ESTIMATEUR DE LA FONCTION DE RÉPARTITION (C5, H)
(14 / 01 / 2020, © Monfort, Dicostat2005, 2005-2020)

On considère un **modèle statistique** de base $(\Omega, \mathcal{T}, \mathcal{P})$ et son **modèle image** $(\mathcal{X}, \mathcal{B}, \mathcal{P}^{\xi})$ par une **vars** $\xi : \Omega \mapsto \mathcal{X}$.

On suppose que l'**espace d'observation** est $(\mathcal{X}, \mathcal{B}) = (\mathbf{R}, \mathcal{B}_{\mathbf{R}})$. On note F la **fonction de répartition** associée à une **loi** $P^{\xi} \in \mathcal{P}^{\xi}$ possible pour ξ et l'on observe un **échantillon iid** $X = (X_1, ..., X_N)$ issu de la **variable parente** ξ.

Pour estimer F à l'aide de X, on cherche une fonction $F_N\tilde{}$ dans l'**ensemble** $\mathcal{T}$ des fr possibles (ensemble associé à $\mathcal{P}^{\xi}$). On considère ainsi que l'ensemble des **paramètres** θ du **problème de décision** (ie ici **problème d'estimation**) considéré est $\Theta = \mathcal{T}$ et que l'espace des **décisions** D est l'ensemble des fr tq $F_N\tilde{}$.

F peut être estimée à partir de la **fonction de répartition empirique** F_N : cet estimateur « naturel » est le plus simple (cf **statistique naturelle**), mais peu pratique pour les calculs analytiques.

(i) Pour résoudre le problème d'estimation, on doit choisir une **fonction de perte** L : Θ x D $\mapsto$ **R**. Celle-ci est souvent définie à l'aide d'une distance : la méthode d'estimation correspondante est appelée **méthode à distance minimum** (cf **estimateur à distance minimum**).

Des exemples classiques de fonction de perte sont les suivants :

(a) $L_p(F, F_N\tilde{}) = \int |F - F_N\tilde{}|^p \, dF$, $\forall\, p \in \mathbf{N}^*$ (norme dans L^p),

avec, en particulier (norme de type quadratique) :

(1) $L_2(F, F_N\tilde{}) = \int (F - F_N\tilde{})^2 \, dF$,

et (norme de type supremum) :

(2) $L_\infty(F, F_N\tilde{}) = \sup_{x \in \mathbf{R}} |F(x) - F_N\tilde{}(x)|$;

(b) la fonction de perte plus générale :

(4) $L_{p,\psi}(F, F_N\tilde{}) = \int |F - F_N\tilde{}|^p \, \psi(F) \, dF$, $\forall\, p \in \mathbf{N}^*$,

où $\psi : [0,1] \mapsto \mathbf{R}_+$ est supposée continue.

(ii) Lorsque $L = L_\infty$, le **théorème de CANTELLI-GLIVENKO** exprime une propriété de **convergence presque sûre** :

(5) $L_\infty (F, F_N) \to_{p.s.} 0$.

(iii) Lorsque $L = L_2$, on considère la **statistique d'ordre** $X^{(.)} = (X^{(1)} ,..., X^{(N)})$ associée à X, le **groupe de transformations** $\mathcal{G}$ définies sur l' « espace » $\mathcal{X}^{(.)}$ des statistiques d'ordre associé à $\mathcal{X}$ et l'on suppose que :

(a) φ : **R** $\mapsto$ **R** est une fonction continue (cf **application continue**) et strictement croissante ;

(b) $g (X^{(.)}) = g (X^{(1)} ,..., X^{(N)}) = (\varphi (X^{(1)}) ,..., \varphi (X^{(N)}))$, $\forall\, g \in \mathcal{G}$.

Une **règle de décision pure** est alors de la forme :

(6) $F_N\tilde{}\,(x) = \Sigma_{n=1}^{N-1} u_n \,.\, \mathbf{1}(A_n (x)) + \mathbf{1}(B_N (x))$, $\forall\, x \in \mathbf{R}$,

où $0 \le u_1 \le ... \le u_{N-1} \le 1$, $A_n = [X^{(n)}, X^{(n+1)}[$ et $B_N = [X^{(N)}, +\infty[$ (en notant **1**(R) la **fonction indicatrice** d'une **partie** R de **R**).