et à la pseudopopulation U_{GR}^* obtenue à l'étape 2. Soit a_i le nombre de fois où le couple (S_i, u_i^*) apparaît dans U_{GR}^*. Notons que $E(a_i|S) = 1/\pi_{Ii}$. L'estimateur bootstrappé du total t_y est égal à

$$\widehat{t_y^*} = \sum_{(S_i, u_i^*) \in S_I^*} \frac{\widehat{t_i^*}}{\pi_{Ii}}$$

avec

$$\widehat{t_i^*} = \epsilon_i \widehat{t_{y\pi}}(S_i^{**}) + (1 - \epsilon_i) \widehat{t_{y\pi}}(S_i)$$

en utilisant les notations de l'algorithme 5.1, où ϵ_i suit une loi de Bernoulli de paramètre π_{Ii}. Pour alléger les notations, nous écrirons simplement

$$\widehat{t_y^*} = \sum_{u_i^* \in S_I^*} \frac{\widehat{t_i^*}}{\pi_{Ii}}.$$

A l'aide de la formule de décomposition de la variance, on a :

$$V(\widehat{t_i^*}|S, U_{GR}^*, S_I^*) = \underbrace{V(E(t_i^*|S, U_{GR}^*, S_I^*, \epsilon_i))}_{V_1} + \underbrace{E(V(t_i^*|S, U_{GR}^*, S_I^*, \epsilon_i))}_{V_2}$$

On a $E(t_i^*|S, U_{GR}^*, S_I^*, \epsilon_i) = \widehat{t_{y\pi}}(S_i)$, ce qui implique que $V_1 = 0$ et

$$E(t_i^*|S, U_{GR}^*, S_I^*, \epsilon_i) = \widehat{t_{y\pi}}(S_i) \tag{5.3}$$

On a également

$$\begin{aligned} V(t_i^*|S, U_{GR}^*, S_I^*, \epsilon_i) &= \epsilon_i^2 V(t_{y\pi}(S_i^{**})|S, U_{GR}^*, S_I^*, \epsilon_i) \\ &= \epsilon_i^2 V(t_{y\pi}(S_i^{**})|S, U_{GR}^*, S_I^*) \end{aligned} \tag{5.4}$$

car S_i^{**} et ϵ_i sont indépendants. On en déduit que :

$$\begin{aligned} V_2 &= E\left(\epsilon_i^2 V(t_{y\pi}(S_i^{**})|S, U_{GR}^*, S_I^*, \epsilon_i)\right) \\ &= E\left(\epsilon_i^2\right) V(t_{y\pi}(S_i^{**})|S, U_{GR}^*, S_I^*, \epsilon_i) \\ &= \pi_{Ii} V(t_{y\pi}(S_i^{**})|S, U_{GR}^*, S_I^*, \epsilon_i) \end{aligned} \tag{5.5}$$

En utilisant à nouveau la formule de décomposition de la variance, on obtient :

$$\begin{aligned} & V\left(\widehat{t_y^*}|S, U_{GR}^*\right) \\ = \; & V\left(E(t_y^*|S, U_{GR}^*, S_I^*)|S, U_{GR}^*\right) + E\left(V(t_y^*|S, U_{GR}^*, S_I^*)|S, U_{GR}^*\right) \\ = \; & V\left(\sum_{u_i^* \in S_I^*} \frac{E(\widehat{t_i^*}|S, U_{GR}^*, S_I^*)}{\pi_{Ii}}|S, U_{GR}^*\right) + E\left(\sum_{u_i^* \in S_I^*} \frac{V(\widehat{t_i^*}|S, U_{GR}^*, S_I^*)}{\pi_{Ii}}|S, U_{GR}^*\right) \\ = \; & V\left(\sum_{u_i^* \in S_I^*} \frac{E(\widehat{t_i^*}|S, U_{GR}^*, S_I^*)}{\pi_{Ii}}|S, U_{GR}^*\right) + E\left(\sum_{u_i^* \in S_I^*} \frac{V(\widehat{t_i^*}|S, U_{GR}^*)}{\pi_{Ii}}|S, U_{GR}^*\right) \\ = \; & V\left(\sum_{u_i^* \in S_I^*} \frac{\widehat{t_{y\pi}(S_i)}}{\pi_{Ii}}|S, U_{GR}^*\right) + E\left(\sum_{u_i^* \in S_I^*} \frac{V(\widehat{t_{y\pi}(S_i^{**})}|S, U_{GR}^*)}{\pi_{Ii}}|S, U_{GR}^*\right) \\ = \; & V\left(\sum_{u_i^* \in S_I^*} \frac{\widehat{t_{y\pi}(S_i)}}{\pi_{Ii}}|S, U_{GR}^*\right) + \sum_{u_i^* \in U_{GR}^*} V(\widehat{t_{y\pi}}(S_i^{**})|S, U_{GR}^*) \\ = \; & \underbrace{V\left(\sum_{u_i^* \in S_I^*} \frac{\widehat{t_{y\pi}(S_i)}}{\pi_{Ii}}|S, U_{GR}^*\right)}_{V_3} + \underbrace{\sum_{u_i \in S_I} a_i V(\widehat{t_{y\pi}}(S_i^{**})|S, U_{GR}^*)}_{V_4} \end{aligned} \tag{5.6}$$

avec $V(\widehat{t_i^*}|S, U_{GR}^*, S_I^*) = V(\widehat{t_i^*}|S, U_{GR}^*)$ par invariance. En utilisant le théorême 3.7 de Hájek (1964), on obtient

$$\begin{aligned} V_3 &= (1+o_p(1)) \sum_{u_i^* \in U_{GR}^*} \frac{\pi_{Ii}(1-\pi_{Ii})}{\pi_{Ii}^2} \left(\widehat{t_{y\pi}}(S_i) - \tilde{t}_{yi}^*\right)^2 \\ &= (1+o_p(1)) \sum_{u_i \in S_I} \frac{a_i \pi_{Ii}(1-\pi_{Ii})}{\pi_{Ii}^2} \left(\widehat{t_{y\pi}}(S_i) - \tilde{t}_{yi}^*\right)^2 \end{aligned} \tag{5.7}$$

avec

$$\begin{aligned} \tilde{t}_{yi}^* &= \pi_{Ii} \frac{\sum_{u_j^* \in U_{GR}^*} \pi_{Ij}(1-\pi_{Ij}) \frac{\widehat{t_{y\pi}}(S_j)}{\pi_{Ij}}}{\sum_{u_j^* \in U_{GR}^*} \pi_{Ij}(1-\pi_{Ij})} \\ &= \pi_{Ii} \frac{\sum_{u_j^* \in S_I} a_j \pi_{Ij}(1-\pi_{Ij}) \frac{\widehat{t_{y\pi}}(S_j)}{\pi_{Ij}}}{\sum_{u_j^* \in S_I} a_j \pi_{Ij}(1-\pi_{Ij})}. \end{aligned} \tag{5.8}$$

Le théorême 1.9 dû à Deville (1999) implique que $E(V_3|S)$ et $\widehat{V}_A^{HAJ} = \sum_{u_i \in S_I} \frac{1-\pi_{Ii}}{\pi_{Ii}^2} \left(\widehat{t_{y\pi}}(S_i) - \tilde{t}_{yi}\right)^2$, convenablement normalisés, ont la même limite.
En utilisant à nouveau le théorême 3.7 de Hájek (1964), on a

$$\begin{aligned} V\left(\widehat{t_{y\pi}}(S_i^{**})|S, U_{GR}^*\right) &= (1+o_p(1)) \sum_{k \in u_i^*} \frac{\pi_{k|i}(1-\pi_{k|i})}{\pi_{k|i}^2} (y_k - \tilde{y}_k^*)^2 \\ &= (1+o_p(1)) \sum_{k \in S_i} \frac{(1-\pi_{k|i})}{\pi_{k|i}^2} (y_k - \tilde{y}_k)^2 \\ &= (1+o_p(1)) \widehat{V}_{HAJ}\left(\widehat{t}_{y\pi}(S_i)\right) \end{aligned} \tag{5.9}$$

$$\Rightarrow E\left(V\left(\widehat{t_{y\pi}}(S_i^{**})|S, U_{GR}^*\right) - \widehat{V}_{HAJ}\left(\widehat{t}_{y\pi}(S_i)\right)|S\right) \to 0$$

avec

$$\begin{aligned} \tilde{y}_k^* &= \pi_{k|i} \frac{\sum_{l \in u_i^*} \pi_{l|i}(1-\pi_{l|i}) \frac{y_l}{\pi_{l|i}}}{\sum_{l \in u_i^*} \pi_{l|i}(1-\pi_{l|i})} \\ &= \pi_{k|i} \frac{\sum_{l \in S_i} (1-\pi_{l|i}) \frac{y_l}{\pi_{l|i}}}{\sum_{l \in S_i} (1-\pi_{l|i})} \quad = \quad \tilde{y}_k. \end{aligned} \tag{5.10}$$

Comme

$$E\left(\sum_{u_i\in S_I} a_i \widehat{V}_{HAJ}\left(\widehat{t_{y\pi}}(S_i)\right)|S\right) = \sum_{u_i\in S_I} \frac{\widehat{V}_{HAJ}\left(\widehat{t_{y\pi}}(S_i)\right)}{\pi_{Ii}} = \widehat{V}_B^{HAJ},$$

V_4 est asymptotiquement équivalent à $\widehat{V}_B^{HAJ}$, d'où le résultat.

□

Cet algorithme se généralise facilement à un plan de sondage à r degrés. Avec r étapes de duplications, on obtient une pseudopopulation U^* image de la population d'origine. La variance est estimée par applications répétées du plan de sondage dans U^*, le tirage étant modulé à chaque degré $d \geq 2$ selon le même procédé que dans l'algorithme 5.1, en fonction des probabilités d'inclusion du degré $d-1$ (on suppose là encore que les probabilités d'inclusion de chaque degré $d \geq 2$ restent faibles).

5.2.4 Une méthode simplifiée de Bootstrap

L'étape 2 de l'algorithme 5.1 est nécessaire si les probabilités d'inclusion au 1er degré sont fortes, ce qui est courant en pratique, voir Funaoka et al. (2006, page155). Dans le cas contraire, on peut introduire une simplification, présentée dans l'algorithme 5.2, qui limite le volume de calcul car elle évite le double Bootstrap nécessaire dans l'algorithme 5.1. Cet algorithme simplifié permet d'intégrer aux données d'enquête les B variables de poids Bootstrap associées au rééchantillonnage, facilitant la production ultérieure d'indicateurs de précision par les utilisateurs. Notons que si les probabilités d'inclusion du 1er degré sont négligées, l'algorithme 5.2 sélectionne quasi-systématiquement à l'étape 4 l'échantillon S_i d'origine : avec un sondage aléatoire simple au premier degré, on retrouve donc dans ce cas la méthode du Bootstrap naïf. Notre algorithme apparaît donc comme une correction de cette méthode, permettant de tenir compte du tirage sans remise des UP.

5.3 Redressement d'un estimateur

On peut disposer à l'étape de l'estimation d'une information auxiliaire, non disponible et/ou non utilisée à l'étape de l'échantillonnage, que l'on utilise

FIG. 5.2 – Bootstrap simplifié pour le tirage à deux degrés

Etape 1. Soit $u_i \in S_I$. Chaque unité k de S_i est dupliquée $\left[1/\pi_{k|i}\right]$ fois, où $[.]$ désigne l'entier le plus proche, pour créer une pseudo UP que l'on note u_i^*.

Etape 2. Chaque couple (S_i, u_i^*) est dupliqué $[1/\pi_{Ii}]$ fois pour créer une pseudopopulation U_{GR}^* de pseudo UP.

Etape 3. On tire un échantillon S_I^* dans U_{GR}^* selon le plan p_I, avec les probabilités d'inclusion π_{Ii}.

Etape 4. Soit $(S_i, u_i^*) \in S_I^*$. On tire un échantillon S_I^{**} de pseudo US dans u_i^* selon le plan de sondage du 2nd degré d'origine. On prend

$\rightarrow$ $S_i^* = S_i^{**}$ avec une probabilité π_{Ii},
$\rightarrow$ $S_i^* = S_i$ avec une probabilité $1 - \pi_{Ii}$.

La même procédure est appliquée pour chaque couple $(S_i, u_i^*) \in S_I^*$. Le rééchantillon S_* est donné par la réunion des S_i^*.

Etape 5. Les étapes 3 et 4 sont répétées B fois, pour obtenir les rééchantillons $S_1^*, \ldots, S_B^*$. $V\left(\widehat{\theta}\right)$ est estimée par $\widehat{V}_{boot}\left(\widehat{\theta}\right) = \frac{1}{B-1}\sum_{b=1}^{B}\left(\widehat{\theta}(S_*^b) - \widehat{\theta}_*^m\right)^2$, où $\widehat{\theta}(S_*^b)$ donne la valeur de l'estimateur sur le rééchantillon S_*^b et $\widehat{\theta}_*^m = \frac{1}{C}\sum_{c=1}^{C}\widehat{\theta}(S_*^c)$.

pour améliorer les estimateurs en modifiant les poids d'extrapolation : on parle alors de redressement. Cette information peut prendre la forme d'effectifs sur des catégories de population, ou de totaux de variables quantitatives : c'est à ce type de redressement, largement utilisé dans les enquêtes, que nous nous intéressons dans la suite de cette section. Plus généralement, l'information auxiliaire peut être donnée par une fonction quelconque des valeurs d'une variable, voire par le détail de cette variable. Le redressement sur une fonction de répartition a été notamment étudié par Ren (2000) et Breidt and Opsomer (2000).

Il existe différentes méthodes de redressement, dépendant de la nature des variables auxiliaires. Ces méthodes peuvent être vues comme un cas particulier de la technique de calage que nous présentons ci-dessous. On trouvera une

présentation plus détaillée des méthodes de redressement dans Tillé (2001), Sautory and Le Guennec (2003) et Ardilly (2006).

5.3.1 Principe

La technique de calage a été proposée par Deville and Särndal (1992) et Deville et al. (1993). Soit y une variable d'intérêt, mesurée sur les individus d'un échantillon S. Chaque individu k est affecté d'un poids de départ $d_k = 1/\pi_k$, donné par l'inverse de sa probabilité de sélection. Ces poids peuvent être utilisés pour estimer sans biais le total t_y.

On suppose que l'on dispose de p variables auxiliaires $\mathbf{x} = x_1, \ldots, x_p$, qualitatives ou quantitatives, dont les totaux sur l'ensemble de la population sont connus. Ces totaux sont utilisés pour modifier les poids de départ, et obtenir un nouveau poids w_k pour l'individu k, permettant de s'ajuster sur les totaux auxiliaires. On cherche donc à vérifier les équations

$$\forall i = 1 \ldots p \qquad \sum_{k \in S} w_k x_{ik} = t_{x_i}$$

Cet ajustement va diminuer la variance, car celle-ci n'est plus donnée que par les résidus de la régression de la variable d'intérêt sur les variables auxiliaires. Cependant, l'ajustement peut conduire à s'éloigner fortement des poids de départ, et donc à augmenter significativement le biais : pour cette raison, on cherche à minimiser l'écart entre anciens et nouveaux poids. Soit G une fonction de distance, positive et convexe, telle que $G(1) = 0$ et $G'(1) = 0$. Alors le calage consiste à résoudre le problème d'optimisation sous contraintes suivant :

$$\min_{w_k \ ; \ k \in S} \sum_{k \in S} d_k G\left(\frac{w_k}{d_k}\right) \text{ sous la contrainte } \sum_{k \in S} w_k x_{ik} = t_{x_i} \ \forall i = 1 \ldots p \tag{5.11}$$

La résolution de ce problème conduit à

$$w_k = d_k F\left(\mathbf{x}_k' \lambda\right) \tag{5.12}$$

où λ désigne un vecteur de multiplicateurs de Lagrange que l'on détermine avec les équations de calage. $\widehat{t_{yw}} = \sum_{k \in S} w_k y_k$ est appelé estimateur calé.

Plusieurs fonctions de distance sont proposées dans Deville and Särndal (1992). L'utilisation de la distance euclidienne permet de retrouver l'estimateur par la régression généralisée (Särndal et al., 1992). L'utilisation d'une fonction de distance exponentielle permet de reproduire la méthode du raking-ratio (Deming and Stephan, 1940; Stephan, 1942). Les méthodes dites de vraisemblance empirique correspondent elles aussi à une fonction de distance particulière, et ont donné lieu à une abondante littérature, voir par exemple Chen and Qin (1993), Chen and Sitter (1993), Chen and Wu (1999) et Wu (2002). L'estimateur par le ratio et l'estimateur post-stratifié (Holt and Smith, 1979) sont également des cas particuliers de la technique de calage.

5.3.2 Prise en compte du calage dans le Bootstrap

Deville and Särndal (1992) et Deville et al. (1993) justifient que, pour une fonction de distance quelconque vérifiant les propriétés précédentes, la variance asymptotique de l'estimateur calé est égale à

$$\begin{aligned} V_{app}\left(\widehat{t_{yw}}\right) &= V\left(\widehat{t_{e\pi}}\right) \\ &= \sum_{k\in U}\sum_{l\in U}\frac{e_k}{\pi_k}\frac{e_l}{\pi_l}\Delta_{kl} \end{aligned} \tag{5.13}$$

où $e_k = y_k - \mathbf{B}'\mathbf{x_k}$, $\mathbf{B} = \left(\sum_{k\in U}\mathbf{x_k}\mathbf{x_k'}\right)^{-1}\sum_{k\in U}\mathbf{x_k}y_k$, et n'est donc donnée que par les résidus de la régression de la variable d'intérêt sur les variables auxiliaires. On peut l'estimer par

$$\widehat{V}\left(\widehat{t_{yw}}\right) = \sum_{k\in S}\sum_{l\in S}\frac{\hat{e_k}}{\pi_k}\frac{\hat{e}_l}{\pi_l}\frac{\Delta_{kl}}{\pi_{kl}} \tag{5.14}$$

où $\hat{e}_k = y_k - \hat{\mathbf{B}}'\mathbf{x_k}$, $\hat{\mathbf{B}} = \left(\sum_{k\in S}\frac{\mathbf{x_k}\mathbf{x_k'}}{\pi_k}\right)^{-1}\sum_{k\in U}\frac{\mathbf{x_k}y_k}{\pi_k}$.

Avec l'algorithme 5.1 de Bootstrap proposé précedemment, le calage est pris en compte par un redressement du rééchantillon S^* sur la pseudopopulation U^*, comme suggéré par Canty and Davison (1999). On note $\widehat{t^*_{yw}}$ l'estimateur Bootstrap ainsi obtenu ; le résultat de Deville and Särndal (1992) implique que

$$V\left(\widehat{t^*_{yw}}|S,U^*\right) \simeq V\left(\widehat{t_{e^*\pi}}|S,U^*\right) \tag{5.15}$$

où $e_k^* = y_k - \mathbf{B}^{*\prime}\mathbf{x_k}$, $\mathbf{B}^* = \left(\sum_{k\in U^*} \mathbf{x_k}\mathbf{x_k'}\right)^{-1} \sum_{k\in U^*} \mathbf{x_k} y_k$. Si le plan de sondage est à entropie forte,

$$\begin{aligned} E\left(V\left(\widehat{t_{e^*\pi}}|S,U^*\right)|S\right) &\simeq E\left(V\left(\widehat{t_{\hat{e}\pi}}|S,U^*\right)|S\right) \\ &\simeq \widehat{V}\left(\widehat{t_{yw}}\right) \end{aligned} \tag{5.16}$$

et le calage est bien pris en compte dans l'estimation Bootstrap de variance.

5.4 Compléments

Le travail réalisé dans ce mémoire s'est essentiellement concentré sur l'estimation de précision sous la randomisation associée au plan de sondage. Le paragraphe précédent montre comment le redressement des estimateurs peut également être pris en compte. Une part non négligeable de la variance provient généralement de la non-réponse, totale ou partielle, et une bonne partie des efforts récents de recherche dans le domaine de l'estimation de précision pour une population finie se sont focalisés sur la mise au point de méthodes permettant de tenir compte de l'alea lié à la non-réponse, voir notamment Rao (1996). Berger and Rao (2006) ont propose un estimateur modifié de type Jackknife permettant de tenir compte de la variance d'imputation. Cette méthode s'apparente à une estimation de variance par linéarisation, voir par exemple Särndal (1992). Saigo et al. (2001) proposent une technique de demi-échantillons équilibrés, et Shao and Sitter (1996) montrent que les méthodes proposées par Sitter (1992) et Gross (1980) peuvent être consistantes pour l'estimation de variance en présence de données imputées, si le processus d'imputation est répété au sein de chaque rééchantillon Bootstrap. L'extension de la méthode de Shao and Sitter (1996) au cas de l'échantillonnage à probabilités inégales ne semble pas poser de difficultés particulières et fera l'objet d'un travail futur.

Une autre voie très active de recherche concerne l'estimation sur petits domaines. Quand on s'intéresse à un domaine particulier de la population (un domaine est pris ici au sens large et peut représenter une zone géographique, une catégorie socioprofessionnelle, ...), l'estimation se fait à partir de l'échantillon tombant dans ce domaine. Si le domaine est de taille restreinte, la taille de l'échantillon concerné est parfois trop faible pour permettre d'inférer sous

le plan de sondage avec une précision raisonnable. On pourra notamment consulter les articles synthétiques proposés par Ghosh and Rao (1994), Rao (1999), Lahiri and Meza (2002), Pfeffermann (2002) ainsi que l'ouvrage de référence sur le sujet (Rao, 2003). Un estimateur de l'erreur quadratique moyenne est donné par Prasad and Rao (1990), et une méthode paramétrique de Bootstrap très générale est proposée par Hall and Maiti (2006b), voir également Hall and Maiti (2006a) et Chatterjee et al. (2007). La méthode de Bootstrap que nous proposons repose sur l'hypothèse d'une randomisation basée sur le plan de sondage, alors que l'estimation sur petits domaines se fonde sur une randomisation assistée par le modèle qui fait appel à des méthodes de Bootstrap plus classiques. La méthode de régression géographique pondérée présentée au chapitre suivant peut être vue comme une méthode particulière d'estimation sur petits domaines.

Conclusion

Nous étudions dans ce chapitre l'utilisation du Bootstrap dans le cadre d'une enquête complexe. Nous montrons que la réduction de variance liée à un calage de l'estimateur est prise en compte avec le principe de plug-in, en reproduisant le calage au niveau de la pseudo-population. Nous montrons également que la méthode de Bootstrap proposée peut être étendue au cas d'un échantillonnage à plusieurs degrés, en reproduisant une pseudopopulation image de la population d'origine, et en modulant le tirage du second degré.

Il manque à ce chapitre des simulations pour comparer les performances de notre algorithme de Bootstrap à un Bootstrap naïf d'une part, et à une approche analytique d'autre part. Nous pouvons cependant donner quelques lignes directrices pour le méthodologue intéressé par un calcul de précision dans le cas d'un tirage multidegrés. Si ce calcul ne concerne que des statistiques linéaires ou faiblement non linéaires, il nous semble plus simple de faire appel à une approche analytique d'estimation de variance. Si on s'intéresse à des statistiques fortement non linéaires, le Bootstrap naïf va vraisemblablement donner de bons résultats dans le cas d'un taux de sondage faible pour le premier degré de tirage. Si le taux de sondage est fort, il sera intéressant de mettre en oeuvre notre méthode.

Bibliographie

Ardilly, P. (2006). *Les Techniques de Sondage*. Technip, Paris.

Berger, Y. (1998). Rate of convergence for asymptotic variance for the Horvitz-Thompson estimator. *Journal of Statistical Planning and Inference*, 74 :149–168.

Berger, Y. and Rao, J. (2006). Adjusted jackknife for imputation under unequal probability sampling without replacement. *Journal of the Royal Statistical Society*, B68 :531–547.

Breidt, F. and Opsomer, J. (2000). Local polynomial regresssion estimators in survey sampling. *Annals of Statistics*, 28 :1026–1053.

Canty, A. and Davison, A. (1999). Resampling-based variance estimation for labour force surveys. *The Statistician*, 48 :379–391.

Chen, J. and Qin, J. (1993). Empirical likelihood estimation for finite populations and the effective usage of auxiliary information. *Biometrika*, 80 :107–116.

Chen, J. and Sitter, R. (1993). Edgeworth expansion and the bootstrap for stratified sampling without replacement from a finite population. *The Canadian Journal of Statistics*, 21 :347–357.

Chen, J. and Wu, C. (1999). Estimation of distribution function and quantiles using the model-calibrated pseudo empirical likelihood method. *Statistica Sinica*, 12 :1223–1239.

Davison, A. and Hinkley, D. (1997). *Bootstrap Methods and Their Application*. Cambridge University Press, Cambridge.

Deming, W. and Stephan, F. (1940). On a least square adjustment of sampled frequency table when the expected marginal totals are known. *Annals of Mathematical Statistics*, 11 :427–444.

Deville, J.-C. (1999). Variance estimation for complex statistics ans estimators : linearization and residual techniques. *Techniques d'enquête, Survey methodology*, 25 :193–204.

Deville, J.-C. and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87 :376–382.

Deville, J.-C., Särndal, C.-E., and Sautory, O. (1993). Generalized raking procedure in survey sampling. *Journal of the American Statistical Association*, 88 :1013–1020.

Funaoka, F., Saigo, H., Sitter, R., and Toida, T. (2006). Bernoulli bootstrap for stratified multistage sampling. *Survey Methodology*, 32 :151–156.

Ghosh, M. and Rao, J. (1994). Small area estimation : An appraisal. *Statistical Science*, 9 :55–93.

Gross, S. (1980). Median estimation in sample surveys. *ASA Proceedings of Survey Research*, pages 181–184.

Hájek, J. (1964). Asymptotic theory of rejective sampling with varying probabilities from a finite population. *Annals of Mathematical Statistics*, 35 :1491–1523.

Hall, P. and Maiti, T. (2006a). Nonparametric estimation of mean-squared prediction errors in nested-error regression models. *Annals of Statistics*, 34 :1733–1750.

Hall, P. and Maiti, T. (2006b). On parametric bootstrap methods for small area prediction. *Journal of the Royal Statistical Society*, B68 :221–238.

Holt, D. and Smith, T. (1979). Post-stratification. *Journal of the Royal Statistical Society*, A142 :Part 1, 33–46.

Lahiri, P. and Meza, J. (2002). Small-area estimation. *Encyclopedia of Environmetrics*, 4 :2010–2014.

Mac Carthy, P. and Snowden, C. (1985). The bootstrap and finite population sampling. Technical report.

Pfeffermann, D. (2002). Small area estimation - new developments and directions. *International Statistical Review*, 70 :125–143.

Prasad, N. and Rao, J. (1990). The estimation of the mean squared error of small-area estimators. *Journal of the American Statistical Association*, 85 :163–171.

Presnell, B. and Booth, J. (1994). Resampling methods for sample surveys. Technical report.

Rao, J. (1965). On two simple schemes of unequal probability sampling without replacement. *Journal of the Indian Statistical Association*, 3 :173–180.

Rao, J. (1996). On variance estimation with imputed survey data. *Journal of the American Statistical Association*, 91 :499–506.

Rao, J. (1999). Some recent advances in model-based small area estimation. *Survey Methodology*, 25 :175–186.

Rao, J. (2003). *Small Area Estimation*. Wiley, New-York.

Rao, J. and Wu, C. (1988). Resampling inference with complex survey data. *Journal of the American Statistical Association*, 83 :231–241.

Ren, R. (2000). *Utilisation d'information auxiliaire par calage sur fonction de répartition*. PhD thesis, Université Paris Dauphine.

Saigo, H., Shao, J., and Sitter, R. (2001). A repeated half-sample bootstrap and balanced repeated replications for randomly imputed data. *Survey Methodology*, 27 :189–196.

Sampford, M. (1967). On sampling without replacement with unequal probabilities of selection. *Biometrika*, 54 :494–513.

Sautory, O. and Le Guennec, J. (2003). La macro calmar2 : redressement d'un échantillon par calage sur marges - documentation de l'utilisateur. Technical report, Insee.

Shao, J. and Sitter, R. (1996). Bootstrap for imputed survey data. *Journal of the American Statistical Association*, 91 :1278–1288.

Shao, J. and Tu, D. (1995). *The Jackknife and The Bootstrap*. Springer, New-York.

Sitter, R. (1992). A resampling procedure for complex survey data. *Journal of the American Statistical Association*, 87 :755–765.

Särndal, C.-E. (1992). Methods for estimating the precision of survey estimates when imputation has been used. *Survey Methodology*, 18 :241–252.

Särndal, C.-E., Swensson, B., and Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer Verlag, New York.

Stephan, F. (1942). An iterative method of adjusting sample frequency data tables when expected marginal totals are known. *Annals of Mathematical Statistics*, 13 :166–178.

Tillé, Y. (2001). *Théorie des sondages : échantillonnage et estimation en populations finies*. Dunod, Paris.

Wu, C. (2002). Empirical likelihood method for finite populations. *Recent Advances in Statistical Methods*, pages 339–351.

Chapitre 6

Application au Nouveau Recensement de la population

Nous proposons dans ce chapitre une application des méthodes de Bootstrap au calcul de précision des estimations du Nouveau Recensement. Jusqu'en 1999, le Recensement de la population française, réalisé en gros tous les 10 ans, consistait à enquêter exhaustivement l'ensemble des ménages du territoire. Depuis 2004, cette opération exhaustive a laissé la place à des enquêtes de Recensement, qui touchent annuellement un échantillon seulement de la population française et permettent de disposer chaque année d'une information récente. La collecte d'une année permet de réaliser des estimations dites globales sur des zones de taille importante (région et France entière) ; l'utilisation de plusieurs années de collecte permet de produire des estimations dites détaillées à un niveau plus fin. Un des enjeux du Nouveau Recensement est de pouvoir disposer d'outils de calcul de précision pour les estimations livrées.

Le chapitre est organisé de la façon suivante. En section 1, nous rappelons le plan de sondage du Nouveau Recensement. En section 2, nous testons les méthodes de Bootstrap proposées dans le chapitre 4 pour le calcul de précision des estimations dans une grande commune. En section 3, nous présentons la méthodologie dite d'estimation sur zones mixtes, et nous proposons une méthode de type Bootstrap d'estimation de précision. Cette méthode est évaluée pour l'estimation sur zones mixtes à l'aide de simulations.

6.1 Le plan de sondage du Nouveau Recensement

Le Nouveau Recensement distingue deux types de communes : les communes de moins de 10 000 habitants au sens du Recensement de 1999, appelées les **petites communes**, et les communes de plus de 10 000 habitants, appelées les **grandes communes**. Nous donnons ci-dessous un aperçu des plans de sondage selon le type de commune ; pour une présentation plus détaillée, on pourra se référer à Dumais and Isnard (2000), Durr and Dumais (2002) et Bertrand et al. (2004).

6.1.1 Les petites communes

Chaque petite commune constitue une unité statistique. Elles sont stratifiées selon la région, et au sein de chaque région ces petites communes sont partitionnées aléatoirement en 5 échantillons, par tirage équilibré à probabilités égales (1/5) sur des variables de type logement et des variables socio-démographiques données par le recensement de 1999 :

→ le nombre de logements,
→ le nombre de logements en immeuble collectif,
→ la population des personnes de moins de 20 ans,
→ la population des personnes de 20 à 39 ans,
→ la population des personnes de 40 à 59 ans,
→ la population des personnes de 60 à 74 ans,
→ la population des personnes de 75 ans et plus,
→ la population des femmes,
→ la population des hommes,
→ la population par département.

Un échantillon est appelé groupe de rotation. Chaque année, l'ensemble des ménages des petites communes d'un groupe de rotation est enquêté : au bout d'un cycle de 5 ans, toutes les petites commune ont donc été recensées.

Le principe est le même pour les départements d'outre-mer, mais le faible nombre de petites communes ne permet pas de les répartir en 5 groupes de rotation par échantillonnage équilibré. Cette partition est donc réalisée de façon déterministe, en cherchant à uniformiser la charge de collecte.

6.1.2 Les grandes communes

Chaque grande commune fait l'objet d'un plan de sondage indépendant. L'unité statistique est l'adresse, et le Répertoire d'Immeubles Localisés (RIL) fournit une cartographie numérisée des adresses de chaque grande commune.

On distingue trois strates d'adresses :

→ Les adresses de grande taille sont celles qui comptent au moins 60 logements et cumulent au plus 10% des logements de la commune. Ces adresses sont réparties en 5 groupes de rotation (par échantillonnage équilibré ou de façon déterministe) ; chaque année, un groupe de rotation est exhaustivement enquêté. Ces adresses sont donc entièrement recensées en cinq ans.

→ Les adresses neuves sont celles qui apparaissent chaque année. Elles sont également réparties en 5 groupes de rotation (par échantillonnage équilibré ou de façon déterministe), et sont exhaustivement recensées en cinq ans.

→ Les autres adresses sont réparties en 5 groupes de rotation par échantillonnage équilibré à probabilités égales sur les variables de type logement et socio-démographique (les mêmes variables que celles utilisées pour les petites communes, hors la population par département). Chaque année, un de ces groupes de rotation est concerné par les enquêtes de recensement ; on prélève dans le groupe de l'année en cours un échantillon aléatoire, avec comme variables d'équilibrage le nombre de logements total, le nombre de logements collectifs et le nombre de logements dans chacun des Iris. Cette dernière variable permet de mieux répartir les échantillons annuels sur l'ensemble de la commune. Cet échantillon aléatoire est prélevé de façon à ce que, annuellement, 8% des adresses (y compris les grandes adresses et les adresses neuves) soient enquêtées

6.2 Estimations basées sur une année de collecte

La collecte d'une année est réalisée sur un cinquième des petites communes et sur 8 % des adresses des grandes communes, soit environ 14% de l'ensemble des adresses (celles-ci se répartissant quasiment équitablement entre petites et grandes communes). Bien que cet échantillon soit d'une taille importante, il ne permet pas de produire des estimations précises à des niveaux

géographiques fins, en raison de la taille des unités (des petites communes notamment). Les estimations sont donc essentiellement produites au niveau national et régional, et portent le nom d'estimations globales.

Comme le plan de sondage est stratifié, entre les petites et les grandes communes tout d'abord, puis par région pour les petites communes, l'estimation et l'estimation de précision pour un domaine quelconque passent par les étapes suivantes :

1. Isoler chaque grande commune de l'échantillon, et réaliser indépendamment l'estimation sur chacune d'elles
2. Isoler région par région les petites communes de l'échantillon
3. Pour chaque région représentée, calculer une estimation sur le champ des petites communes

Compte tenu de l'indépendance, l'estimation sur le domaine et la variance associée s'obtiennent alors par sommation.

Dans la suite de cette section, nous appliquons les méthodes de Bootstrap proposées pour un échantillonnage équilibré au cas du Nouveau Recensement. Nos simulations portent sur une grande commune artificielle, obtenue par extraction de 1 000 adresses d'une grande commune dans le fichier du Recensement de 1999. Nous nous plaçons dans un cadre simplifié et ne nous intéressons qu'à l'échantillonnage de première phase d'une année, en supposant qu'un cinquième des adresses est sélectionné par échantillonnage à probabilités égales, équilibré sur les variables indiquées dans le tableau 6.1.

6.2.1 Estimation sur le champ des grandes communes : étude par simulations

Le plan de sondage consiste à trier aléatoirement la population (afin d'augmenter l'entropie du plan de sondage), puis à sélectionner un échantillon au cinquième, à probabilités égales, équilibré sur les variables mentionnées dans le tableau 6.1. Le plan de sondage est donc analogue à l'échantillonnage de première phase pratiqué par le Nouveau Recensement. Le total des variables d'intérêt données dans le tableau 6.1 est estimé directement à l'aide des poids de sondage.

TAB. 6.1 – Liste des variables disponibles sur la base d'adresses de Bretagne (source : RP 1999)

Variables d'équilibrage
Nombre de logements
Population des moins de 20 ans
Population des 20 à 39 ans
Population des 40 à 59 ans
Population des 60 à 74 ans
Population des 75 ans et plus
Population des hommes
Population des femmes

Variables d'intérêt
Nombre d'actifs
Nombre d'inactifs
Nombre de personnes d'origine française
Nombre de français par acquisition
Nombre d'étrangers de l'Union Européenne
Nombre d'étrangers hors Union Européenne

Pour simplifier, nous supposons

→ l'absence de non-réponse totale, c'est à dire que toutes les adresses échantillonnées sont supposées être effectivement enquêtées,

→ l'absence de non-réponse partielle, c'est à dire qu'au sein des adresses échantillonnées, toutes les variables d'intérêt sont effectivement relevées.

Ces postulats sont assez irréalistes pour une enquête réelle, mais l'objectif de la simulation est avant tout une validation empirique du Bootstrap dans le contexte d'une enquête simplifiée. La prise en compte de l'imputation de la non-réponse partielle dans les méthodes de rééchantillonnage a été étudiée par Rao and Shao (1992), Rao and Sitter (1996) et Shao and Sitter (1996), et fera l'objet de travaux ultérieurs pour la méthode de Bootstrap que nous proposons.

Une approximation de la précision est donnée par 20 000 simulations indépendantes. On calcule également, à l'aide de 20 000 simulations indépendantes,

la précision du plan de sondage obtenu en remplaçant l'étape préalable de tri aléatoire par un tri sur le nombre de logements décroissant.

Deux méthodes de Bootstrap sont utilisées : le Bootstrap de type BWO et la généralisation de la méthode mirror-match de Sitter (1992), voir le chapitre 4, section 5. Dans le cas traité, l'algorithme peut être simplifié car toutes les inverses de probabilités d'inclusion sont entières. A partir d'un échantillon, on constitue donc une seule pseudopopulation (en dupliquant 5 fois chaque individu échantillonné) dans laquelle on rééchantillonne de façon répétée. La précision donnée par chaque méthode de Bootstrap est approchée à l'aide du tirage de 200 échantillons, pour chacun desquels 1 000 rééchantillons Bootstrap sont prélevés. Les intervalles de confiance sont déterminés à l'aide de la méthode des percentiles.
On calcule également l'estimation de précision analytique correspondant à la formule 3 de Deville and Tillé (2005), à l'aide du tirage de 1 000 échantillons. Les intervalles de confiance sont déterminés à l'aide de l'approximation normale.

Le tableau 6.2 donne la précision (approchée par 20 000 simulations) du plan de sondage avec randomisation préalable, et celle (toujours approchée par 20 000 simulations) du même plan de sondage, mais où le tri aléatoire est remplacé par un tri informatif sur le nombre de logements décroissant. Ces précisions sont comparées avec celles données par les deux méthodes de Bootstrap et l'approximation proposée par Deville and Tillé (2005).
Le tri préalable sur le nombre de logements crée un effet de stratification, qui réduit la variance pour les variables d'intérêt bien corrélées au nombre de logements. Cet effet est encore plus sensible avec un équilibrage sur la seule probabilité d'inclusion, c'est à dire avec un simple échantillonnage de taille fixe, voir le tableau 6.3. Dans ce cas, l'échantillonnage équilibré avec tri aléatoire préalable est équivalent au sondage aléatoire simple, quand l'échantillonnage équilibré avec tri sur le nombre de logements est, en utilisant l'algorithme du Fast Cube, équivalent à un tirage stratifié de taille 1 dans chaque strate, les strates étant constituées en regroupant les adresses 5 par 5, par nombre de logements décroissant.
Pour un échantillonnage équilibré de type Recensement, la différence de variance observée avec les deux méthodes est cependant minime. Nous conjecturons que, pour un nombre de variables d'équilibrage important, un tri préalable a peu d'effets sur la précision de l'échantillonnage, et les formules

d'approximation de variance de Deville and Tillé (2005) sont donc largement utilisables.

Les trois méthodes testées donnent une approximation de variance raisonnable, même si la variance est généralement sous-estimée. Le Bootstrap de type BWO fournit la meilleure approximation, alors que le Bootstrap adapté de la méthode de Sitter présente généralement le biais le plus fort.

Le tableau 6.4 compare les taux de couverture effectifs des trois méthodes testées, pour un taux théorique de 10% . Les résultats obtenus pour un taux de couverture théorique de 5% ne présentent pas qualitativement de différence. Les trois méthodes donnent des résultats raisonnables, le résultat le moins bon est obtenu avec le Bootstrap adapté de la méthode Sitter.

Nous avons conduit une simulation analogue afin d'estimer la précision d'un échantillon annuel de petites communes, à l'aide d'un fichier issu du Recensement de 1999 et donnant, pour chaque petite commune de Bretagne, les variables nécessaires à un équilibrage de type Nouveau Recensement et quelques variables d'intérêt. Les résultats obtenus sont médiocres : chacune des trois méthodes sous-estime généralement très largement la variance. Ce problème, qui semble lié à la conjonction de la grande taille des unités et de la très forte corrélation entre les variables d'équilibrage et les variables d'intérêt, est actuellement à l'étude.

Dans la section suivante, nous développons une méthode de calcul de précision sur le champ des petites communes pour l'échantillon issu de trois années de collecte et la méthode d'estimation retenue, de type régression géographique pondérée.

6.3 Utilisation de plusieurs années de collecte : l'estimation sur zones mixtes

6.3.1 Introduction

A la suite des recommandations du COPAR (Comité d'Orientation Pour l'Action Régionale) du 25 mars 2006, six groupes de travail ont été lancés par le département de l'action régionale et l'unité RP (Recensement de la

Population) afin de préparer l'exploitation des enquêtes annuelles de Recensement. L'un de ces groupes avait pour objectif : *Réaliser un investissement "zones mixtes" pour le service spécifique : l'objectif est de valoriser le potentiel d'estimations et d'analyses croisées pour des variables du Recensement de la Population pour répondre aux demandes en région sur des zones à façon de taille suffisante (agglomérations, Etablissements Publics de Coopération Intercommunale - EPCI, pays, départements, etc...). Les possibilités d'exploitation des EAR (Enquêtes Annuelles de Recensement) à l'échelle infracommunale feront l'objet d'une démarche distincte qui sera lancée en 2007 dans le cadre des travaux de la DET (Division des Etudes Territoriales) et du PSAR (Pôle de Service à l'Action Régionale) "Analyse Urbaine".*

Une **zone mixte** est un territoire supracommunal contenant des grandes communes (plus de 10 000 habitants) et des petites communes (moins de 10 000 habitants) recensées ou non recensées avant 2007. Une des priorités du groupe de travail porte sur l'exploitation du Recensement pour estimer les évolutions démographiques récentes sur des zones mixtes.

Le groupe de travail était constitué de François Brunet (Service Etudes et Diffusion-SED Rhône Alpes), Brigitte Baccaïni, Christophe Barret et Pierre Carrelet (PSAR "Analyse territoriale"), Françoise Dupont (Unité Recensement de la Population - URP), David Levy (Service Etudes et Diffusion-SED Bretagne), Jean Laganier (Département de l'Action Régionale-DAR) et Jean-Luc Lipatz (Division Etudes Territoriales-DET). La méthode d'estimation sur zones mixtes a été développée par le groupe de travail, voir Baccaïni and Barret (2006) ; notre contribution a consisté à proposer un outil de calcul de précision pour la méthode d'estimation retenue.

6.3.2 La méthode

Le travail du groupe a essentiellement concerné l'estimation sur le champ des petites communes. Des estimations de population sont fournies par l'Unité RP pour chaque grande commune, ainsi qu'une typologie de précision. Du fait de la stratification du plan de sondage, le problème revient donc à réaliser une estimation sur le champ des petites communes de la zone mixte. Dans ce qui suit, nous ne nous intéressons qu'à l'estimation de population sur chaque zone mixte. L'estimation pour d'autres variables clés (le chômage notamment), et le calcul de précision correspondant, font l'objet de travaux

complémentaires.

Les estimations produites doivent être millésimées 2005. Le groupe s'est orienté vers une solution de type imputation : compte tenu des trois années de collecte disponibles (2004, 2005 et 2006), le problème à résoudre était double :

→ pour les petites communes recensées entre 2004 et 2006, ramener (si nécessaire) les chiffres de population en 2005,
→ pour les petites communes non recensées avant 2006, obtenir une prédiction de leur population en 2005 à l'aide de données. administratives

Cas des communes recensées entre 2004 et 2006

La population des petites communes enquêtées en 2005 est conservée en l'état. Pour les petites communes enquêtées en 2004 ou 2006, plusieurs scénarios ont été envisagés pour tenir compte du décalage temporel. C'est finalement une solution d'extrapolation/interpolation, à l'aide du recensement de 1999, qui a été retenue. Ainsi, pour les communes enquêtées en 2004

$$\hat{y}_{k,05} = y_{k,04} \times \left(\frac{y_{k,04}}{y_{k,99}} \right)^{1/5},$$

et pour les communes enquêtées en 2006

$$\hat{y}_{k,05} = y_{k,06} \times \left(\frac{y_{k,06}}{y_{k,99}} \right)^{-1/7},$$

où $y_{k,A}$ représente la population de la commune k l'année A, et $\hat{y}_{k,05}$ la population estimée de la commune k en 2005.

Cas des communes recensées après 2006

Pour les communes non encore recensées, on dispose d'une source externe, le fichier Revenus Fiscaux Localisés (RFL), constitué à partir du fichier de la taxe d'habitation (TH) et du fichier de l'impôt sur le revenu (IR). Dans les deux principales méthodes d'estimation proposées (voir ci-après), le principe est le même : modéliser le lien entre la population donnée par les EAR et la population donnée par le fichier RFL, afin d'imputer les communes non

encore enquêtées.

La première méthode, proposée par David Levy (et simplement notée dans la suite **méthode DL**), consiste à partitionner chaque région en domaines. Un découpage par tranches de communes a été testé : communes de moins de 6 000 habitants, de 6 000 à 8 000 habitants, de plus de 8 000 habitants (une solution plus fine consistant à réaliser une classification des communes de chaque région sur certaines variables clés). Au sein de chaque domaine, un modèle expliquant la population (éventuellement ramenée en 2005) des petites communes enquêtées entre 2004 et 2006 par la population RFL est ajusté, et la population des petites communes non enquêtées est prédite à l'aide des paramètres estimés et de la population RFL.

La seconde méthode, dite méthode par régression géographique pondérée, a été proposée par Jean-Luc Lipatz. On la notera dans la suite **méthode JLL**. Elle consiste à définir autour de chaque commune à imputer un disque sur lequel le modèle est ajusté comme précedemment ; la population en 2005 de la commune centre du disque est alors prédite à l'aide des paramètres estimés et de la population RFL de cette commune. C'est cette méthode qui a été finalement retenue pour les estimations sur zones mixtes.

6.3.3 Estimation de précision

Pour chacune des deux méthodes d'estimation envisagées, deux techniques d'estimation de précision sont proposées et mises en oeuvre ci-dessous :

→ Si l'on disposait de la base de sondage entière, avec pour chaque individu les variables d'intérêt, il serait possible d'approcher la précision par des simulations, en répétant un grand nombre de fois la stratégie d'échantillonnage et d'estimation. Comme cette base de sondage n'est pas accessible, on a recours à une population artificielle (notée ARTI) proche de la population réelle, et constituée à l'aide des données de la TH. La précision calculée par simulations est supposée "portable" à la précision des enquêtes de recensement sur le champ des petites communes.

→ Une autre méthode consiste à n'utiliser que les données de l'enquête et les sources administratives disponibles. Un algorithme de Bootstrap, reproduisant de façon approchée la stratégie d'échantillonnage et d'estimation, est proposé et mis en oeuvre.

Notre étude concerne la population des 1 237 petites communes de Bretagne. Comme toutes les variables d'intérêt correspondant à l'enquête ne sont pas disponible, on a recours à la population artificielle ARTI évoquée précedemment : on dispose ainsi d'un fichier donnant pour chaque petite commune une approximation de sa population, pour chaque année entre 1999 et 2005. On dispose également de données administratives donnant la population de chaque commune en 2004 et 2005. Les deux techniques d'estimation de précision évoquées ci-dessus sont mises en oeuvre et comparées sur cette population artificielle. Notons que notre étude ne permet pas de statuer sur la portabilité de la précision calculée par simulations de la population artificielle à la population réelle.

Méthode DL : modèle par domaines

La précision est approchée à l'aide de 1 000 simulations indépendantes, selon la technique décrite dans l'algorithme 6.1. Il est important de rappeler que la différence entre l'approximation de précision produite ici et la vraie précision est dûe, d'une part à la différence entre la population utilisée et la population cible, d'autre part au nombre de simulations : pour être proche de la vraie précision, il faudrait un grand nombre de simulations, (au moins 10 000). Nous avons choisi ici de retenir le même nombre de simulations qu'avec la méthode JLL, ce dernier étant limité en raison de la complexité algorithmique de la méthode.

La précision est également estimée en utilisant une technique de Bootstrap qui s'appuie sur les seules données résultant de trois années de collecte du Nouveau Recensement. La technique utilisée, inspirée du Wild Bootstrap de Wu (1986), consiste à reconstituer une pseudopopulation en se basant sur le modèle de prédiction. Cette technique se rapproche davantage d'un Bootstrap classique que des méthodes de Bootstrap en population finie présentées dans les chapitres précédents. Ce choix s'explique par la méthode d'estimation utilisée : celle-ci ne tient pas compte du plan de sondage, et se base entièrement sur un modèle. L'algorithme 6.2 est appliqué 20 fois, avec $B = 100$, afin de comparer l'approximation Bootstrap de variance avec l'approximation donnée par les 1 000 simulations indépendantes.

Les tableaux 6.5, 6.6 et 6.7 comparent pour chacune des deux techniques la précision calculée respectivement au niveau de la région, des départements et

de communautés de communes. Le tableau 6.7 ne donne que les communautés présentant le CV le plus important ; les résultats obtenus sur les autres ne présentent pas qualitativement de différence. Le Bootstrap donne une approximation raisonnable de la précision (même si pour affiner les résultats obtenus, il faudrait réaliser plus de 20 répétitions de Bootstrap). La méthode est ici plutôt conservative (on a tendance à surestimer la variance).

Méthode JLL : la régression géographique pondérée

La précision est à nouveau approchée à l'aide de 1 000 simulations indépendantes, selon la technique décrite dans l'algorithme 6.3. Le nombre de simulations est limité par la complexité algorithmique de la méthode. La précision est également estimée en utilisant une technique adaptée du Wild Bootstrap de Wu (1986) et présentée dans l'algorithme 6.4, qui reconstitue une pseudopopulation en se basant sur le modèle de prédiction. L'algorithme 6.4 est appliqué 20 fois, avec $B = 100$.

Les tableaux 6.8, 6.9 et 6.10 comparent pour chacune des deux techniques la précision calculée respectivement au niveau de la région, des départements et de communautés de communes. Là encore, le tableau 6.10 ne donne que les communautés présentant le CV le plus important. On constate que le Bootstrap a tendance à sous-estimer la variance, et que la méthode JLL présente plus de variabilité que la méthode DL. Bien que le travail réalisé se soit limité au cas d'une seule région, les résultats obtenus ne justifient pas la complexité supplémentaire liée à l'utilisation de la méthode JLL par un gain de précision conséquent.

FIG. 6.1 – Calcul de précision par simulation, méthode de David Levy

Etape 1. Partitionner la population en 5 groupes de rotation, chacun étant sélectionné par échantillonnage à probabilités égales, et équilibré sur les variables utilisées par le Nouveau Recensement.

Etape 2. La population ARTI des petites communes du groupe 1 (respectivement du groupe 3) est extrapolée (respectivement rétropolée) pour être ramenée en 2004. La population ARTI des petites communes du groupe 2 est inchangée.

Etape 3. Les petites communes des groupes de rotation 1 à 3 sont partitionnées en 3 domaines (communes de moins de 6 000 habitants, communes de 6 000 à 8 000 habitants, communes de plus de 8 000 habitants au sens de la population RFL), notés U_1, U_2, U_3. La population ARTI des petites communes du domaine U_h est expliquée par la population RFL, selon le modèle :

$$y^{ARTI} = \alpha_h + \beta_h y^{RFL} + \epsilon_h$$
$$E(\epsilon_h) = 0 \quad V(\epsilon_h) = \sigma_h^2$$
$$\Rightarrow \text{Estimation} : \hat{\alpha_h}, \hat{\beta}_h, \hat{\sigma_h^2},$$

Etape 4. Les petites communes des groupes de rotation 4 et 5 sont partitionnées en 3 domaines (communes de moins de 6 000 habitants, communes de 6 000 à 8 000 habitants, communes de plus de 8 000 habitants au sens de la population RFL), notés V_1, V_2, V_3. La population ARTI des petites communes du domaine V_h est estimée à l'aide des paramètres $\hat{\alpha_h}, \hat{\beta}_h$ précedemment estimés :

$$\hat{y}^{ARTI} = \hat{\alpha_h} + \hat{\beta}_h y^{RFL}$$

Etape 5. A l'aide des étapes 2 à 4, on reconstitue la population ARTI pour l'ensemble des petites communes. La population d'une zone mixte s'obtient en sommant la population reconstituée des petites communes de la zone mixte. Soit $\hat{Y}^{ZM}$ la population ainsi calculée sur la zone mixte

Etape 6. On répète les étapes 1 à 5 un grand nombre de fois, disons B, pour obtenir B estimations $\hat{Y_1}^{ZM}, \ldots, \hat{Y_B}^{ZM}$. La précision est approchée par la variance empirique $\frac{1}{B-1} \sum_{i=1}^{B} \left(\hat{Y_i}^{ZM} - \hat{Y_-}^{ZM} \right)^2$ où $\hat{Y_-}^{ZM} = \frac{1}{B} \sum_{i=1}^{B} \hat{Y_i}^{ZM}$

FIG. 6.2 – Calcul de précision par Bootstrap, méthode de David Levy

Etape 1. On dispose des données des 3 premiers groupes de rotation. La population RP des petites communes du groupe 1 (respectivement du groupe 3) est extrapolée (respectivement rétropolée) pour être ramenée en 2005. La population RP des petites communes du groupe 2 est inchangée.

Etape 2. Les petites communes des groupes de rotation 1 à 3 sont partitionnées en 3 domaines (communes de moins de 6 000 habitants, de 6 000 à 8 000, de plus de 8 000 au sens de la population RFL), notés U_1, U_2, U_3. La population RP des petites communes du domaine U_h est expliquée par la population RFL, selon le modèle :

$$\begin{array}{c} y^{ARTI} = \alpha_h + \beta_h y^{RFL} + \epsilon_h \\ E(\epsilon_h) = 0 \quad V(\epsilon_h) = \sigma_h^2 \end{array} \quad \Rightarrow \quad \text{Estimation : } \hat{\alpha_h}, \hat{\beta_h}, \hat{\sigma_h^2}$$

Etape 3. Les petites communes des groupes de rotation 4 et 5 sont partitionnées en 3 domaines (communes de moins de 6 000 habitants, de 6 000 à 8 000, de plus de 8 000 habitants au sens de la population RFL), notés V_1, V_2, V_3. La population RP des petites communes du domaine V_h est estimée à l'aide des paramètres $\hat{\alpha_h}, \hat{\beta_h}$ calculés précédemment, plus un alea :

$$\hat{y}^{ARTI} = \hat{\alpha_h} + \hat{\beta_h} y^{RFL} + \eta_h \quad \text{où} \quad \eta_h \sim \mathcal{N}(0, \hat{\sigma_h^2})$$

On reconstitue ainsi une pseudo-population U^*. La population reconstituée d'une commune k de U^* est notée y_k^{*ARTI}.

Etape 4. Partitionner la pseudo-population U^* en 5 groupes de rotation, chacun étant sélectionné par échantillonnage à probabilités égales, et équilibré sur les variables utilisées par le Nouveau Recensement

Etape 5. Les petites communes des groupes de rotation 1 à 3 sont partitionnées en 3 domaines (communes de moins de 6 000 habitants, communes de 6 000 à 8 000 habitants, communes de plus de 8 000 habitants au sens de la population RFL), notés U_1^*, U_2^*, U_3^*. La pseudo-population RP des petites communes du domaine U_h^* est expliquée par la population RFL, selon le modèle :

$$\begin{array}{c} y^{*ARTI} = \alpha_h^* + \beta_h^* y^{RFL} + \epsilon_h^* \\ E(\epsilon_h^*) = 0 \quad V(\epsilon_h^*) = \sigma_h^{*2} \end{array} \quad \Rightarrow \quad \text{Estimation : } \hat{\alpha_h^*}, \hat{\beta_h^*}, \hat{\sigma_h^{*2}}$$

Etape 6. Les petites communes des groupes de rotation 4 et 5 sont partitionnées en 3 domaines (communes de moins de 6 000 habitants, communes de 6 000 à 8 000 habitants, communes de plus de 8 000 habitants au sens de la population RFL), notés V_1^*, V_2^*, V_3^*.La pseudo-population RP des petites communes du domaine V_h^* est estimée à l'aide des paramètres $\hat{\alpha_h^*}, \hat{\beta_h^*}$ précédemment estimés :

$$\hat{y}^{*ARTI} = \hat{\alpha_h^*} + \hat{\beta_h^*} y^{RFL}$$

Etape 7. A l'aide des étapes 5 et 6, on obtient ainsi une version Bootstrap complète du fichier. La population Bootstrap d'une zone mixte s'obtient en sommant la population des petites communes de la zone mixte. Soit $Y^{*\hat{ZM}}$ la population ainsi calculée sur la zone mixte ZM.

Etape 8. On répète les étapes 4 à 7 un grand nombre de fois, disons B, pour obtenir B estimations $Y_1^{*\hat{ZM}}, \ldots, Y_B^{*\hat{ZM}}$. La précision est estimée par la variance empirique $\frac{1}{B-1} \sum_{i=1}^{B} \left(\hat{Y_i}^{*ZM} - \hat{Y_-}^{*ZM} \right)^2$ où $\hat{Y_-}^{*ZM} = \frac{1}{B} \sum_{i=1}^{B} \hat{Y_i}^{*ZM}$

FIG. 6.3 – Calcul de précision par simulations, méthode de JL Lipatz

Etape 1. Partitionner la population en 5 groupes de rotation, chacun étant sélectionné par échantillonnage à probabilités égales, et équilibré sur les variables utilisées par le Nouveau Recensement.

Etape 2. La population ARTI des petites communes du groupe 1 (respectivement du groupe 3) est extrapolée (respectivement rétropolée) pour être ramenée en 2004 . La population ARTI des petites communes du groupe 2 est inchangée.

Etape 3. On trace un disque de 30 km autour de chaque commune des groupes de rotation 4 et 5. Dans chaque disque d, la population ARTI des petites communes des groupes de rotation 1 à 3 est expliquée par la population RFL, selon le modèle :

$$\begin{gathered} y^{ARTI} = \alpha_d + \beta_d y^{RFL} + \epsilon_d \\ E(\epsilon_d) = 0 \quad V(\epsilon_d) = \sigma_d^2 \\ \Rightarrow \text{Estimation} : \hat{\alpha_d}, \hat{\beta_d}, \hat{\sigma_d^2} \end{gathered}$$

Etape 4. On estime la population ARTI de la commune centre du disque à l'aide des paramètres précédemment estimés :

$$y^{\hat{ARTI}} = \hat{\alpha_d} + \hat{\beta_d} y^{RFL}$$

Etape 5. A l'aide des étapes 2 à 4, on reconstitue la population ARTI pour l'ensemble des petites communes. La population d'une zone mixte s'obtient en sommant la population reconstituée des petites communes de la zone mixte. Soit $Y^{\hat{ZM}}$ la population ainsi calculée sur la zone mixte

Etape 6. On répète les étapes 1 à 5 un grand nombre de fois, disons B, pour obtenir B estimations $\hat{Y_1}^{ZM}, \ldots, \hat{Y_B}^{ZM}$. La précision est approchée par la variance empirique $\frac{1}{B-1} \sum_{i=1}^{B} \left(\hat{Y_i}^{ZM} - \hat{Y_-}^{ZM} \right)^2$ où $\hat{Y_-}^{ZM} = \frac{1}{B} \sum_{i=1}^{B} \hat{Y_i}^{ZM}$

FIG. 6.4 – Calcul de précision par Bootstrap, méthode de JL Lipatz

Etape 1. On dispose des données des 3 premiers groupes de rotation. La population RP des petites communes du groupe 1 (respectivement du groupe 3) est extrapolée (respectivement rétropolée) pour être ramenée en 2005. La population RP des petites communes du groupe 2 est inchangée.

Etape 2. On trace un disque de 30 km autour de chaque commune des groupes de rotation 4 et 5. Dans chaque disque d, la population RP des petites communes des groupes de rotation 1 à 3 est expliquée par la population RFL, selon le modèle :

$$y^{ARTI} = \alpha_d + \beta_d y^{RFL} + \epsilon_d$$
$$E(\epsilon_d) = 0 \quad V(\epsilon_d) = \sigma_d^2$$
$$\Rightarrow \text{Estimation} : \hat{\alpha_d}, \hat{\beta_d}, \hat{\sigma_d^2}$$

Etape 3. On estime la population RP de la commune centre du disque à l'aide des paramètres $\hat{\alpha_d}, \hat{\beta_d}$ précédemment estimés, plus un alea :

$$\hat{y}^{ARTI} = \hat{\alpha_d} + \hat{\beta_d} y^{RFL} + \eta_d$$
$$\text{où } \eta_d \sim \mathcal{N}(0, \hat{\sigma_d^2})$$

On reconstitue ainsi une pseudo-population U^*. La population reconstituée d'une commune k de U^* est notée y_k^{*ARTI}.

Etape 4. Partitionner la pseudo-population en 5 groupes de rotation, chacun étant sélectionné par échantillonnage à probabilités égales, et équilibré sur les variables utilisées par le Nouveau Recensement

Etape 5. On trace un disque de 30 km autour de chaque commune des groupes de rotation 4 et 5 de U^*. Dans chaque disque d, la population RP des petites communes des groupes de rotation 1 à 3 est expliquée par la population RFL, selon le modèle :

$$y^{*ARTI} = \alpha_d^* + \beta_d^* y^{RFL} + \epsilon_d^*$$
$$E(\epsilon_d^*) = 0 \quad V(\epsilon_d^*) = \sigma_d^{*2}$$
$$\Rightarrow \text{Estimation} : \hat{\alpha_d^*}, \hat{\beta_d^*}, \hat{\sigma_d^{*2}}$$

Etape 6. On estime la pseudo-population RP de la commune centre du disque à l'aide des paramètres $\hat{\alpha_d^*}, \hat{\beta_d^*}$ précédemment estimés :

$$\hat{y}^{*ARTI} = \hat{\alpha_d^*} + \hat{\beta_d^*} y^{RFL}$$

Etape 7. A l'aide des étapes 5 et 6, on obtient ainsi une version Bootstrap complète du fichier. La population Bootstrap d'une zone mixte s'obtient en sommant la population des petites communes de la zone mixte. Soit $\hat{Y}^{*ZM}$ la population ainsi calculée sur la zone mixte

Etape 8. On répète les étapes 4 à 7 un grand nombre de fois, disons B, pour obtenir B estimations $Y_1^{*\hat{Z}M}, \ldots, Y_B^{*\hat{Z}M}$. La précision est estimée par la variance empirique $\frac{1}{B-1} \sum_{i=1}^{B} \left(\hat{Y_i}^{*ZM} - \hat{Y_-}^{*ZM} \right)^2$ où $\hat{Y_-}^{*ZM} = \frac{1}{B} \sum_{i=1}^{B} \hat{Y_i}^{*ZM}$

Conclusion

Ce chapitre présente une application des méthodes de Bootstrap au cas du Nouveau Recensement. Dans une première partie, nous appliquons la méthode de Bootstrap proposée pour un échantillonnage équilibré au cas d'un échantillonnage de type recensement, avec une estimation basée sur les poids de sondage. Cette méthode se compare très favorablement à une approche analytique basée sur les formules de Deville and Tillé (2005) et à l'extension proposée de la méthode mirror-match (Sitter, 1992).

Dans une deuxième partie, nous proposons une nouvelle méthode inspirée du Wild Bootstrap de Wu (1986) afin de prendre en compte une méthode d'estimation complexe non basée sur les poids de sondage. Nous montrons à l'aide de quelques simulations que cette méthode restitue bien l'ordre de grandeur de la variance.

TAB. 6.2 – Approximation de variance avec les trois méthodes testées pour l'estimation du total des variables d'intérêt

	Variance ($\times 10^4$)		Ecart relatif (%) (par rapport à la colonne 2)		
Total	Tri nb de logements	Tri aléatoire	Bootstrap BWO	Bootstrap Sitter	App. Deville Tillé
Actifs	3.79	4.18	-3.26	-5.56	-5.50
Inactifs	3.88	4.25	-2.73	-3.22	-7.50
Français d'origine	9.90	10.73	-3.88	-10.10	-3.94
Français par acquisition	2.05	2.23	-1.22	-2.35	+1.41
Etrangers de l'UE	8.78	8.78	-5.12	-13.38	-3.59
Etrangers hors UE	4.54	4.81	-4.87	-10.19	-2.93

TAB. 6.3 – Approximation de variance pour un échantillonnage de taille fixe, avec tri aléatoire ou tri informatif, pour l'estimation du total des variables d'intérêt

Variable	Variance ($\times 10^4$) Tri nb de logements	Variance ($\times 10^4$) Tri aléatoire	Corrélation au nb de logements
Actifs	15.54	22.56	0.37
Inactifs	21.13	30.66	0.37
Français d'origine	29.34	44.34	0.48
Français par acquisition	2.87	3.91	0.15
Etrangers de l'UE	9.17	9.47	0.06
Etrangers hors UE	5.98	7.68	0.09

TAB. 6.4 – Taux de couverture pour 3 méthodes d'estimation de précision, obtenus avec un IC de type percentile, pour un échantillonnage en petite commune de type recensement (niveau théorique : 10 %)

	Bootstrap BWO			Bootstrap Sitter			App. Deville-Tillé		
Variable	L	U	L+U	L	U	L+U	L	U	L+U
Actifs	6.5	5.5	12.0	7.5	6.0	13.5	7.5	6.0	13.5
Inactifs	7.0	7.0	14.0	6.5	8.0	14.5	6.5	7.5	14.0
Français d'origine	6.5	6.0	12.5	6.5	6.5	13.0	6.0	4.5	10.5
Français par acquisition	4.0	6.5	10.5	5.5	7.0	12.5	4.5	6.5	11.0
Etrangers de l'UE	3.0	5.5	8.5	2.5	5.5	8.0	2.5	6.0	8.5
Etrangers hors UE	6.0	7.5	13.5	6.5	8.0	14.5	4.5	7.5	12.0

Note de lecture : cf. tableau 2.2

Tab. 6.5 – Coefficient de variation estimé à la région

Population au RP99	CV simulations (%)	CV Bootstrap (%)
2 074 612	0.15	0.19

Tab. 6.6 – Coefficient de variation estimé par département

Département	Population au RP99	CV simulations (%)	CV Bootstrap (%)
22	472 855	0.32	0.33
29	553 504	0.21	0.26
35	564 429	0.24	0.28
56	483 824	0.25	0.28

Tab. 6.7 – Coefficient de variation estimé par EPCI (*coefficient de variation médian*)

EPCI	Population au RP99	CV simulations (%)	CV Bootstrap (%)
244400610	4 511	12,50 %	13,26 %
242200533	12 806	9,28 %	7,37 %
242900785	4 366	8,29 %	10,71 %
245614458	6 004	8,04 %	8,30 %
242200822	3 389	5,97 %	6,24 %
242200749	1 961	5,38 %	7,34 %
242913325	16 002	5,03 %	6,79 %
242913317	7 386	4,06 %	3,23 %
242200855	6 285	4,04 %	4,00 %
243500576	8 349	3,98 %	4,38 %
242214450	7 304	3,29 %	4,73 %
243500683	6 800	3,05 %	2,38 %
242900652	9 715	2,94 %	2,84 %
242200830	11 846	2,79 %	2,33 %
245600432	9 331	2,76 %	2,16 %
242200053	*12 806*	*0,94 %*	*1,24 %*

Tab. 6.8 – Coefficient de variation estimé à la région

Population au RP99	CV simulations (%)	CV Bootstrap (%)
2 074 612	0.16	0.14

Tab. 6.9 – Coefficient de variation estimé par département

Département	Population au RP99	CV simulations (%)	CV Bootstrap (%)
22	472 855	0.48	0.37
29	553 504	0.29	0.24
35	564 429	0.31	0.25
56	483 824	0.36	0.31

Tab. 6.10 – Coefficient de variation estimé par EPCI (*coefficient de variation médian*)

EPCI	Population au RP99	CV simulations (%)	CV Bootstrap (%)
244400610	4 511	17,33 %	13,31 %
242200533	12 806	13,08 %	8,23 %
242900785	4 366	12,09 %	12,71 %
245614458	6 004	9,95 %	7,57 %
242200822	3 389	8,37 %	6,53 %
242200749	1 961	6,15 %	4,52 %
242913325	16 002	5,45 %	5,35 %
242913317	7 386	5,25 %	3,50 %
242200855	6 285	5,09 %	4,03 %
243500576	8 349	4,81 %	4,66 %
242214450	7 304	4,26 %	3,40 %
243500683	6 800	4,25 %	3,17 %
242900652	9 715	4,04 %	3,23 %
242200830	11 846	3,94 %	2,75 %
245600432	9 331	3,92 %	3,04 %
242200053	*12 806*	*1,34 %*	*1,03 %*

Bibliographie

Baccaïni, B. and Barret, C. (2006). Zones mixtes : guide méthodologique. Psar analyse territoriale, INSEE, France.

Bertrand, P., Christian, B., Chauvet, G., and Grosbras, J.-M. (2004). Plans de sondage pour le recensement rénové de la population. In *Séries INSEE Méthodes : Actes des Journées de Méthodologie Statistique*, Paris. Paris : INSEE, to appear.

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

Dumais, J. and Isnard, M. (2000). Le sondage de logements dans les grandes communes dans le cadre du recensement rénové de la population. In *Séries INSEE Méthodes : Actes des Journées de Méthodologie Statistique*, volume 100. pp. 37-76, Paris : INSEE.

Durr, J.-M. and Dumais, J. (2002). Design of the french census of population. *Survey Methodology*, 28 :262–269.

Rao, J. and Shao, J. (1992). Jackknife variance estimation with survey data under hot-deck imputation. *Biometrika*, 79 :811–822.

Rao, J. and Sitter, R. (1996). Variance estimation under two-phase sampling with application to imputation for missing data. *Biometrika*, 82 :453–460.

Shao, J. and Sitter, R. (1996). Bootstrap for imputed survey data. *Journal of the American Statistical Association*, 91 :1278–1288.

Sitter, R. (1992). A resampling procedure for complex survey data. *Journal of the American Statistical Association*, 87 :755–765.

Wu, C. (1986). Jackknife, bootstrap and other resampling methods in regression analysis (with discussions). *Annals of Statistics*, 14 :1261–1350.

Conclusion et perspectives

Nous présentons dans ce travail une méthode de Bootstrap pour un échantillonnage à probabilités inégales, qui généralise les algorithmes BWO de Gross (1980) et Booth et al. (1994). Nous montrons que cette méthode est applicable pour un plan de sondage à entropie forte, ce qui inclut le tirage poissonien, le tirage réjectif (Hájek, 1964), les tirages de taille fixe à probabilités inégales proches de l'entropie maximale (Berger, 1998) et le tirage équilibré (Deville and Tillé, 2005). Cette méthode peut aisément s'appliquer au cas d'un échantillonnage stratifié. Une modification permet également de l'étendre au cas d'un échantillonnage à plusieurs degrés. Elle permet de prendre en compte facilement le redressement d'un estimateur.

Dans nos simulations, nous avons produit des intervalles de confiance en utilisant la méthode des percentiles et une méthode de type t-Bootstrap. La méthode du t-Bootstrap nécessite, pour chaque rééchantillon, de disposer d'un estimateur consistant de variance. Une possibilité consiste à effectuer un double Bootstrap à partir du rééchantillon : cette possibilité, qui nécessiterait un temps de calcul prohibitif, n'a pas été testée ici. Une autre possibilité consiste à utiliser la technique de linéarisation pour produire un estimateur approché de variance. Les simulations réalisées ne montrent pas que les taux de couverture théoriques soient mieux respectés avec cette méthode. La méthode des percentiles a l'avantage d'être plus rapide, de fournir des intervalles de confiance facilement y compris pour des statistiques fortement non linéaires, et de permettre la production de poids Bootstrap donnés par le rééchantillonnage qui offrent une bonne portabilité de l'estimation de précision. Nous recommandons donc l'utilisation de la méthode des percentiles.

Nous avons également comparé un algorithme exact de Bootstrap avec un

algorithme plus rapide mais donnant une estimation de variance légèrement biaisé. Nos simulations montrent le bon comportement de ce deuxième algorithme. Si le taux de sondage de l'enquête reste limité, nous recommandons donc également l'utilisation de cette méthode de Bootstrap approché.

Les enquêtes sont généralement entâchées de non-réponse, corrigées par imputation pour la non-réponse partielle et par repondération pour la non-réponse totale. Il est presque impossible de livrer un ordre de grandeur raisonnable de la précision sans prendre en compte ce problème. Cela constitue la prochaine étape de ce travail.

Un autre point important concerne l'estimation de précision dans le cas de deux échantillons. Ce problème apparaît par exemple lorsque l'on veut évaluer l'évolution d'un indicateur entre deux dates, et la précision de cette évolution. Une comparaison entre les méthodes de type Bootstrap et la technique de linéarisation par la fonction d'influence dans le cas de deux échantillons est actuellement à l'étude (Chauvet et al., 2007).

Certains points présentés rapidement dans ce travail mériteraient également une justification plus rigoureuse. D'autres pistes de recherche incluent l'extension de l'approximation de variance dûe à Hájek au cas d'un plan de taille fixe à probabilités inégales randomisé, une approximation des probabilités d'inclusion doubles pour le tirage équilibré sur plusieurs variables, et de façon plus pratique le calcul de précision pour les échantillons de petites communes du Nouveau Recensement.

Bibliographie

Berger, Y. (1998). Variance estimation using list sequential scheme for unequal probability sampling. *Journal of Official Statistics*, 14 :315–323.

Booth, J., Butler, R., and Hall, P. (1994). Bootstrap methods for finite populations. *Journal of the American Statistical Association*, 89 :1282–1289.

Chauvet, G., Goga, C., and Ruiz-Gazen, A. (2007). Estimation de variance entre deux échantillons : comparaison entre linéarisation et bootstrap. *in progress*.

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

Gross, S. (1980). Median estimation in sample surveys. *ASA Proceedings of Survey Research*, pages 181–184.

Hájek, J. (1964). Asymptotic theory of rejective sampling with varying probabilities from a finite population. *Annals of Mathematical Statistics*, 35 :1491–1523.

Résumé : Cette thèse est consacrée aux méthodes de Bootstrap pour une population finie. Le premier chapitre introduit quelques rappels sur l'échantillonnage et propose une présentation synthétique des principales méthodes d'estimation de précision. Le chapitre 2 rappelle les méthodes de Bootstrap proposées pour un sondage aléatoire simple et introduit deux nouvelles méthodes. Le chapitre 3 donne un nouvel algorithme de Bootstrap, consistant pour l'estimation de variance d'un estimateur par substitution dans le cas d'un tirage à forte entropie. Dans le chapitre 4, nous introduisons la notion d'échantillonnage équilibré et proposons un algorithme rapide. Nous montrons que l'algorithme de Bootstrap proposé est également consistant pour l'estimation de variance d'un tirage équilibré à entropie maximale. Le cas d'un échantillonnage complexe et celui d'un redressement est traité au chapitre 5. Une application au Nouveau Recensement de la population est donnée dans le chapitre 6.

English Title : Bootstrap methods for finite population

Abstract : This Phd deals with Bootstrap methods for finite population sampling. The first chapter introduces some bases about sampling and gives an overview of the main variance estimation techniques. A remind on Bootstrap methods for simple random sampling is given in chapter 2, and two new methods are introduced. A Bootstrap algorithm for unequal probability sampling is proposed in chapter 3, and shown to be consistent for variance estimation of plug-in statistics in case of large entropy sampling designs. Balanced sampling is presented in chapter 4, and a fast algorithm is proposed. Former Bootstrap algorithm is shown to be consistent as well in case of variance estimation for a maximum entropy balanced sampling design. Cases of complex sampling designs or reweighting strategies are discussed in chapter 5. An application to the French Renovated Census is given in chapter 6.

Discipline : Mathématiques Appliquées, Statistique

Mots clés : estimation de variance, linéarisation, Bootstrap, fonction d'influence, échantillonnage équilibré, algorithme du Cube, entropie maximale.

Keywords : variance estimation, linearization, Bootstrap, influence function, balanced sampling, Cube algorithm, maximum entropy.

Ecole doctorale, Humanités et Sciences de l'Homme. Université de Rennes 2.