Statistique et analyse des données

Pierre Jacob

Histogrammes, noyaux et densités irrégulières

Statistique et analyse des données, tome 14, n° 1 (1989), p. 33-53
<http://www.numdam.org/item?id=SAD_1989__14_1_33_0>

Statistique et Analyse des Données
1989 - Vol. 14 n° 1 - p. 33 - 53

HISTOGRAMMES, NOYAUX ET DENSITES IRREGULIERES

Pierre JACOB

Laboratoire de Probabilités et Statistique
U.F.R. de Mathématiques
Université de Lille-Flandres-Artois
F-59 655 VILLENEUVE D'ASCQ Cedex

Résumé

Cet article aborde certaines propriétés de l'estimateur de l'histogramme, en dehors du cadre L_1 , vis-à-vis de densités qui possèdent des discontinuités : dans quelle mesure peut-on essayer de préserver les aspects intéressants de la convergence uniforme, aussi bien du point de vue asymptotique qu'à distance finie ? En conclusion le cas de l'estimateur de la méthode du noyau est également examiné.

Mots clés : histogramme, noyau, densité, discontinuités.

Classification AMS : 62 G 05
Classification STMA : 04 180, 04 080

Abstract

In this paper some properties of the histogram estimate are studied, apart from the L_1 view, when the densities are not necessarily continuous : we try to preserve the nice aspects of the uniform distance, for asymptotic results as well as at finite distance. In conclusion, we examine the kernel estimate in the same situation.

Keywords : histogram, kernel, density, discontinuities.

Manuscrit reçu le 29. 6. 88, révisé le 24. 1. 89

1. INTRODUCTION

Soit ($X_1, \ldots, X_n$) un échantillon d'une loi P sur $\mathbb{R}$, de densité f par rapport à la mesure de Lebesgue λ. Soit (h_n) une suite de nombres positifs; pour tout entier relatif i, on pose :

$$\Delta_{i,n} = [\,(i-1)\,h_n\,,\, i\,h_n\,[\tag{1.1}$$

$$f_n = \sum_{i \in \mathbb{Z}} \left[\frac{1}{n\,h_n} \sum_{k=1}^{n} \mathbb{1}_{\Delta_{i,n}}(X_k) \right] \mathbb{1}_{\Delta_{i,n}} \tag{1.2}$$

$$\bar{f}_n = \sum_{i \in \mathbb{Z}} \left[\frac{1}{h_n} P(\Delta_{i,n}) \right] \mathbb{1}_{\Delta_{i,n}} = E\, f_n \tag{1.3}$$

Parmi les modes de convergence usuels de l'histogramme f_n , les plus intéressants sont sans doute la convergence uniforme et la convergence en moyenne. D'après un résultat, d'ailleurs plus général, de Geffroy (1974), si f est uniformément continue, si $h_n \to 0$ et si $(nh_n)/\mathrm{Log}\,n \to \infty$, alors presque complètement sûrement :

$$\| f_n - f \|_\infty = \sup_x | f_n(x) - f(x) | \to 0 \tag{1.4}$$

D'autre part, sans hypothèse spéciale sur f, Abou-Jaoudé (1976) a montré que, si $h_n \to 0$ et $nh_n \to \infty$:

$$\| f_n - f \|_1 = \int | f_n(x) - f(x) |\, dx \to 0 \tag{1.5}$$

presque complètement sûrement.

En fait, pour le résultat (1.4), on constate que l'hypothèse d'uniforme continuité de f sert uniquement à obtenir la convergence vers 0 du biais $\| f - \bar{f}_n \|_\infty$, tandis que la convergence vers 0 de l'aléa $\| f_n - \bar{f}_n \|_\infty$ n'utilise pas cette hypothèse. Or, bien que la convergence en moyenne présente de nombreux avantages [Devroye, Györfi (1984)], la convergence uniforme a quelque chose de particulièrement visuel sur $\mathbb{R}$, que l'on peut souhaiter préserver quand f a des discontinuités.

C'est pourquoi nous allons fixer notre attention sur le biais dans cet article. Un mode de convergence approprié, fondé sur la dilatation des graphes complets des densités, nous permet de rendre compte de ses propriétés asymptotiques, quand f n'est pas nécessairement continue. D'autre part, à distance finie, ces idées permettent la

construction de zones de confiance tenant compte du degré d'irrégularité supposé de la densité à estimer.

2. PRELIMINAIRES

Les densités que nous envisageons dans cet article ont en tout point une limite à droite et une limite à gauche, toutes deux finies, et tendent vers 0 à l'infini : on notera D_0 leur ensemble. Dans ce contexte, les histogrammes sont des estimateurs propres de la densité.

Pour toute densité $g \in D_0$, le graphe complet de g est l'ensemble :

$$(2.1) \qquad \Gamma_g = \left\{ (x,y) : x \in \mathbb{R} \; ; \; g(x^-) \le y \le g(x^+) \text{ ou } g(x^+) \le y \le g(x^-) \right\}$$

Si x est un point de discontinuité de g, la valeur g(x) n'a pas d'intérêt particulier ; on supposera simplement qu'elle est comprise entre $g(x^+)$ et $g(x^-)$, c'est-à-dire que $(x,g(x)) \in \Gamma_g$.

Ensuite, pour tout $\varepsilon > 0$, le dilaté d'ordre ε de Γ_g est l'ensemble :

$$(2.2) \qquad \Gamma_g^{\varepsilon} = \left\{ (u,v) : \exists \, (x_u, y_v) \in \Gamma_g : |u - x_u| < \varepsilon \; ; |v - y_v| < \varepsilon \right\}$$

L'idée est de considérer qu'un histogramme f_n est une bonne approximation d'une densité f dès que Γ_{f_n} est contenu dans Γ_f^{ε}, pour ε fixé. On pose donc :

$$(2.3) \qquad \forall f,g \in D_0 : \; \delta(f,g) = \inf \left\{ \varepsilon > 0 : \Gamma_g \subset \Gamma_f^{\varepsilon} \right\}$$

Bien que δ ne soit pas une distance, on peut déjà obtenir un certain nombre de propriétés intéressantes, dont nous rejetons la démonstration à la partie 6 de cet article.

Propriétés 2.1: Soit f, g, h, g_n, des éléments de D_0 :

(i) $\Gamma_f = \Gamma_g \Rightarrow f = g$, λ-presque partout

(ii) $\delta(f,g) = 0 \Rightarrow \Gamma_f = \Gamma_g$

(iii) $\delta(f,g) \le \delta(f,h) + \delta(h,g)$

(iv) $\delta(f,g) \le \| f - g \|_\infty$

(v) Si f est uniformément continue et si $\delta(f,g_n) \to 0$, alors $\| f - g_n \|_\infty \to 0$.

Inversement, $\delta(f_n,f)$ a un intérêt pratique : en effet, si $\delta(f_n,f) < \varepsilon$, le graphe de la densité inconnue f est dans un dilaté d'ordre ε de celui de l'histogramme f_n : par conséquent, dans la mesure du possible, on utilisera la distance Δ définie par :

$$\forall\ f,g \in D_o\ :\ \Delta(f,g) = \max\left(\delta(f,g);\delta(g,f)\right) \tag{2.4}$$

Il s'agit d'une sorte de distance de Paul Lévy symétrisée, puisque Δ tient compte, comme cette dernière, d'une marge d'erreur sur les abscisses aussi bien que sur les ordonnées. Evidemment la distance de Paul Lévy est adaptée à la propriété de monotonie des fonctions de répartition. On peut donc penser plutôt à utiliser une généralisation à $\mathbb{R}$ de la classique distance de Skorokhod sur D[0,1] (Billingsley ,1968). Mais il s'agit d'une notion trop forte pour exprimer la convergence du biais de l'histogramme :

Contre-exemple 2.1 : Soit $f = 2\ .\ \mathbb{1}_{[1/4\ ,\ 3/4]}$; pour toute suite (λ_n) de fonctions continues strictement croissantes dans [0,1], telles que $\lambda_n(0) = 0$ et $\lambda_n(1) = 1$, la suite des composées $\bar{f}_n \circ \lambda_n$ ne converge pas uniformément vers f sur [0,1].

On s'intéressera aussi dans le paragraphe suivant aux densités tendant vers 0 à l'infini, ayant en tout point une limite à droite et une limite à gauche, finie sauf en un nombre fini de points. On notera D_∞ l'ensemble de ces densités. Les définitions (2.1), (2.2), (2.3) gardent tout leur sens. Cependant, il est clair que si une densité f appartient à D_∞ sans appartenir à D_O, l'histogramme f_n appartient de toute façon à D_O, si bien que :

$$\delta(f_n, f) = \Delta(f_n, f) = +\infty$$

Seul $\delta(f,f_n)$ garde un sens et on doit donc se résoudre à la perte de la symétrie. Les propriétés 2.1 et les propriétés ci-dessous montrent cependant que tout n'est pas perdu ! [§ 7 ; fig. 1].

Propriété 2.2 : Soit $f \in D_\infty$ et (g_n) une suite de D_O, telles que $\delta(f,g_n) \to 0$

(i) si $f(x^+) = +\infty$, $\exists\ (x_n) \downarrow x$ telle que $g_n(x_n) \to +\infty$

(ii) si $f(x^+) = f(x^-) = +\infty$, $g_n(x) \to +\infty$.

3. CONVERGENCE DE L'HISTOGRAMME

Le premier théorème exprime la convergence du biais de l'histogramme pour la distance Δ .

Théorème 3.1 : Soit f_n l'histogramme (1.2) et $\bar{f}_n$ sa moyenne (1.3) ;
si $f \in D_0$ et si $h_n \to 0$, alors $\Delta(\bar{f}_n, f) \to 0$.

Le deuxième théorème est un simple corollaire du précédent, ainsi que du résultat de convergence (1.4) et de la propriété (2.1.iv).

Théorème 3.2 : Soit f_n l'histogramme (1.2) et $f \in D_0$;
si $h_n \to 0$ et si $(nh_n)/\text{Log } n \to \infty$, alors presque complètement sûrement :

$$\Delta(f_n, f) \to 0$$

Dans le cas où $f \in D_\infty$, l'analogue du théorème 3.1 est :

Théorème 3.3 : Soit f_n l'histogramme (1.2) et $\bar{f}_n$ sa moyenne (1.3) ;
si $f \in D_\infty$ et si $h_n \to 0$, alors :

(i) $\delta(\bar{f}_n, f) = +\infty$

(ii) $\delta(f, \bar{f}_n) \to 0$.

Théorème 3.4 : Soit f_n l'histogramme (1.2) et $f \in D_\infty$;
si $h_n \to 0$ et si $(nh_n) / \text{Log } n \to \infty$, alors presque complètement sûrement :

$$\delta(f, f_n) \to 0$$

Corollaire 3.5 : Dans les conditions du théorème 3.4 :

(i) si $f(x^+) = +\infty$, $\exists (X_n) \downarrow x$ p.s. telle que $f_n(X_n) \to +\infty$ p.s.

(ii) si $f(x^+) = f(x^-) = -\infty$, $f_n(x) \to +\infty$ p.s.

4. DILATES DE CONFIANCE

Les résultats de convergence sont d'un intérêt théorique évident, mais sont de peu de secours pour l'utilisateur - spécialement dans une théorie où l'asymptotique signifie généralement d'énormes échantillons. Rappelons par exemple que le critère couramment utilisé pour le choix de h_n est celui qui consiste à minimiser l'erreur quadratique intégrée $E \int |f_n(x) - f(x)|^2 dx$ [Bosq et Lecoutre (1987), Freedman et Diaconis (1981)]. Sous de fortes conditions de régularité, il est recommandé de choisir h_n de l'ordre de $n^{-1/3}$. Cette faible vitesse de convergence fait qu'en pratique, un

histogramme ne comporte que peu de rectangles : on peut donc soupçonner le biais de n'être pas négligeable.

La loi limite de $\left[\inf_x \left(f_n(x) - f(x)\right) \; ; \; \sup_x \left(f_n(x) - f(x)\right) \right]$

a été déterminée pour une densité f lipschitzienne sur [0 , 1] par Bera (1977). Cela permet asymptotiquement de déterminer des bandes de confiance pour f. Cependant, à distance finie il est prudent de maintenir séparés le biais et l'aléa et de leur porter un intérêt égal : il faut garder à l'esprit que f_n est un estimateur sans biais de $\bar{f}_n$ et que $\bar{f}_n$ est une approximation de f. Si la construction de bandes de confiance pour $\bar{f}_n$ se ramène à un problème d'estimation par intervalles des paramètres d'une loi multinomiale, le problème principal semble être, en l'occurence, la construction de zones de tolérance pour f "autour" de $\bar{f}_n$.

Evidemment, il faudra bien faire des hypothèses sur la régularité supposée de f. L'aspect plus ou moins lisse de la densité ne sera pas notre critère de régularité : pour ce qui nous concerne, de faibles variations de f entre deux discontinuités valent mieux qu'une uniforme continuité avec de fortes oscillations. Nous utiliserons donc un critère, que nous appellerons module de régularité, et qui n'est autre que l'analogue de celui couramment utilisé sur D[0 , 1] [Billingsley (1968)].

Pour tout a < b, on pose :

$$w_f[a,b[\; = \; \sup \left\{ \left| f(x) - f(y) \right| \; ; \; (x,y) \in [a,b[\right\} \tag{4.1}$$

et pour tout $\delta > 0$:

$$F_\delta = [\inf\{x : f(x) > \delta\} \; ; \; \sup\{x : f(x) > \delta\}[\tag{4.2}$$

$$w'_f(\delta) = \inf_{\{t_i\}} \max_{1 \le i \le r} w_f \left[t_{i-1} , t_i \right[\tag{4.3}$$

où $\{t_i\}$ décrit les ensembles finis de $(r + 1)$ points pour lesquels on a $t_0 < \ldots < t_r$, $[t_0, t_r [= F_\delta$ et pour tout $i \in \{1,\ldots,r\}$ $(t_i - t_{i-1}) > \delta$.

Avant d'énoncer le théorème sur les dilatés, et si l'on veut vraiment se soucier du problème pratique, il faut bien constater que Δ ne possède pas la belle propriété d'invariance de $\| \; \|_1$, [Devroye, Györfi (1985)] ; or, $(X_1,\ldots,X_n)$ va en général s'exprimer dans un système d'unités, tout comme h_n. Il est donc nécessaire de distinguer l'erreur

verticale de l'erreur horizontale [§7 ; fig.2] ; pour cela on utilise la notion de dilaté bidirectionnel

$$(4.4)\forall\, \varepsilon > 0,\ \forall\, \delta > 0 : \Gamma_f^{\varepsilon,\delta} = \left\{ (x,y) \;\middle|\; \exists\, (u,v) \in \Gamma_f : |x-u| \le \varepsilon \text{ et } |y-v| \le \delta \right\}$$

Théorème 4.1 : Soit f_n l'histogramme (1.2) et f_n sa moyenne (1.3).:

$$\Gamma_f \subset \Gamma_{\bar{f}_n}^{h_n,\, w'_f(2h_n)}$$

Remarques 4.2 :

(i) si h_n est fixé par des considérations probabilistes usuelles [Bosq et Lecoutre (1987), Devroye et Györfi (1985)], $w'_f(2h_n)$ dépend du bon sens de l'utilisateur. En fait, s'intéresser à des classes du type :

$$C_\delta = \{ f \in D_0 : w'_f(2h_n) < \delta \}$$

c'est renoncer à tenir compte d'éventuelles excentricités de f.

(ii) il est instructif de dessiner des bandes de confiance autour d'histogrammes f_n, et des dilatés de confiance autour d'histogrammes $\bar{f}_n$ pris dans les bandes de confiance. Dans ces dilatés on peut tracer n'importe quelle densité de C_δ, pourvu que sa valeur moyenne coïncide avec celle de $\bar{f}_n$ sur chaque cellule. On pourra remarquer que, dans certains cas de figure, on est amené à tracer une densité discontinue plutôt qu'une densité uniformément continue mais trop irrégulière pour appartenir à C_δ.

5. CONCLUSION

L'intérêt de cette étude est essentiellement d'ordre visuel et semblerait sans doute amoindri dans le cas de densités multivariées. Cependant, une autre voie possible est de vérifier que d'autres estimateurs ont, sur $\mathbb{R}$, un bon comportement vis à vis de la distance Δ. C'est le cas, par exemple, de l'estimateur de la méthode du noyau :

Soit g une densité de probabilité bornée ; posons

$$(5.1) \qquad g_n(.) = \frac{1}{h_n}\, g\left(\frac{.}{h_n}\right)$$

(5.2) $$f_n(.) = \frac{1}{n} \sum_{j=1}^{n} g_n(. - x_j)$$

(5.3) $$\bar{f}_n(.) = g_n * f(.) = E f_n(.)$$

On obtient alors la généralisation suivante du lemme de Bochner [Bosq et Lecoutre (1987)] :

Théorème 5.1 : Soit $\bar{f}_n = g_n * f$, où g_n est la fonction (5.1) et f une densité de D_0; si $h_n \to 0$, alors $\Delta(f, \bar{f}_n) \to 0$.

En supposant maintenant pour simplifier que g est un noyau à variations bornées, on obtient :

Théorème 5.2 : Soit f_n l'estimateur à noyau (5.2) et $f \in D_0$; si $h_n \to 0$ et si $\forall \gamma > 0$,

$\sum_{n=1}^{\infty} \exp(-\gamma n h_n^2) < +\infty$ alors, presque complètement sûrement $\Delta(f_n, f) \to 0$.

Il suffit de remarquer que dans le résultat de Nadaraya (1981) la convergence uniforme de $(f_n - \bar{f}_n)$ ne dépend que de propriétés intrinsèques de la fonction de répartition empirique. On utilise alors la propriété (2.1.iv). Selon la remarque d'un rapporteur, que je remercie, le théorème 5.2 peut s'énoncer sous la condition plus simple, et à peine plus forte : $nh_n^2 / \text{Log } n \to \infty$.

6. DEMONSTRATIONS

Dans ce paragraphe, pour toute densité $f \in D_0$, nous noterons $\Gamma_f(x)$ la partie de Γ_f d'abscisse x ; cet ensemble est donc de la forme $\{x\} \times I_f(x)$; si $f(x^-) < f(x^+)$, par exemple, alors $I_f(x) = [f(x^-) ; f(x^+)]$. On notera C_f l'ensemble des points de continuité de f; si $x \in C_f$, $I_f(x) = \{f(x)\}$. Enfin, pour tout $\varepsilon > 0$, $\Gamma_f^{\varepsilon}(x)$ désignera le dilaté d'ordre ε de $\Gamma_f(x)$, c'est-à-dire, si $f(x^-) < f(x^+)$, la bande verticale $]x-\varepsilon ; x+\varepsilon[\times]f(x^-)-\varepsilon ; f(x^+)+\varepsilon[$.

Propriétés 2.1 :

(i) Tout élément de D_0 a au plus une infinité dénombrable de discontinuités ; or, si $x \in C_f \cap C_g$:

$$\{f(x)\} = I_f(x) = I_g(x) = \{g(x)\}$$

(ii) Comme Γ_f est fermé dans $\mathbb{R}\times\mathbb{R}^+$:

$$\Gamma_g \subset \bigcap_{\varepsilon>0} \Gamma_f^{\varepsilon} = \Gamma_f$$

D'autre part, si $x \in C_f$, cette inclusion implique :

$$I_g(x) \subset I_f(x) = \{f(x)\}$$

et nécessairement, $f(x) = g(x)$; enfin, si $x \notin C_f$, il existe $(x_n) \subset C_f$ et $(x'_n) \subset C_f$, telles que $(x_n) \uparrow x$ et $(x'_n) \downarrow x$. Par conséquent $f(x^-) = g(x^-)$, $f(x^+) = g(x^+)$ et $I_f(x) = I_g(x)$.

(iii) Soit $\delta_1 > \delta(f,h)$ et $\delta_2 > \delta(h,g)$; alors $\Gamma_h \subset \Gamma_f^{\delta_1}$ et $\Gamma_g \subset \Gamma_h^{\delta_2}$, donc

$$\Gamma_g \subset \Gamma_h^{\delta_2} \subset (\Gamma_f^{\delta_1})^{\delta_2} \subset \Gamma_f^{\delta_1+\delta_2}.$$

(iv) Supposons que $\| f-g \|_\infty \leq \varepsilon$;

si $x \in C_f \cap C_g$, alors $\forall\, \delta > 0$, il est clair que

$\Gamma_f(x) \subset \Gamma_g^{\varepsilon+\delta}(x)$ et $\Gamma_g(x) \subset \Gamma_f^{\varepsilon+\delta}(x)$.

sinon, on obtient quand même $| f(x^+) - g(x^+) | \leq \varepsilon$ et $| f(x^-) - g(x^-) | \leq \varepsilon$ en considérant des suites (x_n) et (x'_n) comme dans (ii) : si bien qu'à nouveau, $\forall\, \delta > 0, \Gamma_f(x) \subset \Gamma_g^{\varepsilon+\delta}(x)$ et $\Gamma_g(x) \subset \Gamma_f^{\varepsilon+\delta}(x)$.

(v) Soit $0 < \eta \leq \varepsilon$ tels que : $\forall\, x', x'' \in \mathbb{R}$:

$$|x' - x''| < 2\eta \Rightarrow |f(x')-f(x'')| < \varepsilon.$$

Pour n assez grand, $\delta(f,g) < \eta$; alors, pour tout $x \in C_{g_n}$ il existe $z_n(x)$ tel que $|x - z_n(x)| < \eta$ et $|g_n(x)-f(z_n(x))| < \eta$.

$$\sup_{x \in \mathbb{R}} \left| f(x) - g_n(x) \right| = \sup_{x \in C_{g_n}} \left| f(x) - g_n(x) \right|$$

$$\leq \sup_{x \in C_{g_n}} \sup_{x-\eta \leq x' \leq x+\eta} \max\left(|x-x'| \, ; \, \left| g_n(x) - f(x') \right| \right)$$

$$\leq \sup_{x \in C_{g_n}} \sup_{x\ \eta \leq x' \leq x+\eta} \max\Big(|x - z_n(x)| + |z_n(x) - x'| \, ; \left| g_n(x) - f(z_n(x)) \right| + \left| f(z_n(x)) - f(x') \right| \Big)$$

$$\leq \max(\eta + 2\eta \, ; \eta + \varepsilon) \leq 3\varepsilon$$

Contre-exemple 2.1 :
Quand $n = 4p + 2$, $p \in \mathbb{N}^*$, il existe un $\Delta_{i,n} = [\, i / n\ ;\ (i+1) / n\, [$ centré sur 1/4, soit $\Delta_n^{1/4}$ sur lequel $\bar{f}_n = 1$. D'autre part on peut trouver, pour tout n, un nombre t_n de l'intervalle [0 , 1] tel que $\lambda_n(t_n) \in \Delta_n^{1/4}$. La suite $f_n(\lambda_n(t_n))$ possède donc 1 pour valeur d'accumulation.

Ce contre-exemple peut convenir aussi pour l'estimateur à noyau (5.2) : si g est symétrique

$$f * g_n(1/4) = 2 \int_{-1/2h_n}^{0} g(u)\, du \to 1$$

et pour tout n, il existe t_n tel que $\lambda_n(t_n) = \frac{1}{4}$.

Propriétés 2.2 :
(i) supposons que $f(x^-) < +\infty$ et $f(x^+) = +\infty$; pour tout $M > 0$, il existe $\eta > 0$ tel que $f(t) > M+\eta$, $\forall\ t \in]\, x, x+2\eta\, [$. M et η étant fixés, pour n assez grand, Γ_{g_n} est inclus dans $\Gamma_f^{\eta/4}$. On en déduit que :

$$\Gamma_{g_n} \cap \left(\left[x + \frac{1}{2}\eta\ ;\ x + \frac{3}{2}\eta \right] \times [\, 0\, ,\, M\,] \right) = \varnothing$$

Soit x_n un point de continuité de g_n pris dans $\left[x + \frac{1}{2}\eta\ ;\ x + \frac{3}{2}\eta \right]$:

$$g_n(x_n) > M$$

(ii) si $f(x^-) = f(x^+) = +\infty$, on peut faitre une démonstration similaire, en remplaçant $]x\ ;\ x+2\eta[$ par $]x-\eta\ ;\ x+\eta[$ et en posant $x_n = x$; aboutissant ainsi à la conclusion :

$$g_n(x^+) \text{ et } g_n(x^-) > M.$$

Théorème 3.1
a) Pour simplifier cette démonstration, nous supposerons que f et $\bar{f}_n$ sont continues à droite. Comme cela ne modifie pas Γ_f et $\Gamma_{\bar{f}_n}$, la solution n'en est pas moins générale. [§7, fig. 2].

Soit ε un réel positif fixé , $a' = \inf \{x : f(x) > \varepsilon\}$ et $a'' = \sup \{x : f(x) > \varepsilon\}$. L'intervalle $[a',a''[$ peut être partagé en sous-intervalles $[a_j\ ;\ a_{j+1}[$, $j = 1,\ldots,j_\varepsilon-1$, sur

lesquels f a une oscillation plus petite que ε [Billingsley (1968)]. On pose par conséquent $a' = a_1$, $a'' = a_{j_\varepsilon}$ et, par convention, $a_0 = -\infty$ et $a_{j_\varepsilon+1} = +\infty$.

Dans la suite, n est pris assez grand pour que $h_n < \varepsilon$ et pour que chaque intervalle $[a_j ; a_{j+1}[$ contienne au moins un intervalle $\Delta_{i,n}$ (1.1).
On notera $[\, b_j^n ; c_j^n \,[$ la réunion des $\Delta_{i,n}$ contenus dans chaque $[\, a_j ; a_{j+1} \,[$. Par convention $b_0^n = -\infty$ et $c_{j_\varepsilon}^n = +\infty$.

b) Soit $x \in \Delta_{i,n} \subset [a_j ; a_{j+1}[$:

$$\begin{aligned} \left| \bar{f}_n(x) - f(x) \right| &= \left| \frac{1}{\lambda \Delta_{i,n}} \int_{\Delta_{i,n}} f(y)\, dy - f(x) \right| \\ &\le \frac{1}{\lambda \Delta_{i,n}} \int_{\Delta_{i,n}} | f(y) - f(x) | dy \\ &\le w_f [a_j ; a_{j+}[< \varepsilon \end{aligned}$$

Par conséquent, $\Gamma_f(x) \subset \Gamma^{\varepsilon}_{\bar{f}_n}(x)$ et $\Gamma_{\bar{f}_n}(x) \subset \Gamma^{\varepsilon}_f(x)$.

c) Considérons à présent un $\Delta_{i,n}$ tel que a_j appartienne à l'intérieur de $\Delta_{i,n}$, et supposons pour fixer les idées que $f(a_j^-) \le f(a_j)$.

$$\begin{aligned} \frac{1}{h_n} P\left[c_{j\,1}^n ; b_j^n \right[&= \frac{1}{h_n} \int_{a_j}^{b_j^n} f(x)\, dx + \frac{1}{h_n} \int_{c_{j-1}^n}^{a_j} f(x)\, dx \\ &\le \left(f(a_j) + w_f \left( ] a_j ; b_j^n [\right) \right) \left(b_j^n - a_j \right) / h_n \\ &\quad + \left(f(a_j^-) + w_f \left(\left[c_{j-1}^n ; a_j \right[\right) \right) \left(a_j - c_{j\,1}^n \right) / h_n \\ &< f(a_j) + \varepsilon \end{aligned}$$

Et de la même façon :

$$\frac{1}{h_n} P\left[c_{j\,1}^n ; b_j^n \right[> f(a_j) - \varepsilon$$

Par conséquent, $\forall\, j = 1, \ldots, j_\varepsilon$, $\exists\ (a_j, v_j) \in \Gamma_f$ tel que,
pour tout $x \in]\, c_{j-1}^n ; b_j^n \,[$: $| a_j - c_{j-1}^n | < h_n < \varepsilon$ et $| v_j - f_n(x) | < \varepsilon$.

Donc $\Gamma_{\bar{f}_n}(x) \subset \Gamma^{\varepsilon}_f(a_j)$ pour $x \in]c_{j-1}^n ; b_j^n[$.

d) Inversement, pour $x \in [c_{j-1}^n ; b_j^n[- \{a_j\}$:

$$\Gamma_f(x) \subset \Gamma^{\varepsilon}_{\bar{f}_n}\left(c_{j-1}^n\right) \cup \Gamma^{\varepsilon}_{\bar{f}_n}\left(b_j^n\right)$$

En effet :

$$\sup_{\substack{a_j \le x < b_j^n \\ b_j^n \le y \le c_j^n}} |f(x) - f(y)| < \varepsilon \Rightarrow \forall\, x \in \left[a_j, b_j^n\right[,$$

$$\left| f(x) - \bar{f}_n\left(b_j^n\right)\right| \le \frac{1}{h_n}\int_{b_j^n}^{b_j^n + h_n} |f(x) - f(y)|\, dy < \varepsilon$$

et de la même façon :

$$\forall\ x \in \left[c_{j-1}^n , a_j\right[\ : \ \left| f(x) - \bar{f}_n\left(c_{j\,1}^n\right)\right| < \varepsilon$$

e) D'après c) et d) :

$$f(a_j) + \varepsilon > \bar{f}_n\left(c_{j-1}^n\right) = \bar{f}_n\left(b_j^n\right) > f\left(a_j^-\right) - \varepsilon$$
$$f(a_j) + \varepsilon > \bar{f}_n\left(b_j^n\right) > f(a_j) - \varepsilon$$
$$f\left(a_j^-\right) + \varepsilon \ge \bar{f}_n\left(c_{j-1}^{n-}\right) \ge f\left(a_j^-\right) - \varepsilon$$

On en déduit :

$$\Gamma_f(a_j) \subset \Gamma^{\varepsilon}_{\bar{f}_n}\left(c_{j-1}^n\right) \cup \Gamma^{\varepsilon}_{\bar{f}_n}\left(b_j^n\right)$$

$$\text{et} \quad \Gamma_{\bar{f}_n}\left(c_{j\,1}^n\right) \subset \Gamma^{\varepsilon}_f(a_j) \ ; \ \Gamma_{\bar{f}_n}\left(b_j^n\right) \subset \Gamma^{\varepsilon}_f(a_j)$$

Théorème 3.2

La démonstration du théorème (1.4), que l'on pourra consulter dans [Bosq, Lecoutre (1987)], consiste à démontrer séparément que $\|f_n - \bar{f}_n\|_\infty \to 0$ presque complètement sûrement, et n'utilise pas l'hypothèse d'uniforme continuité de f ; on a donc aussi $\Delta\left(f_n, \bar{f}_n\right) \to 0$ presque complètement sûrement d'après la propriété (2.1.iv). Il reste à utiliser le théorème (3.1) et l'inégalité triangulaire.

Théorème 3.3

On suppose qu'il existe un point x unique tel que $f(x^-) < +\infty$ et $f(x^+) = +\infty$; il existe $\varepsilon < 0$ tel que :

$$f(y) > f(x^-) - \varepsilon > 0 \ , \quad \forall\, y \in]x - \varepsilon\ ;\ x + \varepsilon[$$

ε peut être supposé suffisamment petit pour que $[x - \varepsilon\ ;\ x + \varepsilon]$ soit à l'intérieur de l'intervalle $[a_1, a_{j_\varepsilon}[$ hors duquel $f < \varepsilon$; on partage $[a_1\ ,\ x - \varepsilon/4]$ et $[x + \varepsilon/4\ ,\ a_{j_\varepsilon}[$ en sous-intervalles sur lesquels l'oscillation de f est inférieure à ε ; on prend alors h_n assez petit pour que chacun de ces sous-intervalles contienne au moins un $\Delta_{i,n}$, et on suppose de toute façon que $h_n < \varepsilon/8$.

Comme $[x - 3\varepsilon/4\ ;\ x + 3\varepsilon/4]$ est contenu dans la réunion des $\Delta_{i,n}$ inclus dans l'intervalle $[x - \varepsilon\ ;\ x + \varepsilon]$, pour tout $y \in [x - 3\varepsilon/4\ ;\ x + 3\varepsilon/4]$ on a

$$\bar{f}_n(y) > f(x^-) - \varepsilon$$

$$\text{et} \quad \Gamma_{\bar{f}_n}(y) \subset \Gamma_f^{\varepsilon}(x)$$

puisque $\Gamma_f(x) = \{x\} \times]f(x^-)\ ;\ +\infty[$

D'autre part, le $\Delta_{i,n}$ contenant $x+\varepsilon/4$ est suivi d'au moins $\Delta_{i+1,n}$, $\Delta_{i+2,n}$, $\Delta_{i+3,n}$ contenus dans $[x + \varepsilon/4\ ;\ x + 3\varepsilon/4[$ pour lesquels on est ramené à la situation décrite dans la démonstration du théorème 3.1. ; c'est-à-dire que f n'a en aucun cas une oscillation supérieure à ε sur deux intervalles consécutifs de ce type.

Il en est de même sur $[x - 3\varepsilon/4\ ;\ x]$.

Théorème 3.4

C'est une conséquence du théorème (3.3).

Corollaire 3.5

C'est une conséquence de la propriété (2.2).

Théorème 4.1

Si $w'_f(2h_n) < \varepsilon$, on peut aisément se ramener à la situation du a) de la démonstration du théorème 3.1., à ceci près qu'il ne sera pas supposé que $h_n < \varepsilon$. Une démonstration similaire à celle du théorème 3.1 conduit à :

b) $\quad \Gamma_f(x) \subset \Gamma_{\bar{f}_n}^{0,\varepsilon}(x)\ ;\ \Gamma_{\bar{f}_n}(x) \subset \Gamma_f^{0,\varepsilon}(x)$

c) $\Gamma_{\bar{f}_n}(x) \subset \Gamma_f^{h_n,\varepsilon}(a_j)$

d) $\Gamma_f(x) \subset \Gamma_{\bar{f}_n}^{h_n,\varepsilon}\left(c_{j\,1}^n\right) \cup \Gamma_{\bar{f}_n}^{h_n,\varepsilon}\left(b_j^n\right)$

e) $\Gamma_f(a_j) \subset \Gamma_{\bar{f}_n}^{h_n,\varepsilon}\left(c_{j\,1}^n\right) \cup \Gamma_{\bar{f}_n}^{h_n,\varepsilon}\left(b_j^n\right)$

et $\Gamma_{\bar{f}_n}\left(c_{j\,1}^n\right) \subset \Gamma_f^{h_n,\varepsilon}(a_j)$; $\Gamma_{\bar{f}_n}\left(b_j^n\right) \subset \Gamma_f^{h_n,\varepsilon}(a_j)$

Théorème 5.1

a) Supposons d'abord que f soit une densité de la forme :

$$f(x) = \sum_{i=1}^{k} c_i \mathbb{1}_{[a_i, a_{i+1}[}(x)$$

$$\text{et} \quad f * g_n(x) = \sum_{i=1}^{k} c_i \int_{(x-a_{i+1})h_n^{-1}}^{(x-a_i)h_n^{-1}} g(u)\,du$$

Soit $0 < 2\varepsilon < \min_{1\le i\le k} |a_{i+1} - a_i|$ et $j \in \{1,\ldots,k\}$;

pour tout $x \in \left]a_j + \varepsilon\,;\, a_{j+1} - \varepsilon\right[$:

$$|f(x) - f * g_n(x)| \le \left| C_j - C_j \int_{(x-a_{j+1})h_n^{-1}}^{(x\,a_j)h_n^{-1}} g(u)\,du \right| + \sum_{i\ne j} |C_i| \int_{(x-a_{i+1})h_n^{-1}}^{(x\,a_i)h_n^{1}} g(u)\,du$$

$$\le \sum_{i=1}^{k} |C_i| \left(1 - \int_{-\varepsilon h_n^{-1}}^{\varepsilon h_n^{1}} g(u)\,du\right)$$

il existe donc n_0 tel que, $\forall\, n \ge n_0$, $\forall\, x \in \bigcup_{j=1}^{k}]a_j + \varepsilon\,;\, a_{j+1} - \varepsilon[$:

$$|f(x) - f * g_n(x)| < \varepsilon$$

D'autre part, pour tout $x \in [a_j - \varepsilon\,;\, a_j + \varepsilon]$, $j \in \{1,\ldots,k\}$: il existe $n_1 \ge n_0$, tel que pour $n \ge n_1$:

$$\left| f * g_n((x)) - \left[C_{j-1} \int_{(x\ a_j)\,h_n^{-1}}^{+\infty} g(u)\,du + C_j \int_{-\infty}^{(x-a_j)\,h_n^{-1}} g(u)\,du \right] \right| < \varepsilon$$

donc, en supposant que $C_{j-1} \leq C_j$:

$$C_{j-1} - \varepsilon < f * g_n(x) < C_j + \varepsilon$$

Il suffit alors de remarquer que $f * g_n$ est uniformément continue, puisque g est bornée, pour conclure : $\forall\ n \geq n_1$

$$\Gamma_{f * g_n} \subset \Gamma_f^{\varepsilon} \qquad [\S\ 6\ ;\ \text{fig. 3}]$$

b) Inversement, pour $n \geq n_0, \Gamma_f \subset \Gamma_{f * g_n}^{\varepsilon}$: en effet,

si $a_j + \varepsilon \leq x \leq a_{j+1} - \varepsilon$, alors $| f(x) - f * g_n(x) | < \varepsilon$.

D'autre part, si $a_j \leq x \leq a_j + \varepsilon$:

$$C_j - \varepsilon = f(a_j + \varepsilon) - \varepsilon \leq f * g_n(a_j + \varepsilon) \leq f(a_j + \varepsilon) + \varepsilon = C_j + \varepsilon$$

Tandis que si $a_j - \varepsilon \leq x \leq a_j$

$$C_{j-1} - \varepsilon = f(a_j - \varepsilon) - \varepsilon \leq f * g_n(a_j - \varepsilon) \leq f(a_j - \varepsilon) + \varepsilon = C_{j-1} + \varepsilon$$

On peut conclure en utilisant à nouveau l'uniforme continuité de $f * g_n$, qui prend toutes les valeurs entre $f * g_n(a_j+\varepsilon)$ et $f * g_n(a_j-\varepsilon)$.

c) D'après le a) de la démonstration du théorème (3.1), il existe, si $f \in D_0$, une densité h en escalier telle que $\| f-h \|_\infty < \varepsilon$; et donc :

$$\forall\ n,\ \| f * g_n - h * g_n \|_\infty \leq \| f - h \|_\infty < \varepsilon$$

On en déduit, d'après la propriété (2.1.iv), que $\forall\ n \geq n_1$:

$$\Gamma_{f * g_n} \subset \Gamma_{h * g_n}^{\varepsilon} \subset \Gamma_h^{2\varepsilon} \subset \Gamma_f^{3\varepsilon}$$

$$\Gamma_f \subset \Gamma_h^{\varepsilon} \subset \Gamma_{h * g_n}^{2\varepsilon} \subset \Gamma_{f * g_n}^{3\varepsilon}$$

donc, $\forall\ n \geq n_1 : \Delta(f * g_n , f) < 3\varepsilon$.

Théorème 5.2

L'argument est identique à celui du théorème (3.2).

7. FIGURES

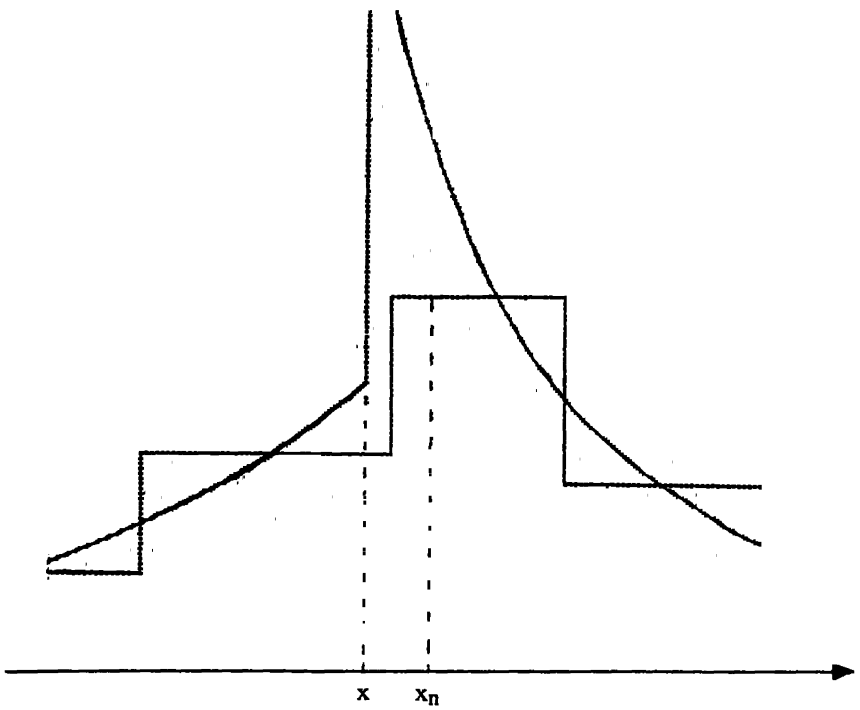

Figure 1

$f(x^-) < +\infty$ et $f(x^+) = +\infty$; Γ_f^ε est la zone grise ; le graphe complet de g_n est représenté sous la forme d'un histogramme.

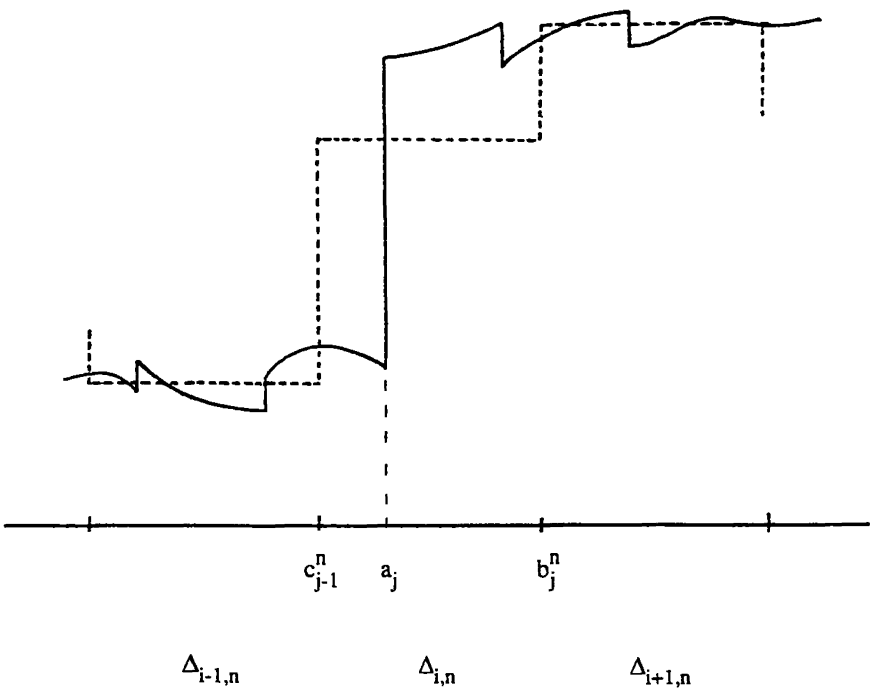

Figure 2

Γ_f est en traits pleins et $\Gamma_{\bar{f}_n}$ en pointillés ; on voit sur $\Delta_{i,n}$ la nécessité de considérer des "dilatations horizontales" des graphes, alors que sur $\Delta_{i+1,n}$ et $\Delta_{i-1,n}$, des "dilatations verticales" suffisent.

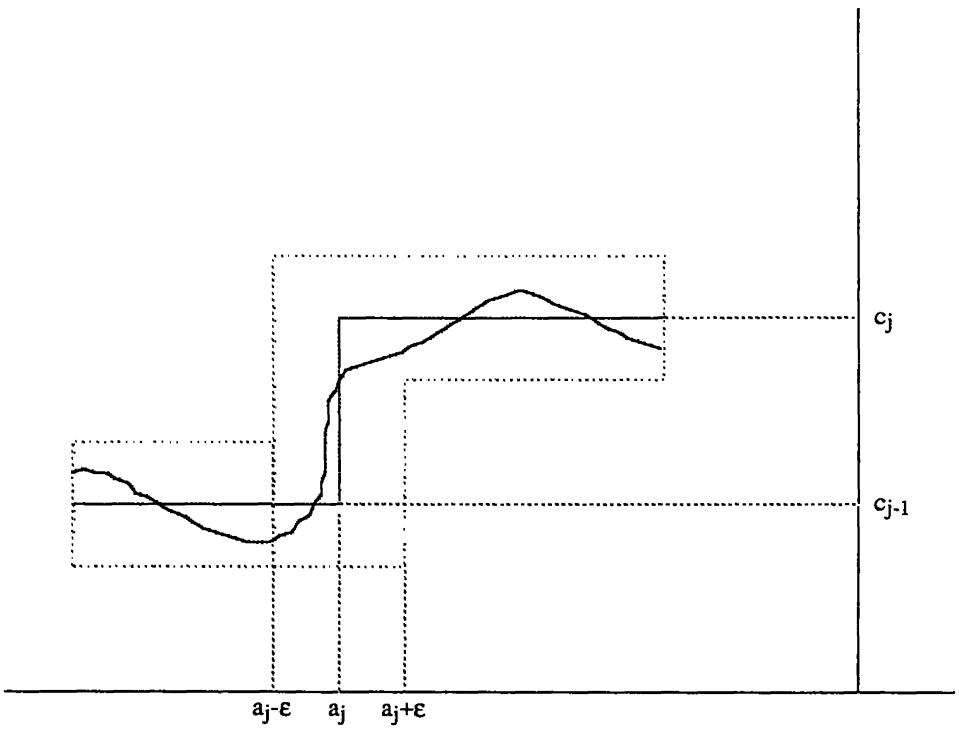

Figure 3

$\Gamma_{f * g_n}$ est contenu dans la zone grise, qui représente Γ_f^{ε}.

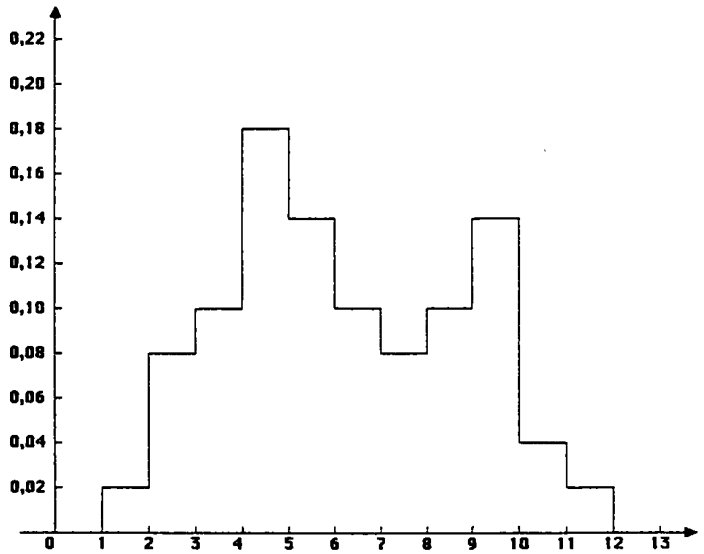

Figure 4 a)

Le graphe Γ_{f_n} d'un certain histogramme.

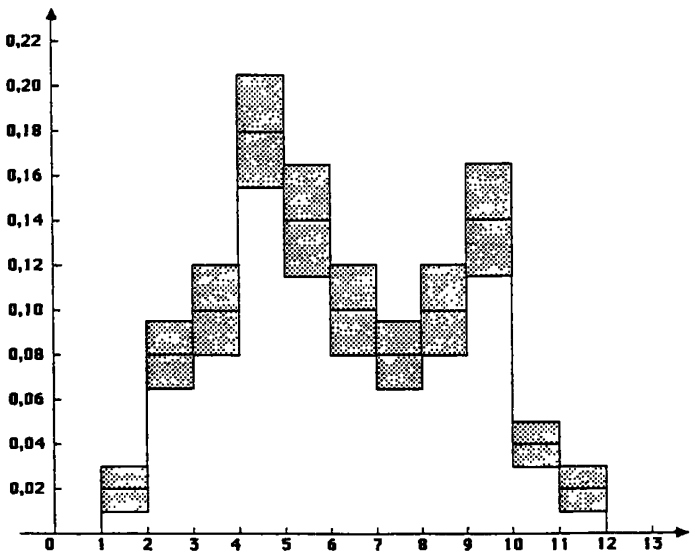

Figure 4 b)

Des zones de confiance pour $\Gamma_{\tilde{f}_n}$ (et non pour Γ_f !), construites autour de Γ_{f_n}, au seuil 5 % sur chaque cellule, avec un échantillon de taille 900.

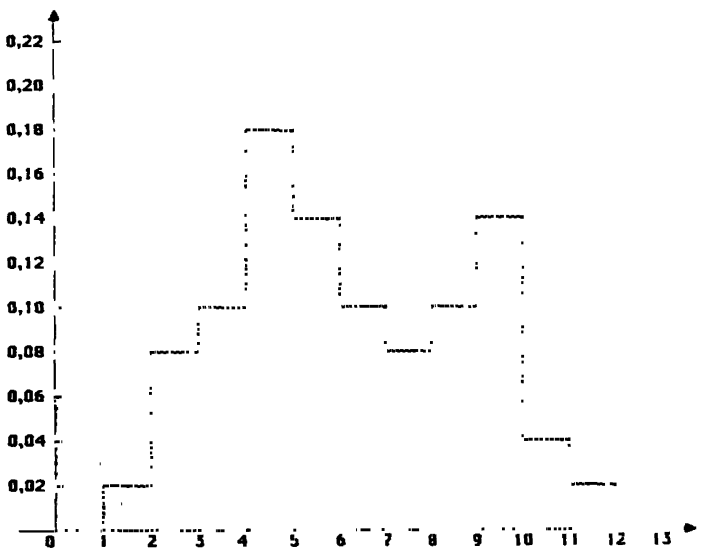

Figure 4 c)

Γ_{f_n} étant pris comme estimateur de $\Gamma_{\bar{f}_n}$, c'est dans la zone grise que peut être tracé Γ_f , si f appartient à une famille de densités A telle que : $\sup_{g \in A} w'_g(2) \leq 0{,}04$.

Bien entendu, il faut prendre garde de tracer le graphe d'une densité de probabilité dont la valeur moyenne sur chaque cellule soit égale à la valeur de $\bar{f}_n$ sur cette cellule.

Références

Abou Jaoudé, S. Sur une C.N.S. de L^1 convergence presque complète de l'estimateur de la partition fixe pour une densité. *C.R.A.S. A*, 283, pp.1107-1110, 1976.

Bera, M. Lois limites des écarts extrêmes associés aux histogrammes et à diverses statistiques d'ordre dans l'estimation d'une densité de probabilité. Thèse de 3ème cycle, Paris,Université Pierre et Marie Curie, 1977.

Billingsley, P *Convergence of Probability Measures*, Wiley, New York, 1968.

Bosq, D. et Lecoutre, J.P. *Théorie de l'estimation fonctionnelle*, Economica, Paris, 1987.

Devroye, L. et Györfi, L. *Nonparametric density estimation : The L^1 view* Wiley, New York,1985.

Freedman, D. et Diaconis, P. On the histogram as a density estimator : L^2-Theory, *Z. W,*.57, pp.453-476, 1981.

Geffroy, J. Sur l'estimation d'une densité dans un espace métrique. *C.R.A.S.* , A, 278, pp.1449-1452, 1974.

Nadaraya, E. On nonparametric estimation of density function and regression. *Theory Prob. Appl.*, 10, pp.186-190, 1981.