ESTIMATION DE LA VARIANCE DU COEFFICIENT DE GINI MESURÉ PAR SONDAGE

Jean-Claude Deville

Le coefficient de GINI mesure la dispersion d'une variable numérique positive dans une population. Il est fréquemment utilisé dans les Instituts de Statistique notamment pour évaluer les disparités et les inégalités de revenus. C'est, par ailleurs, un exemple caractéristique de fonctionnelle "fortement" non linéaire, qu'on ne peut en aucun cas assimiler à une fonction de totaux. Il paraît donc a priori difficile d'évaluer la précision d'un tel indice quand il est calculé à partir des résultats d'une enquête par sondage car la théorie habituelle (WOODRUFF[10]) ne s'applique pas. Cette note s'attaque à un problème qui présente un réel intérêt tant pratique que théorique. On y propose une solution qui résout le problème par des techniques appropriées de linéarisation et qui le ramène à un calcul classique d'estimation de variance d'un total pour l'estimateur de HORVITZ-THOMPSON. La variance de l'indice de GINI peut donc être évaluée à l'aide d'un logiciel de calcul de précision de type standard comme celui auquel travaille l'équipe de l'Unité Méthodes Statistiques.La solution qui est proposée est assez originale bien qu'on puisse en trouver la source dans [1] : on y montre que l'indice de GINI peut-être vu comme un ratio d'estimateurs et qu'une méthode d'estimation de variance peut-être dérivée de cette idée. Toutefois, dans [2], on constate que la technique du ratio donne dans des simulations des résultats extraordinairement mauvais, la variance étant surestimée de 30 à 300 fois. On montre ici qu'il y avait un étrange artefact dans la façon dont on attaquait le problème dans ces articles et qu'une analyse plus fine de la notion de linéarisation permet d'arriver à un résultat correct.On commencera, en prime, par réviser les formulations possibles du coefficient de GINI. On passera ensuite à la question de son estimation dans les sondages complexes, puis à celui de l'estimation de la variance. Une simulation montrera ensuite que l'expérience est en accord très honorable avec les résultats théoriques.

1 - Coefficient de GINI

Comme beaucoup, sans doute, une des seules choses que je savais du coefficient de GINI est ce qu'on trouve dans le livre de G. CALOT [3]. On définit, pour une variable réelle positive Z les objets suivants :

- Sa fonction de répartition F(t) = Fréquence de $Z \leq t$

- Sa moyenne $M = \int t\, dF(t)$

- la fonction de répartition

$$G(t) = \frac{1}{M}\int^{t} z dF(z)$$

- La courbe de concentration de LORENZ définie comme l'ensemble des points :

$$\begin{cases} x(t) = G(t) \\ y(t) = F(t) \end{cases}$$

S'il y a des discontinuités, la courbe de LORENZ est l'enveloppe convexe de ces points auxquels on ajoute (0,0). Cette courbe est convexe, contenue dans le carré $[0,1]\times[0,1]$ passe par $[0,0]$ *et* $[1,1]$. Par suite elle se situe au dessus de la diagonale du carré.

L'indice de GINI est alors le double de la surface comprise entre la courbe et cette diagonale. Il est proche de 0 si Z est presque constante. Il est proche de 1 si F(0) tend vers 1. Cet indice est souvent utilisé pour décrire les inégalités de salaires. Une valeur proche de 1 décrit donc une situation très inégalitaire où une petite fraction de la population dispose d'une grosse partie de l'ensemble des revenus.

La littérature nous fournit d'autres expressions plus commodes sur le plan analytique et permettant de calculer l'indice autrement que graphiquement en comptant le nombre de carreaux sous la courbe. Par ailleurs, le pont entre les diverses expressions de l'indice est rarement indiqué dans les ouvrages (en particulier aucune des références citées ici ne le fait !) bien que (ou parce que ?) ce soit assez peu évident.

Commençons par cela .

La formule d'intégration par partie pour les intégrales de Stieltjes nous sera utile. Rappelons celle-ci (qui est immédiate à établir). Soient F et G sont deux fonctions croissantes définies sur IR^+ (fonctions de répartition de mesures positives). On les supposera continues à droite. Elles admettent en tout point une limite à gauche qu'on notera F(*t-0*). La quantité $\Delta F(t) = F(t) - F(t-0)$ s'appelle le saut de F en *t*. On posera F(0-0) = 0. On a alors la formule d'intégration par parties (pour a $\leq$ b)

$$F(b)\,G(b) - F(a-0)\,G(a-0) = \int_a^b F(t)\,dG(t) + \int_a^b G(t)\,dF(t) - \sum_{t\in[a,b]} \Delta F(t)\,\Delta G(t).$$

Le troisième terme concerne donc les points où F et G ont des sauts communs et indique une famille sommable [9] de termes positifs.

La définition graphique de l'indice de GINI nous conduit à l'expression analytique :

$$GINI = 2 \int F(t)dG(t) - \sum \Delta F(t)\,\Delta G(t) - 1$$

$$= \frac{2}{M} \int t\ F(t)dF(t) - \sum_t t\ \Delta F(t)^2 - 1 \qquad (2)$$

Mais l'intégration par partie nous donne :

$$\int FdG = \left[FG\right]_{0-0}^{+\infty} - \int GdF + \sum_t \Delta G(t)\ \Delta\ F(t)$$

d'où :

$$GINI = \int\ F(t)dG(t) - \int\ G(t)dF(t)$$

$$= \frac{1}{M}\left[\int\ tF(t)dF(t) - \int\ dF(t)\int^t udF(u)\right]$$

$$= \frac{1}{M}\left[\int\ dF(t)\int^t\ (t-u)\ dF(u)\right]$$

et donc, par symétrie :

$$GINI = \frac{1}{2M} \iint\ |t-u|\ dF(t)\ dF(u) \qquad (3)$$

Remarque *: La formule (2) fait apparaître, dans le cas continu en tous cas, la moyenne du supremum de deux variables indépendants suivant la loi F. J'ignore ce que cette remarque apporte de plus qu'un élément de cocasserie.*

2 - Cas d'une population finie

Traduisons les résultats précédents dans le cas d'une population finie d'effectif N. La variable d'intérêt z prend des valeurs z_k *(pour* $k = 1$ *à* N*)* et on supposera que $z_k \leq z_{k+1}$.

La traduction de la formule (3) est alors :

$$GINI = \frac{1}{2} \frac{\sum_{k=1}^{N} \sum_{\ell=1}^{N} |Z_k - Z_\ell|}{N \sum_{i=1}^{N} Z_i}$$

On constate que GINI = 0 quant tous les z_k sont égaux. GINI est maximum quand $z_k=0$ pour k = 1 à N-1 et $z_N>0$. On trouve alors la valeur de $1-\frac{1}{N}$.

Passons maintenant à la formule (2).
Si nous supposons $z_k < z_{k+1}$ pour tout k on aura alors F(t) = k/N pour t dans l'intervalle $[z_k, z_{k+1}[$ (avec la convention $z_0 = 0$ et $z_{N+1} = +\infty$). La formule de l'indice de GINI devient alors :

$$GINI = \frac{2 \sum_{k=1}^{N} k z_k}{N \sum_{k=1}^{N} z_k} - \frac{\sum_{k=1}^{N} z_k}{N^2 \quad M} - 1 = \frac{2 \sum_{k=1}^{N} k z_k}{N \sum_{k=1}^{N} z_k} - \frac{1}{N} - 1$$

$$= \frac{\sum_{k=1}^{N} (2k-1) z_k}{N \sum_{k=1}^{N} z_k} - 1 \qquad (2')$$

Cette forme est particulièrement utile car elle fait apparaître l'indice comme le ratio de deux statistiques assez simple de la population finie : le total au dénominateur et le total de $y_k = \left(\frac{2k-1}{N} - 1\right) z_k$ au numérateur. Cette dernière pose toutefois un problème, car, généralement, le rang k des unités observées est inconnu.

On remarquera, par ailleurs, que la forme (2') reste vraie si plusieurs unités k prennent la même valeur y_k (le classement des ex-aequo étant alors arbitraire). On s'en convaincra soit en remarquant la stabilité par passage à la limite de (2'), soit par un raisonnement direct. Au cas où la valeur z comporte q ex-aequo, l'élément de la formule (2) correspondant à cette valeur s'écrit (à un facteur près) :

$$2z \frac{i}{N} \cdot \frac{q}{N} - z\left(\frac{q}{N}\right)^2 = \frac{z}{N^2}\left(2iq - q^2\right)$$

Or :

$$2iq - q^2 = i^2 - (i-q)^2 = \sum_{r=0}^{q-1} (i-r)^2 - (i-r-1)^2$$

$$= \sum_{r=o}^{q-1} (2(i-r)-1)$$

Ce qui permet de terminer la preuve.

La forme (1) - ou graphique - n'apporte pas véritablement d'information supplémentaire. Le cas fini permet un calcul exact de l'indice défini graphiquement. On doit évaluer l'aire d'une réunion de N trapèzes. Le k^{ème} d'entre eux a pour base z_k/NM et pour demi-hauteur (k-1/2)/N pour k = 1 à N. La surface sous la courbe de LORENZ vaut donc :

$$\sum_{k=1}^{N} \frac{k-\frac{1}{2}}{N} \cdot \frac{z_k}{NM}$$

D'où la formule (2'). Si plusieurs unités k prennent la même valeur z_k il est facile de voir que le calcul précédent reste valide.

3 - Estimation à partir d'une données d'enquêtes par sondage dans le cas où le rang est observable

Commençons par le cas où le rang de classement i est observable. Bien que cela ne soit pas chose courante, voici un exemple qu'on peut imaginer. On veut calculer la dispersion des temps mis par les coureurs du Tour de France à effectuer les étapes successives (vite dit ou veut connaître "scientifiquement" les plus sélectives !). Pour cela on tire avant le départ un échantillon s de coureurs k dont on relèvera, à la fin de chaque étape, le classement i_k et le temps réalisé z_k (pour plus de réalisme ce temps pourrait être l'écart entre le temps d'arrivée du premier et le délai au delà duquel les coureurs sont éliminés).

En notant w_k les poids attribué dans l'enquête, l'indice de GINI sera estimé par un ratio :

$$\hat{GINI} = \frac{\sum_s (2i_k - 1) z_k w_k}{N \sum_s z_k w_k} - 1 \qquad (3-1)$$

Éventuellement, si N n'est pas connu ou si la somme des poids ne vaut pas N on remplace N par $\hat{N} = \sum_s w_k$. La variance de GÎNI s'estime alors comme celle d'un ratio - si N est connu - ou par une formule standard de linéarisation si N est remplacé par son estimateur.

Autrement dit, si on sait calculer (par exemple par un programme standardisé) la variance estimée d'un total de la forme $\hat{Y} = \sum_k w_k y_k$, on remplace y_k dans le calcul par la variable artificielle:

$$U_k = \frac{1}{\hat{Z}}\left(\frac{2i_k - 1}{N} - \left(G\hat{I}NI + 1\right)\right) z_k$$

dans le cas ou N est connu. Dans le cas ou N est estimé on prendra :

$$u_k' = \frac{1}{\hat{N}\,\hat{Z}}\left[\left(2i_k - 1 - \left(G\hat{I}NI + 1\right)\hat{N}\right) z_k - A\right] \quad ,$$

$$avec \qquad A = \sum_s \left(2i_k - 1\right) z_k w_k \Big/ \sum_s w_k$$

Le calcul se ramène donc à une recodification élémentaire.

4 - Estimation dans le cas où le rang n'est pas observable

Malheureusement on ne dispose pas, en général du rang i_k de l'unité k. Celui-ci doit être estimé à partir des données, en se basant sur une estimation $\hat{F}(z)$ de la fonction de répartition de Z.

Cette estimation de $\hat{F}$ peut se faire de diverses façons. Le plus simple et le plus naturel consiste à utiliser les estimateurs de HORVITZ-THOMPSON des proportions de z_k inférieurs ou égaux à z et cela pour tout z. On prend classiquement donc :

$$\hat{F}(z) = \sum_{k:z_k \le z} 1/\pi_k \Big/ \sum_{k \in s} 1/\pi_k$$

Ceci posé, les poids de H.T peuvent être remplacés par n'importe quel autre système de poids associé à des performances connues des estimateurs linéaires qu'ils définissent. La fonction $\hat{F}(z)$ est constante par intervalles, continue à droite et à des sauts pour toutes les valeurs z_k prises par l'échantillon. Pour d'autres principes d'estimation (voir par exemple CHAMBERS et DUNSTAN [5]) cette particularité se maintient. En fait, c'est le rang de z_k que nous voulons estimer, ce qui est un peu malheureux car nous avons justement un saut de $\hat{F}$ en z_k et nous pouvons hésiter entre $\hat{F}(z_k)$ *et* $\hat{F}(z_k - 0)$ comme estimateur de i_k/N. Il est raisonnable d'utiliser la moyenne de ces deux valeurs :

$$\tilde{F}(z_k) = \frac{1}{2}\left(\hat{F}(z_k) + \hat{F}(z_k - 0)\right) = \hat{F}(z_k) - \frac{1}{2}\frac{w_k}{\hat{N}}$$

Ce choix correspond à une fonction de répartition $\tilde{F}$ obtenue par un léger lissage. Graphiquement il correspond à utiliser l'estimateur $\tilde{F}$ dont la courbe représentative est obtenue en joignant les milieux les "contremarches" adjacentes de la courbe représentative de $\hat{F}$. L'inconvénient de cet estimateur est de ne pas être défini (sauf convention particulière) à l'extérieur de l'intervalle $\left[min(z_k),\ max(z_k)\right]$

La quantité $2i_k - 1$ de (3-1) se trouve donc estimée naturellement par $2\hat{N}\tilde{F}_k - 1$ ce qui conduit à la forme suivante de l'estimateur du GINI:

$$\hat{GINI} = \frac{\sum_s \left(2\hat{N}\tilde{F}_k - 1\right) z_k\, w_k}{\hat{N}\sum_s w_k\, z_k}$$

$$= \frac{\sum_s 2\hat{N}\tilde{F}_k\, z_k\, w_k}{\hat{N}\sum_s w_k\, z_k} - \left(1 + \frac{1}{\hat{N}}\right) \qquad (4-1)$$

Dans le cas d'un sondage aléatoire simple (ou plus généralement, d'un sondage de taille fixe à probabilité égales - où N est connu), cette expression prend la forme :

$$\hat{GINI} = \frac{\sum_s (2j_k - 1) z_k}{n\sum_s z_k} - \left(1 + \frac{1}{N}\right) \qquad (4-2)$$

où j_k est le rang (de 1 à n) de l'unité k dans s.

5 - Considérations globales de biais et de variance

Les estimateurs (4-1) ou même (4-2) ont une forme de ratio et n'ont aucune raison d'être sans biais. Dans les applications concrètes, par ailleurs, il est à peu près évident que la variabilité du terme en $\frac{1}{\hat{N}}$ de (4-1) peut-être négligée (le terme lui-même peut l'être !). Le dénominateur du ratio est une quantité simple. Seul le numérateur demande un examen préliminaire un peu précis. Mettons (4-1) sous la forme :

$$\hat{GINI} = \frac{\sum_s 2\,\tilde{F}_k z_k\, p_k}{\sum_s z_k\, p_k} - 1 \qquad où \qquad p_k = w_k / \hat{N}$$

Le remplacement de $\tilde{F}$ *par* $\hat{F}_k$ produit une variation de l'indice égale à :

$$\sum_s z_k\, p_k^2 \Big/ \sum_s z_k\, p_k$$

qui est de l'ordre de *1/n* car p_k est lui même de l'ordre de *1/n*.

Le biais généré par ce remplacement sera donc en général négligeable. Son incidence sur l'erreur quadratique moyenne se traduira par un terme en $1/n^2$ qu'on ne cherchera pas à capturer.

Par ailleurs, le numérateur lui-même n'estime aucune quantité simple sans biais ! Sans reprendre toute l'analyse faite dans [1], examinons une seconde ce qui se passe dans le cas du sondage aléatoire simple où ce numérateur vaut :

$$NUM = \frac{1}{n^2} \sum_s \left(2j_k - 1\right) z_k = \frac{1}{n^2} \sum_U \varepsilon_k z_k \left(2 \sum_{\ell=1}^{k} \varepsilon_\ell - 1 \right)$$

où ε_k est la variable d'appartenance à l'échantillon $\left(\varepsilon_k = 1 \text{ si } k \in s \text{ et } 0 \text{ sinon}\right)$.

On voit alors que :

$$E(NUM) = \frac{1}{n^2} \sum_{k=1}^{N} z_k \left(2k . \frac{n(n-1)}{N(N-1)} - \frac{n}{N} \right)$$

$$= \frac{1}{N-1} \sum_{k=1}^{N} \frac{2k}{N} z_k - \frac{1}{n} \left(\frac{1}{N} \left(\sum_1^N \frac{2k}{N-1} + 1 \right) z_k \right)$$

On découvre donc de nouveau un biais de l'ordre de 1/n. Pour des plans et des estimateurs plus généraux on pourrait montrer qu'on aurait aussi un biais de cet ordre de grandeur. Ce second biais a le même genre d'incidence sur l'écart quadratique moyen que celui qui a déjà été évoqué.

Cette analyse peut-être résumée par la proposition suivante :

Résultat : La quantité

$$\hat{GINI}^* = \frac{\sum_s 2\, z_k\, \hat{F}_k\, p_k}{\sum_s z_k\, p_k} - 1 \qquad (5-3)$$

est un estimateur de l'indice de GINI qui a, à des termes en $\frac{1}{n}$ près, le même biais que l'estimateur (4-1). De plus, il a le même écart quadratique moyen à des termes en $1/n^2$ près.

C'est, donc sur cette quantité que nous allons travailler maintenant.

6 - Ébauche de linéarisation pour GINI*

Posons $\hat{R} = 2\left(1 + \hat{GINI}^*\right) = NUM / \hat{M}$.

En vertu des règles générales de linéarisation et de la confusion habituelle (et justifiée par l'analyse du biais) entre variance et écart quadratique moyen nous pouvons écrire :

$$VAR\left(\hat{R}\right) \simeq \frac{1}{M^2}\left(Var\,(NUM) - 2R\,Cov(NUM, \hat{M}) + R^2\, Var\left(\hat{M}\right)\right)$$

Le traitement de $\hat{M} = \frac{1}{\hat{N}} \sum_s z_k\, w_k$ ne pose pas de problème particulier dès que l'on sait estimer la variance de l'estimateur du total de Z, $\sum_s z_k\, w_k$. L'estimateur de variance utilise la variable linéarisée $\left(z_k - \hat{M}\right)/\hat{N} = z_k'$ comme le montre un calcul élémentaire.

En revanche, comme nous allons le voir, le traitement de NUM demande plus de finesse. En effet, on a :

$$NUM = \sum_s z_k \hat{F}_k p_k$$

On pourrait donc raisonner sommairement en disant que $z_k \hat{F}_k$ estime $z_k F_k$ et qu'on peut traiter le problème comme au paragraphe 3. C'est ce qui est fait dans [2], et conduit à la conclusion étonnée que la variance estimée est de 10 à 300 fois trop forte.

En fait, c'est négliger la covariance entre $\hat{F}_k$ et le poids p_k (aléatoire !!) attribué à l'unité k. Comme, de plus, ces covariances sont sommées sur toute la population on obtient au final un biais d'ordre fini (même pour de très gros échantillons) c'est-à-dire, pour parler clair, un calcul parfaitement faux.

Nous allons, dans la suite, montrer que la variance de NUM est, à des termes négligeables près), celle du total d'une variable artificielle x_k :

$$VAR\,(NUM) \simeq Var\left(\sum_s x_k\, w_k\right).$$

Par suite on aura :

$$Cov\left(NUM, \hat{M}\right) \simeq Cov\left(\sum_s x_k w_k, \sum_s z_k' w_k\right)$$

Il en résultera, d'après (6-1), que :

$$Var\ \hat{R} \simeq \frac{1}{M^2} Var\left(\sum_s \left(x_k - Rz_k'\right) w_k\right)$$

$$= Var\left(\sum_s r_k w_k\right) avec\ r_k = \frac{1}{M}\left(x_k - Rz_k'\right)$$

Ceci marquera la fin de (presque) tous nos problèmes avec la variance de l'indice de GINI.

7 - Linéarisation de la fonctionnelle NUM

Notons qu'on peut écrire :

$$NUM = \sum_s z_k \hat{F}_k p_k = \int_{IR} z\ \hat{F}(z)\ d\hat{F}(z) \qquad (7)$$

avec $\hat{F}$ fonction de répartition estimée de la variable Z. Cette quantité est prise pour estimateur de $\int_{IR} z\ F(z)\ dF(z)$, de sorte qu'on est amené à examiner la quantité :

$$D = \int z\ \hat{F}(z)\ d\hat{F}(z) - \int z\ F(z)\ dF(z)$$

$$= \int z\ F(z)\ d\left[\hat{F} - F\right](z) + \int z \left(\hat{F}(z) - F(z)\right) dF(z) + \int z \left(\hat{F}(z) - F(z)\right) d\left[\hat{F} - F\right]$$

Traitons le dernier terme ; $Max_z\left(\hat{F}(z) - F(z)\right)$ est une variable aléatoire qui est en probabilité de l'ordre de $1/\sqrt{n}$. Pour de gros échantillons le dernier terme sera donc négligeable vis-à-vis du premier.

Examinons maintenant le second terme qui peut s'écrire.

$$\int \left(\hat{F}(z) - F(z)\right) dG(z) \quad avec \quad G(z) = \int^z u\, dF(u).$$

On peut maintenant utiliser l'intégration par partie :

$$\int \left(\hat{F} - F\right) dG = \left[G\left(\hat{F} - F\right)\right]_0^{\infty} - \int G\, d\left[\hat{F} - F\right] + \sum \Delta G.\Delta\left(\hat{F} - F\right)$$

Regardons le troisième terme :

$$\sum \Delta G\ \Delta\hat{F} = \sum_s \frac{z_k}{N}.p_k = \frac{1}{N}\hat{M}$$

$$\sum \Delta G\ \Delta F = \sum_U \frac{z_k}{N}.\frac{1}{N} = \frac{1}{N}M$$

Le troisième terme vaut donc $\frac{1}{N}\left(\hat{M}-M\right)$ et peut donc être considéré comme négligeable.

Le premier est nul car $\hat{F}(o-o)=F(o-o)=0$ *et* $\hat{F}(\infty)=F(\infty)=1$.

On obtient donc que :

$$D \simeq \int\left(z\ F(z)-G(z)\right)\ d\left(\hat{F}(z)-F(z)\right)$$

à des quantités négligeables près.

Il en résulte que la variance de D (ainsi que les covariances de D avec d'autres variables) est donnée au premier ordre, par la variance de la variable :

$$x_k = z_k\ F\left(z_k\right)-G\left(z_k\right).$$

On remarquera qu'on a :

$$G\left(z_k\right)=F\left(z_k\right)z_k^*$$

où z_k^* est la moyenne de la variable Z pour la population des individus ℓ vérifiant $z_\ell \leq z_k$. On écrira donc :

$$x_k = F_k\left(z_k - z_k^*\right),$$

et nous sommes arrivés au but que nous nous étions fixé.

8 - Estimation de la variance de GINI

Synthétisons les résultats :

- *NUM* se linéarise en $\frac{1}{\mathrm{N}}\left(x_k-\bar{x}\right)$
- $\hat{M}$ se linéarise en $\frac{1}{\mathrm{N}}\left(z_k-\bar{z}\right)$
- $\hat{R}$ se linéarise $\frac{1}{\mathrm{M}}.\frac{1}{\mathrm{N}}\left(F_k\left(z_k-z_k^*\right)-Rz_k-\bar{u}\right)$

avec $\bar{u}$ moyenne de $u_k = F_k\left(z_k-z_k^*\right)-Rz_k$

On a alors :

$$Var\left(\hat{GINI}\right) \simeq Var\left[\left(\sum_s \left(u_k - \bar{u}\right) w_k\right) / 2MN\right]$$

On est donc ramené à l'estimation d'un total pour une variable artificielle bien précise. Malheureusement, celle-ci n'est pas connue sur l'échantillon mais peut être, dans le calcul, remplacée par des estimations sans que cela fasse de dégâts. En effet, l'estimateur de variance calcule une forme quadratique finie :

$$\sum\sum_s A_{k\ell}\, x_k\, x_\ell$$

Si les x_k sont remplacés par des estimations introduisant des erreurs numériques de l'ordre de $1/\sqrt{n}$, l'erreur sur l'estimateur de variance sera également de l'ordre de $1/\sqrt{n}$ au plus, ce qui reste parfaitement honorable.

9 - Procédure de calcul

La chaîne de calcul pour l'estimation de GINI et l'estimation de sa variance procède donc comme suit :

1 - Ordonner les données dans le sens des z_k croissant.

2 - Calculer les cumuls W_k *des* w_k *et* Z_k *des* $z_k\, w_k$.

3 - Calculer les estimateurs ponctuels :

$$* \hat{N} = \sum_s w_k\,, \qquad * \hat{Z} = \sum_s z_k\, w_k$$

$$NUM = \sum_s z_k \frac{W_k}{\hat{N}} w_k$$

$$\hat{R} = NUM / \hat{Z}$$

$$\hat{GINI} = 2\,\hat{R} - 1.$$

4 -Calculer pour tout k la variable "artificielle" :

$$u_k = \frac{1}{\hat{Z}} \cdot \left(F_k\left(z_k - \frac{Z_k}{W_k} \right) - \hat{R} z_k \right)$$

5 - Calculer $\overline{u} = \frac{1}{\hat{N}} \sum u_k w_k$.

6 - Calculer la variance de $u_k - \overline{u}$ à l'aide d'un logiciel approprié.

10 - Une petite simulation pour se réconforter

Quand on cherche à établir un résultat de statistique par des voies théoriques, on est souvent conduit à s'interroger sur la validité des approximations qu'on a été amené à faire au cours de l'analyse. Comme en physique, une vérification expérimentale est souvent utile pour se convaincre que tout marche bien comme on le soupçonne (ou comme on le souhaite !). En particulier les résultats de [2] étaient particulièrement intriguants et inquiétants. On a donc réalisé une petite simulation dans l'esprit de cette étude de façon à en avoir le cœur net. Pour des raisons de volume de calcul, et aussi parce que le problème des petits échantillons dans les petites populations est relativement critique pour juger de la validité des approximations, présentons ici les résultats relatifs à une population comptant N = 200 individus dans laquelle on a échantillonné, par sondage aléatoire simple, n = 20 puis 40 individus. On a simulé ainsi 10 000 échantillons indépendants pour établir les résultats.

Le seul point qui pose un réel problème est celui de l'estimation de la fonctionnelle NUM. La simulation qui est présentée dans la suite vérifie que la variance est estimée décemment par linéarisation. Les individus de la population sont numérotés de k=1 à 200, et, à chacun d'eux on associe la variable:

$$z_k = (k/N)^a + 1/a$$

On a donné à a 12 valeurs échelonnées de 0,1 à 12,5 pour simuler des variables de plus en plus concentrées. Vu la définition de ces variables , le rang R_k de l'individu k vaut k . Pour un échantillon s de taille n , on notera i_k(=1 à n) le rang de l'individu k dans s . On considérera deux cas: celui où le rang R_k (= k) de l'individu échantillonné est effectivement observé (voir paragraphe 3) et celui où il est estimé à partir de l'échantillon par $\hat{R}_k = \frac{N}{n}(i_k - 0,5)$.

Autrement dit la quantité :

$$NUM = \sum_{k=1}^{N} (R_k - 0,5)\, z_k$$

a été estimée par :

$$\hat{NUM1} = \frac{N}{n} \sum_{k \in s} (R_k - 0,5) z_k = \frac{N}{n} \sum_{i=1}^{n} (R_{k_i} - 0,5) z_{k_i} \quad (10.1)$$

et :

$$\hat{NUM2} = \frac{N}{n} \sum_{k \in s} \hat{R}_k \; z_k = \frac{N}{n} \sum_{i=1}^{n} \frac{N}{n} (i - 0,5) z_{k_i} \quad (10.2)$$

Les résultats sont présentés dans le tableau annexé. En colonne 1 figure la constante a , en colonne 2 la valeur exacte de NUM.

Les colonnes 3 à 5 sont relatives à NÛM1 et présentent respectivement :

(3) : le biais relatif calculé par $\frac{1}{B} \frac{\sum(\hat{NUM1} - NUM)}{NUM} \times 100$

avec B = 10000 simulations et la somme qui porte sur ces 10000 expériences.

(4) :l'écart-type relatif calculé par $\frac{\sqrt{\frac{1}{B}\sum(\hat{NUM1} - NUM)^2}}{NUM} \times 100$.

(5) : l'écart-type relatif estimé. Si $\hat{V}$ est une estimation (parmi les 10000) de la variance de $\hat{NUM1}$, on calcule la quantité $\frac{\sqrt{\frac{1}{B}\sum \hat{V}}}{NUM} \times 100$. Ici, la variance $\hat{V}$ a donc été calculée en utilisant la procédure du paragraphe 3. Comme on pouvait s'y attendre , on ne décèle aucun biais significatif dans l'estimation de NUM; de même , la variance de NÛM est estimée sans biais apparent.

Les colonnes 6, 7, et 8 sont relatives à NÛM2 et donnent aussi le biais relatif, l'écart-type relatif et l'estimation de l'écart-type relatif. L'estimateur de variance utilisé ans cette colonne 8 est celui qui est décrit au paragraphe 9.

On constate un léger biais négatif pouvant atteindre -1,4% pour des variables très concentrées et des échantillons de taille 20 . La variance (ou l'écart-type relatif) est estimée très honorablement ; une légère sous-estimation apparaît lorsque le biais est important ce qui est bien naturel.

La colonne 9 donne "l'estimateur" de variance (ou d'écart-type relatif) basé sur la variable naïve utilisée dans [1] ou [2]. La variance est alors surestimée de 1,5 à 14000 fois quand on passe d'une variable très concentrée (a= 12,5) à une variable très dispersée. L'ordre de grandeur de 300 fois cité dans [2] pour une variable naturelle apparaît comme plausible.

Numériquement cet estimateur est très voisin de celui qui apparaît en colonne 5 , et on peut donc dire qu'en fait il estime la même quantité, c'est à dire la variance de NÛM1 . Ceci a quelque chose de très surprenant : NÛM1 utilise la variable exacte R_k alors que NÛM2 utilise son estimation $\hat{R}_k$; cependant , le deuxième estimateur est bien meilleur !

Quelle est l'explication de ce mystère ?

Je n'ai pas de réponse définitive mais on peut constater qu'on peut écrire :

$$N\hat{U}M1 = \int z_k \, F_k \, d\hat{F}_k$$

et $$N\hat{U}M2 = \int z_k \, \hat{F}_k \, d\hat{F}_k$$

Si les z_k étaient constants (mais pas les F_k !) , on aurait $N\hat{U}M2 = \overline{Z}/2$, avec une variance nulle ,tandis que NÛM1 conserverait une variance positive. D'un autre point de vue , regardons les formules 10-1 et 10-2 valides pour le sondage aléatoire simple .Dans la première les R_{k_i} diffèrent d'un échantillon à l'autre alors que dans la seconde les quantités iN/n ne varient pas ! On doit donc s'attendre à une variance plus grosse dans le premier cas que dans le second.

Il y a la un artefact statistique aussi étonnant que paradoxal , une véritable curiosité !

11 - Conclusion et voie d'avenir

On arrive ainsi à un traitement relativement honorable du problème de la variance du coefficient de GINI estimé sur des données d'enquête . Il n'a recours qu'à des traitements relativement standardisés et est adaptable, en principe à n'importe quelle enquête ayant un plan de sondage complexe. Il évite, en particulier, un traitement sur mesure par rééchantillonnage (Jacknife, Bootstrap, groupes aléatoires, demi-échantillons équilibrés). Les techniques de rééchantillonnage, en effet, ne sont adaptées qu'à certains plans, demandent toujours une programmation relativement spécifique et lourde, et des calculs qui peuvent atteindre un volume prohibitif.

Il se trouve, que, pendant que cette étude était en gestation, j'ai pu avoir accès à un projet de publication de D. BINDER [8]. Celui-ci s'intéresse au calcul de variance de certains indicateurs de dispersion dont l'indice de GINI. En utilisant la technique de linéarisation des équations estimantes, ils parvient, pour l'indice de GINI au même résultat, que celui qui est établi plus haut. Sa technique semble apparemment plus générale que celle que nous avons utilisée. Un examen soigneux des arguments montre que ce n'est qu'une apparence.

En fait, une véritable généralisation de ce qui vient d'être exposé devrait passer par l'utilisation de la notion de fonction d'influence telle qu'on l'utilise en statistique non paramétrique robuste. En modifiant un peu la notion pour tenir compte du contexte des populations finies, on peut arriver à des résultats extrêmement généraux et simples à la fois. Ceux-ci seront (si les petits cochons ne me mangent pas) exposés dans une prochaine étude.

BIBLIOGRAPHIE

[1] "The Estimation of the GINI and the Entropy Inequality Parameters in Finite Populations", Frédérik NYGARD and Arne SANDSTRÖM, *Journal of Official Statisitics*, Vol 1 n° 4, 1985 pp : 399-412

[2] Arne SANDSTRÖM, Jan WRETMAN and Bertil WALDEN, "Variance Estimators Of the GINI COEFFICIENT" In Simple random Sampling, origine douteuse, vraisemblablement GENUS (1987) pp : 41 à 70

[3] Gérard CALOT : *Statistique Descriptive* (DUNOD 1965)

[4] M. FLEURBAEY et S. LOLLIVIER : "Les mesures des inégalités : Abrégé théorique et pratique", Document de Travail du CREST n° 9408 bis (INSEE, 1994)

[5] R. CHAMBERS and R. DUNSTAN : "Estimating Distribution Functions from Survey Data", - *Biometrika* 1986 - vol 73 pp : 597-604

[6] Jean-Michel HOURRIEZ : "La significativité des variations du coefficient de GINI", Note "manuscrite" INSEE (1995)

[7] Ph. TASSI - B. LECOUTRE : *Statistique non-paramétrique et Robustesse,* Economica (1987).

[8] Milorad S. KOVACEVIC and David A. BINDER : "Variance Estimation for Measures of Income Inequality and Polarization - The Estimating Equations Approach", Document interne de Statistique Canada (1995).

[9] P.R. HALMOS : *An Introduction to Hilbert Spaces and Spectral Multiplicities.*

[10] WOODRUFF, R.S (1971). "A Simple Method for Approximating the Variance of a Complicated Estimate", *Journal of the American Statistical Association* 66, 411-414.

Tableau 1 :
Simulation d'estimation et d'estimation de précision de la variables NUM.

		Le rang est observé			Le rang n'est pas observé			
Constante	Total	biais rel	ect rel	ect rel estimé	biais rel	ect rel	ect rel est	pseudoest
		%	%	%	%	%	%	%
Taille 20								
0.100	219056.6	-0.052	12.325	12.360	-0.018	0.107	0.107	12.617
0.250	97797.7	-0.041	12.599	12.624	-0.083	0.527	0.525	12.886
0.500	56033.2	-0.036	13.308	13.310	-0.218	1.550	1.520	13.538
1.000	33383.2	-0.163	14.736	14.832	-0.495	3.815	3.726	15.031
1.500	24821.9	0.103	16.271	16.408	-0.580	5.974	5.820	16.525
2.000	20066.7	0.101	17.748	17.839	-0.722	7.994	7.718	17.913
3.000	14741.7	0.380	20.449	20.574	-0.698	11.496	11.096	20.514
4.000	11746.8	-0.110	22.838	22.877	-1.062	14.530	13.935	22.742
5.000	9797.9	-0.020	25.048	24.985	-1.161	17.228	16.431	24.792
7.500	6965.9	-0.005	29.946	29.734	-1.227	23.002	21.761	29.366
10.000	5424.9	-0.152	33.874	33.819	-1.392	27.559	26.216	33.326
12.500	4452.1	0.551	37.438	37.664	-0.881	31.581	30.230	37.057
Taille 40								
0.100	219056.6	-0.026	8.150	8.235	-0.008	0.069	0.070	8.312
0.250	97797.7	-0.004	8.364	8.416	-0.036	0.345	0.346	8.493
0.500	56033.2	0.163	8.755	8.864	-0.076	1.002	1.009	8.935
1.000	33383.2	-0.129	9.980	9.895	-0.237	2.567	2.509	9.954
1.500	24821.9	-0.017	10.981	10.927	-0.291	4.029	3.945	10.967
2.000	20066.7	0.004	11.797	11.905	-0.333	5.306	5.264	11.925
3.000	14741.7	-0.162	13.671	13.659	-0.504	7.732	7.571	13.651
4.000	11746.8	0.119	15.413	15.250	-0.398	9.860	9.553	15.213
5.000	9797.9	-0.265	16.683	16.656	-0.676	11.562	11.300	16.601
7.500	6965.9	0.113	19.731	19.855	-0.439	15.311	15.024	19.743
10.000	5424.9	0.167	22.595	22.569	-0.460	18.554	18.099	22.424
12.500	4452.1	0.222	24.930	25.044	-0.399	21.198	20.823	24.864