ÉCHANTILLONNAGE (F1, F4, M)
(18 / 03 / 2020, © Monfort, Dicostat2005, 2005-2020)

(i) Soit $(\Omega_t , \mathcal{T}_t , P_t)_{t \in T}$ une **famille** quelconque d'**espaces probabilisés**, dans lesquels, pour tout $t \in T$,

(a) l'**ensemble** Ω_t représente une **population** constituée d'**événements** élémentaires ou d'**unités statistiques** ;

(b) $\mathcal{T}_t$ est une **tribu de parties** de Ω_t ;

(c) P_t une **mesure de probabilité** définie sur $\mathcal{T}_t$.

On dit que $(\Omega_t , \mathcal{T}_t , P_t)_{t \in T}$ est un **espace d'échantillonnage de base (ou d'événements, ou d'unités)**, et tout élément $\omega = (\omega_t)_{t \in T}$ de la famille $(\Omega_t)_{t \in T}$ est appelé **échantillon** (d'événements ou d'unités), avec $\omega_t \in \Omega_t , \forall\ t \in T$.

On peut donc noter $\Omega = (\Omega_t)_{t \in T}$, $\mathcal{T} = (\mathcal{T}_t)_{t \in T}$ et $\mathcal{P} = (P_t)_{t \in T}$.

Souvent, T est un **ensemble** fini de la forme $N_N^* = \{1 ,..., N\}$ et l'on appelle alors **espace d'échantillonnage « à distance finie »** (de base) (ou d'événements, ou d'unités) la **suite** (finie) résultante $(\Omega_n , \mathcal{T}_n , P_n)_{n=1,...,N}$.

Si T est un ensemble dénombrable (eg T = **N** ou **Z**), on appelle parfois **espace d'échantillonnage « asymptotique »** (de base) (ou d'événements, ou d'unités) la suite (infinie) résultante $(\Omega_n , \mathcal{T}_n , P_n)_{n \in T}$.

(ii) Soit alors $(\mathcal{X}_t , \mathcal{B}_t)_{t \in T}$ une famille d'**espaces d'observations**, aussi indexée par T, et $X = (X_t)_{t \in T}$ une famille de **variables aléatoires** $X_t : \Omega_t \mapsto \mathcal{X}_t$. On note, $\forall\ t \in T$, $P^{X(t)}$ ou $\mathcal{L}(X_t)$ la **loi de probabilité** de X_t (ie l'image de P_t par X_t) (où X(t) désigne par commodité X_t).

On dit que $(\mathcal{X}_t , \mathcal{B}_t , P^{X(t)})_{t \in T}$ est un **espace d'échantillonnage d'observations (ou de mesures**, au second sens**)**, et tout élément $x = (x_t)_{t \in T}$ de la famille $(\mathcal{X}_t)_{t \in T}$ est appelé **échantillon** (d'**observations** ou de mesures, au second sens), avec $x_t = X_t (\omega_t) \in \mathcal{X}_t , \forall\ t \in T$.

On peut donc noter $X = (X_t)_{t \in T}$, $\mathcal{X} = (\mathcal{X}_t)_{t \in T}$, $\mathcal{B} = (\mathcal{B}_t)_{t \in T}$ et $\mathcal{P} = (P^{X(t)})_{t \in T}$.

Lorsque T est assimilable à N_N^* = {1 ,..., N}, la suite $(\mathcal{X}_n , \mathcal{B}_n , P^{X(n)})_{n=1,...,N}$ est appelée **espace d'échantillonnage « à distance finie »** (d'observations ou de mesures).

Si T est dénombrable, on appelle parfois **espace d'échantillonnage « asymptotique » d'observations (ou de mesures)** la suite (infinie) résultante $(\mathcal{X}_n , \mathcal{B}_n , P^{X(n)})_{n \in T}$.

(iii) Les définitions précédentes concernent ainsi des échantillons soit d'unités ω, soit d'observations $x = X(\omega)$ d'une variable X effectuées sur ces unités.

Chaque type d'échantillon est donc défini en sorte que chacun de ses éléments appartient à un ensemble (population ou espace d'observation) distinguable a priori : $\omega = (\omega_t)_{t \in T}$, avec $\omega_t \in \Omega_t$, $\forall$ t $\in$ T, et $X = (X_t)_{t \in T}$, avec $X_t(\omega_t) = x_t \in \mathcal{X}_t$, $\forall$ t $\in$ T.

En pratique, une **situation statistique** courante suppose :

(a) dans le cas des unités, que, pour tout t $\in$ T, $\Omega_t = \Omega_0$, $\mathcal{T}_t = \mathcal{T}_0$ et $P_t = P_0$ (données de base identiques) ;

(b) dans le cas des observations, que, pour tout t $\in$ T, $\mathcal{X}_t = \mathcal{X}_0$, $\mathcal{B}_t = \mathcal{B}_0$ et $P^{X(t)} = P^{X(0)}$ (données d'observation identiques).

On dit alors que l'espace probabilisé $(\Omega, \mathcal{T}, P)$ est :

(a) un **espace d'échantillonnage** (à distance finie N) ssi il peut s'écrire sous forme de « puissance » (cf **produit d'espaces mesurés**, avec mêmes espaces facteurs) :

(1) $(\Omega_0^N, \mathcal{T}_0^{\otimes N}, P_0^{\otimes N})$,

dans laquelle $\mathcal{T}_0^{\otimes N}$ désigne la puissance tensorielle d'ordre N de $\mathcal{T}_0$ et $P_0^{\otimes N}$ celle de la mesure de probabilité P_0 ;

(b) un **espace d'échantillonnage** (asymptotique) ssi il peut s'écrire sous forme de « puissance infinie dénombrable » :

(2) $(\Omega_0^{\mathbf{N}}, \mathcal{T}_0^{\otimes \mathbf{N}}, P_0^{\otimes \mathbf{N}})$.

De même, on dit que l'espace probabilisé $(\mathcal{X}, \mathcal{B}, \mathcal{P})$ est :

(a) un **espace d'échantillonnage** (à distance finie N) ssi il peut s'écrire sous forme de « puissance » (mêmes espaces facteurs) :

(3) $(\mathcal{X}_0^N, \mathcal{B}_0^{\otimes N}, (P^{X(0)})^{\otimes N})$,

dans laquelle $\mathcal{B}_0^{\otimes N}$, désigne la puissance tensorielle d'ordre N de $\mathcal{B}_0$ et $(P^{X(0)})^{\otimes N}$ celle de $P^{X(0)}$;

(b) un **espace d'échantillonnage** (asymptotique) ssi il peut s'écrire sous forme de « puissance infinie dénombrable » :

(4) $(\mathcal{X}_0^{\mathbf{N}}, \mathcal{B}_0^{\otimes \mathbf{N}}, (P^{X(0)})^{\otimes \mathbf{N}})$.

(iv) Souvent (eg en **théorie des sondages**), les « données » du problème considéré sont plus simples et concernent des ensembles finis.

Soit $\Omega = \{\omega_1, ..., \omega_M\}$ un ensemble fini donné (population) (Card Ω = M, avec M $\geq$ 2).

On appelle **(procédé d') échantillonnage** de taille N dans Ω la réalisation d'une **expérience aléatoire** définie par un **espace probabilisé** de la forme :

(2) $(\Omega^N, \mathcal{P}(\Omega^N), \Pi)$,

dans laquelle Π est une **mesure de probabilité** (en général fixée a priori) définie sur $\mathcal{P}(\Omega^N)$. Un élément ω de Ω^N s'écrit donc :

(3) $\omega = (\omega_1, ..., \omega_N)$, avec $\omega_n \in \Omega$, $\forall$ n $\in N_N^*$.

Si M et N sont donnés, Π suffit à décrire l'échantillonnage (**schéma probabiliste** de tirage).

(v) Les deux exemples suivants sont classiques :

(a) l'**échantillonnage bernoullien**, ou **échantillonnage avec remise**, pour lequel (cf **échantillon bernoullien**, **échantillon avec remise**, **tirage bernoullien**) :

(4) $\Pi(\{\omega\}) = M^{-N}$, $\forall \{\omega\} \in \mathcal{P}(\Omega^N)$;

(b) l'**échantillonnage exhaustif**, ou **échantillonnage sans remise**, pour lequel (cf **échantillon sans remise**, **tirage exhaustif**) :

(5) $\Pi(\{\omega\}) = \begin{cases} \{M(M-1)...(M-N+1)\}^{-1} \text{ sur } \Delta_\Omega, \\ 0 \text{ ailleurs,} \end{cases}$

où Δ_Ω est la partie de Ω^N dont les N-uples $\omega = (\omega_1, ..., \omega_N)$ ont des coordonnées distinctes deux à deux. Autrement dit, $\Pi(\{\omega\}) = (A_N^M)^{-1}$ (nombre d'arrangements sans répétition) sur Δ_Ω .

(v) Lorsque des statistiques dérivées d'un échantillon ne possèdent pas certaines propriétés sastisfaisantes (biais, variabilité), le statisticien peut parfois procéder à **ré-échantillonnage**. Deux **méthodes de ré-échantillonnage** importantes sont alors :

(a) la **méthode de EFRON** ;

(b) la **méthode de QUENOUILLE**.