Indice de Gini et évolution de l'indice de Gini: linéarisation versus bootstrap pour estimer la variance

Guillame Chauvet[(1)] et Camelia Goga[(2)]

(1) ENSAI/CREST, Campus de Ker Lann,

(2) IMB, Université de Bourgogne,

Neuchâtel, 15 juin 2012

Définition de l'indice de Gini

- Soit $\mathcal{Y}$ une variable quantitative positive (le revenu, par example) et $F(\cdot)$ sa fonction de répartition ;
- L'indice de Gini (1914) est une mesure de concentration (parmi les plus connues) utilisée dans les études économiques,

$$G = \frac{1}{2} \frac{\int \int |v-u| dF(u) dF(v)}{\int u dF(u)}$$

en supposant que $\int u dF(u) \neq 0$.

- Il mesure la dispersion de $\mathcal{Y}$ à l'intérieur d'une population.

- Le plus souvent, G est utilisé pour évaluer les inégalités de revenu dans un pays à des différents époques ou entre différents pays à la même époque ;
- Durant les dernières années, G a été utilisé dans d'autres domaines que économique : biologie (Graczyk, 2007), environnement (Druckman and Jackson, 2008 ; Groves-Kirkby *et al.*, 2009) or astrophysique (Lisker, 2008) ;
- Sa définition est liée à la courbe de Lorentz (est le double de l'aire compris entre la courbe de Lorentz et la première bissectrice).

Définition de l'indice de Gini en population finie

- On considère une population $\mathbf{U}$ de taille N et $y_1, \ldots, y_N$ les valeurs de $\mathcal{Y}$ mesurées sur U.
- L'indice de Gini devient

$$G = \frac{1}{2} \frac{\sum_{k \in U} \sum_{l \in U} |y_k - y_l|}{N \sum_{k \in U} y_k},$$

  - si $y_k = y_l$, $k, l \in U$ alors $G = 0$:;
  - si $y_1 = \ldots = y_{N-1} = 0$ et $y_N > 0$, alors $G = 1 - \frac{1}{N}$;

- Pour des données distinctes, $y_1 < y_2 < \ldots < y_N$, l'indice de Gini peut s'écrire (Nygård and Sandström, 1985) :

$$G = \frac{\sum_U y_k \left(2F(y_k) - 1\right)}{t_Y} - \frac{1}{N}$$

- La même relation est valable si on a des ex-aequo (Deville, 1997) ;
- Le terme $1/N$ est appelé par Nygård and Sandström (1985) "correction en population finie de G" souvent négligé dans la littérature ;
- Dans la suite de l'exposé, on prend G sans le terme $1/N$.

Estimation de l'indice de Gini

- On sélectionne un échantillon s dans U de taille n selon un plan de sondage $p(\cdot)$; soient π_k et π_{kl} les probabilités de premier et deuxième degré ;

$$G = \frac{\sum_U y_k \left(2F(y_k) - 1\right)}{t_Y}$$

estimé par

$$\hat{G} = \frac{\sum_s \frac{y_k}{\pi_k} \left(2\hat{F}(y_k) - 1\right)}{\sum_s \frac{y_k}{\pi_k}}$$

- Variance $\mathrm{Var}(\hat{G})$?
- L'estimateur de la variance $\widehat{\mathrm{Var}}(\hat{G})$?

L'estimation de la variance de G a été étudiée par

- la méthode de linéarisation : Deville (1999), Kovačević & Binder (1997).
- jackknife généralisé : Berger (2008) ; il montre en particulier que c'est équivalent à la linéarisation.
- bootstrap (pour sondage aléatoire simple avec remise et stratifié) et intervalles de confiance basées sur la vraisemblance empirique (Qin *et al.*, 2010).

Principe des méthodes de linéarisation

1. **Linearisation de Taylor** : Särndal *et al.* 1992
2. **Equations estimantes** : Kovačević & Binder, 1997 ;
3. **Fonction d'influence** : Deville, 1999 ;

Consiste à trouver une variable linéarisée u_k (inconnue) et à approximer

$$\text{Var}(\hat{\phi}) \simeq \text{Var}\left(\frac{\sum_s u_k}{\pi_k}\right) = \sum_U \sum_U \Delta_{kl} \frac{u_k}{\pi_k} \frac{u_l}{\pi_l}$$

$$\widehat{\text{Var}}(\hat{\phi}) = \widehat{\text{Var}}\left(\frac{\sum_s \hat{u}_k}{\pi_k}\right) = \sum_s \sum_s \frac{\Delta_{kl}}{\pi_{kl}} \frac{\hat{u}_k}{\pi_k} \frac{\hat{u}_l}{\pi_l}$$

$\hat{u}_k$ l'estimateur de la variable linéarisée u_k

Estimation de la variance de G par linéarisation par la fonction d'influence

- Soit $M = \sum_U \delta_{y_k}$ et

$$G = T(M) = \frac{\int \{2F(y) - 1\} y dM(y)}{\int y dM(y)}$$

où $F(\cdot)$ est la fonction de répartition empirique définie comme

$$F(y) = \frac{1}{\int dM(y)} \int 1_{\{\xi \leq y\}} dM(\xi)$$

- Soit $\hat{M} = \sum_U w_k \delta_{y_k}$ avec $w_k = 1/\pi_k$ si $k \in s$ et zéro sinon

- par la fonction d'influence (Deville, 1999) on obtient la variable linéarisée de G :

$$u_k = 2F(y_k)\frac{y_k - \overline{y}_{k,<}}{t_y} - y_k\frac{G+1}{t_y} + \frac{1-G}{N}$$

où $\overline{y}_{k,<}$ est la moyenne de y_l inférieurs à y_k.

- pour SAS de taille n, l'estimateur de la variance de $\hat{G}$ est

$$v_{lin}(\hat{G}) = N^2\frac{1-f}{n}S^2_{\hat{u},s}$$

où $\hat{u}$ est l'estimation de u dans l'échantillon s.

L'estimation de la variance : méthodes de rééchantillonnage

1. **jackknife** : calculer de façon répétitive l'estimateur en enlevant une observation (Rao *et al.*, 1992, Berger & Skinner, 2005.)

2. **bootstrap**
   - bootstrap sans remise (Gross, 1980, Chauvet, 2007)
   - "mirror-match" bootstrap (Sitter, 1992)
   - "rescaled" bootstrap (Rao and Wu, 1988)

Algorithme de bootstrap sans remise (BWO) proposé par Gross

- Cet algorithme a été proposé par Gross (1980) pour le sondage alátoire simple sans remise (SAS) et étendu ultérieurement à d'autres plans de sondage ;
- On reproduit les conditions de tirage de départ, en dupliquant $m = N/n$ (supposé entier) fois chaque unité de s. Une pseudo-population U^* de taille N est ainsi créée.

  Soient y_k^* les valeurs répliquées de la variable d'intérêt sur U^*.

  (Plusieurs méthodes ont été proposées pour le cas N/n non entier).

- On sélectionné C échantillons s_c^* dans U^* selon SAS de taille n.
- Soit $\hat{M}_c^* = \sum_{s_c^*} w_k \mathbf{1}_{y_k^*}$ une réplication de la mesure M, avec $w_k = N/n$ pour $k \in s_c^*$ et zéro sinon ; $c = 1, \ldots, C$.
- Une réplication de $G = T(M)$ est obtenue par le principe de "plug-in",

$$\hat{G}_c^* = T(\hat{M}_c^*), \quad c = 1, \ldots, C.$$

- La variance de $\hat{G} = T(\hat{M})$ est estimée par

$$V_*\left(\hat{G}^*\right) \quad = \quad E_*\left\{\hat{G}^* - E_*(\hat{G}^*)\right\}^2,$$

où E_*, resp. V_*, est l'espérance, resp. la variance, par rapport au plan de ré-échantillonnage

- de façon empirique,

$$v_{BWO}(\hat{G}) \quad = \quad \frac{1}{C-1}\sum_{c=1}^{C}\left\{\hat{G}_c^* - \overline{\hat{G}^*}\right\}^2,$$

où $\overline{\hat{G}^*} = C^{-1}\sum_{c=1}^{C}\hat{G}_c^*$.

Intervalles de confiance : linéarisation

On suppose que le plan d'échantillonnage est tel que l'estimateur de Horvitz-Thompson satisfait le théorèm central limite.

L'intervalle de confiance asymptotique $(1-2\alpha)\%$ de $\hat{G}$ est

$$\left[\hat{G} - z_\alpha \sqrt{v_{lin}(\hat{G})}, \hat{G} + z_\alpha \sqrt{v_{lin}(\hat{G})}\right]$$

où z_α est le quantile d'ordre α de la loi $\mathcal{N}(0,1)$.

Intervalles de confiance : bootstrap

- méthode de percentiles : on considère les estimations bootstrapés ordonnées $\hat{G}^*_{(c)},\ c = 1, \ldots, C$ avec C le nombre de réplications et l'intervalle de confiance $(1-2\alpha)\%$ est

$$\left[\hat{G}^*_{(L)}, \hat{G}^*_{(U)}\right]$$

avec $L = \alpha C$ et $U = (1-\alpha)C$.

- bootstrap-t : on utilise les réplications ordonnées de la statistique pivotale $t = \dfrac{\hat{G} - G}{\sqrt{v_{BWO}(\hat{G})}}$. L'intervalle de confiance $(1-2\alpha)\%$ est

$$\left[\hat{G} - t^*_{(U)}\sqrt{v_{BWO}(\hat{G})}, \hat{G} - t^*_{(L)}\sqrt{v_{BWO}(\hat{G})}\right].$$

Estimation de l'évolution de l'indice de Gini entre deux périodes

- On veut estimer l'évolution de l'indice de Gini entre deux dates $d = 1, 2$

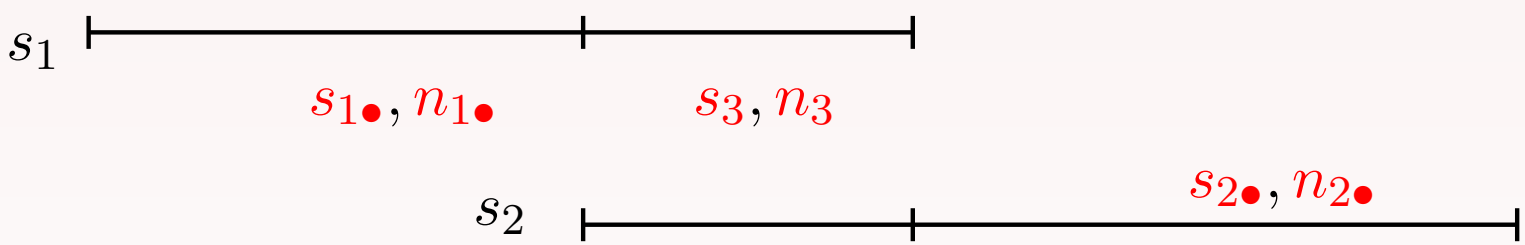
$$\Delta G = G_2 - G_1$$

- Goga *et al.* (2009) suggèrent un estimateur composite de ΔG; ils analysent également l'estimateur de la variance de ΔG obtenue par linéarisation par la fonction d'influence ;
- On propose d'étendre l'algorithme de Gross au cas de deux échantillons ;
- On compare l'estimateur de la variance obtenu par linéarisation et bootstrap dans le cas du plan de sondage aléatoire simple sans remise bi-dimensionnel

Le plan aléatoire simple sans remise bi-dimensionnel (SAS2)

Le plan bidimensionnel : une probabilité $p(s=(s_1,s_2))$ de sélectionner $s=(s_1,s_2)\in[\mathcal{P}(U)]^2$.

$$p(s)\geq 0 \quad \text{and} \quad \sum_s p(s)=1.$$



Le plan aléatoire simple sans remise bidimensionnel (SAS2) :

$$p(s=(s_1,s_2))=\frac{n_{1\bullet}!n_3!n_{2\bullet}!(N-n_{1\bullet}-n_3-n_{2\bullet})!}{N!}$$

$$\pi_k^{1\bullet}=\frac{n_{1\bullet}}{N},\quad \pi_k^{3}=\frac{n_3}{N},\quad \pi_k^{2\bullet}=\frac{n_{2\bullet}}{N}$$

Estimation composite de M_1 and M_2

- Soit $\mathcal{Y}_1$ la variable d'intérêt à l'époque $d = 1$ et $\mathcal{Y}_2$ la même variable mesurée à l'époque $d = 2$;
- L'évolution de Gini peut s'écrire $\Delta G = G_2 - G_1 = T(M_1, M_2)$
- La mesure $M_d = \sum_U \delta_{y_{dk}}$ est estimée sur $s_\diamond$ par (pour SAS2)

$$\hat{M}_{d,\diamond} = N n_\diamond^{-1} \sum_{k \in s_\diamond} \delta_{y_{dk}}$$

pour $d = 1, 2$ et $\diamond \in \{1\bullet, 3, 2\bullet\}$.

- Les estimateurs composites de M_1 et M_2 sont

$$\begin{aligned} \hat{M}_1^{co} &= a\ \hat{M}_{1,1\bullet} + (1-a)\ \hat{M}_{1,3}, \\ \hat{M}_2^{co} &= b\ \hat{M}_{2,2\bullet} + (1-b)\ \hat{M}_{2,3}, \end{aligned}$$

Estimation composite de ΔG

- L'estimation composite de ΔG est

$$\widehat{\Delta G}^{co} = \frac{\int \{2\hat{F}_2^{co}(y) - 1\} y d\hat{M}_2^{co}(y)}{\int y d\hat{M}_2^{co}(y)} - \frac{\int \{2\hat{F}_1^{co}(y) - 1\} y d\hat{M}_1^{co}(y)}{\int y d\hat{M}_1^{co}(y)}$$

où $\hat{F}_d^{co}(y) = \left\{ \int d\hat{M}_d^{co}(y) \right\}^{-1} \int 1_{\{\xi \leq y\}} d\hat{M}_d^{co}(\xi)$

- Pour SAS2, $\widehat{\Delta G}^{co} - \Delta G$ est approximé par (Goga *et al.*, 2009)

$$Nb(\bar{u}_{2,s_{2\bullet}} - \bar{u}_{2,s_3}) - Na(\bar{u}_{1,s_{1\bullet}} - \bar{u}_{1,s_3}) + N(\bar{u}_{2,s_3} - \bar{u}_{1,s_3}).$$

Variance asymptotique et estimateur de la variance pour SAS2

- La variance asymptotique est donnée par

$$V\left(\widehat{\Delta G}^{co}\right) \simeq N^2 \left\{c_1(a) S^2_{u_1,U} - 2c_{12}(a,b) S_{u_1 u_2,U} + c_2(b) S^2_{u_2,U}\right\}$$

avec $c_1(a), c_{12}(ab), c_2(b)$ constantes dépendant de a, b et n_1, n_2, n_3.

- Les estimateurs de la variance sont

$$v_{int}\left(\widehat{\Delta G}^{co}\right) = N^2 \left\{c_1(a) S^2_{\hat{u}_1,s_3} - 2c_{12}(a,b) S_{\hat{u}_1\hat{u}_2,s_3} + c_2(b) S^2_{\hat{u}_2,s_3}\right\}$$

$$v_{uni}\left(\widehat{\Delta G}^{co}\right) = N^2 \left\{c_1(a) S^2_{\hat{u}_1,s_1} - 2c_{12}(a,b) S_{\hat{u}_1\hat{u}_2,s_3} + c_2(b) S^2_{\hat{u}_2,s_2}\right\}$$

Deux cas particuliers

- L'estimateur intersection : $a = b = 0$ et SAS2,

$$\widehat{\Delta G}^{int} - \Delta G \simeq N(\bar{u}_{2,s_3} - \bar{u}_{1,s_3})$$

- L'estimateur union : $a = n_{1\bullet}/n_1$ et $b = n_{2\bullet}/n_2$ et SAS2

$$\widehat{\Delta G}^{uni} - \Delta G \simeq N(\bar{u}_{2,s_2} - \bar{u}_{1,s_1})$$

Algorithme de bootstrap pour SAS2

- on duplique N/n_3 fois pour $k \in s_3$; on obtient U^* de taille N; on note (y^*_{1k}, y^*_{2k}) les valeurs ainsi obtenues ;
- $s^* = (s^*_1, s^*_2)$ est sélectionné dans U^* selon SRS2 de taille $n_{1\bullet}$, n_3 et $n_{2\bullet}$;

  on note $s^*_{1\bullet} = s^*_1 \setminus s^*_2$, $s^*_3 = s^*_1 \cap s^*_2$ et $s^*_{2\bullet} = s^*_2 \setminus s^*_1$.
- Soit $\widehat{M}^{co*}_d,\ d = 1, 2$ la réplication de $\widehat{M}^{co}_d$ et $\widehat{\Delta G}^{co*} = T(\widehat{M}^{co*}_1, \widehat{M}^{co*}_2)$.

- La variance de $\widehat{\Delta G}^{co}$ est estimée par

$$V_*\left(\widehat{\Delta G}^{co*}\right) = E_*\left\{\widehat{\Delta}^{co*} - E_*(\widehat{\theta}^{co*})\right\}^2$$

- version empirique : $s_c^* = (s_{c1}^*, s_{c2}^*), c = 1, \ldots, C$, in U^*. La variance (bootstrap) de $\widehat{\Delta G}^{co}$ est estimée par

$$v_{BWO}(\widehat{\Delta G}^{co}) = \frac{1}{C-1}\sum_{c=1}^{C}\left\{\widehat{\Delta G}^{c,co*} - \overline{\widehat{\Delta G}^{co*}}\right\}^2,$$

où $\widehat{\Delta G}^{c,co*}$ est une réplication de $\widehat{\Delta G}^{co*}$ issue de s_c^* et $\overline{\widehat{\Delta G}^{co*}} = C^{-1}\sum_{c=1}^{C}\widehat{\Delta G}^{c,co*}$.

Cas de l'évolution d'une moyenne : estimateur intersection

On considère $\Delta y = \overline{y}_2 - \overline{y}_1$ et le plan SRS2.

- l'estimateur intersection : $\widehat{\Delta y}^{int} = N(\overline{y}_{2,s_3} - \overline{y}_{1,s_3})$ avec variance

$$V(\widehat{\Delta y}^{int}) = N^2 \frac{1-f_3}{n_3} S^2_{y_2-y_1,U}$$

- estimation par bootstrap

$$V_*(\widehat{\Delta y}^{int}) = \frac{N(n_3-1)}{n_3(N-1)} v_{int}(\widehat{\Delta y}^{int}) = \frac{N(n_3-1)}{n_3(N-1)} \cdot N^2 \frac{1-f_3}{n_3} S^2_{y_2-y_1,s_3}$$

Cas de l'évolution d'une moyenne : estimateur union

- l'estimateur union : $\widehat{\Delta y}^{uni} = N(\overline{y}_{2,s_2} - \overline{y}_{1,s_2})$ avec variance
$V(\widehat{\Delta y}^{uni}) =$
$N^2 \left(\frac{1-f_1}{n_1} \ S^2_{y_1,U} - 2\frac{n_3 - n_1 n_2/N}{n_1 n_2} \ S_{y_1 y_2,U} + \frac{1-f_2}{n_2} \ S^2_{y_2,U} \right)$
- estimation par bootstrap :

$$V_*(\widehat{\Delta y}^{int}) = \frac{N(n_3-1)}{n_3(N-1)} v_{int}(\widehat{\Delta y}^{uni}) \quad \text{avec}$$

$v_{int}(\widehat{\Delta y}^{uni}) =$
$N^2 \left(\frac{1-f_1}{n_1} \ S^2_{y_1,s_3} - 2\frac{n_3 - n_1 n_2/N}{n_1 n_2} \ S_{y_1 y_2,s_3} + \frac{1-f_2}{n_2} \ S^2_{y_2,s_3} \right)$

Une étude par simulations

- On simule 12 populations de taille $N = 1000$ (Deville, 1997)

$$y_{1k} = \left(\frac{k}{N}\right)^{\alpha} + \frac{1}{\alpha}$$

où α est une constante positive permettant de contrôler la concentration de la variable y_1. Plus précisement, l'indice de Gini coefficient croît avec α.

- On a utilisé 12 valeurs de α, comprises entre 0.1 et 12.5 $(0.1; 0.25; 0.5; 1; 1.5; 2; 3; 4; 5; 7.5; 10; 12.5)$.
- Les valeurs de la variable d'intérêt à la deuxième vague sont simulées selon

$$y_{2k} = 1 + y_{1k} + \epsilon_{hk}, \quad h = 1, \ldots, 12$$

$\epsilon_{hk} \simeq \mathcal{N}(0, \sigma_h^2)$ avec σ_h^2 tel que $R^2 = 0.8$

L'indice de Gini G et son évolution ΔG pour les populations simulées

	U_1	U_2	U_3	U_4	U_5	U_6
G_1	0.0039	0.0185	0.0499	0.1111	0.1607	0.2001
ΔG	0.0003	−0.0010	−0.0090	−0.0381	−0.0667	−0.0987

	U_7	U_8	U_9	U_{10}	U_{11}	U_{12}
G_1	0.2574	0.2966	0.3251	0.3709	0.3979	0.4159
ΔG	−0.1473	−0.1842	−0.2159	−0.2701	−0.3074	−0.3304

Estimation de G ainsi que de sa variance ("one-sample" case)

- $B = 1000$ échantillons aléatoires simple sans remise de taille $n = 50$, respectivement $n = 200$.
- On calcule G ainsi que l'estimateur de sa variance :
  - basé sur la linéarisation par la fonction d'influence, $v_{lin}(\hat{G})$;
  - en utilisant l'algorithme de bootstrap, $v_{BWO}(\hat{G})$ obtenu pour $C = 2000$ réplications ;

Mesures de comparaison

- le biais relatif (Monte Carlo) de $v(\hat{G})$

$$RB\{v(\hat{G})\} = 100 \times \frac{B^{-1}\sum_{b=1}^{B} v(\hat{G}_b) - MSE(\hat{G})}{MSE(\hat{G})},$$

$MSE(\hat{G})$ est une approximation obtenue à partir de 20000 simulations.

- l'erreur quadratique moyenne (Monte Carlo) de $v(\hat{G})$,

$$MSE(v(\hat{G})) = B^{-1}\sum_{b=1}^{B} v(\hat{G}_b) - MSE(\hat{G})$$

- $RE = \dfrac{MSE(v_{BWO}(\hat{G}))}{MSE(v_{lin}(\hat{G}))}$

Comparaison pour $n = 50$ et $n = 200$

	Linéarisation	Bootstrap		Linéarisation	Bootstrap	
	n=50				n=200	
	RB	RB	RE	RB	RB	RE
U_1	-4.46	-8.26	0.93	-0.67	-1.48	1.02
U_2	-2.10	-5.39	0.94	-0.21	-0.78	1.03
U_3	-1.37	-3.74	0.92	-1.58	-1.91	1.07
U_4	1.76	1.89	0.86	0.44	0.64	1.08
U_5	3.94	6.92	0.93	0.60	1.67	1.14
U_6	3.45	8.62	0.97	0.90	2.48	1.13
U_7	-0.25	6.86	0.98	-1.67	0.52	1.02
U_8	-1.79	5.60	0.97	-4.35	-2.04	0.99
U_9	-3.16	4.10	0.96	0.53	3.09	1.04
U_{10}	-9.83	-3.15	0.90	-1.42	1.29	1.03
U_{11}	-15.67	-8.22	0.88	-4.48	-1.50	0.99
U_{12}	-18.66	-10.69	0.84	-4.08	-0.66	0.99

Intervalles de confiance : linéarisation versus bootstrap

- pour $n = 50$, bootstrap-t (suivi par la linéarisation) donne des meilleurs résultats que la méthode de percentiles ;
- pour $n = 200$, les méthodes sont équivalentes en ce qui concerne le taux global de couverture ; le bootstrap-t donne des meilleurs taux de couverture uni-lateraux, excepté les populations très concentrées.

Estimation de l'évolution de G ("two-sample" case)

- On sélectionne $B = 1000$ échantillons SRS2 de taille $(n_{1\bullet}, n_3, n_{2\bullet}) = (200, 50, 200)$, respectivement $(n_{1\bullet}, n_3, n_{2\bullet}) = (50, 200, 50)$.
- Dans chaque échantillon, on calcule $\widehat{\Delta G}^{int}$ et $\widehat{\Delta G}^{uni}$.
- Les estimateurs de variance par linéarisation, $v_{lin}(\widehat{\Delta G})$, et bootstrap $v_{BWO}(\widehat{\Delta G})$ avec $C = 2000$ réplications, sont calculés.
- On utilise les mêmes mesures de comparaison : biais relatif (RB), l'erreur quadratique moyenne (MSE) et le rapport $RE = v_{BWO}(\widehat{\Delta G})/v_{lin}(\widehat{\Delta G})$.
- On compare les taux de couverture obtenus selon les deux méthodes (pour le bootstrap-t, on utilise l'approximation de la variance par la linéarisation).

Intersection estimateur : $(n_{1\bullet}, n_3, n_{2\bullet}) = (200, 50, 200)$

	Linearization				Bootstrap				
							t-Bootstrap		
Pop.	RB	L	U	L+U	RB	RE	L	U	L+U
	Sample size $(n_{1\bullet}, n_3, n_{2\bullet}) = (200, 50, 200)$								
U_1	0.15	3.30	3.60	6.90	-3.74	0.98	3.10	2.20	5.30
U_2	-2.17	3.30	2.40	5.70	-3.59	1.06	2.60	3.40	6.00
U_3	-1.75	2.70	3.60	6.30	-3.98	0.92	2.50	3.40	5.90
U_4	3.32	5.30	2.00	7.30	-0.43	0.87	1.90	2.60	4.50
U_5	4.66	3.70	2.00	5.70	2.98	0.91	2.70	3.60	6.30
U_6	1.16	3.60	1.50	5.10	4.73	1.05	2.20	3.70	5.90
U_7	4.77	4.90	2.00	6.90	12.32	1.02	3.40	1.80	5.20
U_8	-2.97	3.80	1.50	5.30	10.70	1.19	4.40	2.60	7.00
U_9	3.18	4.10	3.40	7.50	14.09	0.97	5.10	2.90	8.00
U_{10}	-5.19	4.10	5.40	9.50	5.35	0.79	3.90	1.80	5.70
U_{11}	-9.79	3.30	6.00	9.30	2.20	1.03	5.60	2.10	7.70
U_{12}	-17.23	5.30	7.60	12.90	-0.96	1.03	4.20	1.90	6.10

Intersection estimator : $(n_{1\bullet}, n_3, n_{2\bullet}) = (50, 200, 50)$

	Linearization				Bootstrap				
							t-Bootstrap		
Pop.	RB	L	U	L+U	RB	RE	L	U	L+U
	Sample size $(n_{1\bullet}, n_3, n_{2\bullet}) = (50, 200, 50)$								
U_1	-1.65	3.20	3.00	6.20	-2.24	1.07	2.10	2.50	4.60
U_2	-1.49	2.70	2.60	5.30	-0.99	1.07	2.10	2.20	4.30
U_3	1.82	3.20	2.60	5.80	1.67	0.89	2.40	3.10	5.50
U_4	-1.56	2.30	2.00	4.30	-2.10	1.06	2.90	2.20	5.10
U_5	1.00	3.20	2.10	5.30	1.05	1.00	1.70	2.90	4.60
U_6	-0.80	2.50	1.90	4.40	-0.89	1.01	3.50	2.30	5.80
U_7	0.35	2.70	2.10	4.80	2.79	1.01	3.00	2.70	5.70
U_8	-0.58	1.70	2.70	4.40	5.19	1.11	2.50	2.70	5.20
U_9	-0.92	2.60	3.60	6.20	4.30	1.02	3.10	2.40	5.50
U_{10}	-3.14	2.20	4.30	6.50	3.61	1.08	2.70	1.60	4.30
U_{11}	-2.42	2.70	4.10	6.80	1.04	1.02	2.60	1.50	4.10
U_{12}	-1.56	2.40	2.90	5.30	2.93	0.99	1.60	1.10	2.70

Union estimateur : $(n_{1\bullet}, n_3, n_{2\bullet}) = (200, 50, 200)$

	Linearization				Bootstrap				
						Percentile			
Pop.	RB	L	U	L+U	RB	RE	L	U	L+U
	Sample size $(n_{1\bullet}, n_3, n_{2\bullet}) = (200, 50, 200)$								
U_1	-6.10	4.50	3.30	7.80	-8.96	1.01	3.70	1.80	5.50
U_2	-5.95	4.40	4.30	8.70	-6.63	1.05	3.40	3.60	7.00
U_3	-0.47	2.60	2.80	5.40	-2.96	1.03	6.60	5.50	12.10
U_4	4.46	2.10	3.30	5.40	2.51	0.93	10.50	7.20	17.70
U_5	6.66	3.20	1.90	5.10	4.83	0.89	12.70	9.10	21.80
U_6	5.91	2.80	1.50	4.30	5.62	0.99	15.90	7.90	23.80
U_7	7.71	2.20	2.50	4.70	6.19	0.97	15.40	4.30	19.70
U_8	4.92	3.50	2.40	5.90	5.29	1.10	18.30	4.30	22.60
U_9	5.11	4.30	2.70	7.00	2.74	0.96	17.20	4.80	22.00
U_10	-0.16	4.10	4.20	8.30	-0.71	0.79	18.30	6.90	25.20
U_11	-3.35	5.20	2.70	7.90	-2.60	1.06	21.50	8.30	29.80
U_12	-7.52	4.60	4.30	8.90	-2.50	1.06	24.40	10.30	34.70

Union estimateur : $(n_{1\bullet}, n_3, n_{2\bullet}) = (50, 200, 50)$

	Linearization				Bootstrap				
						Percentile			
Pop.	RB	L	U	L+U	RB	RE	L	U	L+U
	Sample size $(n_{1\bullet}, n_3, n_{2\bullet}) = (50, 200, 50)$								
U_1	-1.30	2.60	3.20	5.80	-1.95	1.06	0.20	0.30	0.50
U_2	1.23	2.30	2.50	4.80	1.05	1.30	0.30	0.60	0.90
U_3	0.59	2.70	1.50	4.20	0.28	1.00	1.00	1.80	2.80
U_4	-1.23	2.40	2.50	4.90	-1.50	1.08	2.90	1.20	4.10
U_5	0.53	2.90	2.00	4.90	1.03	1.11	2.00	1.80	3.80
U_6	-0.54	2.30	1.30	3.60	-0.22	1.07	4.40	1.30	5.70
U_7	0.05	3.40	2.50	5.90	2.09	1.01	3.40	1.40	4.80
U_8	-0.51	2.60	2.50	5.10	3.66	1.09	3.40	1.00	4.40
U_9	-1.93	2.00	3.00	5.00	1.40	1.01	5.00	0.40	5.40
U_{10}	-3.48	2.50	2.70	5.20	1.54	1.08	5.50	1.00	6.50
U_{11}	-3.29	2.70	2.90	5.60	-1.06	1.02	5.30	1.70	7.00
U_{12}	-0.58	3.60	2.60	6.20	2.68	0.99	5.70	1.20	6.90

Courte bibliographie

- Chauvet, G. (2007).Méthodes de Bootstrap en population finie. Ph.D. dissertation, Université de Rennes 2.
- Deville, J.C. (1997). Estimation de la variance du coefficient de Gini mesurée par sondage. Actes des Journées de Méthodologie Statistique, Insee Méthodes.
- Deville, J.C. (1999). Variance estimation for complex statistics and estimators : linearization and residual techniques. Survey Methodology, 25, 193-203.
- Druckman, A. and Jackson, T. (2008). Measuring resource inequalities : The concepts and methodology for an area-based Gini coefficient. Ecological Economics, 65, 242-252.
- Gini, C. (1914). Sulla misura della concentrazione e della variabilità dei caratteri. Atti del R. Istituto Veneto di Scienze Lettere ed Arti.
- Goga, C., Deville, J.C. and Ruiz-Gazen, A. (2009). Composite estimation and linearization method for two-sample survey data. Biometrika, 96, 691-709.
- Graczyk, P.P. (2007). Gini Coefficient : A New Way To Express Selectivity of Kinase Inhibitors against a Family of Kinases. Journal of Medicinal Chemistry, 50, 5773Ű5779.
- Gross, S.T. (1980). Median estimation in sample surveys. ASA Proceedings of Survey Research, 181-184.

- Groves-Kirkby, C.J., Denman, A.R. and Phillips, P.S. (2009). Lorenz Curve and Gini Coefficient : Novel tools for analysing seasonal variation of environmental radon gas. Journal of Environmental Management, 90, 2480-2487.
- Kovačević, M.S. and Binder, D.A. (1997). Variance Estimation for Measures of Income Inequality and Polarization - The Estimating Equation Approach. Journal of Official Statistics.13, 41-58.
- Lisker, T. (2008). Is the Gini coefficient a stable measure on galaxy structure ? The Astrophysical Journal Supplement Series. 179, 319-325.
- Qin, Y., Rao, J.N.K. and Wu, C. (2010). Empirical likelihood confidence intervals for the Gini measure of income inequality. Economic Modelling. 27, 1429-1435.
- Rao, J.N.K. and Wu, C.F.J. (1988). Resampling inference with complex survey data. Journal of the American Statistical Association. 83, 231-241.
- Sitter, R.R. (1992a). A resampling procedure for complex survey data. Journal of the American Statistical Association. 87, 755-765.
- Sitter, R.R. (1992b). Comparing three bootstrap methods for survey data. Canadian Journal of Statistics. 20, 135-154.