ÉCHANTILLON ARTIFICIEL (C2, C12, E, F, M)
(10 / 05 / 2020, © Monfort, Dicostat2005, 2005-2020)

Un **échantillon artificiel** sert à simuler un **phénomène** aléatoire (cf **échantillon aléatoire**). Il s'agit d'un ensemble d' « observations fictives » issues d'un **schéma probabiliste** donné.

(i) Soit $(\Omega, \mathcal{T}, P)$ un **espace probabilisé** et $\xi : \Omega \mapsto \mathbf{R}$ une **vars** de loi P^{ξ}.

On appelle **échantillon « artificiel »**, ou **échantillon « fictif »**, de la **variable parente** ξ, ou de P^{ξ}, un N-**échantillon** $X = (X_1 ,..., X_N)$ engendré à l'aide d'une table de **nombres au hasard** et de la **fr** F de P^{ξ}.

Ce procédé constitue ainsi un **processus générateur de données artificielles**.

(ii) Une méthode générale de création d'échantillons artificiels est fondée sur le **lemme d'uniformisation des lois** : si $\xi \sim P^{\xi}$ et si F est la fr associée à P^{ξ}, alors la va $\eta = F(\xi)$ suit la **loi uniforme** $\mathcal{U}(0, 1)$.

La transformation $\xi \mapsto \eta$ est parfois dite **transformation par répartition de la probabilité**.

La procédure qui en résulte est la suivante :

(a) tirage de N nombres au hasard (uniforme) pour constituer l'échantillon $U = (U_1 ,..., U_N)$ dont la va parente est $\eta \sim \mathcal{U}(0, 1)$;

(b) inversion de F pour chaque observation U_n , d'où $X_n = F^{-1}(U_n)$, $\forall\ n \in N_N^*$;

(c) regroupement en échantillon $X = (X_1 ,..., X_N)$ issu de la va parente $\xi \sim P^{\xi}$.

(iii) Par définition, U est constitué par un **tirage au hasard « uniforme »**. On dit alors que X résulte d'un **tirage au hasard « non uniforme »** (ie selon P^{ξ}).

(iv) Les applications mettant en oeuvre des échantillons fictifs sont variées : **méthodes de MONTE CARLO**, **simulations** diverses. On peut leur ajouter :

(a) l'étude des **changement de variables aléatoires** (pour simuler des lois) et de leurs **propriétés asymptotiques**. Ainsi, avec un **vecteur aléatoire** $\xi \sim P^{\xi}$ et une fonction mesurable non élémentaire donnée φ, l'étude de la loi P^{η} du vecteur aléatoire $\eta = \varphi(\xi)$ peut s'effectuer comme suit :

$(a)_1$ génération, à l'aide d'une table de nombres au hasard, d'un « grand » nombre de copies X_n (où $n \in N_N^*$ et $N >> 1$) (échantillon artificiel) ;

$(a)_2$ calcul des copies $Y_n = f(X_n)$ de η resp associées aux copies X_n de ξ ;

(a)$_3$ étude des propriétés ou **caractéristiques** (**moments**, **densité**, **fr**, etc) de la loi P^η à partir de la **loi empirique** Q_N associée à l'échantillon $Y = (Y_1, ..., Y_N)$. En effet, sous des hypothèses de régularité larges, la **fr empirique** F_N associée à X converge P-p.s. vers la fr théorique F de ξ (cf **théorème de CANTELLI-GLIVENKO**, **théorème fondamental de la statistique**). Sous des **conditions de régularité** (**continuité**, **différentiabilité**, etc), il en va généralement de même de la fr empirique G_N associée à Y (ie déduite de Q_N) pr à la fr théorique G de η ;

(b) l'étude des problèmes de **petits échantillons** : on tire un grand nombre de « petits » échantillon (ie X tq $N << +\infty$) afin de préciser certaines propriétés statistiques qu'il serait impossible d'atteindre à l'aide d'un calcul analytique direct. On parle aussi d'**étude à distance finie**.