ON THE ESTIMATION OF DISTRIBUTION FUNCTION ON INDIRECT SAMPLE

Elizbar Nadaraya[1], Petre Babilua[2], Grigol Sokhadze[3]

Iv. Javakhishvili Tbilisi State University, Tbilisi, Georgia
[1]*elizbar.nadaraya@tsu.ge*, [2]*petre.babilua@tsu.ge*, [3]*grigol.sokhadze@tsu.ge*

1. Let $X_1, X_2, \dots, X_n$ be a sample of independent observations of a non-negative random value *X* with a distribution function $F(x)$. In problems of the theory of censored observations the sample values are pairs of random values $Y_i = (X_i \wedge t_i)$ and $Z_i = I(Y_i = X_i)$, $i = \overline{1,n}$, where t_i are given numbers ($t_i \neq t_j$ for $i \neq j$) or random values independent of X_i, $i = \overline{1,n}$. Throughout the paper $I(\cdot)$ denotes the indicator of the set *A*.

We will consider here several different cases: the observer has an access only to random values $\xi_i = I(X_i < t_i)$, $t_i = c_F \dfrac{2i-1}{2n}$, $i = \overline{1,n}$, $c_F = \inf\{x:\ F(x) = 1\} < \infty$.

The problem consists in estimating distribution functions $F(x)$ by the sample $\xi_1, \xi_2, \dots, \xi_n$. Such a problem arises, for example, in corrosion investigations (see [1] where an experiment connected with corrosion is described).

We will consider estimates for $F(x)$ that are analogous to regression curve estimates of Nadaraya-Watson type and have the form

$$\hat{F}_n(x) = F_{n1}(x) F_{n2}(x), \quad F_{n1}(x) = \sum_{i=1}^{n} K\left(\frac{x - t_i}{h}\right)\xi_i, \quad F_{n2}(x) = \left(\sum_{i=1}^{n} K\left(\frac{x - t_i}{h}\right)\right)^{-1},$$

where $K(x)$ is some weight function (kernel), $\{h = h(n)\}$ is a sequence of positive numbers converging to zero.

Lemma. *Assume that*

1^0. $K(x)$ *is some distribution density with bounded variation and* $K(x) = K(-x)$, $x \in R = (-\infty, \infty)$. *If* $nh \to \infty$, *then*

$$\frac{1}{nh}\sum_{j=1}^{n} K^{m_1-1}\left(\frac{x - t_j}{h}\right) F^{m_2-1}(t_j) = \frac{1}{c_F h}\int_0^{c_F} K^{m_1-1}\left(\frac{x-u}{h}\right) F^{m_2-1}(u)\,du + O\left(\frac{1}{nh}\right) \tag{2}$$

uniformly with respect to $x \in [0, c_F]$, m_1, m_2 *are natural.*

Without loss of generality we assume below that the interval $[0, c_F] = [0,1]$.

Theorem 1. *Let* $F(x)$ *be continuous and the conditions of the lemma be fulfilled. Then the estimate* $\hat{F}_n(x)$ *is asymptotically unbiased and consistent at all points* $x \in [0,1]$. *Moreover,* $\hat{F}_n(x)$ *is distributed asymptotically normally, i.e.*

$$\sqrt{nh}\left(\hat{F}_n(x) - E\hat{F}_n(x)\right)\sigma^{-1}(x) \xrightarrow{d} N(0,1),$$
$$\sigma^2(x) = F(x)(1 - F(x))\int K^2(u)\,du,$$

where d denotes convergence in distribution, and $N(0,1)$ *a random value having a normal distribution with mean 0 and variance* 1.

2. Uniform consistency. We define the conditions under which the estimate $\hat{F}_n(x)$ converges uniformly in probability (almost surely) to a true $F(x)$.

Following E. Parzen [2], we introduce the Fourier transform of $K(x)$

$$\varphi(t)=\int_{-\infty}^{\infty} e^{itx}K(x)dx$$

and assume that

2^0. $\varphi(t)$ is absolutely integrable. Then $F_{n1}(x)$ can be written in the form

$$F_{n1}(x)=\frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-iu\frac{x}{h}}\varphi(u)\frac{1}{nh}\sum_{j=1}^{n}\xi_j e^{iu\frac{t_j}{h}}du .$$

Thus

$$F_{n1}(x)-EF_{n1}(x)=\frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-iu\frac{x}{h}}\varphi(u)\frac{1}{nh}\sum_{j=1}^{n}\left(\xi_j-F(t_j)\right)e^{iu\frac{t_j}{h}}du .$$

Denote

$$d_n=\sup_{x\in\Omega_n}\left|\hat{F}_n(x)-E\hat{F}_n(x)\right|,\quad \Omega_n=\left[h^{\alpha},1-h^{\alpha}\right],\quad 0<\alpha<1 .$$

Theorem 2. *Let $K(x)$ satisfy conditions 1^0 and 2^0.*

(a) *Let $F(x)$ be continuous and $n^{\frac{1}{4}}h(n)\to\infty$, then*

$$D_n=\sup_{x\in\Omega_n}\left|\hat{F}_n(x)-F(x)\right|\xrightarrow{p}0 ;$$

(b) *If $\sum_{n=1}^{\infty}n^{-\frac{p}{4}}h(n)<\infty$, $p>4$, then $D_n\to 0$ almost surely.*

Proof. We have

$$\sup_{x\in\Omega_n}\left(1-\frac{1}{h}\int_0^1 K\left(\frac{x-u}{h}\right)du\right)\le\int_{-\infty}^{-h^{\alpha-1}}K(u)du+\int_{h^{\alpha-1}}^{\infty}K(u)du\to 0 \tag{1}$$

This and this lemma imply

$$\sup_{x\in\Omega_n}\left|F_{n2}(x)-1\right|\to 0 \tag{2}$$

i.e., due to uniform convergence, for any $\varepsilon_0>0$, $0<\varepsilon_0<1$, and sufficiently large $n\ge n_0$ we have $F_{n2}(x)\ge 1-\varepsilon_0$ uniformly with respect to $x\in\Omega_n$.

Denote

$$A_n=\sup_{u\in R}\left|\frac{1}{nh}\sum_{j=1}^{n}\eta_j e^{iu\frac{t_j}{h}}\right|,\quad \eta_j=\xi_j-F(t_j).$$

Then

$$d_n^p\le\frac{(1-\varepsilon_0)^{-p}}{(2\pi)^p}A_n^p\left|\int_{-\infty}^{\infty}|\varphi(u)|du\right|^p,\quad p>4 . \tag{3}$$

Note that

$$EA_n^p = \frac{1}{(nh)^p} E\sup_{u\in R}\left|\sum_{j=1}^{n}\eta_j^2 + \sum_{k\neq j}\eta_k\eta_j \cos\left(\left(\frac{t_k - t_j}{h}\right)u\right)\right|^{\frac{p}{2}} =$$

$$= \frac{1}{(nh)^p} E\sup_{u\in R}\left|\sum_{j=1}^{n}\eta_j^2 + \sum_{k\neq j}\eta_k\eta_j \cos\left(\frac{k-j}{nh}u\right)\right|^{\frac{p}{2}} =$$

$$= \frac{1}{(nh)^p} E\sup_{u\in R}\left|\sum_{j=1}^{n}\eta_j^2 + \sum_{m=-n+1}^{n-1}\sum_{\substack{j=1\\ m\neq 0}}^{n-|m|}\eta_j\eta_{j+m} \cos\left(\frac{m}{nh}u\right)\right|^{\frac{p}{2}}.$$

From this, by the inequality

$$\left|\sum_{j=1}^{n} a_j\right|^q \le n^{q-1}\sum_{j=1}^{n}\left|a_j\right|^q, \quad q\ge 1,$$

we have

$$EA_n^p \le \frac{2^{\frac{p}{2}-1}}{(nh)^p} E\left[\sum_{i=1}^{n}\eta_i^2\right]^{\frac{p}{2}} + \frac{2^{\frac{p}{2}-1}}{(nh)^p} E\sup_{u\in R}\left|\sum_{\substack{m=-n+1\\ m\neq 0}}^{n-1}\cos\left(\frac{m}{nh}u\right)\sum_{j=1}^{n-|m|}\eta_j\eta_{j+m}\right|^{\frac{p}{2}} = C_{n1} + C_{n2}. \qquad (4)$$

Let us estimate C_{n1} and C_{n2}:

$$C_{n1} \le \frac{2^{\frac{p}{2}-1}}{n^{\frac{p}{2}+1}h^p}\sum_{j=1}^{n} E\left|\eta_j\right|^p = \frac{2^{\frac{p}{2}-1}}{n^{\frac{p}{2}+1}h^p}\sum_{j=1}^{n}\left[\left(1-F(t_j)\right)^p F(t_j) + F^p(t_j)\left(1-F(t_j)\right)\right] \le c_2\frac{1}{n^{\frac{p}{2}}h^p}. \qquad (5)$$

Further, using Whittle's inequality [3] for moments of quadratic form, we obtain

$$C_{n2} \le \frac{2^{\frac{p}{2}-1}(2n-3)^{\frac{p}{2}-1}}{(nh)^p}\sum_{\substack{m=-n+1\\ m\neq 0}}^{n-1} E\left|\sum_{j=1}^{n-|m|}\eta_j\eta_{j+m}\right|^{\frac{p}{2}},$$

thus

$$E\left|\sum_{j=1}^{n-|m|}\eta_j\eta_{j+m}\right|^{\frac{p}{2}} \le c(p)\left(n-|m|\right)^{\frac{p}{4}},$$

where $c(p)$ depends only on p and $E\left|\eta_j\right|^p \le 1$.

Thus

$$\sum_{\substack{m=-n+1\\ m\neq 0}}^{n-1} E\left|\sum_{j=1}^{n-|m|}\eta_j\eta_{j+m}\right|^{\frac{p}{2}} \le 2c(p)\sum_{m=1}^{n-1} m^{\frac{p}{4}} = O\left(n^{\frac{p}{4}+1}\right)$$

and

$$C_{n2} = O\left(\frac{1}{n^{\frac{p}{4}}h^p}\right). \qquad (6)$$

After combining the relations (3), (4), (5) and (6), we obtain

$$Ed_n^p = O\left(\frac{1}{n^{\frac{p}{4}} h^p}\right), \quad p > 4.$$

Therefore

$$P\left\{\sup_{x\in\Omega_n}\left|\hat{F}_n(x) - E\hat{F}_n(x)\right| \geq \varepsilon\right\} \leq \frac{c_3}{\varepsilon^p n^{\frac{p}{4}} h^p}. \tag{7}$$

Further we obtain

$$\sup_{x\in\Omega_n}\left|E\hat{F}_n(x) - F(x)\right| \leq \frac{1}{1-\varepsilon_0}\left(\sup_{x\in\Omega_n}\left|EF_{n1}(x) - F(x)\right| + \sup_{x\in\Omega_n}\left|1 - F_{n2}(x)\right|\right). \tag{8}$$

The second summand in the right-hand part of (8) tends, by virtue of (2), to zero, while the first summand is estimated as follows:

$$\sup_{x\in\Omega_n}\left|EF_{n1}(x) - F(x)\right| \leq S_{n1} + S_{n2} + O\left(\frac{1}{nh}\right), \tag{9}$$

$$S_{n1} = \sup_{0\leq x\leq 1}\left|\frac{1}{h}\int_0^1 (F(y) - F(x)) K\left(\frac{x-y}{h}\right) dy\right|,$$

$$S_{n2} = \sup_{x\in\Omega_n}\left(1 - \frac{1}{h}\int_0^1 K\left(\frac{x-y}{h}\right) dy\right),$$

and, by virtue of (1),

$$S_{n2} \to 0. \tag{10}$$

Now let us consider S_{n1}. Note that

$$S_{n1} \leq \sup_{0\leq x\leq 1}\int_0^1 \left|F(y) - F(x)\right| \frac{1}{h} K\left(\frac{x-y}{h}\right) dy = \sup_{0\leq x\leq 1}\int_{x-1}^{x} \left|F(x-u) - F(x)\right| \frac{1}{h} K\left(\frac{u}{h}\right) du \leq$$

$$\leq \sup_{0\leq x\leq 1}\int_{-\infty}^{\infty} \left|F(x-u) - F(x)\right| \frac{1}{h} K\left(\frac{u}{h}\right) du. \tag{11}$$

Assume that $\delta > 0$ and divide the integration domain in (11) into two domains $|u| \leq \delta$ and $|u| > \delta$. Then

$$S_{n1} \leq \sup_{0\leq x\leq 1}\int_{|u|\leq\delta} \left|F(x-u) - F(x)\right| \frac{1}{h} K\left(\frac{u}{h}\right) du + \sup_{0\leq x\leq 1}\int_{|u|>\delta} \left|F(x-u) - F(x)\right| \frac{1}{h} K\left(\frac{u}{h}\right) du \leq$$

$$\leq \sup_{x\in R}\sup_{|u|\leq\delta}\left|F(x-u) - F(x)\right| + 2\int_{|u|\geq\frac{\delta}{h}} K(u) du. \tag{12}$$

By a choice of $\delta > 0$ the first summand in the right-hand part of (12) can be made arbitrarily small. After choosing $\delta > 0$ and making *n* tend to infinity, we obtain that the second summand tends to zero.

Thus

$$\lim_{n\to\infty} S_{n1} = 0. \tag{13}$$

Finally, the proof of the theorem follows from the relations (7)-(10) and (13).

Remarks.

1) If $K(x) = 0$, $|x| \geq 1$ and $\alpha = 1$, i.e., $\Omega_n = [h, 1-h]$, then $S_{n2} = 0$.

2) Under the conditions of Theorem 2,

$$\sup_{x\in[a,b]}\left|\hat{F}_n(x)-F(x)\right|\to 0$$

in probability (almost surely) for any fixed interval $[a,b]\subset[0,1]$ since there exists n_0 such that $[a,b]\subset\Omega_n$, $n\geq n_0$.

Let us assume that $h=n^{-\gamma}$, $\gamma>0$. The conditions of Theorem 2 are fulfilled:

$$n^{\frac{1}{4}}h_n\to\infty \text{ if } \gamma<\frac{1}{4}$$

and

$$\sum_{n=1}^{\infty}n^{-\frac{p}{4}}h_n^{-p}<\infty \text{ if } \gamma<\frac{1}{4}-\frac{1}{p},\ p>4.$$

3. Estimation of moments. In the considered problem there naturally arises the question of estimation of integral functional of $F(x)$, for example, of moments μ_m, $m\geq 1$:

$$\mu_m=m\int_0^1 t^{m-1}(1-F(t))dt.$$

As estimates for μ_m we will consider the statistics

$$\hat{\mu}_{nm}=1-\frac{m}{n}\sum_{j=1}^{n}\xi_j\frac{1}{h}\int_h^{1-h}t^{m-1}K\left(\frac{t-t_j}{h}\right)F_{n2}^{-1}(t)dt.$$

Theorem 3. *Let $K(x)$ satisfy condition 1^0 and, in addition to this, $K(x)=0$ outside the interval $[-1,1]$. If $nh\to\infty$ as $n\to\infty$, then $\hat{\mu}_{nm}$ is an asymptotically unbiased, consistent estimate for μ_m and, moreover,*

$$\frac{\sqrt{n}(\hat{\mu}_{nm}-E\hat{\mu}_{nm})}{\sigma}\xrightarrow{d}N(0,1),\quad \sigma^2=m^2\int_0^1 t^{2m-2}F(t)(1-F(t))dt.$$

References

1. Manjgaladze, K.V. On one estimate of a distribution function and its moments. (Russian) *Bull. Acad. Sci. Georgian SSR* **124** (1980), No. 2, pp. 261-268.
2. Parzen, E. On estimation of a probability density function and mode. *Ann. Math. Statist.* **33** (1962), pp. 1065-1076.
3. Whittle, P. Bounds for the moments of linear and quadratic forms in independent variables. *Teor. Verojatnost. i Primenen.* **5** (1960), pp. 331-335.