**DISPERSION (C5, F3)**
(30 / 04 / 2020, © Monfort, Dicostat2005, 2005-2020)

La **dispersion** est une propriété selon laquelle les valeurs d'une **variable aléatoire** sont, dans l'ensemble, plus ou moins « proche » d'une valeur donnée, ou encore plus ou moins regroupées, ou concentrées, autour d'une **partie centrale**. On parle aussi de **variabilité** ou d'**éparpillement**. La notion ne se définit pas seulement pour une **variable numérique** ; elle peut s'étendre à une **variable qualitative**.

On peut définir un concept de dispersion dès que l'on peut définir :

(a) une notion d'**écart**, ou de **distance**, entre les valeurs prises par une **variable aléatoire** ;

(b) une notion de **valeur de référence**, ou **valeur centrale**, pour apprécier la **distribution** (ou répartition) des écarts ou distances entre les valeurs.

(i) Un indicateur synthétique, permet souvent de « résumer » cette propriété. Ainsi, $(\Omega, \mathcal{T}, P)$ étant un **espace probabilisé** et $\xi : \Omega \mapsto \mathbf{R}^K$ un **vecteur aléatoire** de loi $P^\xi$ et de carré intégrable, un **indicateur de dispersion** par rapport à un vecteur (certain) donné $\alpha \in \mathbf{R}^K$, peut être :

(a) de type scalaire (espérance du carré de la norme, ou moyenne des carrés des écarts) :

(1) $\sigma^2(\alpha) = E\,||\xi - \alpha||^2$ ;

(b) de type matriciel (espérance d'un produit matriciel d'écarts vectoriels, ou moyenne des produits des écarts) :

(2) $\Sigma^2(\alpha) = E\,(\xi - \alpha)(\xi - \alpha)'$.

Le plus souvent, $\alpha$ est une **caractéristique** de **centralité** ou un **paramètre de position** de $P^\xi$ : **espérance mathématique**, **mode**, **quantile**.

L'indicateur de dispersion est alors une **caractéristique**, souvent de même nature (centralité ou position), de la loi d'une va $\eta$ définie à partir de $\xi$ selon un **changement de variable aléatoire** donné :

(3) $\eta = \varphi(\xi - \alpha)$,

où $\varphi : \mathbf{R}^K \mapsto \mathbf{R}^L$ est une fonction donnée.

Dans l'exemple ci-dessus, on considère très souvent le cas où $\alpha = E\,\xi$. D'autre part, dans (1), on a posé L = 1 et $\varphi(e) = ||e||^2$ (**norme** euclidienne) avec $e \in \mathbf{R}^K$ (cf **espace euclidien**) ; dans (2), on a posé L = K ($\mathbf{R}^K$ étant identifié à $M_{K,1}(\mathbf{R})$) et $\varphi(e) = e\,e'$, avec $e \in \mathbf{R}^K$.

Si $P^\xi$ est une **loi unimodale**, on peut aussi considérer le cas modal, dans lequel $\alpha$ = S $\xi$ (**mode** de $\xi$) et définir un indicateur de dispersion analogue à (1) ou (2) selon (mode du carré des écarts pr au mode) :

(4) $\sigma^2 (\alpha) = S \|\xi - \alpha\|^2$,

ou par (mode d'un produit matriciel d'écarts pr au mode) :

(5) $\Sigma^2 (\alpha) = S \{(\xi - \alpha)(\xi - S\ \alpha)'\}$.

La définition d'un indicateur de dispersion n'est pas toujours aisée. Ainsi, on peut mesurer la dispersion du vecteur $\xi$ à l'aide de :

(6) $E \|\xi - \alpha\|^2 = E (\xi - \alpha)'(\xi - \alpha) = \mathrm{tr}\ (V\ \xi) + \|E\ \xi - \alpha\|^2$,

où V $\xi$ est la **matrice des covariances** de $\xi$, du fait de l'**équation triviale** $\xi - \alpha = (\xi - E\ \xi) + (E\ \xi - \alpha)$. Mais cette représentation scalaire possède l'inconvénient de ne pas permettre une comparaison toujours efficace de la dispersion entre deux va. En effet, si $\eta : \Omega \mapsto \mathbf{R}^K$ est une autre va, on peut avoir :

(7) $E \|\xi - \alpha\|^2 < E \|\eta - \alpha\|^2$,

avec, néanmoins, $E(\xi_{k0} - \alpha_{k0})^2 > E(\eta_{k0} - \alpha_{k0})^2$ pour un indice $k_0 \in N_K^*$ (une coordonnée de $\xi$ est, en moyenne, plus dispersée que la coordonnée de même indice de $\eta$).

C'est pourquoi l'on préfère plutôt utiliser des définitions tq :

(8) $E (h'\ \xi - h'\ \alpha)^2 = q (h)$,

**forme quadratique** en $h \in \mathbf{R}^K$ qui s'écrit :

(9) $q (h) = h'\ E (\xi - \alpha)(\xi - \alpha)'\ h = h'\ (D_\alpha\ \xi)\ h$,

où $D_\alpha\ \xi$, aussi notée D $(\xi, \alpha)$, désigne la **matrice de dispersion** de $\xi$ autour de $\alpha$. La **matrice des covariances** de $\xi$ est alors définie comme $V\ \xi = D\ (\xi, E\ \xi)$, aussi notée D $\xi$, ce qui n'est autre que $\Sigma^2 (\alpha)$ définie en (2) avec $\alpha = E\ \xi$. Dans ce cas :

(9) $q (h) = V (h'\ \xi)$.

(ii) Par ailleurs, on peut considérer la **dispersion entre lois** elles-mêmes, au lieu de leurs seules caractéristiques de dispersion. Si $\alpha$ est considéré comme une **variable dégénérée**, de loi $\delta_\alpha$ (**loi de DIRAC** au point $\alpha \in \mathbf{R}^K$), on dit que $\xi$ est une **variable aléatoire dispersée** pr à $\alpha$ dès que $P^x \neq \delta_\alpha$ . Soit $\mathscr{P}$ l'ensemble des **mesures de probabilité** sur $\mathscr{T}$ et $\mathscr{P}^\xi$ l'image de $\mathscr{P}$ par $\xi$. Si l'on munit $\mathscr{P}^\xi$ d'une distance $\delta$, un **indicateur de dispersion** (scalaire) de $\xi$ pr à $\alpha$ sera donc naturellement défini par la distance (cf **distance entre probabilités**) :

(10) $\delta^2(\alpha) = d^2(P^\xi, \delta_\alpha)$, $\forall P^\xi \in \mathscr{P}^\xi$.

Par suite, la **dispersion relative entre lois**, eg $P_1{}^\xi$ et $P_2{}^\xi$, ayant même paramètre de centralité $\alpha$, sera définie par :

(11) $\Delta^2(P_1, P_2) = d^2(P_1{}^\xi, P_2{}^\xi)$.

(iii) On peut relier la notion de dispersion à celle de **centralité**. En effet, soit $(\mathscr{X}, \mathscr{B})$ un **espace probabilisable** dans lequel $\mathscr{X}$ est un **espace topologique** et $\xi : \Omega \mapsto \mathscr{X}$ une va donnée, dont la **lp** $P^\xi$ est supposée absolument continue pr à une **mesure positive** $\sigma$-finie $\mu$ définie sur $\mathscr{B}$. Si $C \in \mathscr{B}$ est une **partie centrale** pour $\xi$ (ou pour sa loi), alors la dispersion (scalaire) de $\xi$ dans C peut être définie à l'aide du « **volume** » de C, mesuré par $\mu$ (C). La notion de dispersion est aussi à rapprocher de celles d'**aplatissement** (cf **coefficient d'applatissement**) ou de **queue d'une loi**.