ÉCHANTILLON (F, K, L, M, N)
(31 / 08 / 2020, © Monfort, Dicostat2005, 2005-2020)

Notion statistique fondamentale, un **échantillon** représente un moyen privilégié par le **statisticien** pour l'**observation** d'un **phénomène** donné. Les investigations menées à l'aide d'un échantillon doivent, en principe, permettre de caractériser ce phénomène ou de répondre à des questions relatives à celui-ci :

(a) recherche de la **loi** qui le gouverne (cf **loi multivariée**, **loi multidimensionnelle**) ;

(b) ou seulement étude de certaines de ses **caractéristiques** ayant un intérêt en rapport avec ce phénomène (cf **relation fonctionnelle**, **régression**).

Autant dire qu'un échantillon constitue, aussi bien dans le cadre d'un **sondage** que celui d'une **expérience aléatoire**, un moyen essentiel d'**inférence statistique** (cf aussi **production statistique**). C'est un « pont » entre le champ de l'**observable** et le champ de ce qui est censé l'avoir généré ou produit, lequel est **inobservable**.

(i) Soit Ω un **ensemble** de base (ou **ensemble fondamental**) appelé **population**, ou parfois **expérience**, constitué d'unités élémentaires $\omega \in \Omega$. On distingue entre :

(a) **échantillon d'unités** : il s'agit d'une **partie** quelconque $A \in \mathscr{P}(\Omega)$. C'est le sens le plus couramment attribué au terme échantillon. Lorsque A est finie (avec Card A = N), on la note $A = \{a_1, ..., a_N\}$;

(b) **échantillon d'observations** : il s'agit de l'ensemble $\xi(A)$ des images $x = \xi(\omega)$ dans un ensemble $\mathscr{X}$, appelé **ensemble d'observation**, lorsque ω parcourt A (cf **espace d'observation**). Autrement dit, à chaque **unité statistique** (individu, ou expérience élémentaire) $\omega \in \Omega$ on associe une grandeur $x \in \mathscr{X}$. Lorsque A est finie (avec Card A = N), on note $X = (X_1, ..., X_N)$, avec $X_n = \xi(a_n)$, $\forall\, n = 1, ..., N$.

En particulier, lorsque $\mathscr{X}$ est un ensemble numérique, on appelle $\xi(A) \subset \mathscr{X}$ un **échantillon de mesures**.

Le language courant ne précise pas toujours s'il s'agit d'un échantillon A d'unités ou d'un échantillon $\xi(A)$ d'observations effectuées sur ces unités.

Par ailleurs :

(a) lorsque Card A = N $< +\infty$ (cas le plus courant), on parle d'**échantillon fini**, et l'on note eg $A = (a_n)_{n = 1,..., N}$; il en est de même pour $\xi(A)$, qui est noté eg $(\xi(a_n))_{n = 1,..., N}$. Le nombre entier $N \in \mathbf{N}^*$ est appelé **taille** de l'échantillon A (ou de l'échantillon $\xi(A)$).

(b) lorsque Card A = Card **N*** = $\aleph_0$ (puissance du dénombrable), on parle d'**échantillon dénombrable** pour A = $(\omega_n)_{n \in \mathbf{N}^*}$, et de **suites d'observations** pour ξ (A) = $(\xi(a_n))_{n \in \mathbf{N}^*}$ (cf aussi **processus stochastique**).

La notion précédente est une conception « vulgaire » : elle est parfois qualifiée d'**échantillon non aléatoire**, ou **échantillon déterministe**, ou encore **échantillon sans modèle**. En effet, dans le contexte précédent, on ignore a priori comment est choisi (ou « tiré ») A dans $\mathscr{P}(\Omega)$ (**représentativité indéterminée**).

(ii) Dans les conceptions « probabiliste » et « statistique », on suppose généralement que Ω est muni d'une **structure** mesurable par adjonction d'une **tribu de parties** notée $\mathscr{T}$. Sur cette tribu est définie une **mesure de probabilité** P qui permet de connaître la probabilité (« chance » ou « vraisemblance ») P (A) de choisir (ou de tirer) A (probabilité d'apparition ou de tirage d'un échantillon A donné).

D'autre part, on définit un **espace d'observation** ($\mathscr{X}$, $\mathscr{B}$) dans lequel $\mathscr{B}$ est une tribu de parties de $\mathscr{X}$. Si X : $\Omega \mapsto \mathscr{X}$ est une application ($\mathscr{T}$, $\mathscr{B}$)-mesurable, ie une **variable aléatoire**, on connaît aussi la probabilité P^X (X (A)) d'obtenir l'ensemble (fini) B = X (A) constitué des **observations** de cette va.

En effet, par définition (cf **loi de probabilité**), il existe alors un ensemble B $\in \mathscr{B}$ tq P^X (X (A)) = P^X (B) = P (X^{-1} (B)) = P (A). La tribu $\mathscr{B}$ peut alors être choisie comme image par X de $\mathscr{T}$ (cf **tribu image**).

Ainsi, formellement :

(a) un échantillon (d'unités) est assimilable à une partie A $\in \mathscr{T}$ dont la probabilité d' « apparition » P (A) est connue ;

(b) un échantillon (d' « observations ») est assimilable à une partie B = X (A) dont la probabilité d'apparition P^X (B) est calculable.

Le but d'un **échantillonnage** est le choix d'un échantillon le plus adapté à une **procédure statistique** donnée.

(iii) Lorsqu'on l'associe à une probabilité P donnée, un échantillon (d'observations) X possède donc la nature intrinsèque d'être une **variable aléatoire**, ie X : $\Omega \mapsto \mathscr{X}$: d'où la dénomination d'**échantillon aléatoire**.

Un échantillon peut aussi être considéré comme une **statistique** S = s (X) particulière, lorsqu'on l'associe à une famille $\mathscr{P}$ de probabilités tq P. Cette statistique n'est autre que l'**application identique** dans $\mathscr{X}$: s = $id_{\mathscr{X}}$.

(iv) Il existe un procédé pratique de construction d'échantillons issus d'une **lp** donnée. Si E est l'**expérience aléatoire** qui consiste à choisir (ou à « tirer ») au **hasard** (selon la loi P^X) un élément $X(\omega) \in \mathcal{X}$, alors (cf **lemme d'uniformisation des lois**) cette expérience revient à tirer (selon un hasard uniforme) un élément aléatoire à valeurs dans un espace auxiliaire (cf **uniformisation**, **échantillon artificiel**).

En pratique, un échantillon est de taille (au plus) dénombrable, parfois « continue » (eg ECG ou EEG, en médecine) (cf **trajectoire**). Le plus souvent, il est de taille finie $N \in \mathbf{N}^*$, même si la population de référence (d'où il est « extrait ») Ω est infinie.

Cependant, le cadre général du **modèle d'échantillonnage** infini (au sens dénombrable) est indispensable à l'étude théorique des **propriétés asymptotiques**.

(v) Diverses **situations statistiques** peuvent se relier à l'étude d'un échantillon X :

(a) les populations considérées peuvent être de cardinalité variable : celle du **fini** lorsqu'il s'agit d'une population entendue au sens courant (populations physique, écologique, biologique, etc), celle du **dénombrable** lorsqu'on observe des **suites** d'unités (indexées eg par $\mathbf{N}^*$, $\mathbf{Z}$, $\mathbf{D}^*_+$ ou $\mathbf{Q}^*_+$), celle du **continu** dans le cas des ensembles numériques usuels $\mathbf{R}^n$ ou $\mathbf{C}^n$, celle des **espaces fonctionnels** lorsqu'il s'agit eg des **trajectoires** d'un processus stochastique ;

(b) l'ensemble d'observation $\mathcal{X}$ parcouru par X est souvent un ensemble produit $\mathcal{X} = \Pi_{n=1,...,N}\, \mathcal{X}_n$ et l'on note $\mathcal{B}_n$ la tribu définie sur $\mathcal{X}_n$. On écrit $X = (X_1, ..., X_N) : \Omega \mapsto \Pi_{n=1}^{N}\, \mathcal{X}_n^N$, où $X_n : \Omega \mapsto \mathcal{X}_n$ est l'une des va coordonnées : on peut donc considérer X comme un **vecteur aléatoire**. On note alors $(\Pi_{n=1}^{N}\, \mathcal{X}_n , \otimes_{n=1}^{N}\, \mathcal{B}_n , P^X)$ le **produit d'espaces probabilisés** correspondant, sachant que :

$(b)_1$ dans le cas général, X est un échantillon quelconque (ie non indépendant), et sa loi P^X n'est autre que la **loi conjointe** de ses composantes X_n (n = 1 ,..., N), ie, par définition, soit l'image de la loi de A par X, soit l'image $P^X = X(P)$ de P par X. Les lois coordonnées sont les **lois marginales** $P^{X(n)} = X_n(P)$ (où X(n) désigne X_n), $\forall\, n \in N_n^*$;

$(b)_2$ si X est indépendant (ie si ses composantes sont indépendantes entre elles), sa loi P^X peut s'écrire sous la forme d'un produit tensoriel $\otimes_{n=1}^{N}\, P^{X(n)}$, où, $\forall\, n \in N_N^*$, $P^{X(n)}$ désigne la loi propre (ou **loi marginale**) de X_n et où X(n) dénote par commodité X_n (cf **échantillon indépendant**).

Très souvent, $\mathcal{X}_n = \mathcal{X}_0$ (ensemble fixe) (eg $\mathcal{X}_n = \mathbf{R}$), pour tout $n \in N_N^*$, d'où $\mathcal{X} = \mathcal{X}_0^N$ (puissance cartésienne ordinaire) (eg $\mathcal{X} = \mathbf{R}^N$). Lorsque X est, en outre, un échantillon indépendant, sa loi P^X s'écrit sous la forme d'une puissance tensorielle $\otimes_{n=1}^{N}\, P^{X(n)} = (P^\xi)^{\otimes N}$, où P^ξ désigne la loi d'une **variable parente** « générant » l'échantillon X (cf **échantillon indépendant équidistribué**) ;

(c) dans les études asymptotiques, $\mathcal{X}$ est un produit infini (dénombrable) $\Pi_{n \in \mathbf{N}^*} \mathcal{X}_n$ (ou $\Pi_{n \in \mathbf{N}^*} \mathcal{X}_n$). Le plus souvent, on a $\mathcal{X}_n = \mathcal{X}_0$ (espace fixe) (eg $\mathcal{X}_n = \mathbf{R}$) pour tout $n \in \mathbf{N}^*$ (ou tout $n \in \mathbf{N}$), d'où $\mathcal{X} = \mathcal{X}_0^{\mathbf{N}}$ (ou $\mathcal{X} = \mathcal{X}_0^{\mathbf{N}^*}$) (eg $\mathcal{X} = \mathbf{R}^{\mathbf{N}}$ ou $\mathbf{R}^{\mathbf{N}^*}$). L'**espace d'échantillonnage** s'écrit alors $(\mathcal{X}, \mathcal{B}, P^X)$, avec eg $\mathcal{X} = \Pi_{n \in \mathbf{N}^*} \mathcal{X}_n$ et $\mathcal{B} = \otimes_{n \in \mathbf{N}^*} \mathcal{B}_n$. L'**échantillon** associé à cet espace est alors représentable par une suite de va notée $X = (X_n)_{n \in \mathbf{N}^*}$.

Des remarques semblables aux précédentes (échantillons finis) se transposent à cette situation ;

(d) la **théorie des processus** permet souvent d'étendre l'étude des échantillons (ou de l'échantillonnage) à des situations plus générales (**espace des temps** quelconque).

(vi) On peut considérer un **échantillon (d'observations) de taille N**, ou **N-échantillon** sous trois angles parfois complémentaires :

(a) soit comme une **partie** d'un **ensemble** (d'observations) :

(3) $X = \{X_1, ..., X_N\} \subset \mathcal{X}_0$,

où $\mathcal{X}_0$ est l'ensemble dans lequel les X_n prennent tous leurs valeurs. Plus précisément, $X \in \mathcal{M}(\Omega, \mathcal{X})$, ensemble des **applications mesurables** (ou **va**) de Ω dans $\mathcal{X} = \mathcal{X}_0^N$;

(b) soit comme un **N-uple** dans un espace produit $\Pi_{n=1,...,N} \mathcal{X}_n$, ie :

(1) $X = (X_1, ..., X_N) : \Omega \mapsto \mathcal{X}$ (eg pour l'étude des **propriétés asymptotiques**) ;

(c) comme un **vecteur** (colonne) d'un **espace vectoriel** $\mathcal{X}$ supposé de dimension finie (Dim $\mathcal{X}$ = N) :

(2) $X = (X_1, ..., X_N)' : \Omega \mapsto \mathcal{X}$ (eg dans les calculs algébriques, avec eg $\mathcal{X} = \mathbf{R}$) ;

Les trois notations précédentes sont souvent implicitement employées de façon alternative.

(vii) Une **convention de notation** consiste à utiliser des majuscules pour désigner des va et des minuscules pour désigner leurs valeurs (ou « observations ») en certains points (unités statistiques). Ainsi, les valeurs $X(\omega) = (X_1(\omega), ..., X_N(\omega))$ au point $\omega \in \Omega$ sont souvent notées $x = (x_1, ..., x_N) = (x_n)_{n=1,...,N}$. La « valeur » $x = X(\omega)$ s'appelle alors **observation** (ou **échantillon observé**).

On définit, de même, les valeurs de X correspondant aux notations (b) et (c) précédentes.

Mais ceci peut créer des difficultés dans certains contextes (eg **conditionnement** de va) et n'est pas toujours nécessaire. Le contexte, ainsi que le type d'opérations mathématiques autorisées sur ces objets, doivent indiquer sans ambiguïté ce qui constitue une va (« avant observation ») et ce qui n'en est pas une (« après observation »).

(viii) Dans les généralités qui précèdent, les échantillons ne sont pas nécessairement supposés indépendants, ie les suites X de va considérées ne sont pas nécessairement constituées de va indépendantes. De ce point de vue, un échantillon est assimilable à un processus stochastique X, souvent supposé en **temps** discret (avec T = **N*** ou T = **Z**).

Si X est un **échantillon indépendant**, on peut le considérer comme un **processus purement aléatoire** particulier. Si, de plus, X est un **échantillon équidistribué**, il s'agit d'un processus aléatoire indépendamment et identiquement distribué (processus iid), ie provenant d'un **espace d'échantillonnage**. Un tel espace, qui fonde la notion de **modèle d'échantillonnage**, se décrit comme suit.

Une va $\xi : \Omega \mapsto \mathcal{X}_0$ (**variable parente**) est « observée » N fois. Si sa **lp** est P^{ξ}, on note P^X celle d'une suite $X = (X_1, ..., X_N)$ constituée de N **copies** de ξ, indépendantes entre elles, lp définie par $P^X = (P^{\xi})^{\otimes N}$ (cf **échantillon iid**).

(viii) En **théorie des sondages**, les populations sont essentiellement finies. Les notions d'**échantillonnage** et de **plan de sondage** peuvent être présentées comme suit.

La population (généralement finie) étant représentée par l'ensemble $\Omega = \{\omega_1, ..., \omega_M\}$, on appelle **échantillon** de taille $N \in N^*$ tout élément ω appartenant à la puissance cartésienne Ω^N. Un **plan de sondage** est alors une (mesure de) probabilité Π définie sur $\mathcal{P}(\Omega^N)$. Le **modèle statistique** correspondant s'écrit $(\Omega^N, \mathcal{P}(\Omega^N), \Pi)$. Cette présentation permet notamment :

(a) d'une part, d'obtenir des échantillons avec répétition (cf **échantillon avec remise**, **échantillon bernoullien**, **tirage bernoullien**) ou sans répétition (cf **tirage exhaustif**, **échantillon sans remise**), selon la définition de la probabilité Π ;

(b) d'autre part, de « tirer » des échantillons de taille aussi grande que l'on veut (notamment dans le cas bernoullien).

(ix) En pratique, un N-échantillon $X = (X_1, ..., X_N)$ peut posséder une structure complexe, qui dépend, généralement, du problème étudié ou de la nature des données observées (cf aussi **disposition**, **indice**, **problème à plusieurs échantillons**). Ainsi, la structuration de X peut être :

(a) à une dimension (ie à un seul indice, comme dans la présentation ci-dessus). Dans ce cas, il peut aussi être partitionné (cf **échantillon partitionné**) ;

(b) à plusieurs dimensions (ie à plusieurs indices). Dans ce cas, si k est le nombre de dimensions considéré, l'indice n précédent est considéré comme un multi-indice (n_1 ,..., n_k), où n_i = 1 ,..., N_i pour tout i = 1 ,..., k (cf aussi **tableau de contingence**, **tableau statistique**). Par suite, on pose $N = \Pi_i N_i$. Lorsque $N_i = N_0$ (entier donné) pour tout i, on parle parfois d'**échantillon équilibré**. Une situation fréquente est celle où k = 2 (cf aussi **modèle multi-indicé**).

Enfin, dans certains contextes, X est un **« échantillon » partiellement observable** (**censure**) ou même totalement inobservable (eg **perturbation aléatoire** d'un modèle de régression, **variable latente**, etc).