MATHÉMATIQUES ET SCIENCES HUMAINES

MARC BARBUT

Diamètres et écarts. Une décomposition du coefficient d'inégalité de C. Gini

Mathématiques et sciences humaines, tome 93 (1986), p. 61-69

<http://www.numdam.org/item?id=MSH_1986__93__61_0>

Math. Sci. hum. (23^{e} année, n°93, *1986*, p.61-69)

DIAMETRES ET ECARTS
UNE DECOMPOSITION DU COEFFICIENT D'INEGALITE DE C. GINI

Marc BARBUT*

Les façons les plus usuelles de mesurer la dispersion (ou la concentration) d'une distribution, consistent à calculer leur écart à une valeur centrale représentative de la distribution.

Il y a quelque temps, Henri Rouanet attirait mon attention sur une autre idée naturelle : celle de sommer (ou de "moyenner") les écarts entre les valeurs observées prises deux à deux; ceci conduit à la notion de diamètre d'une distribution.

La note qui suit montre que les deux procédés conduisent toujours à des mesures de la dispersion ayant le même ordre de grandeur.

Dans le cas particulier des écarts absolus, la relation entre diamètre moyen et écart moyen fournit une décomposition intéressante du coefficient d'inégalité de C. Gini.

1. CENTRES ET ECARTS

Soit α un réel donné, supérieur ou égal à 1 . On sait que, sur $\mathbb{R}^n$, on peut lui associer une norme (dite de Hölder) et une distance définies par :

$$||\vec{x}||_\alpha = \left(\sum_{i=1}^{n} |x_i|^\alpha \right)^{1/\alpha}$$

$$\Delta_\alpha(\vec{x},\vec{y}) = \left(\sum_{i=1}^{n} |x_i - y_i|^\alpha \right)^{1/\alpha} = ||\vec{x} - \vec{y}||_\alpha$$

---

* Centre d'Analyse et de Mathématique Sociales, 54 bd Raspail Paris 6ème.

En particulier, $\alpha = 2$ correspond au cas de la distance (et de la norme) euclidienne, $\alpha = 1$ à celui de la distance parfois appelée "de Manhattan", et α infini à celui de la norme :

$$||\vec{x}||_\infty = \max_i |x_i|$$

parfois appelée "de la convergence uniforme".

Lorsque les coordonnées $x_1, x_2, \ldots, x_n$ des vecteurs $\vec{x}$ de $\mathbb{R}^n$ représentent les valeurs observées d'une variable numérique X (sur n sujets 1,2,...,n par exemple), on peut résumer les observations par une <u>valeur centrale</u> qui soit à distance minimum de celles-ci.

Lorsque $\mathbb{R}^n$ est muni de la distance Δ_α définie supra, ceci conduit à la définition du <u>centre d'ordre</u> α de $\vec{x}$, noté $c_\alpha(\vec{x})$, qui est la valeur de t qui rend minimum :

$$\Delta_\alpha(\vec{x}, t\vec{u}) = \left(\sum_{i=1}^{n} |x_i - t|^\alpha \right)^{1/\alpha}$$

Où $\vec{u} = (1,1,\ldots,1)$.
Pour $\alpha > 1$, $c_\alpha(\vec{x})$ est unique; en particulier,

- pour $\alpha = 2$, la moyenne : $c_2(\vec{x}) = \dfrac{x_1 + x_2 + \ldots + x_n}{n} = m(\vec{x})$

- pour $\alpha = \infty$, le milieu de l'intervalle séparant la plus petite des observations de la plus grande : $c_\infty(\vec{x}) = \dfrac{\max_i x_i + \min_i x_i}{2}$

Pour $\alpha = 1$, $c_1(\vec{x})$ est la médiane (unique) μ, des valeurs observées lorsque n est impair, et n'importe quel nombre de l'intervalle médian sinon.

Une mesure de la <u>dispersion</u> des observations est alors fournie par l'<u>écart d'ordre α</u> de $\vec{x}$:

$$E_\alpha(\vec{x}) = \Delta_\alpha(\vec{x}, c_\alpha\vec{u}) = \left(\sum_{i=1}^{n} |x_i - c_\alpha|^\alpha \right)^{1/\alpha}$$

Et par l'<u>écart-moyen d'ordre α</u> :

$$\ell_\alpha(\vec{x}) = \frac{1}{n^{1/\alpha}} E_\alpha(\vec{x}) = \left(\sum_{i=1}^{n} \frac{|x_i - c_\alpha|^\alpha}{n} \right)^{1/\alpha}$$

En particulier, pour $\alpha = 2$, $c_\alpha(\vec{x})$ est l'écart-type, pour $\alpha = 1$, l'écart moyen absolu, et pour α infini, la demi-étendue

$$\ell_\infty(\vec{x}) = \frac{1}{2}\,(\max_i x_i - \min_i x_i)$$

Pour tout α $(\alpha \geq 1)$, l'écart et l'écart-moyen vérifient des propriétés naturelles pour des mesures de la dispersion :

-1- $\forall \lambda \in \mathbb{R}$, $\ell_\alpha(\lambda\vec{x}) = |\lambda|\, c_\alpha(\vec{x})$, i.e. :
l'écart s'exprime dans la même unité que la variable X observée.

-2- $\forall t \in \mathbb{R}$, $\ell_\alpha(t\vec{u}) = 0$, i.e. :
lorsque toutes les observations sont égales, la dispersion est à son minimum.

-3- $\forall \vec{x} \in \mathbb{R}^n$, $\forall t \in \mathbb{R}$, $\ell_\alpha(\vec{x} + t\vec{u}) = \ell_\alpha(\vec{x})$, i.e. :
un changement d'origine (translation) pour la variable observée ne modifie pas sa dispersion.

-4- $\forall \vec{x}, \forall \vec{y} \in \mathbb{R}^n$, $\ell_\alpha(\vec{x}+\vec{y}) \leq \ell_\alpha(\vec{x}) + \ell_\alpha(\vec{y})$, i.e. :
l'écart est une fonction sous-additive du vecteur des observations.

Enfin, pour chaque vecteur $\vec{x}$ des observations fixé, $\ell_\alpha(\vec{x})$ et $E_\alpha(\vec{x})$ sont des fonctions <u>monotones non décroissantes</u> de α ; en particulier :

$$\forall \vec{x} \ , \quad \ell_1(\vec{x}) = \frac{1}{n} \sum_i |x_i - \mu| \leq \ell_2(\vec{x}) = \sigma(\vec{x}) \leq c_\infty(\vec{x}) = \frac{\max x_i - \min x_i}{2}$$

Un exposé élémentaire détaillé, et les démonstrations des propriétés classiques des normes de Hölder, et des écarts et centres qui s'en déduisent, est disponible dans : M. BARBUT "*Sur deux familles de valeurs centrales généralisant la moyenne arithmétique : moyenne d'ordre* α *, centres d'ordre* α ", <u>Documents du C.A.M.S.</u>, n° P003, Septembre 1983.

2. DIAMETRES

L'écart-moyen d'ordre α mesure la dispersion par la distance du vecteur des observations au vecteur le moins dispersé (le plus concentré) qui est le meilleur "résumé" de celles-ci : distance de $\vec{x}$ à $c_\alpha\vec{u}$, où $c_\alpha = c_\alpha(\vec{x})$ est le centre de $\vec{x}$.

Une autre façon de mesurer la dispersion de $\vec{x} = (x_1, x_2, \ldots, x_n)$ consiste à prendre en compte les différences $|x_i - x_j|$ par paires de valeurs observées; c'est, comme on le verra plus loin, ce qui est fait notamment dans le calcul du coefficient de concentration de C. Gini.

Appelons <u>diamètre d'ordre α</u> de $\vec{x}$ le nombre :

$$D_\alpha(\vec{x}) = \left(\sum_{i<j}\sum |x_i-x_j|^\alpha\right)^{1/\alpha} = \left(\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} |x_i-x_j|^\alpha\right)^{1/\alpha}$$

Le <u>diamètre moyen</u> $d_\alpha(\vec{x})$ est obtenu en prenant la moyenne sur les $\frac{n(n-1)}{2}$ paires $\{x_i,x_j\}$:

$$d_\alpha(\vec{x}) = \left(\frac{2}{n(n-1)}\sum_{i<j}\sum |x_i-x_j|^\alpha\right)^{1/\alpha} = \left(\frac{2}{n(n-1)}\right)^{1/\alpha} D_\alpha(\vec{x})$$

Il est immédiat de vérifier que :

Pour $\alpha = 2$, $D_2(\vec{x}) = n\sigma(\vec{x})$ où σ est l'écart-type

$$d_2(\vec{x}) = \left(\frac{n}{n-1}\right)^{1/2} \sqrt{2}\sigma(\vec{x}) = \left(\frac{n}{n-1}\right)^{1/2} \sqrt{2}\, \ell_2(\vec{x})$$

Pour $\alpha = \infty$, $D_\infty(\vec{x}) = d_\infty(\vec{x}) = \max x_i - \min x_i = 2\ell_\infty(\vec{x})$

Dans ces deux cas, le diamètre moyen est donc du même ordre de grandeur que l'écart; ceci est général.

Quelque soit $\alpha \geqq 1$, on a en effet la double inégalité :

(1) $$\boxed{2\ell_\alpha(\vec{x}) \geqq d_\alpha(\vec{x}) \geqq \left(\frac{n}{n-1}\right)^{1/\alpha} \ell_\alpha(\vec{x})}$$

Montrons d'abord la seconde de ces inégalités.

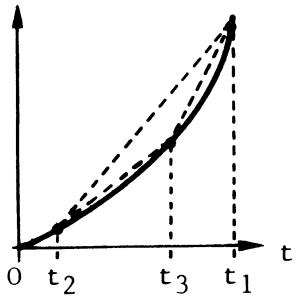
Soit $m = \frac{x_1+x_2+\ldots+x_n}{n}$ la moyenne de $\vec{x}$. Par définition :

$$\ell_\alpha(\vec{x}) \leqq \left(\sum_{i=1}^{n} \frac{|x_i-m|^\alpha}{n}\right)^{1/\alpha} = \varepsilon_\alpha(\vec{x})$$

Or : $$|x_i-m| = \frac{1}{n}\left|nx_i - \sum_1^n x_j\right| = \frac{1}{n}\sum_{j=1}^{n}|x_i - x_j|$$

D'où : $$|x_i-m|^\alpha \leqq \frac{1}{n^\alpha}\left(\sum_{j=1}^{n}|x_i - x_j|\right)^\alpha \leqq \frac{1}{n}\sum_{j=1}^{n}|x_i - x_j|^\alpha$$

La dernière inégalité ci-dessus résulte de ce que la fonction $f(t) = t^\alpha$ ayant, pour $\alpha \geqq 1$, et $t \geqq 0$, sa convexité vers le haut (Fig. 1), elle satisfait, quels que soient $t_1,t_2,t_3,\ldots,t_k$, à l'inégalité de convexité :



$$f\left(\frac{1}{k}\sum_{i=1}^{k} t_i\right) \leqq \frac{1}{k}\sum_{i=1}^{k} f(t_i)$$

On a donc :

$$|x_i - m|^\alpha \leqq \frac{1}{n} \sum_{j=1}^{n} |x_i - x_j|^\alpha \qquad (i \in \{1,2,\ldots,n\})$$

Soit en sommant par rapport à i :

$$\sum_{i=1}^{n} |x_i - m|^\alpha \leqq \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} |x_i - x_j|^\alpha$$

En divisant les deux membres par $\frac{n(n-1)}{2}$, et en les élevant à la puissance $1/\alpha$, il vient finalement :

$$d_\alpha(\vec{x}) \geqq \left(\frac{n}{n-1}\right)^{1/\alpha} \varepsilon_\alpha(\vec{x}) \geqq \left(\frac{n}{n-1}\right)^{1/\alpha} \ell_\alpha(\vec{x})$$

L'inégalité : $2\ell_\alpha(\vec{x}) \geqq d_\alpha(\vec{x})$
se démontre également en utilisant la convexité de $f(t) = t^\alpha$ pour $\alpha \geqq 1$, $t \geqq 0$.

Soit $c = c_\alpha(\vec{x})$ le centre de $\vec{x} = (x_1,x_2,\ldots,x_n)$. Pour toute paire $\{i,j\}$, on a :

$$|x_i - x_j| \leqq |x_i - c| + |x_j - c|$$

Donc $|x_i - x_j|^\alpha \leqq (|x_i - c| + |x_j - c|)^\alpha$

Or $f(t) = t^\alpha$ satisfait en particulier à :

$$\forall u \geqq 0 , \quad \forall v \geqq 0 \qquad f\left(\frac{u+v}{2}\right) \leqq \frac{f(u) + f(v)}{2}$$

D'où : $(|x_i - c| + |x_j - c|)^\alpha \leqq 2^{\alpha-1}(|x_i - c|^\alpha + |x_j - c|^\alpha)$

Et : $|x_i - x_j|^\alpha \leqq 2^{\alpha-1}(|x_i - c|^\alpha + |x_j - c|^\alpha)$

Sommons sur tous les couples (i,j) , $i \neq j$:

$$\sum_{i=1}^{n} \sum_{j=1}^{n} |x_i - x_j|^\alpha \leqq 2^{\alpha-1} \sum_{i=1}^{n} \sum_{j=1, i\neq j}^{n} (|x_i - c|^\alpha + |x_j - c|^\alpha) =$$

$$= 2^{\alpha-1}\left(2(n-1) . \sum_{i=1}^{n} |x_i - c|^\alpha\right) = 2^\alpha (n-1)(E_\alpha(\vec{x}))$$

En divisant par $n(n-1)$, et en élevant à la puissance $1/\alpha$, on obtient l'inégalité annoncée.

La double inégalité (1), ou, en simplifiant :

(1') $2\ell_\alpha \geqq d_\alpha > \ell_\alpha$

montre que les expressions "centre" et "diamètre" utilisées sont cohérentes avec l'interprétation de l'écart comme un "rayon" du vecteur des observations.

N.B. -1- Tout ce qui a été montré dans ce paragraphe reste vrai si l'on remplace $\mathbb{R}^n$ par un espace métrique (E,δ) quelconque, dont $x_1,x_2,\ldots,x_n$ sont n points, et en remplaçant partout les $|x_i - x_j|$ par $\delta(x_i,x_j)$.

-2- On vérifie aisément que le diamètre et le diamètre moyen satisfont aux propriétés des écarts rappelées à la fin du paragraphe 1 ci-dessus.

1. $\forall \lambda \geqq 0$, $d_\alpha(\lambda\vec{x}) = \lambda\, d_\alpha(\vec{x})$
2. $\forall t \in \mathbb{R}$ $\quad d_\alpha(t\vec{u}) = 0$
3. $\forall t \in \mathbb{R}$, $\forall \vec{x} \in \mathbb{R}^n$, $d_\alpha(\vec{x} + t\vec{u}) = d\alpha(\vec{x})$
4. $\forall \vec{x}$, $\forall \vec{y} \in \mathbb{R}^n$, $d\alpha(\vec{x} + \vec{y}) \leqq d_\alpha(\vec{x}) + d_\alpha(\vec{y})$

et $d_\alpha(\vec{x})$ est, pour $\vec{x}$ fixé, une fonction monotone non décroissante de α .

3. LE CAS $\alpha = 1$, ET APPLICATION AU COEFFICIENT D'INEGALITE DE C. GINI

Supposons $\alpha = 1$, le centre est alors la médiane μ pour $n = 2k + 1$, et n'importe quelle valeur de l'intervalle fermé médian pour $n = 2k$.

Supposons les indices i choisis de façon que :

$$x_1 \leqq x_2 \leqq x_3 \leqq \ldots \leqq x_n \ .$$

On a alors, si $n = 2k+1$, $\mu = x_{k+1}$ et (en abrégeant l'écriture) :

$$(2)\quad D_1(\vec{x}) = \sum_{i<j} |x_i - x_j| = \sum_{i<j<k+1} + \sum_{k+1<i<j} + \sum_{i<k+1<j} + \sum_{i=1}^{2k+1} |x_i - x_{k+1}|$$

Si l'on pose : $\vec{x}_I = (x_1,x_2,\ldots,x_k)$ et $\vec{x}_s = (x_{k+2},x_{k+3},\ldots,x_{k+1})$ les deux vecteurs $\vec{x}_I$ et $\vec{x}_s$ de $\mathbb{R}^k$ peuvent être interprétés comme ceux des observations <u>petites</u> (au dessous de la médiane) d'une part, <u>grandes</u> de l'autre; s'il s'agissait de revenus (x_i distribués à des individus i), on dirait : les "pauvres", et les "riches".

D'autre part, pour $i < k+1 < j$, on a :

$|x_i - x_j| = |x_i - \mu| + |\mu - x_j|$, d'où

$$\sum_{i<k+1<j} |x_i - x_j| = k \sum_{i=1}^{2k+1} |x_i - \mu| = k\, E_1(\vec{x})$$

Finalement, la décomposition (2) s'écrit :

(3) $D_1(\vec{x}) = D_1(\vec{x}_I) + D_1(\vec{x}_s) + (k+1)\, E_1(\vec{x})$

Pour $n = 2k$, on obtiendrait :

(3') $D_1(\vec{x}) = D_1(\vec{x}_I) + D_1(\vec{x}_s) + k\, E_1(\vec{x})$

En passant aux diamètres moyens, on obtient :

(4) pour $n = 2k+1$, $d_1(\vec{x}) = \frac{k-1}{2(2k+1)}(d_1(\vec{x}_I) + d_1(\vec{x}_s)) + \frac{k+1}{k}\, \ell_1(\vec{x})$

(4') pour $n = 2k$, $d_1(\vec{x}) = \frac{k-1}{2(2k-1)}(d_1(\vec{x}_I) + d_1(\vec{x}_s)) + \frac{2k}{2k-1}\ell_1(\vec{x})$

Prenons k grand, et omettons l'indice, il vient :

(5) $$d(\vec{x}) \cong \ell(\vec{x}) + \frac{1}{4}\,(d(\vec{x}_I) + d(\vec{x}_s))$$

Le diamètre moyen absolu est approximativement égal à l'écart moyen absolu, majoré du quart de la somme des diamètres moyens des petites valeurs (les "pauvres") et des grandes valeurs (les "riches").

D'autre part, l'inégalité : $d \leqq 2\ell$ montrée supra prouve que :

(6) $$\ell(x) \geqq \frac{1}{4}\,(d(\vec{x}_I) + d(\vec{x}_s))$$

Interprétons (4) et (5) en termes de coefficient d'inégalité de C. Gini. On sait que l'une des expressions de celui-ci est :

$$G(x) = \frac{1}{2n^2\, m(\vec{x})} \sum_{i=1}^{n} \sum_{j=1}^{n} |x_i - x_j| = \frac{D_1(\vec{x})}{n^2 m(\vec{x})}$$

où : $m(\vec{x}) = m = \dfrac{x_1+x_2+\ldots+x_n}{n}$

On déduit de (4) (pour $n = 2k + 1$)

(7) $$G(\vec{x}) = \frac{k^2}{(2k+1)^2}\left[\frac{m(\vec{x}_I)G(\vec{x}_I)+m(\vec{x}_s)G(\vec{x}_s)}{m(\vec{x})}\right] + \frac{k+1}{2k+1}\,\frac{\ell(\vec{x})}{m(\vec{x})}$$

Soit approximativement :

(8) $$G(\vec{x}) \cong \frac{1}{2}\,\frac{\ell(\vec{x})}{m(\vec{x})} + \frac{1}{4}\left[\frac{m(\vec{x}_s)}{m(\vec{x})}\, G(\vec{x}_I) + \frac{m(\vec{x}_s)}{m(\vec{x})}\, G(\vec{x}_s)\right]$$

ou, en écriture abrégée, et compte tenu de : $m_I + m_s \cong 2m$

(9) $$\boxed{G \cong \frac{1}{2}\left(\frac{\ell}{m} + p_I\, G_I + p_s\, G_s\right)} \qquad \left(p_I = \frac{m_I}{m_I+m_s}\ ,\ p_s = \frac{m_s}{m_I+m_s}\right)$$

L'inégalité globale G , au sens du coefficient de C. Gini, est donc la moyenne arithmétique de deux termes; le premier, parfois appelé écart moyen relatif :

$$\frac{\ell}{m} = \frac{E}{X} = \frac{\Sigma|x_i - \mu|}{\Sigma x_i}$$

est l'écart absolu à la médiane, rapporté au total distribué.

Le second :

$p_I G_I + p_s G_s$ (avec $p_I + p_s = 1$)

est la moyenne des coefficients d'inégalités des "pauvres" et des "riches", pondérés par les proportions p_I et p_s du total distribué que chacune de ces deux catégories reçoit respectivement.

Ceci éclaire, me semble-t-il, le lien entre deux mesures (G et ℓ/m) usuelles de l'inégalité.

D'ailleurs, l'inégalité (6) vue ci-dessus, se traduit, compte tenu de (1') et (9), par :

$$(10) \qquad \frac{\ell}{m} \geqq G \geqq \max \begin{cases} \frac{\ell}{2m} \\ p_I G_I + p_s G_s \end{cases}$$

Ce qui signifie notamment que le coefficient de Gini de l'inégalité globale est toujours au moins égal à la moyenne pondérée (par les proportions qu'ils reçoivent respectivement) des coefficients relatifs aux "pauvres" d'une part, et aux "riches" d'autre part, et inférieur à l'écart moyen relatif.

A titre d'exemples :

- la situation la plus inégalitaire, $x_1 = x_2 = \ldots = x_{2k} = 0$, $x_{2k+1} = X$,

  donne = $p_I = G_I = 0$, $p_s = G_s = G = \frac{\ell}{m} = 1$

- la situation la plus égalitaire, $x_1 = x_2 = \ldots = x_{2k} = x_{2k+1} = s$, donne :

  $\frac{\ell}{m} = G = G_I = G_s = 0$

- une situation intermédiaire, $\forall i$, $x_i = is$, donne :

  $G = \frac{1}{3}$, $\frac{\ell}{m} = \frac{1}{2}$, $p_I G_I + p_s G_s = \frac{1}{4} \cdot \frac{1}{3} + \frac{3}{4} \cdot \frac{1}{9} = \frac{1}{6}$

Remarquons enfin qu'en itérant (8), et en divisant l'ensemble $\{1,2,\ldots,i,\ldots,n\}$ des détenteurs de "revenus" x_i en quatre (quartiles), huit (octiles), seize, etc ... successivement, on obtient, dans les notations évidentes

$$G \cong \frac{1}{2m}\left[\ell + \frac{\ell_I+\ell_s}{2} + \frac{1}{2^2}(\ell_{II}+\ell_{Is}+\ell_{sI}+\ell_{ss}) + \frac{1}{2^3}(\ell_{III}+\ell_{IIs}+\ldots+\ell_{sss})+\ldots\right]$$