COEFFICIENT DE CORRÉLATION DES RANGS (D2, F3, F6, I)
(02 / 05 / 2020, © Monfort, Dicostat2005, 2005-2020)

Si, dans un coefficient de corrélation (empirique), on remplace les deux vecteurs d'observations (ou échantillons) par les statistiques de rang associées, on obtient un **coefficient de corrélation des rangs**.

(i) Soit $(\Omega, \mathcal{T}, P)$ un **espace probabilisé** et $(\xi, \eta) : \Omega \mapsto \mathbf{R}^2$ un **couple aléatoire** dont la **loi** est $P^{(\xi, \eta)}$. On observe un **échantillon** $((X_n , Y_n))_{n=1,...,N}$ de ce couple et l'on dissèque cet échantillon en deux échantillons $X = (X_1 ,..., X_N)$ et $Y = (Y_1 ,..., Y_N)$.

On définit le **coefficient de corrélation (linéaire) des rangs de C.E. SPEARMAN** (cf **coefficient de corrélation des rangs de SPEARMAN**)

Dans la définition du **coefficient de corrélation** linéaire (empirique) :

(1) $\quad r_{XY} = (X' P X)^{-1/2} (Y' P Y)^{-1/2} . X' P Y,$

dans laquelle P désigne la **matrice de centrage par rapport à la moyenne**, on remplace les vecteurs X et Y par leurs **statistiques de rang** respectives R et S, ie :

(2) $\quad r_{XY}' = r_{RS} = ||U||^{-1} . ||V||^{-1} . U' V,$

avec $U = R - (1/2) (N + 1) e_N$, $V = S - (1/2) (N + 1) e_N$, $||x||^2 = \Sigma_{n=1}^{N} x_n^2$ (norme euclidienne usuelle), $U = (U_1 ,..., U_N)$ et $V = (V_1 ,..., V_N)$.

Ce coefficient est souvent noté r_N (ou simplement r) : on l'appellle alors **« rho » de SPEARMAN**. C'est le plus utilisé. Il s'écrit aussi :

(2)' $\quad r_{XY}' = 1 - 6 . \{(N - 1) N (N + 1)\}^{-1} . ||R - S||^2,$

ou encore sous la forme d'une **statistique de HOEFFDING** :

(2)'' $\quad r_{XY}' = 3 . \{(N - 1) N (N + 1)\}^{-1} . \Sigma_\alpha \Sigma_\beta w (X_\alpha - X_\beta) . w (Y_\alpha - Y_\beta),$

dans laquelle $w (x) = -1$ si $x < 0$, $w (x) = 1/2$ si $x = 0$ et $w (x) = +1$ si $x > 0$.

(ii) Si l'on note C_N la **statistique de test** associée au **test de SPEARMAN**, le coefficient de SPEARMAN devient :

(2)''' $\quad r_{XY}' = \{12 / (N - 1) N (N + 1)\} C_N - N (N + 1)^2 / 4.$

Dans le **problème de l'indépendance**, l'hypothèse d'**indépendance** entre ξ et η s'écrit :

(3) $\quad H_0 : P^{(\xi,\eta)} = P^\xi \otimes P^\eta.$

On montre que, quelles que soient les lois P^ξ et P^η, la loi de r_{XY}' est une **loi symétrique** pr à 0, d'où :

$$(4)\quad \begin{aligned} &E_0\, r_{XY}' = 0, && \forall\, j \in 2\,\mathbf{N} + 1,\\ &V_0\, r_{XY}' = (N - 1)^{-1}, && \forall\, N \in \mathbf{N} \setminus N_1\ . \end{aligned}$$

L'**approche paramétrique** usuelle se fonde sur la propriété suivante. Si $P^{(\xi,\ \eta)}$ est une **loi normale** (resp si des hypothèses asymptotiques justifiant la **normalité** sont vérifiées) (cf **propriété asymptotique**, **loi asymptotique**, **normalité asymptotique**), dont le coefficient de corrélation noté $\rho_{\xi\eta}$ est estimé par r_{XY} , alors la **statistique** :

$$(5)\quad t_N = (N - 2)^{1/2}\,(1 - r_{XY}^2)^{-1/2}\ r_{XY}$$

admet pour loi (resp admet pour loi asymptotique) $\mathcal{S}_{N-2}$ (**loi de STUDENT** à N - 2 **degrés de liberté**).

L'**approche non paramétrique** utilise, par analogie, la statistique :

$$(6)\quad t_N' = (N - 2)^{1/2}\ r_{XY}',$$

et utilise la propriété de **convergence légale** suivante :

$$(7)\quad \mathcal{L}\,(t_N') \underset{N \to +\infty}{\to}{}^{H0}\ \mathcal{N}(0,\ 1)$$ (**loi normale réduite**).

Si l'on remplace, dans (5), X et Y resp par leur **rang** R et S, on obtient la statistique :

$$(8)\quad t_N'' = (N - 2)^{1/2}\,(1 - r_{XY}'^2)^{-1/2}\ .\ r_{XY}',$$

qui, sous les conditions précédentes, admet la même loi que t_N .

(iii) L'équivalent « théorique » du coefficient de corrélation des rangs r_{XY}' est le **coefficient de corrélation des grades**, ou **coefficient de corrélation des quantiles**, (cf **grade**), ie :

$$(9)\quad \rho_{\xi\eta}' = 3\ .\ \int_{\mathbf{R}2} [2\ F_\xi\,(x) - 1]\ [2\ F_\eta\,(y) - 1]\ dF_{\xi\eta}\,(x,y)$$

(où **R**2 désigne $\mathbf{R}^2$), ou encore :

$$(9)'\quad \rho_{\xi\eta}' = \{E\,(F_\xi(\xi) - 1/2)^2\}^{-1/2}\ .\ \{E\,(F_\eta(\eta) - 1/2)^2\}^{-1/2}\ .\ E\,\{(F_\xi(\xi) - 1/2)\ (F_\eta(\eta) - 1/2)\},$$

où F_ξ (resp F_η) est la **fonction de répartition marginale** de ξ (resp de η) et $F_{\xi\eta}$ la **fr** du couple (ξ, η).

On montre que : $\lim_N r_{XY}' = \rho_{\xi\eta}'$.

Si les échantillons sont gaussiens, on montre que l'**efficacité relative** (au sens de PITMAN) de r_{XY}' pr à r_{XY} est égale à :

$$(10)\quad e\,(r_{XY}' / r_{XY}) = (3 / \pi)^2.$$