Composante du produit nº 12-001-X
Division des méthodes d'enquêtes auprès des entreprises

Article

Estimation de variance pour des statistiques et des estimateurs complexes : linéarisation et techniques des résidus

par Jean-Claude Deville

Décembre 1999

Statistique Canada Statistics Canada

Canada

Techniques d'enquête, décembre 1999
Vol. 25, N° 2, pp. 219-230
Statistique Canada, N° 12-001 au catalogue

Estimation de variance pour des statistiques et des estimateurs complexes : linéarisation et techniques des résidus

Jean-Claude Deville [1]

Résumé

Dans une enquête par sondage, en l'absence d'information externe, on estime le total d'une variable à l'aide de l'estimateur de Horvitz-Thompson. Sa variance est elle même estimée en calculant une forme quadratique assez compliquée, généralement de façon récursive. Dans cet article on suppose ce problème résolu par exemple du fait de l'existence d'un logiciel capable de réaliser automatiquement le calcul. Pour le cas des estimateurs complexes (en fait par calage), ainsi que pour celui des statistiques non linéaires (estimateurs par substitution), on montre qu'on peut toujours utiliser le même outil de calcul à condition d'utiliser une variable artificielle convenablement choisie. Dans tous les cas, cette variable artificielle fournit une estimation de variance approximativement sans biais et est construite en utilisant la technique des fonctions d'influence et quelques postulats asymptotiques. On donne de multiple exemples d'utilisation de cette technique : fonctions compliquées mais explicites de totaux (coefficient de corrélation), fonctions implicites de totaux, estimateurs calés, fractiles et statistiques de rang, statistiques provenant de méthodes factorielles.

Mots clés : Estimation de variance; statistiques complexes; linéarisation; estimateurs par substitution; technique des résidus; fonction d'influence; paramètres implicites; fractiles; statistiques de rang; analyse factorielle.

1. Introduction

La formulation de résultats sous la forme d'intervalles de confiance est le but, (rarement atteint), de toutes les enquêtes par sondage. La procédure la plus habituelle est d'estimer la variance des statistiques mises en cause pour la loi de probabilité induite par le schéma d'échantillonnage (et parfois, pour simplifier, sous des hypothèses assez brutales portant le nom de modèles). Ensuite, sous l'hypothèse, rarement démentie par les faits quand les échantillons sont assez gros, que la statistique suit une loi de Gauss, on déduit selon des procédures simples et standardisées un intervalle de confiance symétrique autour de l'estimation ponctuelle.

Ce problème a été l'objet d'une abondante littérature, avant et après l'ouvrage de référence que constitue le livre de Wolter (1985).

Le but de cet article est de montrer comment des outils simples permettent de réaliser effectivement une estimation de la variance dans des cas complexes en utilisant une technique unique : la linéarisation. On indiquera, en préliminaire, l'état de l'art en ce qui concerne l'estimation de la variance de l'estimateur de Horvitz-Thompson pour un total. Après avoir introduit une définition de « statistique linéarisable » puis la notion de fonction d'influence analogue à celle qu'on utilise en statistique non paramétrique, on introduira la classe des estimateurs par substitution d'une fonctionnelle dont on montrera qu'ils sont linéarisables sous des hypothèses assez générales. On montrera comment les règles habituelles du calcul différentiel s'étendent aux variables linéarisées et permettent, par des procédures pas à pas, de calculer assez facilement les variables linéarisées de statistiques assez complexes. On étudiera avec une attention particulière les statistiques utilisant des quantiles ainsi que celles liées à l'analyse multivariée la plus courante.

Cette démarche est le fil directeur du logiciel POULPE utilisé à l'INSEE :

- Avoir un outil de calcul pour l'estimation de variance d'un total estimé par l'estimateur par expansion simple.
- Se ramener à ce cas, en utilisant des variables construites à dessein, quand on utilise un estimateur complexe et/ou qu'on estime une statistique complexe.

2. Cadre général : l'estimateur par dilatation simple

On considère une population U d'unités $k, l, ...,$ sur laquelle est défini un plan de sondage c'est-à-dire une loi de probabilité p associant à toute partie s de U – l'échantillon – une probabilité $p(s)$ d'être sélectionnée. A partir de celle-ci, on définit les probabilités d'inclusion π_k ($\pi_k = \sum_{s \ni k} p(s)$) et les probabilités d'inclusions à l'ordre 2, π_{kl} des éléments de s. On peut alors utiliser l'estimateur d'Horvitz-Thompson, $\hat{Y} = \sum_{k \in s} y_k/\pi_k$ du total Y d'une variable d'intérêt y. Il a l'avantage d'être (presque) toujours disponible, d'être sans biais et d'avoir une variance facile à calculer :

$$\mathrm{Var}\,(\hat{Y}) = \sum_U \pi_k(1-\pi_k)\left(\frac{y_k}{\pi_k}\right)^2 + 2\sum\sum_U (\pi_{kl} - \pi_k\pi_l)\frac{y_k}{\pi_k}\frac{y_l}{\pi_l} \qquad (2.1)$$

où la seconde somme est étendue à tous les couples (k, l) de la population U.

1. Jean-Claude Deville, Laboratoire de Statistique d'Enquête, Insee/Ensai/Crest, Campus de Ker-Lann, 35170-Bruz (France).

Un estimateur commode de la variance de $\hat{Y}$ est donné par :

$$\widehat{\mathrm{Var}}(\hat{Y}) = \sum_s (1-\pi_k)\left(\frac{y_k}{\pi_k}\right)^2 + 2\sum\sum_s \left(1-\frac{\pi_k\pi_l}{\pi_{kl}}\right)\frac{y_k}{\pi_k}\frac{y_l}{\pi_l}. \qquad (2.2)$$

En pratique, pour les gros échantillons, on calculera cet estimateur de variance de façon récursive car la somme double qui apparaît possède un nombre de termes prohibitif. De plus, les probabilités d'inclusion π_{kl} ne sont calculables facilement que dans de rares cas simples.

En fait, tous les plans de sondage connus se ramènent à quelques schémas simples : sondage Bernoullien ou Poissonnien, sondage aléatoire simple, sondage systématique et sondage à probabilités inégales de taille fixe. Pour les premiers il existe des formules closes donnant l'estimation de variance à partir de sommes de carrés. Il en va de même pour le sondage systématique si on fait quelques hypothèses faciles à vérifier sur l'ordre du tirage. Enfin la variance d'un sondage à probabilités inégales de taille fixe, pour de nombreuses méthodes de tirage, peut être approximée de façon extrêmement fine et générale par la formule suivante, utilisée dans POULPE (Deville 1993) :

$$\hat{V} = \frac{1}{1-\sum a_k^2}\sum_s (1-\pi_k)\left(\frac{y_k}{\pi_k} - A\right)^2 \qquad (2.3)$$

$$\text{où } a_k = \frac{1-\pi_k}{\sum_s (1-\pi_k)} \quad \text{et} \quad A = \sum_s a_k \frac{y_k}{\pi_k}.$$

Ces schémas simples se combinent pour donner des plans arbitrairement compliqués à partir de deux opérations qui sont la stratification et le sondage multidegré (ou sous-sondage).

Pour la première, l'estimation de la variance du total général s'obtient en additionnant les variances des estimateurs des totaux de strate.

Un sondage multidegré s'obtient quand la population est partagée en sous populations U_i nommées « unités primaires ». On sélectionne un échantillon s_1 de celles-ci selon un plan de sondage p_1 sur la population des unités primaires. Ensuite, dans chaque $U_i (i \in s_1)$, on sélectionne un échantillon selon un plan p_i/s_1. Conditionnellement à s_1, ces plans sont indépendants. On en déduit les probabilités d'inclusion et la formule de variance suivante :

$$\mathrm{Var}(\hat{Y}) = \mathrm{Var}\left(\sum_{s_1}\frac{Y_i}{\pi_i}\right) + E\left(\sum_{s_1}\frac{\mathrm{Var}(\hat{Y}_i)}{\pi_i^2}\middle/ s_1\right) \qquad (2.4)$$

ou Y_i est le total de y sur U_i, π_i sa probabilité d'inclusion, $\hat{Y}_i$ l'estimateur de Y_i dans le plan p_{i/s_1}. Enfin, si $V(Y_i;\ i \in s_1)$ est l'estimateur de variance de $\sum_{s_1} Y_i/\pi_i$ et que V_i est un estimateur de la variance de $\hat{Y}_i$ conditionnellement à s_1, alors

$$\widehat{\mathrm{Var}}(\hat{Y}) = V(\hat{Y}_i;\ i \in s_1) + \sum_{i\in s_1}\frac{V_i}{\pi_i} \qquad (2.5)$$

est un estimateur de la variance de $\hat{Y}$ (Durbin 1953).

Naturellement, dans chaque strate h ou dans chaque unité primaire i, le sondage p_h ou p_i/s_1 peut lui-même être stratifié ou être un sondage multidegré. Cependant, dans tous les cas, l'utilisation répétitive et récursive des règles données ci-dessus permet un calcul de l'estimateur de variance à partir d'éléments simples basés sur des sommes de carrés. Dans la pratique des enquêtes auprès des personnes, il est fréquent qu'un plan de sondage comporte de trois à cinq degrés de tirage.

Ceci signifie que la forme quadratique (2.2) est calculable de façon mécanique, sans toutefois qu'on calcule jamais explicitement les termes qui interviennent dans la somme double figurant dans la formule.

Pour compléter ce panorama, il arrive fréquemment qu'un échantillon soit tiré en plusieurs phases, en pratique deux ou trois. Ceci signifie qu'un échantillon s est tiré et sert de population de référence pour le tirage d'un second échantillon r, selon un plan de sondage $q(r/s)$. S'il est contrôlé par le statisticien, ce plan est généralement un plan stratifié avec sondage aléatoire simple dans chaque strate. Sinon, le plan est décrit par un modèle de réponse qui permet de formaliser une procédure de repondération pour non-réponse. Dans tous les cas interviennent des probabilités d'inclusion de deuxième phase P_k et P_{kl} décrivent l'inclusion dans r de l'unité k ou du couple (k, l). L'estimateur par expansion est $\hat{Y}_{\exp} = \sum_r y_k/\pi_k P_k$.

Sa variance se calcule assez facilement et est estimée par l'expression :

$$\widehat{\mathrm{Var}}(\hat{Y}_{\exp}) = \sum\sum_r \left(1-\frac{\pi_k\pi_l}{\pi_{kl}}\right)\frac{1}{P_{kl}}\frac{y_k}{\pi_k}\frac{y_l}{\pi_l} + \sum\sum_r \left(1-\frac{P_kP_l}{P_{kl}}\right)\frac{y_k}{P_k\pi_k}\frac{y_l}{P_l\pi_l} \qquad (2.6)$$

$$\text{avec } \pi_{kk} = \pi_k \text{ et } P_{kk} = P_k.$$

Au prix de quelques difficultés, cet estimateur de variance est calculable de façon mécanique en évitant les sommes doubles prohibitives. Il en va de même pour les plans en trois phases qui surviennent quand une phase de non réponse s'ajoute à une seconde phase contrôlée par les statisticiens. Des procédures de ce genre sont utilisées dans POULPE.

Ainsi, l'estimateur de Horvitz-Thompson (ou son extension, l'estimateur par expansion) a pour variance une forme quadratique $Q(y_k;\ k \in U)$. Celle-ci peut être estimée sans biais (ou éventuellement avec un biais négligeable) par le calcul récursif d'une forme quadratique $\hat{Q}(y_k;\ k \in s)$ (où s désigne maintenant l'échantillon définitif, quel que soit le

nombre de degrés ou de phases après lequel il a été obtenu). Dans la suite, nous supposerons disponible un moyen de calcul « automatique » de cette dernière forme quadratique.

3. Statistiques complexes et postulats asymptotiques

Nous allons voir qu'on peut l'utiliser également quand on adopte un estimateur plus raffiné que HT (intégrant des informations auxiliaires externes), ainsi, et surtout, quand on s'intéresse à des statistiques complexes (moyennes, ratios quantiles, indices complexes du genre GINI, coefficient d'un modèle économétrique, facteur d'une analyse en composantes principales par exemple).

Les résultats que nous allons donner sont des approximations « asymptotiques ». Comme dans Isaki et Fuller (1982) ou Deville et Särndal (1992), on postule le schéma suivant : dans une suite de problèmes de sondage indicés par un entier v (que nous supprimerons pour ne pas surcharger les notations), la taille N de la population tend vers l'infini de même que n taille (ou espérance de la taille) de l'échantillon. Pour chaque v, on a donc aussi un plan de sondage, l'estimateur HT associé, un vecteur de taille fixée invariable de variables x_k de X estimé par $\hat{X}$.

On postule les trois propositions suivantes :

- $N^{-1}X$ à une limite. (3.1)
- $N^{-1}(\hat{X} - X)$ converge en probabilité vers zéro. (3.2)
- $n^{-1/2}N^{-1}(\hat{X} - X)$ a pour limite une loi de Gauss multidimensionnelle. (3.3)

Le premier postulat formalise l'idée d'une suite de populations de tailles croissantes extraites d'une loi continue parente. On peut aussi interpréter cela en disant que la population est un échantillon i.i.d. d'une certaine superpopulation infinie. Les deux autres postulats concernent la convergence de l'estimateur HT et le fait qu'il donne lieu à un théorème central-limite. De façon plus concrète, ces postulats sont satisfaits dans de nombreux cas sous certaines hypothèses techniques : sondage aléatoire simple (Hájek 1964), échantillonnage poissonnien et l'échantillonnage systématique randomisé (Rosen 1972), plan stratifié dont le nombre de strates tend vers l'infini (application immédiate des conditions de Lindeberg).

En pratique, cependant, on ne saura jamais dire si, par exemple, le plan utilisé comporte un nombre de strate tendant vers l'infini ! Ce que traduisent les postulats asymptotiques, c'est simplement le fait que certaines grandeurs (techniquement celles qui seront des $O_p(n^{-1/2})$) sont considérées comme « petites » et que le produit de deux quantités « petites » sera une quantité « négligeable » (et donc négligée !).

A partir de ces postulats nous montrerons comment certains estimateurs et certaines statistiques non linéaires sont approximables par des statistiques de HT de la forme $\sum_s z_k/\pi_k$ pour des variables bien choisies z_k.

4. Fonctionnelles et estimateurs par substitution

Nous allons nous intéresser maintenant à une classe assez générale de statistiques non linéaires de la population finie, basée sur la notion de fonctionnelle d'une mesure, et à leurs estimateurs par substitution.

A chaque unité k de la population U est associé le point x_k de $\boldsymbol{R}^p$ des p variables du problème qui nous intéressent. La population U est donc représentée par la mesure M ayant une masse unité en chacun des points x_k.

Cette mesure est positive, discrète, finie et sa masse totale vaut N. On supposera tous les x_k distincts ce qui n'enlève rien à la généralité (on peut toujours ajouter une dimension qui est le « rang » de k dans une numérotation arbitraire). Pour toute variable $y_k = y(x_k)$, on aura donc $\int y\,dM = \sum_U y_k$.

Dans l'optique asymptotique, la suite de populations est une suite de mesures sur $\boldsymbol{R}^p$. Le premier postulat asymptotique dit que cette suite se comporte comme si on avait affaire à des tirages i.i.d. dans une loi de probabilité fixe sur $\boldsymbol{R}^p$.

Une fonctionnelle $T(M)$ associe à toute mesure d'une classe contenant au moins les mesures ponctuelles, un nombre réel ou un vecteur. On supposera, de plus, que toutes les fonctionnelles auxquelles on s'intéresse sont homogènes, c'est-à-dire qu'il existe un nombre réel positif α dépendant de T tel que $T(tM) = t^{\alpha}T(M)$ pour tout réel positif t. Un total est une fonctionnelle homogène de degré 1, une moyenne de degré 0, une somme à double indice de degré 2. La restriction à des fonctionnelles homogène n'est pas très gênante pour la pratique.

Soit maintenant $\hat{M}$ (estimateur de M) la mesure attribuant le poids w_k à tout point x_k pour k dans s et zéro à tout autre point, quelle que soit l'origine des poids (Horvitz-Thompson ou calage).

Définition : L'estimateur par substitution d'une fonctionnelle $T(M)$ est $T(\hat{M})$.

Dans le cas d'un total cette définition n'apporte aucune surprise car $T(\hat{M}) = \int x d\hat{M}(x) = \sum_s x_k w_k$. Pour des statistiques complexes « ordinaires » (ratios, moyennes ou indices par exemple), cela correspond aux pratiques habituelles de l'exploitation des enquêtes. Pour des statistiques de rang il en va de même, les subtilités touchant plus l'estimation de la fonction de répartition que l'estimation des fractiles (voir par exemple Chambers, Dorfmann et Hall 1992).

Une classe assez générale de paramètres liés à la population finie s'obtient par des équations implicites qui les définissent. C'est le cas, par exemple, de l'ajustement d'un modèle paramétrique au niveau de la population qui conduit à une « équation estimante » déduite d'un grand principe d'ajustement (maximisation de vraisemblance, minimisation de chi-2 ou méthodes des « moments » ou des « moments généralisés »).

Cette écriture fait apparaître le paramètre (éventuellement multidimensionnel) du modèle comme une

fonctionnelle de M. Son estimateur (au sens sondage) est la même fonctionnelle pour $\hat{M}$. Par exemple l'estimation des moindres carrés dans le modèle linéaire s'écrit :

$$B = \arg\ \mathrm{Min} \sum_U q_k\ (y_k - x_k' B)^2.$$

L'estimation – sondage de B est $\hat{B} = \arg\ \mathrm{Min} \sum_s w_k q_k\ (y_k - x_k' B)^2$. L'utilisation de $\hat{B}$ (plutôt qu'un estimateur sous modèle conditionnel à l'échantillon) est beaucoup plus robuste et permet de tenir compte correctement des fluctuations de l'échantillonnage sur le résultat (On pourra se reporter à Binder 1983, Binder et Patak 1994, ou à Binder et Kovačević 1997 sur cette question).

De façon générale, une « équation estimante » au niveau de la population s'écrira sous la forme $T(M, \lambda) = 0$ où T est une fonctionnelle de dimension p paramètrée par le vecteur λ de dimension p également. Cette équation sera supposée avoir une solution unique pour M. fixé. L'estimateur par substitution de λ est la solution de l'équation (estimante) $T(\hat{M}, \hat{\lambda}) = 0$.

5. Statistique linéarisable

On s'intéresse à des statistiques S dépendant des observations $(x_k;\ k \in s)$ (en fait à une suite de statistiques définies dans chacun des problèmes de sondage du cadre asymptotique). On dit que S est en probabilité de l'ordre de $f(n)$ (où f est une fonction positive quelconque de n) et on écrit $S = O_p(f(n))$ si pour tout $\varepsilon > 0$ il existe une constante C telle que

$$\Pr\left(\frac{\| S \|}{f(n)} \geq C \right) \leq \varepsilon.$$

Autrement dit les fonctions de survie des variables $\| S \| / f(n)$ sont uniformément majorées par la fonction de survie de graphe $(C(\varepsilon),\ \varepsilon)$.

Le troisième axiome asymptotique (Axiome central-limite !) peut donc s'écrire $N^{-1}(\hat{X} - X) = O_p(n^{-1/2})$, et le second $N^{-1}\hat{X} = O_p(1)$. On utilisera dans toute la suite et plus ou moins implicitement le résultat bien connu suivant (voir par exemple Billingsley 1969) :

Résultat : Si une statistique S converge vers une certaine loi et que $(S - T) = O_p(f(n))$ avec $f(n) \to 0$, alors T converge vers la même loi. En particulier S et T ont la même variance limite.

On dira que la statistique S est de degré α si $N^{-\alpha} S$ tend vers une limite. Il est clair par exemple qu'un estimateur de HT est de degré 1, qu'un ratio d'estimateurs de HT est de degré zéro (ou homogène). Le troisième postulat asymptotique dit que $nE(\hat{X} - X)^2$ est de degré 2. L'estimateur par substitution d'une fonctionnelle homogène de degré α est une statistique de degré α.

Nous pouvons maintenant poser la définition suivante :

Définition : Une statistique S de degré α est linéarisable s'il existe une variable synthétique z_k (dite variable linéarisée de S) telle que la variance de $\hat{Z}$ soit équivalente à celle de S en ce sens que $n^{1/2}(N^{-\alpha} S - N^{-1}\hat{Z}) = O_p(f(n))$ avec $f(n) \to 0$. On aura en général presque toujours $f(n) = n^{-1/2}$.

En pratique cela signifie qu'on estimera la variance, et donc un intervalle de confiance, pour S à partir de la variance de $\hat{Z}$ (qu'il s'agisse ou non d'un estimateur de HT).

On remarquera, par ailleurs que la définition n'implique pas l'unicité de la variable linéarisant une statistique. En particulier l'approximation contenue dans la définition peut être plus ou moins fine et cela à deux niveaux : celui de la vitesse de convergence $f(n)$, et, à vitesse égale, celui du majorant $C(\varepsilon)$.

En général, cependant, la variable linéarisée z_k n'est pas explicitement calculable quand on utilise les données de l'échantillon. On est alors amené à remplacer z_k par une approximation $\tilde{z}_k$ utilisant certaines statistiques estimées à partir de l'échantillon. Cela survient dans les cas les plus élémentaires. On doit se poser la question de la légitimité de cette approximation, ce qui ne peut se faire que dans un cadre asymptotique.

Résultat : Si les quantités $\tilde{z}_k$ dépendent de façon régulière d'un nombre fini fixe de paramètres estimés, alors les estimateurs de variance $\hat{Q}(z_k;\ k \in s)$ et $\hat{Q}(\tilde{z}_k;\ k \in s)$ sont équivalents, c'est à dire que leur différence normalisée par le facteur n / N^2 est une quantité asymptotiquement négligeable.

Preuve : Par « de façon régulière » nous entendons que $\tilde{z}_k = z_k + \xi_k' (\hat{\Gamma} - \Gamma) + O_p(\| \hat{\Gamma} - \Gamma \|^2)$, où Γ est le p vecteur des paramètres, $\hat{\Gamma}$ son vecteur d'estimateurs et ξ_k une p – variable. Les postulats asymptotiques nous disent que $n / N^2 Q(z_k)$ converge vers une quantité finie de même que $n / N^2 Q(z_k, \xi_k) = \sum_{k,\, l} \Delta_{kl} z_k \xi_l$ si on explicite la forme quadratique, et que $n / N^2 Q(\xi_k)$. On a alors :

$$Q(\tilde{z}_k) = Q(z_k) + 2Q(z_k, \xi_k)'(\hat{\Gamma} - \Gamma) + O_p(\| \hat{\Gamma} - \Gamma \|^2).$$

Comme $\| \hat{\Gamma} - \Gamma \| = O_p(n^{-1/2})$, on obtient le résultat.

Quand le nombre de paramètres estimés tend vers l'infini la situation n'est pas absolument claire. Au plan pratique, évidemment : qu'entendra-t-on par nombre de paramètres estimés tendant vers l'infini ? Au plan théorique, il y a aussi des difficultés comme aussi comme le montrent les deux exemples contradictoires suivants :

Exemple : Poststratification. On suppose les poststrates définies à partir d'une variable numérique, et qu'on fabrique m poststrates contiguës groupant chacune environ n / m unités sondées. Le vecteur Γ est ici celui des m moyennes de poststrates $\bar{Y}_h$, $h = 1$ à m. Si m augmente avec n, chaque paramètre estimé, $\hat{\bar{Y}}_h$ est tel que $\hat{\bar{Y}}_h - \bar{Y} = O_p((n/m)^{-1/2})$.

Par suite $\|\hat{\Gamma} - \Gamma\|$ est de l'ordre de $m^{3/2}n^{-1/2}$. Si on prend $m = n^{\alpha}$ avec $\alpha < 1/3$, le résultat précédent et sa preuve restent valides.

Contre-exemple : Pour l'estimation d'indices d'inégalité on est amené à utiliser des statistiques de la forme $S = \sum_s y_k \hat{F}(y_k)$ où $\hat{F}$ est une estimation de la fonction de répartition de la variable y. Si R_k désigne le rang de y_k dans la population, on pourrait imaginer que $z_k = 1/N\, y_k R_k$ est une statistique linéarisée pour S. C'est entièrement faux (Deville 1997).

La différence avec l'exemple précédent réside dans le fait que S utilise un paramètre estimé par unité échantillonnée, et qu'alors tout peut survenir ! On verra plus loin (section 12), comment on traite de façon générale des statistiques de cette forme.

6. Fonction d'influence d'une fonctionnelle et variance asymptotique de l'estimateur par substitution

Définition : La fonction d'influence (quand elle existe) d'une fonctionnelle T est :

$$IT(M;\ x) = \lim_{t\to 0} \frac{1}{t}(T(M + t\delta_x) - T(M))$$

où δ_x désigne la masse unité posée au point x.

Remarque : Cette définition est une légère modification de celle qui est utilisée dans le domaine de la statistique robuste (Hampel, Ronchetti, Rousseeuw, et Stahel 1985). Elle est rendue nécessaire du fait que la masse totale de M est variable et, dans un problème statistique, souvent inconnue. Elle n'est autre que la différentielle au sens de Gateaux pour une masse de Dirac posée en un point x.

Nous pouvons formuler maintenant le point essentiel de ce travail :

Résultat : Sous de larges hypothèses, l'estimation par substitution d'une fonctionnelle $T(M)$ est linéarisable. Une variable linéarisée est $z_k = IT(M; x_k)$ où IT est la fonction d'influence de T en M.

Remarque : La fonction d'influence peut donc être utilisée pour estimer la variance de $T(\hat{M})$. Ceci dit, très souvent, la fonction d'influence fait intervenir certaines fonctionnelles de M dans sa définition (Exemple : un ratio ou une moyenne). On est donc amené à choisir une estimation de la fonction d'influence elle-même pour calculer l'estimation de la variance. Ce choix n'est pas nécessairement unique.

Preuve du résultat : Munissons l'espace des mesures sur $\boldsymbol{R}^q$ d'une métrique d rendant compte de la convergence : $d(M_1, M_2) \to 0$ si et seulement si $N^{-1}(\int y\, dM_1 - \int y\, dM_2) \to 0$ pour toute variable d'intérêt y. Les postulats asymptotiques entraînent que $d(\hat{M}/N,\ M/N)$ tend vers zéro. On peut visiblement faire en sorte que $d(\hat{M}/N,\ M/N)$ soit $O_p(1/\sqrt{n})$ d'après le troisième postulat. Maintenant, supposons T dérivable au sens de Fréchet, c'est-à-dire pour toute direction de l'accroissement, sur l'espace des mesures « utiles » muni de la métrique ci-dessus évoquée. Cela entraîne qu'on a :

$$N^{-\alpha}(T(\hat{M}) - T(M)) = \frac{1}{N}\sum_U z_k(w_k - 1) + o\left(d\left(\frac{\hat{M}}{M}, \frac{M}{N}\right)\right).$$

Il en résulte que :

$$\sqrt{n}N^{-\alpha}(T(\hat{M} - T(M))) = \frac{\sqrt{n}}{N}\sum_U z_k(w_k - 1) + o_p(1).$$

La variance du second membre tend donc d'après le postulat 3 vers une limite, celle de $n/N^2 \operatorname{Var}(\hat{Z})$ et le résultat est obtenu.

7. Exemples et règles de calcul pour les fonctions d'influence

Exemple 1 : Si T est le total $T = \int x\, dM(x)$ d'une variable, la fonction d'influence de T est cette variable elle-même : $IT(M, x) = x$. En particulier si $x = 1$, $T = N$ effectif de la population. La fonction d'influence est alors constante et vaut 1.

Les règles de composition des fonctions d'influence se calquent sur celles du calcul différentiel :

Règle 1 : Si f est une fonction dérivable définie sur l'espace des valeurs de T fonctionnelle vectorielle, on a :

$$I(f(T)) = Df(T)\, IT$$

(ou Df est les matrices des dérivées partielles de f).
La preuve est immédiate.

Exemple 2 : $f(T) = 1/T$ et $T = \int x\, dM$, total scalaire. La fonction d'influence est $-x/T^2$.

Règle 2 : Si S et T sont deux fonctionnelles on a :

$$I(S + T) = IS + IT \quad \text{et} \quad I(ST) = S\, IT + T\, IS.$$

Si T et S sont à valeurs vectorielles et que H est une matrice, on a, quand les produits sont définis :

$$I(HT) = H\, IT \quad \text{et} \quad I(S'HT) = (IS)'HT + S'H\, IT.$$

Exemple 3 : $R = \int y\, dM / \int x\, dM = Y/X$ ratio de deux totaux. La fonction d'influence est

$$\frac{y}{X} - \frac{Yx}{X^2} = \frac{1}{X}(y - Rx).$$

Pour une moyenne $\overline{Y} = \int y\, dM / \int dM$, la fonction d'influence est donc : $1/N\,(y - \overline{Y})$, ce qui est la définition habituelle, ou presque, donnée dans la théorie de la robustesse (Lecoûtre et Tassi 1987).

Règle 3 : soient $S_i (i = 1, ..., q)$ des fonctionnelles scalaires et $S = \prod_{i=1}^{q} S_i$. On a :

$$IS = S\left(\sum_{i=1}^{q} \frac{IS_i}{S_i} \right).$$

Preuve : $I(\mathrm{Log} S) = \dfrac{IS}{S}$.

Soit maintenant $T(\lambda) = T(M, \lambda)$ une famille de fonctionnelles dépendant régulièrement d'un paramètre λ variant dans un domaine de $\boldsymbol{R}^q$, Λ une mesure sur ce domaine. On a la :

Règle 4 : $I(\int T(\lambda)\, d\Lambda(\lambda)) = \int IT(\lambda)\, d\Lambda(\lambda)$. C'est élémentaire.

NB : Dans les conditions de régularités figurent la possibilité de passage à la limite sous le signe d'intégration.

Si, de plus, φ est une fonction de $\boldsymbol{R}^q$ dans le domaine de T mesurable pour toutes les mesures M qui nous intéressent on procède comme suit :

Règle 5 : $I(\int T(\varphi(x)) dM(x); \xi) = T(\varphi(\xi)) + \int IT(M, \varphi(x); \xi)\, dM(x)$.

Preuve : (donnée à titre d'exemple) : Soit $S(M) = \int T(M, \varphi(x))\, dM(x)$. On a :

$$\frac{1}{t}[s(M + t\delta_\xi) - S(M)] =$$

$$\frac{1}{t} \int [T(M + t\delta_\xi, \varphi(x)) - T(M, \varphi(x))]\, dM(x)$$

$$+ \delta_\xi(T)\,(T(M + t\delta_\xi), \varphi(x)).$$

Le second terme tend vers $T(M, \varphi(\xi))$ quand t tend vers zéro. Le premier peut s'écrire :

$$\left[\int IT(M, \varphi(x); \xi) + R_{M, \varphi(x); \xi}(t) \right] dM(x)$$

où R est une quantité qui tend vers zéro (on peut faire l'hypothèse que la convergence est uniforme en x). Le résultat en découle immédiatement.

Supposons maintenant que $T(\lambda)$ soit une fonctionnelle à valeurs dans $\boldsymbol{R}^q$, régulière en λ. En particulier donc la matrice $\partial T / \partial \lambda$ est inversible pour tout M et, à M fixé, l'application $\lambda \to T(\lambda)$ est injective et admet une fonction réciproque partielle. L'équation $T(\lambda) = T_0$ a donc une solution unique pour tout M, qui définit une fonctionnelle $\lambda(M)$.

Règle 6 : La fonction d'influence de $\lambda(M)$ est :

$$I\lambda(M; \xi) = -\frac{\partial T}{\partial \lambda}(M, \lambda)^{-1}\, IT(M, \lambda; \xi).$$

Preuve : $T(M + t\delta_\xi, \lambda(M + t\delta_\xi) - T(M, \lambda)) = 0$, d'où :

$$IT(M, \lambda; \xi) + \frac{\partial T}{\partial \lambda} I\lambda(M; \xi) = 0.$$

On peut avoir besoin aussi de cette règle :

Règle 7 : Soit S une fonctionnelle dans $\boldsymbol{R}^q$, T_λ une famille de fonctionnelles indicées régulièrement par $\lambda \in \boldsymbol{R}^q$. On a :

$$I(T_s) = IT_{\lambda / \lambda = s} + \left(\frac{\partial T}{\partial \lambda} \right)_{\lambda = s} IS.$$

Preuve : On a, en écrivant tout, $I(T_S) = IT(S(M), M)$. La suite va de soi.

Notons, enfin, un lien intéressant entre la fonction d'influence et la fonctionnelle dont elle dérive :

Résultat : Si T est homogène de degré α on a :

$$\int IT(M; x)\, dM(x) = \sum_U IT(M; x_k) = \alpha T(M).$$

Le cas particulier $\alpha = 0$ montre que toute fonctionnelle homogène a une influence de somme nulle.

Preuve : On a :

$$\frac{T((1 + h) M) - T(M)}{h} = \frac{((1 + h)^\alpha - 1)}{h} T(M)$$

par définition de l'homogénéité. Le résultat suit de part la linéarité de la dérivation au sens de Gateaux et de la définition de la fonction d'influence.

8. Applications : fonctions de totaux

On a déjà vu que pour les fonctionnelles linéaires, $T(M) = \sum_U y_k = \int y dM$, la fonction d'influence est y_k elle même. L'application de la notion de fonction d'influence se ramène à une tautologie. On notera cependant qu'elle n'a rien d'asymptotique.

Fonction de totaux : Si X est un vecteur de totaux, la fonction d'influence de X est, naturellement, le vecteur x_k des variables qui composent X. De ce fait, si $T(M) = f(X) = f(\int x dM)$, la fonction d'influence de T est :

$$IT(M; x_k) = f'(X)\ .\ IT(\int x dM) = f'(X)\ .\ x_k$$

où $f'(X)$ est le vecteur ligne des dérivées partielles de f selon les coordonnées de X prises au point X. On retrouve naturellement, le résultat classique de Woodruff (1971).

Conformément à ce qui a été dit plus haut, l'estimateur par substitution de $f(X)$ est $f(\hat{X})$. Sa variance approchée est celle de $f'(X)\ .\ x_k$, et elle est numériquement approchée par $f'(\hat{X})\ .\ x_k$ ce qui est conforme aux habitudes.

Exemple : Le ratio $R = f(X, Y) = Y/X$ de deux totaux scalaires est estimé par $\hat{R} = \hat{Y}/\hat{X}$. Cette statistique (de degré 0) admet pour variable linéarisée $z_k = 1/X(y_k - Rx_k)$. Pour le calcul numérique de l'estimation de variance de $\hat{R}$, on utilisera l'approximation $\tilde{z}_k = 1/\hat{X}(y_k - \hat{R}x_k)$, expression

qui dépend de $\hat{Y}$ et $\hat{X}$ et donc de s; $\tilde{z}_k$ n'est donc pas une variable linéarisée au sens de la définition.

Exemple : Estimateur par ratio.

C'est $\hat{Y}_{\text{rat}} = (X/\hat{X})\hat{Y}$. Si on se reporte à l'exemple précédent, la variable linéarisée de $\hat{Y}_{\text{rat}}$ est $y_k - Rx_k$, approximée par $y_k - \hat{R}x_k$. Mais on peut aussi se dire que la variance estimée de $\hat{Y}_{\text{rat}}$ doit être égale à X^2 fois la variance estimée de $\hat{R}$, ce qui conduit à l'approximation $X/\hat{X}(y_k - \hat{R}x_k)$ qui a été maintes fois proclamée comme plus intéressante que la précédente. Cet exemple montre que le choix de la variable linéarisée n'est pas nécessairement unique dès qu'on utilise une information externe.

Néanmoins, l'un des intérêts de l'approche par les fonctions d'influence est de permettre un calcul assez facile dans des cas apparemment complexes.

Exemple : Le coefficient de corrélation entre x et y s'écrit :

$$\rho = \frac{\sum_U x_k y_k - \frac{1}{N}\sum_U x_k \sum_U y_k}{\sqrt{\sum_U x_k^2 - \frac{1}{N}\left(\sum x_k\right)^2}\sqrt{\sum_U y_k^2 - \frac{1}{N}\left(\sum y_k\right)^2}} = \frac{V_{XY}}{\sqrt{V_{XX}V_{YY}}}.$$

En utilisant les dérivées logarithmiques (règle 3), on obtient :

$$\frac{(I\rho)_k}{\rho} = \frac{I(V_{XY})_k}{V_{XY}} - \frac{1}{2}\frac{I(V_{XX})_k}{V_{XX}} - \frac{1}{2}\frac{I(V_{YY})_k}{V_{YY}}.$$

L'influence de $A_{XY} = 1/N \sum_U x_k \sum_U y_k$ s'obtient de la même façon par :

$$I(A_{XY})_k = A_{XY}\left(\frac{x_k}{X} + \frac{y_k}{Y} - \frac{1}{N}\right) = \bar{Y}x_k + \bar{X}y_k - \bar{X}\bar{Y}$$

d'où : $I(V_{XY})_k = x_k y_k - I(A_{XY})_k = (x_k - \bar{X})(y_k - \bar{Y})$. Par suite $I(V_{XX})_k = (x_k - \bar{X})^2$ et $I(V_{YY})_k = (y_k - \bar{Y})^2$ d'où il vient, avec

$$S_x^2 = \frac{V_{XX}}{N} \text{ et } \cong_Y^2 = \frac{V_{YY}}{N} :$$

$$N(I\rho)_k = \frac{(x_k - \bar{X})(y_k - \bar{Y})}{S_X S_Y} - \frac{1}{2}\rho\left(\frac{(x_k - \bar{X})^2}{S_X^2} + \frac{(y_k - \bar{Y})^2}{S_Y^2}\right).$$

Et voilà le travail.

9. Application : paramètre implicite

Supposons que B, paramètre à q composantes, soit solution d'une équation de la forme :

$$H(B) = \sum_U l_k(B) = 0 \qquad (9.1)$$

où les l_k sont des fonctions régulières de $\boldsymbol{R}^q$ dans $\boldsymbol{R}^q$. Cette situation survient fréquemment lorsque B est le paramètre d'un modèle qu'on suppose valide dans la population U. L'équation (9.1), sous les hypothèses d'indépendance habituelles peut résulter de l'application du principe d'estimation par le maximum de vraisemblance. C'est alors l'équation du score. Dans le cas d'un modèle linéaire avec résidus gaussiens, ceci conduit aux équations normales qu'on peut aussi déduire du principe des moindres carrés :

$$l_k(B) = \frac{1}{\sigma_k^2} x_k (y_k - x_k' B)$$

avec des notations évidentes.
Si la famille de fonctionnelle $H(B)$ dépend régulièrement de B, on a :

$$I(B_0)_k = -\left(\frac{\partial H}{\partial B}\right)_{B=B_0}^{-1} IT(B_0)_k$$

soit :

$$I(B_0)_k = -\left(\sum_U \frac{\partial l_k}{\partial B}\right)^{-1} l_k(B_0).$$

Dans le cas de la régression cela donne :

$$I(B)_k = -\left(\sum_U \frac{x_k x_k'}{\sigma_k^2}\right)^{-1} \frac{1}{\sigma_k^2} x_k e_k = -T^{-1}\frac{1}{\sigma_k^2} x_k e_k$$

avec e_k. résidu de la régression. On retrouve ainsi simplement la variable linéarisée du vecteur des coefficients de régression. Pour le calcul numérique de l'estimateur de variance, on utilisera l'approximation

$$\tilde{z}_k = \hat{T}_s^{-1}\frac{1}{\sigma_k^2} x_k \tilde{e}_k \text{ où } \hat{T}_s = \sum_s \frac{x_k\, x_k'}{\sigma_k^2\, \pi_k}$$

et $\tilde{e}_k = y_k - \hat{B}x_k$. Cette expression dépend donc de s par l'intermédiaire de $\hat{T}_s$ et

$$\hat{A} = \sum_s \frac{x_k y_k}{\sigma_k^2 \pi_k}.$$

Exemple : Estimateur par régression.

Quand la constante (ou la variable σ_k^2) fait partie des régresseurs, c'est-à-dire qu'il existe un vecteur λ tel que $x_k'\lambda = 1$ (ou σ_k^2) pour tout x, l'estimateur par régression prend la forme simple $\hat{Y}_{\text{reg}} = X'\hat{B}$ ou X est le vecteur connu du total des x_k.

La théorie de l'estimation par régression (Cochran 1977, Särndal, Swensson, Wretman 1992) nous dit que les résidus e_k sont la variable linéarisée de cet estimateur et qu'on peut les approximer par les résidus empiriques

estimés $\tilde{e}_k$ (on observera que ceux ci ne dépendent que d'un nombre fini de paramètres).

Cependant, ce qui précède nous incline à penser qu'on devrait avoir :

$$\widehat{\mathrm{Var}}(\hat{Y}_{\mathrm{Reg}}) = X' \widehat{\mathrm{Var}}(\hat{B}) X$$

et qu'une variable « linéarisée » naturelle pour $\hat{Y}_{\mathrm{Reg}}$ devrait être $X' \hat{T}_s^{-1} 1/\sigma_k^2 \, x_k \, \tilde{e}_k$. Si on remplace $\hat{T}_s$ par son espérance T, on s'aperçoit que $X' T^{-1} = \lambda'$ et que $1' x_k = \sigma_k^2$ et on retombe sur l'approximation précédente. On remarquera, enfin, que les quantités $X' \hat{T}_s^{-1}$ sont exactement les corrections de poids (où $g-$poids) utilisées dans l'estimateur par régression et dont l'usage est souvent conseillé dans le cadre de l'estimation de variance.

Il est bien clair que les deux variables linéarisées mènent asymptotiquement au même résultat. Le choix devrait donc se baser sur d'autres critères. La notion d'estimation conditionnelle justifie, dans quelques cas particuliers, l'usage de ces $g-$poids, notamment pour l'estimateur poststratifié dans le cas d'un sondage aléatoire simple. Le cas général reste assez mystérieux.

Dans le cas de l'ajustement d'une régression logistique, la variable dépendante y_k vaut 0 ou 1 et les équations du score s'écrivent :

$$H(B) = \sum_U x_k (y_k - f(x_k' B)) \text{ avec } f(u) = \exp u/(1+\exp u).$$

On a donc :

$$I(B)_k \equiv \left(\sum_U x_k x_k' f(x_k' B)(1 - f(x_k' B)) \right)^{-1} x_k (y_k - f(x_k' B)).$$

L'usage de cette variable permet de calculer correctement la précision d'une régression logistique, c'est à dire en prenant en compte le schéma d'échantillonnage.

10. La technique des résidus pour les estimateurs complexes

Beaucoup d'estimateurs complexes couramment utilisés de nos jours s'intègrent dans le cadre général du calage sur des données externes (Deville, Särndal 1992 et Deville, Särndal, Sautory 1993). Un vecteur X de totaux de variables auxiliaires x_k est connu et on cherche de nouveaux poids w_k vérifiant les équations de calage $\sum_s w_k x_k = X$ pour tout échantillon s. Si on cherche ces poids aussi proches que possible des poids de HT, on trouve qu'ils sont nécessairement de la forme $w_k = 1/\pi_k \, F_k(x_k' \lambda)$ où λ est un vecteur de même dimension que X solution des équations de calage. Les fonctions F_k dépendent de la distance choisie et admettent un développement limité de la forme $F_k(u) = 1 + q_k u + O(u^2)$. La forme la plus fréquente est $F_k(u) = F(q_k u)$, F fonction unique. Souvent aussi les q_k sont tous égaux à 1. On retrouve ainsi dans cette famille l'estimateur par ratio (F arbitraire et $q_k = 1/x_k$), l'estimateur poststratifié (x_k vecteur indicateur de strate), l'estimateur par raking-ratio (x_k vecteur indicateur des marges et F fonction exponentielle) et l'estimateur par régression ($F(u) = 1 + u$).

La variance asymptotique de ces estimateurs peut s'obtenir naturellement par application des règles de linéarisation. Les équations de calage définissent λ par :

$$T(\hat{M}, \lambda) = \int x_k \, F_k(x_k' \lambda) \, d\hat{M}(k) = \sum_s \frac{1}{\pi_k} x_k \, F_k(x_k' \lambda) = X.$$

Comme on a $T(M, 0) = X$, l'application de la règle 6 conduit à :

$$I\lambda(M, x) = -T^{-1} x_k$$

avec

$$T = \int x_k x_k' F_k(0) \, dM(k) = \sum_U q_k x_k x_k'.$$

Par ailleurs l'estimateur calé apparaît comme l'estimateur par substitution de la fonctionnelle $S(M, \lambda) = \int y_k F_k(x_k' \lambda) \, dM(k)$, qui admet, d'après la règle 7, la variable linéarisée $y_k - x_k' T^{-1} \int q_i y_i x_i \, dM(i) = y_k - x_k' B$ en introduisant le vecteur B des paramètres de la régression des moindres carrés dans la population pour les poids q.

La variance de l'estimateur calé s'obtient donc en remplaçant dans la formule (2.1) les y_k par les résidus $e_k = y_k - x_k' B$ de la régression de y sur x avec les poids q. Pour l'estimation de variance, on utilisera dans la formule (2.2) soit $\tilde{e}_k = y_k - x_k' \hat{B}$ soit, comme pour le cas de l'estimateur par régression, $F(x_k' \hat{\lambda}) \, \tilde{e}_k$.

Si on s'intéresse maintenant à un paramètre $T(M)$ estimé par substitution en utilisant les poids w_k issu d'un calage, on a le résultat important suivant :

Résultat : Si $T(M)$ admet une variable linéarisée z_k, et que $T(\hat{M}_w)$ est l'estimateur de $T(M)$ utilisant des poids w_k issus d'un calage sur un vecteur $X = \sum_U x_k$, alors e_k, résidu de la régression z_k sur x_k est une variable linéarisée pour $T(\hat{M}_w)$.

Preuve : La variance de $T(\hat{M}_w)$ est équivalente à celle de $\hat{Z}_w = \sum w_k z_k$ d'après la démonstration précédente. Mais la variance de $\hat{Z}_w$ est équivalente à celle de $\sum_s 1/\pi_k \, e_k$.

Commentaire : Très souvent, par exemple dans le cas d'une fonction explicite de totaux, z_k est une forme linéaire $\sum_{i=1}^p A_i \, y_k^i$. On a alors :

$$\mathrm{Var} \sum_s w_k z_k = \mathrm{Var} \sum_{i=1}^p A_i \, \hat{Y}_w^i.$$

Ceci suggère la procédure suivante :

- calculer les résidus ε_k^i des régressions des y_k^i sur les x_k.
- former la variable synthétique $\sum_i A_i \, \varepsilon_k^i$
- calculer la variance de cette variable.

Il est bien clair que cela revient à calculer directement les résidus des z_k, ce qui est tout de même plus simple.

Remarque : En dépit de son caractère trivial , ce résultat est peut-être le plus utile de ce travail, d'où cette remarque destinée uniquement à ce qu'il ne passe pas trop inaperçu.

11. Application : fractiles

La fonction de répartition $F(x) = 1/N \,\mathrm{Card}(k;\, x_k \le x)$ est une famille de fonctionnelle $1/N \int 1(\xi \le x)\, dM(\xi)$. L'influence $IF(x)_k$ vaut donc

$$\frac{1}{N}(\mathbf{1}(x_k \le x) - F(x)). \qquad (11.1)$$

Pour $\alpha \in]0,1[$, le fractile t_α est défini par $F(t_\alpha) = \alpha$ si on est brutal, par t_α: $F(t_\alpha - 0) < \alpha \le F(t_\alpha)$ si on tient compte du caractère escaliériforme de F. Si on est brutal, la variable linéarisée ad hoc est donc :

$$I(t_\alpha)_k = -\left(\frac{\partial F}{\partial x}\middle| x = t\alpha\right)^{-1} IF(t_\alpha)_k$$

$$= -\frac{1}{F'(t_\alpha)} \cdot \frac{1}{N}(\mathbf{1}(x_k \le t_\alpha) - \alpha). \qquad (11.2)$$

Le problème vient du fait que $F'(x)$ idéalise une densité de la variable au point x laquelle n'existe pas à cause des escaliers.

On peut circonvenir cette difficulté par la construction suivante :

Un noyau régularisant est, par définition, une fonction positive $K(x, t)$, vérifiant pour tout x, $\int K(x, t)dt = 1$, et régulière (suffisamment dérivable par exemple). Pour tout x, $K(x, .)$ est une fonction en « cloche » autour de x, par exemple l'indicatrice normalisée d'un intervalle entourant x. Plus généralement le support de $K(x,.)$ sera un intervalle contenant x. On notera $G(x, t) = \int^t K(x, u)\, du$ et $\bar{G}(x, t) = 1 - G(x, t)$. $G(x; .)$ est une fonction de répartition. Dans une optique asymptotique, le noyau K dépendra de la taille N de la population; la « largeur de bande », c'est-à-dire la largeur « moyenne » du support de $K(x, .)$ diminuera avec N.

Remplaçons la fonction de répartition par son lissage $F_K(x) = \int F(t)\, K(x, t)\, dt$. Pour un choix raisonnable de K, F_K est strictement croissante en tout point où elle ne vaut ni 0 ni 1, et très voisine de F de sorte que tous les fractiles $t_{K\alpha}$ sont définis de façon univoque et voisins des t_α quelle que soit la façon dont on les définisse. Après une intégration par partie on notera qu'on a aussi :

$$F_K(x) = \int \bar{G}(x, t)\, dF(t) = \frac{1}{N}\sum_U \bar{G}(x, x_k).$$

On a donc :

$$I(F_K(x),\, \xi) = \frac{1}{N}(\bar{G}(x, \xi) - F_K(x)) \qquad (11.3)$$

ce qui est tout à fait analogue à (11.1).

Comme F_K est dérivable ($\bar{G}$ l'étant) on a :

$$It_{K\alpha}(x) = -\frac{1}{F'_K(t_{K\alpha})}\frac{1}{N}(\bar{G}(t_{K\alpha},\, x) - \alpha). \qquad (11.4)$$

Cette formule est complètement analogue à (11. 2) à ceci près que $F'_K(t_{K\alpha})$ est parfaitement défini. La linéarisation de $t_{K\alpha}$ ne pose donc aucun problème particulier et pourra être utilisée de façon approximative pour la linéarisation de t_α lui-même. Une voie mixte consiste à utiliser la variable linéarisée

$$z_k = -\frac{1}{F'_K(t_\alpha)}(\mathbf{1}(x_k \le t_\alpha) - \alpha)$$

avec

$$K(x, t) = \frac{1}{b - a}\mathbf{1}(a \le t < b)$$

(où $[a, b]$ est un intervalle contenant x plus ou moins arbitraire). Une variable linéarisé pratiquement correcte sera :

$$z_k = -\frac{b - a}{F(b) - F(a)}(\mathbf{1}(x_k \le t_\alpha) - \alpha).$$

L'intervalle $[a, b]$ devra être suffisamment large pour que dans

$$\tilde{z}_k = -\frac{b - a}{\hat{F}(b) - \hat{F}(a)}(\mathbf{1}(x_k \le \hat{t}_\alpha) - \alpha)$$

le premier facteur soit suffisamment insensible aux fluctuations d'échantillonnage.

12. Indices de concentration et autres fonctionnelles liées aux rangs

Nous prendrons quelques exemples.

a) Indice de GINI.

Avec $T_x = \int \mathbf{1}(\xi < x)\, dM(\xi)$, on peut définir l'indice de Gini par :

$$\mathrm{GINI} = \frac{\int xT_x dM(x)}{NX}.$$

En appliquant la règle 5, on trouve pour influence du numérateur $xT_x + \int \xi \mathbf{1}(x \le \xi)\, dM(\xi)$. Or $\int \xi \mathbf{1}(x \le \xi)\, dM(\xi) = X - \int \xi \mathbf{1}(\xi \le x)\, dM(\xi) = X - T_x\bar{x}_<$ où $\bar{x}_<$ est la moyenne des x_k inférieurs à x. Comme X est une constante, la linéarisée de numérateurs est donc $T_x(x - \bar{x}_<)$. Une linéarisée pour GINI est donc :

$$\mathrm{IGINI_k} = F(x_k)\frac{x_k - \bar{x}_<}{X} - \mathrm{GINI}\frac{x_k}{X}.$$

b) Population en dessous du seuil de pauvreté.

Elle est définie comme la proportion (des revenus) inférieurs à la moitié de la médiane de la distribution. Pour faire bon poids, soit α et β deux nombres compris entre 0 et 1 et étudions l'indicateur $J_{\alpha\beta} = F(\beta\, t_\alpha)$. L'indicateur habituel correspond à $\alpha = \beta = 1/2$.

La linéarisation va de soi en utilisant les règles du paragraphe 6 et la convention pour l'écriture de la dérivée de la fonction de répartition :

$$IJ_{\alpha\beta}(x) = IF_{\beta q_\alpha}(x) + F'(\beta t_\alpha)\,\beta I_{t_\alpha}(x)$$
$$= \frac{1}{N}(\mathbf{1}(x \le \beta t_\alpha) - F(\beta t_\alpha)) - \frac{1}{N}\frac{F'(\beta t_\alpha)}{F'(t_\alpha)}(1(x \le t_\alpha) - \alpha)$$
$$= \frac{1}{N}\left[\mathbf{1}(x \le \beta t_\alpha) - \frac{F'(\beta t_\alpha)}{F'(t_\alpha)}1(x \le t_\alpha) + (\alpha - F(\beta t_\alpha))\right].$$

Pour $\beta = 1$ on trouve bien $IJ_{\alpha 1} = 0$.

La variance de l'indicateur se calculera donc simplement à l'aide de la variable artificielle qui vaut 1 si $x_k \le \beta\hat{t}_\alpha$,

$$1 - \frac{F'(\beta\hat{t}_\alpha)}{F'(\hat{t}_\alpha)}$$

si $\beta\hat{t}_\alpha < x_k \le \hat{t}_\alpha$ et 0 si $x_k > \hat{t}_\alpha$.

c) Coefficient de corrélation de rang de Kendall.

Deux variables numériques x_k et y_k sont attachées à l'individu k. Les rangs de x_k et y_k respectivement peuvent s'écrire $R_k^X = \int_{x \le x_k} dM(x, y)$ et $R_k^Y = \int_{y \le y_k} dM(x, y)$. Le coefficient de corrélation de rang est le coefficient de corrélation entre R_k^X et R_k^Y, c'est-à-dire, après des simplifications élémentaires :

$$r = 12\left(\frac{1}{N^3}\int R_\xi^X R_\eta^Y dM(\xi, \eta) - \frac{1}{2}\right).$$

Cette expression se linéarise en appliquant les règles relatives aux fonctions d'influence. Pour : $T = \int R_\xi^X R_\eta^Y\, dM(\xi, n)$, on a :

$$IT(x, y) = R_x^X R_y^Y + \int \mathbf{1}(x \le \xi)\, R_\eta^Y\, dM(\xi, \eta)$$
$$+ R_\xi^X \mathbf{1}(y \le \eta)\, dM(\xi, \eta)$$
$$= R_x^X\, R_y^Y + A_x + B_y$$

où on a posé

$$A_x = \sum_{k \in U:\, x \le x_k} R_K^Y$$

et

$$B_y = \sum_{k \in sU:\, y \le x_k} R_k^X,$$

d'où enfin :

$$Ir(x, y) = \frac{12}{N}\left(F_x^X F_y^Y + \frac{A_x}{N^2} + \frac{B_y}{N^2} - \frac{1}{4}\left(r + \frac{1}{2}\right)\right).$$

Le calcul de variance s'effectue de la façon suivante :

- La variable linéarisée est $z_k = Ir(x_k, y_k)$.
- $\hat{F}_{x_k}^X$ et $\hat{F}_{y_k}^Y$ sont les estimateurs des fonctions de répartition de x et y respectivement.
- A_x est estimé par

$$\hat{A}_x = \sum_{k \in s:\, x_k \ge x} w_k$$

et B_y pareillement.

- On utilise, dans le calcul, l'approximation de

$$z_k,\ \tilde{z}_k = \frac{12}{\hat{N}}\left(\hat{F}_{x_k}^X \hat{F}_{y_k}^Y + \frac{A_{x_k}}{\hat{N}_2} + \frac{B_{y_k}}{\hat{N}_2} - \frac{1}{4}\left(\hat{r} + \frac{1}{2}\right)\right)$$

et on calcule la variance du total de cette variable estimée par l'estimateur de HT (formule 2.2).

13. Méthodes factorielles

Les composantes principales de la variable vectorielle x_k sont les vecteurs propres u de la matrice des covariances $C = \sum_U x_k x_k' - X\bar{X}'$. Ils vérifient donc :

$$\begin{aligned} Cu &= \lambda u \\ u'u &= 1 \end{aligned} \quad \text{avec } \lambda \text{ valeur propre.}$$

La variance de λ et celle des composantes du u s'obtiennent assez simplement. L'influence de C est $IC(x) = (x - \bar{X})(x - \bar{X})'$. L'influence de $Cu - \lambda u$ est

$$ICu + CIu - I\lambda u - \lambda Iu = 0. \qquad (13.1)$$

Mais $(Iu)'u = 0$, de même que $u'CIu = 0$ parce que C est une matrice symétrique. En multipliant (13.1) à gauche par u' il vient :

$$u'ICu = I\lambda = (u'(x - \bar{X}))^2.$$

Or $u'(x - \bar{X})$ est égal à $\lambda\xi^2$ où ξ est la composante principale associé à l'élément propre (λ, u). On en déduit le calcul de la variance de $\hat{\lambda}$, solution de $\hat{C}\hat{u} - \hat{\lambda}\hat{u} = 0$.

La variance des composantes de u s'obtient de façon analogue. Soit (λ_v, v) un autre élément propre de C. Multiplions l'équation (13.1) à gauche par v. Il vient :

$$v'ICu + \lambda_v(v'Iu) - \lambda(v'Iu) = 0$$

d'où :

$$(v'Iu) = \frac{(\lambda\lambda_v)^{1/2}\xi\xi_v}{\lambda - \lambda_v}$$

et donc :

$$Iu = \sum_{v \ne u}\frac{(\lambda\lambda_v)^{1/2}\xi\xi_v}{\lambda - \lambda_v}\, v.$$

L'analyse des correspondances (AFC) ou l'analyse des correspondances multiples (ACM) sont susceptibles d'un traitement analogue.

Pour l'ACM (cas le plus général), chaque individu est caractérisé par le vecteur x_k qui « empile » les indicatrices d'appartenance aux modalités de p variables qualitatives (2 dans le cas de l'AFC). Si $\underline{1}$ désigne le vecteur dont toutes les composantes valent 1, on a $x_k'\,\underline{1} = p$ pour tout k. On cherche alors les vecteurs u normés par $1/p\,N\,u'Du = 1$, avec $D = \text{diag}\sum_U x_k = \sum_U \text{diag}\, x_k$ tels que la variance de $\xi_k = 1/p x_k' u$ soit stationnaire. Ceci conduit à résoudre le

problème aux valeurs propres : $Cu - p\lambda Du = 0$ où $C = \sum_U x_k x_k'$.

La recherche d'une linéarisée pour λ et u suit le même chemin que précédemment. On a la relation entre influences :

$$(IC - pI\lambda D - p\lambda ID)\,u - (C - p\lambda D)\,Iu = 0.$$

Comme $IC = xx'$, $ID = \text{diag}\ x$, et $u'DIu = 0$, on obtient en prémultipliant par u':

$$I\lambda(x) = \frac{1}{N}\left(\left(\frac{x'}{p}u\right)^2 - \lambda\frac{x'}{p}uOu\right)$$

où uOv désigne le produit d'Hadamard (c'est-à-dire composante par composante) de u et v. On sait que $u = \underline{1}$ est vecteur propre associé à la valeur propre 1 qui est aussi la plus grande. On vérifiera que pour $u = \underline{1}$ on a $I\lambda = 0$!

De la même façon on obtient les composantes de Iu sur les autres vecteurs propres v :

$$v'DIu = \frac{1}{N}\left[\left(\frac{x'}{p}\cdot u\right)\left(\frac{x'}{p}\cdot v\right) - \lambda\frac{x'}{p}\cdot uOv\right].$$

On peut continuer l'analyse en calculant la variabilité d'une projection sur un plan factoriel. Si A est une sous-population d'effectif N_A, les coordonnées de son point représentatif sur les plans factoriels sont

$$\alpha_u = \left(\frac{\sum_A x_k}{\sum_A 1}\right)'\cdot u = \bar{X}_A'\cdot u.$$

On linéarise α_u en utilisant la relation $I\alpha_u = (I\bar{X}_A)'u + \bar{X}_A' Iu$ et le reste est simple.

14. Conclusion

La linéarisation des statistiques complexes est considérée, depuis longtemps, comme la méthode la plus souple et la plus compréhensible pour en obtenir une estimation de variance. Cette méthode s'applique, en particulier, à n'importe quel plan de sondage et à n'importe quel type d'estimateur. La vogue des méthodes basées sur des réplications d'échantillon est due surtout au fait qu'on considère certaines statistiques comme trop complexes pour être linéarisées. Cependant, pour la large classe des estimateurs par substitution, l'usage des fonctions d'influence et des règles algébriques qui gouvernent leur construction permet d'obtenir assez simplement des variables linéarisées grâce auxquelles l'estimation de variance se ramène à celle d'un total estimé par l'estimateur de Horvitz-Thompson.

Remerciements

Je tiens à remercier Olivier Sautory et Yves Tillé pour leurs commentaires critiques et bienveillants de diverses versions préliminaires de ce texte. La démonstration du début de la section 7 est due à ce dernier. Je remercie également les deux arbitres de « *Technique d'Enquête* » dont les remarques ont fortement contribué à la lisibilité de cet article et la rédaction du journal pour sa compréhension.

Bibliographie

Billingsley, P. (1969). *Convergence of Probability Measures*. New York : John Wiley & Sons, Inc.

Binder, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *Revue Internationale de Statistique*, 51, 279-292.

Binder, D.A., et Kovačević, M.S. (1997). Variance estimation for measures of income inequality and polarization: The estimating equations approach. *Journal of Official Statistics*, 13, 41-58.

Binder, D.A., et Patak, Z. (1994). Use of estimating functions for interval estimation from complex surveys. *Journal of the American Statistical Association*, 89, 1035-1043.

Chambers, R.L., Dorfman, A.H. et Hall, P. (1992). Properties of estimators of the finite population distribution function. *Biometrika*, 79, 577-582.

Cochran, W. (1977). *Sampling Techniques*, 3rd Edition. New York : John Wiley & Sons, Inc.

Deville, J.-C. (1993). Une formule universelle d'estimation de variance. Document interne, INSEE-UMS.

Deville, J.-C. (1997). Estimation de la variance du coefficient de Gini mesuré par sondage. Dans *Actes des Journées de Méthodologie Statistiques*, INSEE METHODES, 69-70-71.

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

Deville, J.-C., Särndal, C.-E. et Sautory, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association*, 88, 1013-1020.

Durbin, J. (1953). Some results in sampling theory when units are selected with unequal probabilities. *Journal of the Royal Statistical Society B*, 15, 262-269.

Hájek, J. (1960). Limiting distributions in simple random sampling from a finite population. *Publications of the Mathimatical Institute of the Hungarian Academy of Sciences*, 5, 361-374.

Hampel, F.R., Ronchetti, E., Rousseeuw, P.J. et Stahel, W. (1985). *Robust Statistics*: *The Approach Based on the Influence Function*. New York : John Wiley & Sons, Inc.

Isaki, C.T., et Fuller, W.A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77, 89-96.

Lecoûtre, J.P., et Tassi, PH. (1987). Statistique non-paramétrique et robustesse. *Economica*.

Rosen, B. (1972). Asymptotic theory for successive sampling I et II. *Annals of Mathematical Statistics*, 43, 373-397 et 748-776.

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

Wolter, K.M. (1985). *Introduction to Variance Estimation*. New York : Springer.

Woodruff, R.S. (1971). A simple method for approximating the variance of a complicated estmate. *Journal of the American Statist ical Association*, 66, 411-414.