

MISE EN ŒUVRE DU CALCUL DE VARIANCE PAR LINÉARISATION

Fabien DELL(*), *Xavier d'HAULTFOEUILLE* (*), *Philippe FÉVRIER* (*),
Emmanuel MASSÉ (**)

(*) *Insee, Direction des statistiques démographiques et sociales,*
(**) *Ministère de l'Environnement*

Introduction

Une statistique n'a pas de valeur sans son écart-type. Partant de cette constatation, ce document a pour objet de faire le point sur les méthodes de calcul de la variance des statistiques estimées par sondage. Dans une première partie nous examinons le cas classique d'un total. Après avoir rappelé les formules de variance dans le cadre général, nous montrons quel estimateur peut être mis en oeuvre. Dans une deuxième partie, nous montrons comment la méthode de linéarisation proposée par Deville ([6] et [8]) peut être utilisée pour calculer des variances d'estimateurs complexes (non-linéaires). Nos applications portent principalement sur des indicateurs d'inégalité (THEIL, GINI, ATKINSON,...). La linéarisation est également comparée au bootstrap, qui est une autre méthode pour estimer les variances d'estimateurs complexes. Dans une troisième partie, le code des macros de linéarisation écrites sous SAS est détaillé. Enfin, nous donnons un exemple de calcul de variance dans un cas pratique: celui du taux de pauvreté calculé sur l'*Enquête Revenus Fiscaux*. La syntaxe et les programmes complets des macros SAS sont présentés en annexe.

Ce document ainsi que les macros qui l'accompagnent sont sujets à mises à jour. Les commentaires, critiques et propositions de nouvelles macros sont bienvenues¹.

1. Variance d'un total

Nous nous plaçons ici dans le cas le plus simple, à savoir celui de l'étude du total d'une variable. Notre objet n'est pas ici de faire un cours général de sondages² mais d'explicitier les déterminants de la variance de ce total. Ces facteurs sont principalement le plan de sondage, la non-réponse totale, le calage de l'enquête, les erreurs de mesure et l'imputation.

¹S'adresser à Fabien DELL, fabien.dell@insee.fr.

²Le lecteur intéressé pourra se reporter par exemple à l'ouvrage de Tillé ([12]) pour une présentation détaillée des résultats évoqués ici.

1.1 Prise en compte du plan d'échantillonnage

1.1.1 Résultats généraux

Soit U la population totale composée de N individus. La variable étudiée est une variable Y qui prend la valeur Y_k pour l'individu k : elle est observée sur un échantillon S de n personnes. On note \mathbf{p}_k les probabilités d'inclusion simples (*i.e.* la probabilité qu'un individu k soit dans l'échantillon) et \mathbf{p}_{kl} les probabilités d'inclusion doubles (*i.e.* la probabilité que les individus k et l soient tous deux dans l'échantillon).

On s'intéresse au total d'une variable Y sur la population : $Y = \sum_{k \in U} Y_k$. Ce total peut être estimé sur l'échantillon par l'estimateur de Horvitz-Thompson :

$$\hat{Y} = \sum_{k \in S} \frac{Y_k}{\mathbf{p}_k}$$

Cette statistique estime sans biais le total dans la population, comme le montre la démonstration classique que nous redonnons ici. Il est très important de noter que dans le cadre formel de la théorie des sondages, seul l'échantillon est aléatoire et non la variable elle-même. L'individu k a un Y_k non aléatoire (une « étiquette »), seul est aléatoire le fait qu'il soit ou non interrogé. Nous obtenons donc :

$$E \left[\sum_{k \in S} \frac{Y_k}{\mathbf{p}_k} \right] = E \left[\sum_{k \in U} \frac{Y_k \mathbf{1}(k \in S)}{\mathbf{p}_k} \right] = \sum_{k \in U} Y_k \frac{E[\mathbf{1}(k \in S)]}{\mathbf{p}_k} = \sum_{k \in U} Y_k$$

Comme nous l'avons déjà dit dans l'introduction, la seule donnée de l'estimation n'a pas vraiment de sens sans son complément indispensable qu'est la variance de cette estimation. Dans le cadre présent, la variance de l'estimation d'un total se calcule facilement (en utilisant la même méthode que précédemment pour le calcul de l'espérance) et l'on obtient :

$$\hat{\hat{Y}} = \frac{\hat{Y}}{N}, \quad \hat{V}(\hat{\hat{Y}}) = \frac{\hat{V}(\hat{Y})}{N^2}$$

$$V \left[\sum_{k \in S} \frac{Y_k}{\mathbf{p}_k} \right] = \sum_{k \in U} \frac{Y_k^2}{\mathbf{p}_k} (1 - \mathbf{p}_k) + \sum_{k \in U} \sum_{\substack{l \in U \\ l \neq k}} \frac{Y_k Y_l}{\mathbf{p}_k \mathbf{p}_l} (\mathbf{p}_{kl} - \mathbf{p}_k \mathbf{p}_l)$$

Cette variance ne peut évidemment pas être calculée puisqu'elle nécessite de l'information (dont on ne dispose pas) sur toute la population U . Il est néanmoins possible de construire un estimateur sans biais de cette variance à partir de l'échantillon :

$$\hat{V} \left[\sum_{k \in S} \frac{Y_k}{\mathbf{p}_k} \right] = \sum_{k \in S} \frac{Y_k^2}{\mathbf{p}_k} (1 - \mathbf{p}_k) + \sum_{k \in S} \sum_{\substack{l \in S \\ l \neq k}} \frac{Y_k Y_l}{\mathbf{p}_k \mathbf{p}_l} (\mathbf{p}_{kl} - \mathbf{p}_k \mathbf{p}_l)$$

1.1.2 Plans à un seul degré

Sondage aléatoire simple ou poissonnien

Cette formule se simplifie dans le cas des sondages aléatoires simples et poissonnien³ et prend les formes suivantes :

$$\hat{V}_{SAS} \left[\sum_{k \in S} \frac{Y_k}{n/N} \right] = \frac{N(N-n)}{n} \frac{1}{n-1} \sum_{k \in S} (Y_k - \hat{Y})^2 \quad \text{avec} \quad \hat{Y} = \frac{1}{n} \sum_{k \in S} Y_k$$

$$\hat{V}_{POIS} \left[\sum_{k \in S} \frac{Y_k}{p_k} \right] = \sum_{k \in S} \frac{Y_k^2}{p_k^2} (1 - p_k)$$

Autres plans à un degré

Dans le cas général, l'estimateur de la variance est difficile à calculer car le calcul des probabilités d'inclusion doubles p_{kl} s'avère trop complexe.

De nombreux auteurs ont cherché des approximations de la variance n'utilisant que les probabilités d'inclusion simples. Nous avons retenu ici la formule de Deville⁴ :

$$\hat{V}(\hat{Y}) = \frac{1}{1 - \sum_{k \in S} a_k^2} \sum_{k \in S} (1 - p_k) \left(\frac{Y_k}{p_k} - A \right)^2$$

$$\text{où } a_k = \frac{1 - p_k}{\sum_{k' \in S} (1 - p_{k'})} \text{ et } A = \sum_{k \in S} a_k \frac{Y_k}{p_k}$$

Cette approximation est facilement calculable dans les enquêtes puisqu'elle n'utilise que les poids et la variable sur l'échantillon. C'est cette formule qui a été retenue dans le logiciel Poulpe (voir la partie sur les plans complexes).

1.1.3 Plans stratifiés

Une stratification de la population consiste à utiliser une information auxiliaire pour partitionner la population globale en H sous-ensembles les plus homogènes possibles, appelés strates. On tire ensuite de manière indépendante dans chaque strate h un sous-échantillon S_h .

L'estimateur du total s'écrit alors (avec des notations évidentes) :

$$\hat{Y} = \sum_{h=1}^H \sum_{k \in S_h} \frac{Y_k}{p_k}$$

³Le sondage poissonnien correspond à un tirage indépendant de chaque unité k avec une probabilité p_k .

⁴Cette formule n'est théoriquement valable que lorsque l'échantillon a été tiré suivant un plan dont l'entropie est pratiquement maximale (cf. Tillé, [12] pour la définition de l'entropie).

tandis que sa variance prend la forme :

$$V(\hat{Y}) = \sum_{h=1}^H V_H \left(\sum_{k \in S_h} \frac{Y_k}{p_k} \right)$$

On peut alors appliquer les remarques et techniques précédentes pour le calcul de la variance à l'intérieur de chaque strate i.e. pour le calcul de $V_h \left(\sum_{k \in S_h} \frac{Y_k}{p_k} \right)$.

1.1.4 Plans à deux degrés

Le plan à deux degrés met en œuvre un double échantillonnage. On échantillonne dans un premier temps m unités primaires parmi M à l'aide d'un sondage S_1 . Puis, indépendamment pour chaque unité primaire $i \in S_1$, on tire n_i unités secondaires parmi N_i à l'intérieur de l'unité primaire considérée à l'aide de sondages S_{2i} .

En utilisant des notations évidentes, l'estimateur du total s'écrit :

$$\hat{Y} = \sum_{i \in S_1} \sum_{k \in S_{2i}} \frac{Y_{ik}}{p_i p_{k|i}}$$

La variance de cet estimateur peut se décomposer à l'aide de la formule $V(\hat{Y}) = V(E(\hat{Y} | S_1)) + E(V(\hat{Y} | S_1))$ et l'on obtient finalement :

$$V(\hat{Y}) = V \left(\sum_{i \in S_1} \frac{\sum_{k=1}^{N_i} Y_{ik}}{p_i} \right) + \sum_{i=1}^M V \left(\sum_{k \in S_{2i}} \frac{Y_{ik}}{p_{k|i}} \mid S_1 \right)$$

Cette variance peut être estimée par :

$$V(\hat{Y}) = \hat{V} \left(\sum_{i \in S_1} \frac{\sum_{k \in S_{2i}} Y_{ik} / p_{k|i}}{p_i} \right) + \sum_{i \in S_1} \frac{\hat{V} \left(\sum_{k \in S_{2i}} \frac{Y_{ik}}{p_{k|i}} \mid S_1 \right)}{p_i}$$

Chaque variance \hat{V} qui apparaît dans cette formule correspond à un échantillonnage à un degré (soit sur l'échantillon S_1 , soit sur l'échantillon S_{2i} conditionnellement à S_1) pour lequel les outils présentés précédemment s'appliquent.

1.1.5 Plans complexes

Pour des plans d'échantillonnages plus complexes, deux options sont possibles.

La première consiste à estimer la variance théorique en décomposant le plan de sondage au niveau le plus fin. On calcule alors les variances en agrégeant les résultats à chaque étape. Cette démarche est retenue dans le logiciel Poulpe (maintenu à l'U.M.S., voir Caron et al. pour son utilisation, [3]) et qui permet les calculs de variance les plus performants.

La deuxième méthode consiste à simplifier le plan de sondage en ne prenant en compte que les étapes qui semblent les plus importantes en terme de variance. C'est cette démarche que nous avons utilisée pour calculer des variances sur l'*Enquête Revenus Fiscaux* dans la dernière partie de ce document.

1.2 Prise en compte de la non-réponse totale

Les calculs menés ci-dessus supposent que toutes les unités échantillonnées répondent à l'enquête. Ce cas de figure ne se produit malheureusement jamais. Pour tenir compte de la non-réponse totale, on considère le plus souvent que la décision de répondre est aléatoire, auquel cas l'échantillon des répondants est obtenu par un tirage en deux phases : une première phase d'échantillonnage et une deuxième d'acceptation de l'enquête. La deuxième phase est généralement modélisée par un tirage poissonnien : autrement dit, les unités sont supposées décider de répondre indépendamment les unes des autres.

A ce stade, les probabilités de réponse (c'est-à-dire $Pr[k \in R | k \in S]$, notées \mathbf{p}_{kR}) restent inconnues et il s'agit de les estimer. Plusieurs solutions sont possibles. La plus simple est de considérer que la probabilité de réponse est uniforme, auquel cas elle peut être estimée simplement par le taux de réponse global. Cette hypothèse est souvent trop restrictive : dans les enquêtes ménages, les taux de réponse sont par exemple fortement corrélés à l'âge (ainsi, les jeunes ménages répondent moins que les autres). Il est plus raisonnable de supposer que les probabilités de réponse sont constantes par strate, auquel cas on aura, avec des notations évidentes :

$$\mathbf{p}_k^{R_h} = f_h \cong \frac{\# R_h}{\# S_h} = \widehat{\mathbf{p}}_k^{R_h}$$

Enfin, on peut supposer que la probabilité de réponse \mathbf{p}_{kR} est fonction d'une combinaison linéaire de variables explicatives X_k :

$$\mathbf{p}_{kR} = h(X_k \mathbf{b})$$

Si les X_k sont disponibles sur S , on peut estimer \mathbf{b} par un logit ; d'autres solutions sont également possibles (voir Tillé, [12], pour une présentation complète).

Sous l'hypothèse classique que la variance due à l'estimation de la probabilité de réponse est négligeable, on obtient alors :

$$V \left[\sum_{k \in R} \frac{y_k}{\mathbf{p}_k \widehat{\mathbf{p}}_{kR}} \right] \cong V \left[\sum_{k \in S} \frac{y_k}{\mathbf{p}_k} \right] + \sum_{k \in U} \frac{y_k^2 (1 - \widehat{\mathbf{p}}_{kR})}{\mathbf{p}_k \widehat{\mathbf{p}}_{kR}}$$

Le premier terme est la variance due à la phase d'échantillonnage (qui se calcule à l'aide des méthodes présentées dans la partie précédente) et le deuxième la variance due à la phase de réponse.

1.3 Impact du calage

Lorsque certains totaux sont connus sur la population, il est possible d'améliorer les estimations à l'aide d'un calage sur ces totaux. Dans le cas sans non-réponse⁵ On cherche à transformer les poids

⁵Le cas avec non-réponse se traite de la même manière en remplaçant \mathbf{p}_k par $\mathbf{p}_k \mathbf{p}_{kR}$

$1/p_k$ de départ en des poids w_k tels que les estimations des totaux coïncident avec les vrais totaux connus par ailleurs. Ainsi, en notant X_k les variables dont le total X sur la population est connu, on cherche des poids w_k les plus proches de $1/p_k$ au sens d'une certaine distance, et qui vérifient en outre :

$$\sum_{k \in S} w_k X_k = X$$

De tels poids peuvent par exemple être obtenus à l'aide de la macro calmar.

Considérons donc une enquête où les poids ont été transformés par calage. Si on cherche à estimer le total d'une variable Y sur la population : $Y = \sum_{k \in U} Y_k$, l'estimateur calé est donné par :

$$\hat{Y} = \sum_{k \in S} w_k Y_k$$

Deville et Särndal ([4]) ont montré que la variance de cet estimateur est asymptotiquement équivalente à la variance de l'estimateur de Horvitz-Thompson du total des résidus de la régression des Y_k sur X_k . En d'autres termes, si l'on note e_k le résidu associé à l'individu k , on a l'approximation suivante :

$$V \left[\sum_{k \in S} w_k Y_k \right] \cong V \left[\sum_{k \in S} \frac{e_k}{p_k} \right]$$

Les e_k considérés ici sont les « vrais » résidus de la régression sur U des Y_k sur X_k . En pratique, ils sont estimés sur S à partir d'une régression pondérée par les $\frac{1}{p_k}$.

La précision d'un estimateur calé est toujours meilleure que celle d'un estimateur non calé, et le gain de variance est d'autant plus important que les Y_k sont bien corrélés aux X_k . Par exemple, sur le premier trimestre 2003 de l'enquête emploi en continu, l'écart type du nombre de chômeurs est estimé à 52718 lorsque le calage est pris en compte et à 59661 dans le cas contraire. De même, l'écart type du taux de chômage diminue de 0.21% à 0.19% lorsqu'il y a un calage.

1.4 Impact des erreurs de mesure et de l'imputation de la non-réponse partielle

Jusqu'à présent, on a supposé que la variable d'intérêt Y_k était parfaitement mesurée sur les répondants de l'enquête. Cette hypothèse est rarement vérifiée, et ce pour deux raisons :

- les ménages interrogés ne connaissent pas toujours précisément la réponse à la question posée ou ne souhaitent pas donner l'information exacte (par exemple sur les revenus totaux qu'ils ont perçus au cours des douze derniers mois). Nous sommes alors en présence d'erreurs de mesure e_k , la variable mesurée vérifiant $Y_k^* = Y_k + e_k$

- les ménages ne savent pas ou refusent de répondre à la question (sur les questions des revenus, on a ainsi en général 10% de non-réponse). Dans ce cas de non-réponse partielle, Y_k est en général imputé.

On peut également considérer que la variable imputée vérifie $Y_k^* = Y_k + e_k$.

Les erreurs de mesure et l'imputation soulèvent les deux problèmes du biais et de la variance de l'estimateur. S'il n'existe pas de réponse générale, on peut toutefois citer le théorème suivant, dû à Deville ([7]).

Proposition 1. Si pour tout (k,l) de la population, $E(\mathbf{e}_k) = 0$, $E(\mathbf{e}_k \mathbf{e}_l) = 0$ et $V(\mathbf{e}_k) = \mathbf{s}_k^2$, alors le biais de l'estimateur $\sum_{k \in S} Y_k^*$ de $\sum_{k \in U} Y_k$ est nul et sa variance peut être estimée (pratiquement) sans biais par :

$$\sum_{k, l \in S} \frac{Y_k^* Y_l^*}{\mathbf{p}_k \mathbf{p}_l \mathbf{p}_{kl}} (\mathbf{p}_{kl} - \mathbf{p}_k \mathbf{p}_l)$$

Démonstration : On a :

$$E\left(\sum_{k \in S} Y_k^*\right) = E\left(E\left(\sum_{k \in S} Y_k^* \mid S\right)\right) = E\left(\sum_{k \in S} Y_k\right) = \sum_{k \in U} Y_k$$

Par ailleurs :

$$\begin{aligned} V\left(\sum_{k \in S} Y_k^*\right) &= E\left(V\left(\sum_{k \in S} Y_k^* \mid S\right)\right) + V\left(E\left(\sum_{k \in S} Y_k^* \mid S\right)\right) \\ &= E\left(\sum_{k \in S} \frac{\mathbf{s}_k^2}{\mathbf{p}_k}\right) + V\left(\sum_{k \in S} Y_k\right) \\ &= \sum_{k \in U} \frac{\mathbf{s}_k^2}{\mathbf{p}_k} + V\left(\sum_{k \in S} Y_k\right) \end{aligned}$$

Or :

$$\begin{aligned} E\left(\sum_{k, l \in S} \frac{Y_k^* Y_l^*}{\mathbf{p}_k \mathbf{p}_l \mathbf{p}_{kl}} (\mathbf{p}_{kl} - \mathbf{p}_k \mathbf{p}_l)\right) &= E\left(E\left(\sum_{k, l \in S} \frac{Y_k^* Y_l^*}{\mathbf{p}_k \mathbf{p}_l \mathbf{p}_{kl}} (\mathbf{p}_{kl} - \mathbf{p}_k \mathbf{p}_l) \mid S\right)\right) \\ &= E\left(\sum_{k \in S} E(Y_k^{*2} \mid S) \frac{\mathbf{p}_k (1 - \mathbf{p}_k)}{\mathbf{p}_k^3} + \sum_{k \neq l \in S} \frac{Y_k Y_l}{\mathbf{p}_k \mathbf{p}_l \mathbf{p}_{kl}} (\mathbf{p}_{kl} - \mathbf{p}_k \mathbf{p}_l)\right) \\ &= E\left(\sum_{k \in S} \mathbf{s}_k^2 \frac{(1 - \mathbf{p}_k)}{\mathbf{p}_k^2} + \sum_{k, l \in S} \frac{Y_k Y_l}{\mathbf{p}_k \mathbf{p}_l \mathbf{p}_{kl}} (\mathbf{p}_{kl} - \mathbf{p}_k \mathbf{p}_l)\right) \\ &= \sum_{k \in U} \mathbf{s}_k^2 \frac{(1 - \mathbf{p}_k)}{\mathbf{p}_k} + \sum_{k, l \in U} \frac{Y_k Y_l}{\mathbf{p}_k \mathbf{p}_l} (\mathbf{p}_{kl} - \mathbf{p}_k \mathbf{p}_l) \\ &= \sum_{k \in U} \frac{\mathbf{s}_k^2 (1 - \mathbf{p}_k)}{\mathbf{p}_k} + V\left(\sum_{k \in S} Y_k\right) \end{aligned}$$

Si l'on suppose que $1 - \mathbf{p}_k \cong 1$, on a alors :

$$E\left(\sum_{k, l \in S} \frac{Y_k^* Y_l^*}{\mathbf{p}_k \mathbf{p}_l \mathbf{p}_{kl}} (\mathbf{p}_{kl} - \mathbf{p}_k \mathbf{p}_l)\right) \cong V\left(\sum_{k \in S} Y_k^*\right)$$

Ce résultat est très intéressant puisqu'il signifie que lorsque les résidus sont centrés et indépendants, la variance de $\sum_{k \in S} Y_k^*$ est correctement estimée par la formule habituelle de la variance donnée en (1.1.1). Il n'y a pas besoin d'ajouter de termes supplémentaires pour prendre en compte la variance due aux erreurs de mesure.

Néanmoins, dans le cadre général, on peut penser que la formule standard d'estimation de la variance sous-estime la variance réelle en présence d'erreurs de mesure.

2. Estimateurs complexes et linéarisation

2.1 Théorie de la linéarisation

Les calculs de la partie précédente ne peuvent être appliqués que pour des estimations de variance de totaux. En effet, pour des statistiques plus complexes, il devient très difficile de faire un calcul direct et la formule exacte de la variance reste inconnue. Néanmoins, Deville ([6]) a montré qu'il était possible de se ramener à un calcul du type précédent au prix d'un développement limité d'ordre 1. Nous ne donnerons ici que l'intuition de la méthode de linéarisation et les principaux résultats. Nous renvoyons à la lecture de l'article de Deville pour de plus amples informations, en particulier sur les hypothèses techniques assurant la validité des résultats.

Le calcul d'une statistique sur la population U consiste à mettre un poids égal à 1 sur chaque individu k dans U . L'estimation de cette même statistique sur l'échantillon consiste à mettre un poids égal à $\frac{1}{p_k}$ sur les individus k de S et un poids nul pour les autres. Il est donc intéressant de se demander quel est l'effet d'une variation de poids associée à l'individu k sur la statistique que l'on cherche à calculer.

D'un point de vue mathématique, notons M la mesure qui met un poids égal à 1 sur chaque individu, \hat{M} la mesure associée au sondage et $M + t\mathbf{d}_k$ la mesure qui met un poids égal à 1 pour tous les individus sauf pour l'individu k qui a un poids égal à $1+t$. Nous noterons $T(M)$, $T(\hat{M})$ et $T(M + t\mathbf{d}_k)$ les statistiques associées. La dérivée de T liée à une variation infinitésimale de poids associée à l'individu k est appelée fonction d'influence et est égale à :

$$z_k = \lim_{t \rightarrow 0} \frac{T(M + t\mathbf{d}_k) - T(M)}{t} = \text{lin}_k(T)$$

Deville ([6]) a démontré qu'il était possible d'approximer la variance de la statistique $T(\hat{M})$ par la variance du total $\sum_{k \in S} \frac{z_k}{p_k}$:

Proposition 2.1 La variance du total $\sum_{k \in S} \frac{z_k}{p_k}$ est un estimateur de la variance de la statistique complexe $T(\hat{M})$.

Ce résultat est fondamental puisqu'il nous permet pour calculer la variance d'une statistique complexe de se ramener au cas d'un total, cas que l'on sait traiter d'après la partie précédente. La difficulté consiste donc à être capable de calculer cette fonction d'influence. Les règles de calcul sont en fait très proches de celles du calcul différentiel classique.

Nous donnons ici quelques exemples de linéarisation ainsi que des règles de calcul général.

Exemple 1. Soit $T(M) = \sum_{k' \in U} x_{k'}$ un total. On a :

$$T(M + t\mathbf{d}_k) = \sum_{k' \in U} x_{k'} + tx_k$$

Donc :

$$\text{lin}_k = x_k$$

C'est le cas trivial où la linéarisée est égale à la variable d'intérêt elle-même.

Exemple 2. Soient $T(M)$ et $S(M)$ deux statistiques. On a

$$\begin{aligned}\text{lin}_k(T + S) &= \text{lin}_k(T) + \text{lin}_k(S) \\ \text{lin}_k(TS) &= S(M)\text{lin}_k(T) + T(M)\text{lin}_k(S)\end{aligned}$$

et

$$\text{lin}_k\left(\frac{T}{S}\right) = \frac{\text{lin}_k(T)}{S(M)} - \frac{T(M)\text{lin}_k(S)}{S(M)^2}$$

On retrouve les résultats classiques de différentiation.

Exemple 3. Soit $R(M) = \frac{Y(M)}{X(M)}$ le ratio de deux totaux $X(M)$ et $Y(M)$. La formule précédente nous permet d'écrire

$$\text{lin}_k(R) = \frac{1}{X(M)}(y_k - R(M)x_k)$$

On retrouve l'estimation habituelle dans le cas d'un ratio.

D'une manière générale, en cas de doute, il est important de revenir à la définition de la fonction d'influence à savoir le changement infinitésimal du poids de l'individu k . Ce changement infinitésimal porte sur les sommes utilisées dans la statistique : chaque somme apporte une contribution à la fonction d'influence. Pour bien comprendre ce phénomène, nous finirons par un exemple instructif.

Exemple 4. Considérons tout d'abord la statistique $T(M) = \sum_{k' \in U} x_{k'} \log(x_{k'})$. On a :

$$T(M + t\mathbf{d}_k) = \sum_{k' \in U} x_{k'} \log(x_{k'}) + tx_k \log(x_k)$$

La fonction d'influence s'écrit donc :

$$z_k = x_k \log(x_k)$$

Il n'y a qu'un terme dans la fonction d'influence car il n'y qu'une seule somme. Il faut bien voir que c'est le poids de l'individu k qui est modifié et non la variable x_k .

Considérons maintenant la statistique $S(M) = \sum_{k' \in U} x_{k'} \log(\sum_{k'' \in U} y_{k''})$. On a :

$$S(M + t\mathbf{d}_k) = \sum_{k' \in U} x_{k'} \log(\sum_{k'' \in U} y_{k''} + ty_k) + tx_k \log(\sum_{k'' \in U} y_{k''} + ty_k)$$

Donc :

$$\frac{S(M + t\mathbf{d}_k) - S(M)}{t} = \sum_{k' \in U} x_{k'} (\log(\sum_{k'' \in U} y_{k''} + ty_k) - \log(\sum_{k'' \in U} y_{k''})) + x_k \log(\sum_{k'' \in U} y_{k''} + ty_k)$$

En passant à la limite, on obtient la linéarisée :

$$z_k = x_k \log\left(\sum_{k'' \in U} y_{k''}\right) + y_k \frac{\sum_{k' \in U} x_{k'}}{\sum_{k'' \in U} y_{k''}}$$

Le premier terme correspond au terme classique dû à la première somme, mais il ne faut pas oublier le deuxième terme correspondant à la deuxième somme : le changement de poids de l'individu k affecte les deux sommes.

Une dernière difficulté provient du fait que la variable linéarisée z_k n'est pas forcément calculable comme le montre le dernier exemple. En effet, dans la plupart des cas, la variable z_k nécessite d'avoir une information sur la population U dans son entier. Ce n'est évidemment pas le cas, et il faut alors remplacer z_k par son estimation à partir de l'échantillon \hat{z}_k . Le résultat suivant nous assure que cette estimation ne pose pas de difficulté :

Proposition 2.2 La variance du total $\sum_{k \in S} \frac{\hat{z}_k}{p_k}$ est un bon estimateur de la variance de la statistique complexe $T(\hat{M})$.

Ainsi, dans le dernier exemple,

$$z_k = x_k \log\left(\sum_{k'' \in U} y_{k''}\right) + y_k \frac{\sum_{k' \in U} x_{k'}}{\sum_{k'' \in U} y_{k''}} \text{ et } \hat{z}_k = x_k \log\left(\sum_{k'' \in S} \frac{y_{k''}}{p_{k''}}\right) + y_k \frac{\sum_{k' \in S} \frac{x_{k'}}{p_{k'}}}{\sum_{k'' \in S} \frac{y_{k''}}{p_{k''}}}$$

Enfin, l'existence d'un calage ne pose pas de problèmes particuliers dans l'estimation de la variance de statistiques complexes. Si on note \hat{M} la mesure associée aux poids w_k après calage, on peut énoncer le résultat suivant :

Proposition 2.3 La variance du total $\sum_{k \in S} \frac{e_k}{p_k}$ est un estimateur de la variance de la statistique complexe $T(\hat{M})$ où e_k est le résidu associé à l'individu k dans la régression de la variable \hat{z}_k sur les variables x_k de calage.

Deville ([6]) donne des conditions sur les statistiques pour que la théorie de la linéarisation s'applique. Dans la plupart des cas, ces conditions sont vérifiées et on peut donc calculer, grâce à cette technique, les variances de la plupart des estimations réalisées à l'INSEE, même si les statistiques utilisées sont hautement non linéaires.

2.2 Application aux indicateurs d'inégalités

Nous allons appliquer la méthode de linéarisation pour calculer la variance des différents indicateurs d'inégalité⁶. Les calculs de linéarisées seront menés pas à pas pour familiariser le lecteur avec cette technique.

Le lecteur pourra trouver dans l'annexe (5 ?) toutes les macros SAS calculant les estimateurs des indicateurs étudiés (GINI, THEIL, ATKINSON et taux de pauvreté) ainsi que leurs linéarisées.

⁶On ne redonnera pas les propriétés de ces indicateurs, expliquées en détail par exemple par Fleurbaey et Lollivier [9]).

Notons que ces indicateurs, excepté le taux de pauvreté, s'appliquent à des variables Y strictement positives comme le revenu.

2.2.1 Etude du GINI

Définition et estimation du GINI :

l'indice de GINI (G) peut s'écrire comme suit :

$$G(M) = \frac{\sum_{k' \in U} (2r(k') - 1)Y_{k'}}{N \sum_{k' \in U} Y_{k'}} - 1$$

où $r(k')$ est le rang de l'individu k' dans la distribution des Y (triés par ordre croissant). Notons que $r(k')$ se réécrit comme suit :

$$r(k') = \sum_{k'' \in U} 1_{Y_{k''} \leq Y_{k'}}$$

On estime donc $G(M)$ par :

$$G(\hat{M}) = \frac{\sum_{k' \in S} (2\hat{r}(k') - 1)w_{k'}Y_{k'}}{\sum_{k' \in S} w_{k'} \sum_{k' \in S} w_{k'}Y_{k'}} - 1$$

où $\hat{r}(k')$ est l'estimation par substitution de $r(k')$:

$$\hat{r}(k') = \sum_{k'' \in S} w_{k''} 1_{Y_{k''} \leq Y_{k'}}$$

Notons que d'autres choix sont possibles pour l'estimation de $r(k')$ (cf. Deville, [5]), mais leur impact sur l'estimation devient négligeable dès que la taille de l'échantillon est grande (supérieure à 1000 en pratique).

Linéarisation de l'indice de GINI:

nous allons linéariser le numérateur et le dénominateur puis utiliser la règle de linéarisation d'un quotient.

Commençons par le dénominateur qui a la forme la plus simple :

$$\text{lin}_k(\text{dén}) = \text{lin}_k(N) \sum_{k' \in U} Y_{k'} + N \text{lin}_k\left(\sum_{k' \in U} Y_{k'}\right)$$

$$\text{lin}_k(\text{dén}) = \sum_{k' \in U} Y_{k'} + NY_k$$

La linéarisation du numérateur est un peu plus ardue. Remarquons d'abord que :

$$\text{lin}_k(r(k')) = 1_{Y_k \leq Y_{k'}}$$

En raison des deux sommes présentes, on obtient alors deux termes pour le numérateur :

$$\text{lin}_k(\text{num}) = (2r(k) - 1)Y_k + 2 \sum_{k' \in U} 1_{Y_k \leq Y_{k'}} Y_{k'}$$

Or on a :

$$\text{lin}_k \left(\frac{\text{num}}{\text{dén}} \right) = \frac{\text{lin}_k(\text{num})}{\text{dén}} - \frac{\text{num} \times \text{lin}_k(\text{dén})}{\text{dén}^2} = \frac{\text{lin}_k(\text{num}) - (G(M) + 1)\text{lin}_k(\text{dén})}{\text{dén}}$$

Ainsi, la linéarisée du GINI vaut :

$$\text{lin}_k(G) = \frac{2(Y_k r(k) + \sum_{k' \in U} 1_{Y_k \leq Y_{k'}} Y_{k'}) - Y_k - (G(M) + 1) \left(\sum_{k' \in U} Y_{k'} + N Y_k \right)}{N \sum_{k' \in U} Y_{k'}}$$

2.2.2 Etude de l'indicateur d'ATKINSON

Définition et estimation de l'indicateur d'ATKINSON :

L'indicateur d'Atkinson $A_a(M)$, avec $a < 1$, est défini par :

$$A_a(M) = 1 - \left(\frac{1}{N} \sum_{k' \in U} \left(\frac{Y_{k'}}{Y} \right)^a \right)^{\frac{1}{a}}$$

Cet indicateur se réécrit sous la forme suivante :

$$A_a(M) = 1 - \frac{1}{Y} \left(\frac{1}{N} \sum_{k' \in U} Y_{k'}^a \right)^{\frac{1}{a}}$$

Dans le cas où $a = 0$ la formule est indéterminée. On prolonge alors $A_a(M)$ par continuité :

$$A_0(M) = 1 - \frac{\left(\prod_{k' \in U} Y_{k'} \right)^{\frac{1}{N}}}{Y}$$

Pour $a \neq 0$, l'indicateur d'ATKINSON s'estime par :

$$A_a(\hat{M}) = 1 - \left(\frac{1}{\hat{N}} \sum_{k' \in S} w_{k'} \left(\frac{Y_{k'}}{\hat{Y}} \right)^a \right)^{\frac{1}{a}}$$

Une alternative se présente lorsque N est connu : faut-il utiliser cette information sur N dans le calage ou se contenter de l'inclure dans les estimateurs ? La première solution semble préférable en général dans la mesure où les estimateurs sont plus cohérents. L'exemple de l'indice d'Atkinson est à cet égard éloquent.

Dans un premier temps, supposons que N est connu et qu'on utilise les poids de sondage :

$$\hat{A}_{1a} = 1 - \frac{1}{\hat{Y}} \left(\frac{1}{N} \sum_{k' \in S} \frac{Y_{k'}^a}{p_{k'}} \right)^{\frac{1}{a}}$$

Lorsque a tend vers 0 on a : $Y_{k'}^a \longrightarrow 1$. Donc :

$$\frac{1}{N} \sum_{k' \in S} \frac{Y_{k'}^a}{p_{k'}} \xrightarrow{a \rightarrow 0} \frac{1}{N} \sum_{k' \in S} \frac{1}{p_{k'}}$$

Or cette valeur est prise à une puissance qui devient infinie. Donc dès qu'elle est différente de 1, \hat{A}_{1a} tend soit vers $-\infty$, soit vers 1. Cela n'est pas cohérent avec le fait que le « vrai » indice d'ATKINSON se prolonge par continuité en 0 (en une valeur différente de 1). L'estimateur proposé n'a donc aucun sens pour des a proches de 0.

A l'inverse, considérons l'estimateur utilisant les poids obtenus par calage sur N :

$$\hat{A}_{2a} = 1 - \frac{1}{\bar{Y}} \left(\frac{1}{N} \sum_{k' \in S} w_{k'} Y_{k'}^a \right)^{\frac{1}{a}}$$

Cet estimateur est cohérent car :

$$\hat{A}_{2a} \xrightarrow{a \rightarrow 0} 1 - \frac{1}{\bar{Y}} \left(\prod_{k' \in S} Y_{k'}^{w_{k'}} \right)^{\frac{1}{N}}$$

On obtient ici un « bon » estimateur de $A_0(M)$.

Notons enfin que l'estimateur de l'indice d'ATKINSON obtenu en remplaçant N par $\sum_{k' \in S} \frac{1}{p_{k'}}$ est également meilleur que le premier puisqu'il tend également vers un bon estimateur de $A_0(M)$. En d'autres termes, mieux vaut « oublier » l'information sur N plutôt que de mal l'utiliser.

A priori, cette réflexion sur l'indice d'ATKINSON a une certaine valeur de généralité : les meilleurs estimateurs sont ceux obtenus avec des poids de calage. Si un total X est connu, il est préférable de ne pas l'utiliser et de le remplacer par son estimation \hat{X} sur l'échantillon : les estimateurs ainsi obtenus sont plus cohérents.

Linéarisation de l'indicateur :

pour linéariser cet indicateur, on utilise les règles de linéarisation d'un produit et d'une fonction d'une statistique. Cette dernière règle s'énonce comme ceci :

$$\text{lin}_k (f(T)) = f'(T(M)) \text{lin}_k (T)$$

On obtient ici :

$$\text{pour } a \neq 0 : \text{lin}_k (A_a) = \frac{1 - A_a(M)}{N} \left(\frac{Y_k}{\bar{Y}} - 1 - \frac{1}{a} \left(\frac{N Y_k^a}{\sum_{k' \in U} Y_{k'}^a} - 1 \right) \right)$$

Et si les Y_i sont tous non nuls :

$$\text{pour } a = 0 : \text{lin}_k (A_0) = \frac{(1 - A_0(M))}{N} \left(\frac{Y_k}{\bar{Y}} - 1 - \log(Y_k) + \frac{\sum_{k' \in U} \log(Y_{k'})}{N} \right)$$

2.2.3 Etude du THEIL

Définition et estimation du THEIL:

le THEIL, noté T s'exprime de la manière suivante :

$$T(M) = \sum_{k' \in U} \frac{Y_{k'}}{NY} \log \left(\frac{Y_{k'}}{Y} \right)$$

On l'estime donc par :

$$T(\hat{M}) = \frac{\sum_{k' \in S} w_{k'} Y_{k'} \log \left(\frac{Y_{k'}}{Y} \right)}{\sum_{k' \in S} w_{k'} Y_{k'}}$$

Linéarisation du THEIL:

le THEIL peut se réécrire :

$$T(M) = \frac{\sum_{k' \in U} Y_{k'} \log(Y_{k'}) - \sum_{k' \in U} Y_{k'} \log \left(\sum_{k'' \in U} Y_{k''} \right) + \left(\sum_{k' \in U} Y_{k'} \right) \log(N)}{\sum_{k' \in U} Y_{k'}}$$

Pour linéariser cette statistique, nous allons procéder pas à pas en linéarisant chaque morceau puis en utilisant les règles de composition présentées précédemment.

Commençons par le numérateur :

$$\text{lin}_k \left[\sum_{k' \in U} Y_{k'} \log(Y_{k'}) \right] = Y_k \log(Y_k)$$

$$\text{lin}_k \left[\sum_{k' \in U} Y_{k'} \log \left(\sum_{k'' \in U} Y_{k''} \right) \right] = Y_k \log \left(\sum_{k'' \in U} Y_{k''} \right) + Y_k$$

$$\text{lin}_k \left[\left(\sum_{k' \in U} Y_{k'} \right) \log(N) \right] = Y_k \log(N) + \frac{1}{N} \sum_{k' \in U} Y_{k'}$$

On peut donc en déduire la linéarisation du numérateur :

$$\text{lin}_k(\text{num}) = Y_k [\log(Y_k) - \log(\bar{Y})] + \bar{Y} - Y_k$$

La linéarisation du dénominateur est plus simple :

$$\text{lin}_k(\text{dén}) = Y_k$$

et en combinant ces deux résultats, nous pouvons enfin obtenir la linéarisation du THEIL :

$$\boxed{\text{lin}_k(T) = \frac{1}{\sum_{k \in U} Y_k} [Y_k (\log(Y_k) - \log(\bar{Y})) + \bar{Y} - Y_k - T(M) Y_k]}$$

2.4 Etude du Taux de pauvreté (*Poverty Headcount*)

Définition

Le taux de pauvreté se définit comme la proportion d'individus dont le revenu est inférieur au seuil de pauvreté.

Deux situations se distinguent, suivant que le seuil de pauvreté est connu de façon exogène (par exemple par une enquête suffisamment grande pour négliger la variance de l'estimateur) ou estimé à partir de l'enquête. Par ailleurs, le seuil peut être évalué sur une population plus large que le champ de l'étude : ainsi le taux de pauvreté des personnes âgées est calculé à partir du seuil de pauvreté définie sur la France entière.

Seuil de pauvreté exogène :

Traisons tout d'abord le cas le plus simple du seuil de pauvreté s exogène. Le taux de pauvreté de la population A ⁷ s'écrit alors :

$$J(M) = F_A(M, s)$$

F_A est la fonction de répartition de Y sur la population A considérée :

$$F_A(y) = \frac{1}{N_A} \sum_{k \in A} \mathbf{1}_{Y_k \leq y}$$

F_A peut également être considéré comme un ratio sur U :

$$F_A(y) = \frac{\sum_{k \in U} \mathbf{1}_{Y_k \leq y} \mathbf{1}_{k \in A}}{\sum_{k \in U} \mathbf{1}_{k \in A}}$$

$J(M)$ s'estime simplement par :

$$J(\hat{M}) = F_A(M, s) = \frac{1}{\sum_{k \in S_A} w_k} \sum_{k \in S_A} w_k \mathbf{1}_{Y_k \leq s}$$

F_A étant un simple ratio sur U , la linéarisation de $J(M)$ s'écrit directement :

$$\boxed{\lim_k (J) = \frac{\mathbf{1}_{k \in A}}{N_A} (\mathbf{1}_{Y_k \leq s} - J(M))}$$

Seuil de pauvreté endogène :

On considère maintenant le cas où le seuil de pauvreté est estimé à partir de l'enquête. Le seuil le plus souvent considéré est la moitié du revenu médian (ou parfois 60% de ce revenu médian, notamment dans les travaux réalisés par EUROSTAT). On va par la suite considérer l'indicateur plus général suivant :

$$J_{(a,b)}(M) = F_A(M, \mathbf{b}F^{-1}(M, \mathbf{a}))$$

où $F^{-1}(M, \mathbf{a})$ est le quantile d'ordre \mathbf{a} défini sur la population totale. Le taux de pauvreté « classique » correspond à $\mathbf{a} = 0,5$ et $\mathbf{b} = 0,5$ voire $0,6$.

⁷ A étant éventuellement différent de U champ de l'enquête, cf. l'exemple des personnes âgées.

F étant une fonction en escalier, F^{-1} n'est a priori pas définie. Ici, on adoptera la définition par défaut de SAS :

$$F^{-1}(M, \mathbf{a}) = Y_{(\lfloor N\mathbf{a} \rfloor + 1)} \text{ si } \lfloor N\mathbf{a} \rfloor \neq N\mathbf{a}$$

$$F^{-1}(M, \mathbf{a}) = \frac{Y_{(\lfloor N\mathbf{a} \rfloor)} + Y_{(\lfloor N\mathbf{a} \rfloor + 1)}}{2} \text{ si } \lfloor N\mathbf{a} \rfloor = N\mathbf{a}$$

où $\lfloor \cdot \rfloor$ est la fonction partie entière.

$J_{(a,b)}(M)$ est alors estimée par :

$$J_{(a,b)}(\hat{M}) = F_A(\hat{M}, \mathbf{b}F^{-1}(\hat{M}, \mathbf{a}))$$

avec, par exemple lorsque $\lfloor N\mathbf{a} \rfloor \neq N\mathbf{a}$:

$$F^{-1}(\hat{M}, \mathbf{a}) = Y_{(\lfloor \hat{N}\mathbf{a} \rfloor + 1)}$$

$Y_{(k)}$ correspondant ici aux Y ordonnés sur l'échantillon.

La linéarisation de cet indicateur est délicate. Pour mieux comprendre le calcul qui va suivre, réécrivons cet indicateur :

$$J_{(a,b)}(M) = F_A(M, G(M))$$

où $G(M) = \mathbf{b}F^{-1}(M, \mathbf{a})$. On note également F la fonction de répartition de Y sur l'ensemble de la population.

On va supposer dans la suite, pour faciliter les calculs, que F_A et F sont continues dérivables (et donc que F^{-1} est défini sans ambiguïté). En revenant à la définition de la fonction d'influence, on obtient la formule de linéarisation d'une composée de statistiques dépendant de M :

$$\text{lin}_k(J_{(a,b)}) = \text{lin}_k(F_A(M, G(M))) = \text{lin}_k(F_A)(G(M)) + F_A(M, G(M))\text{lin}_k(G)$$

Le premier terme a été calculé précédemment :

$$\text{lin}_k(F_A)(G(M)) = \frac{1_{k \in A}}{N_A} (1_{Y_k \leq G(M)} - J_{(a,b)}(M))$$

Dans le deuxième terme, on doit calculer $\text{lin}_k(G)$. Pour cela remarquons que :

$$F^{-1}(M, F(M, y)) = y$$

valeur qui ne dépend pas de M . Donc :

$$\text{lin}_k(F^{-1}(M, F(M, y))) = 0$$

On utilise alors la formule précédente de linéarisation d'une composée :

$$0 = \text{lin}_k(F^{-1}(M, F(M, y))) = \text{lin}_k(F^{-1})(F(M, y)) + (F^{-1})'(M, F(M, y))\text{lin}_k(F)(y)$$

Et la formule de dérivation de la fonction réciproque fournit :

$$\text{lin}_k (F^{-1})(y) = -\frac{\text{lin}_k (F)(F^{-1}(y))}{F'(M, F^{-1}(M, y))}$$

Ainsi, on a :

$$\text{lin}_k (G) = -\mathbf{b} \frac{1_{Y_k \leq F^{-1}(M, \mathbf{a})} - \mathbf{a}}{NF'(M, F^{-1}(M, \mathbf{a}))}$$

Finalement, la linéarisée de J s'écrit :

$$\boxed{\text{lin}_k (J_{(a,b)}) = \frac{1_{k \in A}}{N_A} \left[1_{Y_k \leq \mathbf{b}F^{-1}(M, \mathbf{a})} - J_{(a,b)}(M) \right] - \frac{\mathbf{b}F'_A(M, \mathbf{b}F^{-1}(M, \mathbf{a}))}{NF'(M, F^{-1}(M, \mathbf{a}))} (1_{Y_k \leq F^{-1}(M, \mathbf{a})} - \mathbf{a})}$$

Dans le cas particulier où $A=U$, c'est-à-dire lorsqu'on s'intéresse au taux de pauvreté sur la population entière, on obtient :

$$\text{lin}_k (J_{(a,b)}) = \frac{1}{N} \left[1_{Y_k \leq \mathbf{b}F^{-1}(M, \mathbf{a})} - J_{(a,b)}(M) - \frac{\mathbf{b}F'(M, \mathbf{b}F^{-1}(M, \mathbf{a}))}{F'(M, F^{-1}(M, \mathbf{a}))} (1_{Y_k \leq F^{-1}(M, \mathbf{a})} - \mathbf{a}) \right]$$

Pour le calcul proprement dit des linéarisées, on adopte la définition précédente de $F^{-1}(M, \mathbf{a})$ et de $F^{-1}(\hat{M}, \mathbf{a})$.

Enfin, la fonction F_A n'existe pas a priori⁸. Cependant, on peut considérer qu'elle n'est qu'un estimateur d'une "vraie" fonction F_A dérivable. Une approximation de sa dérivée est alors obtenue via l'estimateur par le noyau :

$$F_A(M, y) = \frac{1}{N_A} \sum_{k \in A} K(y - Y_k)$$

où K est un fonction de densité centrée en 0.

Dans la macro de linéarisation, le choix s'est porté sur un noyau gaussien :

$$F'_A(M, y) = \frac{1}{N_A h_A} \sum_{k \in A} \mathbf{f}\left(\frac{y - Y_k}{h_A}\right)$$

où \mathbf{f} est la densité d'une loi normale centrée réduite.

Plusieurs méthodes sont possibles pour choisir h_A , fenêtres du noyau. La macro adopte par défaut la "règle du pouce"⁹ :

$$h_A = \frac{\mathbf{s}_A}{N_A^{1/5}}$$

avec :

$$\mathbf{s}_A^2 = \frac{1}{N_A} \sum_{k \in A} (y_k - \bar{y}_A)^2$$

⁸ce qui suit s'applique à F également.

⁹En pratique, le choix de h_A n'a qu'une influence mineure : l'estimation de la linéarisée n'est pratiquement pas affectée par une multiplication ou division par 10 du h_A obtenu par la règle du pouce.

2.3 Comparaison des techniques de linéarisation avec le *bootstrap*

Déterminer un intervalle de confiance par une technique de linéarisation pose un certain nombre de problèmes. En particulier, pour des statistiques autres que des fonctions de moyennes, les lois des estimateurs ne sont pas connues, même asymptotiquement. On postule en général qu'elles sont normales, sans jamais pouvoir le vérifier. Les hypothèses fortes qui assurent la validité de la linéarisation sont également difficiles à tester.

Il est donc légitime de se tourner vers d'autres méthodes. Le *bootstrap* est l'une d'entre elles, sans doute la plus connue.

2.3.1 Principe du *bootstrap*

Le cadre premier du *bootstrap* est celui de la statistique classique : l'échantillon est constitué de n variables aléatoires (X_1, \dots, X_n) i.i.d. Leur fonction de répartition F est supposée inconnue. On cherche à estimer la loi de $T_n = T(X_1, \dots, X_n)$, T_n étant un estimateur d'une grandeur \mathbf{q} . Cela permettrait en particulier d'estimer sa variance, son biais, des intervalles de confiance... L'idée pour cela est de remplacer F par la fonction de répartition empirique F_n obtenue à partir de l'échantillon :

$$F_n(x) = \frac{1}{n} \sum_{k=1}^n 1_{X_k \leq x}$$

Ce principe de substitution (ou plug-in) est justifiée asymptotiquement puisqu'on a presque-sûrement, d'après le théorème Givenko-Cantelli :

$$\sup |F_n(x) - F(x)| \longrightarrow 0$$

On étudie donc les propriétés de T_n par l'intermédiaire de $T_n^* = T(X_1^*, \dots, X_n^*)$ où les X_k^* sont des variables i.i.d de fonction de répartition F_n .

L'espérance de T_n :

$$E[T_n] = \int \dots \int T(x_1, \dots, x_n) dF(x_1) \dots dF(x_n)$$

est ainsi estimée par :

$$E[T_n^*] = \int \dots \int T(x_1, \dots, x_n) dF_n(x_1) \dots dF_n(x_n)$$

Cela s'écrit également :

$$E[T_n^*] = \frac{1}{n^n} \int \dots \int T(x_1, \dots, x_n) \sum_{l_1=1}^n d_{X_{l_1}}(x_1) \dots \sum_{l_n=1}^n d_{X_{l_n}}(x_n) dx_1 \dots dx_n$$

$$E[T_n^*] = \frac{1}{n^n} \sum_{l_1=1}^n \dots \sum_{l_n=1}^n T(X_{l_1}, \dots, X_{l_n})$$

Il s'agit tout simplement de la moyenne de T sur l'ensemble des échantillons de taille n tirés avec remise à partir de l'échantillon initial (X_1, \dots, X_n) . Sur ce même principe, on peut donc, théoriquement en tout cas, fournir une estimation de la loi de T_n .

Malheureusement, la formule précédente montre qu'il est nécessaire en général de calculer n^n termes (C_{2n-1}^n en réalité si T est symétrique), ce qui est en pratique impossible dès que n dépasse la dizaine.

Pour s'en sortir, on a recours à des simulations dites de Monte-Carlo. En clair, on effectue un grand nombre de tirages avec remise à partir de l'échantillon initial, et on calcule l'estimateur pour chacun de ceux-ci. Au prix de cette double approximation (estimation de F par F_n , puis approximation du calcul analytique par simulation), on dispose alors d'un estimateur pour l'ensemble des caractéristiques de T_n .

Par exemple, on estime $E[T_n]$ par la moyenne empirique sur l'ensemble des tirages :

$$E_B[T_n^*] = \frac{1}{B} \sum_{k=1}^B T_n^{*(k)}$$

où B est le nombre de tirages avec remise effectués et $T_n^{*(k)}$ la statistique obtenue au k -ième tirage.

De même, la fonction de répartition G de T_n est estimée par :

$$G_B^*(x) = \frac{1}{B} \sum_{k=1}^B 1_{T_n^{*(k)} \leq x}$$

L'avantage de cette méthode est donc qu'il est possible, moyennant un jeu d'hypothèses très faibles (et en particulier, aucune paramétrisation de la loi des X_k), de fournir une estimation convergente de la loi de T_n . Cela permet d'obtenir des intervalles de confiance sur q sans postulat de normalité.

2.3.2 Application aux sondages

Comment bootstraper en sondages ? : en statistique inférentielle, la méthode du *bootstrap* repose donc sur l'équivalence asymptotique entre tirage avec remise dans l'échantillon et tirage avec remise dans la population (que celle-ci soit finie ou non). La statistique étudiée se comporte alors de la même manière que l'on tire dans l'échantillon ou dans la population. En sondages, cette idée n'est pas applicable directement dans la mesure où les échantillons issus de la population sont tirés sans remise : un tirage avec remise dans l'échantillon ne mimera donc pas correctement le tirage initial. Ainsi, pour T_n moyenne de la variable Y et dans le cas d'un SAS de taille n , on aura :

$$V(T_n^*) = \left(1 - \frac{1}{n}\right) \frac{s^2}{n}$$

Cet estimateur est biaisé car

$$E[V(T_n^*)] = \left(1 - \frac{1}{n}\right) \frac{S^2}{n} \neq \left(1 - \frac{n}{N}\right) \frac{S^2}{n}$$

Le biais peut être élevé : par exemple, pour un SAS au quart de la population, la variance est en moyenne surestimée d'un facteur $4/3$ environ¹⁰. Le bootstrap ne tient pas non plus compte des différences dans les poids de tirage : si ceux-ci sont bien corrélés avec la variable d'intérêt, la variance estimée par *bootstrap* sera très supérieure à la variance réelle.

Plusieurs solutions ont été proposées pour tenter de pallier cet inconvénient¹¹. Une des méthodes consiste à dupliquer les individus de l'échantillon autant de fois que leur poids afin de construire une population fictive d'une taille proche de N . On tire alors dans cette population un grand nombre

¹⁰Ce cas n'est pas irréaliste, si l'on songe par exemple aux tirages dans des petites strates.

¹¹On trouvera dans Cabeça ([2]) une revue de la littérature sur la question.

d'échantillons suivant la procédure suivie pour tirer l'échantillon initial. Le calcul de T_n sur chacun de ces échantillons permet notamment d'obtenir, comme dans le cadre classique, un estimateur de la fonction de répartition de T_n . Il est donc possible d'obtenir des intervalles de confiance de q sans hypothèse de normalité sur T_n .

Cette idée séduisante de réplication des individus et du mode de tirage n'est cependant pas appuyée mathématiquement. A distance finie, l'estimateur de la variance d'une moyenne est ainsi biaisé. Dans le cas d'un SAS, on obtient par exemple :

$$V(T_n^*) = \frac{1-1/n}{1-1/N} \left(1 - \frac{n}{N}\right) \frac{s^2}{n}$$

Cette méthode reste malgré tout préférable à la première dans la mesure où, dans cet exemple précis de la variance de l'estimateur de la moyenne dans un SAS, l'estimation est asymptotiquement convergente. Mais il n'existe pas, à notre connaissance, de théorème assurant cette convergence de façon générale, c'est-à-dire pour des statistiques ou des plans de sondage complexes.

Un autre inconvénient du *bootstrap* en sondage réside dans la lourdeur des procédures informatiques. Ainsi, dans le cas d'un échantillon tiré avec des probabilités inégales et redressé ensuite par calage (c'est-à-dire le cas habituel des enquêtes ménages), il est nécessaire d'effectuer les opérations suivantes :

1. répliquer chaque ménage de l'échantillon suivant l'approximation entière de son poids de calage w_k . On obtient alors une population d'environ 23 millions de ménages
2. puis, un grand nombre de fois (pour un échantillon de plusieurs milliers d'individus, on conseille au moins 100000) :
 - (a) tirer dans la population fictive un échantillon suivant la procédure adoptée pour l'échantillon initial
 - (b) redresser les poids de tirage par un calage identique au calage de l'échantillon initial
 - (c) calculer la statistique T_n

Cette procédure est donc très coûteuse en temps machine.

Exemple de comparaison avec la linéarisation:

la comparaison porte sur l'estimateur de l'indice de GINI des revenus simulés issus de l'enquête PCV de mai 2002. L'objectif ici n'est pas de calculer un « vrai » estimateur de précision sur l'enquête PCV mais bien de comparer *bootstrap* et linéarisation sur des données d'enquête. On a donc supposé que l'échantillon des répondants était issu d'un sondage aléatoire simple. Plus précisément :

1. côté *bootstrap*, la méthode utilisée est un tirage avec remise. 65 000 échantillons environ ont ainsi été simulés.
2. côté linéarisation, les linéarisés ont été obtenus en utilisant des poids tous égaux (l'inverse du taux de sondage), et la variance a été calculée comme si le tirage était un SAS.

le graphique 1 présente les estimations de la densité de l'estimateur obtenus par *bootstrap* et linéarisation.

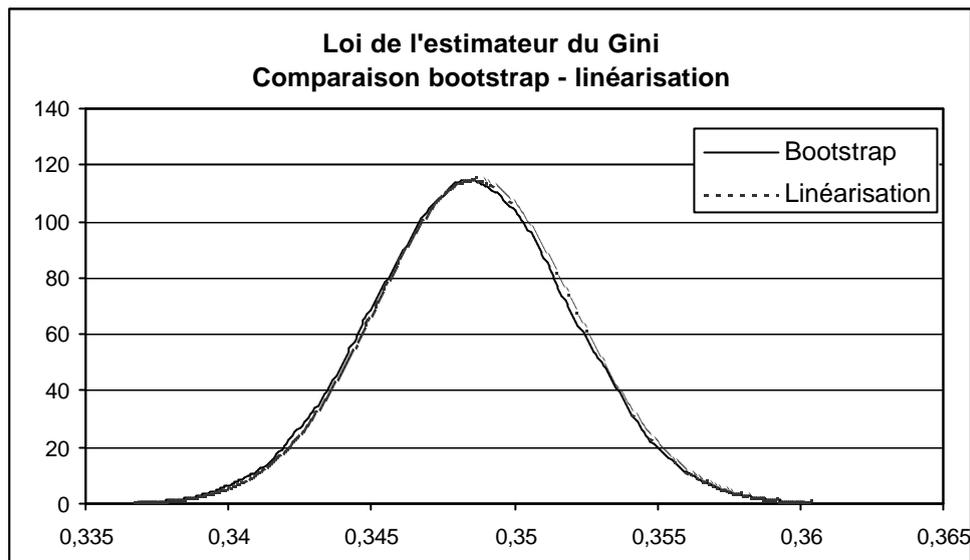


Figure 1. Comparaison des estimations de la densité de l'estimateur par *bootstrap* et par linéarisation

Comme on peut le constater, sur l'exemple considéré, les résultats sont très proches. Deux points en particulier sont très satisfaisants. D'une part, le biais de l'estimateur, estimé par *bootstrap*¹², est négligeable devant l'écart-type : $8,6 \times 10^{-5}$ contre $3,5 \times 10^{-3}$. Ceci est rassurant car la méthode de la linéarisation ne permet pas d'estimer un biais éventuel. D'autre part, l'hypothèse de normalité, nécessaire pour construire des intervalles de confiance dans le cadre de la linéarisation, est acceptable au vu de la distribution obtenue par *bootstrap*. En particulier, les tests de Kolmogorov-Smirnov et d'Anderson-Darling ne sont pas rejetés au seuil de 5%. Ces deux points, ainsi que des estimateurs d'écart-type très proches ($3,48 \times 10^{-3}$ d'après le *bootstrap*, $3,47 \times 10^{-3}$) conduisent à des intervalles de confiance pratiquement confondus : [0,343,0,354] selon le *bootstrap*, [0,342,0,355] selon la linéarisation.

Pour conclure, ces deux méthodes se confirment l'une l'autre dans cet exemple empirique. Le *bootstrap*, qui n'est pas justifié théoriquement, semble donner des résultats satisfaisants. Par ailleurs, le postulat de normalité des estimateurs semble vérifié et valide donc la méthode de linéarisation. Au final, cette dernière est préférable au *bootstrap* étant donné la rigueur du cadre théorique et la simplicité de sa mise en oeuvre informatique.

Caractéristiques	Bootstrap	Linéarisation
Estimation du biais de l'estimateur	Oui	Non (supposé négligeable devant l'écart-type)
Estimation de la loi de l'estimateur	Oui	Non (supposée normale)
Respect du plan de sondage	Non en général	Oui
Justifié asymptotiquement	Non	Oui
Rapidité des calculs	Non	Oui

Table 1. Avantages comparés de la linéarisation et du *bootstrap*

¹²Ce biais est calculé en soustrayant l'estimateur obtenu sur l'échantillon initial à la moyenne des estimateurs obtenus sur les boucles *bootstrap*.

3. Exemple de calcul de variance d'une statistique complexe

3.1 Introduction : démarche générale adoptée

Le but de cette section est d'illustrer les méthodes exposées en parties 1 et 2 de ce document par une application concrète sur une enquête particulière : connaître la précision des taux de pauvreté estimés sur l'*Enquête Revenus Fiscaux*. Nous cherchons donc à estimer la variance de l'estimateur du taux de pauvreté employé.

Pour le statisticien qui s'attèle à cette tâche, trois problèmes surviennent:

1. le taux de pauvreté est une statistique non linéaire : l'estimation de la variance de son estimateur nécessite la *linéarisation* de ce dernier.
2. l'*Enquête Revenus Fiscaux* est adossée à l'*Enquête Emploi*, qui est réalisée suivant un plan de sondage particulier qui ne peut a priori être assimilé à un sondage aléatoire simple (effets de grappe liés au plan de sondage aréolaire et au calcul des taux de pauvreté sur la distribution des revenus individuels¹³).
3. outre le phénomène de non-réponse propre à l'*Enquête Emploi*, l'*Enquête Revenus Fiscaux* ne repose cependant pas sur l'intégralité de l'échantillon issu de l'*Enquête Emploi*¹⁴. Des redressements pour « non-réponse totale » sont réalisés d'une part sur l'*Enquête Emploi* avant appariement, et d'autre part sur l'*Enquête Revenus Fiscaux* après appariement, redressements qui doivent être pris en compte dans l'estimation de la précision des estimations réalisées.

La linéarisation du taux de pauvreté ayant été réalisée et implémentée dans un cadre général, il faut cependant l'adapter à une particularité récente des calculs de taux de pauvreté à l'INSEE: l'utilisation de distributions « individus » issues d'enquêtes « ménages ». Reste ensuite à prendre en compte plan de sondage et calage dans ce qu'ils ont de spécifique à l'*Enquête Revenus Fiscaux*.

3.2 Prise en compte de l'effet grappe spécifique aux ménages

Les taux de pauvreté calculés par l'INSEE, comme ceux publiés par EUROSTAT d'ailleurs, reposent sur la notion de revenu disponible (*i.e.* revenu après prestations et impôts). Le revenu disponible ne peut donc être connu qu'au niveau d'un ménage. Or c'est également le ménage qui constitue l'unité statistique que l'on retient dans les enquêtes.

Toutefois, les taux de pauvreté calculés par l'INSEE comme par EUROSTAT reposent sur des distributions de niveaux de vie estimées sur des populations d'individus: on déduit de l'échantillon ménage un échantillon des individus qui y vivent¹⁵. Cette convention de calcul a cependant une conséquence importante sur la précision des estimations réalisées: elle introduit artificiellement des «grappes » supplémentaires puisque dans un même ménage, tout le monde est pauvre ou personne ne l'est. Le gain en précision qui découle de l'augmentation mécanique des observations (passage de 70.000 ménages environ à 160.000 individus dans *ERF*) a donc une contrepartie négative qu'il faut prendre en compte.

¹³L'INSEE a en effet pris le parti de calculer dorénavant les taux de pauvreté sur la distribution des niveaux de vie des individus. Ce choix a pour conséquence la génération de grappes (la pauvreté est mise en évidence au niveau du ménage et tous les individus du ménage sont ensuite déclarés pauvres).

¹⁴Rappel sur la constitution de l'enquête *Revenu Fiscaux*: les individus répondants à l'*Enquête Emploi* sont appariés avec leur déclaration de revenus (2042) fournies par la DGI. Pour certains ménages présents dans l'enquête, on ne retrouve aucune déclaration d'impôts; ces échecs d'appariement sont assimilés ici à de la non-réponse totale.

¹⁵Chacun dispose du niveau de vie du ménage (revenu disponible qui a été déflaté par une échelle d'équivalence pour prendre précisément en compte la taille et la structure du ménage).

Deux approches équivalentes peuvent être adoptées pour résoudre ce problème. Soit rajouter un degré au plan de sondage (la dernière grappe est alors le ménage), soit prendre en compte dans le calcul de la linéarisée le fait que la distribution utilisée est déduite de la « vraie » distribution « ménage » par une transformation simple. Pour des raisons de simplicité, nous avons adopté cette seconde démarche que nous détaillons ici.

On définit un taux de pauvreté « individus » J_{ind} de la manière suivante:

$$J_{\text{ind}}(\mathbf{a}, \mathbf{b}) = F_{\text{ind}}\left(\mathbf{b}F_{\text{ind}}^{-1}(\mathbf{a})\right)$$

où F_{ind} et F_{ind}^{-1} sont définies sur la population des individus. Par exemple :

$$F_{\text{ind}}(x) = \frac{1}{N_{\text{ind}}} \sum_{i \in U_{\text{ind}}} 1_{(y_i \leq x)} = \frac{1}{\sum_{i \in U_{\text{men}}} n_i} \sum_{i \in U_{\text{men}}} n_i 1_{(y_i \leq x)}$$

F_{ind} (et de même F_{ind}^{-1} et $J_{\text{ind}}(\mathbf{a}, \mathbf{b})$) peut donc être estimé au niveau ménage en utilisant non pas les pondérations « ménages » mais les pondérations « individus » (c'est-à-dire les pondérations ménages multipliées par le nombre d'individus dans le ménage). En d'autres termes :

$$\widehat{J}_{\text{ind}}(x) = \widehat{J}_{\text{men}}(\mathbf{a}, \mathbf{b}, nw)$$

Le même raisonnement tient également pour la linéarisée de $J_{\text{ind}}(\mathbf{a}, \mathbf{b})$. Ainsi :

$$\text{lin}_k F_{\text{ind}}(x) = \sum_{i \in U_{\text{men}}} \frac{n_k}{n_i} (1_{Y_k \leq y} - F_{\text{ind}}(y))$$

Et donc:

$$\text{lin}_k F_{\text{ind}}^{-1}(\mathbf{a}) = -\frac{\text{lin}_k F_{\text{ind}}(F_{\text{ind}}^{-1}(\mathbf{a}))}{F'_{\text{ind}}(F_{\text{ind}}^{-1}(\mathbf{a}))} = -\frac{n_k (1_{y_k \leq F_{\text{ind}}^{-1}(\mathbf{a})} - \mathbf{a})}{\left(\sum_{i \in U_{\text{men}}} n_i\right) F'_{\text{ind}}(F_{\text{ind}}^{-1}(\mathbf{a}))}$$

Soit :

$$\text{lin}_k J_{\text{ind}}(\mathbf{a}, \mathbf{b}) = \sum_{i \in U_{\text{men}}} \frac{n_k}{n_i} \left[1_{Y_k \leq \mathbf{b}F_{\text{ind}}^{-1}(\mathbf{a})} - J_{\text{ind}}(\mathbf{a}, \mathbf{b}) - \mathbf{b} \frac{F'_{\text{ind}}(\mathbf{b}F_{\text{ind}}^{-1}(\mathbf{a}))}{F'_{\text{ind}}(F_{\text{ind}}^{-1}(\mathbf{a}))} (1_{Y_k \leq \mathbf{b}F_{\text{ind}}^{-1}(\mathbf{a})} - \mathbf{a}) \right]$$

Et finalement, en passant à l'estimation:

$$\widehat{\text{lin}}_k J_{\text{ind}}(\mathbf{a}, \mathbf{b}, w) = n_k \widehat{\text{lin}}_k J_{\text{men}}(\mathbf{a}, \mathbf{b}, nw)$$

Autrement dit, l'estimation de la linéarisée du taux de pauvreté individuel est égale à l'estimation de la linéarisée du taux de pauvreté ménage (obtenu avec des poids individus) pondérée par la taille du ménage.

3.3 Variance sous le plan de sondage de l'Enquête Emploi Annuelle

3.3.1 Rappels sur le plan de sondage de l'Enquête Emploi Annuelle

Si l'on fait abstraction du détail de sa mise en œuvre (plan à 3 degrés), le plan de sondage de l'Enquête Emploi est fondamentalement un sondage aréolaire *i.e.* un sondage par grappes géographiquement déterminées (cf. Roth, [11] et également [10]).

Les unités primaires sont des *groupes d'aires*¹⁶ qui comprennent théoriquement $4 \times 40 = 160$ logements dans le cas des *Unités urbaines* (dorénavant *UU*) de moins de 100.000 habitants et $4 \times 20 = 80$ logements dans le cas des *UU* de plus de 100.000 habitants¹⁷. On tire d'abord les *groupes d'aires* avec un taux de sondage de $1/75$ pour avoir finalement des aires tirées avec un taux de $1/300$. On a en 1999 3.270 aires (soit environ $1.425 \times 40 + 1.845 \times 20 = 93.900$ logements). Le tirage des groupes d'aires se fait après stratification par région (21 régions: PACA et la Corse étant rassemblées) et par tranche d'unité urbaine (10 tranches, Paris étant à part).

On décide alors de « modéliser » le plan de sondage l'*Enquête Revenus Fiscaux* comme un sondage de grappes: on tire un échantillon S_a de m aires parmi M , par un SAS dans chaque strate (avec un taux de sondage de $1/300$ en moyenne¹⁸).

Dans chacune de ces unités primaires tirées, on réalise un sondage poissonnien R au taux f_s . Cette deuxième phase rend compte de la non-réponse à l'*Enquête Emploi* ainsi que des échecs d'appariement avec les données DGI qui peuvent survenir lors de la construction de l'*Enquête Revenus Fiscaux*.

3.3.2 Calcul de la variance

La variance s'estime alors de la façon suivante: on note $Y_{g,i}$ la variable d'intérêt, et $Y_g = \sum_{i=1}^{N_g} Y_{g,i}$ le total sur la grappe.

L'estimateur d'HORVITZ-THOMPSON du total s'écrit dans chaque strate (on pondère ensuite par le poids relatif des strates dans la population):

$$Y_{HT} = \sum_{g \in S_a} \sum_{i \in R} \frac{1}{p_{g,i}} Y_{g,i} = \frac{M}{mf_s} \sum_{g \in S_a} \sum_{i \in R} Y_{g,i}$$

En effet, la probabilité de l'individu (ou du ménage) i de l'aire g s'écrit:

$$p_{g,i} = \mathbf{P}[\text{être tiré}] \times \mathbf{P}[\text{répondre}] = \frac{m}{M} f_s$$

On calcule ensuite la variance en conditionnant par S_a :

$$\mathbf{V}[Y_{HT}] = \mathbf{E}[\mathbf{V}(Y_{HT} | S_a)] + \mathbf{V}[\mathbf{E}(Y_{HT} | S_a)]$$

Calcul du premier terme:

$$\mathbf{V}(Y_{HT} | S_a) = \mathbf{V}\left(\frac{M}{mf_s} \sum_{g \in S_a} \sum_{i \in R} Y_{g,i} | S_a\right) = \frac{M^2}{m^2} \sum_{g \in S_a} \mathbf{V}\left(\frac{1}{f_s} \sum_{i \in R} Y_{g,i} | S_a\right) = \frac{M^2}{m^2} \sum_{g \in S_a} \mathbf{V}[\hat{Y}_{HT, \text{poissonnien}}]$$

$$\text{or } \mathbf{V}[\hat{Y}_{HT, \text{poissonnien}}] = \sum_{i=1}^{N_g} \frac{Y_k^2}{f_s} (1 - f_s)$$

¹⁶Elles sont au nombre de 4: 3 à cause de la rotation sur 3 ans, plus une aire de *réserve*.

¹⁷Pour limiter les effets de grappe cette diminution de la taille des grappes n'a été opérée cependant que pour ces *UU* de grande taille de façon à limiter les coûts.

¹⁸Pour des raisons pratiques, nous avons dû supposer ce taux de sondage identique dans chaque strate, égal à $1/300$.

on a donc :

$$\mathbf{V}(Y_{HT} | S_a) = \frac{M^2}{m^2} \frac{1-f_s}{f_s} \sum_{g \in S_a} \left(\sum_{i=1}^{N_g} Y_{g,i}^2 \right)$$

finalemt:

$$\mathbf{E}[\mathbf{V}(Y_{HT} | S_a)] = \mathbf{E} \left[\frac{M^2}{m^2} \frac{1-f_s}{f_s} \sum_{g \in S_a} \left(\sum_{i=1}^{N_g} Y_{g,i}^2 \right) \right] = \frac{M}{m} \frac{1-f_s}{f_s} \sum_{g=1}^M \sum_{i=1}^{N_i} Y_{g,i}^2$$

en effet $\frac{M}{m} \sum_{g \in S_a} \left(\sum_{i=1}^{N_i} Y_{g,i}^2 \right)$ estime sans biais (sous un SAS) $\sum_{g=1}^M \left(\sum_{i=1}^{N_g} Y_{g,i}^2 \right)$

Cette quantité s'estime par:

$$\frac{M^2}{m^2} \frac{1-\hat{f}_s}{\hat{f}_s^2} \sum_{g \in S_a} \sum_{i \in R} Y_{g,i}^2$$

Calcul du second terme :

$$\mathbf{E}(Y_{HT} | S_a) = \left[\frac{M}{mf_s} \sum_{g \in S_a} \sum_{i \in R} Y_{g,i} | S_a \right] = \frac{M}{m} \sum_{g \in S_a} \mathbf{E} \left[\frac{1}{f_s} \sum_{i \in R} Y_{g,i} | S_a \right] = \frac{M}{m} \sum_{g \in S_a} Y_g$$

d'où :

$$\mathbf{V}[\mathbf{E}(Y_{HT} | S_a)] = \mathbf{V} \left[\frac{M}{m} \sum_{g \in S_a} Y_g \right]_{\text{sous un SAS}} = M^2 \left(1 - \frac{m}{M} \right) \frac{S_g^2}{m}$$

où

$$S_g^2 = \frac{1}{M-1} \sum_{i=1}^M (Y_i - \bar{Y})^2$$

qu'on estime par

$$s_g^2 = \frac{1}{m-1} \sum_{g \in S_a} (\hat{Y}_g - \hat{\bar{Y}})^2 \quad \text{où} \quad \hat{Y}_g = \sum_{i=1}^{N_g} \frac{Y_{g,i}}{\hat{f}_s}$$

Notons que les estimateurs du premier et du second terme sont biaisés mais que leur somme est sans biais¹⁹ (si l'on néglige le biais dû à l'estimation de f_s).

3.4 Prise en compte du calage

Pour tenir compte des calages successifs, on calcule la variance sur les résidus de la régression de notre variable d'intérêt (en l'occurrence la linéarisée du taux de pauvreté) sur les variables de calage. Dans la réalité deux grandes phases de calages ont lieu qui sont indépendantes: d'une part le calage de l'*Enquête Emploi* sur la structure démographique, et d'autre part le calage après appariement avec les données DGI sur des masses (salaires, retraites, revenus des indépendants) fournies par la DGI ainsi que sur certaines variables de l'*Enquête Emploi* avant appariement²⁰.

¹⁹voir Tillé ([12])

²⁰Il s'agit des variables suivantes: tymen (type de ménage; 5 modalités), reg (la région; 8 modalités), pretud (dummy: personne de référence étudiante ou non), wact (activité au sens du BIT; 4 modalités), l'âge croisé avec le sexe (10 modalités au total comme dans le calage de l'*Enquête Emploi*) et enfin la cs au niveau individuel (7 modalités).

Bien que cela ne soit pas équivalent en principe, nous faisons ici comme si une seule étape de calage avait eu lieu et nous régressons simultanément la variable d'intérêt sur toutes les variables utilisées pour l'un ou l'autre des calages.

3.5 Résultats

3.5.1 Précision des estimateurs

Le tableau 2 présente les précisions obtenues sur les taux de pauvreté à 50% et 60% pour les années 1996 à 2000.

Tableau 2. Intervalles de confiance sur l'ensemble du « champ pauvres » de l'Enquête Revenus Fiscaux

Année	Taille échantillon	Taille échantillon	Taux de pauvreté : seuil à 50%				Taux de pauvreté : seuil à 60%			
	individus	ménages	demi IC95 (en points)	demi IC95 SAS (en points)	demi IC95 naïf (en points)	demi IC95 (en points)	demi IC95 SAS (en points)	demi IC95 naïf (en points)		
1996	55335	22001	7,2%	1,98%	0,42%	0,22%	13,5%	2,49%	0,52%	0,28%
1997	110311	44097	6,9%	0,62%	0,29%	0,15%	13,4%	0,81%	0,36%	0,20%
1998	166261	67046	6,7%	0,31%	0,23%	0,12%	12,8%	0,39%	0,30%	0,16%
1999	166321	67539	6,4%	0,29%	0,23%	0,12%	12,3%	0,38%	0,29%	0,16%
2000	163890	67340	6,5%	0,29%	0,22%	0,12%	12,7%	0,38%	0,29%	0,16%

Source: *Enquêtes Revenus Fiscaux* 1996-2000, INSEE-DGI; « champ pauvre »: *i.e.* individus vivant dans des ménages d'étudiants et ménages ayant déclaré des revenus négatifs exclus.

Lecture: le demi IC95 naïf est calculé en utilisant un estimateur de la variance qui ignore le plan de sondage et considère le

$$\text{taux de pauvreté comme une simple proportion : } \hat{V}(\hat{J}) = \frac{\hat{J}(1-\hat{J})}{n_{\text{ind}}}$$

La précision du taux de pauvreté est donc très faible en 1996. Il s'agit de la première année de l'enquête Revenus Fiscaux, et seules les aires entrantes de l'enquête Emploi ont été appariées avec les données fiscales. Le faible nombre d'aires entraîne alors un très fort « effet grappes ». En 1996, l'approximation du tirage de l'enquête par un SAS s'avère par conséquent très mauvaise : le design effect est proche de 5. En 1998, 1999 et 2000, l'« effet grappes » est plus réduit puisque le nombre d'aires est 3 fois plus important. L'approximation par un SAS conduit cependant encore à une surestimation de 30% de la précision. Enfin, les estimateurs naïfs de la précision sont particulièrement biaisés. Leur problème principal est de ne pas prendre en compte le fait qu'il s'agit d'un tirage de ménages et non d'individus.

3.5.2 Analyse de l'évolution du taux de pauvreté

Pour tester la significativité de la baisse du taux, on effectue le test suivant :

$$H_0 : J_{1996} = J_{1997} = \dots = J_{2000} \quad \text{contre} \quad H_1 : \exists(k, l) / J_k \neq J_l$$

J est observé avec une erreur :

$$\hat{J}_{a,b}(t) = J_{a,b}(t) + \mathbf{e}(t)$$

Du postulat de normalité asymptotique de l'estimateur, on tire :

$$\mathbf{e}(t) \sim N(0, \mathbf{s}_t^2)$$

et on suppose que les $\mathbf{e}(t)$ sont indépendants²¹.

²¹ On pourrait introduire des erreurs dépendantes pour prendre en compte la spécificité du plan de sondage rotatif.

Sous l'hypothèse nulle H_0 :

$$\Delta = \begin{pmatrix} \hat{J}_{1996} - \hat{J}_{1997} \\ \hat{J}_{1997} - \hat{J}_{1998} \\ \hat{J}_{1998} - \hat{J}_{1999} \\ \hat{J}_{1999} - \hat{J}_{2000} \end{pmatrix} \sim N(0_4, \Sigma)$$

avec :

$$\Sigma = \begin{pmatrix} \frac{s_{1996}^2}{n_{1996}} + \frac{s_{1997}^2}{n_{1997}} & -\frac{s_{1997}^2}{n_{1997}} & 0 & 0 \\ -\frac{s_{1997}^2}{n_{1997}} & \frac{s_{1997}^2}{n_{1997}} + \frac{s_{1998}^2}{n_{1998}} & -\frac{s_{1998}^2}{n_{1998}} & 0 \\ 0 & -\frac{s_{1998}^2}{n_{1998}} & \frac{s_{1998}^2}{n_{1998}} + \frac{s_{1999}^2}{n_{1999}} & -\frac{s_{1999}^2}{n_{1999}} \\ 0 & 0 & -\frac{s_{1999}^2}{n_{1999}} & \frac{s_{1999}^2}{n_{1999}} + \frac{s_{2000}^2}{n_{2000}} \end{pmatrix}$$

et par conséquent $\Delta' \Sigma^{-1} \Delta$ suit, sous H_0 , une loi du χ^2 à 4 degrés de liberté qui nous permet de tester notre hypothèse de constance du taux de pauvreté sur la période.

Tableau 3 : significativité de l'évolution du taux de pauvreté

Taux de pauvreté	Estimation	Seuil à 5% ($\chi^2(4)$)
50% de la médiane	3,8	9,49
60% de la médiane	7,9	9,49

L'hypothèse de constance du taux de pauvreté est donc acceptée sur la période pour les deux seuils considérés. La diminution du taux de pauvreté n'est donc pas significative (au seuil de 5%).

Bibliographie

- [1] Berger, Y.G. (1999), Approximation de la variance de l'estimateur de Horvitz-Thomson, *Enquêtes et sondages*, Dunod, 223-229.
- [2] Cabeça, J.C.S. (1999), Une procédure de réplcation d'échantillons appliquée aux sondages, *Enquêtes et sondages*, Dunod, 252-260.
- [3] Caron, N., Deville, J.C., Sautory, O. (1998), Estimation de précision de données issues d'enquêtes : document méthodologique sur le logiciel POULPE. *Document de travail INSEE, série Méthodologie Statistique*, n°9806.
- [4] Deville, J.C., Särndal, C.E. (1992), Calibration estimators in survey sampling, *Journal of the American Statistical Association*, 87, 11, 376-382.
- [5] Deville, J.C. (1996), Estimation de la variance du coefficient de Gini estimé par sondage, *Actes des journées de méthodologie statistique*, INSEE méthodes n° 69-70-71, 269-288.
- [6] Deville, J.C. (1998), Estimation de variance pour des statistiques complexes: technique des résidus et de linéarisation, *Document de travail INSEE, série Méthodologie Statistique*, n°9802.
- [7] Deville, J.C. (1998), Variance et estimation de variance en cas d'erreurs de mesure non corrélées ou de l'intrusion d'un individu Kish *Document de travail INSEE, série Méthodologie Statistique*, n°9805.
- [8] Deville, J.C. (1999), Estimation de variance pour des statistiques et des estimateurs complexes : techniques de résidus et de linéarisation, Variance estimation for complex statistics and estimators : linearization and residual techniques, *Techniques d'enquête / Survey methodology* 25:219-230 (fr.), 193-204 (angl.).
- [9] Fleurbaey, M., Lollivier, S. (1994), Les mesures des inégalités : abrégé théorique et pratique , *Document de travail CREST n° 9408 bis*.
- Insee Résultats, Enquête Emploi* (1999).
- [10] Roth, N. (1991), L'enquête emploi: échantillon 1992 et années suivantes, *Insee Méthodes* n°29-30-31.
- [11] Tillé, Y. (2001), *Théorie des sondages*, Dunod.

ANNEXE : Syntaxe des macros

On explicite ici la syntaxe des macros SAS permettant de calculer plusieurs statistiques complexes et leurs linéarisées²². Cela concerne l'indice de GINI, les indices d'ATKINSON, le THEIL, les quantiles et rapports interquantiles et le taux de pauvreté. Notons que les macros, sous leur forme actuelle, ne fonctionnent pas sous SAS 6 : certains noms de tables et de variables comprennent plus de 8 caractères.

Deux types de macros ont été écrites : celles calculant la statistique proprement dite, comme la macro **%indice_Gini** qui calcule l'indice de GINI, et les macros qui calculent les linéarisées correspondant à ces statistiques, comme la macro **%linearisation_Gini** qui calcule les linéarisées relatives à cet indice de GINI. Enfin, une macro calculant les résidus d'une régression a été implementée afin de prendre en compte le calage des enquêtes.

Il est à noter que le calcul des linéarisées fait pratiquement toujours intervenir les statistiques elle-même. Il est donc nécessaire de lancer la macro calculant cette statistique avant de lancer la macro de calcul des linéarisées²³. Pour que les calculs soient valides, les paramètres des deux macros doivent être compatibles.

Exemple : On s'intéresse à la précision du taux de pauvreté à 60% estimé à partir de l'*Enquête Revenus Fiscaux* qui a été calée. la table SAS s'appelle ERF, la variable de revenus Y, les pondérations W et les variables de calage X1 à X10. La syntaxe est alors la suivante :

```
%taux_Pauvrete(variable=Y,ponderation=W, table=ERF, beta=0.6);  
%linearisation_Pauvrete(variable=Y, ponderation=W, table=ERF, beta=0.6, linearise=LIN);  
%residus(variable=LIN, ponderation=W, liste_quant=X1-X10, table=ERF, residus=RES);
```

La première macro calcule le taux de pauvreté à 60 %, et l'affiche dans l'OUTPUT. Le taux est également enregistré dans la macro-variable pauvrete.

Cette macro-variable est réutilisée (par défaut) dans %linearisation_Pauvrete. Notons qu'il est nécessaire de rappeler dans cette deuxième macro la valeur de beta (0.6) car il ne s'agit pas de la valeur par défaut.

A l'issue de cette seconde macro, une variable LIN est créée dans la table ERF. Pour tenir compte du calage, cette variable est régressée sur X1-X10 (avec les poids W) dans %residus. Cette troisième macro ajoute la variable RES dans la table ERF.

A l'issue de ces trois étapes, il ne reste plus qu'à calculer la variance du total sur RES pour obtenir une estimation de la précision du taux de pauvreté.

²²Les programmes complets sont présentés dans l'annexe suivante.

²³cette solution n'est pas strictement obligatoire : on peut également indiquer dans la macro de linéarisation la valeur de la statistique

A.1 Syntaxe des macros de calcul des estimateurs

Les estimateurs calculés ci-dessous reprennent les formules présentées en 2.2. Les instructions identiques d'une macro à l'autre ne sont pas réexpliquées.

A.1.1 Indice de GINI

On utilise la macro %indice_Gini qui s'appelle par l'instruction générale suivante :

%indice_Gini(variable=,ponderation=, librairie=, table=, sortie=, Gini_Prov=);

variable : nom de la variable sur laquelle on souhaite calculer le Gini.

ponderation : nom des poids de sondage.

librairie : nom de la librairie dans laquelle se situe la table. Par défaut, la librairie choisie est la work.

table : nom de la table exploitée.

sortie : indique si l'on souhaite obtenir l'indice de GINI dans l'OUTPUT de SAS. Deux possibilités :

- 0 : pas de sortie de l'indice de GINI dans l'OUTPUT

- tout autre valeur : sortie

Par défaut, la valeur de sortie est 1

Gini_Prov : nom de la macro-variable dans laquelle est stockée la valeur de l'indice de GINI. Par défaut, le nom de cette macro-variable est GINI.

Ainsi, par défaut, cette macro sort en OUTPUT et enregistre dans la macro variable GINI la valeur de l'estimateur du GINI (en d'autres termes, &Gini renvoie la valeur de l'indice).

A.1.2 Indice d'ATKINSON

Syntaxe générale :

%indice_ATKINSON(variable=,ponderation=, librairie=, table=, a=, sortie=, ATKINSON_Prov=);

a : coefficient de l'indice d'ATKINSON A(a).

ATKINSON_Prov : par défaut, le nom de la macro-variable est Atkinson.

A.1.3 Indice de THEIL

%indice_Theil(variable=,ponderation=, librairie=, table=, sortie=, Theil_Prov=);

Theil_Prov : par défaut, le nom de la macro-variable est Theil.

A.1.4 Quantile

%Quantile(variable=,ponderation=, librairie=, table=, q=, sortie=, quantile_Prov=);

q : valeur du quantile (centile en fait : par exemple 25 donne le premier quartile)

quantile_Prov : par défaut, le nom de la macro-variable est quantile.

A.1.5 Rapport Interquantile

Cette macro calcule les rapports interquantile de la forme $\frac{q_{100-x}}{q_x}$ ($x \in]0;50[$). La syntaxe générale est la suivante :

%Interquantile(variable=,ponderation=, librairie=, table=, q=, sortie=, interquantile_Prov=);

q : le x de la définition précédente. Par exemple $q = 10$ correspond au rapport interdécile. Il s'agit de

la valeur par défaut.

interquantile_Prov : par défaut, le nom de la macro-variable est interquantile.

A.1.6 Taux de pauvreté

`%taux_Pauvrete(variable=,ponderation=, librairie=, table=, alpha=, beta=, seuil_connu=, seuil=, sous_champ=, indicatrice_champ=,sortie=, pauvrete_Prov=);`

alpha : il s'agit du **a** du taux de pauvreté tel qu'il est défini en 2.2. Par défaut, il vaut 0.5 (indicateur correspondant à la moitié de la médiane ; pour avoir 60% de la médiane, indiquer alpha=0.6).

beta : **b** du taux de pauvreté défini en 2.2. Par défaut, $beta = 0.5$ (calcul à partir de la médiane).

seuil_connu : indicatrice indiquant si le seuil est connu ou estimé à partir de l'enquête. 0 par défaut (seuil estimé par l'enquête).

seuil : valeur du seuil (à renseigner uniquement si `seuil_connu = 1`).

sous_champ : indicatrice indiquant si l'on souhaite calculer le taux de pauvreté sur un sous-champ (par exemple les personnes de 65 ans ou plus). 0 par défaut (taux calculé sur la population entière).

indicatrice_champ : nom de la variable indicatrice correspondant au sous-champ considéré (à renseigner uniquement si `sous_champ = 1`).

pauvrete_Prov : par défaut, le nom de la macro-variable est pauvrete.

A.2 Syntaxe des macros de calcul des linéarisées

Les macros de linéarisation ajoutent dans la table d'origine une colonne comprenant les linéarisées.
subsubsection30mm-0.80.4Linéarisation de l'Indice de GINI

`%linearisation_Gini(variable=, ponderation=, librairie=, table=, Gini_Prov=, linearise=);`

Gini_Prov : valeur de l'indice de GINI. Par défaut, valeur de la macro variable GINI.

linearise : nom de la variable comprenant les linéarisées. Par défaut, la variable s'appelle `Linearise_Gini`.

A.2.1 Linéarisation de l'Indice d'Atkinson

`%linearisation_Atkinson(variable=, ponderation=, a=, librairie=, table=, Atkinson_Prov=, linearise=);`

a : même signification que dans `%indice_Atkinson`.

Atkinson_Prov : par défaut, valeur de la macro variable Atkinson.

linearise : par défaut, `Linearise_Atkinson`.

A.2.2 Linéarisation de l'Indice de Theil

`%linearisation_Theil(variable=, ponderation=, librairie=, table=, Theil_Prov=, linearise=);`

Theil_Prov : par défaut, valeur de la macro variable Theil.

linearise : par défaut, `Linearise_Theil`.

A.2.3 Linéarisation du Quantile

`%linearisation_Quantile(variable=, ponderation=, librairie=, table=, q=, quantile_Prov=, h=, sortie_h=, linearise=);`

q : même signification que dans `%Quantile`.

quantile_Prov : par défaut, valeur de la macro variable quantile.

h : valeur de la fenêtre du noyau utilisé pour calculer $F'(M, y)$ (cf. 2.2). Par défaut, $h = \frac{s}{N^{1/5}}$ (en d'autres termes, h est calculé selon la règle du pouce.

sortie_h : indique, lorsque h a été laissé à sa valeur par défaut, si l'on souhaite obtenir la valeur de h dans l'OUTPUT de SAS.

0 : pas de sortie de h dans l'OUTPUT

tout autre valeur : sortie de h.

Par défaut, sortie_h vaut 0.

linearise : par défaut, Linearise_quantile.

subsubsection30mm-0.80.4Linéarisation du Rapport Interquantile

%linearisation_Quantile(variable=, ponderation=, librairie=, table=, q=, h=, linearise=);

q : même signification que dans %Interquantile.

linearise : par défaut, Linearise_interquantile.

Notons que contrairement aux autres, cette macro ne nécessite pas la valeur du rapport interquantile.

Par ailleurs, elle ne dispose pas d'option de sortie pour la valeur de h.

2.4 Linéarisation du Taux de pauvreté

%linearisation_Pauvrete(variable=, ponderation=, librairie=, table=, alpha=, beta=, seuil_connu=, seuil=, sous_champ=, indicatrice_champ=, pauvrete_Prov=, h=, sortie_h=, linearise=);

alpha, beta, seuil_connu, seuil, sous_champ, indicatrice_champ : même signification que dans %taux_Pauvrete.

pauvrete_Prov : par défaut, valeur de la macro variable pauvrete.

linearise : par défaut, Linearise_pauvrete.

A.3 Prise en compte du calage

Lorsqu'il y a calage, la variance d'un total est égale à la variance du total des résidus de la régression linéaire sur les variables de calage (cf. 1.3). Ce résultat vaut également pour les statistiques complexes : on régresse alors la linéarisée sur les variables de calage. La macro suivante ajoute dans la table d'origine les résidus de la régression :

%residus(variable=, ponderation=, liste_quali=, liste_quanti=, librairie=, table=, residus=);

variable : nom de la variable qu'on souhaite régresser.

ponderation : nom de la variable de pondération à utiliser dans la régression. Les poids de sondages sont conseillés mais on peut ne pas renseigner ce paramètre, auquel cas la régression sera non pondérée.

liste_quali : liste des variables de calage qualitatives. Les variables doivent être séparées par des espaces.

liste_quanti : liste des variables de calage quantitatives. Les variables doivent être séparées par des espaces.

librairie : nom de la librairie dans laquelle se situe la table. Par défaut, la librairie choisie est la work.

table : nom de la table.

residus : nom de la variable correspondant aux résidus. Par défaut, la variable s'appelle residus_cal.