LOI D'ÉCHANTILLONNAGE (F4, F5)
(26 / 04 / 2020, © Monfort, Dicostat2005, 2005-2020)

La loi de probabilité d'une statistique définie (ou calculée) à partir d'un échantillon est appelée **loi d'échantillonnage** ou **distribution d'échantillonnage**.

(i) Soit $(\Omega, \mathscr{T}, \mathscr{P})$ un **modèle statistique**, $(\mathscr{X}^N, \mathscr{B}^{\otimes N})$ un **espace d'échantillonnage** et $X : \Omega \mapsto \mathscr{X}^N$ un **échantillon** donné. A toute **mesure de probabilité** $P \in \mathscr{P}$ correspond une **loi de probabilité** possible $P^X = X(P)$ pour l'échantillon X. Par suite, si $(\mathscr{Y}, \mathscr{C})$ est un **espace mesurable** auxiliaire et $s : \mathscr{X}^N \mapsto \mathscr{Y}$ une application mesurable, la **statistique** S définie par $S = s(X) = s \circ X$ suit une loi définie selon :

(1) $\quad P^S = P^{s(X)} = s(P^X) = (s \circ X)(P), \qquad \text{où } P \in \mathscr{P}.$

On dit que P^S est la **loi d'échantillonnage** associée à X : P^X elle-même est l'exemple le plus simple d'une telle loi, avec $(\mathscr{Y}, \mathscr{C}) = (\mathscr{X}^N, \mathscr{B}^{\otimes N})$ et $s = id(\mathscr{X}^N)$.

(ii) Ainsi, la **loi du chi-deux** ou la **loi de FISHER-SNEDECOR** sont des lois d'échantillonnage classiques.

Comme toute lp, P^S possède des **caractéristiques** (**moment**, **mode**, **quantile**, **fr**, **fc**, **densité**, etc) appelées **caractéristiques d'échantillonnage**.

(iii) Selon le **contexte statistique** :

(a) dans un **problème d'estimation**, S est généralement un **estimateur** : sa loi permet alors de définir des **régions de confiance** ;

(b) dans un **problème de test**, S est une **statistique de test** : sa loi permet de définir des **régions critiques d'un test**.

(iv) La notion de « loi d'échantillonnage » (tq P^S) se distingue de deux autres notions :

(a) celle de **loi empirique**, notée P_N , associée à l'échantillon X : P_N est une distribution aléatoire car elle dépend de X ;

(b) lorsque $\mathscr{Y} = \mathscr{X}^N$ et $s \neq id_{\mathscr{Y}}$, celle de **loi de l'échantillon**, ou **loi d'échantillon**, qui est la lp de X. Notée P^X, cette loi est, par définition, l'image d'une **mesure de probabilité** $P \in \mathscr{P}$. Les deux notions coïncident lorsque S = X.

(v) La présentation précédente est de forme non paramétrée. Elle se particularise directement au cas d'un **modèle paramétrique**.

(vi) Lorsqu'une loi d'échantillonnage est de forme analytique complexe, on peut souvent en simplifier l'étude :

(a) soit par approximation : cf **approximation**, **approximation des lois**, **approximation d'une densité**, **loi approchée** ;

(b) soit en adoptant un point de vue asymptotique (ie $N \to +\infty$) : cf **loi asymptotique**.

(vii) Une situation classique concerne un N-**échantillon iid** $X = (X_1 ,..., X_N)$ issu d'une **vars** ξ (**variable parente**), de loi P^ξ, avec $\xi \in L_{\mathbf{R}}^2 (\Omega, \mathcal{T}, \mathcal{P})$ (variable de carré intégrable). On note $\mu = E\,\xi$ et $\sigma^2 = V\,\xi$ (**moyenne** et **variance** théoriques).

Ainsi, X peut être un **échantillon avec remise**, ou encore un échantillon tiré dans une **population infinie**.

Par suite :

(a) si $\overline{X}_N$ désigne la **moyenne empirique**, on montre que :

(a_1) $E\ \overline{X}_N = \mu$ (**espérance** de la moyenne empirique) : absence de **biais** ;

(a_2) $V\ \overline{X}_N = N^{-1}\,\sigma^2$ (**variance** de la moyenne empirique) ;

(a_3) si $P^\xi = \mathcal{N}(\mu, \sigma^2)$, alors $\mathcal{L}(\ \overline{X}_N\ ) = \mathcal{N}(\mu, N^{-1}\,\sigma^2)$, ie $\mathcal{L}(U_N) = \mathcal{N}(0, 1)$, avec $U_N = (\sigma / N^{1/2})^{-1} (\ \overline{X}_N - \mu\ )$;

(a_4) si P^ξ est quelconque et si le **théorème de la limite centrale** est applicable, alors (mêmes notations) :

(2) $\mathcal{L}(U_N) \ \to_{N \to +\infty}\ \mathcal{N}(0, 1)$;

(b) si F_N désigne la **fréquence empirique** de l'ensemble des coordonnées X_n de X vérifiant une propriété donnée, on remplace dans ce qui précède μ par p et σ par $(p\ q)^{1/2}$ (cf **loi binômiale**) ;

(c) si $S_N^2 = N^{-1} \Sigma_{n=1}^N (X_n - \overline{X}_N)^2$ désigne la **variance empirique** (brute, ie non corrigée), alors :

(c_1) $E\ S_N^2 = N^{-1} (N - 1)\,\sigma^2$, ie $(N-1)^{-1}\,N\,S_N^2$ est un **estimateur sans biais** de σ^2 ;

(c_2) $V\ S_N^2 = E\ S_N^4 - N^{-2} (N - 1)^2\,\sigma^4$;

(c_3) si $P^\xi = \mathcal{N}(\mu, \sigma^2)$, alors $\mathcal{L}(N\,\sigma^{-2}\,S_N^2) = \mathcal{X}_{N-1}^2$ (**loi du chi-deux** à N-1 **dl**) ;

(c_4) si P^{ξ} est quelconque et si le **théorème de la limite centrale** est applicable, alors l'égalité précédente est asymptotiquement vraie (cf **loi asymptotique**).