MÉTHODE DE LAMBERT-WILLIAMS (K9)
(16 / 12 / 2019, © Monfort, Dicostat2005, 2005-2019)

La **méthode de LAMBERT - WILLIAMS** est une méthode de **classification automatique** descendante, qui procède par **dichotomies successives** de l'ensemble d'**unités** à analyser (**population**, **échantillon**).

(i) Soit $X \in M_{NK}$ ({0, 1}) une **matrice** constituée de N **observations** (en lignes) relatives à H **variables qualitatives** η_h (en colonnes), la h-ième variable η_h possédant les modalités {1,..., M_h}, $\forall$ h $\in N_H^*$, avec $\Sigma_{h=1}^{H} M_h = K$ (nombre total de colonnes).

Autrement dit, X est un (N,K)-**tableau statistique** mis sous **forme disjonctive complète**, son élément général x_{nk} étant défini selon :

(1) x_{nk} = 1 si l'observation d'**indice** n de la variable n° h possède la modalité m_h ,
0 sinon,

avec $k = M_1 + ... + M_{h-1} + m_h$.

La matrice X peut être représentée par **blocs** disposés en lignes, le h-ième bloc X (h) étant relatif à la variable η_h :

(2) $X = [X(1) /// X(H)]$, avec $X(h) \in M_{N,M(h)}$ ({0,1}),

où M(h) désigne M_h et /// des changements de lignes.

Par suite, $\forall$ (g, h), la matrice C (g, h) = X (g)' X (h) est un **tableau de contingence** de dimensions (M_g , M_h) dont le terme général $c_{m(g),m(h)}$ (g, h) est le nombre d'observations n possédant à la fois la modalité m_g de g et la modalité m_h de h. Il en est de même de la matrice d'ensemble X' X, dont les blocs sont les C (g, h), $\forall$ (g, h).

(ii) La **méthode de J.M. LAMBERT - W.T. WILLIAMS** procède par étapes :

(a) calcul, $\forall$ (g, h), des **statistiques du chi-deux** empiriques :

(3) $\mathcal{X}_{gh}^2 = c..(g,h) \cdot \{(\Sigma_{m(g)=1}^{M(g)} \Sigma_{m(h)=1}^{M(h)} R_{gh}^2) - 1\}$,

dans lesquelles $R_{gh}^2 = (C_{g,.} C_{.,h})^{-1} \cdot C_{gh}^2$, avec $C_{gh} = c_{m(g),m(h)}(g,h)$, $C_{g,.} = c_{m(g),.}(g,h)$ et $C_{.,h} = c_{.,m(h)}(g,h)$;

(b) calcul des **statistiques** :

(4) $I(g) = \Sigma_{h \neq g} \mathcal{X}_{gh}^2$, $\forall$ g.

On retient la variable η_0 d'indice h_0 tq : $I(h_0) = \max_g I(g)$ (**distance** maximale) ;

(c) partitionnement en deux classes de l'ensemble des modalités {1,..., $M_{h(0)}$} de la variable η_0 (où h(0) désigne h_0). On peut, eg :

$(c)_1$ soit considérer l'ensemble des **partitions** en deux classes de cet ensemble. Il existe exp_2 ($M_{h(0)}$ - 1) - 1 partitions de ce type (avec exp_2 x = 2^x) ;

$(c)_2$ soit ne considérer que les partitions en deux classes consécutives de cet ensemble, en conservant l'ordre {1,..., $M_{h(0)}$} des modalités de η_0 (cas d'une **variable ordinale** η_0). Il n'existe alors que $M_{h(0)}$ - 1 partitions en deux classes ;

En notant {$\mathcal{T}_{h(0)}'$, $\mathcal{T}_{h(0)}''$} une partition quelconque du type précédent, on choisit celle pour laquelle I (h_0) est maximum ;

(d) en notant {$(\mathcal{T}_{h(0)}')\tilde{}$, $(\mathcal{T}_{h(0)}'')\tilde{}$} cette dernière partition, on assimile ses deux classes à deux modalités d'une nouvelle variable qualitative (qui remplace alors η_0). On définit alors deux (N,1)-sous-matrices de X (h_0) : dans la première (resp la seconde), on réunit les observations pour lesquelles la variable prend la modalité $(\mathcal{T}_{h(0)}')\tilde{}$ (resp $(\mathcal{T}_{h(0)}'')\tilde{}$) ;

(e) on itère le procédé pour chacun des tableaux considérés. Chaque itération aboutit ainsi à partitionner progressivement l'ensemble E constitué des N observations.