VARIABLE CATÉGORIELLE (C1, J2, J3, K1, L, M)
(20 / 03 / 2020, © Monfort, Dicostat2005, 2005-2020)

Une **variable catégorielle** est une **variable qualitative**, en général « multiple » (eg multivariée). L'étymologie provient de la notion ordinaire de « catégorie », et se rapporte notamment à la notion de **classification** ou de typologie (eg nomenclatures, tables, etc).

(i) Soit $(\Omega , \mathcal{T}, P)$ un **espace probabilisé**, $((\mathcal{K}_h , \mathcal{D}_h))_{h = 1,...,H}$ une suite d'**espaces mesurables** auxiliaires, $\mathcal{K} = \Pi_{h=1}^{H} \mathcal{K}_h$, $\mathcal{D} = \otimes_{h=1}^{H} \mathcal{D}_h$, et $\kappa : \Omega \mapsto \mathcal{K}$ une **variable aléatoire**.

On dit que κ est une **variable catégorielle** ssi $\mathcal{K}$ n'est pas un ensemble numérique, ni une puissance d'un tel ensemble. Autrement dit, il n'existe pas d'entier $L \geq 1$ ni de partie $R \subset \mathbf{R}^L$ tq $\mathcal{K} = R$. En particulier, aucun $\mathcal{K}_h$ ne peut être identifié à **R**, $\forall$ h = 1 ,..., H. Aucune opération algébrique n'est donc concevable sur $\mathcal{K}$. Ainsi, κ est une **variable qualitative**.

On appelle alors **catégorie multivariée**, **modalité multivariée**, ou **modalité multiple**, ou encore **modalité multidimensionnelle**, de κ tout H-uple de valeurs $k = (k_1 ,..., k_H) \in \mathcal{K}$ de κ.

(ii) En général, $\mathcal{K}_h$ est (pour tout h = 1,..., H) un ensemble fini non numérique, avec Card $\mathcal{K}_h = M_h$, et l'on pose $\mathcal{K}_h = \{k_{h,1} ,..., k_{h,M(h)}\}$, en notant M(h) pour désigner commodément M_h . On dit que $\mathcal{K}_h$ possède un **nombre fini de modalités** ou de catégories.

On peut rendre quantitative une variable catégorielle tq κ en définissant des **variables indicatrices**, aussi appelées **variables catégorielles**, **variables polytomiques**, ou parfois **variables de comptage** (ou **variables de dénombrement**), $\xi_{m(1)...m(H)}$. On définit ainsi les **variables numériques** suivantes :

$$(1) \quad x_{m(1)...m(H)} = \xi_{m(1)...m(H)} (\omega) = \begin{cases} 1 \text{ ssi } \kappa (\omega) = (k_{m(1)} ,..., k_{m(H)}) \in \mathcal{K} , \\ 0 \text{ sinon,} \end{cases}$$

où ω parcourt Ω, et où l'on note encore, par commodité, m(h) au lieu de m_h (pour tout h = 1 ,..., H).

Si l'on suppose, de plus, que $\Omega = \{1 ,..., M\} = N_M^*$ est fini, le **tableau statistique** à H+1 dimensions :

$$(2) \quad T = (\xi_I (\omega))_{\omega,I} ,$$

dans lequel $\xi_I(\omega) = \xi_{m(1)...m(H)}(\omega)$, $I = (m_1, ..., m_H)$ ou $(m(1), ..., m(H)) \in \mathcal{T} = \Pi_{h=1}^{H} \mathcal{T}_h$ (ensemble des indices) et $m_h \in \mathcal{T}_h = \{1, ..., M_h\}$ (pour tout $h = 1, ..., H$), est défini par l'ensemble des valeurs (ou image) de la fonction $t : \Omega \times \mathcal{K} \mapsto \mathbf{N}$ tq :

(3) $t(\omega, u_1, ..., u_H) = \Sigma_{I \in \mathcal{T}} \xi_I(\omega) u_{m(1)} ... u_{m(H)}$,

fonction qui est H-linéaire pr à ses H derniers arguments (cf **application multilinéaire**).

(iii) On appelle parfois aussi **donnée catégorielle** toute observation portant sur une variable entière naturelle (ie à valeurs dans **N**) et figurant dans un **tableau de contingence** dont les **critères** (ou « variables ») de croisement sont des **variables qualitatives** (**attributs**) : eg un tableau répartissant, par sexe et couleur des cheveux, les personnes d'une ville. Lorsque le nombre de modalités de ces critères (resp de ces variables) est supérieur ou égal à 2, on parle de **critères** (resp **variables**) **polytomiques**, et l'on parle de **tableau de contingence multivarié**, ou de **tableau de contingence multidimensionnel**, lorsque $H \geq 2$ (cf **tableau statistique multidimensionnel**).

(iv) Comme toute variable qualitative, une variable catégorielle κ peut être de l'un des deux types suivants :

(a) soit une **variable nominale**, ie non ordinale (cas général a priori) ;

(b) soit une **variable ordinale**, ou **variable « ordonnée »**, κ, ce qui suppose les ensembles $\mathcal{K}_h$ dotés d'une relation de préordre ou d'ordre $\leq$ permettant de comparer les modalités : eg $\mathcal{K}_h$ = {« oui », « abstention », « non »}, ou $\mathcal{K}_h$ = {« beaucoup », « moyennement », « peu »}.

Une variable nominale peut cependant parfois être considérée comme ordinale lorsque le **statisticien** exprime des préférences entre ses modalités, à l'aide d'une **fonction d'utilité** (cf **variable qualitative**).

(v) Le concept de variable catégorielle s'associe ainsi naturellement à celui de tableau de contingence multivarié ou multidimensionnel.

Il s'associe aussi au **modèle de régression** multi-indicé (cf **indice**) à variable endogène catégorielle : **modèle Logit**, **modèle Probit**, etc.

La théorie du **codage** permet de généraliser ces concepts (cf **codage d'un modèle statistique**).