

COEFFICIENT DE CORRÉLATION PONCTUELLE (C5, D2, F3)

(19 / 11 / 2019, © Monfort, Dicostat2005, 2005-2019)

Le **coefficient de corrélation ponctuelle** est un **coefficient de corrélation** défini pour deux **variables qualitatives** (ou « **attributs** ») à deux modalités chacune : **variables indicatrices** ou variables dichotomiques.

(i) Soit $(\Omega, \mathcal{F}, P)$ un **espace probabilisé** et $(\xi', \xi'') : \Omega \mapsto \mathcal{X}' \times \mathcal{X}''$ un **couple aléatoire** (qualitatif ou quantitatif) dont chaque coordonnée possède deux modalités (ou valeurs) : $X' = \{a', b'\}$ et $X'' = \{a'', b''\}$. On effectue un **codage** de chacune d'elles en posant : $\eta' = 1_{[\xi' = a']}$, ce qui définit un couple (η', η'') à valeurs dans $\{0,1\}^2$. La variable entière $\eta' + \eta''$ mesure donc le nombre de coordonnées tq $\xi = a$.

On appelle **coefficient de corrélation ponctuelle (théorique)** entre ξ' et ξ'' le **coefficient de corrélation linéaire** (théorique) entre leurs codifiées, ie :

$$(1) \quad \rho_p(\xi', \xi'') = \rho_{\eta'\eta''} = C(\eta', \eta'') / \{\sigma(\eta') \cdot \sigma(\eta'')\}.$$

(ii) Soit $((X_1', X_1''), \dots, (X_N', X_N''))$ un **N-échantillon** constitué de **copies** de (ξ', ξ'') et $((Y_1', Y_1''), \dots, (Y_N', Y_N''))$ l'échantillon correspondant aux variables codées. On note resp (sous forme matricielle) :

$$X = [(X_1', X_1'') / \dots / (X_N', X_N'')] \in M_{N,2}(\mathbf{R}) \quad \text{ou} \quad X = [x_1', x_1''] \in M_{N,2}(\mathbf{R})$$

$$Y = [(Y_1', Y_1'') / \dots / (Y_N', Y_N'')] \in M_{N,2}(\mathbf{R}) \quad \text{ou} \quad Y = [y_1', y_1''] \in M_{N,2}(\mathbf{R})$$

les matrices (aléatoires) constituées des 2 colonnes et N lignes (où / indique un saut de ligne matriciel).

La **matrice des coïncidences** est définie par :

$$(2) \quad M_c = Y' Y \in S_2(\mathbf{N}),$$

et son terme général est $m_{kl} = y_k' y_l$ (nombre d'observations possédant une modalité de ξ' et une modalité de ξ'') (où ' dénote aussi la transposition matricielle).

On appelle alors **coefficient de corrélation ponctuelle (empirique)** de y_k et y_l le nombre:

$$(3) \quad r_{kl} = \frac{(y_k' P_N y_k)^{-1/2} \cdot (y_l' P_N y_l)^{-1/2} \cdot y_k' P_N y_l}{\{m_{kk}(N - m_{kk}) m_{ll}(N - m_{ll})\}^{-1/2} \cdot (N m_{kl} - m_{kk} m_{ll})}.$$

(iii) Le coefficient de corrélation ponctuelle tient souvent lieu, notamment en **classification**, de **coefficient d'association** mesurant des « liens » ou des **indices de similarité** entre variables qualitatives.

(iv) Les notions précédentes s'étendent directement à un nombre $K \geq 2$ de variables $\xi_1, \dots, \xi_K$, chacune d'elles étant à valeurs dans un ensemble $\mathcal{X}_k = \{a_k, b_k\}$ ($\forall k \in \mathbf{N}_K^*$). Alors, $M_c \in S_K(\mathbf{N})$ et (3) est toujours valide, avec une écriture en colonnes $[y_1, \dots, y_K]$ de Y .