COEFFICIENT DE CORRÉLATION BISÉRIELLE (C5, D2)
(14 / 11 / 2019, © Monfort, Dicostat2005, 2005-2019)

Un **coefficient de corrélation « bisériel »** est un **coefficient de corrélation linéaire** entre deux **variables** dont l'une est (partiellement) inobservable (cf **censure**).

(i) Soit $(\Omega, \mathcal{T}, P)$ un **espace probabilisé** et $(\xi, \eta) : \Omega \mapsto \mathbf{R}^2$ un **couple aléatoire**. On suppose que η est inobservable, mais que l'on observe la **vars** définie par :

(1) $\zeta = \mathbf{1}_{[\eta \geq \alpha]}$,

où α représente un « seuil » (en général inconnu).

On appelle **coefficient de corrélation bisérielle** (théorique) entre ξ et η le coefficient de corrélation linéaire ordinaire $\rho_{\xi\zeta}$ entre les **va** ξ et ζ, ie :

(2) $\beta_{\xi\eta} = \rho_{\xi\zeta} = (V\,\xi)^{-1/2} (V\,\zeta)^{-1/2} . C(\xi, \zeta)$.

(ii) En pratique, θ et $\rho_{\xi\zeta}$ représentent des **paramètres** à estimer. On suppose donc disponible un N-**échantillon** $X = ((X_1 , Z_1),..., (X_N , Z_N))$, **échantillon iid** selon (ξ, ζ).

On appelle **coefficient de corrélation bisérielle** (empirique) entre ξ et η le coefficient de corrélation linéaire entre $X = (X_n)_{n=1,..., N}$ et $Z = (Z_n)_{n=1,..., N}$, ie :

(3) $b_{XY} = r_{XZ} = (||X - P\,X|| . ||Z - P\,Z||)^{-1} . X' P Z$,

où $Z_n = \mathbf{1}[Y_n \geq \theta]$, P est la **matrice de centrage par rapport à la moyenne** empirique, $||x||^2 = \Sigma_{n=1}^N x_n^2$ (carré de la norme euclidienne) et Y symbolise l'échantillon (inobservable) généré par η.

Le coefficient r_{XZ} est un **estimateur** de $\beta_{\xi\eta}$. On utilise aussi la **méthode du maximum de vraisemblance** pour estimer le couple $(\rho_{\xi\zeta} , \alpha)$.

(iii) On appelle **coefficient (de corrélation bisérielle) de H.E. BROGDEN** la **statistique** :

(4) $B_{XZ} = \Sigma_{n=1}^N Z_n (X_n - \bar{X}_N) / \Sigma_{n=1}^N U_n (X_n - \bar{X}_N)$,

avec $\bar{X}_N = e_N' X / e_N' e_N$ (**moyenne empirique** de ξ), et $U = (U_1 ,..., U_N)'$ est défini comme la **suite** constituée des termes :

(5) $U_n = \begin{cases} 1, & \forall\, n \in \{1 ,..., [N\,p]\}, \\ 0, & \forall\, n \in \{[N\,p] + 1 ,..., N\}, \end{cases}$

avec N p = $\Sigma_{n=1}^{N} Z_n$ (ie p est la **proportion** de 1 parmi les N). On suppose, dans ce qui précède, que P ($\eta \geq \alpha$) = P ($\zeta = 1$) ne prend pas ses valeurs dans {0, 1} et que 0 $<$ N p $<$ N (strictement).