ANNALES D'ÉCONOMIE ET DE STATISTIQUE. – N° 46 — 1997

Bootstrap généralisé d'un sondage

Patrice BERTAIL, Pierre COMBRIS*

RÉSUMÉ. – Le bootstrap généralisé introduit par Lo [1991] et Mason et Newton [1992] est appliqué aux méthodes de sondages. L'idée est de choisir des poids aléatoires (ou plan de rééchantillonnage) adéquats, imitant les fluctuations initiales du plan de sondage et pouvant tenir compte des critères de choix du sondeur, pour construire un estimateur de la distribution des statistiques d'intérêt. Après avoir donné quelques résultats de simulations dans le cas de la construction d'intervalles de confiance pour un ratio, la méthode est appliquée à la construction d'intervalles de confiance pour des moyennes, des ratios et des fractiles de consommations de produits alimentaires tirées des panels de ménages de Secodip.

Weighted Bootstrap in Survey Sampling

ABSTRACT. – The generalized bootstrap, introduced by Lo [1991] and further by Mason and Newton [1992] is applied to survey sampling. The idea is to choose adequate random weights (or resampling plans) so as to imitate the initial fluctuations of the probability sampling, maybe taking into account the priorities of the survey statistician, to obtain an estimator of the distribution of the statistics. The method is applied to the construction of confidence intervals for means, ratios and fractiles of food consumption data from the household panel surveys of Secodip.

* P. BERTAIL, P. COMBRIS : INRA-CORELA. Cette recherche a été réalisée dans le cadre du programme de mise en place de l'Observatoire des Consommations Alimentaires, financé par la Direction Générale de la Concurrence, de la Consommation et de la Répression des Fraudes (DGCCRF), par la Direction Générale de l'Alimentation (DGAL), et par la Direction Générale de la Santé (DGS).

1 Introduction

Le calcul de la précision d'une statistique dans un sondage est un problème délicat, qui ne peut être résolu théoriquement que pour des statistiques linéaires, sous des hypothèses restrictives sur les plans de sondage utilisés. De plus la construction d'intervalles de confiance repose sur l'hypothèse qu'on peut, lorsque la taille de la population étudiée, N, et la taille de l'échantillon observé, n, sont très grandes, admettre un principe de normalité asymptotique. A N et n fixés, une telle hypothèse est parfois inadaptée et peut conduire à des intervalles de confiance inappropriés. Lorsque le paramètre étudié est non linéaire, par exemple un fractile ou un ratio, la question devient encore plus ardue car on se ramène dans ce cas à appliquer les résultats parfois peu convaincants obtenus sur l'estimateur d'un total à une version linéarisée de la statistique.

Les méthodes de rééchantillonnage parmi lesquels le jackknife (QUENOUILLE [1949]) et le bootstrap (EFRON [1979]), à l'origine utilisées pour estimer la variabilité d'un estimateur dans un modèle d'échantillonnage, ont été transposées de diverses manières au cas des sondages (voir DEVILLE [1987]). La validation des méthodes proposées a été essentiellement obtenue sur des statistiques linéaires (CHAO et LO [1985]). On est généralement conduit à proposer une forme de bootstrap adaptée au plan de sondage, dont la validité asymptotique repose sur le comportement asymptotique relatif de N et n. Nous proposons dans cet article une méthodologie basée sur la méthode du bootstrap généralisé, introduite par LO [1991] (sous le nom de « clones bootstrap bayesiens ») puis généralisée par MASON et NEWTON [1992] et développée par PRAESTGAARD [1992] et par BARBE et BERTAIL [1995]. La philosophie du bootstrap généralisé est d'imiter les fluctuations dues au tirage (tirage de v.a. i.i.d., plan de sondage, etc.) en pondérant les observations par un système de poids aléatoires (ou plan de rééchantillonnage), pouvant dépendre des observations et de la statistique étudiée. Le choix de ces poids dépend bien entendu des hypothèses faites sur les variables aléatoires observées mais aussi des critères d'estimation choisis par le statisticien ou le sondeur : estimation sans biais, convergence asymptotique, efficacité, validité au second ordre des intervalles de confiance (au sens où les intervalles de confiance construits possèdent de meilleures propriétés que ceux construits avec des méthodes asymptotiques traditionnelles), etc.

Dans la première partie, nous décrivons rapidement dans le contexte du modèle d'échantillonnage les méthodes du bootstrap usuel (au sens d'EFRON [1979]-[1982]) et du bootstrap généralisé (au sens de MASON et NEWTON [1992]). Dans une seconde partie, nous adaptons ces idées au cas de sondages aléatoires simples : nous montrons comment il est possible, en choisissant bien le plan de rééchantillonnage, ou système de poids, d'obtenir un estimateur bootstrap de la variance sans biais, mais aussi des propriétés asymptotiques de validité au second ordre de la distribution bootstrap généralisée. Cette approche permet d'inclure un grand nombre de méthodes déjà proposées dans la littérature sur les sondages et les réplications

d'échantillons. Nous donnons une version générale de la méthode pour un plan de sondage à probabilités inégales. Des résultats asymptotiques sont obtenus dans le cadre du sondage poisonnien. Nous montrons en particulier qu'il est possible d'obtenir des résultats au second ordre en choisissant le moment d'ordre 3 des poids en fonction de l'échantillon observé et des probabilités d'inclusion. Nous abordons rapidement le problème de la généralisation de ces résultats à une fonctionnelle quelconque de l'échantillon.

Les méthodes proposées permettent, entre autre, de construire des intervalles de confiance plus précis que les intervalles de confiance basés sur l'approximation asymptotique : ceci est particulièrement intéressant pour des petits échantillons issus d'une population dissymétrique. Dans une dernière partie nous étudions ces méthodes par simulation dans le cas de l'estimation d'un ratio, puis nous donnons une application à la construction d'intervalles de confiance pour des moyennes, des ratios et des fractiles de consommations, estimés à partir des données des panels de ménages de Secodip.

2 Une généralisation du bootstrap

Le bootstrap d'EFRON [1979]

Soit $\mathbf{X}_n = (X_1, \ldots, X_n)$ n variables aléatoires i.i.d. de loi de probabilité P, supposée inconnue, on définit la mesure empirique par

$$P_n = n^{-1} \sum_{i=1}^{n} \delta_{X_i}$$

où les δ_{X_i} sont des masses de dirac en X_i. Conditionnellement à $(X_1, ..., X_n)$ une suite d'échantillons bootstrap de taille k_n, lorsque n varie, est un tableau triangulaire $\mathbf{X}_{k_n}^{(n)} = (X_1^{(n)}, \ldots, X_{k_n}^{(n)})$ de variables aléatoires i.i.d. de loi de probabilité P_n. A n fixé un échantillon bootstrap s'obtient simplement par **tirage équiprobable avec remise** d'un échantillon de taille k_n dans l'ensemble des observations. Considérant une statistique $T_n(\mathbf{X}_n)$ estimant un paramètre θ, il est généralement possible de donner sa distribution limite quand n tend vers l'infini, néanmoins cette distribution asymptotique ne rend pas toujours bien compte de la distribution à distance finie. Le principe du bootstrap est de dire que, à n fixé, puisque P_n est proche de P, résultat découlant des propriétés du processus empirique et du théorème de Kolmogorov, la distribution et les caractéristiques de $T_n(\mathbf{X}_n)$, sous la loi P des observations doivent être proches de la distribution et des caractéristiques de $T_{k_n}(\mathbf{X}_{k_n}^{(n)})$ sous la loi P_n , conditionnellement aux observations.

La validité de cette conjecture à été démontrée dans de très nombreux cas, en un sens asymptotique *i.e.* lorsque n tend vers ∞ (voir

BERTAIL ([1992]) chap. 1 pour une revue de la littérature et de plus amples références). Le résultat fondamental est que la distribution de $T_{k_n}(\mathbf{X}_{k_n}^{(n)})$ convenablement standardisée, dite distribution bootstrap, s'interprète comme un développement d'Edgeworth empirique (EDGEWORTH [1907], FELLER [1971]) et possède donc de meilleurs propriétés que la distribution asymptotique. L'utilisation des quantiles de la distribution bootstrap permet d'obtenir des intervalles de confiance pour le paramètre θ, possédant des propriétés au second ordre. En effet pour un seuil α visé, l'utilisation des quantiles de la loi asymptotique conduit à faire, à distance finie, une erreur sur le seuil de l'ordre de $n^{-1/2}$, cette erreur étant d'autant plus grande que les observations et la statitistique étudiée présentent des dissymétries, alors que l'erreur n'est plus que de l'ordre de n^{-1} avec la distribution bootstrap (HALL [1986]). D'un point de vue pratique, la distribution bootstrap (que l'on pourrait entièrement calculer si l'on avait le temps et la puissance informatique nécessaire, puisque P_n est connue) est estimée par une méthode de Monte-Carlo (d'où la terminologie « méthodes de calculs intensifs sur ordinateurs » (DIACONIS et EFRON [1983]). Ayant observé $(X_1, \ldots, X_n)$, on tire avec remise dans cet ensemble, B échantillons de taille k_n , on calcule la valeur de la statistique T_n, sur chacun de ces échantillons. La distribution empirique des B valeurs obtenues est une approximation (aussi fine que l'on veut, pourvu que B soit grand) de la distribution bootstrap, elle-même approximation de la distribution exacte.

Les travaux sur le bootstrap et la modélisation montrent qu'il convient parfois d'adapter la méthode aux modèles, de manière à pouvoir obtenir de bonnes propriétés (ceci est particulièrement important pour les sondages si l'on veut construire des estimateurs sans biais de la variance). De nombreuses adaptations du bootstrap ont été proposées dans la littérature pour tenir compte des spécificités de certains modèles statistiques. La plupart de ces méthodes sont des cas particuliers du bootstrap généralisé introduit par LO [1991] et MASON et NEWTON [1992], qui remet par ailleurs en cause l'intuition initiale du bootstrap d'EFRON, à savoir que son fondement est la proximité entre P et P_n.

Le bootstrap pondéré

Soit $W^{(n)} := \{W_i^{(n)} : 1 \leq i \leq n\}$ des variables aléatoires échangeables (*i.e.* dont la loi jointe est invariante par permutation, voir CHOW et TEICHER [1988]) conditionnellement à l'échantillon $(X_1, \ldots, X_n)$ et telles que

$$\sum_{i=1}^{n} W_i^{(n)} \; = \; k_n.$$

Ce système de poids peut aussi s'interpréter comme un plan de rééchantillonnage (voir EFRON [1979]). La probabilité bootstrap au sens de MASON-NEWTON est donnée par

$$P_{W,k_n}^{(n)} := k_n^{-1} \sum_{i=1}^{n} W_i^{(n)} \delta_{X_i}, \tag{1}$$

pondération des masses de dirac par des poids aléatoires.

Lorsque les $(W_i^{(n)})_{1\leq i\leq n}$ prennent des valeurs entières, (1) est en fait, l'analogue de la probabilité empirique de l'échantillon bootstrap,

$$P_{k_n}^{(n)} := k_n^{-1} \sum_{i=1}^{k_n} \delta_{X_i^{(n)}},$$

dans lequel X_i apparaitrait $W_i^{(n)}$ fois. Le bootstrap d'Efron consistant à effectuer un tirage avec remise dans les $(X_i)_{1\leq i\leq n}$ s'obtient en choisissant des poids $(M_i^{(n)})_{1\leq i\leq n}$ de loi multinomiale $Mult(k_n, 1/n)$; il n'est donc qu'un cas particulier de cette méthode.

Bootstrap généralisé d'une fonctionnelle

Si le paramètre θ est identifiable on peut généralement le mettre sous la forme d'une fonction dépendant de la vraie loi P :

$$\theta := T(P).$$

L'approche fonctionnelle consiste à s'intéresser à $T(P_n)$ la valeur empirique de la fonctionnelle en tant qu'estimateur de $T(P)$. On peut montrer que, sous des conditions de régularité sur $T(P)$ (continuité, différentiabilité pour une métrique adaptée), $T(P_n)$ est un estimateur consistant, asymptotiquement gaussien (HUBER [1981]). Dans ce cadre la statistique bootstrap associée est simplement $T(P_{k_n}^{(n)})$, la statistique bootstrap généralisée est $T(P_{W,k_n}^{(n)})$. Ainsi si $T(P) := E_P X = \int x \mathrm{dP}(x)$ est l'espérance, alors on a clairement :

$$T(P_n) = \int x \mathrm{dP}_n(x) = n^{-1} \sum_{i=1}^{n} X_i$$

$$T(P_{k_n}^{(n)}) = \int x \mathrm{dP}_{k_n}^{(n)}(x) = k_n^{-1} \sum_{i=1}^{k_n} X_i^{(n)} = k_n^{-1} \sum_{i=1}^{k_n} M_i^{(n)} X_i$$

$$T(P_{W,k_n}^{(n)}) = \int x \mathrm{dP}_{W,k_n}^{(n)}(x) = k_n^{-1} \sum_{i=1}^{n} W_i^{(n)} X_i$$

La statistique bootstrap est dans ce cas simplement une somme pondérée (les poids étant aléatoires) des observations, dont on étudie le comportement conditionnellement à $(X_1, \ldots, X_n)$, sous la loi du plan de rééchantillonnage. Là encore, d'un point de vue pratique, l'étude de la distribution se fait par du calcul de Monte-Carlo, en tirant B fois un système de poids, dans une loi donnée (la multinomiale pour le bootstrap d'Efron).

MASON et NEWTON [1992], EINMAHL et MASON [1992] ont montré la consistance du bootstrap généralisé de la moyenne, de la fonction de répartition, de la fonction de quantile et de certaines fonctionnelles lorsque les X_i sont à valeurs réelles. Dans une certaine mesure, le bootstrap généralisé fonctionne parce que le plan de rééchantillonnage imite bien les fluctuations du premier tirage. BARBE et BERTAIL [1995] ont montré dans

le cadre très général des fonctionnelles Fréchet différentiables pour une métrique indexée par une classe de fonctions que, suivant le critère choisi (à savoir la validité au second ordre, la bonne représentation des queues de distribution, la précision de la probabilité de couverture des intervalles de confiance ou un comportement identique à la vraie loi en terme de grande déviation), on devait choisir des poids spécifiques, fonctions des observations et de la statistique étudiée. En particulier, dans certain cas, l'erreur commise, sur le seuil initialement choisi, en utilisant la distribution bootstrap généralisé peut être encore diminuée jusqu'à l'ordre $O(n^{-3/2})$ au lieu de $O(n^{-1/2})$ pour l'approximation asymptotique.

2 Bootstrap généralisé d'un sondage

L'optique bootstrap généralisé apparaît, par sa souplesse, particulièrement adaptée au sondage. On notera d'ailleurs que, adoptant un critère de sans biais de l'estimateur bootstrap de la variance, MACCARTHY et SNOWDEN [1983] ainsi que CHAO et LO [1985] proposent des formes de bootstrap différentes du bootstrap d'Efron, que l'on peut inclure dans ce shéma plus général. Dans les paragraphes suivants nous étudions la validité du bootstrap généralisé d'abord dans le cas de sondages aléatoires simples puis dans le cas de sondages à probabilités inégales.

2.1. Bootstrap généralisé d'un sondage aléatoire simple sans remise

2.1.1. L'échec du bootstrap d'Efron

Considérons le cas d'un tirage sans remise d'un échantillon de taille n dans une population de taille N. On s'intéresse ici à l'estimation de la moyenne d'une variable Y. Dans tout ce qui suit on supposera pour simplifier les notations et les démonstrations que Y est à valeurs réelles mais tous les résultats se généralisent aisément à $\mathbb{R}^k,\ k > 1$.

On cherche donc à estimer

$$Y_N = N^{-1} \sum_{i=1}^{N} Y_i,$$

à partir d'un échantillon $(y_1, \ldots, y_n)$. Un estimateur sans biais de Y_N est donné par

$$\overline{y}_n = n^{-1} \sum_{i=1}^{n} y_i.$$

Le critère de sans biais du sondeur conduit à rejeter la procédure usuelle du bootstrap qui consisterait à tirer avec remise dans l'ensemble $(y_1, \ldots, y_n)$. En effet, on peut aisément montrer que si $(y_1^{(n)}, \ldots, y_n^{(n)})$ sont i.i.d. de loi

$$P_n \;=\; n^{-1} \sum_{i=1}^{n} \delta_{y_i}$$

alors la variance bootstrap de la moyenne bootstrap $\overline{y}_n^{(n)} = n^{-1} \sum_{i=1}^{n} y_i^{(n)}$ vaut

$$\mathrm{Var}(\overline{y}_n^{(n)} \mid P_n) := \; \mathrm{Var}_{P_n}\left(\bar{n}^1 \sum_{i=1}^{n} y_i^{(n)} \mid P_n \right) = n^{-2} \sum_{i=1}^{n} (y_i - \bar{y}_n)^2 = n^{-1} s_n^2$$

avec

$$s_n^2 = n^{-1} \sum_{i=1}^{n} (y_i - \bar{y}_n)^2$$

est un estimateur biaisé de la vraie variance de $\bar{y}_n$:

$$\mathrm{Var}(\bar{y}_n) = n^{-1}(1 - n/N)(1 - 1/N)^{-1} s_N^2, \text{ avec}$$

$$s_N^2 = N^{-1} \sum_{i=1}^{N} (Y_i - \bar{Y}_N)^2,$$

Rappelons qu'un estimateur sans biais de $\mathrm{Var}(\bar{y}_n)$ est donné par

$$\hat{V}(\bar{y}_n) = (1 - n/N)(n-1)^{-1} s_n^2. \tag{2}$$

Plusieurs solutions ont été proposées pour donner une version adaptée du bootstrap donnant directement un estimateur bootstrap sans biais de la variance. MACCARTHY et SNOWDEN [1983] proposent de choisir une taille de rééchantillonnage k_n différente de n. Lorsque $k_n = N/n$ est grand, ils préconisent l'utilisation de $k_n = n-1$. Ce résultat n'est pas étonnant dans la mesure où déjà dans le modèle d'échantillonnage usuel, la variance bootstrap avec $k_n = n$, égale à $n^{-2} \sum_{i=1}^{n} (X_i - X_n)^2$, (avec les notations de la partie précédente) est un estimateur biaisé à distance finie de $\mathrm{Var}(\overline{X}_n)$, mais est sans biais avec une taille de rééchantillonnage $k_n = n-1$. Lorsque $k = N/n$ est entier, la solution adoptée par CHAO et LO [1985] est de répliquer chacun des individus observés k fois de manière à obtenir un échantillon de taille N et de rééchantillonner sans remise dans cet échantillon fictif. Ceci revient à proposer un bootstrap de type généralisé dans lequel les poids auraient une loi multidimensionnelle hypergéométrique. On voit bien que ce plan de rééchantillonnage restitue les fluctuations du plan de sondage initial. Lorsque k n'est pas entier, CHAO et LO [1985] proposent une procédure de randomisation entre deux types d'échantillons, l'un où les individus sont répliqués $[k]$ fois et l'autre $[k] + 1$ fois. La méthode que nous proposons, incluant celle de CHAO et LO [1985] pour k entier, permet d'éviter cette randomisation et donne des résultats exacts à distance finie.

2.1.2. **Bootstrap généralisé et critère de sans biais**

Soit $(W_i^{(n)})_{1\leq i\leq n}$ un vecteur de variables aléatoires échangeables de loi $\mathcal{W}_n$ telles que

$$\sum_{i=1}^{n} W_i^{(n)} = k_n$$

ce qui induit entre autres par interchangeabilité des variables aléatoires $W_i^{(n)}$,

$$\mathrm{E}W_i^{(n)} = n^{-1}k_n$$

$$\mathrm{Cov}\big(W_1^{(n)}, W_2^{(n)}\big) = -\mathrm{Var}(W_1^{(n)})/(n-1) \tag{3}$$

On posera dans la suite $W_{i,n} := k_n^{-1}W_i^{(n)}$, de sorte que

$$\sum_{i=1}^{n} W_{i,n} = 1$$

L'estimateur Bootstrap généralisé de la moyenne est de la forme

$$\bar{y}_{w,n} = k_n^{-1}\sum_{i=1}^{n} W_i^{(n)} y_i = \sum_{i=1}^{n} W_{i,n} y_i,$$

de variance conditionnellement à $(y_1, \ldots y_n)$ (ou encore P_n)

$$\mathrm{Var}(\bar{y}_{w,n} \mid P_n) := \mathrm{Var}(k_n^{-1}\sum_{i=1}^{n}(W_i^{(n)} - n^{-1}k_n)(y_i - \bar{y}_n) \mid P_n) \tag{4}$$

$$= k_n^{-2}\big(\mathrm{Var}(W_i^{(n)}) - \mathrm{Cov}(W_1^{(n)}, W_2^{(n)})\big)\sum_{i=1}^{n}(y_i - \bar{y}_n)^2$$

$$= k_n^{-2}\,\mathrm{Var}(W_i^{(n)})(1 + (n-1)^{-1})\sum_{i=1}^{n}(y_i - y_n)^2.$$

L'égalité de l'estimateur bootstrap généralisé de la variance (4) avec l'estimateur sans biais (2), s'écrit donc

$$k_n^{-2}\,\mathrm{Var}(W_i^{(n)})(1 + (n-1)^{-1}) = n^{-1}(n-1)^{-1}(1 - N^{-1}n) \tag{5}$$

Si on choisit $k_n = n$ comme dans Chao et Lo [1985] alors la condition (5) s'interprète comme une contrainte sur la variance du plan de rééchantillonnage,

$$\mathrm{Var}(W_i^{(n)}) = 1 - n/N, \tag{6}$$

condition qui est asymptotiquement réalisée par le plan hypergéométrique multidimensionnel qu'ils proposent lorsque n et N sont grands, puisque, dans ce cas, la variance du plan est

$$\text{Var}(W_i^{(n)}) = (1 - n/N)(1 - n^{-1})/(1 - N^{-1}).$$

Remarquons que le rapport $(1 - n^{-1})/(1 - N^{-1})$ est inférieur à 1 pour des valeurs de $n \leq N$. Il s'ensuit que l'utilisation du plan proposé par CHAO et LO [1985] a tendance à sous-estimer la variance de la statistique, ce qu'observent empiriquement MACCARTHY et SNOWDEN [1983] par simulation.

Inversement si on choisit un plan de rééchantillonnage multinomial alors $\text{Var}(W_i^{(n)}) = k_n n^{-1}(1 - n^{-1})$ et il faut choisir une taille d'échantillonnage

$$k_n = (n - 1)/(1 - n/N).$$

Si $n/N \to 0$ quand N et $n \to \infty$ (voir infra pour plus de précision sur ces notions asymptotiques), on retrouve la proposition de MACCARTHY et SNOWDEN [1983]. Il est clair avec cette formulation que de nombreux choix sont possibles pour le plan de rééchantillonnage et l'on peut très bien en exhiber plusieurs donnant une correction de population finie **exacte** quelque soit la valeur du rapport N/n. Cette méthode permet en outre d'éviter la procédure de randomisation sur deux échantillons de tailles respectives $n[N/n]$ et $n([N/n]+1)$, suggérée par CHAO et LO [1985] lorsque le rapport N/n n'est pas entier.

Il suffit en effet de générer des variables aléatoires $(W_{i,n})_{1 \leq i \leq n}$ de somme 1 et de variance fixée

$$\text{Var}(W_{i,n}) = n^{-2}(1 - n/N) \tag{7}$$

Ceci peut être fait en utilisant les clones bootstrap bayesien de LO [1991] *i.e.* en choisissant, pour $1 \leq i \leq n$

$$W_{i,n} = Z_{i,n} / \sum_{i=1}^{n} Z_{i,n} \ , \tag{8}$$

où les $(Z_{i,n})_{1 \leq i \leq n}$ sont des variables aléatoires i.i.d. indépendantes des y_i, dont on fixe le moment d'ordre 2 pour obtenir une variance adéquate pour $W_{i,n}$ (voir aussi BARBE et BERTAIL [1995], I.5).

2.1.3. **Validité asymptotique de la méthode**

Quelques remarques préliminaires sur les convergences utilisées s'imposent car les notions asymptotiques usuelles ne peuvent pas être utilisées directement dans le cadre des sondages. En effet, la taille totale de la population N est fixée et il n'est donc pas possible de faire tendre la taille n de la population observée vers l'infini. Lorsque N est très grand et que $\frac{n}{N}$ n'est pas trop petit, il est possible de plonger le sondage étudié dans une suite de sondages (indexée par un paramètre $\alpha \in \mathbb{N}$) effectués dans

une suite de population $(Y_1, ..., Y_{N_\alpha})$ de taille croissante N_α, $N_\alpha \underset{\alpha\to\infty}{\longrightarrow} \infty$. On note alors $\{(y_1, y_2, ..., y_{n_\alpha}),\ n_\alpha \leq N_\alpha\ \}_{\alpha\in\mathbb{N}}$ le tableau triangulaire des échantillons observés. Il est clair que lorsque N_α et n_α augmentent, la valeur du paramètre d'intérêt, ici la moyenne sur la population, $\bar{Y}_{N_\alpha}$, change. Par ailleurs, le plan de sondage $\mathcal{P}_{N_\alpha}$ change aussi, de sorte que les probabilités d'inclusion sont aussi indexées par α. On dit alors que l'estimateur $\overline{y}_{n_\alpha}$ de $\bar{Y}_{N_\alpha}$ est asymptotiquement sans biais si

$$\lim_{\alpha\to\infty}(E_{\mathcal{P}_{N_\alpha}}\overline{y}_{n_\alpha} - \bar{Y}_{N_\alpha}) = 0,$$

est asymptotiquement convergent si

$$\lim_{\alpha\to\infty} P_{\mathcal{P}_{N_\alpha}}\{|\,\overline{y}_{n_\alpha} - \bar{Y}_{N_\alpha}| > \epsilon\} = 0$$

(où $E_{\mathcal{P}_{N_\alpha}}$ et $P_{\mathcal{P}_{N_\alpha}}$ désignent respectivement l'espérance et la probabilité sous la loi du plan de sondage α), est asymptotiquement gaussien de vitesse de convergence $\tau_{(n_\alpha,N_\alpha)}$si

$$\lim_{\alpha\to\infty} P_{\mathcal{P}_{N_\alpha}}\{\tau_{(n_\alpha,N_\alpha)}|\,\overline{y}_{n_\alpha} - \bar{Y}_{N_\alpha}| \leq x\} - \Phi(x) = 0.$$

Pour alléger les notations et toutes ambiguités étant écartées, la référence au plan de sondage $\mathcal{P}_{N_\alpha}$ sera supprimée dans la suite. En terme de convergence en loi, nous noterons ce résultat sous la forme abusive traditionnellement utilisée:

$$\tau_{(n_\alpha,N_\alpha)}(\bar{y}_{n_\alpha} - \bar{Y}_{N_\alpha}) \xrightarrow{\mathcal{L}} N(0,1).$$

L'obtention de tels résultats sur la moyenne dépend en grande partie du plan de sondage choisi, mais aussi du comportement respectif de n_α et de N_α.

On supposera, pour le plan de sondage considéré (n_α pouvant être aléatoire), que pour $1 > \eta > 0$ et $1 > \eta' > 0$,

$$\eta < \frac{n_\alpha}{N_\alpha} < 1 - \eta',$$

condition qui premunit contre les sondages de taille trop petite ou exhaustifs. Nous renvoyons à SEN [1988] pour un survey des principaux résultats obtenus dans un tel contexte.

On considèrera donc désormais une suite de tirages sans remise indexés par un paramètre α; $\{(W_i^{(n_\alpha)})_{1\leq i\leq n_\alpha},\ \alpha\in\mathbb{N}\}$ désigne une suite triangulaire de variables échangeables de loi $\mathcal{W}_{n_\alpha}$. On peut alors obtenir des résultats de validité asymptotique au second ordre du bootstrap généralisé. On notera $\xrightarrow{\mathcal{L}|P_{n_\alpha}}$ la convergence en loi par rapport à la loi du vecteur des variables échangeables $\mathcal{W}_{n_\alpha}$, conditionnellement à l'échantillon initial observé. Si X_{W,n_α} est une statistique dépendant des poids $(W_i^{(n_\alpha)})_{1\leq i\leq n_\alpha}$ et de l'échantillon observé (et donc de $P_{n_\alpha} = n_\alpha^{-1}\sum_{i=1}^{n_\alpha}\delta_{y_i}$), la notation $X_{W,n_\alpha} \xrightarrow{\mathcal{L}|P_{n_\alpha}} N(0,1)$ signifie donc ici que

$$\lim_{\alpha\to\infty} P_{\mathcal{W}_{n_\alpha}}\{X_{W,n_\alpha} \leq x|P_{n_\alpha}\} - \Phi(x) = 0.$$

On note dans ce qui suit

$$K^{\epsilon}_{p,N_\alpha} := N_\alpha^{-1} \sum_{i=1}^{N_\alpha} \mid Y_i - \bar{Y}_{N_\alpha}|^{p+\epsilon}, \quad \epsilon > 0$$

$$K_{p,N_\alpha} := K^0_{p,N_\alpha}$$

respectivement les moments absolus centrés d'ordre $p+\epsilon$ et p des Y_i, $1 \leq i \leq N_\alpha$, et enfin

$$\beta_{w,n_\alpha} := E(n_\alpha W_{i,n_\alpha} - 1)^3 / \text{ Var}(n_\alpha W_{i,n_\alpha})^{3/2}$$

le coefficient d'asymétrie de $n_\alpha W_{i,n_\alpha}$.

Le théorème suivant montre que si l'on choisit convenablement le coefficient d'asymétrie des poids, on peut obtenir la validité au second ordre du bootstrap pondéré, dans le cadre du tirage sans remise. En général, dans le cadre du modèle d'échantillonnage, la valeur de β_{w,n_α} qui donne la correction au second ordre est

$$\beta_{w,n_\alpha} = \sigma_{w,n_\alpha} = 1 + o(1)$$

condition qui est réalisée par exemple par des poids multinomiaux, *i.e.* par le bootstrap d'Efron. Dans le cas du rééchantillonnage d'un sondage nous avons déjà vu avec la variance qu'il fallait tenir compte d'une correction de population finie. La proposition suivante montre que le coefficient d'asymétrie β_{w,n_α} qui permet d'obtenir la correction au second ordre dépend aussi du rapport $N_\alpha^{-1} n_\alpha$.

PROPOSITIONS 1 : Soit $(W_{i,n_\alpha})_{1 \leq i \leq n_\alpha}$ une suite de v.a. échangeables définies par (2.7) et satisfaisant (2.6), si $K^{\epsilon}_{4,N_\alpha}$ est une suite bornée quand $N_\alpha \to \infty$, pour $\epsilon > 0$ alors on a :

$$\text{Var}(\bar{y}_{n_\alpha})^{-1/2}(\bar{y}_{n_\alpha} - \bar{Y}_{N_\alpha}) \xrightarrow{L} N(0,1) \tag{9}$$

$$\text{Var}(\bar{y}_{w,n_\alpha}|P_{n_\alpha})^{-1/2}(\bar{y}_{w,n_\alpha} - \bar{y}_{n_\alpha}) \xrightarrow{L|P_{n_\alpha}} N(0,1) \quad \text{p.s.} \tag{10}$$

Si de plus $K^{\epsilon}_{6,N_\alpha}$ est bornée quand $N_\alpha \to \infty$ et si P_{N_α} vérifie la condition de Cramer

$$\lim_{N_\alpha \to \infty} \lim_{t \to \infty} \mid E_{P_{N_\alpha}} e^{\mathrm{itY}} \mid < \eta < 1, \ \eta > 0, \tag{11}$$

si W_{i,n_α} est telle que

$$\lim_{n_\alpha \to \infty} \lim_{t \to \infty} \mid E e^{itW_{i,n_\alpha}} \mid < \eta' < 1, \quad \eta' > 0, \tag{12}$$

et si

$$\beta_{w,n_\alpha} \coloneqq (1 - 2N_\alpha^{-1} n_\alpha)/(1 - N_\alpha^{-1} n_\alpha)^{1/2} + o(1), \text{ quand } N_\alpha \to \infty, \tag{13}$$

alors on a

$$\begin{aligned} \| \, \mathrm{P}(\mathrm{Var}(\bar{y}_{w,n_\alpha} \mid P_{n_\alpha})^{-1/2} (\bar{y}_{w,n_\alpha} - \bar{y}_{n_\alpha}) \leq x \mid P_{n_\alpha}) \\ - \mathrm{P}(\mathrm{Var}(\bar{y}_{n_\alpha})^{-1/2} (\bar{y}_{n_\alpha} - \bar{Y}_{N_\alpha}) \leq x) \|_\infty = O(n_\alpha^{-1}) \end{aligned} \tag{14}$$

Démonstration: Voir annexe 1.

Ce résultat montre donc que la distribution bootstrap pondéré donne, lorsqu'on choisit des poids adéquats, un estimateur sans biais de la variance et une meilleure approximation de la vraie distribution (et donc de meilleurs intervalles de confiance) que l'approximation asymptotique.

La condition (11) est une condition de type Cramer sur la distribution empirique de la population. Elle signifie entre autre que la variable Y ne doit pas être distribuée sur un treillis (en d'autres termes que les $(Y_i)_{1 \leq i \leq N_\alpha}$ ne doivent pas être équidistants). La condition de Cramer (12) portant sur la distribution des poids est réalisée, par exemple, si les variables aléatoires $Z_{i,n_\alpha},\ 1 \leq i \leq n_\alpha$ définissant les poids en (8) ont une distribution continue.

2.1.4. **Généralisation aux fonctionnelles statistiques**

Utilisant les résultats de BARBE et BERTAIL [1995], il est aisé de montrer que, sous quelques conditions de régularité, ces résultats se généralisent à toute statistique associée à une fonctionnelle Fréchet différentiable pour une métrique adaptée. En effet, si on s'intéresse à la valeur d'une fonctionnelle T sur la population initiale *i.e.* au paramètre $T(P_{Y,N})$, où

$$P_{Y,N} := N^{-1} \sum_{i=1}^{N} \delta_{Y_i}$$

et à son estimateur naturel $T(P_{y,n})$ avec

$$P_{y,n} \ := \ n^{-1} \sum_{i=1}^{n} \delta_{y_i},$$

on peut introduire la version bootstrap généralisé de ces quantités en définissant

$$P_{w,y,n} := \sum_{i=1}^{n} W_{i,n} \delta_{y_i}, \quad \sum_{i=1}^{n} W_{i,n} = 1$$

On peut donc chercher à approximer la distribution de $T(P_{y,n}) - T(P_{Y,N})$ par la distribution de $T(P_{w,y,n}) - T(P_{y,n})$ conditionnellement aux $y_i, 1 \leq i \leq n$.

Considérant une suite de tirages indexés par α, on suppose que

$$P_{Y,N_\alpha} \text{ converge vers une probabilité P quand } N_\alpha \to \infty \tag{15}$$

ce qui signifie que, asymptotiquement, la distribution de la variable Y sur la population de taille N_α peut être approximée par une distribution fixe.

Si T est deux fois Fréchet différentiable en P pour une métrique d, alors P_{Y,N_α}, P_{y,n_α} et P_{w,y,n_α} étant des lois de probabilité, on peut, en suivant SERFLING [1981] chap. 6, à partir d'un développement de Taylor donner les approximations linéaires suivantes :

$$T(P_{y,n_\alpha}) - T(P_{Y,N_\alpha}) = n_\alpha^{-1} \sum_{i=1}^{n_\alpha} T^{(1)}(y_i, P) - N_\alpha^{-1} \sum_{i=1}^{N_\alpha} T^{(1)}(Y_i, P) + r_{n_\alpha}$$

$$T(P_{w,y,n_\alpha}) - T(P_{y,n_\alpha}) = \sum_{i=1}^{n_\alpha} (W_{i,n_\alpha} - n_\alpha^{-1}) T^{(1)}(y_i, P) + r_{w,n_\alpha},$$

où $T^{(1)}(., P)$ est la fonction d'influence en P de la fonctionnelle T (voir HUBER [1981]) et où r_{n_α} et r_{w,n_α} sont des restes de la forme :

$$r_{n_\alpha} = d(P_{y,n_\alpha}, P)\xi(d(P_{y,n_\alpha}, P)) + d(P_{Y,N_\alpha}, P)\xi(d(P_{Y,N_\alpha}, P))$$

$$r_{w,n_\alpha} = d(P_{y,n_\alpha}, P)\xi(d(P_{y,n_\alpha}, P)) + d(P_{w,y,n_\alpha}, P)\xi(d(P_{w,y,n_\alpha}, P)).$$

où ξ est une fonction continue nulle en 0.

Sous quelques hypothèses de régularité sur la fonctionnelle, la distance d et l'existence de moment d'ordre $2 + \epsilon$, pour $\epsilon > 0$, de $T^{(1)}(Y, P)$ sous la loi P (voir BARBE et BERTAIL [1995] pour des conditions exactes), il est possible de contrôler le comportement des restes, pour $n_\alpha \to \infty$ de sorte que le comportement asymptotique de $T(P_{y,n_\alpha}) - T(P_{Y,N_\alpha})$ et de $T(P_{w,y,n_\alpha}) - T(P_{y,n_\alpha})$ est donné par celui de la partie linéaire, à laquelle s'appliquent les résultats de la proposition précédente. Un développement jusqu'à l'ordre 2, faisant intervenir la fonction d'influence bivariée permet d'aboutir à un résultat de type (14), sous une condition de Cramer portant sur $T^{(1)}(X, P)$, sous la loi P. Ces résultats s'appliquent en particulier à une large catégorie de M et L estimateurs. Notons cependant, pour les applications qui vont suivre, que le résultat de validité au second ordre n'est pas valable si $T(P) = \inf(x,\ P(X \leq x) \geq \alpha)$ est le quantile d'ordre α de la distribution. En effet la fonction d'influence de $T(P)$ est une variable de treillis qui ne satisfait pas la condition de Cramer, ce qui explique le comportement atypique au second ordre des estimateurs associés (voir par exemple FALK et REISS [1989]). La validité asymptotique de la méthode pour les fractiles découle cependant directement de la différentiabilité de la fonctionnelle associée, pour la métrique en p-variation (voir DUDLEY [1994]). Par ailleurs BERTAIL [1997] a récemment proposé dans un cadre i.i.d. et/ou $\alpha-$mélangeant une méthode de basée sur l'extrapolation de Richardson de la distribution bootstrap sans remise, qui permet d'obtenir automatiquement des intervalles de confiance correct au second ordre sous des conditions minimales. Elle s'applique en particulier au fractile : l'adaptation de ce résultat aux sondages semble ne pas poser de problème et fera l'objet de travaux ultérieurs.

2.2. **Bootstrap pondéré et sondage à probabilités inégales**

2.2.1. **Le cadre d'analyse**

On considère un plan de sondage à probabilités inégales dans une population de taille N. On note $(\pi_i)_{1\leq i\leq N}$ les probabilités d'inclusion d'ordre 1, et $(\pi_{i,j})_{1\leq i\leq j\leq \leq n}$ les probabilités d'inclusion d'ordre 2. On suppose le plan de taille n aléatoire (éventuellement fixe) et on s'intéresse au bootstrap généralisé de l'estimateur de Horvitz-Thompson d'une moyenne sur une variable Y. On note $(Y_1, Y_2, \ldots, Y_N)$ les valeurs de la variable Y sur la population totale et $(y_1, \ldots, y_n)$ les valeurs observées sur l'échantillon tiré, noté s, $(\epsilon_1, \ldots, \epsilon_N)$ les variables indicatrices valant 1 ou 0 selon que l'individu caractérisé par $i \in \{1, \ldots, N\}$ est tiré ou non. On rappelle que, sous la loi du plan de sondage, on a par définition

$$E\epsilon_i = \pi_i, \quad E\epsilon_i\epsilon_j = \pi_{i,j}.$$

Il est encore possible, et souhaitable dans une perspective de généralisation, de conserver une optique fonctionnelle: la fonctionnelle d'intérêt $T(P)$ est supposée définie sur un espace de mesure contenant les masses de dirac, à valeur scalaire. La distribution de la variable Y sur la population totale est

$$P_{Y,N} = N^{-1} \sum_{i=1}^{N} \delta_{Y_i}$$

son estimateur de Horvitz-Thompson

$$P_{y,n} = N^{-1} \sum_{i=1}^{N} \pi_i^{-1} \epsilon_i \delta_{Y_i} = N^{-1} \sum_{j=1}^{n} \pi_i^{-1} \delta_{y_j}$$

On notera que de façon générale $\mathbf{P_{y,n}}$ **n'est pas une probabilité** (sauf par exemple pour un tirage équiprobable sans remise ou encore un tirage en grappes de même taille). Ceci permet de mieux comprendre certaines mauvaises propriétés de cet estimateur (voir GOURIEROUX [1987], p. 60). Par exemple pour $T(P) = \int dP$, la masse totale de la mesure $P_{y,n}$ est

$$N^{-1} \sum_{i=1}^{N} \pi_i^{-1} \epsilon_i,$$

qui est elle-même une variable aléatoire.

La difficulté du bootstrap réside ici dans le fait que l'aléa provient des ϵ_i qui ne sont pas i.i.d. Dans ce type de problème, il est bien connu que le bootstrap usuel ne fonctionne pas, car le rééchantillonnage conduit à détruire l'hétéroscédasticité ou les corrélations entre les variables de l'échantillon. Certaines adaptations, proposées dans ces cas (ZHENG et TU [1988], KUNSCH [1989]) peuvent s'interpréter comme des formes de bootstrap

généralisé. Une solution naturelle proposée par DEVILLE [1987], lorsque les π_i^{-1} sont entiers et le sondage à taille fixe n, est, conditionnellement à $(y_1, \ldots, y_n)$, de dupliquer π_i^{-1} fois chaque y_i puis de constituer des échantillons bootstrap en utilisant le plan de sondage initial à une modification près des probabilités d'inclusion d'ordre 2 maintenant égales à $\pi_{i,j}^{(n)}$ de manière à obtenir un échantillon bootstrap de taille n. La procédure fonctionne en terme d'estimation de la variance, si les $\pi_{i,j}^{(n)}$ et $\pi_{i,j}$ sont proches de $\pi_i \pi_j$, *i.e.* si on a approximativement indépendance des tirages. Ceci le conduit à introduire des algorithmes de tirages qui sont « bootstrappables » (par exemple l'algorithme de Sunter) d'autres qui ne le sont pas (algorithme de tirage systématique sur un fichier trié par π_i croissant). Nous proposons dans le paragraphe suivant une forme de bootstrap généralisé pour des sondages de taille aléatoire, sous des condition de « bootstrappabilité » plus faibles puis nous en étudions les propriétés asymptotiques pour un sondage poissonnien.

2.2.2. **Bootstrap généralisé du sondage**

Soient

(16) $(W_{i,N})_{1 \leq i \leq N}$ *un ensemble de v.a. de même espérance*

$$EW_{i,N} = N^{-1} \tag{17}$$

On définit **la mesure bootstrap généralisé du sondage** par

$$P_{w,y,n} = \sum_{i=1}^{N} W_{i,N} \pi_i^{-1} \epsilon_i \delta_{Y_i} = \sum_{i \in s} W_{i,N} \pi_i^{-1} \delta_{Y_i},$$

L'intérêt d'introduire des poids $(W_{i,N})_{1 \leq i \leq N}$ sans aucune normalisation est de laisser des degrés de liberté supplémentaires, en particulier sur la masse $\sum_{i \in s} W_{i,N} \pi_i^{-1}$ qui peut être aléatoire, mais aussi sur la covariance des poids (voir BARBE et BERTAIL [1995] annexe 3, pour les contraintes induites sur les moments des poids par la normalisation à 1). En espérance, conditionnellement à l'échantillon s observé, on a:

$$E(P_{w,y,n} \mid s) = N^{-1} \sum_{i=1}^{N} \pi_i^{-1} \epsilon_i \delta_{Y_i} = P_{y,n}$$

Étudions alors les conditions sur le plan de rééchantillonnage pour obtenir un estimateur bootstrap sans biais de la variance d'une moyenne $T(P) = E_P \mathrm{Y}$. $T(P_{y,n})$ est un estimateur sans biais de $T(P_{Y,N})$. On a sous les conditions précédentes :

$$E_W(T(P_{w,y,n}) \mid s) = T(P_{y,n})$$

en terme de variance,

$$\text{Var}(T(P_{w,y,n}) \mid s) = \text{Var}\{\sum_{i\in s}(W_{i,N} - N^{-1})\pi_i^{-1}Y_i\}$$

$$= \sum_{i\in s} \text{Var}(W_{i,N})\pi_i^{-2}Y_i^2 + \sum_{i\in s}\sum_{j\in s} \text{Cov}(W_{i,N}, W_{j,N})\pi_i^{-1}\pi_j^{-1}Y_iY_j$$

Un estimateur sans biais de la variance de $T(P_{y,n})$ étant donné par

$$\hat{V}_n = N^{-2}\{\sum_{i\in s}(1-\pi_i)\pi_i^{-2}Y_i^2$$

$$+\sum_{i\in s}\sum_{j\in s}(\pi_{i,j} - \pi_i\pi_j)\pi_{i,j}^{-1}\pi_i^{-1}\pi_j^{-1}Y_iY_j\},$$

un choix évident s'impose pour les variances et covariances des poids :

$$\text{Var}(NW_{i,N}) = (1-\pi_i) \tag{18}$$

$$\text{Cov}(NW_{i,N}, NW_{j,N}) = (1 - \pi_i\pi_j/\pi_{i,j}). \tag{19}$$

Lorsque les tirages sont indépendants alors $\pi_i\pi_j = \pi_{i,j}$ et on doit avoir

$$Cov(W_{i,N}, W_{j,N}) = 0$$

i.e. non corrélation des variables du plan de rééchantillonnage. Si le tirage est équiprobable sans remise on retrouve le plan défini dans la partie précédente puisque dans ce cas, $\pi_i = n/N$ et $\pi_{ij} = n(n-1)/(N(N-1))$ redonne (3) et (6).

De façon générale, pour qu'un plan satisfaisant (18), (19) existe, il faut d'après Cauchy-Schwarz que

$$\forall(i,j) \in \{1,..N\}^2, \quad (1 - \pi_i\pi_j/\pi_{i,j})^2 \leq (1-\pi_i)(1-\pi_j), \tag{20}$$

condition de « bootstrappabilité » du plan de sondage. Cette condition laisse cependant plus de liberté que celle imposée par DEVILLE [1987].

DEVROYE [1986] donne différents algorithmes permettant de générer des variables aléatoires ayant une matrice de variance-covariance ou ses premiers moments fixés: la génération de tels poids ne pose donc pas de problèmes algorithmiques. D'un point de vue pratique et dans tout ce qui suit nous supposerons que les poids sont construits de la manière suivante :
soient

$$Z = (Z_{i,N})_{1\leq i\leq N}, \quad N \text{ v.a. indépendantes i.i.d. d'espérance } 0 \tag{21}$$

et de variance $\sigma_z = 1$, de distribution F_Z, tel que $E_{F_z}Z^4 < \infty$.

Soit

$$\Omega = [\omega_{i,j}]_{\substack{1\le i\le n\\ 1\le j\le n}}, \quad \omega_{i,i} = (1-\pi_i) \quad \text{et} \quad \omega_{\mathbf{i,j}} = (1-\pi_{\mathbf{i}}\pi_{\mathbf{j}}/\pi_{\mathbf{i,j}})$$

Soit H la « racine carrée » de Ω, *i.e.* telle que $H^2 = \Omega$. On pose

$$W = (W_{i,N})_{1\le i\le N} = N^{-1}(I + HZ), \tag{22}$$

où I est un vecteur de 1 de taille N, alors $(W_{i,N})_{1\le i\le N}$ vérifie (16), (17), (18) et (19).

2.2.3. **Convergence asymptotique de la méthode pour un sondage poissonnien**

On se place dans le même cadre asymptotique qu'en 2.1.3. ROSEN [1972] a démontré la convergence asymptotique de l'estimateur de Horvitz-Thompson et obtenu un théorème central limite, pour des plans de sondage réjectifs: voir aussi SEN [1988]. On a sous des conditions analogues pour une suite de tirages sans remise indexés par α, le résultat de validité asymptotique du bootstrap pondéré d'un sondage à probabilité inégale suivant :

PROPOSITION 2 : Lorsque $T(P) = E_P Y$, sous les hypothèses de ROSEN [1972] si le plan de sondage vérifie (20) avec

$$N_\alpha \hat{V}_{n_\alpha} \to S^2, \quad 0 < S^2 < +\infty \tag{23}$$

$$K^c_{\mathbf{4},N_\alpha} \textit{ est une suite bornée quand } N_\alpha \to \infty, \textit{ pour } \epsilon > 0, \tag{24}$$

si de plus les v.a. $(W_{i,N_\alpha})_{1\le i\le N_\alpha}$ sont définies par (21) et (22) alors

$$\mathrm{Var}(T(P_{y,n_\alpha}))^{-1/2}(T(P_{y,n_\alpha}) - T(P_{Y,n_\alpha})) \xrightarrow{L} N(0,1) \tag{25}$$

et

$$\mathrm{Var}(T(P_{w,y,n_\alpha})|P_{y,n_\alpha})^{-1/2}(T(P_{y,w,n_\alpha}) - T(P_{y,n_\alpha})) \xrightarrow{L|P_{n_\alpha}} N(0,1) \tag{26}$$

Démonstration: voir annexe 2.

On notera que si le vecteur Z est gaussien alors le résultat (26) est trivialement vérifié puisque par construction des poids on a directement

$$T(P_{w,y,n_\alpha}) - T(P_{y,n_\alpha}) \rightsquigarrow N(0, \hat{V}_{n_\alpha})$$

Cependant ce choix n'est pas judicieux si l'on désire obtenir des résultats de validité au second ordre. Le résultat de la proposition 3 montre que, pour la classe de poids considérée, le choix adéquat du moment d'ordre 3 dépend de l'échantillon observé.

Les résultats au second ordre semblent plus difficiles à obtenir dans le cadre général. En effet l'obtention d'un développement d'Edgeworth pour

$T(P_{y,w,n_\alpha}) - T(P_{y,n_\alpha})$ pose déja de sérieux problèmes à cause de la nature non i.i.d. des v.a. ϵ_i. De façon à donner néanmoins une indication sur le choix des moments d'ordre 3 des poids et pour de nombreuses applications, nous supposons dans la proposition suivante que les tirages sont indépendants de sorte que l'on a $\pi_{i,j} = \pi_i \pi_j$.

PROPOSITION 3 : Sous les hypothèses de la proposition 2, si de plus les tirages sont indépendants et que l'on a pour $\eta > 0$ et $\eta' > 0$

$$\lim_{N_\alpha \to \infty} \lim_{t \to \infty} \mid E_{P_{Y,N_\alpha}} e^{\mathrm{itY}} \mid < \eta < 1, \tag{27}$$

$$\lim_{N_\alpha \to \infty} \lim_{t \to \infty} \mid E e^{itW_{i,N_\alpha}} \mid < \eta' < 1, \tag{28}$$

$$K^{\epsilon}_{6,N_\alpha} \; est\ une\ suite\ bornée\ quand\ N_\alpha \to \infty,\ pour\ \epsilon > 0 \tag{29}$$

et si le coefficient d'asymétrie des poids est choisi tel que

$$\begin{aligned} \beta_{Z,n_\alpha} :&= EZ^3_{i,n_\alpha} / (EZ^2_{i,n_\alpha})^{3/2} \\ &= \{\sum_{i \in s} (1-\pi_i)(1-2\pi_i) Y_i^3 / \pi_i^2\} \\ &\quad / \{\sum_{i \in s} (1-\pi_i)^{3/2} Y_i^2 / \pi_i^3\} + o(1) \end{aligned} \tag{30}$$

alors

$$\begin{aligned} &\| \text{Prob}\{\text{Var}(T(P_{y,n_\alpha}))^{-1/2} (T(P_{y,n_\alpha}) - T(P_{Y,n_\alpha})) < x\} \\ &\quad - \text{Prob}(\text{Var}(T(P_{w,y,n_\alpha}) | P_{y,n_\alpha})^{-1/2} (T(P_{y,w,n_\alpha}) \\ &\quad - T(P_{y,n_\alpha})) < x \mid P_{y,n_\alpha}) \| \\ &\qquad = o(N_\alpha^{-1/2} K_{3,Y,N_\alpha}) \end{aligned} \tag{31}$$

avec

$$\begin{aligned} K_{3,Y,N_\alpha} = &\left\{ N_\alpha^{-1} \sum_{i=1}^{N_\alpha} \pi_i^{-1} (1-\pi_i) Y_i^2 \right\}^{-3/2} \\ &\times \left\{ N_\alpha^{-1} \sum_{i=1}^{N_\alpha} \pi_i^{-2} (1-\pi_i)(1-2\pi_i) Y_i^3 \right\} \end{aligned}$$

Démonstration: voir annexe 3.

(27) et (28) sont les conditions de Cramer usuelles qui permettent d'obtenir les développements d'Edgeworth jusqu'à l'ordre 1 des quantités étudiées. La forme de β_{Z,n_α} donnée en (30) permet de faire coïncider les deux développements jusqu'à l'ordre $o(n_\alpha^{-1/2})$. On notera que si le tirage est à probabilités égales alors on doit avoir

$$\beta_{Z,n_\alpha} = (1-2\pi_i)/(1-\pi_i)^{1/2} = (1-2n/N)/(1-n/N)^{1/2}$$

On retrouve la condition (13) obtenue pour un sondage aléatoire simple sans remise.

On notera par ailleurs que K_{3,Y,N_α} n'est pas forcément bornée, ce qui explique la forme du reste dans le développement d'Edgeworth; par exemple dans le cas du sondage aléatoire simple avec remise

$$K_{3,Y,N_\alpha} = (N_\alpha/n_\alpha)^{1/2}\left\{N_\alpha^{-1}\sum_{i=1}^{N_\alpha} Y_i^2\right\}^{-3/2}\left\{N_\alpha^{-1}\sum_{i=1}^{N_\alpha} Y_i^3\right\} = O((N_\alpha/n_\alpha)^{1/2})$$

le d.e. dans ce cas est valide jusqu'à l'ordre $o(n_\alpha^{-1/2})$.

La généralisation de (31) à des fonctions de la moyenne et par extension à des fonctions de moments ne pose pas de problèmes si les fonctions en jeu sont suffisamment continuement différentiables. La généralisation directe de ces résultats à des fonctionnelles autres, qui présenterait un réel intérêt pratique, pose néanmoins de sérieux problèmes méthodologiques.

2.2.4. **Le problème de la généralisation à des fonctionnelles statistiques**

La difficulté d'une généralisation vient de ce que les estimateurs de Horvitz-Thompson de la distribution de probabilité sont des **mesures et non des probabilités** et que, de ce fait, **les théorèmes de décomposition linéaire et quadratique des fonctionnelles sur la base des gradients** (fonction d'influence au premier ordre et au second ordre) **ne sont plus valides** (voir BARBE et BERTAIL [1995], remarque 3.2); une solution possible est de renormaliser les estimateurs de Horvitz Thompson en utilisant un estimateur par différence [1].

Pour cela, on peut considérer une variable auxilliaire X que l'on observe sur toute la population (par exemple la constante C), et on définit l'estimateur par différence de la distribution $P_{Y,N}$ des Y_i par

$$P_{y,n}^X = N^{-1}\sum_{i=1}^N \delta_{X_i} + N^{-1}\sum_{i=1}^N \pi_i^{-1}\epsilon_i(\delta_{Y_i} - \delta_{X_i})$$

$$= P_{X,N} + N^{-1}\sum_{i=1}^N \pi_i^{-1}\epsilon_i(\delta_{Y_i} - \delta_{X_i})$$

En particulier, si X est la variable aléatoire constante C, ceci revient à donner un poids (aléatoire) $1 - N^{-1}\sum_{i=1}^N \pi_i^{-1}\epsilon_i$ à un point arbitraire C de manière à renormaliser la mesure $P_{y,n}$ pour lui donner une masse totale égale à 1.

1. Une autre solution serait de renormaliser à 1 les inverses des probabilités d'inclusion intervenant dans l'estimateur de Horvitz-Thompson ce qui revient en fait à introduire un estimateur par le ratio (avec une variable de contrôle constante) de la distribution $P_{Y,N}$. Il convient alors de renormaliser les $W_{i,N}$ de sorte que $\sum_{i=1}^N W_{i,N}\pi_i^{-1}\epsilon_i = 1$. L'étude des propriétés asymptotiques de ces estimateurs semble néanmoins plus difficile, en particulier si la taille n du sondage n'est pas fixe.

La mesure bootstrap généralisé (signée de masse 1) du sondage, associée à la variable auxilliaire X est définie par

$$P^X_{w,y,n} = P_{X,N} + \sum_{i=1}^{N} W_{i,N}\pi_i^{-1}\epsilon_i(\delta_{Y_i} - \delta_{X_i})$$

le choix des poids qui permet d'obtenir un estimateur bootstrap généralisé sans biais de la variance de $T(P^X_{y,n})$ n'est pas affecté par cette transformation (il suffit de remplacer Y_i par $Y_i - X_i$ dans les expressions). L'introduction d'une variable externe n'induit pas forcément une perte d'efficacité de l'estimateur si la corrélation entre X et Y est grande (le cas extrême étant $X_i = Y_i$ auquel cas $T(P_{Y,N}) = T(P_{X,N})$ est connu de variance nulle!).

S'il est assez aisé de montrer un résultat analogue à (25), sous des hypothèses de différentiabilité sur T , la généralisation de ces résultats au second ordre *i.e.* les analogues de (26) et (31) semblent beaucoup plus difficile à obtenir et reste un problème ouvert.

3 Simulations et application à des données de consommation

3.1. Simulations

Afin de comparer les performances des techniques de bootstrap pondéré avec les techniques asymptotiques traditionnelles, nous étudions dans ce paragraphe par simulation le comportement des deux méthodes dans l'estimation d'un ratio. Des données fictives (X_i, Y_i) $i = 1, ...N = 1000$ ont été générées de la façon suivante :

$$X_i = e_i \; ; Y_i = X_i\beta + f_i$$

où e_i et f_i sont des variables aléatoires indépendantes log-normales, translatée et standardisée pour f_i de sorte que $Ef_i = 0$ et $\mathrm{V}f_i = 1$. Le paramètre β permet de faire varier la corrélation ρ entre les variables X_i et Y_i. Les comportements respectifs des deux méthodes sont fortement dépendants de cette corrélation. Le paramètre d'intérêt est

$$R = \frac{\sum_{i=1}^{N} X_i}{\sum_{i=1}^{N} Y_i}$$

3.1.1. Tirage à probabilités égales sans remise (PESR)

On réalise alors un sondage à probabilités égales, sans remise de taille n, les tailles retenues pour n sont respectivement $n = 20$ et $n = 50$. Les tableaux suivants donnent les intervalles de confiance obtenus par la méthode asymptotique et par la méthode du bootstrap pondéré. La distribution de la

version pondérée du ratio est obtenue par calcul de Monte-Carlo en simulant $B = 1000$ fois un système de poids de variance et de skewness fixées selon la proposition 3. Les poids sont générés à partir de la multiplication de la combinaison de deux bernouillis multipliées par une loi uniforme, dont on ajuste les paramètres pour avoir les moments adéquats. Cette méthode permet de générer facilement des poids de moments fixés, ayant une distribution continue et unimodale. D'autres méthodes de génération de poids à moments fixés sont proposées dans BARBE et BERTAIL [1995] et DEVROYE [1986].

Cette procédure a été répétée 10 000 fois ce qui permet, d'une part d'étudier le comportement moyen de la procédure et d'estimer les probabilités de couverture des intervalles de confiance construits par les deux méthodes (il suffit de compter pour cela le nombre de fois où la vraie valeur du paramètre tombe dans l'intervalle de confiance construit à chaque itération de la procédure). Tous les calculs ont été effectués en double précision avec le logiciel Splus sous station Sun475 Sparc2.

Les tableaux 1, 2 et 3 sont tous construits sur le même modèle: ils indiquent les bornes des intervalles de confiance correspondant aux niveaux 2.5, 5, 95 et 97.5% respectivement pour l'asymptotique BA et pour le bootstrap BB et la valeur exacte de la probabilité de couverture (entre parenthèses sous la borne). Nous donnons en entête de chaque tableau les caractéristiques de la simulation ainsi que la vraie valeur du paramètre.

Ces résultats montrent que, dans tous les cas, le bootstrap pondéré améliore la précision car la méthode permet de tenir compte des phénomènes de dissymétrie. Néanmoins lorsque le coefficient de corrélation entre les deux variables est grand, l'approximation asymptotique même pour une petite taille ($n = 20$) donne déjà des intervalles de confiance satisfaisants et la méthode du bootstrap pondéré ne paraît pas s'imposer dans ce cas. Lorsque le coefficient de corrélation est faible voire moyen, l'approximation asymptotique donne des intervalles de confiance moins fiables que la distribution du bootstrap pondéré. On notera néanmoins que si l'on s'intéresse à la probabilité de couverture de l'intervalle de confiance bilatéral est beaucoup plus proche du niveau de départ fixé. Par exemple pour $n = 20$ et $\rho = 0.00$, au niveau 95% (resp. 90%), le niveau effectivement atteint par l'asymptotique est de 94.6% (resp. 91.4%). Ce phénomène s'explique aisément à partir des développements d'Edgeworth des statistiques symétriques (voir BERTAIL [1997]). En effet, un intervalle de confiance bilatéral (basé sur une fonctionnelle statistique suffisamment différentiable) est toujours correct au second ordre (*i.e.* jusqu'à l'ordre $O(n^{-1})$ en terme de probabilité de couverture bien que l'erreur commise sur les quantiles eux-mêmes puisse être grande de l'ordre de $O(n^{-1/2})$. De sorte que même si l'erreur de première espèce est petite, l'erreur de seconde espèce peut être grande conduisant à des intervalles de confiance fortement biaisés. Le bootstrap et, dans le cadre choisi ici, le bootstrap pondéré permettent de contourner cette difficulté. Par ailleurs, pour des intervalles de confiance bilatéraux, il est possible théoriquement dans certains cas de choisir les poids de telle sorte que les intervalles de confiance bilatéraux soient corrects jusqu'à l'ordre $O(n^{-2})$ voir $O(n^{-5/2})$, mais les procédures de mise en oeuvre deviennent alors très complexes.

TABLEAU 1

Comparaison des intervalles de confiance et des probabilités de couverture associées, cas PESR.

	$n=20, \rho=0.90, R=0.34$				$n=50, \rho=0.90, R=0.34$			
α	2.5%	5%	95%	97.5%	2.5%	5%	95%	97.5%
BA	0.269 (2.7)	0.280 (5.1)	0.391 (95.1)	0.401 (98.2)	0.301 (3.8)	0.308 (7.0)	0.376 (96.7)	0.382 (99.4)
BB	0.267 (2.4)	0.279 (5.0)	0.389 (94.6)	0.397 (97.4)	0.296 (2.4)	0.303 (4.8)	0.371 (94.8)	0.378 (97.3)
	$n=20, \rho=0.40, R=0.721$				$n=50, \rho=0.40, R=0.721$			
α	2.5%	5%	95%	97.5%	2.5%	5%	95%	97.5%
BA	0.412 (1.4)	0.454 (3.5)	1.015 (94.6)	1.070 (97.1)	0.526 (1.8)	0.559 (4.4)	0.905 (94.5)	0.940 (97.1)
BB	0.442 (2.4)	0.479 (4.9)	1.021 (94.9)	1.082 (97.5)	0.536 (2.4)	0.565 (4.9)	0.909 (95.0)	0.946 (97.5)
	$n=20, \rho=0.00, R=0.937$				$n=50, \rho=0.00, R=0.937$			
α	2.5%	5%	95%	97.5%	2.5%	5%	95%	97.5%
BA	0.715 (0.0)	0.751 (0.4)	1.124 (91.8)	1.160 (94.6)	0.816 (1.2)	0.838 (3.6)	1.062 (94.1)	1.083 (96.6)
BB	0.782 (2.4)	0.801 (5.0)	1.166 (94.9)	1.224 (97.5)	0.831 (2.6)	0.845 (5.0)	1.068 (95.0)	1.096 (97.5)

3.1.2. **Tirage à probabilités inégales (PPI)**

Nous présentons ici une étude par simulation similaire pour des sondages à probabilité inégales. Les probabilité d'inclusion associées aux individus (générés comme dans (i)) ont été choisies de deux manières différentes.

Dans les deux cas les v.a. Z_{i,n_α} définies par (21) et (30), avec trois premiers moments fixés, ont été générées à partir de combinaison de puissance de variables aléatoires gaussiennes (voir BARBE et BERTAIL [1995], I.5), soit pour $1 \leq i \leq n_\alpha$

$$Z_i = X_{i,n_\alpha} + \alpha(X^2_{i,n_\alpha} - 1) + \beta(X^3_{i,n_\alpha} - 3X_{i,n_\alpha})$$

$$\text{avec } X_{i,n_\alpha} \text{ i.i.d. de loi } N(0,1).$$

Les constantes α et β sont ajustées de manière à obtenir les moments adéquats des poids. La taille du rééchantillonnage a été fixée à 1000.

Cas 1 : Probabilités d'inclusion proportionnelles à Y_i, PPI-1.

Dans le premier type de simulation, les probabilités d'inclusion ont été choisies proportionnelles aux Y_i de sorte que l'estimateur de Horvitz-Thompson du ratio ainsi obtenu $n^{-1}\sum_{i=1}^{n} \frac{X_i}{Y_i}\epsilon_i$ est sans biais. Les résultats de cette simulation pour différentes tailles et coefficients de corrélations sont consignés dans le tableau 2 construit sur le même modèle que le tableau 1.

On remarque que lorsque la corrélation entre X et Y est assez forte, l'intervalle de confiance donné par l'asymptotique est beaucoup trop conservatif (le niveau atteint pour $n = 20$ est de l'ordre de 99% au lieu du 95% fixé initialement) et conduit à des intervalles de confiance trop larges. Les intervalles de confiance construits avec le bootstrap pondéré donnent d'excellents résultats en terme de couverture et ne sont pas centrés sur la vraie valeur du paramètre, ceci étant dû aux dissymétries des distributions sous-jacentes.

TABLEAU 2

Comparaison des intervalles de confiance et des probabilités de couverture associées, cas PPI-1.

	$n = 20, \rho = 0.40, R = 0.69$				$n = 50, \rho = 0.40, R = 0.69$			
α	2.5%	5%	95%	97.5%	2.5%	5%	95%	97.5%
BA	0.291 (0.5)	0.356 (1.0)	1.021 (99.8)	1.086 (100.0)	0.480 (0.4)	0.524 (1.4)	0.971 (99.2)	1.014 (99.9)
BB	0.405 (2.7)	0.463 (5.2)	0.904 (95.4)	0.965 (97.8)	0.538 (2.4)	0.590 (5.1)	0.898 (95.3)	0.953 (97.6)
	$n = 20, \rho = 0.00, R = 0.96$				$n = 50, \rho = 0.00, R = 0.96$			
α	2.5%	5%	95%	97.5%	2.5%	5%	95%	97.5%
BA	0.240 (5.4)	0.362 (7.8)	1.615 (99.6)	1.738 (100.0)	0.580 (8.7)	0.649 (11.5)	1.358 (96.6)	1.437 (98.8)
BB	0.115 (2.8)	0.221 (5.6)	1.459 (95.4)	1.574 (97.8)	0.272 (2.7)	0.443 (5.3)	1.328 (95.2)	1.419 (97.5)

Lorsque le coefficient de corrélation est nul, les intervalles de confiances sont naturellement moins précis que dans le cas précédent: dans ce cas, l'asymptotique tend à donner des intervalles d'un niveau plus faible que celui désiré. Le bootstrap donne encore de bons résultats.

Cas 2 : Probabilités d'inclusion décroissantes, PPI-2

Dans le second type, les probabilités d'inclusion sont générées aléatoirement à partir d'une loi log-normale puis classées par ordre décroissant. Ces probabilités de sélection ont été associées aux X_i triés par ordre croissant (ce qui signifie que la probabilité de tirer une valeur petite de X_i est plus grande) ce plan de sondage (qui induit de forts biais de sélection) conduit à des intervalles de confiances de moins bonne qualité. Les résultats sont consignés dans le tableau 3.

TABLEAU 3

Comparaison des intervalles de confiance et des probabilités de couverture associées, cas PPI-2.

	$n = 20, \rho = 0.40, R = 0.97$				$n = 50, \rho = 0.40, R = 0.97$			
α	2.5%	5%	95%	97.5%	2.5%	5%	95%	97.5%
BA	0.355 (9.2)	0.461 (13.7)	1.554 (97.9)	1.660 (99.6)	0.545 (6.0)	0.617 (8.8)	1.358 (96.4)	1.429 (98.5)
BB	0.075 (2.7)	0.23 (5.2)	1.46 (95.1)	1.60 (97.4)	0.353 (2.2)	0.509 (4.9)	1.326 (95.2)	1.412 (97.6)
	$n = 20, \rho = 0.00, R = 1.02$				$n = 50, \rho = 0.00, R = 1.02$			
α	2.5%	5%	95%	97.5%	2.5%	5%	95%	97.5%
BA	0.409 (13.1)	0.537 (17.6)	1.842 (98.4)	1.967 (99.5)	0.651 (10.7)	0.735 (14.5)	1.542 (93.9)	1.626 (96.9)
BB	−0.10 (3.6)	0.13 (7.3)	1.684 (95.6)	1.838 (98.1)	0.362 (2.7)	0.491 (5.3)	1.564 (94.8)	1.821 (97.8)

Là encore il est clair que, pour l'ensemble de ces simulations, la distribution asymptotique donne une approximation beaucoup moins fiable de la distribution à distance finie, ce d'autant plus que le coefficient de corrélation entre les variables est faible: pour $n = 20$ et $n = 50$, le niveau de l'intervalle est de l'ordre de 80% au lieu du 90% fixé initialement, les intervalles de confiance sont beaucoup trop « optimistes » et ne tiennent pas compte des dissymétries qui sont flagrantes ici. Le bootstrap donne

des résultats très satisfaisants dans tous les cas et permet de construire des intervalles de confiance unilatéraux d'un niveau plus satisfaisant : les intervalles de confiance bilatéraux sont néanmoins beaucoup plus larges que ceux de l'asymptotique alors que le gain global n'est pas toujours important (par exemple pour $n = 50$ et $\rho = 0,4$, la couverture de l'intervalle bilatéral est de 92,5 pour l'asymptotique et de 95,4 pour le bootstrap pondéré) : ceci s'explique par le choix symétrique des quantiles qui n'est pas adapté, lorsque la distribution exacte est dissymétrique, et fournit des intervalles biaisés, de faibles puissances (voir aussi le commentaire dans le cas PESR). Il est bien sûr possible à partir de la distribution bootstrap qui rend compte de cette dissymétrie de calibrer les niveaux de telle sorte que l'intervalle bilatéral soit approximativement de niveau exact et de longueur minimale.

3.2. **Intervalles de confiances pour des ratios et des fractiles de consommation par ménage**

Les techniques étudiées dans les paragraphes précédents ont été mises en oeuvre de manière systématique pour évaluer la variabilité et construire des intervalles de confiance pour des moyennes, des ratios et des fractiles de consommation. Ces estimations destinées à établir les caractéristiques des distributions des consommations alimentaires des ménages, et en particulier à évaluer leurs valeurs extrêmes, ont été réalisées, à la demande de l'Observatoire des Consommations Alimentaires, à partir des données des panels de ménages de Secodip [*cf.* COMBRIS *et al.*, 1995].

Les panels de ménages de Secodip sont constitués sur la base d'un échantillonnage aléatoire stratifié par région, et par type d'habitat. Dans chacune des 21 régions (la Corse est exclue), les strates d'habitat distinguent les communes rurales et les communes urbaines, ces dernières étant classées en fonction de la taille de l'unité urbaine à laquelle elles appartiennent. Selon les régions, jusqu'à huit strates peuvent être constituées. La sélection des ménages résulte d'un tirage à deux degrés : le premier degré détermine un échantillon de communes dans chacune des strates, le deuxième degré permet de sélectionner des logements dans les communes retenues. La première étape aboutit à la constitution d'un échantillon de communes qui reste pratiquement stable pendant toute la période inter-censitaire. Ces communes constituent l'ensemble des points de sondage dans lesquels seront recrutés des ménages en fonction des besoins de renouvellement des panels. La deuxième étape de l'échantillonnage combine en proportions équivalentes le tirage aléatoire d'adresses, et la sélection, sur quotas simples, de ménages présentant des caractéristiques particulières. Sans cette combinaison de méthodes, certaines caractéristiques socio-démographiques seraient fortement sous-représentées dans l'échantillon des ménages participant effectivement au panel.

Les panels sont renouvelés de façon quasi-continue : chaque année un quart de l'échantillon est remplacé. Les recrutements sont effectués par vagues trimestrielles, les remplacements se faisant en bloc la fin de chaque trimestre. Chaque ménage participant est donc observé de façon continue pendant une longue période (quatre ans en moyenne). Pour diminuer la pression de collecte, deux panels permettent de suivre deux

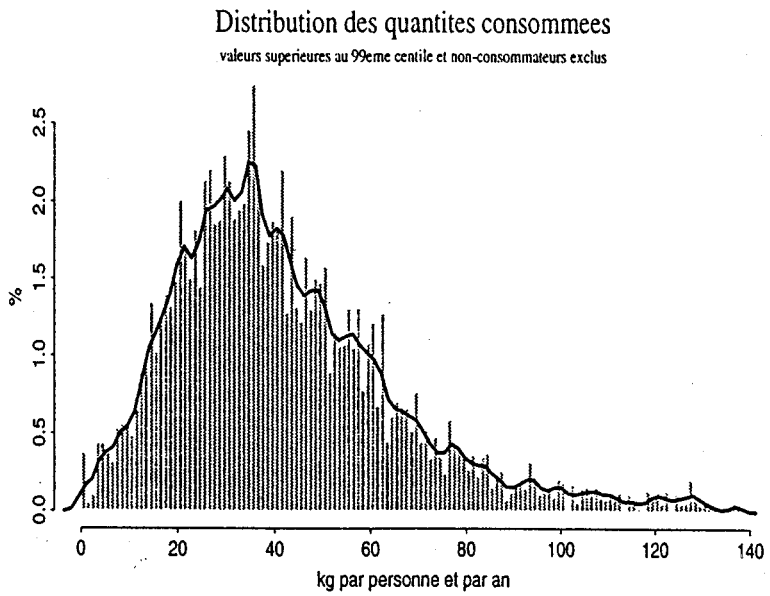
sous-ensembles d'achats distincts couvrant un grand nombre de produits de consommation courante et en particulier près de 70 % des dépenses alimentaires [*cf.* COMBRIS, 1993]. Chaque ménage inscrit dans un relevé hebdomadaire tous ses achats relatifs aux produits couverts par le panel auquel il appartient, en indiquant précisément la désignation de chaque produit, la quantité achetée, la dépense correspondante et la nature du point de vente.

Pour le calcul des statistiques périodiques (mensuelles, trimestrielles ou annuelles), Secodip procède à des redressements spécifiques. L'échantillon redressé se compose des seuls ménages ayant été « actifs » pendant la période considérée (par exemple les ménages ayant renvoyé au moins trois relevés au cours d'une période de quatre semaines pour les statistiques mensuelles, et les ménages ayant été « actifs » pendant douze périodes de quatre semaines sur treize pour les statistiques annuelles soit au moins trente six semaines). Les critères géographiques et socio-démographiques de redressement (région, type d'habitat, taille du ménage, âge de la ménagère, présence d'enfants...) sont actualisés par extrapolation à partir des résultats de la dernière enquête « emploi » disponible à l'Insee. Le calcul des poids associés à chaque individu se fait selon une procédure itérative de calage sur marges, qui impose des contraintes sur les valeurs maximum et minimum des poids acceptables. Nous utiliserons dans la suite ces poids comme s'ils s'interprétaient directement comme l'inverse des probabilités d'inclusion du plan de sondage : ceci se justifie par des considérations asymptotiques mais il serait néanmoins intéressant dans une étape ultérieure de voir quelle est en terme de variance l'influence du calage. Nous donnons ici quelques résultats concernant la consommation de viande, toutes catégories confondues et sur quelques produits peu ou moyennement consommés, chapon et oie. C'est sur ces produits que les différences entre les méthodes asymptotiques et les méthodes de type bootstrap sont les plus nettes.

Le graphique 1 donne la distibution $P_{y,n}$ de la consommation annuelle de viande. Les tableaux 4 à 6 sont des tableaux de synthèse donnant l'estimateur, l'écart-type de l'estimateur et permettant de comparer les estimations respectives des intervalles de confiance au seuil $\alpha = 5\%$ obtenues avec l'approximation asymptotique et la distribution bootstrap pondéré.

Les consommations ont été calculées après redressement du nombre de semaines pendant lesquelles le ménage est observé. Cons/pop désigne le ratio de la consommation totale sur la taille de la population. C/T est la consommation par tête des ménages, %Cons le pourcentage de consommateurs du produit, Cons/tête la consommation des ménages consommant effectivement le produit. Les fractiles d'ordre β, Fr cons $\beta\%$, sont eux aussi calculés sur la population dont la consommation est non nulle. Ces dernières statistiques sont fortement non linéaires d'où l'intérêt du bootstrap même si dans ce cas l'obtention de propriétés au second ordre est utopique.

Ces résultats donnent une bonne idée de la variabilité des estimateurs. Pour les fractiles, les variances asymptotiques ont été calculées en construisant un estimateur à noyau, pondéré par les probabilités d'inclusion des individus, de la densité de la distribution. Les intervalles bootstrap



GRAPHIQUE 1

Consommation de viande en 1989.

semblent plus réalistes que les intervalles asymptotiques en particulier pour les fractiles extrêmes: ceci est dû au fait que l'estimateur à noyau de la densité donne une mauvaise approximation des queues de la distribution

TABLEAU 4

Consommation de viande de chapon : Estimation, Ecart-type, Intervalles de confiance à 5% bootstrap et asymptotique.

	Estim.	*Ect*	*Int. Asympt.*		*Int. Bootst.*	
Cons/Pop (g)	20.6	3.4	14.0	27.2	15.6	24.4
C/T (g)	23.0	4.0	15.1	30.9	15.9	27.8
% Cons	1.7	0.3	1.1	2.3	1.2	2.0
Cons/Tête (kg)	1.47	0.11	1.26	1.70	1.27	1.64
Fr Cons 1%	0.11	0.03	0.05	0.17	0.10	0.63
Fr Cons 5%	0.35	0.02	0.30	0.39	0.11	0.80
Fr Cons 50%	1.38	0.01	1.35	1.42	1.12	1.74
Fr Cons 95%	2.52	0.37	1.78	3.25	1.95	3.60
Fr Cons 99%	3.60	6.92	0	17.16	2.10	3.61

TABLEAU 5

Consommation de viande d'oie: Estimation, Ecart-type, Intervalles de confiance à 5% bootstrap et asymptotique.

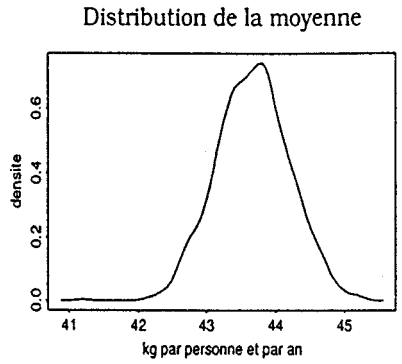
	Estim.	*Ect*	*Int. Asympt.*		*Int. Bootst.*	
Cons/Pop (g)	45.1	8.1	29.1	61.0	30.3	52.2
C/T (g)	51.1	8.9	33.7	68.5	35.6	59.1
% Cons	2.6	0.3	1.9	3.2	2.2	2.9
Cons/Tête (kg)	2.06	0.22	1.62	2.50	1.55	2.32
Fr Cons 1%	0.06	0.01	0.03	0.08	0.06	0.21
Fr Cons 5%	0.19	0.02	0.15	0.23	0.12	0.30
Fr Cons 50%	1.56	0.3	1.50	1.60	1.38	1.87
Fr Cons 95%	6.00	0.21	5.58	6.42	4.05	8.47
Fr Cons 99%	9.93	0.12	9.69	10.02	6.00	9.93

TABLEAU 6

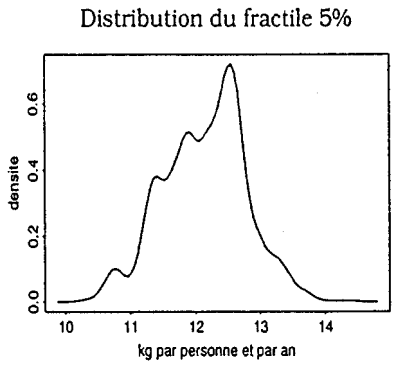
Consommation de viande: Estimation, Ecart-type, Intervalles de confiance à 5% bootstrap et asymptotique.

	Estim.	*Ect*	*Int. Asympt.*		*Int. Bootst.*	
Cons/Pop (kg)	39.48	0.46	38.52	40.42	38.91	40.53
C/T (kg)	43.57	0.53	42.53	44.61	42.82	44.32
% Cons	99.9	0.1	99.7	100.0	99.6	100.0
Cons/Tête (kg)	43.67	0.53	42.63	44.71	42.93	44.41
Fr Cons 1%	41.77	0.50	31.90	51.64	31.49	52.79
Fr Cons 5%	12.19	0.44	11.34	13.06	10.76	13.25
Fr Cons 50%	38.11	0.46	37.20	39.01	36.74	39.13
Fr Cons 95%	94.08	2.18	89.80	98.36	89.79	99.55
Fr Cons 99%	137.72	5.35	127.29	148.24	127.84	151.93

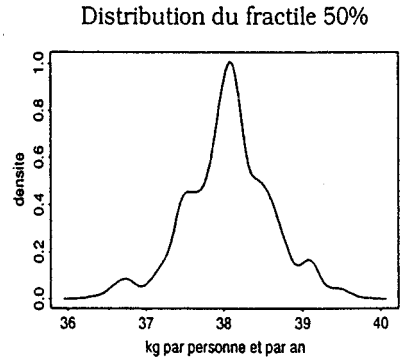
surtout pour des petites tailles d'échantillon. Lorsque le produit est très fortement non-consommé, les intervalles de confiance donnés par l'asymptotique sont en général beaucoup trop optimistes. Pour l'ensemble des produits, les résultats sont assez similaires avec cependant une légère dissymétrie des intervalles de confiance sans doute due à la non-symétrie des distributions sous-jacentes et à la nature non-linéaire des statistiques étudiées. Les graphiques 2 à 5 donnent par ailleurs les distributions bootstrap pondéré de la moyenne par tête, de la moyenne de la consommation totale de viande, et des fractiles à 5, 50 et 95%. Les formes des distributions



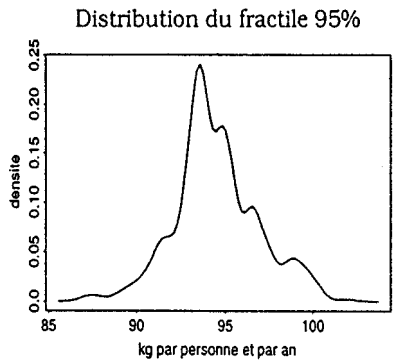
GRAPH. 2



GRAPH. 3



GRAPH. 4



GRAPH. 5

GRAPHIQUE 2 à 5

Distributions bootstrap de la moyenne et des fractiles de la consommation de viande.

montrent clairement que, pour les fractiles, la normalité asymptotique donne un reflet erroné du comportement à distance fini.

ANNEXE 1

Preuve de la proposition 1

Les résultats des propriétés 1 et 2 reposent sur un théorème de MASON et NEWTON [1993] dont nous rappelons l'énoncé :

THÉORÈME 1 : MASON et NEWTON [1993].

Soit $\{X_{k,n}, k = 1, 2, .., k_n, n \geq 1\}$ un tableau triangulaire de variables aléatoires définies sur un espace (Ω_1, A_1, P_1) et $\{W_{k,n}, k = 1, 2, .., k_n, n \geq 1\}$ un tableau triangulaire de variables aléatoires échangeables définies sur (Ω_2, A_2, P_2).

On définit pour toute séquence k_n réelle les quantités

$$Z_n = k_n^{1/2} \sum_{k=1}^{k_n} X_{k,n}(W_{k,n} - \bar{W}_n) / \left(\sum_{k=1}^{k_n} (X_{k,n} - \bar{X}_n)^2 \sum_{k=1}^{k_n} (W_{k,n} - \bar{W}_n)^2 \right)^{1/2}$$

$$U_{k,n} = (X_{k,n} - \bar{X}_n) / \left(\sum_{k=1}^{k_n} (X_{k,n} - \bar{X}_n)^2 \right)^{1/2}$$

$$V_{k,n} = (W_{k,n} - \bar{W}_n) / \left(\sum_{k=1}^{k_n} (W_{k,n} - \bar{W}_n)^2 \right)^{1/2}$$

Si on a

(A.1) $\max_{1 \leq k \leq k_n} U_{k,n}^2 \to 0 \ a.s. \ P_1$

(A.2) $\max_{1 \leq k \leq k_n} V_{k,n}^2 \to 0 \ en \ P_2$-probabilité

(A.3) $D_n(\tau) = \sum_{i=1}^{k_n} \sum_{j=1}^{k_n} U_{i,n}^2 V_{j,n}^2 WWW_{\{k_n U_{i,n}^2 V_{j,n}^2 > \tau\}}$

converge en P_2-probabilité vers 0, p.s. pour toute suite $\{U_{k,n}, \ k = 1, 2, .., k_n, \ n \geq 1\}$ (ou encore p.s. dans la suite)

alors uniformément en t

$$P(Z_n \leq t \mid \{X_{k,n}, k = 1, 2, .., k_n\}) \to \Phi(t), \text{ p.s. dans la suite.}$$

Démonstration de (19) :

(19) se démontre en vérifiant les hypothèses (A.1) et (A.3) et en appliquant le théorème à la suite triangulaire des v.a. $X_{i,n_\alpha} = y_i, \ i \in s$ et aux variables aléatoires échangeables W_{i,n_α}, avec $k_n = n_\alpha$. (A.2) appliqué aux poids échangeables W_{i,n_α} définis par (18) découle directement du corollaire 2.3 de MASON et NEWTON [1993] (voir aussi leur exemple 2.1). En toute rigueur, lorsque n_α et N_α tendent vers ∞, les variables Y_i, π_i sont aussi indexées par n_α; toute ambiguité étant écartée, nous conservons néanmoins ces notations pour alléger les démonstrations.

(i) *Vérification de* (A.1)

Sous $K^{\epsilon}_{4,N_\alpha} < \infty$, $s^2_{n_\alpha}$ étant un estimateur sans biais convergent de $s^2_{N_\alpha} = N_\alpha^{-1} \sum_{i=1}^{N_\alpha} (Y_i - \bar{Y}_{N_\alpha})^2 < \infty$, sous les hypothèses de la proposition 1, le sondage est convergent et (9) est une conséquence du théorème de Lindeberg-Feller. De plus (A.1) s'écrit

$$U_{k,n_\alpha} = s^{-1}_{n_\alpha}(y_k - \bar{y}_{n_\alpha})/n_\alpha^{1/.2}$$

mais $0 < s^2_{n_\alpha} < \infty$ p.s. implique $\forall k$, $|y_k - \bar{y}_{n_\alpha}|/\sqrt{n_\alpha} \xrightarrow{\text{p.s.}} 0$ d'où (A.1)

(ii) *Vérification de* (A.3)

Sous (A.2), pour tout $\nu > 0$, il existe N tel que $\forall n_\alpha > N$,

$$\max_{1 \le k \le n_\alpha} V^2_{k,n_\alpha} \le \nu$$

On en déduit que

$$D_n(\tau) \le \sum_{i=1}^{n_\alpha} U^2_{i,n_\alpha \{n_\alpha U^2_{i,n_\alpha} > \tau/\nu\}} \sum_{j=1}^{n_\alpha} V^2_{j,n_\alpha}$$

$$= s^{-2}_{n_\alpha} n_\alpha^{-1} \sum_{i=1}^{n_\alpha} (y_i - \bar{y}_{n_\alpha})^2_{\{(y_i - \bar{y}_{n_\alpha})^2 > s^2_{n_\alpha} \tau/\nu\}}$$

$$\le s^{-2-\epsilon}_{n_\alpha} \left(\frac{\nu}{\tau}\right)^{\epsilon} \left\{ n_\alpha^{-1} \sum_{i=1}^{n_\alpha} (y_i - \bar{y}_{n_\alpha})^{2+\epsilon} \right\}$$

Comme $n_\alpha^{-1} \sum_{i=1}^{n_\alpha} (y_i - \bar{y}_{n_\alpha})^{2+\epsilon}$ est stochastiquement borné, (A.3) s'en déduit immédiatement.

Démonstration de (14):

La validité du développement d'Edgeworth pour la moyenne dans le cas d'un sondage équiprobable sans remise a été montré par BABU et SINGH [1985]. On a directement avec leur résultat:

$$P\{(\bar{y}_{n_\alpha} - \bar{Y}_{N_\alpha})/\sqrt{\text{Var}(\bar{y}_{n_\alpha})} < x\} \tag{E.1}$$

$$= \Phi(x) - n_\alpha^{-1/2}(1 - 2n_\alpha/N_\alpha)(1 - n_\alpha/N_\alpha)^{-1/2} K_{3,N_\alpha}/S^3_{N_\alpha}(x^2 - 1)\phi(x) + o(n_\alpha^{-1/2})$$

La validité du développement d'Edgeworth pour la version bootstrap pondéré

$$\text{Var}(\bar{y}_{w,n_\alpha} \mid P_{n_\alpha})^{-1/2}(\bar{y}_{w,n_\alpha} - \bar{y}_{n_\alpha})$$

découle de ZHIDONG et LINCHENG [1986].

La forme du développement est donné par le calcul du moment centré d'ordre 3 de $\bar{y}_{w,n_\alpha}$ conditionellement à P_{y,n_α}. On a par un calcul direct

$$E(\bar{y}_{w,n_\alpha} - \bar{y}_{n_\alpha} \mid P_{y,n_\alpha})^3 = n_\alpha^{-3} \sum_{i=1}^{n_\alpha} E(n_\alpha W_{i,n_\alpha} - 1)^3 (y_i - \bar{y}_{n_\alpha})^3 + o(n_\alpha^{-3})$$

en posant $\beta_{w,n_\alpha} = E(n_\alpha W_{i,n_\alpha} - 1)^3/\{E(n_\alpha W_{i,n_\alpha} - 1)^2\}^{3/2}$ on en déduit immédiatement le D.E. :

$$\text{(E.2)} \qquad P\{\mathrm{Var}(\bar{y}_{w,n_\alpha} \mid P_{n_\alpha})^{-1/2}(\bar{y}_{w,n_\alpha} - \bar{y}_{n_\alpha}) < x \mid P_{y,n_\alpha}\}$$

$$= \Phi(x) - n_\alpha^{-1/2}\beta_{w,n}K_{3,n_\alpha}/S_{n_\alpha}^3(x^2-1)\phi(x) + o(n_\alpha^{-1/2})$$

En particulier si β_{w,n_α} satisfait (2.12), sous la condition $K_{6,N_\alpha}^{\epsilon} < \infty$, on a

$$K_{3,n_\alpha}/S_{n_\alpha}^3 - K_{3,N_\alpha}/S_{N_\alpha}^3 \xrightarrow{\text{p.s.}} 0$$

et donc les deux développements d'Edgeworth coïncident jusqu'à l'ordre $o(n_\alpha^{-1/2})$, d'où (14).

Preuve de la proposition 2

La preuve de (26) repose sur le théorème de Lindeberg Feller. On a

$$\bar{y}_{w,n_\alpha} - \bar{y}_{n_\alpha} = N_\alpha^{-1} \sum_{i=1}^{N_\alpha} (N_\alpha W_{i,N_\alpha} - 1)\pi_i^{-1} Y_i \epsilon_i$$

En posant $H = (h_1, .., h_{N_\alpha})$ avec $h_i \in I\!R^{N_\alpha}$ on a $N_\alpha W_{i,N_\alpha} - 1 = h_i^{'} Z$ d'où

$$\bar{y}_{w,n_\alpha} - \bar{y}_{n_\alpha} = N_\alpha^{-1} Z^{'} HY = N_\alpha^{-1} \sum_{i=1}^{N_\alpha} (h_i^{'} Y) Z_{i,N_\alpha}$$

Les $(Z_{i,N_\alpha}), 1 \leq i \leq N_\alpha$, sont des v.a. indépendantes d'espérance nulle et on a

$$N_\alpha^{-1} \sum_{i=1}^{N_\alpha} (h_i^{'} Y)^2 \mathrm{Var}(Z_{i,N_\alpha}) = N_\alpha^{-1} \sum_{i=1}^{N_\alpha} (h_i^{'} Y)^2 = N_\alpha \hat{V}_{n_\alpha}$$

qui converge vers $S^2 < \infty$, d'après (23). D'après le théorème de Lindeberg-Feller (voir SERFLING [1981] théorème B, p. 31) il suffit de vérifier que

$$D_{N_\alpha}(\epsilon) := N_\alpha^{-1} \sum_{i=1}^{N_\alpha} (h_i^{'} Y)^2 \int_{|h_i^{'} Y||x| > \epsilon N_\alpha^{1/2}} x^2 dF_{Z_i}(x) \to 0 \text{ quand } N_\alpha \to \infty$$

Or comme $N_\alpha \hat{V}_{n_\alpha} = N_\alpha^{-1} \sum_{i=1}^{N_\alpha} (h_i^{'} Y)^2$ converge vers S^2, pour $N_\alpha > N_O$, on a

$$\max_{1 \leq i \leq N_\alpha} (h_i^{'} Y)^2 < A$$

d'où

$$D_{N_\alpha}(\epsilon) \leq \hat{V}_{n_\alpha} \int_{|x| > \epsilon A^{-1} N_\alpha^{1/2}} x^2 dF_Z(x)$$

mais $E \mid Z \mid^3 < \infty$ implique

$$\int_{|x| > \epsilon A^{-1} N_\alpha^{1/2}} x^2 dF_{Zi}(x) \to 0 \text{ quand } N_\alpha \to \infty$$

d'où le résultat.

ANNEXE 3

Preuve de la proposition 3

Les v.a. ϵ_i sont indépendantes et donc $T(P_{y,n_\alpha}) - T(P_{Y,N_\alpha})$ s'exprime comme une somme pondérée des ϵ_i. Sous les conditions (27) et (29), il suffit d'appliquer les résultats de ZHIDONG et LINCHENG [1986] sur la validité du développement d'Edgeworth pour la moyenne pondérée. Comme pour la démonstration de (24) il suffit de calculer le moment centré d'ordre 3 de $T(P_{y,n_\alpha})$ conditionellement à $\mid P_{y,n_\alpha}$, soit par indépendance des ϵ_i,

$$m_{3,n_\alpha} := E(\bar{y}_{n_\alpha} - \bar{Y}_{N_\alpha} \mid P_{y,n_\alpha})^3 = N_\alpha^{-3} E\left\{\sum_{i=1}^{N_\alpha} (\pi_i^{-1}\epsilon_i - 1) Y_i\right\}^3$$

$$= N_\alpha^{-3} \sum_{i=1}^{N_\alpha} E(\pi_i^{-1}\epsilon_i - 1)^3 Y_i^3$$

avec

$$E(\pi_i^{-1}\epsilon_i - 1)^3 = \pi_i^{-2}(1-\pi_i)(1-2\pi_i).$$

Avec la définition de K_{3,Y,N_α}, on en déduit immédiatement que

$$P\{(\bar{y}_{n_\alpha} - \bar{Y}_{N_\alpha})/\text{Var}(\bar{y}_{n_\alpha}) < x\} = \Phi(x) - N_\alpha^{-1/2} K_{3,Y,N_\alpha}(x^2-1)\phi(x)$$

$$+ o(N_\alpha^{-1/2} K_{3,Y,N_\alpha})$$

Le développement d'Edgeworth de $T(P_{y,n_\alpha}) - T(P_{Y,N_\alpha})$ se déduit des mêmes arguments car on a

$$N_\alpha W_{i,N_\alpha} = 1 + (1-\pi_i)^{1/2} Z_{i,N_\alpha},\ 1 \le i \le N_\alpha,$$

donc

$$T(P_{y,n_\alpha}) \ - \ T(P_{Y,N_\alpha}) = N_\alpha^{-1} \sum_{i=1}^{N_\alpha} (1-\pi_i)^{1/2} \pi_i^{-1} \epsilon_i Y_i Z_{i,N_\alpha}$$

qui est une somme pondérée des v.a. indépendantes Z_{i,N_α}, $1 \le i \le N_\alpha$.

Son moment d'ordre 3 conditionnellement à P_{n_α} est donné par

$$m_{w,3,n_\alpha} = N_\alpha^{-3} E Z^3 \sum_{i=1}^{N_\alpha} (1-\pi_i)^{3/2} \pi_i^{-3} Y_i^3 \epsilon_i.$$

Sous l'hypothèse (30) en posant

$$k_{3,w,n_\alpha} = \left\{N_\alpha^{-1} \sum_{i=1}^{N_\alpha} \pi_i^{-2}(1-\pi_i) Y_i^2 \epsilon_i\right\}^{-3/2}$$

$$\left\{N_\alpha^{-1} \sum_{i=1}^{N_\alpha} \pi_i^{-3}(1-\pi_i)(1-2\pi_i) Y_i^3 \epsilon_i\right\},$$

on en déduit que

$$P\{(\bar{y}_{w,n_\alpha} - \bar{y}_{n_\alpha})/\text{Var}(\bar{y}_{w,n_\alpha} \mid P_{n_\alpha}) < x \mid P_{n_\alpha}\} = \Phi(x)$$

$$-N_\alpha^{-1/2} k_{3,w,n_\alpha}(x^2-1)\phi(x) + o(N_\alpha^{-1/2} k_{3,w,n_\alpha}).$$

Sous les hypothèses de ROSEN [1972] et l'hypothèse (29), le numérateur et le dénominateur de k_{3,w,n_α} convergent respectivement vers les valeurs associées de $K_{3,Y,N}$, de sorte que l'on a

$$k_{3,w,n_\alpha} - K_{3,Y,N} = o(1).$$

On en déduit le résultat de la proposition 3.

• Références bibliographiques

BABU, G. J., SINGH, K. (1985). – "Edgeworth Expansion for Sampling Without Replacement from a Finite Populations." *J. Multivariate Anal.*, 17, pp. 261-278.

BARBE, Ph., BERTAIL, P. (1995). – *The Weighted Bootstrap*, Monographie, 250 p., Lectures Notes in Statistics n° 98, Springer Verlag, N.Y.

BERTAIL, P. (1992). – *La Méthode du Bootstrap, Quelques Applications et Résultats Théoriques*, Thèse, Université Paris IX.

BERTAIL, P. (1997). – "Second Order Properties of an Extrapolated Bootstrap Without Replacement Under Weak Asumptions: the I.I.D. and Strong Mixing Cases," Bernouilli, 3, pp. 1-33.

CHAO, H., LO, K. Y. (1985). – "A Bootstrap Method for Finite Populations",*Sankhya*, 47, pp. 399-405.

CHOW, Y., TEICHER, H. (1988). – *Probability Theory, Independence, Interchangeability*, Martingales, Springer, New York.

COMBRIS, P. (1993). – "Méthodologie Statistique pour l'Estimation des Niveaux et de la Dispersion des Consommations", *Bulletin d'Information et de Documentation, DGCCRF*, Ministère de l'Economie et des Finances, 3, pp. 80-93.

COMBRIS, P., BERTAIL, P., BOIZOT, C., POUPA, J. C. (1995). – "La Consommation Alimentaire en 1991: Distribution des Quantités Consommées à Domicile," *INRA, Observatoire des Consommations Alimentaires*.

DEVILLE, J. C. (1987). – "Réplications d'Echantillons, Demi-échantillons, Jackknife, bootstrap", dans *Les Sondages*, Economica, Ed. Droesbeked, Fichet, Tassi.

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

DIACONIS, P., EFRON, B. (1983). – "Méthodes de Calculs Statistiques Intensifs sur Ordinateurs", *Pour la Science* (traduit de Computer Intensive Methods paru dans *The American Scientist*).

DUDLEY, R. M. (1994). – "The Order of the Remainder in Derivatives of Composition and Inverse Operators, *Ann. Stat.*, 22, pp. 1-20.

EDGEWORTH, F.(1907). – "On the Representation of a Statistical Frequency by a Series", *JRSS*(A), 70, pp. 102-106.

EFRON, B. (1979). – "Bootstrap Methods: An Other Look at the Jackknife", *Ann. Statist.*, 7, pp. 1-26.

EFRON, B. (1982). – *The Jackknife, the Bootstrap, and Other Resampling Plans*, CBMS-NF SIAM, 38.

EINMAHL, U., MASSON, D. (1992). – "Approximations to Permutation and Exchangeable Process", *J. Theor. Probab.*, 5, pp. 101-126.

FALK, M., REISS, R.D. (1989). – "Weak Convergence of Smoothed and Non-smoothed bootstrap Quantile Estimate", *Ann. Probab*., 17, pp. 362-371.

FELLER, W. (1971). – *An Introduction to Probability Theory and its Application*, Wiley, New-york.

GOURIEROUX, C. (1987). – "Généralités sur la Théories des Sondages", *Les Sondages*, Economica, Ed. Droesbeke, Fichet, Tassi.

HAEUSLER, E., MASON, D., NEWTON, M. (1992). – "Weighted Bootstrapping of Means", *CWI Quaterly*, pp. 213-228.

HALL, P. (1986). – "On the Bootstrap and Confidence Intervals", *Ann. Statist.*, 14, pp. 1431-1452.

HÄRDLE, W. (1989). – "Resampling from Inference Curve", *47th* P*roc. I.S.I. S*e*ssion*.

HUBER, P.J. (1981). – *Robust Statistics*, Wiley, New York.

KÜNSCH, H. R. (1989). – "The Jackknife and the Bootstrap for General Stationary Observations", *Ann. Statist*., 17, pp. 1217-1241.

LO, A.Y. (1991). – "Bayesian Bootstrap Clones and a Biometry Function. *Sankhya A*, 53, pp. 320-333.

MACCARTHY, P., SNOWDEN, H. (1984). – "The Bootstrap and Finite Population Sampling", *Technical Report*, National Center for Health Statistics.

MASON, D., NEWTON, M. A. (1992). – "A Rank Statistic Approach to the Consistency of a General bootstrap", *Ann. Statist.*, 20, pp. 1611-1624.

PRAESTGAARD, J. (1992). – *Bootstrap with General Weights and Multiplier Central Limit Theorems*, Ph.D. Thesis, University of Washington.

QUENOUILLE, M. H. (1949). – "Approximate Tests of Correlation in Time-Series", *JRSS(B)*, 11, pp. 68-84.

ROSEN, P. (1972). – "Asymptotic Theory for Successive Sampling", AMS, 43, pp. 373-397.

SCHNELLER, W.(1989). – "Edgeworth Expansion for the Linear Rank Statistic", *Ann. Statist.*, 17, pp. 1103-1123

SEN, P. K. (1988). – "Asymptotics in Finite Population Sampling", in *Handbook of Statistics*, vol. 6 (Sampling), pp. 291-331.

SERFLING, J. (1981). – *Approximation Theorems of Mathematical Statistics*, Wiley, New-York.

ZHENG, Z., TU, D. (1988). – "Random Weighting Method in Regression Model", *Scienta Sinica A*, XXXI, pp. 1442-1459.

ZHIDONG, B., LINCHENG, Z. (1986). – "Edgeworth Expansion of Distribution Function of Independant Random Variables", *Scienta Sinica(A)*, 29, pp. 1-22.