DISTANCE DE MAHALANOBIS (A4, C, G, H, I, K7)
(06 / 05 / 2020, © Monfort, Dicostat2005, 2005-2020)

La **distance « généralisée » de MAHALANOBIS** est une **métrique** qui intègre les relations (**covariances**) existant entre des **variables**.

(i) Soit $(\Omega, \mathcal{T}, P)$ un **espace probabilisé** et $\mathcal{L}_{\mathbf{R}(K)}^2 (\Omega,\mathcal{T},P)$ l'espace des **vecteurs aléatoires** à valeurs dans $\mathbf{R}^K$ (aussi noté par commodité $\mathbf{R}(K)$) et de carré intégrable. On se restreint au sous-ensemble des **vecteurs aléatoires homogènes**, au sens où :

(1) $V\,\xi = \Sigma, \quad \forall\, \xi \in \mathcal{L}_{\mathbf{R}K}^2 (\Omega, \mathcal{T}, P)$ (même **matrice de dispersion**).

La **norme de P.C. MAHALANOBIS** $||.||_M$ est définie sur $\mathcal{L}_{\mathbf{R}K}^2 (\Omega, \mathcal{T}, P)$ à partir de la **forme quadratique** (aléatoire) :

(2) $||\xi||_M^2 = \xi' \Sigma^{-1} \xi.$

La **distance de P.C. MAHALANOBIS** (théorique) est alors la distance associée à cette **norme** :

(3) $\Delta^2 = \delta_M^2 (\xi, \eta) = ||\xi - \eta||_M^2 = (\xi - \eta)' \Sigma^{-1} (\xi - \eta).$

(ii) L'équivalent empirique se définit à partir d'un **échantillon aléatoire** $X = (X_1, ..., X_N)$, **échantillon iid** selon la loi P^ξ de la **variable parente** ξ. On note X la (N,K)-**matrice** $\{X_1 ... X_N\}$ dont la n-ième ligne X_n correspond à la n-ième observation de ξ et $\bar{X}_N = e_N' X / e_N' e_N$ la **moyenne empirique** vectorielle.

On appelle **distance de P.C. MAHALANOBIS** (empirique) entre $\bar{X}_N$ et $E\,\xi$ la va (ou **statistique**) :

(4) $D^2 \text{ ou } \Delta^2_N = \delta^2 (\bar{X}_N , E\,\xi) = (\bar{X}_N - E\,\xi)' \Sigma^{-1} (\bar{X}_N - E\,\xi).$

(iii) En pratique, pour effectuer le **test** d'une **hypothèse statistique** (eg H_0 : égalité entre des moyennes), Σ doit être estimée. Si S_N est la **matrice de covariance** empirique de X, ie $S_N = X' P X / e_N' e_N$ (où P est la **matrice de centrage par rapport à la moyenne** empirique), un **estimateur sans biais** de Σ est :

(5) $S_N^{\wedge} = (N - 1)^{-1} . N . S_N .$

On remplace alors Δ^2_N de (4) par :

(6) $(\Delta_N^2)^{\wedge} = (\bar{X}_N - E\,\xi)' (S_N^{\wedge})^{-1} (\bar{X}_N - E\,\xi),$

expression qui est aussi notée D^2 ou D_N^2.

Par suite, dans le cas où X est un **échantillon iid** de loi parente $P^\xi = \mathcal{N}(\mu, \Sigma)$ (**loi normale multidimensionnelle**), avec $\mu = E\,\xi$, on montre que :

$N\ (\Delta_N^2)^\wedge \sim \mathcal{K}_{K,N}$ (**loi de HOTELLING** à K et N **degré de liberté**),

(7)

$K^{-1}\ (N - 1)^{-1}\ N\ (N - K)\ (\Delta_N^2)^\wedge \sim \mathcal{T}_{K,N-K}$ (**loi de FISHER-SNEDECOR** à K et N-K degrés de liberté),

ce qui permet de définir des **régions critiques** ainsi que des tests d'égalité (eg $E\ \xi = \mu_0$ donné).

(iv) A titre d'exemple, si l'on dispose de deux **échantillons aléatoires** iid X^1 et X^2, indépendants entre eux, on pose :

(8) $(\Delta_{N(1)N(2)}^2)^\wedge = (\ \bar{X}_{N1} - \bar{X}_{N2}\ )'\ S_{N1N2}^{-1}\ (\ \bar{X}_{N1} - \bar{X}_{N2}\ )$,

où $S_{N(1)N(2)}$ représente la (matrice de) **dispersion** empirique d'ensemble :

(9) $S_{N(1)N(2)} = (N_1 + N_2 - 2)^{-1}\ (N_1\ S_{N1} + N_2\ S_{N2})$,

expression dans laquelle $S_{N(i)}$ est la **matrice des covariances** empirique de X^i et N_i (notée N(i)) la taille de X^i ($\forall\ i \in N_2^*$).

Sous la même hypothèse de **normalité** précédente, on montre que :

(10) $(N_1 + N_2)^{-1}\ N_1\ N_2\ .\ (\Delta_{N1N2}^2)^\wedge \sim \mathcal{K}(K,\ N_1 + N_2 - 2)$

(**loi de HOTELLING** à K et $N_1 + N_2 - 2$ degrés de liberté), ce qui fonde des tests d'égalité (eg d'**homogénéité** des moyennes des deux **populations**, $H_0 : E\ \xi_1 = E\ \xi_2$).

Lorsque la taille des échantillons est grande, l'hypothèse de **normalité** précédente peut être remplacée par des hypothèses qui fondent le **théorème de la limite centrale**. La conduite des tests est analogue.

La distance de MAHALANOBIS permet ainsi :

(a) de comparer les moyennes de deux populations ;

(b) ou de discriminer entre deux populations de même dispersion, à travers leurs **lois de probabilité** (cf **analyse discriminante**, **fonction discriminante de FISHER**).