MODÈLE D'ÉCHANTILLONNAGE (F1, G2)
(25 / 03 / 2020, © Monfort, Dicostat2005, 2005-2020)

(i) Soit $(\Omega, \mathcal{T}, \mathcal{P})$ un **modèle statistique** de base (ou modèle fondamental) $(\mathcal{X}, \mathcal{B})$ un **espace d'observation** et $X : \Omega \mapsto \mathcal{X}$ une **statistique** donnée (eg un **échantillon**). On note $(\mathcal{X}, \mathcal{B}, \mathcal{P}^X)$ le **modèle image** par X du modèle initial.

On dit que $(\mathcal{X}, \mathcal{B}, \mathcal{P}^X)$ est un **modèle d'échantillonnage** (ou, parfois, un **modèle de sondage**) ssi il peut s'écrire sous la forme d'un modèle puissance $(\mathcal{X}_0, \mathcal{B}_0, \mathcal{P}^\xi)^{\otimes N}$, dans lequel le nombre entier $N \in \mathbf{N}^*$ est appelé **taille de l'échantillon** X (comparer avec **modèle de sondage**).

Autrement dit (cf **mesure produit**, **modèle produit**, **produit**, **produit d'espaces mesurables**, **produit d'espaces mesurés**) :

(a) $\mathcal{X}$ est le produit cartésien de N **copies**, ou **répliques**, de $\mathcal{X}_0$ (ie la puissance cartésienne de degré N de $\mathcal{X}_0$) : $\mathcal{X} = \Pi_{n=1}^N \mathcal{X}_0 = \mathcal{X}_0^N$;

(b) $\mathcal{B}$ est le **produit tensoriel** de N copies de $\mathcal{B}_0$ (ie puissance tensorielle de degré N de $\mathcal{B}_0$) : $\mathcal{B} = \otimes_{n=1}^N \mathcal{B}_0 = \mathcal{B}_0^{\otimes N}$;

(c) $\mathcal{P}^X$ est le produit tensoriel de N copies de $\mathcal{P}^\xi$ (ie puissance tensorielle de degré N de $\mathcal{P}^\xi$) : $\mathcal{P}^X = \otimes_{n=1}^N \mathcal{P}^\xi = (\mathcal{P}^\xi)^{\otimes N}$, expression dans laquelle $(\mathcal{P}^\xi)^{\otimes N}$ représente symboliquement l'ensemble des produits tensoriels de N copies de la loi $P^\xi \in \mathcal{P}^\xi$. Ainsi :

(1) $\mathcal{P}^X = \{(\mathcal{P}^\xi)^{\otimes N} : P^\xi \in \mathcal{P}^\xi\}$.

Il existe donc une **va** $\xi : \Omega \mapsto \mathcal{X}_0$ dont la **lp** P^ξ appartient à $\mathcal{P}^\xi$, et le **modèle d'échantillonnage** ainsi défini décrit une **expérience aléatoire** constituée de N **expériences « élémentaires »**, identiques et indépendantes, $(\mathcal{X}_0, \mathcal{B}_0, P^\xi)$.

La va ξ est parfois appelée **variable parente**, ou **variable mère**, ou **variable source**, ou **variable génératrice**, ou encore **variable générique**, de l'échantillon X (cf **échantillon équidistribué**, **échantillon indépendant**, **échantillon iid**).

Comme $\mathcal{X}$ est une puissance cartésienne, on peut écrire $X = (X_1, ..., X_N)$ (cf **échantillon**). Ses coordonnées $X_n = pr_n X$ sont des copies, ou des répliques, indépendantes et équidistribuées (iid), de ξ (cf **suite iid**).

(ii) Le modèle précédent, très courant en pratique, est associé à un **problème d'échantillonnage à « distance finie »**, ou **problème statistique à « distance finie »** : le terme de « **distance** » est un abus de language qui désigne l'entier $N \in \mathbf{N}^*$ (cf aussi **échantillonnage**, **espace d'échantillonnage**, **loi d'échantillonnage**).

Lorsque N est « petit » (ie $N << +\infty$), on désigne par **problème de petit échantillon** un problème statistique fondé sur un modèle d'échantillonnage à « distance » finie et « petite ».

(iii) Le modèle d'échantillonnage doit généralement être étendu pour en permettre l'étude des **propriétés asymptotiques**, ie l'étude des propriétés statistiques des **procédures** mises en oeuvre lorsque $N \to +\infty$ en parcourant **N**. C'est aussi le cas des problèmes d'**analyse séquentielle**.

Ceci conduit à étendre le modèle puissance « finie » ci-dessus dans un modèle puissance plus vaste, qui en constitue un « **plongement** ».

On dit alors que $(\mathcal{X}, \mathcal{B}, \mathcal{P}^X)$ est un **modèle d'échantillonnage asymptotique** ssi il peut s'écrire $(\mathcal{X}_0, \mathcal{B}_0, \mathcal{P}^\xi)^{\otimes \mathbf{N}}$, ie ssi :

(a) $\mathcal{X} = \mathcal{X}_0^{\mathbf{N}}$ (espace des **suites** $X = (X_n)_{n \in \mathbf{N}^*}$ sur $\mathcal{X}_0$) ;

(b) $\mathcal{B} = \otimes_{n \in \mathbf{N}^*} \mathcal{B}_0 = \mathcal{B}_0^{\otimes \mathbf{N}}$ (puissance tensorielle infinie dénombrable) ;

(c) $\mathcal{P}^X$ est constituée des probabilités P^X définies sur $\mathcal{B}$ dont les **restrictions** P_N^X à toute sous-**tribu** de $\mathcal{B}$ de la forme $\otimes_{n=1}^N \mathcal{B}_0 = \mathcal{B}_0^{\otimes N}$ (où $N \in \mathbf{N}^*$) s'exprime en fonction des **lois** $P^\xi \in \mathcal{P}^\xi$ selon (puissance tensorielle des P^ξ) :

(2) $\quad P_N^X = \otimes_{n=1}^N P^{X(n)} = \otimes_{n=1}^N P^\xi = (P^\xi)^{\otimes N}$,

où X(n) désigne par commodité X_n (tout n), les **conditions de « compatibilité »** étant de même nature que celles du **théorème de KOLMOGOROV** relatif aux **processus stochastiques** (existence de la « **loi jointe** » d'un processus) (cf **système projectif de probabilités**).

(iv) Lorsque le **modèle d'observation** $(\mathcal{X}, \mathcal{B}, \mathcal{P}^X)$ peut se mettre sous l'une des formes précédentes (à distance finie ou infinie), ie sous forme d'un **modèle puissance**, on dit que X est un **échantillon (aléatoire) indépendant identiquement distribué** (iid), ou **indépendant équidistribué**, ou **indépendant homogène**, ou parfois un **échantillon simple**, ou même un **échantillon** (cf **échantillon iid**).

On précise, le cas échéant, que X est un **échantillon** issu de (ou généré par) ξ. A cette notion sont associées, dans le cas de va réelles, deux notions importantes : celle de **loi empirique** et cell de **fonction de répartition empirique**.

(v) Un exemple de modèle d'échantillonnage classique est celui du **modèle d'échantillonnage gaussien**, dans lequel $\mathcal{X}_0 = \mathbf{R}^K$, $\mathcal{B}_0 = \mathcal{B}(\mathbf{R}^K)$ (**tribu des boréliens**) et $P^\xi = \mathcal{N}_K(\mu, \Sigma)$. Alors, P^X admet la **densité de probabilité** suivante pr à la **mesure de LEBESGUE** puissance $\lambda_K^{\otimes N}$:

(3) $f(x_1, ..., x_N) = (2\pi)^{-N K / 2} . |\Sigma|^{-N/2} . \exp\{- (1/2) \Sigma_{n=1}^N (x_n - \mu)' \Sigma^{-1} (x_n - \mu)\}$.

(vi) Lorsque le modèle d'observation précédent ne peut pas se mettre sous la forme d'une puissance, on parle d'**échantillon non indépendant**, ou d'**échantillon complexe**. Un **processus de MARKOV** ou une **martingale** peuvent être considérés comme tels.

Une première extension de la notion d'échantillon iid est ainsi celle d'**échantillon indépendant non identiquement distribué**, ou d'**échantillon indépendant non homogène** : le modèle considéré s'écrit alors sous la forme d'un produit $(\Pi_{n=1}^N \mathcal{X}_n , \otimes_{n=1}^N \mathcal{B}_n , \otimes_{n=1}^N P^{X(n)})$, en remplaçant $N_N^* = \{1, ..., N\}$ par **N** dans le cas non fini (ou asymptotique). Ce schéma correspond à la situation dans laquelle les expériences se réalisent indépendamment entre elles, mais où les « conditions » de l'observation (**expérimentation**, etc) varient avec N.

(vii) Enfin, selon la nature du problème concret, des données mises en oeuvre (**variables** ou **observations**) ou selon la **structure** imposée à l'échantillon, ce dernier peut être noté (ou se présenter) sous diverses formes, dont les plus courantes sont les suivantes :

(a) **forme « ensembliste »** $X = \{X_1, ..., X_N\} \subset \mathcal{X}_0$ (même **ensemble** de valeurs), dans laquelle aucune structure particulière ni aucun ordre ne sont imposés (**échantillon « amorphe »**). Aucune « répétition » des observations n'est a priori possible ici, sauf si deux **unités statistiques** α et β sont à l'origine d'une même valeur, ie être tq $X_\beta = X_\alpha$ (eg cas de **variables discrètes**). Les **observations** peuvent être quantitatives ou qualitatives ;

(b) **forme « produit »** $X = (X_1, ..., X_N) \in \mathcal{X}_0^N$ ou $\Pi_{n=1}^N \mathcal{X}_n$ (selon le cas), dans laquelle chaque « coordonnée » $X_n \in \mathcal{X}_n$ (espace facteur). Le produit cartésien $\Pi_{n=1}^N \mathcal{X}_n$ des espaces facteurs constitue l'**espace d'observation** (cf aussi **espace d'échantillonnage**) produit $\mathcal{X}$. Ici, l'**ordre** des coordonnées intervient. Les espaces facteurs peuvent être identiques ou non (puissances cartésiennes ou produits cartésiens). Ce formalisme correspond souvent à un modèle d'échantillonnage. Ici, des **coordonnées multiples** X_n sont possibles. Les observations peuvent être quantitatives ou qualitatives ;

(c) **forme « vectorielle »** ou **forme « matricielle »** $X = (X_1, ..., X_N)'$ (vecteur colonne), ou encore $X = (X_1, ..., X_N)$ (vecteur ligne), dans laquelle les coordonnées X_n du **vecteur aléatoire** X sont supposées à valeurs dans un **espace vectoriel** (réel), eg dans $\mathbf{R}^K$.

Lorsque X prend ses valeurs dans $\mathbf{R}^N$ (N-échantillon réel scalaire), on l'assimile souvent à un vecteur colonne. On peut alors effectuer des opérations algébriques de façon simple (concepts « empiriques ») : eg le total des coordonnées de X s'écrit $T = e_N' X$, sa **moyenne empirique** $\bar{X}_N = N^{-1} T$, sa **variance** $S_N^2 = X' P X / e_N' e_N$, où P désigne la **matrice de centrage par rapport à la moyenne**.

Lorsque X prend ses valeurs dans M_{NK} (**R**) (**ev** des **matrices** réelles d'ordre (N, K)), on considère qu'il s'agit d'une **matrice aléatoire** à N lignes, notée $(X_1 /// X_N)$ (les observations des variables), ou à K colonnes, notée $[x_1 ,..., x_K]$ (les vecteurs des variables observées) (cf **matrice d'observation**). On note alors :

(4) $X = (x_{nk})_{(n,k)} = (X_1 /// X_N) = [x_1 ,..., x_K]$.

où /// dénote des sauts de lignes.

Ici encore, on peut commodément effectuer des opérations algébriques simples : eg le vecteur des totaux en ligne ($T = e_N' X$), celui des totaux en colonne ($t = X e_K$), la **matrice de covariance** $S_N^2 = X' P X / e_N' e_N$, etc ;

(d) **forme « non équilibrée »**, notée eg :

(5)
$X_{1,1} ,..., X_{1,N(1)}$,
...
$X_{i,1} ,..., X_{i,N(i)}$,
...
$X_{k,1} ,..., X_{k,N(k)}$,

où les N_i (aussi notés, par commodité, N(i)) ne sont pas nécessairement égaux (cf aussi **bloc aléatoire**, **problème à plusieurs échantillons**). Une telle « **disposition** » provient souvent d'un **plan d'expérience** (avec eg des observations perdues ou détruites), d'**échantillonnages** répétés (dans le temps ou dans l'espace) avec des tailles différentes (**plans de sondage** différents, **non réponses**, etc), d'une **classification** en «**strates** » appelées « classes », etc ;

(e) **forme « multidimensionnelle »**, au sens où X est à valeurs dans un espace tensoriel (cf eg **tableau statistique**).

(viii) Deux **contextes statistiques** peuvent affecter la « qualité » d'un échantillon (cf **modèle de génération de données**, **observabilité**) :

(a) un échantillon X peut ne pas être entièrement **observable** (ou observé) (cf **censure**, **inobservable**) ;

(b) de même, un **échantillon simple** peut être engendrée par une **va** ξ dont la **loi** P^ξ ne « charge » pas nécessairement toutes les régions de l'espace d'observation $\mathcal{X}_0$ (cf **troncature**).