ᲡᲐᲥᲐᲠᲗᲕᲔᲚᲝᲡ ᲛᲔᲪᲜᲘᲔᲠᲔᲑᲐᲗᲐ ᲔᲠᲝᲕᲜᲣᲚᲘ ᲐᲙᲐᲓᲔᲛᲘᲘᲡ ᲛᲝᲐᲛᲑᲔ, ტ. 5, №2, 2011
BULLETIN OF THE GEORGIAN NATIONAL ACADEMY OF SCIENCES, vol. 5, no. 2, 2011

Mathematics

On the Estimation of a Distribution Function by an Indirect Sample. II

Elizbar Nadaraya*, Petre Babilua, Grigol Sokhadze****

* *Academy Member, I. Javakhishvili Tbilisi State University*
** *I. Javakhishvili Tbilisi State University*

ABSTRACT. In this paper the limit theorems are proved for continuous functionals related to the estimate of $\hat{F}_n(x)$ in the space $C[a,1-a]$. *© 2011 Bull. Georg. Natl. Acad. Sci.*

Key words: *distribution function estimate, unbiased, consistency, asymptotic normality, estimate of time moments, Wiener process, random process.*

Here as a sample we consider a sequence of random indicators $\xi_1 = I(X_1 < t_1), \xi_2 = I(X_2 < t_2), \dots, \xi_n = I(X_n < t_n)$, where $X_1, X_2, \dots, X_n$ are independent, equally distributed nonnegative random values with a distribution function $F(x)$, $t_i = c_F \ \dfrac{2i-1}{2n}$, $i = \overline{1,n}$, $c_F = \inf\{x \ge 0: \ F(x) = 1\} < \infty$. The problem consists in estimation of the distribution function $F(x)$ by using the sample $\xi_1, \xi_2, \dots, \xi_n$.

As an estimate for $F(x)$ we consider an expression of the form

$$\hat{F}_n(x) = \begin{cases} 0, & x \le 0, \\ F_{1n}(x) \cdot F_{2n}^{-1}(x), & 0 < x < c_F, \\ 1, & x \ge c_F, \end{cases}$$

$$F_{1n}(x) = \frac{1}{nh} \sum_{j=1}^{n} K\left(\frac{x - t_j}{h}\right) \xi_j,$$

$$F_{2n}(x) = \frac{1}{nh} \sum_{j=1}^{n} K\left(\frac{x - t_j}{h}\right),$$

where $\{h = h(n)\}$ is a sequence of positive numbers tending to zero, while the kernel $K(x) \ge 0$ is chosen so that it would be a function of finite variation and satisfy the conditions

$$K(-u) = K(u), \quad \int K(u) \ du = 1, \quad K(u) = 0 \quad for \quad |u| \ge 1. \tag{1}$$

Lemma 1 ([1]). *If $nh \to \infty$ as $n \to \infty$, then*

$$\frac{1}{nh}\sum_{j=1}^{n} K^{\nu_1-1}\left(\frac{x-t_j}{h}\right) F^{\nu_2-1}\left(t_j\right) = \frac{1}{c_F h}\int_0^{c_F} K^{\nu_1-1}\left(\frac{x-u}{h}\right) F^{\nu_2-1}(u)\,du + O\left(\frac{1}{nh}\right)$$

uniformly with respect to $x \in [0, c_F]$; ν_1, ν_2 are natural numbers. In the sequel, it is assumed that the interval $[0, c_F] = [0,1]$.

Theorem 1. *Let $g(x) \ge 0$, $x \in [a, 1-a]$, $0 < a < \frac{1}{2}$, be a measurable and bounded function.*

(a) *If $F(a) > 0$ and $nh^2 \to \infty$ as $n \to \infty$, then*

$$\overline{T}_n = \sqrt{n}\int_a^{1-a} g_1(x)\left[\hat{F}_n(x) - E\hat{F}_n(x)\right] dx \xrightarrow{d} N\left(0, \sigma^2\right), \tag{2}$$

$$g_1(x) = g(x)\psi\left(F(x)\right), \quad \psi(t) = \frac{1}{\sqrt{t(1-t)}}.$$

(b) *If $F(a) > 0$, $nh^2 \to \infty$, $nh^4 \to 0$ as $n \to \infty$ and $F(x)$ has bounded derivatives up to second order, then*

$$T_n = \sqrt{n}\int_a^{1-a} g_1(x)\left[\hat{F}_n(x) - F(x)\right] dx \xrightarrow{d} N\left(0, \sigma^2\right)$$

as $n \to \infty$, where $N\left(0, \sigma^2\right)$ is a random value having a normal distribution with zero mean and variance

$$\sigma^2 = \int_a^{1-a} g^2(u)\,du.$$

Remark 1. We have introduced $a > 0$ in (2) to avoid the boundary effect of the estimate $\hat{F}_n(x)$, i.e., the estimate $\hat{F}_n(x)$ being a kernel type estimate behaves near the boundary of the interval [0,1] worse in the sense of bias order than inside any interval $[a, 1-a] \subset [0,1]$, $0 < a < \frac{1}{2}$.

Proof of Theorem 1. We have

$$\overline{T}_n = \frac{1}{\sqrt{n}}\sum_{j=1}^{n}\left(\xi_j - F\left(t_j\right)\right)\frac{1}{h}\int_a^{1-a} K\left(\frac{u-t_j}{h}\right) g_{2n}(u)\,du,$$

where $g_{2n}(u) = g_1(u) F_{2n}^{-1}(u)$.

Hence

$$\sigma_n^2 = Var\overline{T}_n = \frac{1}{n}\sum_{j=1}^{n}\psi^{-2}\left(F\left(t_j\right)\right)\left(\frac{1}{h}\int_a^{1-a} K\left(\frac{u-t_j}{h}\right) g_{2n}(u)\,du\right)^2. \tag{3}$$

Since $K(u)$ has $[-1,1]$ as a support and $0 < a \le u \le 1-a$, we have $F_{2n}(u) = 1 + O\left(\frac{1}{nh}\right)$ and $g_2(u) = g_1(u) + O\left(\frac{1}{nh}\right)$ uniformly with respect to $u \in [a, 1-a]$ [1]. Therefore from (3) we have

$$\sigma_n^2 = \frac{1}{n}\sum_{j=1}^{n}\psi^{-2}\left(F\left(t_j\right)\right)\left(\frac{1}{h}\int_a^{1-a} K\left(\frac{u-t_j}{h}\right)g_1\left(u\right)du\right)^2 + O\left(\frac{1}{nh}\right).$$

By virtue of Lemma 1 it can be easily shown that

$$\frac{1}{n}\sum_{j=1}^{n}\psi^{-2}\left(F\left(t_j\right)\right)\left(\frac{1}{h}\int_a^{1-a} K\left(\frac{u-t_j}{h}\right)g_1\left(u\right)du\right)^2 = \int_0^1 \psi^{-2}\left(F\left(t\right)\right)dt\left(\frac{1}{h}\int_a^{1-a} K\left(\frac{u-t}{h}\right)g_1\left(u\right)du\right)^2 + O\left(\frac{1}{nh^2}\right).$$

Therefore

$$\sigma_n^2 = \int_a^{1-a} \psi^{-2}\left(F\left(t\right)\right)dt\left(\frac{1}{h}\int_a^{1-a} K\left(\frac{u-t}{h}\right)g_1\left(u\right)du\right)^2 + \varepsilon_n^{(1)} + \varepsilon_n^{(2)} + O\left(\frac{1}{nh^2}\right), \tag{4}$$

$$\varepsilon_n^{(1)} = \int_0^a \psi^{-2}\left(F\left(t\right)\right)dt\left(\frac{1}{h}\int_a^{1-a} K\left(\frac{u-t}{h}\right)g_1\left(u\right)du\right)^2,$$

$$\varepsilon_n^{(2)} = \int_{1-a}^1 \psi^{-2}\left(F\left(t\right)\right)dt\left(\frac{1}{h}\int_a^{1-a} K\left(\frac{u-t}{h}\right)g_1\left(u\right)du\right)^2.$$

Since from $F\left(u\right)\left(1-F\left(u\right)\right) \le \frac{1}{4}$ and from inequalities $g\left(u\right) \le c_1$, $\psi\left(F\left(u\right)\right) \le \dfrac{1}{\sqrt{F\left(a\right)\left(1-F\left(1-a\right)\right)}}$, $a \le u \le 1-a$, it follows that $g_1\left(u\right) \le c_2$, we have

$$\varepsilon_n^{(1)} \le c_3 \int_0^a dt\left(\int_{\frac{a-t}{h}}^{\frac{1-a-t}{h}} K\left(u\right)\ du\right)^2, \tag{5}$$

with $a-t \ge 0$ and $1-a-t \ge 0$. The first inequality is obvious, while the second one follows from the inequalities $0 \le t \le a$ and $0 < a < \frac{1}{2}$.

Thus

$$\lim_{n\to\infty}\int_{\frac{a-t}{h}}^{\frac{1-a-t}{h}} K\left(u\right)\ du = \begin{cases} 0, & 0 \le t < a \\ \frac{1}{2}, & t = a \end{cases}.$$

By the Lebesgue theorem on bounded convergence, from the latter formula and (5) we obtain

$$\varepsilon_n^{(1)} \to 0 \quad \text{for} \quad n \to \infty. \tag{6}$$

Analogously,

$$\varepsilon_n^{(2)} \to 0 \quad \text{for} \quad n \to \infty. \tag{7}$$

Now let us establish that

$$\int_a^{1-a} \psi^{-2}\left(F\left(t\right)\right)dt\left(\frac{1}{h}\int_a^{1-a} K\left(\frac{u-t}{h}\right)g_1\left(u\right)du\right)^2 \to \sigma^2 = \int_a^{1-a} g^2\left(u\right)du$$

as $n \to \infty$.

We have

$$\left|\int_a^{1-a}\psi^{-2}(F(t))dt\left(\frac{1}{h}\int_a^{1-a}g_1(u)K\left(\frac{u-t}{h}\right)du\right)^2-\int_a^{1-a}\psi^{-2}(F(t))g_1^2(t)dt\right|\le$$

$$\le c_4\int_a^{1-a}\psi^{-2}(F(t))dt\left|\frac{1}{h}\int_a^{1-a}g_1(u)K\left(\frac{u-t}{h}\right)du-g_1(t)\right|\le$$

$$\le c_5\int_a^{1-a}dt\left|\frac{1}{h}\int_a^{1-a}g_1(u)K\left(\frac{u-t}{h}\right)du-g_1(t)\int_a^{1-a}\frac{1}{h}K\left(\frac{u-t}{h}\right)du\right|+c_6\int_a^{1-a}\left|\int_a^{1-a}\frac{1}{h}K\left(\frac{u-t}{h}\right)du-1\right|dt=A_{1n}+A_{2n}. \tag{8}$$

Since

$$\int_a^{1-a}\frac{1}{h}K\left(\frac{u-t}{h}\right)du\to 1$$

for all $t\in(a,1-a)$, we have

$$A_{2n}\to 0 \quad \text{as} \quad n\to\infty. \tag{9}$$

Further we extend the function $g_1(u)$ and assume that outside $[a,1-a]$ it has zero values. Denote this extended function by $\overline{g}_1(u)$. Then

$$A_{1n}\le c_7\int_0^1\left(\int_{-\infty}^{\infty}\left|\overline{g}_1(x+y)-\overline{g}_1(y)\right|dy\right)\ \frac{1}{h}K\left(\frac{x}{h}\right)dx\le$$

$$\le c_8\int_{-1}^1\left(\int_{-\infty}^{\infty}\left|\overline{g}_1(y+uh)-\overline{g}_1(y)\right|\ dy\right)K(u)du=c_8\int_{-1}^1\omega(uh)K(u)du\to 0 \quad \text{for} \quad n\to\infty, \tag{10}$$

where

$$\omega(y)=\int_{-\infty}^{\infty}\left|\overline{g}_1(y+x)-\overline{g}(x)\right|dx.$$

(10) holds by virtue of the Lebesgue theorem on majorized convergence and the fact that $\omega(uh)\le 2\|\overline{g}_1\|_{L_1(-\infty,\infty)}$ and $\omega(uh)\to 0$ as $n\to\infty$. Thereby, taking (4)-(10) into account, we have proved that

$$\sigma_n^2\to\sigma^2=\int_a^{1-a}g^2(u)du. \tag{11}$$

Let us now verify the fulfillment of the conditions of the central limit theorem for the sums

$$\overline{T}_n=\frac{1}{\sqrt{n}}\sum_{j=1}^n a_{jn}\left(\xi_j-F(t_j)\right),$$

$$a_{jn}=\int_a^{1-a}\frac{1}{h}K\left(\frac{x-t_j}{h}\right)g_2(x)dx.$$

We have

$$L_n = \frac{n^{-\left(1+\frac{\delta}{2}\right)}\sum_{j=1}^{n} a_{jn}^{2+\delta} E\left|\xi_j - F\left(t_j\right)\right|^{2+\delta}}{\left(\sqrt{Var\ \overline{T}_n}\right)^{2+\delta}} = O\left(n^{-\frac{\delta}{2}}\right)$$

since $a_{jn} \le c_9$, $E\left|\xi_j - F\left(t_j\right)\right|^{2+\delta} \le 1$ for all $1 \le j \le n$ and $Var\overline{T}_n \to \sigma^2$.

Finally, the statement b) of the theorem follows from (a) if we take into account that

$$\sqrt{n}\int_a^{1-a} g_1(x)\left[E\hat{F}_n(x) - F(x)\right]dx = \sqrt{n}\int_a^{1-a} g_{2n}(x)\left[\int_{-1}^{1} K(u)\left(F(x-uh) - F(x)\right)du\right]dx = O\left(\sqrt{n}\,h^2\right) + O\left(\frac{1}{\sqrt{n}\,h}\right). \quad (12)$$

The theorem is proved.

Lemma 2. 1) *In the conditions of the item* (a) *of Theorem* 1,

$$E\left|\overline{T}_n\right|^s \le c_{10}\left(\int_a^{1-a} g(u)\,du\right)^{\frac{s}{2}}, \quad s > 2. \quad (13)$$

2) *In the conditions of the item* (b) *of Theorem* 1,

$$E\left|T_n\right|^s \le c_{11}\left(\int_a^{1-a} g(u)\,du\right)^{\frac{s}{2}}, \quad s > 2. \quad (14)$$

Proof. Since $\overline{T}_n$ is the linear form of $\eta_j = \xi_j - F\left(t_j\right)$, $E\eta_j = 0$, $1 \le j \le n$, to prove (13) we will use Whittle's inequality [2].

It is obvious that $E\left|\eta_j\right|^s \le 1$, $j = \overline{1,n}$. Therefore by virtue of Whittle's inequality we have

$$E\left|\overline{T}_n\right|^s \le c(s)2^s\left[\frac{1}{nh^2}\sum_{j=1}^{n}\left(\int_a^{1-a} K\left(\frac{u-t_j}{h}\right)g_{2n}(u)\,du\right)^2\right]^{\frac{s}{2}},$$

where

$$g_{2n}(u) = g_1(u)F_{2n}^{-1}(u), \quad c(s) = \frac{2^{\frac{s}{2}}}{\sqrt{\pi}}\Gamma\left(\frac{s+1}{2}\right).$$

By virtue of Lemma 1 this inequality implies

$$E\left|\overline{T}_n\right|^s \le c(s)2^s\left[\int_0^1\left(\frac{1}{h}\int_a^{1-a} K\left(\frac{u-t}{h}\right)g_{2n}(u)\,du\right)^2 dt + O\left(\frac{1}{nh^2}\right)\left(\int_a^{1-a} g_{2n}(u)\,du\right)^2\right]^{\frac{s}{2}}. \quad (15)$$

Taking into account that

$$g_{2n}(u) \le g(u)\left[\frac{1}{F(a)\left(1-F(1-a)\right)}\right]\left[1 + O\left(\frac{1}{nh}\right)\right] \le c_{12}\,g(u), \quad a \le u \le 1-a,$$

from (15) we obtain

$$E\left|\bar{T}_n\right|^s \le c_{13}\left[\sup_{0\le t\le 1}\left(\frac{1}{h}\int_a^{1-a} K\left(\frac{u-t}{h}\right)g_{2n}(u)\,du\right)\int_0^1 dt\int_a^{1-a}\frac{1}{h}\quad K\left(\frac{u-t}{h}\right)g_{2n}(u)\,du\right]^{\frac{s}{2}}+O\left(\frac{1}{nh^2}\right)^{\frac{s}{2}}\left(\int_a^{1-a} g_{2n}(u)\,du\right)^{\frac{s}{2}}\le$$

$$\le c_{14}\left(\int_a^{1-a} g(u)\,du\right)^{\frac{s}{2}}\left[1+o(1)\right]\le c_{15}\left(\int_a^{1-a} g(u)\,du\right)^{\frac{s}{2}}, \qquad s>2.$$

Further, we have

$$E\left|T_n\right|^s \le 2^{s-1}\left(E\left|\bar{T}_n\right|^s+\left|\sqrt{n}\int_a^{1-a} g_1(u)\left[E\hat{F}_n(u)-F(u)\right]du\right|^s\right)\le$$

$$\le c_{16}\left(\int_a^{1-a} g(u)\,du\right)^{\frac{s}{2}}+\left|O\left(\sqrt{n}\quad h^2\right)\int_a^{1-a} g(u)\,du\right|^s\le$$

$$\le c_{17}\left(\int_a^{1-a} g(u)\,du\right)^{\frac{s}{2}}.$$

The lemma is proved.

Let us introduce the following random processes:

$$\bar{T}_n(t)=\sqrt{n}\int_a^t\left(\hat{F}_n(u)-E\hat{F}_n(u)\right)\psi\left(F(u)\right)du,$$

$$T_n(t)=\sqrt{n}\int_a^t\left(\hat{F}_n(u)-F(u)\right)\psi\left(F(u)\right)du.$$

Theorem 2. 1^0. *Let the conditions of the item* (a) *of Theorem* 1 *be fulfilled. Then for all continuous functionals* $f(\cdot)$ *on* $C[a,1-a]$, $0<a<\frac{1}{2}$ *the distribution of* $f\left(\bar{T}_n(t)\right)$ *converges to the distribution of* $f\left(W(t-a)\right)$, *where* $W(t-a)$, $a\le t\le 1-a$, *is a Wiener process.*

2^0. *Let the conditions of the item* (b) *of Theorem* 1 *be fulfilled. Then for all continuous functionals* $f(\cdot)$ *on* $C[a,1-a]$ *the distribution of* $f\left(T_n(t)\right)$ *converges to the distribution of* $f\left(W(t-a)\right)$.

Proof. We will first show that the finite-dimensional distributions of processes $\bar{T}_n(t)$ converge to the finite-dimensional distributions of a process $W(t-a)$, $t\ge a$ .We begin by considering one moment of time t_1. We must show that

$$\bar{T}_n(t_1)\xrightarrow{d}W(t_1-a). \tag{16}$$

To prove (16), it suffices to take $g(x)=I_{[a,t_1)}(x)$ in (2). Then, by virtue of Theorem 1,

$$\bar{T}_n(t_1)\xrightarrow{d}N\left(0,\int_a^{1-a} I_{[a,t_1)}(x)\,dx\right)=N(0,t_1-a).$$

Let us now consider two moments of time t_1, t_2, $t_1<t_2$. We must show that

$$\left(\bar{T}_n(t_1),\bar{T}_n(t_2)\right)\xrightarrow{d}\left(W(t_1-a),W(t_2-a)\right). \tag{17}$$

To prove (17), it suffices to take

$$g(x) = (\lambda_1 + \lambda_2) I_{[a,t_1)}(x) + \lambda_2 I_{[t_1,t_2)}(x)$$

in (2). Here λ_1 and λ_2 are arbitrary finite numbers. Then, by virtue of Theorem 1,

$$\lambda_1 \overline{T}_n(t_1) + \lambda_2 \overline{T}_n(t_2) \xrightarrow{d} N\left(0, (\lambda_1 + \lambda_2)^2 (t_1 - a) + \lambda_2^2 (t_2 - t_1)\right).$$

On the other hand,

$$\lambda_1 W(t_1 - a) + \lambda_2 W(t_2 - a) = (\lambda_1 + \lambda_2)\left[W(t_1 - a) - W(0)\right] + \lambda_2 \left[W(t_2 - a) - W(t_1 - a)\right]$$

is distributed like $N\left(0, (\lambda_1 + \lambda_2)^2 (t_1 - a) + \lambda_2^2 (t_2 - t_1)\right)$. Therefore (17) is true.

The case with three or more moments of time is considered analogously. Thus the finite-dimensional distributions of processes $\overline{T}_n(t)$ converge to the finite-dimensional distributions of a Wiener process $W(t-a)$, $a \le t \le 1-a$.

Now let us show that the sequence $\{\overline{T}_n(t)\}$ is tight, i.e. that the sequence of respective distributions is tight. For this it suffices to show that for any $t_1,\ t_2 \in [a, 1-a]$,

$$E\left|\overline{T}_n(t_1) - \overline{T}_n(t_2)\right|^s \le c_{18} |t_1 - t_2|^{\frac{s}{2}}, \quad s > 2.$$

Indeed, this inequality is obtained from (13) for $g(x) = I_{[t_1,t_2]}(x)$.

Further, using (12), (14) and the statements of the item b) of Theorem 1, we easily make sure that the finite-dimensional distributions of processes $T_n(t)$ converge to the finite-dimensional distributions of the Wiener process $W(t-a)$ and also that

$$E\left|T_n(t_1) - T_n(t_2)\right|^s \le c_{19} |t_1 - t_2|^{\frac{s}{2}}, \quad s > 2.$$

Thus the proof of the theorem follows from Theorem 2 of the monograph [3] (chapter IX, section 2).

Application. By virtue of Theorem 2 and the Corollary of Theorem 1 from [3] (chapter VI, section 5) we can write that

$$P\left\{T_n^+ = \max_{a \le t \le 1-a} T_n(t) > \lambda\right\} \to G(\lambda) = \frac{2}{\sqrt{2\pi(1-2a)}} \int_\lambda^\infty \exp\left\{-\frac{x^2}{2(1-2a)}\right\} dx$$

(a is a number given in advance, $0 < a < \frac{1}{2}$) as $n \to \infty$.

This result makes it possible to construct tests of a level α, $0 < \alpha < 1$, for testing the hypothesis H_0 according to which

$$H_0: \lim_{n\to\infty} E\hat{F}_n(x) = F_0(x), \quad a \le x \le 1-a,$$

when the alternative hypothesis is

$$H_1: \int_a^{1-a} \psi(F_0(x)) \left(\lim_{n\to\infty} E\hat{F}_n(x) - F_0(x)\right) dx > 0.$$

Let λ_α be a critical value, $G(\lambda_\alpha) = \alpha$. If as a result of the experiment it turns out that $T_n^+ \ge \lambda_\alpha$, then the hypothesis H_0 must be rejected.

მათემატიკა

განაწილების ფუნქციის შეფასება არაპირდაპირი შერჩევით. II

ე. ნადარაია*, პ. ბაბილუა**, გ. სოხაძე**

** აკადემიის წევრი, ი. ჯავახიშვილის სახ. თბილისის სახელმწიფო უნივერსიტეტი*
*** ი. ჯავახიშვილის სახ. თბილისის სახელმწიფო უნივერსიტეტი*

ნაშრომში დამტკიცებულია $\hat{F}_n(x)$ შეფასებასთან დაკავშირებული $C[a,1-a]$ სივრცეში უწყვეტი ფუნქციონალებისათვის ზღვარითი თეორემები.

REFERENCES

1. *E. Nadaraya, P. Babilua, G. Sokhadze* (2010), Bull. Georgian National Acad. Sci.*,* **4**, 3:
2. *P. Whittle* (1960), Teor. Veroyatnost. i Primenen. **5:** 331-335 (in Russian).
3. *I. I. Gihkman, A. V. Skorokhod* **(**1965), Vvedenie v teoriyu sluchainykh protsessov. M. (in Russian).

Received October, 2010