

VARIABLE MULTIDIMENSIONNELLE (C1, C6, F, G11, H7, I9, J, K, L)

(27 / 04 / 2020, © Monfort, Dicostat2005, 2005-2020)

L'**observation** statistique d'un **phénomène** concerne généralement des **unités statistiques** sur chacune desquelles on observe plusieurs **variables**, ce qui est plus « informatif » qu'une variable « isolée ».

(i) En **Statistique**, une **variable aléatoire** ou une **statistique** peut prendre ses valeurs dans un **espace** (**espace mesurable**) qui est :

(a) soit une **suite**, ou « **liste** », non structurée d'**ensembles**, en nombre généralement fini. Ces ensembles peuvent être soit amorphes, soit dotés d'une **structure** particulière. On qualifie cette liste de « **liste de variables simples** » ;

(b) soit dans un **produit** cartésien d'ensembles (cf **produit d'espaces mesurables**), dont le nombre est en général fini (cf cependant **processus stochastique**). La variable considérée est alors un « **élément d'un produit** » ;

(c) soit dans un **espace vectoriel** (mesurable) de dimension généralement finie) (cf **espace vectoriel topologique**, **espace vectoriel mesurable**). La variable est alors un « **vecteur** » (cf **vecteur aléatoire**).

Une **variable multivariée**, ou « **variable multiple** », ou encore **variable multidimensionnelle** est ainsi composée de plusieurs « variables simples ».

Il en va de même pour une statistique : on parle ainsi de **statistique multidimensionnelle** ou encore de **statistique multivariée**.

(ii) La terminologie courante ne distingue pas toujours entre variable multidimensionnelle, variable multiple ou variable multivariée.

La notion de variable multivariée (cf **loi multivariée**) recouvre à la fois la notion de multitude, et celle de différenciation : eg une variable multivariée peut comporter des variables univariées qualitatives aussi bien que numériques.

Une **représentation statistique** comportant des variables ou des statistiques de cette nature est alors elle-même appelée **modèle multidimensionnel**, ou **modèle multivarié**.

(iii) Dans ce dictionnaire, on distingue les notions de **dimensionnalité** et de **multivariation** :

(a) une variable multidimensionnelle ou un modèle multidimensionnel seront associés à une structure d'**ev** de dimension donnée (souvent finie), et défini sur un corps **K**, le plus souvent « numérique » (eg **K = Q**, **K = R** ou **K = C**) ;

(b) une variable multivariée ou un modèle multivarié seront, par contre, associés à une structure, plus générale, de produit cartésien, combinant des **variables quantitatives** (variables ou vecteurs numériques) et des **variables**

qualitatives, simples ou multiples. La prise en compte de la notion de **variable morphologique** est aussi réalisable. Autrement dit, les opérations algébriques usuelles ne sont définies que pour certaines variables.

C'est pourquoi on distingue aussi entre **loi multidimensionnelle** et **loi multivariée**, la seconde notion admettant la première comme cas particulier. Autrement dit, la notion de multivariation englobe celle de dimensionalité.

(iv) Dans l'**espace des variables**, un modèle multivarié est le **modèle paramétrique (modèle image)** $(\mathcal{X}, \mathcal{B}, (P_{\theta}^{\xi})_{\theta \in \Theta})$ dans lequel $\mathcal{X} = \prod_{k=1}^K \mathcal{X}_k$, $\mathcal{B} = \bigotimes_{k=1}^K \mathcal{B}_k$ et $\xi : \Omega \mapsto \mathcal{X}$ est une **variable aléatoire** multivariée, ou vecteur aléatoire, ie $\xi = (\xi_1, \dots, \xi_K)$.

Chaque coordonnée $\xi_k : \Omega \mapsto \mathcal{X}_k$ ($k \in N_K^*$) de ξ est une variable observable sur chaque élément ω d'un ensemble Ω donné (**population**).

En particulier, lorsque $(\mathcal{X}, \mathcal{B}, (P_{\theta}^{\xi})_{\theta \in \Theta}) = (\mathbf{R}^K, \mathcal{B}(\mathbf{R}^K), (P_{\theta}^{\xi})_{\theta \in \Theta})$, ξ est à valeurs dans l'espace vectoriel réel $\mathbf{R}^K$.

(v) Dans l'**espace des observations** (supposées en nombre fini), chaque variable $\xi_k : \Omega \mapsto \mathcal{X}_k$ du type précédent est observée sur N unités. Un **échantillon aléatoire** $X = (X_1, \dots, X_N)$ est alors une va multiple structurée comme un tableau : ie $X_n : \Omega \mapsto \mathcal{X}$ est la valeur de $\xi = (\xi_1, \dots, \xi_K)$ observée sur l'unité $n \in N_N^*$. Par suite, l'**élément aléatoire** X prend ses valeurs dans l'ensemble produit $\mathcal{X}^N = (\prod_{k=1}^K \mathcal{X}_k)^N$, avec pour tribu la puissance tensorielle $\mathcal{B}^{\otimes N} = (\bigotimes_{k=1}^K \mathcal{B}_k)^{\otimes N}$.

La variable générique s'écrit donc X_{kn} , avec $(n, k) \in \{1, \dots, N\} \times \{1, \dots, K\}$.

Une **situation** élémentaire est celle où X est constitué de **copies** d'une variable multiple (**variable parente**) ξ qui sont indépendantes entre elles (cf **échantillon iid**), ce qui se représente généralement sous la forme d'une matrice aléatoire $X = (X_{nk})_{(n,k)}$, où $(n, k) \in N_N^* \times N_K^*$.

(vi) Un modèle multivarié s'obtient souvent par généralisation d'un modèle univarié ($K = 1$ dans l'exemple précédent). Il peut aussi résulter d'une transformation de la représentation initiale (cf **changement de variable, transformation des données**).

Ainsi, le **modèle de régression linéaire** multiple classique (comportant 1 variable endogène et K variables exogènes) se généralise en un modèle multivarié (avec G variables endogènes et K variables exogènes) (cf **régressions multiples**).

Ce procédé se retrouve couramment eg en **analyse de la variance** (analyse de la variance multiple) ou en **analyse des données**.

(ii) Un exemple de modèle multidimensionnel (ou à K dimensions) est le **modèle paramétrique** image (cf **modèle image**) $(\mathbf{R}^K, \mathcal{B}(\mathbf{R}^K), (P_{\theta}^{\xi})_{\theta \in \Theta})$ dans lequel $\Theta \subset \mathbf{R}^Q$. En pratique, chaque coordonnée ξ_k ($k = 1, \dots, K$) de la **variable aléatoire** multidimensionnelle $\xi: \Omega \mapsto \mathbf{R}^K$ (ou **vecteur aléatoire**) constitue l'une des K **mesures** effectuées sur les éléments ω d'un ensemble Ω donné (**population**).

Un **échantillon aléatoire** $X = (X_1, \dots, X_N)$, constitué de copies de ξ indépendantes entre elles et équidistribuées (cf **échantillon iid**), se représente généralement sous la forme d'une matrice $X^* = (x_{nk})_{(n,k)}$ où $(n, k) \in \{1, \dots, N\} \times \{1, \dots, K\}$. Cet échantillon engendre le **modèle produit** $(\mathbf{R}^K, \mathcal{B}(\mathbf{R}^K), (P_{\theta}^{\xi})_{\theta \in \Theta})^{\otimes N}$.

(iii) Un modèle multidimensionnel est souvent la généralisation d'un modèle unidimensionnel (tq $K = 1$ dans l'exemple précédent). Ainsi, le **modèle de régression linéaire** multiple classique (une variable endogène et K variables exogènes, toutes numériques) se généralise ainsi en un modèle multidimensionnel (G variables endogènes et K variables exogènes) (cf **régressions multiples**).