Évaluation de la précision d'estimateurs de fonctionnelles : l'exemple de la consommation alimentaire

Patrice Bertail, Christine Boizot, Pierre Combris

Évaluation de la précision d'estimateurs de fonctionnelles : l'exemple de la consommation alimentaire

Patrice BERTAIL
Christine BOIZOT
Pierre COMBRIS

Cahiers d'économie et sociologie rurales, n° 67, 2003

Patrice BERTAIL*, Christine BOIZOT**, Pierre COMBRIS**

Precision performance of functional estimators: the case of food consumption

Summary *– This paper proposes several methods for computing precise confidence intervals or evaluating the precision of some statistics related to individual food consumption, based on complex household food survey datas. We show how it is possible to obtain asymptotic confidence intervals for non-linear functionals thanks to the delta method and the notion of Hadamard differentiability. However asymptotic confidence intervals may not be very precise and to not take into account the dissymetries of the statistics or the underlying distributions. We develop two different methods based on resampling ideas to obtain precise confidence intervals. The first one is a transposition of the weighted bootstrap to survey sampling. The second uses the universal properties of subsampling and extrapolation methods to obtain rapidly accurate results. We compare and apply these methods to the construction of confidence intervals for means, fractiles, dispersion indexes of individual food consumptions (with or without null consumptions). We apply these methods to several products from the 1994 Secodip french panel.*

Key-words:
Hadamar differentiability, bootstrap, consumption survey, confidence intervals

Évaluation de la précision d'estimateurs de fonctionnelles : l'exemple de la consommation alimentaire

Résumé – Cet article propose plusieurs méthodes pour calculer des intervalles de confiance précis ou évaluer la précision de statistiques relatives à la consommation individuelle, lorsqu'on dispose de données de consommation (ou d'achats) par ménages issues de sondages complexes. Nous montrons comment il est possible d'obtenir des intervalles de confiance asymptotiques pour des statistiques non-linéaires complexes grâce à la méthode delta et la notion de différentiabilité au sens de Hadamard. Les intervalles de confiance asymptotiques peuvent ne pas être très précis et ne permettent pas de prendre en compte la dissymétrie des statistiques et des distributions sous-jacentes. Nous développons deux méthodes différentes basées sur du ré-échantillonnage permettant d'obtenir des intervalles de confiance plus précis. La première utilise une transposition du bootstrap pondéré dans le cadre des sondages. La seconde méthode utilise les propriétés universelles de distributions de sous-échantillonnage et les méthodes d'extrapolation pour obtenir rapidement des résultats précis. Nous comparons et appliquons ces méthodes à la construction d'intervalles de confiance pour des moyennes de consommation individuelle, à partir de données sur les ménages, ainsi que d'intervalles de consommations pour les fractiles de consommation par produit et pour les indices de dispersion (avec ou sans consommation nulle). Ces résultats sont appliqués à plusieurs produits du panel Secodip 1994.

Mots-clés :
Hadamar-différentiabilité, bootstrap, sondage, consommation, intervalles de confiance

* *CREST, LSA, J340, 3 avenue Pierre Larousse, 92405 Malakoff et INRA, CORELA, 65, Bd de Brandebourg, 94205 Ivry sur Seine*
e-mail : Patrice.Bertail@ensae.fr
** *INRA, CORELA, 65, Bd de Brandebourg, 94205 Ivry sur Seine*
e-mail : boizot@ivry.inra.fr
combris@ivry.inra.fr

Nous remercions vivement un des rapporteurs anonymes et le secrétariat de rédaction des *Cahiers* pour leurs remarques qui ont considérablement amélioré la forme de cet article. Ce travail a bénéficié du soutien financier de l'Observatoire des consommations alimentaires et de l'INRA : nous les en remercions.

CET article a pour but de faire le point sur les méthodes utilisées dans le laboratoire de recherche sur la consommation (INRA-CORELA) pour calculer des intervalles de confiance ou la précision de statistiques, relatives à la consommation individuelle, lorsque l'on dispose de données de consommation (ou d'achats) par ménage. Ces travaux, réalisés dans le cadre de contrats avec l'Observatoire de la consommation alimentaire (voir Bertail *et al.*, 1995, 1996, 1997, 1998a; Bertail *et al.*, 1998b), peuvent également s'appliquer à d'autres types de données. Nous avons toutefois souhaité les introduire dans le cadre de cette application, qui peut intéresser des économistes et également les personnes travaillant sur l'évaluation des risques alimentaires. Le problème de la construction d'intervalles de confiance précis pour des fonctionnelles fortement non linéaires (par exemple des fractiles) dans le cadre des sondages est, en effet, un problème difficile et peu abordé dans la littérature statistique. L'objectif de cet article est de proposer plusieurs méthodes, de fournir les expressions exactes des intervalles de confiance utilisés et de les comparer à la fois théoriquement et pratiquement, dans le cadre de la description de la consommation alimentaire en France, à partir des données des panels Secodip. Ces données et les problèmes qu'elles peuvent poser sont décrits dans le travail statistique préliminaire de Bertail et Combris (1997).

Dans une première partie, nous décrivons rapidement le contexte statistique en nous référant à la théorie des sondages. Nous montrons comment il est possible d'obtenir des intervalles de confiance asymptotiques pour des statistiques non linéaires complexes grâce à la méthode delta, qui consiste à prendre la partie principale linéaire de la statistique considérée comme une fonctionnelle dans l'espace des probabilités. Nous donnons brièvement quelques résultats théoriques sur la différentiabilité au sens de Hadamard, résultats qui peuvent également être utilisés dans de nombreuses applications économétriques. Nous appliquons ces méthodes à la construction d'intervalles de confiance pour des moyennes de consommation individuelle, à partir de données sur les ménages, ainsi que pour les fractiles de consommation par produit et des indices de dispersion. Pour certaines statistiques, il est important de tenir compte des consommations nulles (soit que l'on s'intéresse à elles en tant que telles ou que l'on cherche à les éliminer): les méthodes présentées permettent de tenir compte de ce facteur supplémentaire.

Les intervalles de confiance asymptotiques ne permettent cependant pas de prendre en compte la dissymétrie des distributions sous-jacentes ni la courbure du problème statistique. Il s'avèrent souvent (des études par simulations peuvent le montrer aisément) très éloignés des intervalles de confiance exacts que l'on pourrait construire si tous les paramètres du sondage et les distributions sous-jacentes étaient connus (ceci est impossible dans le cadre des sondages puisqu'alors, toute l'information étant disponible, il ne serait pas utile de recourir à une statis-

tique...). Son principal avantage est de fournir très rapidement des intervalles de confiance raisonnables pour de grandes tailles d'échantillons. Nous proposons deux méthodes différentes à base de ré-échantillonnages, permettant d'obtenir des intervalles de confiance plus précis. Chacune fait l'objet d'une nouvelle partie. La première méthode utilise une transposition du bootstrap pondéré (voir Barbe et Bertail,1995, pour une bibliographie complète). Mise au point par Bertail et Combris (1997), elle a pour but de généraliser au sondage les méthodes du bootstrap usuel ou du bootstrap généralisé qui ne fonctionnent pas dans ce cadre. Cette méthode, qui consiste à reproduire l'aléa du sondage à partir de système de poids aléatoires choisis de manière adéquate, permet d'avoir d'excellentes approximations en termes de précision des intervalles de confiance (sauf pour les fractiles). La mise en œuvre de la méthode est cependant très coûteuse en temps informatique. Elle ne se justifie que pour des échantillons de petite taille. La deuxième méthode proposée est, elle aussi, basée sur des méthodes de calcul intensif, plus exactement sur l'extrapolation de distribution de sous-échantillonnages. Cette idée, exposée dans Bertail (1997) dans le cadre des champs aléatoires, est très proche du jackknife et consiste à estimer la valeur de la statistique sur des sous-échantillons de taille b_n de structures identiques pour avoir une idée de la distribution à b_n fixé. Une méthode d'extrapolation ou d'interpolation (c.-à-d. de prédiction de la distribution pour une autre valeur n) permet alors de construire des intervalles de confiance qui possèdent de meilleures propriétés en termes de précision que les intervalles asymptotiques, mais moins bons que ceux obtenus par bootstrap pondéré. En termes de rapport précision/temps de calcul, cette méthode l'emporte cependant sur la méthode du bootstrap pondéré, y compris pour les tailles moyennes d'échantillons (pour les très petites tailles, le bootstrap pondéré donne de bien meilleurs résultats). Par ailleurs, elle ne nécessite aucun calcul préalable, quelle que soit la statistique considérée et peut même être utilisée dans des problèmes où la dynamique du sondage joue un rôle important (séries temporelles, enquêtes répétées). Dans la dernière partie, nous donnons des éléments de comparaison entre les différentes méthodes sur plusieurs exemples de produits, à partir du panel Secodip de 1994. Une description succincte des programmes utilisés, disponibles sur demande aux auteurs, est présentée en annexe. Nous discutons brièvement de la généralisation de ces résultats à d'autres modèles économétriques dans le cadre des sondages.

STATISTIQUES DE CONSOMMATION INDIVIDUELLE

Nous observons (au minimum) dans nos données des vecteurs x_i, ici $x_i = (c_i, n_i)$, $i = 1,..., q$ affectés de poids p_i, où c_i est la consommation totale du ménage i d'un certain produit, n_i le nombre de personnes dans le

ménage et p_i le poids du ménage calculé à partir du plan de sondage et de divers redressements. Les poids attribués aux ménages dans de nombreuses enquêtes, et en particulier dans le panel Secodip utilisé ici, sont souvent la résultante de divers redressements des probabilités d'inclusion du plan de sondage et de calage sur certaines marges connues (Deville et Särndal, 1992). Ceci rend leur interprétation difficile comme l'inverse de probabilités d'inclusion (dans un sondage, l'aléa ne vient que de la probabilité qu'un ménage soit ou non tiré). Nous serons donc amenés à supposer d'emblée que le sondage réalisé est convergent, c.-à-d. qu'il rend compte des distributions réelles des consommations et des tailles des ménages (hypothèses que nous préciserons dans le paragraphe suivant). Nous supposerons, dans la suite, que le sondage a été effectué dans une population de taille Q très grande et l'on note $X_i = (C_i, N_i)$, $i = 1,...,Q$, les valeurs de la consommation du ménage et du nombre d'individus dans le ménage, sur l'ensemble de la population. Quitte à re-normaliser les poids, on peut toujours supposer que $\sum_{i=1}^{q} p_i = 1$.

Différentiabilité compacte et méthode Delta : un outil essentiel en économétrie et en sondage

La méthode delta (Green, 2000 ; Manski, 1988, pour des références en économétrie) est une généralisation de la méthode dite de Slutsky en économétrie. Nous montrons comment celle-ci se généralise dans le cadre des sondages, mais aussi pour de très nombreuses fonctionnelles en économétrie *via* la notion de différentiabilité au sens de Hadamard (ou différentiabilité compacte). Pour de plus amples références sur ce concept, nous renvoyons aux travaux récents de Van der Vaart (1998). Nous conservons le cadre de l'estimation de paramètre de la distribution de la consommation, mais ces résultats sont bien sûr valides pour des fonctionnelles plus complexes et de nombreux modèles économétriques.

Nous ferons l'hypothèse que le sondage a été fait de manière correcte, de telle sorte que la probabilité empirique des couples $x_i = (c_i, n_i)$, $i = 1,..., q$ pondérée par les p_i, c.-à-d., si l'on note δ_{x_i} la masse de Dirac en $x_i = (c_i, n_i)$, la probabilité empirique des observations pondérées par les poids p_i,

$$P_q = \sum_{i=1}^{q} p_i \, \delta_{x_i}$$

converge asymptotiquement (au moins en moyenne quadratique) vers

$$P = \lim_{Q \to \infty} P_Q,$$

où P_Q est la probabilité empirique des X_i, $i = 1,...,Q$

$$P_Q = Q^{-1} \sum_{i=1}^{Q} \delta_{X_i}$$

Une telle hypothèse ne permet cependant pas la construction d'intervalles de confiance. Nous supposerons en outre que le sondage réalisé satisfait les conditions usuelles des sondages, à savoir :

– que le sondage tient compte de tous les individus de la base de sondage,

– que le sondage ne charge pas une catégorie particulière de la population (pas de biais de sélection),

– que le sondage n'équivaut pas à un recensement et est de taille suffisante ($\frac{q}{Q} > \mu > 0$),

– que les distributions des variables d'intérêt ne sont pas dégénérées (c.-à-d. qu'il n'y a pas constance d'un facteur),

de telle sorte qu'il y a normalité asymptotique du sondage. Plus exactement, nous supposons que la probabilité P_q satisfait un principe d'invariance asymptotique, c.-à-d.

$$\sqrt{q}\,(P_q - P_Q) \rightarrow B(P)$$

quand $q \rightarrow \infty$ où $B(P)$ est un pont brownien.

Voir Rosen (1972) pour plus de précision sur la convergence asymptotique en sondage. On pourra aussi se référer à l'excellent article de Sen (1988) et au chapitre V de Gouriéroux (1981).

La construction d'intervalles de confiance pour des statistiques complexes relève alors de théorèmes de composition et repose essentiellement sur la différentiabilité de la fonctionnelle associée au paramètre d'intérêt (envisagée comme une fonction de la probabilité P_Q). Cette approche, due à von Mises (1936), est à la base de la robustesse (Hampel, 1974 ; Huber, 1981) et est connue en économétrie sous le nom de principe d'analogie (Manski, 1988).

Définition 1 : Une fonctionnelle $T(P)$ définie d'un espace de probabilité $\boldsymbol{P}$ (contenant les masses de Dirac et donc les probabilités empiriques) dans R, R^q ou plus généralement tout espace de Banach séparable $\boldsymbol{B}$ est dit différentiable au sens de Fréchet en P pour une métrique d sur P, de premier gradient $T^{(1)}(x, P)$, appelée aussi fonction d'influence, si et seulement si on peut écrire pour tout $R \in \boldsymbol{P}$ dans un voisinage de P

$$T(R) - T(P) = \int T^{(1)}(x, P)\,(R - P)\,(dx) + d(R, P)\,\varepsilon\,(d\,(R,P)),$$

où

$$T^{(1)}(x, P) = \lim \left(\frac{T((1-t)\,P + t\delta_x) - \mathrm{T}(P)}{t} \right)$$

et ε est continue nulle en 0. On notera que par construction $E_p T^{(1)}(x, P) = 0$ (voir Hampel, 1974).

Cette notion permet de généraliser le théorème de Slutsky (qui repose essentiellement sur la linéarisation de la statistique étudiée) et d'établir aisément des théorèmes centraux limites. Il est en effet aisé de voir que si d est une métrique telle que $\sqrt{qd}(P_q, P_Q) = O_P(1)$, alors on a pour une fonctionnelle différentiable en P_Q

$$T(P_q) - T(P_Q) = \sum_{i=1}^{q} p_i T^{(1)}(X_i, P_Q) + {}_{OP}(q^{-1/2}),$$

de sorte que l'on peut étudier la fonctionnelle en étudiant sa partie linéaire, c'est-à-dire sa fonction d'influence. Pour un sondage convergent et si $T^{(1)}(X_i, P_Q)^2$ est équi-intégrable par rapport à P, on aura alors :

$$q^{1/2}(T(P_q) - T(P_Q)) \rightarrow N(0, VP(T^{(1)}(X, P))$$

quand $q,Q \rightarrow \infty$. Le principal problème est en fait le choix de la métrique qui doit rendre simultanément la fonctionnelle différentiable et être compatible avec la condition $\sqrt{qd}(P_q, P_Q) = O_P(1)$. Cet aspect est étudié dans un cadre i.i.d. dans Barbe et Bertail (1995) où plusieurs métriques (de la famille des métriques indexées par des classes de fonctions) sont proposées. Il existe néanmoins une notion plus faible que la différentiabilité au sens de Fréchet qui permet d'obtenir par l'utilisation de simple propriété de continuité des théorèmes centraux limites sans avoir à choisir de métrique : la différentiabilité au sens de Hadamard par rapport à un espace tangent approprié (voir Van der Vaart, 1998 pour de plus amples références et une introduction moderne à la statistique). La différentiabilité au sens de Hadamard est la notion de différentiabilité la plus faible qui conserve la continuité de la composition (c.-à-d. telle que la composée de fonctions Hadamard différentiables soit Hadamard différentiable) et l'efficacité (la transformée d'une statistique efficace par une fonction Hadamard différentiable est efficace). De très nombreuses fonctionnelles considérées en statistique et en économétrie dans les applications courantes (sauf la densité par rapport à une certaine mesure) sont Hadamard différentiables, ce qui devrait en faire un outil privilégié de l'analyse de ces problèmes. Nous en donnons une définition simplifiée ci-dessous.

Définition 2 : T est Hadamard différentiable en P par rapport à l'espace tangent B_P si et seulement si il existe une fonctionnelle dT_P linéaire continue telle que pour toutes suites h_n telles que $h_n \rightarrow h \in B_P$,

$$\frac{T((P + t_n h_n)) - T(P)}{t_n} - dT_P.h \rightarrow 0 \text{ quand } t_n \rightarrow 0$$

Cette notion permet en particulier de généraliser la méthode delta au sondage de la manière suivante.

Propriété 1 : Si le sondage est convergent au sens où $\sqrt{q}d(P_q, P_Q) \to B(P)$, quand $q \to \infty$ et si T est continûment Hadamard différentiable en P par rapport à un espace B_p contenant les trajectoires de $B(P)$, alors si $\frac{q}{Q} \to 0$, on a :

$$q^{1/2}(T(P_q) - T(P_Q)) \to dT_P.B(P)$$

quand Q et $q \to \infty$. En particulier si la fonction d'influence existe et est non dégénérée, c.-à-d. $V_P T^{(1)}(X, P) > 0$, alors $dT_P.h = \int T^{(1)}(x, P)h(dx)$ et la limite est gaussienne $N(0, V_P T^{(1)}(X,P))$.

Idée de la démonstration : Il suffit de choisir respectivement $n = q$, $t_n = q^{-1/2}$, $h_n = \sqrt{q}(P_q - P_Q)$. Si $\sqrt{q}(P_q - P_Q) \to B(P)$, alors on a :

$$q^{1/2}(T(P_Q + (P_q - P_Q)) - T(P_Q)) = q^{1/2}(T(P_q) - T(P_Q)) \to dT_P.B(P)$$

lorsque Q et $q \to \infty$, par continuité de la différentielle. Par ailleurs si $dT_P.h = \int T^{(1)}(x, P)h(dx)$ l'intégrale stochastique $\int T^{(1)}(x, P)B(P)(dx)$ se réduit à une gaussienne de variance $V_P T^{(1)}(X, P)$. Une démonstration rigoureuse nécessiterait d'introduire des notions de convergence (Hoffmann-Jorgensen convergence), de contrôle de la mesurabilité des événements concernés, qui dépassent largement le cadre de cet article. Voir Van der Vaart (1998) pour de plus amples références.

Nous donnons, dans la suite, des applications à la construction d'intervalles de confiance pour des fonctionnelles importantes en analyse de la consommation et des risques potentiellement liés à la consommation. Ces calculs montrent que la technique utilisée permet très simplement d'obtenir des estimateurs de la variance et des intervalles de confiance pour des quantités parfois complexes.

Intervalle de confiance pour la moyenne des consommations individuelles

En faisant l'hypothèse (certes réductrice, mais que l'on pourra facilement relâcher par la suite) que la consommation du produit considéré se répartit uniformément entre les différents membres du ménage, un estimateur naturel de la consommation individuelle est donné par :

$$\hat{r}_q = \frac{\sum_{i=1}^{q} p_i c_i}{\sum_{i=1}^{q} p_i n_i} = \frac{\bar{c}}{\bar{n}}$$

où $\bar{c}$ et $\bar{n}$ ne sont rien d'autre que la consommation et la taille moyennes des ménages, calculées sur l'échantillon

$$\bar{c} = \sum_{i=1}^{q} p_i c_i$$
$$\bar{n} = \sum_{i=1}^{q} p_i n_i$$

Avec les notations introduites, la consommation individuelle moyenne peut aussi se réécrire comme le ratio de deux espérances mathématiques.

$$\hat{r}_q = \frac{E_{P_q} C}{E_{P_q} n}$$

est un estimateur convergent (sous les hypothèses *ad hoc* faites auparavant) de la fonctionnelle

$$r_Q = \frac{E_{P_Q} C}{E_{P_Q} N} = \frac{Q\, E_{P_Q} C}{Q\, E_{P_Q} C} = \frac{\sum_{i=1}^{Q} C_i}{\sum_{i=1}^{Q} N_i}.$$

Le paramètre fonctionnel associé à r_Q, noté

$$r(P) = \frac{E_P C}{E_P N},$$

pour toute probabilité jointe du couple (C, N), est un ratio qui présente de fortes non-linéarités mais qui est Hadamard différentiable par composition (la moyenne est Hadamard différentiable et la fonction $(x,y) \to \frac{x}{y}$ Hadamard différentiable en tout point tel que $y \neq 0$). La linéarisation de cette quantité est facilitée par l'approche explicitée précédemment et le calcul de la fonction d'influence de cette fonctionnelle. Celle-ci est définie par :

$$r^{(1)}(c, n, P) = \lim_{t \to 0} t^{-1} \left(\frac{E_{(1-t)P + t\delta_{(c,n)}} C}{E_{(1-t)P + t\delta_{(c,n)}} N} - \frac{E_P C}{E_P N} \right) = \frac{d}{dt}\, r((1-t)P + t\delta_{(c,n)}) \Big|_{t=0}$$

et vaut par un calcul immédiat :

$$r^{(1)}(c, n, P) = \frac{(c - E_P C)}{E_P N} - \frac{E_P C}{(E_P N)^2} \quad (n - E_P N).$$

Cette quantité donne la contribution d'une observation à la statistique d'intérêt. On peut alors approcher $\hat{r}_q$ par sa partie linéaire sous la forme :

$$\hat{r}_q \approx r(P) + \sum_{i=1}^{q} p_i r^{(1)}(c, n, P)$$

et obtenir sa variance asymptotique grâce à la **propriété** 1

$$V_{r_q} = E_p(r^{(1)}(C, N, P)^2).$$

En remplaçant P par son estimateur naturel à savoir P_q, on obtient un estimateur convergent de V_{r_q} (en utilisant simplement la continuité des fonctionnelles en jeu)

$$\begin{aligned}\widehat{V_{r_q}} &= E_{p_q}(r^{(1)}(C,\, N,\, P)^2) \\ &= \sum_{i=1}^{q} p_i\, r^{(1)}(c_i,\, n_i,\, P_q)^2 \\ &= \sum_{i=1}^{q} p_i \left(\frac{(c_i - \bar{c})}{\bar{n}} - \frac{\bar{c}}{\bar{n}^2}(n - \bar{n}) \right)^2,\end{aligned} \tag{1}$$

qui est l'estimateur usuel de la variance d'un ratio.

Un intervalle de confiance asymptotique de niveaux 95 % pour r_Q est alors donné par la traditionnelle formule « plus ou moins deux fois (ou plus exactement 1,96) l'écart-type »

$$\left[r_q - 1{,}96 * \sqrt{\widehat{V_{r_q}}\,/q}\;;\; r_q + 1{,}96 * \sqrt{\widehat{V_{r_q}}\,/q} \right].$$

Naturellement, ce calcul se justifie uniquement par des considérations asymptotiques qui peuvent s'avérer inadéquates si les tailles q et Q des populations considérées sont trop petites ; d'autres méthodes du type bootstrap peuvent alors s'avérer plus réalistes et meilleures (Bertail et Combris, 1997) ; nous y reviendrons ultérieurement.

Intervalle de confiance pour un fractile d'ordre α de la consommation individuelle

Un calcul similaire permet d'obtenir des intervalles de confiance pour les fractiles de consommation individuelle. On notera dans la suite $1\{A\}$ l'indicatrice de l'événement A : $1\{A\} = 1$ si l'événement A est réalisé et 0 sinon. La distribution théorique des consommations individuelles est définie par :

$$H_P(x) = \frac{E_P\left(N\, 1\left\{ \frac{C}{N} < x \right\} \right)}{E_P N}$$

et vaut

$$H_Q(x) = \frac{\sum_{i=1}^{Q} N_i\, 1\left\{ \frac{C_i}{N_i} < x \right\}}{\sum_{i=1}^{Q} N_i}$$

sur l'ensemble de la population (il s'agit donc de la proportion d'individus qui ont une consommation par tête inférieure à un niveau x sous l'hypothèse de répartition uniforme de la consommation au sein du ménage). Elle est estimée par :

$$H_q(x) = \frac{\sum_{i=1}^{q} p_i n_i\, 1\left\{ \frac{c_i}{n_i} < x \right\}}{\sum_{i=1}^{q} p_i\, n_i},$$

sur la sous-population observée.

On notera que dans la plupart des enquêtes, N_i est une variable discrète. Le calcul étant similaire (et plus général) dans le cas continu (il suffit dans le cas discret de remplacer les intégrales sur la seconde composante par des sommes finies dans les calculs suivants), nous ferons l'hypothèse que N_i est continu et indiquerons les modifications à faire dans le cas discret.

En utilisant les propriétés de différentiabilité au sens de Hadamard des fractiles (Dudley, 1994), on en déduit que le quantile d'ordre α de la distribution des consommations individuelles

$$F_Q(\alpha) = H_Q^{-1}(\alpha) = \inf\{x,\ H_Q(x) \geq \alpha\}$$

est Hadamard différentiable. En utilisant la proposition 1, il est donc estimé de manière convergente par :

$$F_q(\alpha) = H_q^{-1}(\alpha) = \inf\{x,\ H_q(x) \geq \alpha\}$$

La partie linéaire de la fonctionnelle associée $F_P(\alpha) = H_P^{-1}(\alpha)$ est donnée par la fonction d'influence

$$F_\alpha^{(1)}(c,\ n,\ P) = \frac{n\left(1\left\{\frac{c}{n} \leq H_P^{-1}(\alpha)\right\} - \alpha\right)}{\int n\, f_{(C,\ N)}(n H_P^{-1}(\alpha),\ n)\, dn},$$

où $f_{(C,\ N)}$ désigne la densité jointe du couple $(C,\ N)$ (voir le calcul en annexe 1). La variance asymptotique est donc :

$$VF_P(\alpha) = \frac{E_P N^2\left(1\left\{\frac{C}{N} \leq H_P^{-1}(\alpha)\right\} - \alpha\right)^2}{\left(\int n\, f_{(C,\ N)}(n H_P^{-1}(\alpha),\ n)\, dn\right)^2}.$$

Comme dans le cas de l'estimation de la variance d'un fractile estimé dans un modèle d'échantillonnage, il est aisé d'obtenir un estimateur convergent du numérateur, l'estimateur du dénominateur nécessite l'estimation préalable de la densité jointe $f_{(C,\ N)}$ par une méthode de type non paramétrique. Par exemple, un estimateur à noyau gaussien de $f_{(C,\ N)}$ est donné par :

$$\hat{f}_q(u,\ v) = (2\pi)^{-1}\, h_q^{-2} \sum_{i=1}^{q} p_i \exp\left(-\frac{(u-c_i)^2 + (v-n_i)^2}{2h_q^2}\right), \qquad (2)$$

où h_q est le paramètre de lissage, que l'on peut choisir de manière quasi-optimale (au sens de l'erreur quadratique) en prenant h_q proportionnelle à $q^{-1/6}$ voire de façon optimale par des procédures classiques d'itérations (Bosq et Lecoutre, 1987). Un estimateur de la variance de $F_q(\alpha)$ est alors donné par :

$$\widehat{VF}_q(\alpha) = \frac{\sum_{i=1}^{q} p_i n_i^2\left(1\left\{\frac{c_i}{n_i} \leq H_q^{-1}(\alpha)\right\} - \alpha\right)^2}{\left(\int n \hat{f}_q(n H_q^{-1}(\alpha),\ n)\, dn\right)^2}.$$

Le choix d'un noyau gaussien permet de calculer explicitement la valeur de l'intégrale au dénominateur (voir annexe 2), de sorte que l'on aboutit à une expression explicite de l'estimateur de la variance :

$$\widehat{VF}_q(x) = 2\pi(1+H_q^{-1}(\alpha)^2)^3 \frac{\sum_{i=1}^q p_i\, n_i^2 \left(1\{\frac{c_i}{n_i} \le H_q^{-1}(\alpha)\} - \alpha\right)^2}{\left(h_q^{-1}\sum_{i=1}^q p_i(c_i H_q^{-1}(\alpha) + n_i)\exp\left(-\frac{(c_i - H_q^{-1}(\alpha)n_i)^2}{2h_q^2(1+H_q^{-1}(\alpha)^2)}\right)\right)^2} \qquad (3)$$

On en déduit alors un intervalle de confiance pour le fractile d'ordre α de la forme :

$$\left[F_q(\alpha) - 1{,}96\sqrt{\widehat{VF}_q(\alpha)/q}\,;\; F_q(\alpha) + 1{,}96\sqrt{\widehat{VF}_q(\alpha)/q}\right].$$

Lorsque N_i est discret à valeur dans {1,...,K}, la densité marginale de N_i peut être simplement donnée par les proportions empiriques $\hat{f}_j = \sum_{i=1}^q p_i 1_{\{N_i = j\}}$ et l'on peut simplifier en utilisant l'estimateur de la densité par lissage simple de la distribution conditionnelle donné par :

$$\tilde{f}_q(nH_q^{-1}(\alpha)\backslash n = j) = (2\pi)^{-1/2} h_q^{-1} \sum_{i=1}^q p_i \exp\left(-\frac{(c_i - j\,H_q^{-1}(\alpha))^2}{2h_q^2}\right)\hat{f}_j 1_{\{n_i = j\}},$$

de sorte que le dénominateur devient

$$\left(\int n\,\tilde{f}_q(nH_q^{-1}(\alpha), n)\,dn\right)^2 = (2\pi)^{-1}h_q^{-2}\left(\sum_{j=1}^K \sum_{i=1}^q j\hat{f}_j p_i \exp\left(-\frac{(c_i - j\,H_q^{-1}(\alpha))^2}{2h_q^2}\right)1_{\{n_i = j\}}\right)^2.$$

Montrer la validité asymptotique de ces estimateurs relève de techniques classiques (voir par exemple Bosq et Lecoutre, 1987) et ne sera pas abordé ici. On notera cependant que des arguments de différentiabilité au sens de Hadamard ne peuvent être utilisés ici pour l'estimation de la variance, la densité n'étant pas Hadamard différentiable.

Intervalle de confiance pour des indices de dispersion

La construction d'un intervalle de confiance pour un indice de dispersion d'ordre α, défini comme le rapport du fractile d'ordre α à la moyenne, s'obtient en utilisant l'ensemble des résultats précédents. Considérons respectivement les quantités théoriques et estimées suivantes, associées à l'indice d'ordre α :

$$i(P)(\alpha) = \frac{H_P^{-1}(\alpha)}{r(P)}$$

est construit sur la probabilité théorique P,

$$i_Q(\alpha) = \frac{H_Q^{-1}(\alpha)}{r_Q},$$

sur l'ensemble de la population,

$$i_q(\alpha) = \frac{H_q^{-1}(\alpha)}{r_q},$$

sur l'échantillon observé.

La fonction d'influence de $i(P)(\alpha)$ s'obtient aisément à partir de celles de $H_P^{-1}(\alpha)$ et de $r(P)$, calculées dans les paragraphes précédents et vaut :

$$i^{(1)}(c, n, P) = \frac{F_\alpha^{(1)}(c, n, P)}{r(P)} - \frac{H_P^{-1}(\alpha)}{r(P)^2} r^{(1)}(c, n, P),$$

(on notera la similitude avec la fonction d'influence d'un ratio, déjà calculée dans le premier paragraphe).

On en déduit alors immédiatement la forme de l'estimateur de la variance de $i_q(\alpha)$:

$$\widehat{V} i_q(\alpha) = \sum_{i=1}^{q} p_i \, i^{(1)}(c_i, n_i, P_q)^2$$

$$= \frac{\widehat{VF}_q(\alpha)}{r_q^2} + \frac{H^{-1}(\alpha)^2}{r_q^4} \widehat{V_{r_q}} - 2\frac{H_q^{-1}(\alpha)}{r_q^3} * CORR_q(\alpha)$$

avec

$$CORR_q(\alpha) = (2\pi)^{1/2}(1 + H_q^{-1}(\alpha))^{3/2}$$

$$\frac{\sum_{i=1}^{q} p_i n_i \left(1\{\frac{c_i}{n_i} \le H_q^{-1}(\alpha)\} - \alpha\right)\left(\frac{(c_i - \overline{c})}{\overline{n}} - \frac{\overline{c}}{\overline{n}^2}(n - \overline{n})\right)}{h_q^{-1} \sum_{i=1}^{q} p_i (c_i H_q^{-1}(\alpha) + n_i) \exp\left(-\frac{(c_i - H_q^{-1}(\alpha) n_i)^2}{2h_q^2(1 + H_q^{-1}(\alpha)^2)}\right)}.$$

L'intervalle de confiance asymptotique de niveau 95 % pour i_Q est alors donné par :

$$\left[i_q(\alpha) - \sqrt{\widehat{V} i_q(\alpha)/q} \; ; \; i_q(\alpha) + \sqrt{\widehat{V} i_q(\alpha)/q} \right].$$

INTERVALLE DE CONFIANCE BOOTSTRAP PONDÉRÉ

Une méthode complexe mais performante

Lorsque la taille de la population totale ou la taille de la population observée est trop petite (le problème de savoir où commence l'asymptotique, c'est-à-dire à partir de quelles valeurs de q, est un problème déli-

cat. La règle que l'on rencontre parfois dans les manuels selon laquelle l'asymptotique commence à 30 est non seulement ridicule mais dangereuse), les intervalles de confiance asymptotiques peuvent se révéler complètement inadéquats, surtout si la distribution de la population sous-jacente possède une très forte dissymétrie et/ou si la statistique étudiée possède des non-linéarités. En effet, les quantités asymptotiques, en tant qu'effet moyen sur une grande population, ne tiennent pas compte de ces dissymétries. On peut montrer, sous certaines conditions de régularité, que l'erreur commise sur le niveau initialement choisi et les bornes d'un intervalle de confiance en prenant une méthode de type asymptotique est de l'ordre de $q^{-1/2}$ fois un certain coefficient d'asymétrie. Cette erreur peut donc être très grande pour des petites tailles d'échantillons. Pour pallier ce problème, il est possible d'utiliser des techniques de type bootstrap, techniques de ré-échantillonnage qui permettent d'obtenir des distributions des estimateurs et des intervalles de confiance à distance finie. Il existe une très vaste littérature sur le bootstrap en statistique : nous renvoyons à Shao et Tu (1996) pour un panorama très large de cette technique et de plus amples références. Malheureusement, ni le bootstrap usuel ni les méthodes de double bootstrap (voir par exemple Letson et McCullough, 1998) ne fonctionnent dans le cadre des sondages (Deville, 1987). Bertail et Combris (1997) ont proposé d'adapter une procédure de type bootstrap généralisé au problème de sondage aléatoire simple ou au sondage poissonien, qui permet d'obtenir des intervalles de confiance plus précis dans les cas évoqués précédemment. L'idée principale est de pondérer les observations (c_i , n_i) par un ensemble de poids aléatoires, appelé plan de ré-échantillonnage, qui restitue en quelque sorte la variabilité du sondage initial, et de déterminer par du calcul de Monte-Carlo la distribution des estimateurs sous la loi du plan de ré-échantillonnage. Il est possible de caractériser, selon différents critères d'estimation (critères de sans biais, critère de maximisation de la probabilité de couverture, etc.), la forme optimale des poids. Nous renvoyons à Bertail et Combris (1997), Bertail et Barbe (1995) pour plus de précisions sur le sujet. Cette méthode est lourde d'un point de vue informatique et nécessite des calculs préalables pour chaque statistique d'intérêt. Son principal avantage est de permettre d'obtenir des intervalles de confiance très précis, en particulier pour des moyennes et des rapports. Pour donner un ordre de grandeur, l'erreur commise peut être réduite de $q^{-1/2}$ à $q^{-3/2}$ voire q^{-2}. Par exemple, sur une population de 100 individus avec un coefficient d'asymétrie de 2 (par exemple pour une loi $\gamma(1)$), l'erreur commise en utilisant l'intervalle asymptotique est de l'ordre de 20 % alors qu'elle est inférieure à 1‰ avec la méthode du bootstrap pondéré. Le cas des fractiles est plus complexe : en effet la fonction d'influence d'un fractile (qui donne une idée du saut que peut faire la statistique si on rajoute un nouveau point) est une variable qui prend des valeurs discrètes de sorte que les approximations usuelles (développements d'Edgeworth) à la base de la méthode du bootstrap ne peuvent être utilisées directement. Le bootstrap usuel donne dans ce cas des

résultats très mauvais (des erreurs de l'ordre $n^{-1/4}$) bien pire que les résultats asymptotiques traditionnels. Le bootstrap pondéré permet de résoudre ce problème par l'utilisation de poids de distribution continue. Le choix adéquat des poids étant un problème délicat pour ce type de statistique, nous n'entrerons pas dans les détails : la difficulté de la mise en oeuvre du bootstrap pondéré qui doit être adapté à chaque situation pour obtenir des résultats optimaux fait de lui un outil utile pour une analyse fine mais peu pratique pour sa mise en oeuvre. Pour ces raisons et pour réduire les temps de calculs informatiques trop importants associés à cette méthode, nous avons mis au point une méthode plus flexible, donnant certes des résultats moins spectaculaires que le bootstrap pondéré mais qui peuvent s'appliquer à n'importe quelle statistique sous des hypothèses minimales dans des temps très raisonnables. Les propriétés de cette méthode sont étudiées en détails dans Bertail (1997), et Bertail et Politis (2001).

LES MÉTHODES DE SOUS-ÉCHANTILLONNAGE

Une méthode très générale

L'idée du sous-échantillonnage par bloc, ou le bootstrap par bloc, est très proche de l'estimateur du jackknife. L'idée est d'ôter systématiquement des blocs de population pour obtenir différentes valeurs de la statistique et ainsi avoir une idée de sa volatilité. Son origine remonte aux travaux de Carlstein (1986) sur l'estimation de la variance de statistiques complexes en séries temporelles. Ce dernier a en effet proposé d'obtenir un estimateur de la variance d'une statistique générale T_n en considérant la variance empirique des valeurs de la statistique calculées sur des sous-périodes adjacentes et a montré que, sous certaines conditions dépendances des variables aléatoires en jeu, l'estimateur ainsi obtenu était convergent, asymptotiquement gaussien. Künsch (1989) (et aussi Liu et Singh, 1992) a proposé une forme du bootstrap basée sur le ré-échantillonnage de blocs d'observations. L'idée est que pour garder la structure de corrélation entre les observations, il suffit de ré-échantillonner des blocs d'observations successives, c.-à-d. les variables $Y_t = (X_t,...,X_{t+b})$ (où b est « bien choisi ») au lieu des observations X_t, puis de reconstituer une population de taille n, ce qui permet d'obtenir une nouvelle estimation. Künsch (1989) a montré que cette procédure permet d'obtenir une approximation asymptotiquement valide de la vraie distribution. Cependant Lahiri (1992) a souligné que le fait de ré-échantillonner des blocs d'observations donnait un poids différent aux extrêmes de la série observée et que, de ce fait, la distribution bootstrap ne possédait pas de propriétés au second ordre, c.-à-d. n'améliorait pas l'approximation par rap-

port à l'asymptotique. Une correction explicite (un recentrage adéquat de la distribution) est possible dans le cas de la moyenne, mais peut s'avérer plus délicate dans le cas de fonctionnelles plus complexes. Le choix de la standardisation adéquate rend quant à lui la méthode difficilement utilisable en pratique même pour une moyenne (voir Bertail et Politis, 2001).

L'approche de Politis et Romano (1994) et Bertail (1997), dans le cadre très général des champs aléatoires, est de construire la distribution bootstrap directement sur les sous-blocs et non à partir d'une reconstruction artificielle de la population. Ceci revient à construire la distribution empirique de toutes les valeurs possibles de la statistique sur tous les sous-blocs de populations (pouvant se chevaucher) de taille fixe b_n avec $\frac{b_n}{n} \to 0$ (distribution qui peut aussi s'interpréter comme l'histogramme du « $n - b_n$ delete jackknife » (c.-à-d. un jackknife dans lequel non pas 1 observation mais $n - b_n$ observations sont détruites), introduit par Shao et Wu, 1989 et Wu, 1990). Cette forme de sous-échantillonnage est asymptotiquement valide sous des conditions minimales sur la statistique et le paramètre considérés, lorsque la vitesse de convergence de la statistique est connue. Ces résultats ont été généralisés par Bertail, Politis et Romano (1999) à des statistiques dont la vitesse de convergence est inconnue. Néanmoins Bertail (1997) a montré que les distributions obtenues par sous-échantillonnage ne possédaient pas de bonnes propriétés au second ordre. Ceci s'explique par le fait que cet histogramme est construit sur des sous-échantillons de taille b_n beaucoup trop petite par rapport à n (puisque $\frac{b_n}{n} \to 0$). Il est néanmoins possible, lorsque l'on connaît la distribution asymptotique de construire une extrapolation de Richardson de la distribution qui permet d'obtenir des propriétés au second ordre. Nous donnons dans la suite les éléments essentiels des résultats obtenus et indiquons brièvement les techniques utilisées.

Extrapolation de Richardson des distributions de sous-échantillonnages

Nous expliquons d'abord brièvement les principes de base de la méthode dans le cas usuel d'une statistique (quelconque) avec des variables aléatoires indépendantes identiquement distribuées. Nous montrons ensuite comment la méthode s'applique aux cas des sondages.

Le cas indépendant identiquement distribué

Soit $\{X_1, X_2,...,X_n\}_{n \in N}$ une séquence de variables aléatoires indépendantes de même loi P. Considérons une statistique $T_n = T_n(X_1, X_2,...,X_n)$ estimant un paramètre réel $\theta(P)$ (le cas général $\theta(P)$ à valeur dans un

espace fonctionnel est similaire) de vitesse de convergence τ_n (dans les cas les plus usuels $\tau_n = n^{1/2}$). Soit $S_n^2 = S_n^2 (X_1, X_2, ..., X_n)$ un estimateur de la variance asymptotique $\sigma^2 > 0$.

La taille de sous-échantillonnage est notée $b_n < n$. Il est possible de construire $N_n = \binom{n}{b_n}$ sous-échantillons de taille b_n, $\chi_{b_{n,i}}$, $i = 1, ..., N_n$. Il est alors possible de construire l'histogramme de l'ensemble des valeurs de la statistique construite sur tous les sous-échantillons (centrés en la valeur de la statistique calculée sur l'échantillon global et correctement standardisée). Cette distribution est appelée distribution bootstrap sans remise ou distribution de sous-échantillonnage. Sa fonction de répartition est donnée par :

$$\bar{K}_{b_n}(x \backslash X_n) = N_n^{-1} \sum_{i=1}^{N_n} I\{\tau_{b_n}(T_{b_n}(\chi_{b_{n,i}}) - T_n)/S_{b_n}(\chi_{b_{n,i}}) \leq x\},$$

où $I\{B\}$ est l'indicateur de l'événement B. Bertail (1997) montre que cette distribution peut aussi s'interpréter comme une distribution bootstrap pondérée.

Comme il n'est pas possible en général de construire la distribution complète car N_n peut être très grand, on peut se contenter d'une approximation stochastique, c.-à-d. de sélectionner non pas tous les sous-échantillons de taille b_n, mais de choisir au hasard un nombre raisonnable d'échantillons de cette taille (Bertail (1997) montre qu'un choix B_n de l'ordre de b_n^2 est suffisant). Politis et Romano (1994) ont montré que, sous des conditions minimales (même pour des cas non standard avec non-normalité asymptotique), cette distribution est asymptotiquement correcte pourvu que b_n soit choisi petit devant n, $b_n = n^{1/3}$ étant dans les cas standard un choix optimal.

Il est cependant préférable, dans le cas où la distribution asymptotique est gaussienne (notée Φ), d'utiliser les approximations normales usuelles. Pour la plupart des statistiques, l'erreur commise en utilisant la loi normale est généralement de l'ordre de $f_1(n)^{-1}$, où f_1 est une fonction croissante. Comme nous l'avons vu précédemment, dans le cas de la moyenne ou pour des fonctions de moments (variance, coefficient d'asymétrie etc.), $f_1(n) = n^{1/2}$. L'erreur commise en utilisant les quantiles de la distribution bootstrap sans remise pour construire un intervalle de confiance est typiquement de l'ordre de $f_1(b_n)^{-1}$, ce qui se comprend aisément car les statistiques sont construites sur des sous-échantillons de taille b_n au lieu de n. Cette erreur lui ôte tout avantage pour des échantillons de petite taille. Cependant, la distribution bootstrap sans remise permet de capter des phénomènes de dissymétrie de la statistique que l'on ignore totalement avec la distribution gaussienne. Il est clair que chacune de ces distributions possède des propriétés propres d'où l'idée de considérer une combinaison linéaire des deux. Cette combinaison s'interprète comme une interpolation au sens de Richardson de la distribution bootstrap sans remise (Bickel et Yahav, 1988). Elle est donnée par :

$$\overline{K}_n^*(x \backslash \boldsymbol{X}_n) = \frac{f_1(b_n)}{f_1(n)} \overline{K}_{b_n}(x \backslash \boldsymbol{X}_n) + \left(1 - \frac{f_1(b_n)}{f_1(n)}\right) \Phi(x).$$

Cette méthode permet d'obtenir des intervalles de confiance plus précis que l'asymptotique, mais moins précis que le bootstrap pondéré. Il suffit de prendre les quantiles de cette distribution (c.-à-d. de l'inverser, ce qui est possible avec n'importe quel outil statistique standard) pour obtenir un intervalle de confiance possédant de bonnes propriétés, y compris pour des statistiques complexes. Pour des moyennes, des ratios, il est même possible d'apporter un facteur correctif de population finie de la forme $(1 - \frac{b}{n})^{1/2}$ à la statistique standardisée pour avoir de meilleures approximations (Bertail, 1997).

Le second avantage par rapport à d'autres techniques de ré-échantillonnage est qu'elle nécessite beaucoup moins de calculs informatiques, puisqu'il suffit de recalculer la valeur de la statistique sur des échantillons de taille beaucoup plus faible et en beaucoup moins grand nombre. Pour comparaison, dans le cas simple de la moyenne, pour une taille d'échantillon $n = 100$, le bootstrap usuel ou le bootstrap pondéré nécessitent le calcul d'au moins $B_n = 10\,000$ valeurs de la statistique T_n, calculée sur des échantillons de taille n tirés avec remise si l'on souhaite avoir une correction effective du bootstrap qui ne soit pas perturbée par la phase de Monte-Carlo. Alors que la technique de sous-échantillonnage introduite ne nécessite (en théorie) que le calcul de 25 valeurs de la statistique calculée sur des sous-échantillons de taille inférieure à 100 ! (Bertail, 1997).

Ces résultats se généralisent aisément aux champs aléatoires fortement mélangeants, c.-à-d. présentant une structure de dépendance asymptotiquement faible. Ceci inclut, entre autres, le cas des séries temporelles stationnaires univariées et multivariées, ainsi que le cas des sondages (avec ou sans remise). Pour cela, il convient de construire les sous-échantillons de taille b_n en tenant compte de la structure de dépendance.

L'idée est de découper l'échantillon en sous-blocs de taille b_n, ou encore dans l'optique jackknife, de supprimer $k_n = n - b_n$ observations formant un ensemble cohérent (un cluster dans le cas des sondages, une suite dans le cas des séries temporelles). Sur chacun des sous-blocs présentant une structure de dépendance similaire à l'ensemble de la série, il est alors possible de calculer les valeurs de la statistique et sa variance. On peut alors de manière similaire au cas indépendant construire un histogramme des valeurs obtenues : là encore cet histogramme, bien qu'asymptotiquement convergent, ne possède pas de bonnes propriétés à distance finie, mais l'extrapolation de Richardson permet de résoudre ce problème sous des hypothèses minimales sur les statistiques et les séries utilisées.

Exemple 1. *Cas d'un sondage poissonien (tirages indépendants)*

Soit $(x_1, p_1),....(x_n, p_n)$, les individus observés affectés de poids que l'on supposera standardisés à 1 : $\sum_i p_i = 1$ (on notera que, dans ce cas, les poids ne s'interprètent plus comme les inverses des probabilités d'inclusion). Soit

$$p_n = \sum_i p_i \delta_{x_i}$$

la distribution pondérée des individus. On s'intéresse à une statistique différentiable $T(P_n)$, par exemple la moyenne :

$$m_n = \sum_i p_i x_i$$

ou un fractile de la distribution.

Dans le cas du sondage poissonien, il n'est pas nécessaire de construire des groupes d'individus à cause de l'indépendance mais il convient de conserver la structure des poids.

Soit (x_{i1}, p_{i1}), ... (x_{ib_n}, p_{ib_n}) un sous-échantillon de taille b_n obtenu en choisissant b_n individus (on peut montrer que si b_n est suffisamment petit devant n, alors le fait de tirer avec ou sans remise n'influe pas sur la technique). On peut alors calculer la distribution :

$$P_{b_n} = \sum_{j=1}^{b_n} p_{ij} \delta_{x_{ij}} / \sum_{j=1}^{b_n} p_{ij}$$

ainsi que ses caractéristiques moyennes, fractiles, etc.

Cas de la moyenne :

En répétant cette opération $B_n = b_n^2$ fois, on obtient B_n valeurs de la moyenne de la forme :

$$m_{b_n}^{(k)} = \sum_{j=1}^{b_n} p_{ij} x_{ij} / \sum_{j=1}^{b_n} p_{ij}, \; K = 1, \ldots B_n$$

et B_n estimateurs de l'écart-type associé de la forme :

$$\sigma_{b_n}^{(k)} = \left(b_n^{-1} \sum_{j=1}^{b_n} p_{ij} (x_{ij} - m_{b_n})^2 / \sum_{j=1}^{b_n} p_{ij} \right)^{1/2}.$$

Il suffit alors de construire l'histogramme des valeurs de $b_n^{1/2} (m_{b_n}^{(k)} - m_n) / \sigma_{b_n}^{(k)}$ puis de les mixer avec la normale dans les proportions $\frac{b_n^{1/2}}{n^{1/2}}$ et $1 - \frac{b_n^{1/}}{n^{1/2}}$. Soit alors $c_n(\alpha/2)$ et $c_n(1 - \alpha/2)$, respectivement les quantiles d'ordre $\alpha/2$ et $1 - \alpha/2$ de cette distribution. Alors l'intervalle de confiance :

$$[T_n - c_n(1 - \alpha/2) S_n/\sqrt{n}; \; T_n - c_n(\alpha/2) S_n/\sqrt{n}]$$

est un intervalle de confiance pour θ correct au second ordre.

Cas des fractiles :

La procédure est similaire. Les calculs des variances estimées sont donnés par (3).

Exemple 2. Cas d'un sondage stratifié

Les procédures d'évaluation sont similaires à ce qui précède, mais afin de tenir compte de la possible corrélation au sein de chaque strate, on élimine successivement non pas $k_n = n - b_n$ individus de façon aléatoire, mais des groupes d'individus dans chaque strate. Supposons que l'on ait H strates indexées par $h = 1, ..., H$ de taille respective N_h et que l'on ait sélectionné dans chaque strate n_h clusters (typiquement des communes de tailles quasi-identiques dans le panel Secodip) où n_h est suffisamment grand.

Soit n_{hi} le nombre d'unités tirées dans le groupe i, un individu j caractérisé par le triplet (h, i, j) est affecté d'un poids $w_{h,i,j}$. La distribution estimée d'une variable X sur la population observée est maintenant :

$$P_N = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{n_{hi}} w_{h,i,j} \delta_{x_{h,i,j}} / \sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{n_{hi}} w_{h,i,j}$$

qui permet de calculer les caractéristiques usuelles.

Soit maintenant $b_h < n_h$, $h = 1, ..., H$. Pour constituer les sous-échantillons, on retient à présent de manière aléatoire seulement b_h clusters, au lieu de n_h dans chaque strate, et on construit la distribution des individus ainsi obtenus :

$$P_q = \left(\sum_{h=1}^{H} \sum_{k=1}^{b_h} \sum_{j=1}^{n_{h,i_k}} w_{h,i_k,j} \delta_{x_{h,i_k,j}} \right) / \sum_{h=1}^{H} \sum_{k=1}^{b_h} \sum_{j=1}^{n_{h,i_k}} w_{h,i_k,j} .$$

En répétant cette opération $B_n = \Pi b_h^2$ fois, on obtient une fois de plus B_n valeurs de la statistique sur des sous-échantillons possédant une structure identique à la structure de l'échantillon de départ. L'extrapolation avec un taux convenable (cf. exemple 1) permet une fois de plus de construire des intervalles de confiance corrects au second ordre.

COMPARAISON PRATIQUE DES DIFFÉRENTES MÉTHODES

Afin de donner quelques ordres de grandeur sur l'intérêt et les performances relatives des méthodes étudiées précédemment, nous comparons dans les tableaux ci-dessous les intervalles de confiance obtenus par les trois méthodes, d'abord sur des produits dont la consommation est

très faible, puis sur des produits plus fortement consommés. Il est clair que dans les deux premiers cas la méthode du bootstrap pondéré apporte une très nette amélioration de la précision alors que dans les suivants, les intervalles de confiance asymptotiques et bootstrap donnent des ordres de grandeurs similaires. Après de nombreuses simulations, nous avons choisi de faire figurer dans l'annuaire (Bertail *et al.*, 1995, 1996, 1997, 1998) les intervalles de confiance bootstrap au lieu des intervalles asymptotiques pour les produits consommés par moins de 500 ménages.

Tous les calculs ont été effectués en Splus 2001, sur une station biprocesseur, avec 512 Mo de mémoire (dans une situation optimale d'utilisateur unique). Les procédures sont décrites en annexe 3. Pour donner des ordres de grandeurs des temps de calculs, pour le tableau 1, la procédure *asymptotic* (calcul des intervalles de confiance asymptotiques) s'exécute en 1 à 2 s, tandis que les procédures *wboot* (bootstrap pondéré) et *extrapol* (extrapolation de distribution de sous-échantillonnage) mettent en moyenne respectivement 10mn32s et 2mn10s. Il est donc clair, au vu de ces comparaisons, que la procédure asymptotique est la plus efficace en termes de temps de calcul, mais nous voyons dans les tableaux suivants que les résultats qu'elle donne ne sont pas très satisfaisants lorsque les tailles des populations de consommateurs effectifs sont très petites. Dans ce cas, c'est la procédure bootstrap pondéré qui donne les résultats les plus satisfaisants, mais au prix d'une longue attente. Ces mauvaises performances s'expliquent par la lourdeur du calcul mais aussi par le fait que le langage Splus, parfaitement adapté au calcul statistique, se prête mal aux boucles (à cause d'une vectorisation systématique des boucles). La procédure *extrapol* se présente comme une alternative moyenne, approximativement équivalente au bootstrap pondéré pour des tailles d'échantillons moyennes ; elle est nettement moins intéressante pour des petits échantillons.

Nous donnons ci-dessous quelques éléments de comparaison sur trois types de produits, les biscottes, le mouton par quartier entier et les maquereaux frais (le premier étant un produit de consommation courante du panel Secodip P1, le second étant très faiblement consommé et le troisième faiblement consommé, ces deux produits étant extraits du panel Secodip P2).

La partie entre « Cons. Tot./Pop. Tot » et « % consommateur » concerne l'ensemble de la population. Fract α % désigne donc le fractile d'ordre α dans l'ensemble de la population (non-consommateurs compris). S'il y a très peu de consommateurs, comme c'est le cas pour le mouton, tous ces fractiles peuvent être égaux à zéro. Les résultats suivants ne concernent que la population consommant effectivement le produit. Nous donnons la valeur de l'estimateur (Est.) et les intervalles de confiance bilatéraux à 95 % de la forme [Wb2,5 %, Wb97,5 %] pour le bootstrap pondéré, [Ext2,5 %, Ext97,5 %] pour l'extrapolation, et [As2,5 %, As97,5 %] pour l'asymptotique.

Tableau 1. **Comparaison des intervalles de confiance pour la consommation de biscottes : 1994, données Secodip**

	Est.	Wb2,5%	Wb97,5%	Ext2,5%	Ext97,5%	As2,5%	As97,5%
Cons.Tot./Pop.Tot	1,404	1,349	1,460	1,336	1,465	1,328	1,481
Fract 50%	0,596	0,550	0,634	0,509	0,691	0,465	0,727
Fract 75%	1,629	1,505	1,760	1,425	1,836	1,327	1,931
Fract 90%	3,625	3,343	3,840	3,224	4,002	3,038	4,212
Fract 95%	5,505	5,174	5,946	4,919	6,028	4,682	6,328
Fract 97,5%	8,160	7,355	8,875	7,125	8,916	6,973	9,347
% Consommateurs	80,4	79,4	81,4	79,1	81,7	78,8	82,0
Cons.Conso/Pop.Conso	1,747	1,681	1,813	1,663	1,826	1,645	1,848
Fract 25% Conso	0,375	0,336	0,400	0,308	0,449	0,272	0,478
Fract 50%	0,862	0,796	0,923	0,753	0,969	0,694	1,030
Fract 75%	2,025	1,911	2,158	1,799	2,250	1,667	2,383
Fract 90%	4,176	3,840	4,503	3,705	4,597	3,524	4,828
Fract 95%	6,412	5,642	6,828	5,680	7,070	5,425	7,399
Fract 97,5%	9,070	8,177	9,967	7,921	9,878	7,757	10,383

Tableau 2. **Comparaison des intervalles de confiance pour la consommation de mouton par quartier entier : 1994, données Secodip**

	Est.	Wb2,5%	Wb97,5%	Ext2,5%	Ext97,5%	As2,5%	As97,5%
Cons.Tot./Pop.Tot	0,051	0,038	0,064	0,028	0,068	0,030	0,071
Fract 50%	0,000	0,000	0,000	0,000	0,000	0,000	0,000
Fract 75%	0,000	0,000	0,000	0,000	0,000	0,000	0,000
Fract 90%	0,000	0,000	0,000	0,000	0,000	0,000	0,000
Fract 95%	0,000	0,000	0,000	0,000	0,000	0,000	0,000
Fract 97.5%	0,000	0,000	0,000	0,000	0,000	0,000	0,000
% Consommateurs	1,2	0,9	1,6	0,8	1,5	0,8	1,7
Cons.Conso/Pop.Conso	4,085	3,567	4,602	0,000	11,408	0,000	11,639
Fract 25% Conso	2,362	1,950	3,500	0,631	3,933	0,000	4,915
Fract 50%	3,656	3,250	3,825	2,097	5,223	1,241	6,071
Fract 75%	4,767	3,667	5,500	2,376	7,215	1,096	8,438
Fract 90%	6,000	4,875	8,500	2,918	8,886	1,524	10,476
Fract 95%	8,500	5,500	10,800	1,145	14,137	0,000	17,393
Fract 97,5%	9,000	6,500	12,000	2,245	14,274	0,841	17,159

Tableau 3. **Comparaison des intervalles de confiance pour la consommation de maquereaux frais: 1994, données Secodip**

	Est.	Wb2,5%	Wb97,5%	Ext2,5%	Ext97,5%	As2,5%	As97,5%
Cons.Tot./Pop.Tot	0,103	0,093	0,112	0,089	0,115	0,088	0,118
Fract 50%	0,000	0,000	0,000	0,000	0,000	0,000	0,000
Fract 75%	0,000	0,000	0,000	0,000	0,000	0,000	0,000
Fract 90%	0,238	0,185	0,289	0,000	0,000	0,201	0,275
Fract 95%	0,595	0,532	0,723	0,472	0,653	0,506	0,684
Fract 97,5%	1,148	1,000	1,317	0,745	1,267	0,940	1,356
% Consommateurs	14,2	13,2	15,1	13,0	15,3	12,7	15,6
Cons.Conso/Pop.Conso	0,724	0,667	0,782	0,568	0,872	0,522	0,927
Fract 25% Conso	0,208	0,183	0,244	0,120	0,305	0,062	0,354
Fract 50%	0,445	0,362	0,500	0,321	0,554	0,263	0,627
Fract 75%	0,920	0,737	1,000	0,717	1,115	0,631	1,209
Fract 90%	1,600	1,390	1,850	1,288	1,876	1,187	2,013
Fract 95%	2,261	2,000	2,530	1,732	2,653	1,661	2,861
Fract 97,5%	2,975	2,500	3,291	2,137	3,476	2,202	3,748

CONCLUSION

Cet article présente et compare plusieurs méthodes statistiques de construction d'intervalles de confiance pour des fonctionnelles non linéaires dans le cadre de sondages pouvant être complexes. Ces méthodes ont été mises en oeuvre systématiquement depuis plusieurs années pour construire des intervalles de confiance permettant de décrire précisément la consommation à partir du panel Secodip. La méthode asymptotique, basée sur la linéarisation des fonctionnelles (méthode delta), s'avère souvent mauvaise pour des échantillons de taille moyenne. Par ailleurs, l'utilisation du bootstrap naïf d'Efron n'est pas valide dans le cadre des sondages. Nous présentons deux méthodes permettant de généraliser les résultats du bootstrap à ce cadre: le bootstrap pondéré et les méthodes d'extrapolation de distribution de sous-échantillonnage. Ces méthodes peuvent aisément se généraliser à des fonctionnelles et des modèles plus complexes que ceux utilisés ici. Avant de poursuivre, il convient de noter que l'approche économétrique est très différente de l'approche sondage, les « variables » n'étant en elles-mêmes jamais aléatoires dans un sondage, l'aléa ne venant que du tirage de l'échantillon. Il est cependant possible d'introduire un modèle économétrique dans le cadre de sondage, en considérant ce dernier comme un sur-modèle auxiliaire (on parle de manière incorrect de modèle bayésien, dans la mesure

où l'on peut voir les « variables » non observées comme des paramètres à estimer pouvant être prédits par le sur-modèle). L'effet d'un sondage sur un modèle linéaire même simple peut se révéler catastrophique, si l'on ne sait pas comment il a été réalisé et en particulier, pour un sondage stratifié, si l'on ne connaît pas les variables utilisées pour la stratification (voir, par exemple, Gouriéroux, 1987). Lorsque les variables ayant servi au plan de sondage sont introduites par défaut dans le modèle économétrique, il n'y a généralement pas lieu de tenir compte du plan de sondage (ceci dépend bien évidemment de la forme du modèle et des hypothèses retenues). Une telle pratique n'est possible que si l'on sait comment le sondage a été réalisé (et redressé) et si la variable explicative n'a pas elle-même servi à l'élaboration du plan de sondage, auquel cas il y a un problème de sélection endogène. Il convient alors de redresser, lorsque c'est possible, l'effet de cette sélection par des méthodes appropriées. Une solution envisageable est l'utilisation de méthodes de type vraisemblance empirique (voir l'excellent ouvrage d'Owen, 2001) que l'on peut considérer comme une généralisation des méthodes de calage sur marge, employées en sondage pour tenir compte de l'information externe (Deville et Särndal, 1992). L'idée des vraisemblances empiriques est d'incorporer non seulement les contraintes du sondage et l'information extérieure (par exemple, les valeurs de marges déterministes sur des caractères observés de manière exhaustive), mais aussi les contraintes induites par le modèle. On cherche alors à pondérer les observations par des poids proches des poids du plan de sondage (pour une certaine métrique, distance de Kullback ou distance du χ^2) qui vont réaliser simultanément les contraintes d'informations et les contraintes du modèle (Bertail, 2002). L'utilisation de techniques de type bootstrap pondéré et/ou d'extrapolation par strates peut alors s'avérer intéressante dans ce cadre et devrait faire l'objet de développements ultérieurs.

BIBLIOGRAPHIE

Barbe P., Bertail P. (1995). *The Weighted Bootstrap. Lecture Notes in Statistics*, New-York, Springer Verlag.

Bertail P. (2002). Empirical likelihood in some semi-parametric models, document de travail CREST, soumis *Bernoulli*.

Bertail P. (1997). Second order properties of an extrapolated bootstrap without replacement under weak assumptions : the i.i.d. and strong mixing case, *Bernoulli*, 3, pp. 149-179.

Bertail P., Boizot C. et Combris P. (1995) (1996) (1997) (1998a). *La consommation alimentaire : distribution des quantités consommées à domicile*, Observatoire des consommations alimentaires, 270 p.

Bertail P., Boizot C., Combris P., Hébel P. et Volatier J.-L. (1998b). CD-ROM de consultation de données de consommation, rapport final pour le ministère chargé de la Recherche, et CD-ROM, *La consommation alimentaire en France*, CREDOC et INRA-CORELA.

Bertail P., Combris P. (1997). Bootstrap généralisé d'un sondage, application à l'estimation de distributions de consommations alimentaires, *Annales d'Économie et de Statistiques*, 46, pp. 50-83.

Bertail P., Politis D.N. (2001). Extrapolation of subsampling distribution estimators : the i.i.d. and strong mixing cases, *Canadian Journal of Statistics*, 29, 4, pp. 667-680.

Bertail P., Politis D.N. et Romano J.-P. (1999). On subsampling estimators with unknown rate of convergence, *Journal of the American Statistical Association*, 94, 446, pp. 569-579.

Bickel P.J., Yahav J.A. (1988). Richardson extrapolation and the bootstrap, *Journal of the American Statistical Association*, 83, pp. 387-393.

Bosq D., Lecoutre J.-P. (1987). *Théorie de l'estimation fonctionnelle*, Paris, Economica.

Carlstein E. (1986). The use of subseries values for estimating the variance of a general statistics from a stationary sequence, *The Annals of Statistics*, 14, pp. 1171-1179.

Deville J.-C. (1987). Replications d'échantillons, demi-échantillons, jackknife, bootstrap dans les sondages, *in : Les sondages*, Droesbeke J.J., Tassi B. et Fichet P. (eds), Paris, Economica.

Deville J.-C., Särndal C.E. (1992). Calibration estimators in survey sampling, *Journal of the American Statistical Association*, 87, pp. 376-382.

Dudley R.M. (1994). The order of the remainder in derivatives of composition and inverse operators, *The Annals of Statistics*, 22, pp. 1-20.

Green W.H. (2000). *Econometric Analysis*, 4rd ed., London, Prentice Hall International Inc.

Gourieroux C. (1987). Effets d'un sondage : cas du χ^2et de la régression, *in : Les sondages*, Droesbeke J.J., Tassi B. et Fichet P. (eds), Paris, Economica.

Gourieroux C. (1981). *Théorie des sondages*, Paris, Economica.

Hampel F. (1974). The influence curve and its role in robust estimation, *Journal of the American Statistical Association*, 69, pp. 383-393.

Huber P.J. (1981). *Robust Statistics*, New-York, Wiley.

Künsch H.R. (1989). The jackknife and the bootstrap for general stationary observations, *The Annals of Statistics*, 17, pp. 1217-1241.

Lahiri S.N. (1992). Edgeworth correction by "moving block" bootstrap for stationary and non stationary data, *in* : *Exploring the Limits of the Bootstrap*, Ed. Le Page R., Billard L. (eds), New York, John Wiley.

Letson D., McCullough B.D. (1998). Better confidence intervals : the double bootstrap with no pivot, *American Journal of Agricultural Economics*, vol. 80, pp. 552-559.

Liu R., Singh K. (1992). Moving blocks jackknife and bootstrap capture weak dependence, *in* : *Exploring the Limits of the Bootstrap*, Ed. Le Page R., Billard L. (eds), New York, John Wiley.

Manski C. (1988). *Analog Estimation Methods in Econometrics*, London, Chapman and Hall.

Mises R. von (1936). Les lois de probabilités pour les fonctions statistiques, *Annales de l'Institut Hubert Poincaré*, 6, pp. 185-212.

Owen A.B. (2001). *Empirical Likelihood*, London, Chapman and Hall/ CRC.

Politis D.N., Romano J.P. (1994). A general theory for large sample confidence regions based on subsamples under minimal assumptions, *Annals of Statistics*, 20, pp. 2031-2050.

Rosen P. (1972). Asymptotic theory for successive sampling, *Annals of Mathematical Statistics*, 43, pp. 373-397.

Sen P.K. (1988). Asymptotics in finite sampling, *in : Handbook of Statistics*, Krishnaiah P.R., Rao C.R. (eds), vol. 6, pp. 291-331.

Shao J., Tu D. (1996). *The Jackknife and Bootstrap*, New York, Springer-Verlag.

Shao J., Wu C.F.J. (1989). A general theory for jackknife variance estimation, *Annals of Statistics*, 17, pp. 1176-1197.

Van der Vaart A.W. (1998). Asymptotic Statistics, *Cambridge series in Statistical and Probabilistic Mathematics*.

Wu C.F.J (1990). On the asymptotic properties of the jackknife histogram, *Annals of Statistics*, 18, pp. 1438-1452.

ANNEXE 1

Un exemple de calcul du gradient d'ordre 1

En tout point P la fonctionnelle $H_P^{-1}(\alpha)$ vérifie

$$H_P(H_P^{-1}(\alpha)) = \frac{E_P(N\ 1\{\frac{C}{N} < H_P^{-1}(\alpha)\})}{E_P N} = \alpha.$$

On en déduit que pour tout $t \in [0,1]$, la fonctionnelle prise en la contaminée $(1-t)\ P + t\delta_{(c,\ n)}$, satisfait

$$E_{(1-t)P+t\delta_{(c,n)}}(N\ 1\{\frac{C}{N} < H^{-1}_{(1-t)P+t\delta_{(c,n)}}(\alpha)\}) = \alpha E_{(1-t)P+t\delta_{(c,n)}}(N).$$

On en déduit que

$$(1-t)E_P(N\ 1\{\frac{C}{N} < H_P^{-1}(\alpha)\}) + tn1\{\frac{c}{n} < H^{-1}_{(1-t)P+t\delta_{(c,n)}}(\alpha)\} = \alpha(1-t)E_P(N) + \alpha tn$$

et par dérivation en $t = 0$:

$$\begin{aligned} &E_P(N\ 1\{\tfrac{C}{N} < H_P^{-1}(\alpha)\}) + n1\{\tfrac{c}{n} < H_P^{-1}(\alpha)\} \\ &+E_P N\ \frac{\partial\ 1\{\frac{C}{N} < H^{-1}_{(1-t)P+t\delta_{(c,n)}}(\alpha)\}}{\partial t}\bigg|_{t=0} = \alpha(n - E_P N) \end{aligned} \quad , \tag{1}$$

mais, on a par un calcul formel sur les distributions

$$D_\alpha = \frac{\partial\ 1\{\frac{C}{N} < H^{-1}_{(1-t)P+t\delta_{(c,n)}}(\alpha)\}}{\partial t}\bigg|_{t=0} = -\delta_{\frac{C}{N}}(H_P^{-1}(\alpha))\ F_\alpha^{(1)}(c,n,P), \tag{2}$$

d'où

$$\begin{aligned} E_P N\ \delta_{\frac{C}{N}}(H_P^{-1}(\alpha)) &= \int n\delta_{\frac{c}{n}}(H_P^{-1}(\alpha))\ f_{(C,N)}(c,n)\ dc\ dn \\ &= \int n\ f_{(C,N)}(nH_P^{-1}(\alpha), n)\ dn \end{aligned} \tag{3}$$

On déduit de (1), (2) et (3) que

$$F_\alpha^{(1)}(c,n,P) = \frac{n\left(1\{\frac{c}{n} \le H_P^{-1}(\alpha)\} - \alpha\right)}{\int n\ f_{(C,N)}(nH_P^{-1}(\alpha), n)dn}.$$

ANNEXE 2

Une expression explicite de l'estimateur de la variance de $H_P^{-1}(\alpha)$

Un estimateur de $f_{(C,\,N)}(u, v)$ est donné par $\hat{f}_q(u, v)$ en (2). On a alors par un calcul direct

$$D \doteq \int n\, \hat{f}_q(nH_P^{-1}(\alpha), n)dn = h_q^{-2} \sum_{i=1}^{q} p_i\, I_i,$$

avec

$$I_i \doteq (2\pi)^{-1} \int u \exp\left(-\tfrac{(uH_P^{-1}(\alpha)-c_i)^2+(u-n_i)^2}{2h_q^2}\right) =$$

$$(2\pi)^{-1} \int u \exp\left(-\tfrac{(H_P^{-1}(\alpha)^2+1)}{2h_q^2}\left(u^2 - 2\tfrac{(c_iH_P^{-1}(\alpha)+n_i)}{(H_P^{-1}(\alpha)^2+1)}u + \tfrac{c_i^2+n_i^2}{H_P^{-1}(\alpha)^2+1}\right)\right) du$$

$$= (2\pi)^{-1/2}\sigma_q \int (2\pi\sigma_q^2)^{-1/2} u \exp(-\tfrac{(u-m_i)^2}{2\sigma_q^2}) \exp(-\delta_i)$$
$$= (2\pi)^{-1/2}\sigma_q\, m_i \exp(-\delta_i),$$

avec les notations

$$\sigma_q^2 = \frac{h_q^2}{H_P^{-1}(\alpha)^2 + 1},$$

$$\delta_i = \tfrac{c_i^2+n_i^2}{h_q^2} - \tfrac{(c_iH_P^{-1}(\alpha)+n_i)^2}{h_q^2(H_P^{-1}(\alpha)^2+1)}$$

$$= (2h_q^2)^{-1}\tfrac{(H_P^{-1}(\alpha)n_i-c_i)^2}{(H_P^{-1}(\alpha)^2+1)}$$

et

$$m_i = \frac{c_iH_P^{-1}(\alpha) + n_i}{H_P^{-1}(\alpha)^2 + 1}.$$

On en déduit que

$$D = h_q^{-2} \textstyle\sum_{i=1}^{q} p_iI_i$$

$$= (2\pi)^{-1/2}\, h_q^{-2} \tfrac{h_q}{(H_P^{-1}(\alpha)^2+1)^{1/2}} \textstyle\sum_{i=1}^{q} p_i \tfrac{c_iH_P^{-1}(\alpha)+n_i}{H_P^{-1}(\alpha)^2+1} \exp(\left(-(2h_q^2)^{-1}\tfrac{(H_P^{-1}(\alpha)n_i-c_i)^2}{(H_P^{-1}(\alpha)^2+1)}\right),$$

d'où le résultat annoncé.

ANNEXE 3

Description des programmes Splus

Les méthodes décrites ici étant utilisées par plusieurs organismes de recherche, nous ne décrivons que succinctement les procédures Splus utilisées. Ces procédures sont disponibles auprès du laboratoire de Recherche sur la consommation, INRA-CORELA, Ivry-sur-Seine.

Intervalles de confiance asymptotiques

Le programme *asymptotic* permet de calculer les intervalles de confiance pour des moyennes de consommations totales ou non nulles et pour des fractiles à partir des données du panel Secodip. Les paramètres du programme en sont *quant*, la matrice des quantités, *code*, le code du (des) produit(s) sélectionné(s), *fichpoids*, le fichier contenant les caractéristiques du ménage, à savoir le poids attribué après calage sur marge, le nombre de semaines d'activité (utilisé pour le redressement annuel) ainsi que le nombre de personnes dans le ménage, *cc*, un paramètre de lissage utilisé dans le calcul des intervalles de confiance des fractiles par la méthode du noyau. Ce paramètre, par défaut égal à 1, permet de tester la robustesse des résultats lorsque les échantillons sont de tailles très petites ; la plupart du temps, il n'est pas nécessaire de le modifier. L'appel du programme se fait de la manière suivante : *asymptotic(quant, code, fichpoids, cc)* ou *asymptotic(quant, code, fichpoids)*. Les quantités sont d'abord corrigées par une simple règle de trois pour rapporter la consommation Secodip annuelle du ménage à une activité de 52 semaines ou de 13 périodes selon les années. La procédure fait essentiellement appel à trois sous-programmes qui peuvent être utilisés indépendamment : *moyb*, calcul de la moyenne pondérée des consommations individuelles, ayant pour paramètres *fich*, le fichier des consommations redressées, *nbp*, le nombre de personnes dans le ménage, et *poids* la pondération du ménage (la somme des poids est rapportée à 1). Le programme *sbp* calcule la variance associée à *moyb* avec les mêmes paramètres en utilisant la formule (1). Les fractiles associés et leur écart-type sont calculés grâce à la procédure *smpond* qui utilise la relation (3). La constante de lissage h_q est choisie de la forme $cc*q^{-1/6}$. Le reste du programme affiche les résultats sous forme d'un tableau donnant l'estimation et l'intervalle de confiance à 95% associé, en utilisant la formule standard « plus ou moins 1,96 fois l'écart-type ».

Intervalles de confiance par bootstrap pondéré

Le programme *wboot* utilise les estimateurs asymptotiquement convergents étudiés dans la première partie. Il s'utilise sous la forme *wboot*(*nboot, quant, code, fichpoids*), où *nboot* est le nombre de ré-échantillonnages utilisés (en général supérieur à 1 000). Les autres paramètres sont identiques à *asymptotic*. Le programme se présente sous la forme d'une boucle sur le nombre de rééchantillonnages. À chaque itération est généré, grâce à la procédure *mgs*, un nouveau système de poids calculé en fonction du système original. Puis les statistiques asymptotiques sont recalculées avec ce nouveau système de poids, en faisant appel aux programmes *moyb*, *sbp*, *smpond*. Les quantités sont toutes standardisées

par leur écart-type (cette méthode est connue dans la littérature sur le bootstrap comme méthode *t*-percentile). La collection des *nboot* valeurs obtenues permet de construire une distribution dont les quantiles d'ordre 97,5 et 2,5 sont obtenus grâce à la procédure *fractile*. Ces valeurs sont alors utilisées à la place du traditionnel +-1,96 pour construire l'intervalle de confiance sur la base de l'estimateur initial de la statistique et de son écart-type. La forme finale des résultats est identique à celle d'*asymptotic*, dont les résultats sont aussi fournis comme des éléments de comparaison.

Intervalles de confiance par extrapolation de distributions de sous-échantillonnage

La fonction *extrapol* qui met en œuvre la méthode de construction d'intervalles de confiance par extrapolation se présente sous la forme :

extrapol(quant,code,fichpoids,nboot,sousech,cc), où les paramètres *quant, code, fichpoids* sont, comme dans les procédures précédentes, respectivement la matrice des quantités, le code du produit et le fichier des caractéristiques des ménages (poids, nombre de semaines d'activité, nombre de personnes dans le ménage et éventuellement la région et la commune pour tenir compte de la stratification). *nboot* est le nombre de répétitions de la procédure de ré-échantillonnage. Lorsque l'on tient compte du facteur correctif de population finie, la taille optimale du sous-échantillonnage pour des fonctions de moments est de la forme $sousech*q^{2/3}$. Le paramètre *sousech* permet donc de moduler la taille du sous-échantillonnage dans une optique de robustesse. Le paramètre *cc* est lui aussi utilisé dans cette perspective lors du lissage des fractiles (cf. la procédure *asymptotic*). Les paramètres *nboot*, *sousech* et *cc* sont optionnels et peuvent être omis, la procédure peut donc être utilisée directement sous la forme *extrapol(quant, code, fichpoids)*, auquel cas les valeurs par défaut des paramètres optionnels sont *nboot* = 1000, *sousech* = 1 et *cc* = 1. La forme finale des résultats (estimation et intervalles de confiance) est similaire à celle de *asymptotic* et *wboot*.