MODÈLE DES CLASSES LATENTES (B4, K12, O24)
(25 / 04 / 2020, © Monfort, Dicostat2005, 2005-2020)

Le **modèle des classes latentes** est un modèle d'analyse de **structures latentes**, dont l'objet est la décomposition d'une **probabilité** (cf aussi **modèle de structures latentes**, **loi multivariée**).

(i) Soit $(\Omega, \mathcal{T}, \mathcal{P})$ un **modèle statistique** de base. On considère, $\forall\ h \in N_H^*$, un « **caractère** », ou une **variable qualitative**, $\kappa_h : \Omega \mapsto \mathcal{K}_h$ comportant un nombre fini H de modalités.

$\mathcal{K}_h$ désigne ainsi l'**ensemble** fini des valeurs de κ_h , ie :

(1) $\mathcal{K}_h = \{k_{h,1}, ..., k_{h,M(h)}\}, \quad \forall\ h \in N_H^*$.

où l'on note, par commodité, M(h) au lieu de M_h le nombre de modalités de κ_h .

Chaque variable κ_h est un **« attribut polytomique »** observable (cf **observabilité**), appelé **variable manifeste** du modèle (cf **tableau polytomique**).

On définit, $\forall\ P \in \mathcal{P}$, les **probabilités** :

(2) $p_I = P([(\kappa_1, ..., \kappa_H) = (k_{1,m(1)}, ..., k_{k,m(H)})])$,

dans lesquelles :

(a) les **indices** sont notés $I = (m_1, ..., m_H) \in \mathcal{J}$, $\mathcal{J} = \prod_{h=1}^{H} \mathcal{J}_h$, $\mathcal{J}_h = \{1, ..., M_h\}$ (en notant m(h) au lieu de m_h l'indice d'une modalité courante) ;

(b) les valeurs d'un **échantillon** sont notées $(K_1, ..., K_H) : \Omega \mapsto \prod_{h=1}^{H} \mathcal{K}_h$, où $(k_{1,m(1)}, ..., k_{H,m(H)})$ est un H-uple parmi tous les H-uples possibles.

Chaque probabilité p_I est celle d'une valeur correspondant à une « case » $(k_{1,m(1)}, ..., k_{H,m(H)})$ (cf **tableau de contingence**).

(ii) On suppose qu'il existe un **caractère qualitatif** à L modalités, appelées **classes latentes**, ou **classes sous-jacentes**, $\mathcal{Z} = \{z_1, ..., z_L\}$ définissant une va qualitative $\zeta : \Omega \mapsto \mathcal{Z}$, appelée **variable latente**.

Le **modèle de(s) classes latentes** (de P.F. LAZARSFELD) est défini comme un modèle tq le précédent, dans lequel, $\forall\ P \in \mathcal{P}$, on suppose que les p_I admettent une décomposition de la forme :

(3) $p_I = \sum_{\lambda=1}^{L} p_I(\lambda)$,

avec :

(4) $p_I(\lambda) = q_\lambda . \{\Pi_{h=1}^{H} r_{m(h),\lambda}\}, \quad \forall \lambda \in N_L^*,$

expression dans laquelle :

(5)
$$q_\lambda = P(\zeta = z_\lambda), \quad \forall \lambda \in N_L^*,$$
$$r_{m(h),\lambda} = P(K_h = k_{h,m(h)} / \zeta = z_\lambda), \quad \forall m_h \in \mathcal{J}_h, \forall h \in N_H^*, \forall \lambda \in N_L^*.$$

Autrement dit :

(a) l'équation (3) exprime que les probabilités p_I peuvent se décomposer selon L **catégories** mutuellement exclusives et exhaustives (cf **partition**, **théorème de BAYES**, **théorème des probabilités composées**) ;

(b) l'équation (4) exprime que, dans chaque catégorie $\lambda \in N_L^*$ (donc pour chaque classe latente z_λ), les variables K_h sont indépendantes (cf **indépendance**).

(iii) Les $L^2 . (\Pi_{h=1}^{H} M_h)$ **paramètres** du modèle {(3),(4)} peuvent être estimés de plusieurs façons (eg par la **méthode du maximum de vraisemblance**), à l'aide du N-**échantillon** $K = (K_1, ..., K_H)$ observé pour les variables manifestes κ (les **« observables »**).

Ces paramètres vérifient donc les **contraintes** suivantes :

(6)
$$\Sigma_{\lambda=1}^{L} q_\lambda = 1,$$
$$\Sigma_{m(h)=1}^{M(h)} r_{m(h),\lambda} = 1, \quad \forall (h, \lambda) \in N_H^* \times N_L^*.$$

En pratique, les **informations** qui permettent une telle **estimation** sont souvent disposées (ou disponibles) (cf **disposition**) sous la forme :

(a) d'un **tableau de contingence** multidimensionnel (empirique), noté eg $N = (n_I)_{I \in \mathcal{J}}$, dans lequel n_I désigne le nombre d'observations (ou d'**unités statistiques** : individus, etc) élémentaires ayant la **modalité multiple** $(k_{1,m(1)}, ..., k_{H,m(H)})$;

(b) ou d'un tableau de **fréquences relatives**, défini à partir du précédent :

(7) $f_I = n_I / \Sigma_{J \in \mathcal{J}} n_J , \quad \forall I \in \mathcal{J}.$