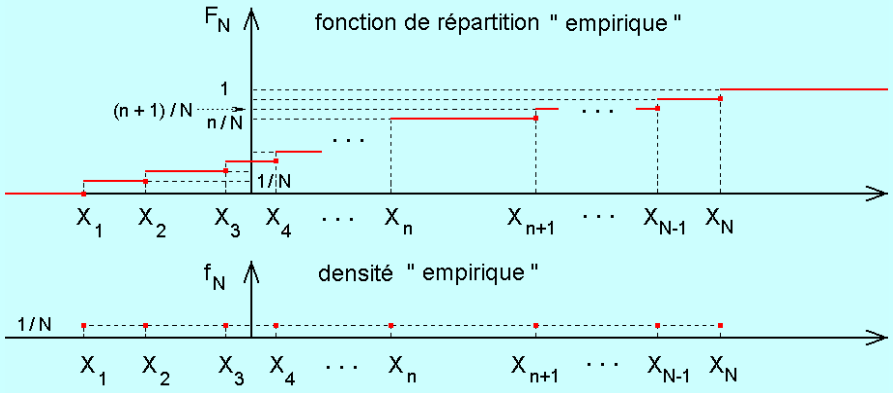
DENSITÉ EMPIRIQUE (F3)
(28 / 11 / 2019, © Monfort, Dicostat2005, 2005-2019)

Soit ξ une **vars** dont la **lp** est P^{ξ}, et soit f la densité de P^{ξ} pr à une **mesure positive** μ. Tout **histogramme** constitué à partir d'**observations** de ξ générées par P^{ξ} constitue un **estimateur** (naturel) de f (cf **méthode du noyau**).

La notion intuitive de **densité empirique** peut alors se référer à toute fonction en escalier (donc non partout dérivable) associée à cet histogramme.

(i) Soit $(\Omega, \mathcal{T}, P)$ un **espace probabilisé**, $(\mathcal{X}, \mathcal{B})$ un **espace d'observation** et $X : \Omega \mapsto \mathcal{X}_0^N$ un **échantillon aléatoire** de **loi** P^X (image de P par X). On note P_N la **loi empirique** associée à X et, le cas échéant (ie lorsque $\mathcal{X}_0 \subset \mathbf{R}$), F_N sa **fr empirique**. Soit ν_N la **mesure de comptage** (aléatoire) qui dénombre les coordonnées de X prenant leurs valeurs dans tout **borélien** (élément de $\mathcal{B}$).

F_N n'est pas différentiable, mais il est possible de définir une notion de **densité empirique** comme **dérivée de NIKODYM-RADON** f_N de la mesure P_N pr à la mesure ν_N . Si $\mathcal{X}_0 = \mathbf{R}$ (muni de la **relation d'ordre** usuelle $\leq$) et si X est ordonné selon $X_1 \leq ... \leq X_N$ (cf **statistique d'ordre**), un « **graphe** » de f_N peut être représenté à l'aide des couples $(X_n , f_N (X_n))$: la première coordonnée X_n est donc aléatoire et la seconde $f_N (X_n) = n/N$ est constante entre les abscisses X_n et X_{n+1} . Chaque point X_n est un point de discontinuité (eg à gauche) de f_N (cf schéma ci-dessous).

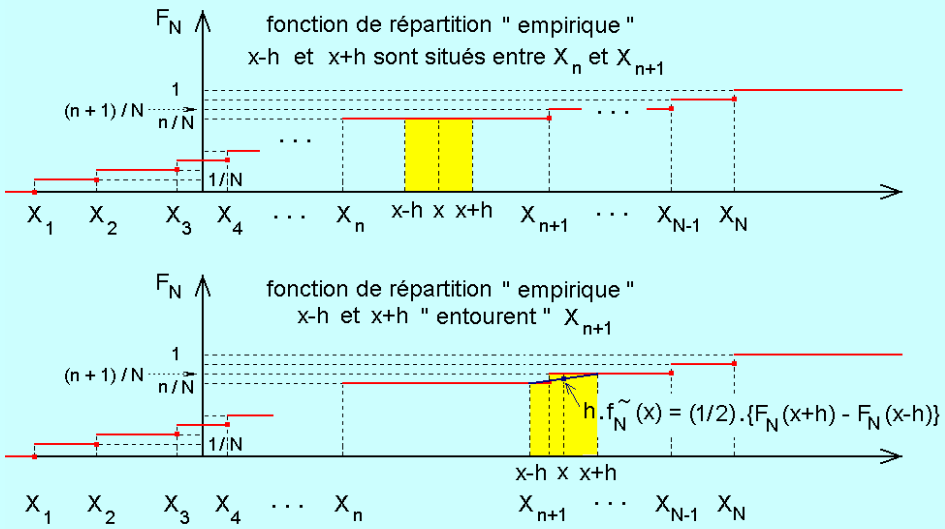


Dans les questions d'**estimation non paramétrique** de la densité (cf **estimateur de la densité**), on préfère généralement des estimateurs plus simples, ou plus « réguliers », que l'estimateur « naturel » f_N .

(ii) Dans le même cadre que le précédent, on appelle aussi parfois (cf **méthode du noyau**) **densité empirique** de pas $h > 0$ de X la fonction $f_N^{\sim}$ définie sur **R** par :

$$(1) \quad f_N^{\sim}(x) = \begin{cases} (2h)^{-1} . \{F_N (x + h) - F_N (x) + F_N (x) - F_N (x - h)\} \\ h^{-1} . \{(F_N (x + h) - F_N (x - h)) / 2\}, \end{cases} \quad \forall x \in \mathbf{R},$$

où F_N désigne la **fr empirique**. Autrement dit, en tout point $x \in \mathbf{R}$, la valeur $h \, . \, \tilde{f_N}(x)$ est égale au demi-écart entre $F_N(x + h)$ et $F_N(x - h)$ (cf schéma ci-dessous).



Cette « densité » f_N n'est pas la **dérivée** (au sens usuel) de F_N , mais plutôt une variation relative en termes de **différence finie**. Cependant, si les coordonnées X_n (n = 1 ,..., N) de X suivent une même lp P^{ξ} de densité f pr à μ (cf **échantillon iid**), alors f_N constitue un estimateur (simple) de la densité f par la méthode du noyau.

La définition (1) peut s'étendre directement en remplaçant h par un pas variable en fonction de la taille N de l'échantillon, ie en utilisant une suite $(h_n)_{n \in \mathbf{N}^*}$ constituée de pas $h_n \in \mathbf{R}_+^*$, $\forall \, n \in N^*$, et tq $\lim_n h_n = 0+$.