Studia Scientiarum Mathematicarum Hungarica 8 (1973) 347—361.

ON A NON-PARAMETIC ESTIMATION OF THE REGRESSION FUNCTION

by

P. MAJOR

The regression function is usually estimated by a polynomial. In this paper an other method is investigated, similar to one of the methods of density estimation.

It is the following method: Let (ξ_i, η_i) $i=1, 2, \dots$ be independent identically distributed two-dimensional random variables, $\mathsf{P}(0 \leqq \eta_i \leqq 1)=1$. Let us divide the interval [0, 1] into a_n equal parts. We estimate the function $R(x)=\mathsf{E}(\xi|\eta=x)$ by the following $R_n(x)$

$$(1) \qquad R_n(x) = \frac{\displaystyle\sum_{\substack{\eta_i \in \left[\frac{k}{a_n}, \frac{k+1}{a_n}\right] \\ (i \leqq n)}} \xi_i}{\text{the number of } \eta_i - \text{s } (i \leqq n) \text{ falling into the interval } \left[\dfrac{k}{a_n}, \dfrac{k+1}{a_n}\right]}$$

$$x \in \left[\frac{k}{a_n}, \frac{k+1}{a_n}\right] \qquad (k = 1, 2, \dots, a_n).$$

We investigate the order of magnitude of $R_n(x)-R(x)$ in supremum norm. We introduce some notations. $f(x)$ is the density function of η_i,

$$\sigma^2(x) = \mathsf{E}([\xi_i - \mathsf{E}(\xi_i|\eta_i)]^2|\eta_i = x).$$

Our first statement is the following

THEOREM 1. *Let us suppose that $f(x)$, $\sigma(x)$, $R(x)$ are differentiable, $|f'(x)| \leqq K$, $|\sigma'(x)| \leqq K$, $|R'(x)| \leqq K$ and $f(x) > K'$, $\sigma(x) > K'$ for appropriate $K, K' > 0$ and for every x. Let us suppose further that $\mathsf{E}(e^{t_0|\xi|}|\eta = x) \leqq c$ for appropriate $t_0 > 0$ and c for every $0 \leqq x \leqq 1$. If $n^\alpha < a_n < n^\beta$, $\frac{1}{3} < \alpha < \beta < 1$ then*

$$(2) \qquad \mathsf{P}\left(\sup_x \sqrt{\frac{nf(x)}{a_n}} \cdot \frac{|R_n(x)-R(x)|}{\sigma(x)} < \sqrt{2\log a_n - \log\log a_n + y}\right) \to \exp\left(-\frac{e^{-y/2}}{\sqrt{\pi}}\right)$$

$$(3) \qquad \mathsf{P}\left(\sup_x \sqrt{\frac{nf(x)}{a_n}} \frac{R_n(x)-R(x)}{\sigma(x)} < \sqrt{2\log a_n - \log\log a_n + y}\right) \to \exp\left(\frac{-e^{-y/2}}{2\sqrt{\pi}}\right)$$

as $n \to \infty$.

If the functions $f(x)$, $R(x)$, $\sigma(x)$ are smooth enough, it is worth substituting the estimation $R_n(x)$ by

$$\bar{R}_n(x) = R_n\left(\frac{k-\frac{1}{2}}{a_n}\right) + \left[R_n\left(\frac{k+\frac{1}{2}}{a_n}\right) - R_n\left(\frac{k-\frac{1}{2}}{a_n}\right)\right]\left(a_n x - k - \frac{1}{2}\right) \tag{4}$$

for

$$k-\frac{1}{2} < a_n x < k+\frac{1}{2}.$$

Then we have our.

THEOREM 1′. *If the conditions of Theorem 1 are fulfilled and, in addition $f(x)$, $R(x)$ and $\sigma(x)$ are differentiable twice with bounded second derivatives, then the relations* (2) *and* (3) *hold for* $n^{\alpha}<a_n<n^{\beta}$, $\frac{1}{5}<\alpha<\beta<1$, *if we substitute* $R_n(x)$ *by* $\bar{R}_n(x)$.

Now if we want to construct a "confidence strip" i.e. we want to construct a region T in the plane such that $\mathsf{P}(R(x)\in T)>1-\varepsilon$ with a prescribed ε, then we are interested in whether $f(x)$ and $\sigma(x)$ in the formulae (2) and (3) can be substituted by their estimations. The answer is in the affirmative.

THEOREM 2. *Let*

$$f_n(x) = \frac{a_n}{n}\left\{\text{the number of } \eta_i - \text{s } (i \leqq n) \text{ for wich } \eta_i \in \left[\frac{k}{a_n}, \frac{k+1}{a_n}\right]\right\}$$

$$\sigma_n^2(x) = \frac{a_n}{nf_n(x)} \sum_{\substack{\eta_i \in \left[\frac{k}{a_n}, \frac{k+1}{a_n}\right] \\ i \leqq n}} \xi_i^2 - \left[\frac{a_n}{nf_n(x)} \sum_{\substack{\eta_i \in \left[\frac{k}{a_n}, \frac{k+1}{a_n}\right] \\ i \leqq n}} \xi_i\right]^2$$

if the condition of Theorem 1 holds

$$x \in \left[\frac{k}{a_n}, \frac{k+1}{a_n}\right] \qquad k = 0, 1, \dots, a_n - 1.$$

Then

$$\mathsf{P}\left(\sup_x \sqrt{\frac{nf_n(x)}{a_n}} \cdot \frac{|R_n(x)-R(x)|}{\sigma_n(x)} < \sqrt{2\log a_n - \log\log a_n + y}\right) \to \exp\left(\frac{-e^{-y/2}}{\sqrt{\pi}}\right) \tag{2′}$$

and

$$\mathsf{P}\left(\sup_x \sqrt{\frac{nf_n(x)}{a_n}} \cdot \frac{R_n(x)-R(x)}{\sigma_n(x)} < \sqrt{2\log a_n - \log\log a_n + y}\right) \to \exp\left(\frac{-e^{-y/2}}{2\sqrt{\pi}}\right) \tag{3′}$$

if.

As an immediate consequence of Theorem 1 we get that

$$\sup_x \sqrt{\frac{nf(x)}{2a_n \log a_n}} \cdot \frac{|R_n(x)-R(x)|}{\sigma(x)} \Rightarrow 1$$

where the symbol $\Rightarrow$ means convergence in probability. The question arises whether this expression is also convergent with probability 1.

THEOREM 3. *Under the conditions of theorem* 1 *we have*

$$\mathsf{P}\left(\limsup_{n\to\infty}\sup_x \sqrt{\frac{nf(x)}{2a_n\log a_n}}\,\frac{|R_n(x)-R(x)|}{\sigma(x)}=1\right)=1$$

$$\mathsf{P}\left(\limsup_{n\to\infty}\sup_x \sqrt{\frac{nf(x)}{2a_n\log a_n}}\,\frac{R_n(x)-R(x)}{\sigma(x)}=1\right)=1$$

if

$$n^{\alpha}<a_n<n^{\beta},\quad \frac{1}{3}<\alpha<\beta<1.$$

For proving our theorems we need some lemmas:

LEMMA 1. *Let* $\xi_{n1}, \xi_{n2}\dots\xi_{nn}$, $n=1,2\dots$ *be a double array of random variables independent and identically distributed within a row. Let* $\mathsf{E}\xi_{n1}=0$, $\mathsf{D}^2\xi_{n1}=1$, $\mathsf{E}e^{t|\xi_{n1}|}<c$ *for appropriate* $t>0$ *and* c. *Then*

$$\mathsf{P}(\xi_{n1}+\dots+\xi_{nn}>\sqrt{n}\cdot x_n)\sim 1-\Phi(x_n)\quad for\quad x_n=o\left(\sqrt[6]{n}\right)$$

where $\Phi(x)$ *is the standard normal distribution function.*

The proof is essentially the same as in [2] page 517, and we omit it.

LEMMA 2. *Let* $(\xi_i^{(n)}, \eta_i^{(n)})$ *be pairs of random variables independent and identically distributed for fixed* n,

$$\mathsf{P}(\eta_i^{(n)}=k)=p_k^{(n)},\qquad k=1,2,\dots,a_n,$$

$$\sum_{k=1}^{a_n}p_k^{(n)}=1,\quad n^{\varepsilon}<a_n<n^{1-\varepsilon},\qquad \varepsilon>0,$$

$$\mathsf{E}(\xi_i^{(n)}|\eta_i^{(n)}=k)=0,\quad \mathsf{E}(\xi_i^{(n)2}|\eta_i^{(n)}=k)=\sigma_k^{(n)2},\quad \frac{c_1}{a_n}<p_k^{(n)}<\frac{c_2}{a_n}$$

$$\mathsf{E}(e^{t|\xi_i^{(n)}|}|\eta_i^{(n)})\leq c,\quad \sigma_k^{(n)}>c'$$

for appropriate c_1, c_2, c, $t>0$, c' *for every* k *and* n. *Then*

$$\mathsf{P}\left(\max_k\frac{\sqrt{np_k^{(n)}}}{\sigma_k^{(n)}}\cdot\frac{\sum_{i=1}^{n}\xi_i^{(n)}\cdot I_{(\eta_i^{(n)}=k)}}{\sum_{i=1}^{n}I_{(\eta_i^{(n)}=k)}}<\sqrt{2\log a_n-\log\log a_n+y}\right)\to\exp\left(\frac{-e^{-y/2}}{2\sqrt{\pi}}\right)$$

and

$$\mathsf{P}\left(\max_k \frac{\sqrt{np_k^{(n)}}}{\sigma_k^{(n)}}\cdot\left|\frac{\sum_{i=1}^n \xi_i^{(n)}\cdot I_{(\eta_i^{(n)}=k)}}{\sum_{i=1}^n I_{(\eta_i^{(n)}=k)}}\right| < \sqrt{2\log a_n-\log\log a_n+y}\right) \to \exp\left(\frac{-e^{-y/2}}{\sqrt{\pi}}\right)$$

where $I_{(\eta_i^{(n)}=k)}=1$ *if* $\eta_i^{(n)}=k$ *and* 0 *otherwise.*

PROOF OF LEMMA 2. In the following we omit the superscript n. Let us introduce the notations $s_j=\sum_{i=1}^n \xi_i I_{(\eta_i=j)}$ and $v_j=\sum_{i=1}^n I_{(\eta_i=j)}$. Then

$$\mathsf{P}(s_1<x_1,\dots,s_{a_n}<x_{a_n}|v_1=t_1\dots,v_{a_n}=t_{a_n})=$$

$$=\mathsf{P}(s_1<x_1|v_1=t_1)\dots\mathsf{P}(s_{a_n}<x_{a_n}|v_{a_n}=t_{a_n})$$

Let $\tilde{\xi}_{ij}$, $i=1,2\dots j=1,2\dots a_n$ be — for fixed j — independent identically distributed random variables with distribution function $\mathsf{P}(\tilde{\xi}_{ij}<x)=\mathsf{P}(\xi_1=x_1|\eta_1=j)$. Then

$$\mathsf{P}(s_j<x_j|v_j=t_j)=\mathsf{P}(\tilde{\xi}_{1j}+\dots+\tilde{\xi}_{tj,j}<x_j)$$

Now we prove that

$$\mathsf{P}(|v-np_j|>\sqrt{np_j}\,n^{\frac{1-\varepsilon}{3}})\leqq 2\exp\left(-\frac{n^{\frac{2}{3}(1-\varepsilon)}}{3}\right) \tag{5}$$

Really

$$\mathsf{P}(v_j-np_j>\sqrt{np_j}\,n^{\frac{1-\varepsilon}{3}})=\mathsf{P}\left(\sum_{i=1}^n[I_{(\eta_i=j)}-p_j]>\sqrt{np_j}\,n^{\frac{1-\varepsilon}{3}}\right)=$$

$$=\mathsf{P}\left(\exp\left(n^{\frac{1-\varepsilon}{3}}\sum_{i=1}^n\frac{I_{(\eta_i=j)}-p_j}{\sqrt{np_j}}\right)>\exp(n^{\frac{2}{3}(1-\varepsilon)})\right)\leqq$$

$$\leqq\frac{\left(\mathsf{E}\left[\exp\left(n^{\frac{1-\varepsilon}{3}}\frac{I_{(\eta_i=j)}-p_j}{\sqrt{np_j}}\right)\right]\right)^n}{\exp(n^{\frac{2}{3}(1-\varepsilon)})}\leqq\frac{\left[1+\frac{2}{3}n^{\frac{2}{3}(1-\varepsilon)}\mathsf{E}\left(\frac{I_{(\eta_1=j)}-p_j}{\sqrt{np_j}}\right)^2\right]^n}{\exp(n^{\frac{2}{3}(1-\varepsilon)})}\leqq$$

$$\leqq\frac{\exp\left[\frac{2}{3}(1-p_j)n^{\frac{2}{3}(1-\varepsilon)}\right]}{\exp\quad(n^{\frac{2}{3}(1-\varepsilon)})}\leqq\exp\left(-\frac{n^{\frac{2}{3}(1-\varepsilon)}}{3}\right)$$

and similarly

$$\mathsf{P}(v_j-np_j<-\sqrt{np_j}\,n^{\frac{1-\varepsilon}{3}})=\exp\left(-\frac{n^{\frac{2}{3}(1-\varepsilon)}}{3}\right).$$

Thus (5) is valid.

Now, by lemma 1 and by the relation $\frac{1}{\sqrt{2\pi}\,x} e^{-x^2/2} \sim 1-\Phi(x)$ for $x\to\infty$ we have

$$\mathsf{P}\left(\frac{\sqrt{np_j}}{\sigma_j}\cdot\frac{s_j}{v_j} > \sqrt{2\log a_n - \log\log a_n + y}\,\Big|\, v_j = t\right) =$$

$$= \mathsf{P}\left(\frac{\tilde{\xi}_{1j}+\ldots+\tilde{\xi}_{tj}}{\sigma_j\sqrt{t}} > \sqrt{\frac{t}{np_j}}\sqrt{2\log a_n - \log\log a_n + y}\right) \sim$$

$$\sim \sqrt{\frac{np_j}{2\pi t}}\cdot\frac{1}{\sqrt{2\log a_n - \log\log a_n + y}}\cdot e^{-\frac{np_j}{2t}(2\log a_n - \log\log a_n + y)} \sim \frac{e^{-y/2}}{2\sqrt{\pi}\,a_n}$$

if

$$|t-np_j| < \sqrt{np_j}\, n^{\frac{1-\varepsilon}{3}}$$

On the other hand from (5) we have

$$\mathsf{P}\left(|v_k - np_k| > \sqrt{np_k}\, n^{\frac{1-\varepsilon}{3}} \text{ for some } k\right) \leqq 2a_n \exp\left(-\frac{n^{\frac{2}{3}(1-\varepsilon)}}{3}\right) = o(1)$$

Thus

$$\mathsf{P}\left(\max_k \frac{\sqrt{np_k}}{\sigma_k}\cdot\frac{s_k}{v_k} < \sqrt{2\log a_n - \log\log a_n + y}\right) =$$

$$= \left[1 - \frac{e^{-y/2}}{2\sqrt{\pi}\,a_n}(1+o(1))\right]^{a_n} \mathsf{P}\left(|v_k - np_k| < \sqrt{np_k}\, n^{\frac{1-\varepsilon}{3}} \text{ for every } k\right) + o(1) =$$

$$= \exp\left(-\frac{e^{-y/2}}{2\sqrt{\pi}}\right) + o(1)$$

as we have stated.

The second relation of lemma 2 can be proved in the same way.

LEMMA 3. *Under the conditions of lemma 2*

$$\mathsf{P}\left(\max_k \frac{\sqrt{np_k^{(n)}}}{\sigma_k^{(n)}}\cdot\frac{\sum_{i=1}^n \xi_i^{(n)} I_{(\eta_i^{(n)}=k)}}{\sum_{i=1}^n I_{(\eta_i^{(n)}=k)}} < \sqrt{2\log a_n - 3\log\log a_n - 2\log 2\sqrt{\pi}\,c}\right) = \frac{1}{a_n^{c+o(1)}}$$

$$\mathsf{P}\left(\max_k \frac{\sqrt{np_k^{(n)}}}{\sigma_k^{(n)}}\cdot\left|\frac{\sum_{i=1}^n \xi_i^{(n)} I_{(\eta_i^{(n)}=k)}}{\sum_{i=1}^n I_{(\eta_i^{(n)}=k)}}\right| < \sqrt{2\log a_n - 3\log\log a_n - 2\log\sqrt{\pi}\,c}\right) = \frac{1}{a_n^{c+o(1)}}$$

The proof is similar to that of Lemma 2.

LEMMA 4. *Under the conditions of Lemma 2 we have*

$$\mathsf{P}\left(\max_k \frac{\sqrt{np_k^{(n)}}}{\sigma_k^{(n)}}\left(1+o\left(\frac{1}{\log a_n}\right)\right)\cdot \frac{\sum\limits_{i=k}^{n} I_{(\eta_i^{(n)}=k)}\left[\xi_i^{(n)}+o\left(\sqrt{\frac{a_n}{n\log a_n}}\right)\right]}{\sum\limits_{i=1}^{n} I_{(\eta_i^{(n)}=k)}} < \right.$$

$$\left. < \sqrt{2\log a_n - \log\log a_n + y}\right) \to \exp\left(-\frac{e^{-y/2}}{2\sqrt{\pi}}\right)$$

$$\mathsf{P}\left(\max_k \frac{\sqrt{np_k^{(n)}}}{\sigma_k^{(n)}}\left(1+o\left(\frac{1}{\log a_n}\right)\right)\cdot \left|\frac{\sum\limits_{i=1}^{n} I_{(\eta_i^{(n)}=k)}\left[\xi_i^{(n)}+o\left(\sqrt{\frac{a_n}{n\log a_n}}\right)\right]}{\sum\limits_{i=1}^{n} I_{(\eta_i^{(n)}=k)}}\right| < \right.$$

$$\left. < \sqrt{2\log a_n - \log\log a_n + y}\right) \to \exp\left(-\frac{e^{-y/2}}{\sqrt{\pi}}\right).$$

Lemma 4 is an easy consequence of Lemma 2.

PROOF OF THEOREM 1. Let us divide the interval [0,1] into a_n equal parts. Obviously

$$m_k^{(n)} = \mathsf{E}\left(\xi_i \left| \frac{k-1}{a_n} \leqq \eta_i < \frac{k}{a_n}\right.\right) = \frac{\int\limits_{\frac{k-1}{a_n}}^{\frac{k}{a_n}} R(y)f(y)\,dy}{\int\limits_{\frac{k-1}{a_n}}^{\frac{k}{a_n}} f(y)\,dy}$$

and

$$D_k^{2(n)} = D^2\left(\xi_i \left| \frac{k-1}{a_n} \leqq \eta_i < \frac{k}{a_n}\right.\right) = \frac{\int\limits_{\frac{k-1}{a_n}}^{\frac{k}{a_n}} [\sigma^2(y)+R^2(y)]f(y)\,dy}{\int\limits_{\frac{k-1}{a_n}}^{\frac{k}{a_n}} f(y)\,dy} - [m_k^{(n)}]^2.$$

Let us define the random variables $(\xi_i^{(n)}, \eta_i^{(n)})$ $i=1, 2 \dots n$ in the following way:

$$\eta_i^{(n)} = k \quad\text{and}\quad \xi_i^{(n)} = \frac{\xi_i - m_k^{(n)}}{D_k^{(n)}}\cdot\sigma\left(\frac{k-\frac{1}{2}}{a_n}\right)$$

if

$$\frac{k-1}{a_n} \leqq \eta_i < \frac{k}{a_n} \qquad i=1, 2, \dots, n \quad k=1, 2, \dots, a_n.$$

Then the conditions of Lemma 2 are satisfied for the sequence $(\xi_i^{(n)}, \eta_i^{(n)})$. Indeed, $p_k^{(n)} = \int_{\frac{k-1}{a_n}}^{\frac{k}{a_n}} f(y)dy$, thus $\frac{c_1}{a_n} < p_k^{(n)} < \frac{c_2}{a_n}$ since $f(x)$ is bounded both from above and below. $\mathsf{E}(\xi_i^{(n)}|\eta_i^{(n)}=k)=0$, $(\sigma_k^{(n)})^2 = \mathsf{D}^2(\xi_i^{(n)}|\eta_i^{(n)}=k)=\sigma^2\left(\frac{k-\frac{1}{2}}{a_n}\right)$ is bounded from below. $\frac{D_k^{(n)}}{\sigma\left(\frac{k-\frac{1}{2}}{a_n}\right)}$ is bounded from below and

$$\mathsf{E}\left(\exp\left(t_0\frac{D_k^{(n)}}{\sigma\left(\frac{k-\frac{1}{2}}{a_n}\right)}|\xi_i^{(n)}|\right)\middle|\eta_i^{(n)}=k\right) \leqq c\exp(t_0 m_k^{(n)})$$

is also bounded as required. By the Taylor formula

$$f(y) = f(x)+(y-x)\cdot f'(x+\theta(y-x))$$

$$R(y) = R(x)+(y-x)\cdot R'(x+\theta'(y-x)) \qquad 0<\theta, \theta', \theta''<1$$

$$\sigma(y) = \sigma(x)+(y-x)\cdot\sigma'(x+\theta''(y-x))$$

The first derivatives of $f(x)$, $R(x)$, $\sigma(x)$ are bounded and so it can be seen easily that $m_k^{(n)} = R(x)+O\left(\frac{1}{a_n}\right)$, $p_k^{(n)} = \frac{f(x)}{a_n}+O\left(\frac{1}{a_n^2}\right)$, $\mathsf{D}^2(\xi_i^{(n)}|\eta_i^{(n)}=k) = \sigma^2(x)+O\left(\frac{1}{a_n}\right)$ if $x\in\left[\frac{k-1}{a_n}, \frac{k}{a_n}\right]$. Since $n^\alpha<a_n$ and $\alpha>\frac{1}{3}$, $\frac{1}{a_n} = o\left(\sqrt{\frac{a_n}{n\log a_n}}\right)$ and

$$\sqrt{\frac{nf(x)}{a_n}}\frac{1}{\sigma(x)} = \frac{\sqrt{np_k^{(n)}}}{\sigma_k^{(n)}}\left(1+o\left(\frac{1}{\log a_n}\right)\right) \quad \text{for} \quad x\in\left[\frac{k-1}{a_n}, \frac{k}{a_n}\right].$$

Hence Lemma 4 can be applied, and we get

$$\mathsf{P}\left(\sup_k \sup_{k\in\left[\frac{k-1}{a_n}, \frac{k}{a_n}\right]} \sqrt{\frac{nf(x)}{a_n}} \frac{\sum_{i=1}^n I_{(\eta_i^{(n)}=k)}[\xi_i-R(x)]}{\sigma(x)\sum_{i=1}^n I_{(\eta_i^{(n)}=k)}} < \sqrt{2\log a_n-\log\log a_n+y}\right) \to$$

$$\to \exp\left(-\frac{e^{-y/2}}{2\sqrt{\pi}}\right)$$

which is identical to formula (3). The second statement of lemma 4 gives formula (4).

The PROOF of Theorem 1′ is almost the same. By the Taylor expansion up to 2 terms

$$R(x)=R\left(\frac{k-\frac{1}{2}}{a_n}\right)+R'\left(\frac{k-\frac{1}{2}}{a_n}\right)\left(x-\frac{k-\frac{1}{2}}{a_n}\right)+$$

$$+\frac{1}{2}R''\left(\frac{k-\frac{1}{2}}{a_n}+\theta\left(x-\frac{k-\frac{1}{2}}{a_n}\right)\right)\left(x-\frac{k-\frac{1}{2}}{a_n}\right)^2$$

$$f(x)=f\left(\frac{k-\frac{1}{2}}{a_n}\right)+f'\left(\frac{k-\frac{1}{2}}{a_n}\right)\left(x-\frac{k-\frac{1}{2}}{a_n}\right)+$$

$$+\frac{1}{2}f''\left(\frac{k-\frac{1}{2}}{a_n}+\theta'\left(x-\frac{k-\frac{1}{2}}{a_n}\right)\right)\left(x-\frac{k-\frac{1}{2}}{a_n}\right)^2$$

$$\sigma(x)=\sigma\left(\frac{k-\frac{1}{2}}{a_n}\right)+\sigma'\left(\frac{k-\frac{1}{2}}{a_n}\right)\left(x-\frac{k-\frac{1}{2}}{a_n}\right)+$$

$$+\frac{1}{2}\sigma''\left(\frac{k-\frac{1}{2}}{a_n}+\theta''\left(x-\frac{k-\frac{1}{2}}{a_n}\right)\right)\left(x-\frac{k-\frac{1}{2}}{a_n}\right)^2$$

where $0<\theta, \theta', \theta''<1$. It implies that

$$m_k^{(n)}=R\left(\frac{k-\frac{1}{2}}{a_n}\right)+O\left(\frac{1}{a_n^2}\right),\quad p_k^{(n)}=\frac{f\left(\frac{k-\frac{1}{2}}{a_n}\right)}{a_n}+O\left(\frac{1}{a_n^3}\right)$$

and

$$D^2(\xi_i^{(n)}|\eta_i^{(n)}=k)=\sigma^2\left(\frac{k-\frac{1}{2}}{a_n}\right)+O\left(\frac{1}{a_n^2}\right),$$

and these imply, just as in Theorem 1, the limit relation

$$\mathsf{P}\left(\sup_k\sqrt{\frac{nf\left(\frac{k-\frac{1}{2}}{a_n}\right)}{a_n}}\,\frac{R_n\left(\frac{k-\frac{1}{2}}{a_n}\right)-R\left(\frac{k-\frac{1}{2}}{a_n}\right)}{\sigma\left(\frac{k-\frac{1}{2}}{a_n}\right)}<\sqrt{2\log a_n-\log\log a_n+y}\right)\to$$

$$\to\exp\left(\frac{-e^{-y/2}}{2\sqrt{\pi}}\right).$$

Now by

$$R(x)=R\left(\frac{k-\frac{1}{2}}{a_n}\right)+\left[R\left(\frac{k+\frac{1}{2}}{a_n}\right)-R\left(\frac{k-\frac{1}{2}}{a_n}\right)\right]\left(a_nx-k-\frac{1}{2}\right)+O\left(\frac{1}{a_n^2}\right)$$

for $x\in\left[\frac{k-\frac{1}{2}}{a_n},\ \frac{k+\frac{1}{2}}{a_n}\right]$ and (4) the statement of Theorem 1′ is valid.

PROOF OF THEOREM 2. Since we have already proved Theorem 1, it is sufficient to prove that

$$\mathsf{P}\left(\sup_x \frac{\sqrt{f(x)}}{\sigma(x)}\frac{\sigma_n(x)}{\sqrt{f_n(x)}} > 1+\frac{\delta_n}{\sqrt{\log a_n}}\right) \to 0$$

and

$$\mathsf{P}\left(\inf_x \frac{\sqrt{f(x)}}{\sigma(x)}\frac{\sigma_n(x)}{\sqrt{f_n(x)}} > 1-\frac{\delta_n}{\sqrt{\log a_n}}\right) \to 0$$

for an appropriate sequence $\delta_n \to 0$, i.e. $\frac{\sqrt{f(x)}}{\sigma(x)}$ and $\frac{\sqrt{f_n(x)}}{\sigma_n(x)}$ are near to each other. We shall estimate $f_n(x)-f(x)$ and $\sigma_n(x)-\sigma(x)$. This will be similar to the proof of Lemma 2.

$$f(x)-a_n p_k^{(n)} = O\left(\frac{1}{a_n}\right) \qquad x\in\left[\frac{k+1}{a_n}, \frac{k}{a_n}\right]$$

and

$$f_n(x)-a_n p_k^{(n)} = \frac{a_n}{n}\sum_{i=1}^{n}[I_{(\eta_i^{(n)}=k)}-\mathsf{E}I_{(\eta_i^{(n)}=k)}].$$

Thus, exactly the same way as in proving estimation (5) we get that

$$\mathsf{P}\left(|f_n(x)-f(x)| > n^{\frac{1-\alpha}{3}}\frac{\sqrt{n}}{a_n\sqrt{p_k^{(n)}}}\right) = O\left(\exp\left(-n^{\frac{1-\alpha}{3}}\right)\right).$$

$n^{\frac{1-\alpha}{3}}\frac{\sqrt{n}}{a_n\sqrt{p_k^{(n)}}} < K\cdot n^{-\frac{1-\alpha}{6}}$ for an appropriate constant K. Thus

$$\mathsf{P}\left(\sup_x |f_n(x)-f(x)| > K\cdot n^{-\frac{1-\alpha}{6}}\right) \leqq a_n O\left(\exp\left(-n^{\frac{1-\alpha}{3}}\right)\right) \to 0$$

We have

$$\sigma^2(x)-D_k^{(n)})^2 = O\left(\frac{1}{a_n}\right) \quad \text{and} \quad \sigma_n^2(x) = \frac{\sum_{i=1}^{n}\xi_i^2 I_{(\eta_i^{(n)}=k)}}{\sum_{i=1}^{n} I_{(\eta_i^{(n)}=k)}} - \left[\frac{\sum_{i=1}^{n}\xi_i I_{(\eta_i^{(n)}=k)}}{\sum_{i=1}^{n} I_{(\eta_i^{(n)}=k)}}\right]^2$$

for

$$x\in\left[\frac{k-1}{a_n}, \frac{k}{a_n}\right].$$

Now we estimate

$$\mathsf{P}\left(\left|\frac{\sum_{i=1}^{n}\xi_i I_{(\eta_i^{(n)}=k)}}{\sum_{i=1}^{n} I_{(\eta_i^{(n)}=k)}} - m_k^{(n)}\right| > \frac{n^{\frac{1-\alpha}{3}}}{\sqrt{np_k^{(n)}}}\right).$$

Let $\tilde{\xi}_1, \dots \tilde{\xi}_n \dots n=1, 2, \dots$ be independent identically distributed random variables having the distribution

$$\mathsf{P}(\tilde{\xi}_i < x) = \mathsf{P}(\xi_1 - m_k^{(n)} < x | \eta_1^{(n)} = k)$$

Then

$$\mathsf{P}\left(\left|\frac{\sum_{i=1}^{n} \xi_i I_{(\eta_i^{(n)}=k)}}{\sum_{i=1}^{n} I_{(\eta_i^{(n)}=k)}} - m_k^{(n)}\right| > \frac{n^{\frac{1-\alpha}{3}}}{\sqrt{np_k^{(n)}}} \,\middle|\, \sum_{i=1}^{n} I_{(\eta_i^{(n)}=k)} = t\right) =$$

$$= \mathsf{P}\left(\left|\frac{\tilde{\xi}_1 + \dots + \tilde{\xi}_t}{\sqrt{t}}\right| < n^{\frac{1-\alpha}{3}} \sqrt{\frac{t}{np_k^{(n)}}}\right) = O(\exp(-n^{\frac{1-\alpha}{3}}))$$

for $|t - np_k^{(n)}| < n^{\frac{1-\alpha}{3}} \sqrt{np_k^{(n)}}$. On the other hand

$$\mathsf{P}\left(\left|\sum_{i=1}^{n} I_{(\eta_i^{(n)}=k)} - np_k^{(n)}\right| > n^{\frac{1-\alpha}{3}} \sqrt{np_k^{(n)}}\right) = O(\exp(-n^{\frac{1-\alpha}{3}}))$$

Thus

$$\mathsf{P}\left(\left|\frac{\sum_{i=1}^{n} \xi_i I_{(\eta_i^{(n)}=k)}}{\sum_{i=1}^{n} I_{(\eta_i^{(n)}=k)}} - m_k^{(n)}\right| > \frac{n^{\frac{1-\alpha}{3}}}{\sqrt{np_k^{(n)}}} \text{ for some } k\right) \leqq a_n O(\exp(-n^{\frac{1-\alpha}{3}})) \to 0$$

The relation

$$\mathsf{P}\left(\left|\frac{\sum_{i=1}^{n} \xi_i^2 I_{(\eta_i^{(n)}=k)}}{\sum_{i=1}^{n} I_{(\eta_i^{(n)}=k)}} - \frac{\int_{\frac{k-1}{a_n}}^{\frac{k}{a_n}} [\sigma^2(y) + R^2(y)] f(y)\, dy}{\int_{\frac{k-1}{a_n}}^{\frac{k}{a_n}} f(y)\, dy}\right| > \frac{n^{\frac{1-\alpha}{3}}}{\sqrt{np_k^{(n)}}} \text{ for some } k\right) \to 0$$

can be proved in the same way. From these estimations it follows that $|f_n(x) - f(x)|$ and $|\sigma_n(x) - \sigma(x)|$ are less than n^{-c} for appropriate $c > 0$ with a probability near 1. Thus Theorem 2 is valid.

Proof of Theorem 3. First we prove that

$$\mathsf{P}\left(\varliminf_{x} \sup \sqrt{\frac{nf(x)}{2a_n \log a_n}} \frac{|R_n(x) - R(x)|}{\sigma(x)} > 1 - \varepsilon\right) = 1$$

for arbitrary $\varepsilon > 0$.

It is easy to check the validity of the conditions of Lemma 2 for the sequence $(\xi_i^{(n)}, \eta_i^{(n)})$, therefore by Lemma 3 we can state that

$$\mathsf{P}\left(\max_k \sqrt{np_k^{(n)}}\left|\frac{\sum_{i=1}^{n} \xi_i^{(n)} I_{(\eta_i^{(n)}=k)}}{\sigma\left(\frac{k-\frac{1}{2}}{a_n}\right)\sum_{i=1}^{n} I_{(\eta_i^{(n)}=k)}}\right| > \sqrt{2\log a_n - 3\log\log a_n - 2\log\sqrt{\pi}\, c}\right) =$$

$$= \frac{1}{a_n^{c+o(1)}}$$

Thus, choosing n_0 large enough, by the Borel—Cantelly lemma we obtain

$$\max_k \sqrt{np_k^{(n)}}\left|\frac{\sum_{i=1}^{n} \xi_i^{(n)} I_{(\eta_i^{(n)}=k)}}{\sigma\left(\frac{k-\frac{1}{2}}{a_n}\right)\sum_{i=1}^{n} I_{(\eta_i^{(n)}=k)}}\right| <$$

$$< \sqrt{2\log a_n - 3\log\log a_n - 2\log\sqrt{\pi}\, c} \quad \text{for} \quad n > n_0(\omega)$$

with probability 1. Since

$$\sqrt{np_k^{(n)}} = \sqrt{\frac{nf(x)}{a_n}}\left(1+O\left(\frac{1}{a_n}\right)\right), \quad \sigma\left(\frac{k-\frac{1}{2}}{a_n}\right) = \sigma(x) + O\left(\frac{1}{a_n}\right)$$

and

$$\xi_i^{(n)} I_{(\eta_i^{(n)}=k)} = \left[\xi_i - R(x) + O\left(\frac{1}{a_n}\right)\right] I_{(\eta_i^{(n)}=k)}\left[1+O\left(\frac{1}{a_n}\right)\right]$$

if $x \in \left[\frac{k-1}{a_n}, \frac{k}{a_n}\right]$, our statement follows.

The proof of the relation

$$\mathsf{P}\left(\overline{\lim}\sup_x \sqrt{\frac{nf(x)}{2a_n\log a_n}}\,\frac{|R_n(x)-R(x)|}{\sigma(x)} < 1-\varepsilon\right) = 1$$

is more involved.

Set

$$\bar{\xi}_i = \frac{\xi_i - \mathsf{E}(\xi_i|\eta_i)}{\mathsf{D}(\xi_i|\eta_i)}.$$

Let us define the random variable

$$T_n(x) = \sum_{\substack{i\le n\\ \eta_i \le x}} \bar{\xi}_i \qquad 0 \le x \le 1.$$

We want to prove that

(6)
$$\mathsf{P}(A_n)=\mathsf{P}\left(\sup_{\substack{y-x\le\frac{1}{a_n}\\ 0\le x<y\le 1}}\frac{1}{\sqrt{f(x)}}|T_n(x)-T_n(y)|>\left(1+\varepsilon+\frac{\sqrt{\varepsilon}}{2}\right)\sqrt{\frac{2n\log a_n}{a_n}}\right)\le\frac{12}{\varepsilon\cdot a_n^{\varepsilon}}.$$

First we estimate, for fixed x, the probability

$$\mathsf{P}\left(\sup_{x\le y\le x+\frac{3+\varepsilon}{3a_n}}\frac{1}{\sqrt{f(x)}}|T_n(x)-T_n(y)|>(1+\varepsilon)\sqrt{\frac{2n\log a_n}{a_n}}\right).$$

Let us order the η_i-s falling into the interval $\left[x, x+\frac{3+\varepsilon}{3a_n}\right]$ in increasing order $x\le\eta_1^*<\eta_2^*\dots<\eta_t^*<x+\frac{3+\varepsilon}{3a_n}$ and denote by ξ_i^* the ξ_i which is the pair of η_j^*. Then

$$\sup_{x<y<x+\frac{3+\varepsilon}{3a_n}}|T_n(y)-T_n(x)|=\sup_{k\le t}\left|\sum_{j=1}^{k}\xi_j^*\right|$$

We can state that

(7)
$$\mathsf{P}\left(\sup_{k\le t}\sum_{j=1}^{k}\xi_j^*>(1+\varepsilon)\sqrt{\frac{2nf(x)\log a_n}{a_n}}\,\middle|\,\begin{matrix}\text{the number of } \eta_i\text{-s } i\le n \text{ falling}\\ \text{into the interval } \left[x,x+\frac{3+\varepsilon}{3a_n}\right]\\ \text{is } t,\ \eta_1^*=x_1,\dots\eta_t^*=x_t\end{matrix}\right)=$$

$$=\mathsf{P}\left(\sup_{k\le t}\sum_{j=1}^{k}\zeta_j>(1+\varepsilon)\sqrt{\frac{2nf(x)\log a_n}{a_n}}\right)$$

where $\zeta_1,\zeta_2,\dots,\zeta_t$ are independent and $\mathsf{P}(\zeta_j<y)=\mathsf{P}(\xi_1<y\,|\,\eta_1=x_j)$. Now the sequence $\sum_{j=1}^{k}\zeta_j$, $k=1,2\dots,t$ forms a martingale, so $\exp\left(t\sum_{j=1}^{k}\zeta_j\right)$ is a semimartingale for arbitrary t, and by a well-known martingale inequality see [1] p. 314.,

(8)
$$\mathsf{P}\left(\sup_{k\le t}\sum_{j=1}^{k}\zeta_j>(1+\varepsilon)\sqrt{\frac{2nf(x)\log a_n}{a_n}}\right)=$$

$$=\mathsf{P}\left(\sup_{k\le t}\exp\left(\sqrt{\frac{2a_n\log a_n}{nf(x)}}\sum_{j=1}^{k}\zeta_j\right)>\exp\left(2(1+\varepsilon)\log a_n\right)\right)\le$$

$$\le\frac{\prod_{i=1}^{t}\mathsf{E}\exp\left(\sqrt{\frac{2a_n\log a_n}{nf(x)}}\zeta_i\right)}{a_n^{2(1+\varepsilon)}}$$

Now

$$\text{(9)}\qquad \mathsf{E}\exp\left(\sqrt{\frac{2a_n\log a_n}{nf(x)}}\zeta_i\right) \leqq 1+\mathsf{E}\sqrt{\frac{2a_n\log a_n}{nf(x)}}\zeta_i+$$

$$+\mathsf{E}\left[\frac{a_n\log a_n}{nf(x)}\zeta_i^2\exp\left|\sqrt{\frac{2a_n\log a_n}{nf(x)}}\zeta_i\right|\right] \leqq 1+\left(1+\frac{\varepsilon}{2}\right)\frac{a_n\log a_n}{nf(x)} \leqq$$

$$\leqq \exp\left[\left(1+\frac{\varepsilon}{2}\right)\frac{a_n\log a_n}{nf(x)}\right]$$

It can be proved in the same way as relation (5) that

$$\mathsf{P}\left(\left|t-\frac{n\left(1+\frac{\varepsilon}{3}\right)}{a_n}\cdot f(x)\right| > n^{\frac{1-\alpha}{3}}\sqrt{\frac{n}{a_n}f(x)}\right) = O(\exp(-n^{\frac{1}{3}(1-\alpha)}))$$

$\left(t\right.$ is the number of η_i-s ($i\leqq n$) falling into the interval $\left.\left[x, x+\frac{3+\varepsilon}{3a_n}\right]\right)$, and so by (7), (8) and (9)

$$\mathsf{P}\left(\sup_{x<y<x+\frac{3+\varepsilon}{3a_n}}(T_n(y)-T_n(x)) > (1+\varepsilon)\sqrt{\frac{2nf(x)\log a_n}{a_n}}\right) < \frac{1}{a_n^{1+\varepsilon}}$$

Using this relation for the sequence $(-\xi_1, \eta_1), \dots, (-\xi_n, \eta_n)$ we get that

$$\mathsf{P}\left(\sup_{x<y<x+\frac{3+\varepsilon}{3a_n}}|T_n(x)-T_n(y)| > (1+\varepsilon)\sqrt{\frac{2nf(x)\log a_n}{a_n}}\right) < \frac{2}{a_n^{1+\varepsilon}}$$

In the same way (x is fixed) we find

$$\mathsf{P}\left(\sup_{x<y<x+\frac{\varepsilon}{3a_n}}|T_n(y)-T_n(x)| > \frac{\sqrt{\varepsilon}}{2}\sqrt{\frac{2nf(x)\log a_n}{a_n}}\right) < \frac{2}{a_n^{1+\varepsilon}}$$

Thus

$$\mathsf{P}\left(\sup_{\substack{x<y<x+\frac{3+\varepsilon}{3a_n}\\ x=\frac{k\varepsilon}{3a_n}\left(k=0, 1, \dots, \frac{3a_n}{\varepsilon}\right)}}\frac{1}{\sqrt{f(x)}}|T_n(y)-T_n(x)| > (1+\varepsilon)\sqrt{\frac{2n\log a_n}{a_n}}\right) < \frac{6}{\varepsilon a_n^{\varepsilon}}$$

and

$$\mathsf{P}\left(\sup_{\substack{x<y<x+\frac{\varepsilon}{3a_n}\\ x=\frac{k\varepsilon}{3a_n}\left(k=0, 1, \dots, \frac{3a_n}{\varepsilon}\right)}}\frac{1}{\sqrt{f(x)}}|T_n(y)-T_n(x)| > \frac{\sqrt{\varepsilon}}{2}\sqrt{\frac{2n\log a_n}{a_n}}\right) < \frac{6}{\varepsilon a_n^{\varepsilon}}.$$

Let us now consider an arbitrary interval $[u, v]$ such that $0<v-u\leqq\frac{1}{a_n}$. Choose the integer k in such a way that for the number $x=\frac{k\varepsilon}{3a_n}$ the inequality $0<u-x<\frac{\varepsilon}{3a_n}$ hold. Then

$$|T_n(v)-T_n(u)|\leqq \sup_{x<y<x+\frac{3+\varepsilon}{3a_n}} |T_n(y)-T_n(x)|+\sup_{x<y<x+\frac{\varepsilon}{3a_n}} |T_n(y)-T_n(x)|,$$

and therefore (6) follows from the last two inequalities. Choosing a $c>\frac{6}{\varepsilon}$ we get $\mathrm{P}(A_{l^c})<\frac{1}{l^2}$, and by the Borel—Cantelli lemma

$$\lim_{l\to\infty} A_{l^c}=\emptyset \tag{10}$$

Given a number n it determines a unique l for which $l^c\leqq n<(l+1)^c$. Then $n-l^c<c(l+1)^{c-1}<2cn^{c-1}$. By similar calculations as before we get

$$\mathrm{P}\left(\sum_{i=l^c}^{n}\xi_i I_{(\eta_i^{(n)}=k)}>\varepsilon\sqrt{\frac{2nf(x)\log a_n}{a_n}}\right)<\exp(-Dn^{1/c})$$

with appropriate $D>0$ for $k=1, 2, \ldots, a_n$, $x\in\left[\frac{k-1}{a_n}, \frac{k}{a_n}\right]$ and

$$\mathrm{P}(B_n)=\mathrm{P}\left(\sup_{\substack{k\\ x\in\left[\frac{k-1}{a_n}, \frac{k}{a_n}\right]}}\left|\frac{1}{\sqrt{f(x)}}\sum_{i=l^c}^{n}\xi_i I_{(\eta_i^{(n)}=k)}\right|>\varepsilon\sqrt{\frac{2n\log a_n}{a_n}}\right)<a_n\exp(-Dn^{1/c})$$

This yields the relation

$$\lim B_n=\emptyset. \tag{11}$$

By the definition of A_n and B_n and relations (10) and (11) we have with probability 1

$$\left|\sup\frac{1}{\sqrt{f(x_k)}}\sum_{i=1}^{n}\xi_i I_{(\eta_i^{(n)}=k)}\right|<\left(1+2\varepsilon+\frac{\sqrt{\varepsilon}}{2}\right)\sqrt{\frac{2n\log a_n}{a_n}} \tag{12}$$

for

$$x_k\in\left[\frac{k-1}{a_n}, \frac{k}{a_n}\right]$$

for every large n.

An easy computation shows that with probability one

$$\sum_{i=1}^{n} I_{(\eta_i^{(n)}=k)}>(1-\varepsilon)\frac{nf(x_k)}{a_n},\ x_k\in\left[\frac{k-1}{a_n}, \frac{k}{a_n}\right] \tag{13}$$

for every large n and $k=1, 2, \ldots, a_n$. Since $R(x)$, $\sigma(x)$ and $f(x)$ are smooth enough

$$\mathrm{P}\left(\overline{\lim_n}\sup_x \sqrt{\frac{nf(x)}{2a_n \log a_n}}\,\frac{|R_n(x)-R(x)|}{\sigma(x)} < \frac{1+2\varepsilon+\frac{\sqrt{\varepsilon}}{2}}{1-\varepsilon}\right) = 1$$

from the relations (12) and (13). Since it is true for arbitrary small $\varepsilon>0$, the first statement of theorem 3 is valid. The second statement can be proved in the same way.

LITERATURE

[1] Doob, J. L.: *Stochastic Processes,* Wiley, New York (1953).
[2] Feller, W.: *An Introduction to Probability Theory and Its Applications* 2. Wiley, New York.
[3] Révész, P.: On empirical density function, *Periodica Math. Hungar.* **2** (1972) 85—110.
[4] Надарая, Э. А.: Замечания о непаратмерических оценках плотности вероятности и кривой регрессии. *Теория вероят. и ее примен.* XV. 1 (1970) 139—142.

Mathematical Institute of the Hungarian Academy of Sciences, Budapest

(Received April 15, 1973)