COEFFICIENT D'ASSOCIATION (C5, D2)
(24 / 12 / 2019, © Monfort, Dicostat2005, 2005-2019)

Un **coefficient d'association** est un scalaire, généralement sans dimension pr aux **unités de mesure**, qui résume la « liaison » existant entre des **variables aléatoires** (cf **association**, **corrélation**, **dissimilarité**).

En général, les variables considérées sont des **variables qualitatives** : le coefficient d'association joue alors un rôle analogue à celui du **coefficient de corrélation** pour des **variables numériques**.

(i) Soit $(\Omega, \mathscr{T}, P)$ un **espace probabilisé**, $(\mathscr{Y}', \mathscr{C}')$ et $(\mathscr{Y}'', \mathscr{C}'')$ deux espaces non numériques, $\eta' : \Omega \mapsto \mathscr{Y}'$ et $\eta'' : \Omega \mapsto \mathscr{Y}''$ deux variables qualitatives.

On dit que η' et η'' sont des **variables associées** ssi le **coefficient de corrélation linéaire** ρ, liant un **codage** (numérique) ξ' de η' à un codage (numérique) ξ'' de η', n'est pas nul. S'il est positif (resp négatif), l'association est qualifiée conformément : resp **association positive**, **association négative**.

On obtient le schéma suivant :

$$(1) \qquad \left. \begin{array}{l} (\Omega, \mathscr{T}, P) \rightarrow (\eta') \rightarrow (\mathscr{Y}', \mathscr{C}') \rightarrow (c') \rightarrow (\mathscr{X}', \mathscr{B}') \\ \\ (\Omega, \mathscr{T}, P) \rightarrow (\eta'') \rightarrow (\mathscr{Y}'', \mathscr{C}'') \rightarrow (c'') \rightarrow (\mathscr{X}'', \mathscr{B}'') \end{array} \right. \qquad \rightarrow \ \rho,$$

(où les triplets symboliques $\rightarrow (\eta) \rightarrow$ et $\rightarrow (c) \rightarrow$ désignent les **applications** correspondantes), ainsi que la formule du **coefficient d'association (théorique)** :

$$(2) \qquad \rho \ = \ \begin{array}{l} (V\, \xi')^{-1/2}\, (V\, \xi'')^{-1/2}\, .\, C\, (\xi', \xi'') \\ \\ (V\, c' \circ \eta')^{-1/2}\, (V\, c'' \circ \xi'')^{-1/2}\, .\, C\, (c' \circ \eta', c'' \circ \eta''). \end{array}$$

Le coefficient précédent dépend des codages c' et c'' choisis. Dans certains cas, le codage est « naturel » (eg utilise des **variables indicatrices** ou des **variables de comptage**). Dans d'autres cas, on doit chercher des codages adaptés au problème à résoudre.

(ii) La définition (2) s'applique de même à des observations $(Y_n')_n$ et $(Y_n'')_n$ relatives resp aux variables η' et η'' et à des observations $(X_n')_n$ et $(X_n'')_n$ relatives resp aux variables ξ' et ξ''. Elles conduisent directement à la notion de **coefficient d'association (empirique)** (cf **statistique naturelle**).

(iii) L'étude d'un **tableau de contingence** conduit souvent à définir des **coefficients de contingence** (cf **contingence**, **test du chi-deux**) : ces derniers sont des coefficients d'association relatifs aux variables qualitatives (« critères ») définissant le tableau.

(iv) Dans le cas d'un (2,2)-tableau T = {(a, b), (c, d)}, croisant les variables η' et η'', supposées chacune à deux modalités et codées avec le même codage c, à valeurs dans N_1 = {0, 1}, on peut interpréter l'élément courant de l'ensemble {a, b, c, d} comme la **fréquence absolue** de la case (i, j) de T, ie comme le nombre d'observations $n \in N_N^*$ tq $\xi' = u$ et $\xi'' = v$, avec $(u, v) \in (N_1)^2$, ie :

T =

	$\xi'' = 0$	$\xi = 1$
$\xi' = 0$	a	b
$\xi' = 1$	c	d

On peut, dans ce **contexte statistique**, définir plusieurs coefficients d'association :

(a) le **coefficient de G.U. YULE** :

(3) $Y = \text{dét}\, T / (a\,d + b\,c) = (a\,b + b\,c)^{-1} \,.\, (a\,d - b\,c)$;

(b) le **coefficient de « colligation »** :

(4) $C = \{1 - (a\,d)^{-1/2}\,(b\,c)^{1/2}\} / \{1 + (a\,d)^{-1/2}\,(b\,c)^{1/2}\}$;

(c) le **carré moyen de contingence** :

(5) $\Phi^2 = N^{-1} \,.\, \mathcal{X}_2^{\,2}$,

où $\mathcal{X}_2^{\,2}$ est la **statistique du chi-deux** calculée sur T.