

Mathematics

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

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

* Academy Member, I. Javakhishvili Tbilisi State University

** I. Javakhishvili Tbilisi State University

ABSTRACT. The problem of estimation of a distribution function is considered when the observer has access only to some indicator random values. Some basic asymptotic properties of the constructed estimates are studied. © 2010 Bull. Georg. Natl. Acad. Sci.

Key words: distribution function estimate, unbiased, consistency, asymptotic normality, estimate of time moments.

Let X_1, X_2, \dots, X_n be a sample of independent observations of a nonnegative 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 $I(A)$ denotes the indicator of the set A .

We will consider here several different cases: the observer has access only to random values $\xi_i = I(X_i < t_i)$, $t_i = c_F \frac{2i-1}{2n}$, $i = \overline{1, n}$, $c_F = \inf\{x \geq 0: F(x) = 1\} < \infty$.

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).

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

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

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

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

where $K(x) \geq 0$ is some weight function (kernel) and $K(x) = K(-x)$, $-\infty < x < \infty$. $\{h = h(n)\}$ is a sequence of positive numbers converging to zero.

1. In this section, we give asymptotic unbiased and consistency conditions and theorem on a limiting distribution $\hat{F}_n(x)$.

Lemma. Assume that

1⁰. $K(x)$ is some distribution density with bounded variation. If $nh \rightarrow \infty$, then

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

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

Proof. Let $P(x)$ be a uniform distribution function on $[0, c_F]$ and $P_n(x)$ be an empirical distribution function of the "sample" t_1, t_2, \dots, t_n , i.e. $P_n(x) = n^{-1} \sum_{j=1}^n I(t_j < x)$. It is obvious that

$$\sup_{0 \leq x \leq c_F} |P_n(x) - P(x)| = \sup_{0 \leq x \leq c_F} \left| \frac{1}{n} \left[n \frac{x}{c_F} + \frac{1}{2} \right] - \frac{x}{c_F} \right| \leq \frac{1}{2n}. \quad (3)$$

We have

$$\begin{aligned} & \frac{1}{nh} \sum_{i=1}^n K^{m_1-1} \left(\frac{x-t_i}{h} \right) F^{m_2-1}(t_i) - \frac{1}{c_F h} \int_0^{c_F} K^{m_1-1} \left(\frac{x-u}{h} \right) F^{m_2-1}(u) du = \\ & = \frac{1}{h} \int_0^{c_F} K^{m_1-1} \left(\frac{x-u}{h} \right) F^{m_2-1}(u) d(P_n(u) - P(u)). \end{aligned} \quad (4)$$

Applying the integration by parts of formula and taking (3) into account, we obtain (2) from (4).

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 (1) is asymptotically unbiased and consistent at all points $x \in [0, 1]$. Moreover, $\hat{F}_n(x)$ is distributed asymptotically normally, i.e.

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

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

Proof. From the lemma we have

$$E\hat{F}_n(x) = \int_{\frac{x-1}{h}}^{\frac{x}{h}} K(t) F(x+ht) dt + O\left(\frac{1}{nh}\right), \quad F_{2n}(x) = \frac{1}{h} \int_0^1 K\left(\frac{x-u}{h}\right) du + O\left(\frac{1}{nh}\right), \quad (5)$$

and, as $n \rightarrow \infty$,

$$\frac{1}{h} \int_0^1 K\left(\frac{x-u}{h}\right) du \rightarrow F_2(x) = \begin{cases} 1, & x \in (0,1) \\ \frac{1}{2}, & x = 0, \quad x = 1 \end{cases}$$

$$\int_{\frac{x-1}{h}}^{\frac{x}{h}} K(t) F(x+th) dt \rightarrow F(x) \cdot F_2(x).$$

Hence it follows that $E\hat{F}_n(x) \rightarrow F(x)$, $x \in [0,1]$ as $n \rightarrow \infty$.

Analogously, it is not difficult to show that

$$Var \hat{F}_n(x) = \left[\frac{1}{nh^2} \int_0^1 K^2\left(\frac{x-u}{h}\right) F(u)(1-F(u)) du + O\left(\frac{1}{(nh)^2}\right) \right] F_{2n}^{-2}(x).$$

This readily implies that

$$nh \text{ Var } \hat{F}_n(x) \sim \sigma^2(x) = F(x)(1-F(x)) \int K^2(u) du \tag{6}$$

as $x \in [0,1]$.

Thus $\hat{F}_n(x)$ is a consistent estimate for $F(x)$, $x \in [0,1]$, and therefore, $P\{\hat{F}_n(x_1) \leq \hat{F}_n(x_2)\} \rightarrow 1$, $x_1 < x_2$, $x_1, x_2 \in [0,1]$.

Now we will establish that $\hat{F}_n(x)$ is distributed asymptotically normally. Since by virtue of (5), $F_{2n}(x) \rightarrow F_2(x)$, it remains for us to verify the condition of Lyapunov's Central Limit Theorem for $F_{1n}(x)$. Let us denote

$\eta_i = \eta_i(x) = (nh)^{-1} K\left(\frac{x-t_i}{h}\right) \xi_i$ and show that

$$L_n = \sum_{j=1}^n E|\eta_j - E\eta_j|^{2+\delta} (Var F_{1n}(x))^{-1-\frac{\delta}{2}} \rightarrow 0, \quad \delta > 0. \tag{7}$$

We have

$$\sum_{j=1}^n E|\eta_j - E\eta_j|^{2+\delta} \leq 2M^{1+\delta} (nh)^{-(2+\delta)} \sum_{j=1}^n K\left(\frac{x-t_j}{h}\right) F(t_j), \quad M = \max_{x \in R} K(x).$$

Taking (2) into account, from this inequality we find

$$\sum_{j=1}^n E|\eta_j - E\eta_j|^{2+\delta} \leq c_1 (nh)^{-(1+\delta)}. \tag{8}$$

Using the relation (6) and the inequality (8) we obtain $L_n = O\left((nh)^{-\frac{\delta}{2}}\right)$, which means that (7) holds.

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

We introduce the Fourier transform of $K(x)$:

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

and assume that

2⁰. $\varphi(t)$ is absolutely integrable. Following E. Parzen [2] $F_{1n}(x)$ can be written in the form

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

Thus

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

Denote

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

Theorem 2. Let $K(x)$ satisfy conditions 1⁰ and 2⁰.

(a) Let $F(x)$ be continuous and $\frac{1}{n^2 h_n} \rightarrow \infty$, then

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

(b) If $\sum_{n=1}^{\infty} n^{-\frac{p}{2}} h^{-p} < \infty$, $p > 2$, then $D_n \rightarrow 0$ almost surely.

Proof. We have

$$\sup_{x \in \Omega_n} \left(1 - \frac{1}{h} \int_0^1 K\left(\frac{x-u}{h}\right) du \right) \leq \int_{-\infty}^{-h^{\alpha-1}} K(u) du + \int_{h^{\alpha-1}}^{\infty} K(u) du \rightarrow 0. \quad (9)$$

This and (5) imply

$$\sup_{x \in \Omega_n} |F_{2n}(x) - 1| \rightarrow 0 \quad (10)$$

i.e., due to uniform convergence, for any $\varepsilon_0 > 0$, $0 < \varepsilon_0 < 1$, and sufficiently large $n \geq n_0$ we have $F_{2n}(x) \geq 1 - \varepsilon_0$ uniformly with respect to $x \in \Omega_n$.

Therefore

$$d_n \leq (1 - \varepsilon_0)^{-1} \sup_{x \in \Omega_n} |F_{1n}(x) - EF_{1n}(x)| \leq (1 - \varepsilon_0)^{-1} \cdot \frac{1}{2\pi} \int |\varphi(u)| \frac{1}{nh} \left| \sum_{j=1}^n \bar{\eta}_j e^{iu \frac{t_j}{h}} \right| du, \quad \bar{\eta}_j = \xi_j - F(t_j),$$

From here owing to Gelder's inequality, we have

$$d_n^p \leq (1 - \varepsilon_0)^{-p} \frac{1}{(2\pi)^p} \frac{1}{(nh)^p} \int |\varphi(u)| \left| \sum_{j=1}^n \bar{\eta}_j e^{iu \frac{t_j}{h}} \right|^p du \left(\int |\varphi(u)| du \right)^{\frac{p}{q}}, \quad \frac{1}{p} + \frac{1}{q} = 1, \quad p > 2.$$

Therefore

$$Ed_n^p \leq c(\varepsilon, p, \varphi) \frac{1}{(nh)^p} \int |\varphi(u)| E \left| \sum_{j,k} \cos\left(\left(\frac{t_j - t_k}{h}\right)u\right) \bar{\eta}_j \bar{\eta}_k \right|^{\frac{p}{2}} du, \tag{11}$$

where

$$c(\varepsilon, p, \varphi) = (1 - \varepsilon_0)^{-p} \frac{1}{(2\pi)^p} \left(\int |\varphi(u)| du \right)^{\frac{p}{2}}.$$

Denote

$$A(u) = \sum_{j,k} \cos\left(\left(\frac{t_j - t_k}{h}\right)u\right) \bar{\eta}_j \bar{\eta}_k.$$

Then from (11) we can write

$$Ed_n^p \leq 2^{\frac{p}{2}-1} c(\varepsilon_0, p, \varphi) \frac{1}{(nh)^p} \left[\int |\varphi(u)| |EA(u)|^{\frac{p}{2}} du + \int |\varphi(u)| E|A(u) - EA(u)|^{\frac{p}{2}} du \right]. \tag{12}$$

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

$$E|A(u) - EA(u)|^{\frac{p}{2}} \leq 2^{\frac{3}{2}p} c\left(\frac{p}{2}\right) [c(p)]^{\frac{1}{2}} \left(\sum_{i,j} \cos^2\left(\left(\frac{t_j - t_k}{h}\right)u\right) \gamma_j^2(p) \lambda^2(p) \right)^{\frac{p}{4}},$$

where

$$\gamma_k(p) = \left(E|\bar{\eta}_k|^p \right)^{\frac{1}{p}} \leq 1, \quad c(s) = \frac{2^{\frac{s}{2}}}{\sqrt{\pi}} \Gamma\left(\frac{s+1}{2}\right).$$

From here follows

$$E|A(u) - EA(u)|^{\frac{p}{2}} = O\left(n^{\frac{p}{2}}\right), \tag{13}$$

uniformly with respect to $u \in (-\infty, \infty)$, and also clear that,

$$|EA(u)|^{\frac{p}{2}} = O\left(n^{\frac{p}{2}}\right), \tag{14}$$

uniformly with respect to $u \in (-\infty, \infty)$. After combining the relations (12), (13) and (14), we obtain

$$Ed_n^p = O\left(\frac{1}{(\sqrt{n} h)^p}\right), \quad p > 2.$$

Therefore

$$P\left\{ \sup_{x \in \Omega_n} |\hat{F}_n(x) - EF_n(x)| \geq \varepsilon \right\} \leq \frac{c_3}{\varepsilon^p (\sqrt{n} h)^p}. \tag{15}$$

Further we obtain

$$\sup_{x \in \Omega_n} |E\hat{F}_n(x) - F(x)| \leq \frac{1}{1 - \varepsilon_0} \left(\sup_{x \in \Omega_n} |EF_{1n}(x) - F(x)| + \sup_{x \in \Omega_n} |1 - F_{2n}(x)| \right). \tag{16}$$

The second summand in the right-hand part of (16) tends, by virtue of (10), to zero, while the first summand is estimated as follows:

$$\sup_{x \in \Omega_n} |EF_{1n}(x) - F(x)| \leq S_{1n} + S_{2n} + O\left(\frac{1}{nh}\right), \quad (17)$$

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

$$S_{2n} = \sup_{x \in \Omega_n} \left(1 - \frac{1}{h} \int_0^1 K\left(\frac{x-y}{h}\right) dy \right),$$

and, by virtue of (9),

$$S_{2n} \rightarrow 0. \quad (18)$$

Now let us consider S_{1n} . Note that

$$\begin{aligned} S_{1n} &\leq \sup_{0 \leq x \leq 1} \int_0^1 |F(y) - F(x)| \frac{1}{h} K\left(\frac{x-y}{h}\right) dy = \sup_{0 \leq x \leq 1} \int_{x-1}^x |F(x-u) - F(x)| \frac{1}{h} K\left(\frac{u}{h}\right) du \leq \\ &\leq \sup_{0 \leq x \leq 1} \int_{-\infty}^{\infty} |F(x-u) - F(x)| \frac{1}{h} K\left(\frac{u}{h}\right) du. \end{aligned} \quad (19)$$

Assume that $\delta > 0$ and divide the integration domain in (19) into two domains $|u| \leq \delta$ and $|u| > \delta$. Then

$$\begin{aligned} S_{1n} &\leq \sup_{0 \leq x \leq 1} \int_{|u| \leq \delta} |F(x-u) - F(x)| \frac{1}{h} K\left(\frac{u}{h}\right) du + \sup_{0 \leq x \leq 1} \int_{|u| > \delta} |F(x-u) - F(x)| \frac{1}{h} K\left(\frac{u}{h}\right) du \leq \\ &\leq \sup_{x \in R} \sup_{|u| \leq \delta} |F(x-u) - F(x)| + 2 \int_{|u| \geq \frac{\delta}{h}} K(u) du. \end{aligned} \quad (20)$$

By a choice of $\delta > 0$ the first summand in the right-hand part of (20) 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

$$\lim_{n \rightarrow \infty} S_{1n} = 0. \quad (21)$$

Finally, the proof of the theorem follows from the relations (15)-(18) and (21).

Remarks.

- 1) If $K(x) = 0$, $|x| \geq 1$ and $\alpha = 1$, i.e., $\Omega_n = [h, 1-h]$, then $S_{2n} = 0$.
- 2) Under the conditions of Theorem 2,

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

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:

$$\frac{1}{n^2 h_n} \rightarrow \infty \text{ if } \gamma < \frac{1}{2}$$

and

$$\sum_{n=1}^{\infty} n^{-\frac{p}{2}} h_n^{-p} < \infty \text{ if } 0 < \gamma < \frac{p-2}{2p}, \quad p > 2.$$

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

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

As estimates for μ_m we will consider the statistics

$$\hat{\mu}_{nm} = 1 - \frac{m}{n} \sum_{j=1}^n \xi_j \frac{1}{h} \int_h^{1-h} t^{m-1} K\left(\frac{t-t_j}{h}\right) F_{2n}^{-1}(t) dt.$$

Theorem 3. Let $K(x)$ satisfy condition 1⁰ and, in addition to this, $K(x)=0$ outside the interval $[-1,1]$. If $nh \rightarrow \infty$ as $n \rightarrow \infty$, then $\hat{\mu}_{nm}$ is an asymptotically unbiased, consistent estimate for μ_m and, moreover,

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

Proof. Since $K(x)$ has $[-1,1]$ as a carrier, from (5) it follows that

$$F_{2n}(n) = 1 + O\left(\frac{1}{nh}\right)$$

uniformly with respect to $x \in [h, 1-h]$.

From this and the lemma we have

$$\begin{aligned} E\hat{\mu}_{nm} &= 1 - \frac{m}{n} \sum_{j=1}^n F(t_j) \frac{1}{h} \int_h^{1-h} t^{m-1} K\left(\frac{t-t_j}{h}\right) F_{2n}^{-1}(t) dt = 1 - m \int_h^{1-h} \left[\frac{1}{h} \int_0^1 K\left(\frac{t-u}{h}\right) F(u) du \right] t^{m-1} dt + O\left(\frac{1}{nh}\right) = \\ &= 1 - m \int_h^{1-h} \left(\int_{-1}^1 K(v) F(t+vh) dv \right) t^{m-1} dt + O\left(\frac{1}{nh}\right) = 1 - m \int_0^1 t^{m-1} \left[\int_{-1}^1 K(u) F(t+vh) dv \right] dt + O(h) + O\left(\frac{1}{nh}\right). \end{aligned} \quad (22)$$

By Lebesgue's theorem on majorized convergence, from (22) it follows that

$$E\hat{\mu}_{nm} \rightarrow 1 - m \int_0^1 F(t) t^{m-1} dt = m \int_0^1 t^{m-1} (1-F(t)) dt = \mu_m, \quad m \geq 1. \quad (23)$$

Therefore $\hat{\mu}_{nm}$ is an asymptotically unbiased estimate for μ_m .

Further, analogously to (22) it can be shown that

$$Var \hat{\mu}_{nm} = \frac{m^2}{n} \int_0^1 F(t)(1-F(t)) t^{2m-2} \left[\mathcal{K}\left(\frac{1-t}{h} - 1\right) - \mathcal{K}\left(1 - \frac{t}{h}\right) \right]^2 dt + O\left(\frac{h}{n}\right) + O\left(\frac{1}{(nh)^2}\right),$$

where

$$\mathcal{K}(v) = \int_{-\infty}^v K(u) du.$$

By the same Lebesgue's theorem we see that

$$n \operatorname{Var} \hat{\mu}_{nm} \sim \sigma^2 = m^2 \int_0^1 t^{2m-2} F(t)(1-F(t)) dt. \quad (24)$$

Therefore (23) and (24) imply that $\hat{\mu}_{nm} \xrightarrow{P} \mu_m$.

To complete the proof of the theorem it remains to show that the statistics $\sqrt{n}(\hat{\mu}_{nm} - E\hat{\mu}_{nm})$ are asymptotically distributed normally with mean 0 and variance σ^2 . For this it suffices to show that the Lyapunov fraction $L_n \rightarrow 0$. Indeed,

$$\begin{aligned} L_n &= n^{-(2+\delta)} m^{2+\delta} \sum_{j=1}^n |\xi_j - F(t_j)|^{2+\delta} \left| \frac{1}{h} \int_h^{1-h} t^{m-1} K\left(\frac{t-t_j}{h}\right) F_{2n}^{-1} dt \right|^{2+\delta} (\operatorname{Var} \hat{\mu}_{nm})^{-\left(1+\frac{\delta}{2}\right)} \leq \\ &\leq c_\delta n^{-(2+\delta)} \sum_{j=1}^n |\xi_j - F(t_j)|^{2+\delta} (\operatorname{Var} \hat{\mu}_{nm})^{-\left(1+\frac{\delta}{2}\right)} \leq c_7 n^{-1-\delta} (\operatorname{Var} \hat{\mu}_{nm})^{-\left(1+\frac{\delta}{2}\right)} = O\left(n^{-\frac{\delta}{2}}\right). \end{aligned}$$

The theorem is proved.

მათემატიკა

განაწილების ფუნქციის შეფასება არაპირდაპირი შერჩევით. I

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

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

** ი. ჯავახიშვილის სახ. თბილისის სახელმწიფო უნივერსიტეტი

ნაშრომში აგებულია განაწილების ფუნქციის შეფასება, როდესაც დამკვირვებლისთვის მისაწვდომია ზოგიერთი ინდიკატორული შემთხვევითი სიდიდის მნიშვნელობები. შესწავლილია აგებული შეფასებების ზოგიერთი ძირითადი თვისება.

REFERENCES

1. K. V. Manjgaladze (1980), Bull. Acad. Sci. Georg. SSR, **124**, 2: 261-268 (in Russian).
2. E. Parzen (1962), Ann. Math. Statist., **33**: 1065-1076.
3. P. Whittle (1960), Teor. Veroyatnost. i Primenen., **5**: 331-335.

Received July, 2010