ESPACE DES VARIABLES (A5, B1)
(06 / 05 / 2020, © Monfort, Dicostat2005, 2005-2020)

Un **espace de variables** décrit les « valeurs » prises par une (ou plusieurs) variable(s) d'intérêt. C'est un **espace mesurable** particulier, dont le rôle est fondamental en **science**, et notamment en **Statistique**.

(i) Soit $(\Omega, \mathcal{T})$ un **espace probabilisable fondamental**, dans lequel :

(a) Ω est un **ensemble fondamental** (eg une **population**) dont les éléments (ou événements élémentaires) sont des **unités statistiques** (ou individus) ;

(b) $\mathcal{T}$ est une **tribu de parties**, ou **tribu d'événements**. Ces parties représentent, pour le **statisticien**, des réalisations ou des « résultats », etc, qu'il considère, de façon générale, comme « **aléatoires** ».

L'espace $(\Omega, \mathcal{T})$ peut notamment permettre de décrire une **expérience aléatoire**

Soit $\mathcal{X}$ un ensemble dont les éléments x sont les valeurs x prises par une application $\xi : \Omega \mapsto \mathcal{X}$ et sont ainsi observées sur les individus $\omega \in \Omega$. Cette application ξ représente une **variable d'intérêt** (eg un **attribut**) et $\mathcal{X}$ est appelé **ensemble des valeurs** ou **ensemble de valeurs** de la variable ξ.

Des « **observations** » (**mesures**, etc) $x = \xi(\omega)$ sont effectuées sur les éléments ω. Si l'application $\xi : \Omega \mapsto \mathcal{X}$ est mesurable (ie $(\mathcal{T},\mathcal{B})$-mesurable), elle définit une **variable aléatoire** qui transporte sur $\mathcal{X}$ la tribu $\mathcal{T}$, ce qui conduit à la notion de **tribu image** $\mathcal{B}$ de $\mathcal{T}$ par ξ.

L'**espace probabilisable** $(\mathcal{X}, \mathcal{B})$ ainsi défini est appelé **espace de la variable** ξ.

(ii) La variable ξ précédente peut être :

(a) une variable « simple » : eg variable « unique », ou variable « scalaire » ;

(b) une variable « complexe » : eg variable « multiple », ou variable « vectorielle » (**vecteur aléatoire**) ;

(c) une variable complexe « mixte », combinant des **variables endogènes** et des **variables exogènes** ;

(d) une variable complexe « mixte », combinant des **variables qualitatives** et des **variables quantitatives**, etc.

Le plus souvent, on considère que $\mathcal{X}$ est un ensemble de la forme $\mathcal{X} = \Pi_{k=1}^{K} \mathcal{X}_k$, ie un produit (cartésien) d'ensembles $\mathcal{X}_k$ ou encore un espace vectoriel à K dimensions. On définit alors des variables associées $\xi_k : \Omega \mapsto \mathcal{X}_k$ (k = 1 ,..., K) et l'on note $\xi = (\xi_1 ,..., \xi_K)$ la suite (ou « liste ») ou le vecteur des variables considérées.

L'ensemble $\mathcal{X}$ est alors appelé **espace des variables** ou **espace de variables**.

On suppose qu'il existe un espace (éventuellement de type produit) $(\mathcal{Y}, \mathcal{C})$ et une va $(\mathcal{B}, \mathcal{C})$-mesurable $\eta : \Omega \mapsto \mathcal{Y}$. Par suite (cf **relation fonctionnelle**) :

(a) s'il existe une relation $f : \mathcal{X} \mapsto \mathcal{Y}$ tq (forme explicite) :

(1) $f (\xi_1 ,..., \xi_K) = \eta$

entre ξ et η (image des ξ_k par f), alors la représentation de f dans l'espace $\mathcal{X}$ x $\mathcal{Y}$ est appelée **représentation dans l'espace des variables** ;

(b) s'il existe une relation $g : \mathcal{X}$ x $\mathcal{Y} \mapsto \mathcal{Y}$ tq (forme implicite) :

(1)' $g (\xi_1 ,..., \xi_K , \eta) = e$

entre ξ et η, dans laquelle $e \in \mathcal{Y}$ est un élément spécifique de $\mathcal{Y}$, alors la représentation de g dans l'espace $\mathcal{X}$ x $\mathcal{Y}$ x $\mathcal{Y}$ est aussi appelée **représentation dans l'espace des variables**.

(iii) Il importe ainsi de distinguer entre espace des variables et **espace des observations** :

(a) l'espace des variables décrit des relations entre « concepts » (variables) ;

(b) l'espace d'observation décrit la structure existant entre les « informations » observable portant sur ces concepts, et généralement appelées « **données** » : ce sont les observations des variables effectuées sur des unités.

Ce **contexte statistique** est très général, et se retrouve notamment dans le contexte d'une **relation fonctionnelle**, d'une **régression** (cf **modèle de régression**) ou d'une **interdépendance** (cf **modèle d'interdépendance**).

Ainsi, la relation entre variables de l'exemple (1) peut se représenter dans l'espace des variables $((\xi_1 ,..., \xi_K) , \eta)$: on l'appelle « relation théorique » (ou conceptuelle) entre variables (cf **loi**, **loi scientifique**).

En pratique, on observe un **échantillon** des variables précédentes, relatif à diverses unités 1 ,..., N. Si $X = (x_1 ,..., x_K)$ représente une suite constituée de N observations

x_{nk} (n = ,..., N) de chacune des variables ξ_k (ie x_k prend ses valeurs dans $\mathcal{X}_k^{\otimes N}$), et si les N observations y_n (n = ,..., N) de η prennent leurs valeurs dans $\mathcal{Y}^{\otimes N}$, la réécriture de (1) pour chacune des observations (x_{nk} , y_n), aussi bien que pour l'ensemble des ces équations, s'appelle « **relation observée** » ou « **forme observée** », ie :

(2) $f(x_{n1}, ..., x_{nK}) = y_n , \quad \forall\, n = 1, ..., N,$

Il en va de même de la relation implicite :

(2)' $g(x_{n1}, ..., x_{nK}, y_{n1}, ..., y_{nG}) = e,$

dans laquelle y_g prend ses valeurs dans $\mathcal{Y}_g^{\otimes N}$ (g = 1 ,..., G) et e est un élément spécifique de $\mathcal{Y} = \Pi_{g=1}^{G} \mathcal{Y}_g$.

Les relations f ou g peuvent ne pas être connues, ni être observables, ni être estimées.

(iv) Soit P une **mesure de probabilité** définie sur $\mathcal{T}$. On peut associer à l'**espace probabilisé** (fondamental) défini par le triplet (Ω, $\mathcal{T}$, P) un « espace probabilisé image » et, plus généralement, une **représentation statistique** (Ω, $\mathcal{T}$, $\mathcal{P}$) dans laquelle $\mathcal{P}$ est une famille de **mesures de probabilité** tq P. En effet, la probabilité image de P par ξ est, par définition, la **loi de probabilité** P^{ξ} de la **va** ξ (cf **mesure image**). L'espace probabilisé ($\mathcal{X}$, $\mathcal{B}$, P^{ξ}) obtenu est alors l'**espace des variables image** de (Ω, $\mathcal{T}$, P) par ξ.

C'est dans cet espace que s'effectue généralement l'analyse statistique « conceptuelle », le modèle correspondant s'écrivant ($\mathcal{X}$, $\mathcal{B}$, $\mathcal{P}^{\xi}$), où $\mathcal{P}^{\xi}$ désigne une **famille de lois**.

Les développements et terminologies précédentes se transposent à ce type d'espaces.