MÉTHODE DU NOYAU (C5, H2)
(18 / 04 / 2020, © Monfort, Dicostat2005, 2005-2020)

La **méthode du noyau** est une **méthode non paramétrique** d'**estimation** d'une densité de probabilité ou d'une caractéristique légale (cf **estimation de la densité**).

(i) Soit $(\Omega, \mathcal{T}, \mathcal{P})$ une **représentation statistique**, $\xi : \Omega \mapsto \mathbf{R}$ une **vars** et X un N-**échantillon iid** dont la **variable parente** est ξ. On suppose que la **loi** P^{ξ} de ξ possède une **densité** $f = d P^{\xi} / d\lambda_1$ pr à la **mesure de LEBESGUE** (on assimile f et sa classe de λ_1-équivalence).

La **méthode du noyau de M. ROSENBLATT** consiste à définir un **estimateur ponctuel** de f (x), appelé **estimateur par (le) noyau**, selon :

(1) $\quad f_{p,N}\tilde{}\,(x) = (N \,.\, h_N)^{-1} \Sigma_{n=1}^{N}\ p\{(x - X_n) / h_N\}, \qquad \forall\, x \in \mathbf{R}, \forall\, N \in \mathbf{N}^*,$

dans lequel :

(a) la **suite** $h = (h_N)_{N \in \mathbf{N}^*}$ est constituée de nombres réels positifs $h_N \in \mathbf{R}_+^*$, $\forall\, N \in N^*$, chacun d'eux étant appelé **largeur de bande**, ou **largeur de fenêtre** ;

(b) la fonction $p : \mathbf{R} \mapsto \mathbf{R}_+$ est appelée **noyau de l'estimateur**. Elle est souvent notée N, ou K (de l'anglais « *kernel* ») ou encore W (de l'anglais « *weight* »).

La notation $f_{p,N}\tilde{}$ indique la dépendance de l'estimateur pr au noyau et à la taille N d'échantillon. Elle est souvent simplifiée.

(ii) Les propriétés de $f_{p,N}\tilde{}$ dépendent des hypothèses faites sur ses composantes. Celles-ci sont choisies afin d'obtenir des propriétés optimales, eg :

(a) la suite h est décroissante et tend vers zéro. Par exemple, $h_N^2 > N^{\alpha - 1}$, avec $\alpha > 0$, ou $h_N = \gamma \,.\, N^{-\beta}$, avec $\gamma > 0$ et $\beta = 1 / 5$) ;

(b) la fonction p peut être :

(b_1) soit une **densité de probabilité de type N**, ie symétrique pr à zéro, unimodale en zéro et tq $p'(x) \geq 0$, $\forall\, x \in \mathbf{R}_-$ (croissance à gauche de zéro), et $p'(x) \leq 0$, $\forall\, x \in \mathbf{R}_+$ (décroissance à droite de zéro) (cf **loi symétrique**, **loi unimodale**) ;

(b_2) soit une **fonction de poids**, ie une fonction tq $\int p\, d\lambda_1 = 1$, $p(-x) = p(x)$ (**symétrie**), $p \in \mathcal{L}_{\mathbf{R}+}^{1}(\mathbf{R}, \mathcal{B}_{\mathbf{R}}, \lambda_1)$ (**fonction intégrable**), $\sup_x p(x) < \infty$ et $\lim_{x \to \infty} x \,.\, p(x) = 0$;

(b_3) soit encore une fonction mesurable bornée tq $\lim_{|x| \to \infty} p(x) = 0+$ (cf **application mesurable**).

En pratique, on choisit souvent pour p une **densité de probabilité** : eg celle de la **loi normale** ou celle de la **loi de CAUCHY**.

Le principal problème de la méthode du noyau est le choix de h ou celui de p. L'estimateur paraît cependant peu sensible au choix de p.

(iii) On montre que :

(a) l'estimateur par le noyau est un **estimateur strict** (ie est une densité de probabilité) ;

(b) si $h_N \to 0+$ et $N\, h_N \to +\infty$ lorsque $N \to +\infty$, on établit la **convergence en moyenne quadratique** (ou **convergence dans L^p** avec p = 2) :

(2) $\quad f_{p,N}\tilde{} \to^{m.q.} f$;

(c) si p est une **fonction à variation bornée**, si f est une **fonction uniformément continue** et si h est une **suite** tq $\Sigma_{N \in \mathbf{N}^*} \exp(-\gamma\, N\, h_N^2) < +\infty$, $\forall\, \gamma > 0$, on établit la **convergence uniforme** presque sûre (cf **convergence presque sûre**) :

(3) $\quad \sup_{x \in \mathbf{R}} |f_{p,N}\tilde{}(x) - f(x)| \to_{N \to +\infty} 0$, $\quad$ P-p.s., $\quad \forall\, P \in \mathscr{P}$;

(d) sous des conditions assez larges, $f_{p,N}\tilde{}$ converge :

(d_1) ponctuellement vers f au sens de l'**écart quadratique moyen** ;

(d_2) globalement vers f au sens de l'**écart quadratique moyen intégré** ;

(e) pour estimer $f^{(j)}$ (**dérivée** d'ordre $j \in \mathbf{N}^*$ de f) par la **méthode du noyau**, on choisit un noyau p de classe C^j et l'on calcule la dérivée d'ordre j de $f_{p,N}\tilde{}$, notée $f_{p,N}\tilde{}^{(j)}$. Sous certaines conditions, on établit que :

(4) $\quad \sup_{x \in \mathbf{R}} |f_{p,N}\tilde{}^{(j)}(x) - f^{(j)}(x)| \to_{N \to +\infty} 0$, $\quad$ P-p.s., $\quad \forall\, P \in \mathscr{P}$.

De plus, $f_{p,N}\tilde{}^{(j)}$ admet une **loi asymptotique** gaussienne (cf **loi gaussienne**).

(iv) Un **histogramme** est un exemple le plus élémentaire d'estimateur par le noyau.

L'**estimateur de E. PARZEN** est aussi un cas particulier d'estimateur par noyau, dans lequel $h_N = h_0$, $\forall\, N \in \mathbf{N}^*$ (suite constante).

(v) La méthode s'étend à K dimensions. Si $\xi : \Omega \mapsto \mathbf{R}^K$ est un **vecteur aléatoire** réel et X un N-échantillon issu de la va parente ξ, un noyau adapté p : $\mathbf{R}^K \mapsto \mathbf{R}_+$ est tq :

(5)
$$\int p\, d\lambda_K = 1, \qquad p(-x) = p(x), \qquad \forall\, x \in \mathbf{R}^K,$$
$$\int x_k\, p(x)\, d\lambda_K(x) = 0, \qquad \forall\, k \in N_K^* \text{ (avec } x_k = pr_k\, x).$$

On suppose aussi que :

(6) $\int ||x||^{\alpha} \, p(x) \, d\lambda_K(x) < +\infty, \qquad \forall \, \alpha \in]0, +\infty[,$

où $||x||^2 = x' \, x$ (norme euclidienne).

L'**estimateur par le noyau** s'écrit alors :

(7) $f_{p,N}\tilde{}(x) = (N \, . \, h_N^K)^{-1} \, \Sigma_{n=1}^N \, p \{(x - X_n) / h_N\}, \qquad \forall \, x \in \mathbf{R}^K,$

avec $h_N = \gamma \, . \, N^{-\beta}$, $\gamma > 0$ et $\beta = (K + 4)^{-1}$.

(vi) La définition de l'estimateur par le noyau admet diverses variantes. Ainsi, dans le cas scalaire :

(7) $f_{p,N}\tilde{}(x) = N^{-1} \, . \, \Sigma_{n=1}^N \, p_N(x, X_n), \quad \forall \, x \in \mathbf{R},$

où $(p_N)_{N \in \mathbf{N}^*}$ est une suite de noyaux tq ceux définis précédemment. La forme (1) s'en déduit comme cas particulier, avec $p_N(x, u) = h_N^{-1} \, . \, p \{(x - u) / h_N\}$, $\forall \, (x, u) \in \mathbf{R}^2$.

L'estimateur (7) peut aussi s'écrire sous la forme (cf **statistique naturelle**) :

(8) $f_{p,N}\tilde{}(x) = \int p_N(x, u) \, dF_N(u),$

où F_N est la **fonction de répartition empirique** associée à X.