**COEFFICIENT DE CONCENTRATION (C5, F3)**
(15 / 01 / 2020, © Monfort, Dicostat2005, 2005-2020)

La notion de **concentration** peut se « résumer » à l'aide d'une grandeur (en général scalaire) associée à une **loi de probabilité** : cette grandeur est donc une **caractéristique légale**.

Ainsi, on considère un **espace probabilisé** $(\Omega, \mathcal{T}, P)$ et un **espace d'observation** $(\mathcal{X}, \mathcal{B})$ dans lequel $\mathcal{X}$ est un **espace topologique** muni de sa **tribu borélienne** $\mathcal{B}$. On suppose que la **lp** $P^{\xi}$ de la **va** $\xi : \Omega \mapsto \mathcal{X}$ est une **loi unimodale**. Alors, si $C \in \mathcal{B}$ est une **partie mesurable** donnée qui est aussi une **partie centrale** de $P^{\xi}$, la **probabilité** $P^{\xi}(C)$ peut constituer un indicateur de concentration élémentaire relatif à $P^{\xi}$. Si $Q = \mathcal{X} \setminus C$ est une **queue** de $P^{\xi}$, le rapport $P^{\xi}(C) / P^{\xi}(Q) = (1 - P^{\xi}(C))^{-1} . P^{\xi}(C)$ représente un autre indicateur de concentration.

Un **coefficient de concentration**, ou parfois **coefficient d'inégalité**, est une mesure scalaire de la **dispersion** ou du « **degré d'uniformité** » d'une **va** positive ou de sa **loi**. Il décrit donc la « répartition » d'une va positive (taillle des individus, revenu des ménages, chiffre d'affaires des entreprises, etc). On parle aussi d'**indice de concentration**, ou parfois d'**indice d'inégalité**.

Il existe de nombreux coefficients de ce type, qui sont souvent sans dimension (ie de dimension nulle) pr aux **unités de mesure**.

(i) **Coefficient théoriques**. On considère un **espace probabilisé** $(\Omega, \mathcal{T}, P)$ et une **vars** positive $\xi : \Omega \mapsto \mathbf{R}_+$ de **loi** $P^{\xi} = \xi(P)$.

On appelle :

(a) **coefficient de C. GINI**, ou **coefficient de C. GINI - M.O. LORENZ**, (théorique) le nombre :

$$(1) \qquad \gamma_G = 2 . \int_{\mathbf{R}+} F(x)\, dG(x) - 1,$$

où F est la **fr** associée à $P^{\xi}$ et $G(x) = (E\,\xi)^{-1} \int_0^x dF(u)$, $\forall\, x \in \mathbf{R}_+$ . L'indice $\gamma_G$ dérive de la courbe $(u = F(x), v = G(x))_{x \in \mathbf{R}+}$, paramétrée par x (cf **coefficient de GINI**) ;

(b) **coefficient d'entropie généralisée**, ou **coefficient d'entropie d'ordre** $\alpha$, (théorique) le nombre :

$$(2) \qquad \varepsilon(\alpha) = \int_{\mathbf{R}+} \{1 - (f(x))\}^{\alpha}\, dF(x), \qquad \forall\, x \in \,]-1, +\infty[,$$

dont l'**entropie** constitue un cas particulier (avec $\alpha \to 0$), où l'on note $f = P^{\xi} / d\lambda_1$ la **densité** de $P^{\xi}$ pr à la **mesure de LEBESGUE**.

(ii) **Coefficient empiriques**. Les coefficient précédents possèdent des analogues empiriques (cf **statistique naturelle**) : ces derniers sont obtenus en remplaçant F (resp $P^{\xi}$) par la **fonction de répartition empirique** $F_N$ (resp par la **loi empirique** $P_N$) associée à un N-**échantillon**.

Si $X = (X_1, ..., X_N) : \Omega \mapsto \mathbf{R}_+^N$ est un **échantillon aléatoire** (non nécessairement indépendant), les **statistiques** suivantes sont des exemples de **coefficients de concentration empiriques**, ou **coefficients d'inégalité empiriques** :

(a) le **coefficient de C. GINI** (empirique) :

(3) $$G_N = (2 N^2)^{-1} . (1 / \bar{X}_N) . \Sigma_{\alpha = 1}^{N} \Sigma_{\beta = 1}^{N} |X_\alpha - X_\beta|,$$

généralement associé à la **« courbe » de concentration empirique** du même nom (ou **courbe de C. GINI - O. LORENZ**), définie à partir des points (cf **courbe de LORENZ**) :

(4) $$\begin{cases} U_\alpha = \alpha / N \\ V_\alpha = \Sigma_{n=1}^{\alpha} X_n / \Sigma_{n=1}^{N} X_n , \end{cases} \qquad \forall \alpha = 1,..., N,$$

où $X_n$ désigne la coordonnée d'indice n de l'échantillon X ordonné de façon croissante (cf **statistique d'ordre**). Ainsi, $U_\alpha$ représente le nombre relatif cumulé d'**unités statistiques**, et $V_\alpha$ la valeur relative cumulée de la variable $\xi$ observée sur ces unités ;

(b) l'**indice de O.C. HERFINDAHL - A.O. HIRSCHMAN** (empirique), défini selon :

(5) $$H_N^2 = \Sigma_{n=1}^{N} p_n^2, \quad \text{avec } p_n = (\Sigma_{\alpha=1}^{N} X_\alpha)^{-1} . X_n = X_n / e_N' X \ \text{(poids des unités)},$$

qui s'écrit aussi $H_N^2 = N^{-1} (C_N^2 + 1)$, où $C_N^2$ est le carré du **coefficient de variation** empirique : $C_N^2 = S_N^2 / \bar{X}_N^2$, avec $S_N^2 = N^{-1} \Sigma_{n=1}^{N} (X_n - \bar{X}_N)^2$. Le coefficient $(H_N^2)' = 1 - H_N^2$ est aussi appelé **indice de A.O. HIRSCHMAN** ;

(c) l'**indice de A.F. SHORROKS**, ou **indice d'entropie généralisée**, (empirique), qui consiste en une famille d'indices de la forme :

(6) $$S_N (\alpha) = N^{-1} \alpha^{-1} (\alpha - 1)^{-1} . \Sigma_{n=1}^{N} \{(X_n / \bar{X}_N )^{\alpha} - 1\}, \qquad \forall \alpha \in \mathbf{R}^*_+ .$$

Lorsque $\alpha \to 0+$, on obtient la **déviation logarithmique moyenne** :

(7) $$S_N (0) = N^{-1} . \Sigma_{n=1}^{N} \mathrm{Log} ( \bar{X}_N / X_n ).$$

Lorsque $\alpha \to 1-$, on obtient le **coefficient de H. THEIL** (cf **entropie**) :

(8) $$S_N (1) = N^{-1} \Sigma_{n=1}^{N} (X_n / \bar{X}_N) \mathrm{Log} (X_n / \bar{X}_N) ;$$

lorsque $\alpha = 2$, on obtient une relation simple avec le coefficient de variation empirique :

(9) $S_N(2) = (1/2)\, C_N^2$ ;

(d) le **coefficient de PARETO** (cf **loi de PARETO**).

(iii) Si les poids $p_n$ définis précédemment sont ordonnés par valeurs décroissantes, on appelle **distribution de poids** le vecteur $p = (p_1, ..., p_N)' \in S_N$ (**simplexe** de dimension N - 1 de $\mathbf{R}^N$). La concentration des unités n, appréciée à travers ces poids, peut alors être définie (empiriquement) à l'aide d'une fonction $c_N : S_N \mapsto \mathbf{R}$.

Pour définir une notion de **concentration relative**, on dit qu'une distribution p' fondée sur N unités est **plus concentrée** qu'une distribution p" fondée sur le même nombre d'unités ssi :

(10) $c_N(p') > c_N(p")$.

La fonction $C_N$ doit vérifier des propriétés interprétables :

(a) **symétrie** : $c_N(\sigma(p)) = c_N(p), \forall \sigma \in \boldsymbol{\sigma}_N$ (groupe des **permutations** de $N_N^*$) ;

(b) non décroissance par **agrégation** quelconque. Si deux unités $\alpha$ et $\beta \neq \alpha$ sont « regroupées », l'indicateur de concentration $c_N$ ne doit pas décroître, ie :

(11) $c_N(p_1, ..., p_\alpha + p_\beta, ..., p_{\beta-1}, 0, p_{\beta+1}, ..., p_N) \geq c_N(p)$,

pour tout couple $(\alpha, \beta)$ tq $\beta \neq \alpha$ et toute distribution p fondée sur X ;

(c) non décroissance par **agrégation** ordonnée. Si l'on regroupe une unité $\beta$ avec une unité antérieure $\alpha < \beta$ (ie une unité strictement plus importante), l'indicateur de concentration $c_N$ doit croître :

(12) $c_N(p_1, ..., p_\alpha + p_\beta, ..., p_{\beta-1}, 0, p_{\beta+1}, ..., p_N) > c_N(p)$,

pour tout $(\alpha, \beta)$ tq $\beta \neq \alpha$ et toute distribution p fondée sur X ;

(d) décroissance avec N dans le cas « uniforme » : si les unités sont de même importance, avec $X_n = X_0$ (donné), $\forall$ n, (ie si $p_n = p_0 = N^{-1}, \forall n$), la fonction $c_N$ doit diminuer lorsque le nombre N d'unités augmente :

(13) $N' < N" \Rightarrow c_{N'}(e_{N'} C / N') > c_{N"}(e_{N"} C / N")$,

où $e_N = (1, ..., 1)' \forall \mathbf{R}^N, \forall N \in \mathbf{N}^*$.

(iv) A titre d'exemple, si f est une fonction adéquate, on peut définir un **indicateur de A. JACQUEMIN** de la forme :

(14) $c_N(p) = \Sigma_{n=1}^N p_n\, f(p_n)$.

Ainsi :

(a) si f est une fonction **constante** discontinue tq $f(p_n) = 1$, $\forall n \leq \alpha$, et $f(p_n) = 0$, $\forall n > \alpha$, on définit le poids des $\alpha$ premières unités $C_\alpha$ (avec $\alpha / N \leq C_\alpha \leq 1$) ;

(b) si f est la fonction linéaire $f(p_n) = p_n$, on définit le coefficient de HERFINDAHL-HIRSCHMAN précédent (avec $1 / N \leq H_N^2 \leq 1$) ;

(c) si f est la fonction logarithmique $f(p_n) = - \text{Log } p_n$, $\forall n$, on définit l'entropie (empirique) de la distribution p, ie $E_N = - \sum_{n=1}^{N} p_n \text{ Log } p_n$ (avec $0 \leq E_N \leq \text{Log } N$).