Slajdy do wykładu

Ryszard Zieliński

ESTYMATORY DYSTRYBUANTY Z ZADANĄ PRECYZJĄ OPARTE NA FUNKCJACH WIELOMIANOWYCH I SKLEJANYCH

Wydział Matematyki i Informatyki UAM 27 listopada 2008

Zakład Statystyki Matematycznej IMPAN 18 grudnia 2008

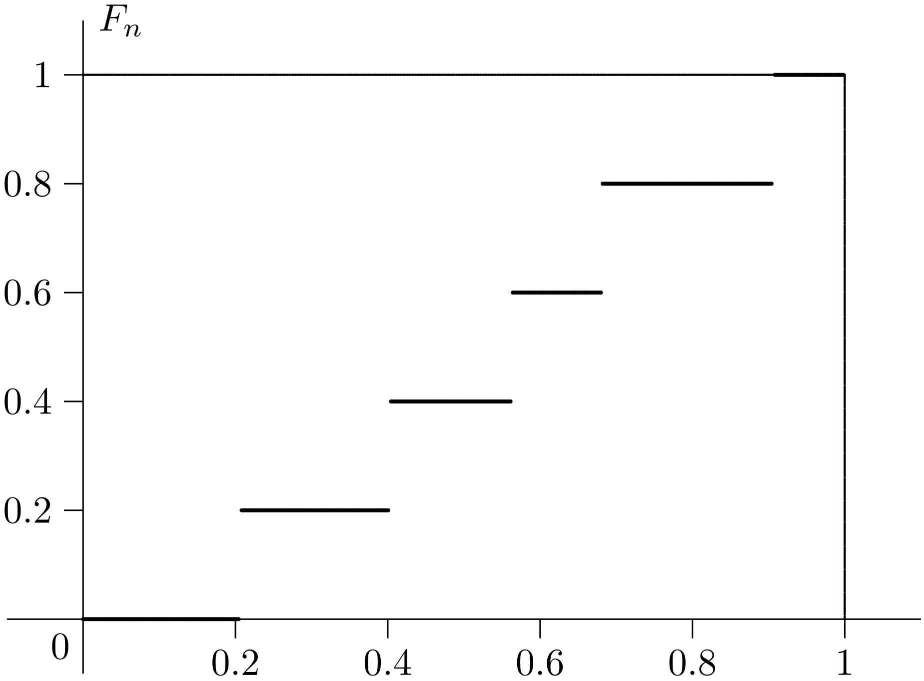
Wyniki tu zaprezentowane pochodzą z pracy:

Zbigniew Ciesielski i Ryszard Zieliński: Polynomial and spline estimators of the distribution function with prescribed accuracy.

Praca przyjęta do *Applicationes Mathematicae*

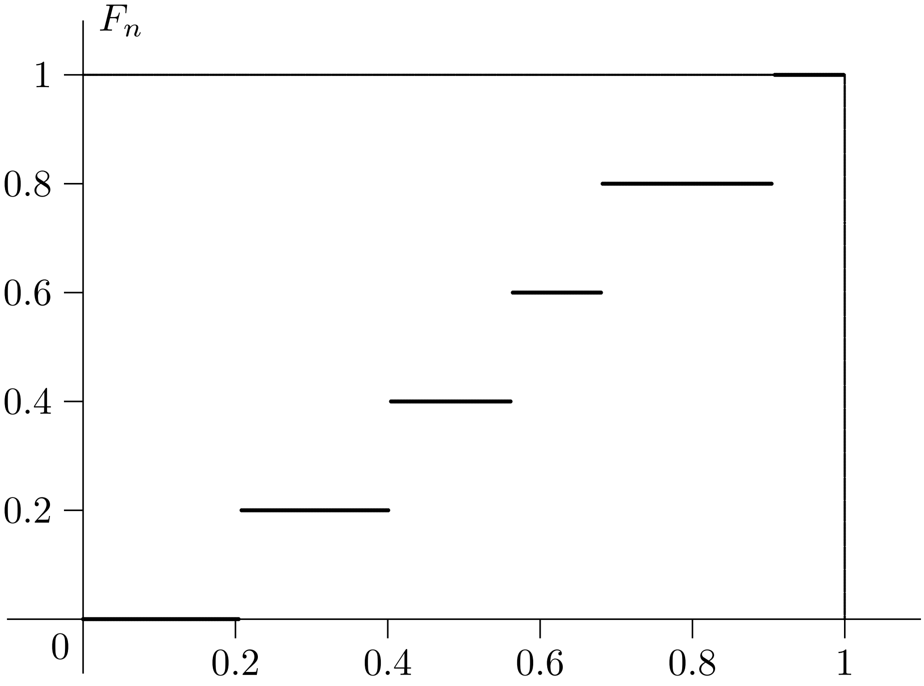
$X_1, X_2, \ldots, X_n$ – iid rv $\sim F$

$$F_n(x) = \frac{1}{n}\sum_{j=1}^{n} 1_{(-\infty,x]}(X_j)$$



$X_1, X_2, \dots, X_n$ – iid rv $\sim F$

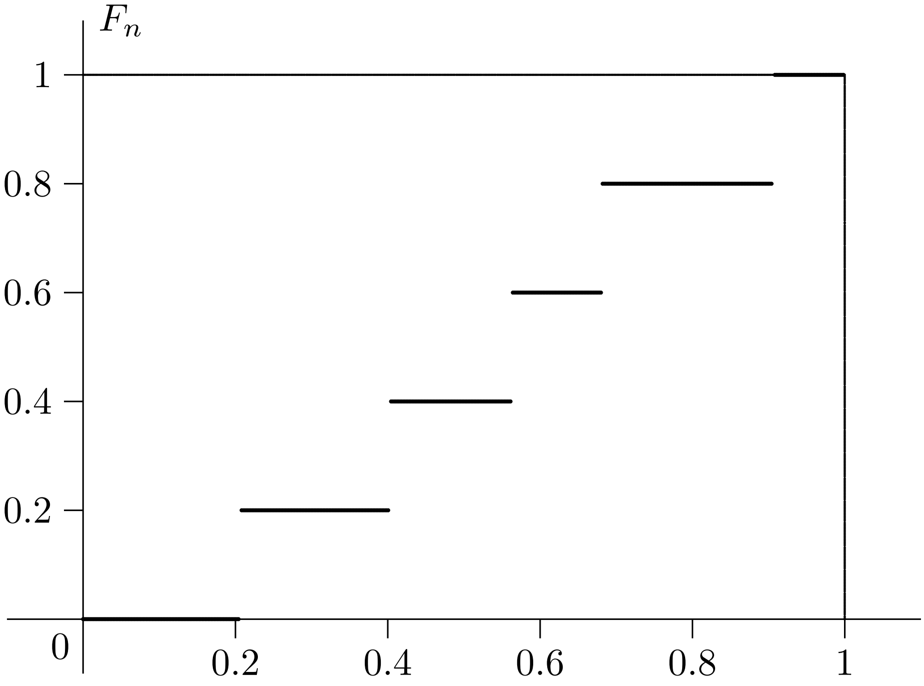
$$F_n(x) = \frac{1}{n} \sum_{j=1}^{n} 1_{(-\infty,x]}(X_j)$$



$$(\forall \varepsilon) \quad P_F\{\sup_{x\in\mathbf{R}} |F_n(x) - F(x)| \geq \varepsilon\} \to 0$$

$X_1, X_2, \ldots, X_n$ – iid rv $\sim F$

$$F_n(x) = \frac{1}{n}\sum_{j=1}^{n} 1_{(-\infty,x]}(X_j)$$

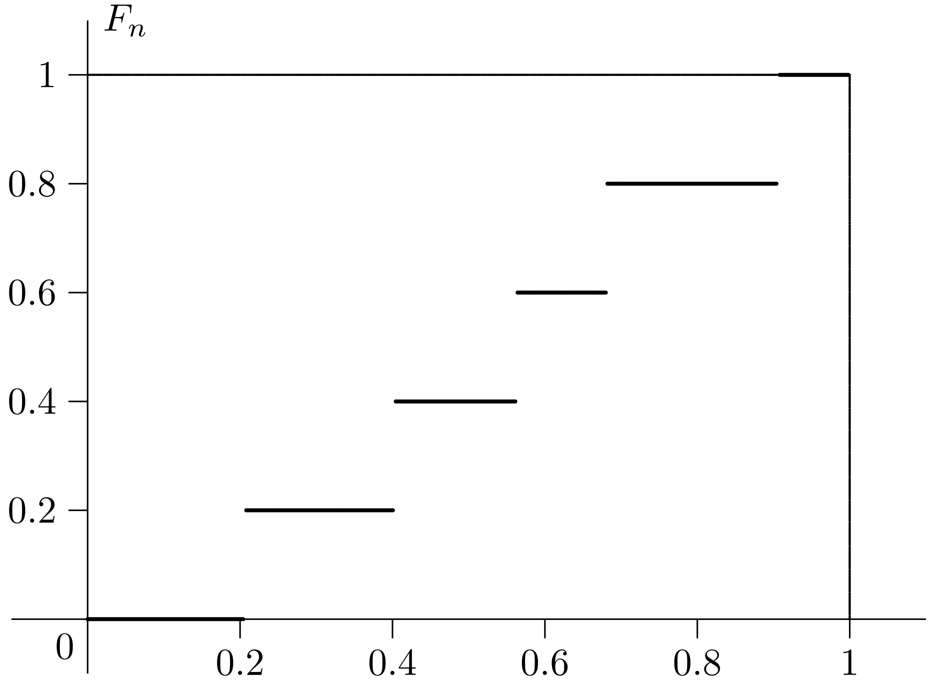


$$(\forall \varepsilon) \quad P_F\{\sup_{x\in\mathbf{R}} |F_n(x) - F(x)| \geq \varepsilon \to 0$$

$$(\forall \varepsilon)(\forall \eta)(\exists N)(\forall n \geq N)(\forall F \in \mathcal{F}) \quad P_F\{\sup_{x\in\mathbf{R}} |F_n(x) - F(x)| \geq \varepsilon\} \leq \eta$$

$X_1, X_2, \ldots, X_n$ – iid rv $\sim F$

$$F_n(x) = \frac{1}{n} \sum_{j=1}^{n} 1_{(-\infty, x]}(X_j)$$

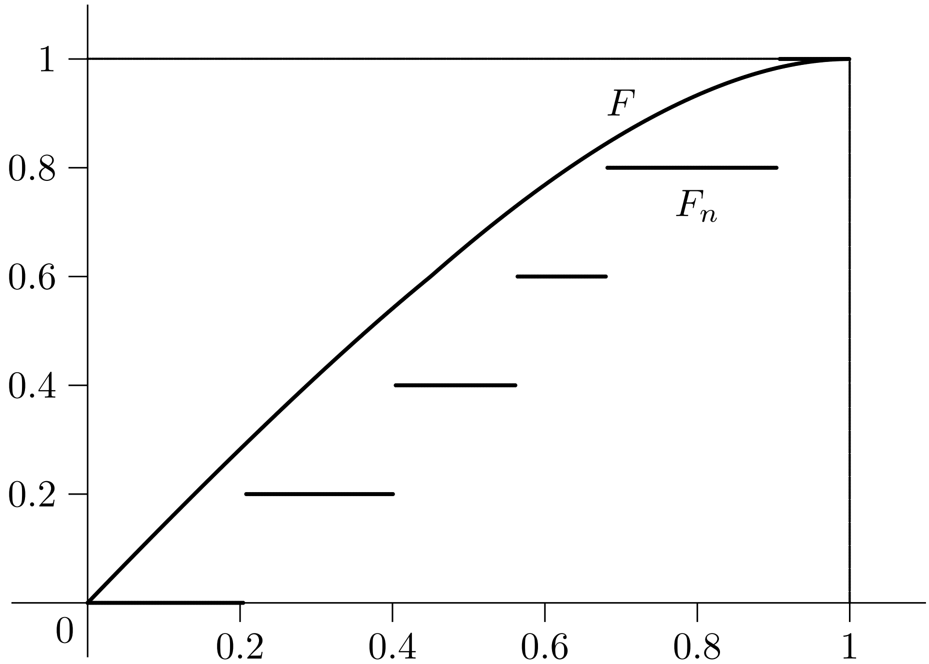


$$(\forall \varepsilon) \quad P_F\{\sup_{x \in \mathbf{R}} |F_n(x) - F(x)| \geq \varepsilon \to 0$$

$$(\forall \varepsilon)(\forall \eta)(\exists N)(\forall n \geq N)(\forall F \in \mathcal{F}) \quad P_F\{\sup_{x \in \mathbf{R}} |F_n(x) - F(x)| \geq \varepsilon\} \leq \eta$$

$$(\forall \varepsilon)(\forall \eta)(\exists N)(\forall n \geq N) \quad \sup_{F \in \mathcal{F}} P_F\{\sup_{x \in \mathbf{R}} |F_n(x) - F(x)| \geq \varepsilon\} \leq \eta$$

DVORETZKY-KIEFER-WOLFOWITZ



$$(\forall \varepsilon > 0)(\forall n)(\forall F \in \mathcal{F})$$

$$P_F\{\sup_{x \in \mathbf{R}} |F_n(x) - F(x)| \geq \varepsilon\} \leq 2e^{-2n\varepsilon^2}$$

$\mathcal{F}$ - rodzina wszystkich dystrybuant ciągłych i ściśle rosnących

DVORETZKY-KIEFER-WOLFOWITZ

$$(\forall \varepsilon > 0)(\forall n)(\forall F \in \mathcal{F})$$

$$P_F\{\sup_{x \in \mathbf{R}} |F_n(x) - F(x)| \geq \varepsilon\} \leq 2e^{-2n\varepsilon^2}$$

Birnbaum (1952) - tablice i przykład: *We wish to approximate $F(x)$ empirically by $F_N(x)$ so that the error is everywhere less than .15, on the 90% probability level. How large must be the sample N? To answer this question, we find by interpolation in Table 1 that $P\{D_{65}\} > .900$, so that $N = 65$ is sufficient.*

Dvoretzky at al. (1956):

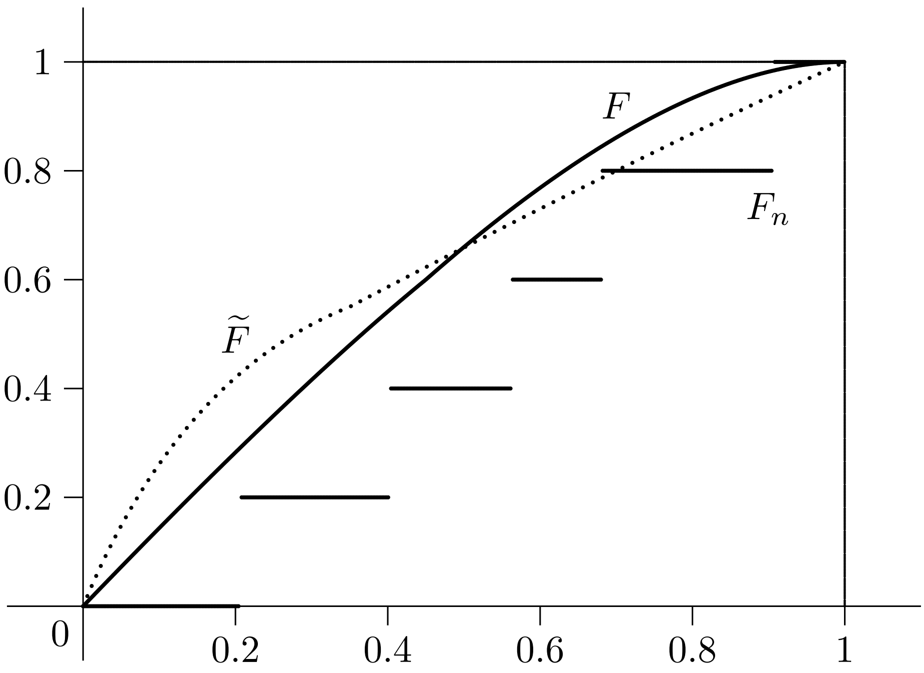
$$(*) \qquad P\{\sup_x |F_n(x) - F(x)| \geq \lambda\} \leq C \cdot \exp(-2n\lambda^2), C - \text{stała uniwersalna},$$

Gaensler at al. (2006): "$C = 58$ works; the smallest C for which $(*)$ holds is still unknown"

Massart, P. (1990). The tight constant in the Dvoretzky–Kiefer–Wolfowitz inequality. *Annals of Probability*, 18: 1269–1283

$$\varepsilon = 0.1, \quad \eta = 0.1 \quad \Rightarrow \quad n = 150$$

$$\varepsilon = 0.01, \quad \eta = 0.01 \quad \Rightarrow \quad n = 26492$$

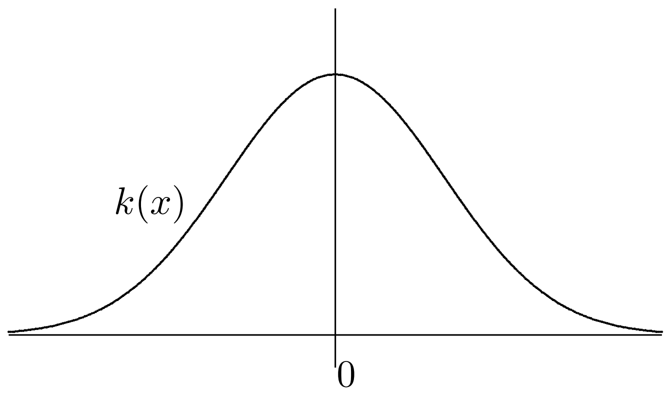


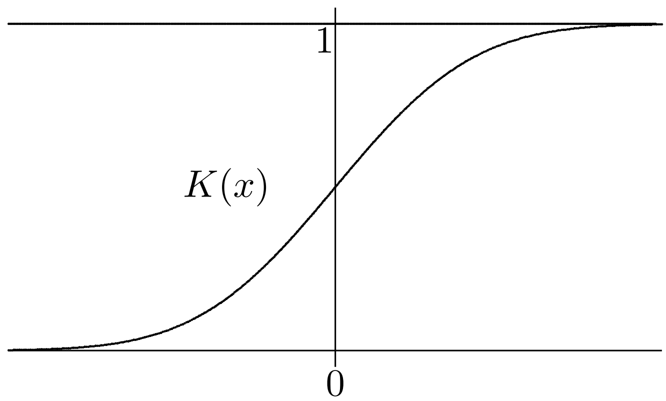
1. ESTYMATORY JĄDROWE
2. ESTYMATORY WIELOMIANOWE
3. ESTYMATORY SPLAJNOWE

1. ESTYMATORY JĄDROWE

$$\widehat{F}_n(x) = \frac{1}{n}\sum_{j=1}^{n} K\Big(\frac{x - X_j}{h_n}\Big).$$

$$K(x) = \int_{-\infty}^{x} k(t)dt, \qquad \widehat{f}_n(x) = \frac{1}{n}\sum_{j=1}^{n} \frac{1}{h_n} k\Big(\frac{x - X_j}{h_n}\Big)$$





TWIERDZENIE NEGATYWNE (Zieliński 2007): *Jeżeli*

– *K jest dowolnym jądrem (zcałkowanym) takim, że* $0 < K(0) < 1$ *oraz* $K^{-1}(t) < 0$ *dla pewnego* $t \in (0, K(0))$

– $(h_n, n = 1, 2, \ldots)$ *jest dowolnym ciągiem liczb dodatnich*

to

istnieją takie $\varepsilon > 0$ *oraz* $\eta > 0$,

że dla każdego n *znajdzie się rozkład* $F \in \mathcal{F}$ *taki, że*

$$P_F\{\sup_{x\in\mathbf{R}} |\widetilde{F}_n(x) - F(x)| \geq \varepsilon\} \geq \eta$$

DOWÓD

Wystarczy udowodnić, że

$$(\exists\varepsilon)(\exists\eta)(\forall n)(\exists F \in \mathcal{F}) \quad P\{\widehat{F}_n(0) > F(0) + \varepsilon\} \geq \eta$$

Dowodzę, że

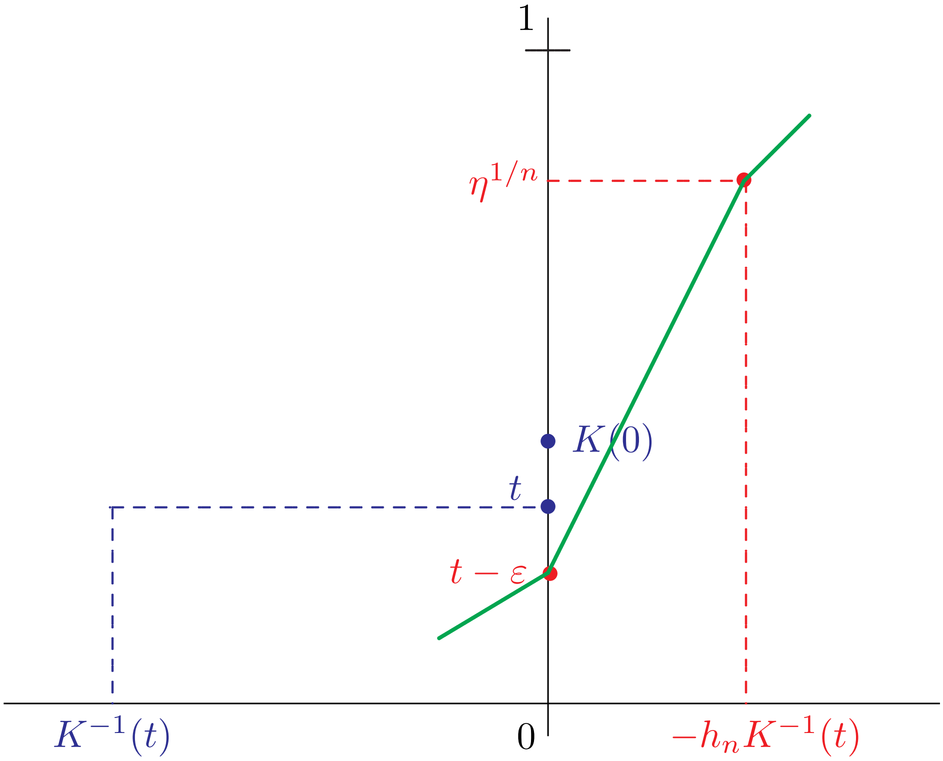
$$(\exists\varepsilon)(\exists\eta)(\forall n)(\exists F \in \mathcal{F}) \quad P\{\widehat{F}_n(0) > F(0) + \varepsilon\} \geq \eta$$

Ustalamy $t \in (0, K(0))$, takie że $K^{-1}(t) < 0$

Ustalamy $\varepsilon \in (0, t)$ oraz $\eta \in (t - \varepsilon, 1)$.

Dla ustalonych ε, η oraz n biorę F takie, że

$F(0) = t - \varepsilon$ and $F\big(-h_n K^{-1}(t)\big) > \eta^{1/n}$.



Wtedy

$$P_F\{X_j < -h_n K^{-1}(t)\} > \eta^{1/n} \quad czyli \quad P_F\{K\left(-\frac{X_j}{h_n}\right) > t\} > \eta^{1/n}$$

Ponieważ

$$\bigcap_{j=1}^{n}\left\{K\left(-\frac{X_j}{h_n}\right) > t\right\} \subset \left\{\frac{1}{n}\sum_{j=1}^{n} K\left(-\frac{X_j}{h_n}\right) > t\right\}$$

to

$$P_F\left\{\frac{1}{n}\sum_{j=1}^{n} K\left(-\frac{X_j}{h_n}\right) > t\right\} = P_F\left\{\underbrace{\frac{1}{n}\sum_{j=1}^{n} K\left(-\frac{X_j}{h_n}\right)}_{\widehat{F}_n(0)} > F(0) + \varepsilon\right\} > \eta$$

cbdo

1. ESTYMATORY JĄDROWE (c.d.)

TWIERDZENIE POZYTYWNE (Zieliński 2007)
(konstrukcja estymatora z losową szerokością okna)

$X_1, X_2, \ldots, X_n$ - próba z rozkładu $F \in \mathcal{F}$
$X_{1:n} \leq X_{2:n} \leq \ldots \leq X_{n:n}$ - statystyka pozycyjna z tej próby

$$H_n = \min\{X_{j:n} - X_{j-1:n},\ j = 2, 3, \ldots, n\}$$

Estymator jądrowy

$$\widetilde{F}_n(x) = \frac{1}{n} \sum_{j=1}^{n} K\Big(\frac{x - X_j}{H_n}\Big)$$

gdzie

$$K(t) = \begin{cases} 0, & \text{for } t \leq -1/2, \\ 1, & \text{for } t \geq 1/2, \end{cases}$$

$K(0) = 1/2$, $K(t)$ ciągłe i niemalejące w $(-1/2, 1/2)$.

Wtedy

$$P_F\{\sup_{x \in \mathbf{R}} |\widetilde{F}_n(x) - F(x)| \geq \varepsilon\} \leq 2e^{-2n(\varepsilon - 1/2n)^2}, \quad n > \frac{1}{2\varepsilon}, \quad F \in \mathcal{F}$$

TWIERDZENIE POZYTYWNE dowód

Dla k oraz $j = 1, 2, \ldots, n$ mamy

$$K\Big(\frac{X_{k:n} - X_{j:n}}{H_n}\Big) =$$

$$= \begin{cases} 0, & \text{for } \frac{X_{k:n} - X_{j:n}}{H_n} \le -\frac{1}{2} \Leftrightarrow X_{j:n} > X_{k:n} + \frac{1}{2}H_n \Leftrightarrow j > k \\ \frac{1}{2}, & \text{for } t = 0 \\ 1, & \text{for } j < k \end{cases}$$

Zatem

$$\widetilde{F}_n(X_{k:n}) = \frac{1}{n}\sum_{j=1}^{n} K\Big(\frac{X_{k:n} - X_{j:n}}{H_n}\Big) = \frac{k-1}{n} + \frac{1}{2n}$$

$$= F_n(X_{k-1:n}) + \frac{1}{2n} = F_n(X_{k:n}) - \frac{1}{2n}$$

Więc, dla $k = 1, 2, \ldots, n$, mamy $|\widetilde{F}_n(X_{k:n}) - F_n(X_{k:n})| = \dfrac{1}{2n}$.

Dla $k = 1, 2, \ldots, n$, mamy $|\widetilde{F}_n(X_{k:n}) - F_n(X_{k:n})| = \dfrac{1}{2n}$.

Estymator jądrowy $\widetilde{F}_n(x)$ jest ciągły i rosnący, dystrybuanta empiryczna $F_n(x)$ jest funkcją schodkową, więc $|\widetilde{F}_n(x) - F_n(x)| \leq \dfrac{1}{2n}$ dla wszystkich $x \in (-\infty, \infty)$.

Na mocy nierówności trójkąta

$$|\widetilde{F}_n(x) - F(x)| \leq |F_n(x) - F(x)| + \frac{1}{2n}$$

otrzymujemy

$$P\{\sup_{x \in \mathbf{R}} |\widetilde{F}_n(x) - F(x)| \geq \varepsilon\} \leq P\{\sup_{x \in \mathbf{R}} |F_n(x) - F(x)| + \frac{1}{2n} \geq \varepsilon\}$$

stąd nierówność DKW:

$$P\{\sup_{x \in \mathbf{R}} |\widetilde{F}_n(x) - F(x)| \geq \varepsilon\} \leq 2e^{-2n(\varepsilon - 1/2n)^2}, \quad n > \frac{1}{2\varepsilon}$$

$$\varepsilon = 0.1, \quad \eta = 0.1 \quad \Rightarrow \quad n = 150(F_n) \quad n = 160(\widetilde{F}_n)$$

$$\varepsilon = 0.01, \quad \eta = 0.01 \quad \Rightarrow \quad n = 26,492(F_n) \quad n = 26,592(\widetilde{F}_n)$$

2. ESTYMATORY WIELOMIANOWE (na $[0,1]$)

Wielomiany podstawowe na $[0,1]$:

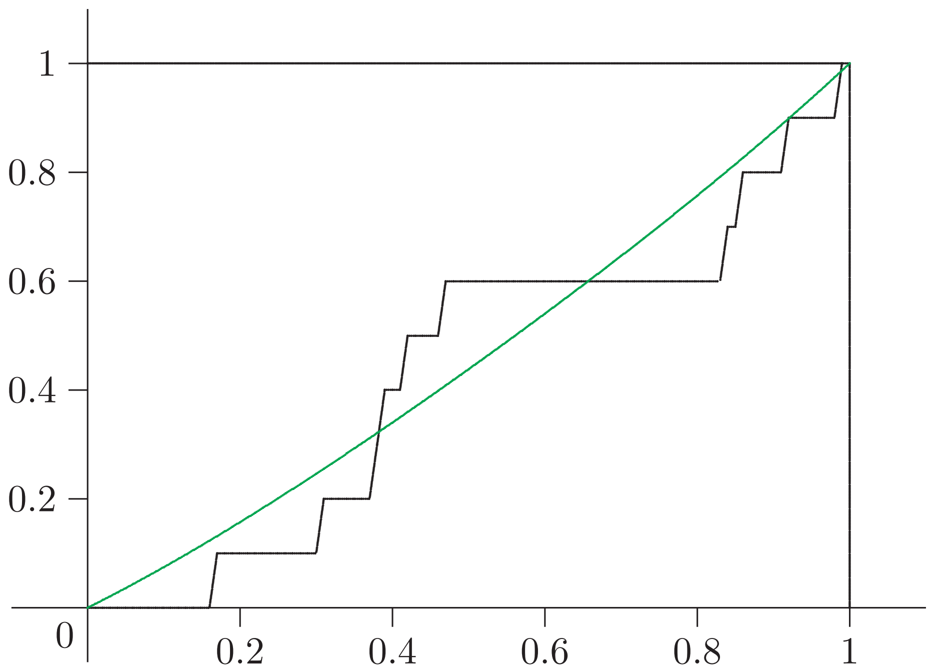
$$N_{i,m}(x) = \binom{m}{i} x^i (1-x)^{m-i}, \quad 0 \le x \le 1, \quad i = 0, 1, \ldots, m; \quad m \ge 1$$

Operator (Ciesielski 1988)

$$T_m F(x) = \sum_{i=0}^{m} \int_0^1 (m+1) N_{i,m}(y) dF(y) \int_0^x N_{i,m}(z) dz$$

T_m przekształca dystrybuanty na $[0,1]$, ciągłe lub nie, w dystrybuanty na $[0,1]$, które są wielomianami stopnia $m+1$

Estymatorem dystrybuanty F jest $F_{m,n} = T_m F_n$



TWIERDZENIE NEGATYWNE:

$$(\exists \varepsilon > 0)(\exists \eta > 0)(\forall m)(\forall n)(\exists F \in \mathcal{F})$$

$$P_F\{\sup_{x \in \mathbf{R}} |F_{m,n}(x) - F(x)| > \varepsilon\} > \eta$$

DOWÓD

Z definicji operatora T_m:

$$F_{m,n}(x) = \frac{1}{n} \sum_{j=1}^{n} \sum_{i=0}^{m} N_{i,m}(X_j) \int_0^x (m+1) N_{i,m}(z) dz$$

Oznaczamy:

$$b(i,m,q) = \binom{m}{i} q^i (1-q)^{m-i}, \quad B(i,m,q) = \sum_{j=0}^{i} b(j,m,q)$$

oraz

$$I_x(p,q) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} \int_0^x t^{p-1}(1-t)^{q-1} dt.$$

Zapisuję estymator w postaci:

$$F_{m,n}(x) = \frac{1}{n} \sum_{j=1}^{n} \sum_{i=0}^{m} b(i,m,X_j) I_x(i+1, m-i+1).$$

Udowodnię, że dla wybranych ε oraz η, dla każdego n oraz dla każdego nieparzystego m istnieje rozkład F taki, że

(#) $$P_F\Big\{F_{m,n}(\frac{1}{2}) > F(\frac{1}{2}) + \varepsilon\Big\} > \eta.$$

Wystarczą do tego bardzo grube oszacowania.

Weźmy dowolne $\varepsilon \in (0, \frac{1}{16})$ oraz $\eta \in (\varepsilon, 1)$.

Jeżeli $i \leq \frac{m-1}{2}$, to $I_{1/2}(i+1, m-i+1) \geq 1/2$, więc

$$F_{m,n}\Big(\frac{1}{2}\Big) > \frac{1}{2}\frac{1}{n}\sum_{j=1}^{n}\sum_{i=0}^{(m-1)/2} b(i, m, X_j) = \frac{1}{2}\frac{1}{n}\sum_{j=1}^{n} B(\frac{m-1}{2}, m, X_j).$$

Ponieważ $B\Big(\frac{m-1}{2}, m, q\Big)$ jest funkcją ciągłą argumentu q, malejącą, $B\Big(\frac{m-1}{2}, m, \frac{1}{2}\Big) = \frac{1}{2}$, więc dla danego $\varepsilon \in (0, \frac{1}{16})$ istnieje $\delta > 0$ takie, że

$$B(\frac{m-1}{2}, m, \frac{1}{2} + \delta) > 4\varepsilon.$$

Niech F będzie dystrybuantą taką, że

$$F(\frac{1}{2}) < \varepsilon \qquad oraz \qquad F(\frac{1}{2} + \delta) > \eta^{1/n}.$$

Ostatnia nierówność oznacza, że dla każdego $j = 1, 2, \ldots, n$,

$$P_F\{X_j < \frac{1}{2} + \delta\} > \eta^{1/n}$$

więc

$$P_F\{B(\frac{m-1}{2}, m, X_j) > 4\varepsilon\} > \eta^{1/n}.$$

Ale

$$\bigcap_{j=1}^{n}\Big\{B(\frac{m-1}{2},m,X_j)>4\varepsilon\Big\}\subset\Big\{\frac{1}{n}\sum_{j=1}^{n}B(\frac{m-1}{2},m,X_j)>4\varepsilon\Big\}$$

więc

$$P_F\Big\{\frac{1}{n}\sum_{j=1}^{n}B(\frac{m-1}{2},m,X_j)>4\varepsilon\Big\}>\eta,$$

a stąd wynika (#), q.e.d.

A. Ograniczenie klasy $\mathcal{F}$

B. Wykorzystanie operatora T_m do modyfikacji estymatora F_n

A. Ograniczenie klasy $\mathcal{F}$

===============

$\mathcal{F}_M$ - rodzina dystrybuant $F \in \mathcal{F}$ o absolutnie ciągłych gęstościach $f = F'$, spełniających warunek

$$\int_0^1 |f'(x)|^2 dx \leq M$$

Przykład: jeżeli f jest pdf $N(0, \sigma^2)$, *to* $\int_0^1 |f'(x)|^2 dx = (4\pi\sigma^4)^{-1}$

TWIERDZENIE

Dla danych $\varepsilon > 0$, $\eta > 0$ oraz M istnieją m oraz n takie, że

$$(\forall F \in \mathcal{F}_M) \quad P_F\{||F_{m,n} - F||_\infty > \varepsilon\} < \eta$$

Liczby m oraz n można wyznaczyć efektywnie z wzorów:

$$\frac{2M}{m^{1/4}} < \varepsilon \quad \text{oraz} \quad 2\exp\Big(-2\frac{nM^2}{m^{1/2}}\Big) < \eta$$

Dowód (szkic)

$$F_{m,n} - F = T_m F_n - T_m F + T_m F - F = T_m(F_n - F) + (T_m F - F)$$

$$\begin{aligned} ||F_{m,n} - F|| &\leq ||T_m|| \cdot ||F_n - F|| + ||(T_m F - F)|| \\ &\leq ||F_n - F|| + M \cdot m^{-1/4} \end{aligned} \tag{$\bullet$}$$

$$P_F\{||F_{m,n} - F|| > \varepsilon\} \leq P_F\{||F_n - F|| + \frac{M}{m^{1/4}} > \varepsilon\}$$

$$m: \quad \frac{2M}{m^{1/4}} < \varepsilon \tag{!}$$

$$\begin{aligned} P_F\{||F_{m,n} - F|| > \varepsilon\} &\leq P_F\{||F_n - F|| > \frac{M}{m^{1/4}}\} \\ &\leq 2\exp\Big(-2n\frac{M^2}{m^{1/2}}\Big) \end{aligned}$$

$$n: \quad 2\exp\Big(-2\frac{nM^2}{m^{1/2}}\Big) < \eta$$

B. Wykorzystanie wielomianów $N_{i,n}(x)$ do modyfikacji estymatora F_n

====================================

Wielomiany podstawowe na $[0,1]$:

$$N_{i,m}(x) = \binom{m}{i} x^i (1-x)^{m-i}, \quad 0 \le x \le 1, \quad i = 0, 1, \ldots, m; \quad m \ge 1$$

Definiujemy

$$\varphi_m(x) = (2m+1) N_{m,2m}(x) = (2m+1) \binom{2m}{m} x^m (1-x)^m, \quad 0 \le x \le 1, \quad m = 1, 2, \ldots$$

$$\Phi_m(x) = \int_0^x \varphi_m(y) dy.$$

$$\Phi_m(x; [a,b]) = \Phi_m\left(\frac{x-a}{b-a}\right), \quad -\infty < a < b < +\infty$$

Definiujemy

$$X_{0:n} = max\{0, X_{1:n} - (X_{2:n} - X_{1:n})\} = max\{0, 2X_{1:n} - X_{2:n}\}$$

$$X_{n+1:n} = min\{X_{n:n} + (X_{n:n} - X_{n-1:n}), 1\} = min\{2X_{n:n} - X_{n-1:n}, 1\}$$

Konstruujemy estymator wielomianowy

$$\Phi_{m,n}(x) = \begin{cases} 0, & \text{for } x < X_{0:n}, \\ \frac{1}{n}\Phi_m(x; [X_{i-1:n}, X_{i:n}]) + F_n(X_{i-1:n}) - \frac{1}{2n}, & \text{for } X_{i-1:n} \le x < X_{i:n}, \\ & i = 1, 2, \ldots, n+1 \\ 1, & \text{for } x \ge X_{n+1:n}. \end{cases}$$

Własności estymatora $\Phi_{m,n}(x)$:

1. $\Phi_{m,n}(x)$ jest dystrybuantą na $[X_{0:n}, X_{n+1:n}]$

2. $\Phi_{m,n}(X_{i:n}) = \dfrac{i}{n} - \dfrac{1}{2n} = F_n(X_{i:n}) - \dfrac{1}{2n}, \quad i = 1, 2, \ldots, n$

3. $\Phi_{m,n}(x) \in C^m(\mathbf{R})$

4. $D^k \Phi_{m,n}(X_{i:n}) = 0$ for $k = 1, 2, \ldots, m$ and $i = 1, 2, \ldots, n$

5. $\sup\limits_{x \in \mathbf{R}} |\Phi_{m,n}(x) - F(x)| \leq \sup\limits_{x \in \mathbf{R}} |F_n(x) - F(x)| + \dfrac{1}{2n}$

Nierówność DKW

$$P_F\{\sup_{x \in \mathbf{R}} |\Phi_{m,n}(x) - F(x))| \geq \varepsilon\} \leq 2e^{-2n(\varepsilon - 1/2n)^2}$$

$$n > \frac{1}{2\varepsilon}, \quad F \in \mathcal{F}$$

3. ESTYMATORY SPLAJNOWE

Definiujemy $B^{(r)}$ jako

podstawowy symetryczny B-splajn stopnia r

(podstawowa symetryczna funkcja gięta stopnia r, a symmetric cardinal B-spline of order r) z węzłami $\{i + r/2, i \in \mathcal{Z}\}$,

jeżeli

$B^{(r)}(x) \geq 0, \; x \in R$,

supp $B^{(r)} = [-r/2, r/2]$,

$B^{(r)}$ jest wielomianem stopnia $r - 1$ na każdym przedziale

$$[j - r/2, j + 1 - r/2], j = 0, 1, \ldots, r - 1,$$

$B^{(r)} \in C^{(r-2)}(R)$ (dla $r = 1$ jest to lewostronnie ciągła funkcja schodkowa)

Definicja probabilistyczna:

$B^{(r)}$ jest gęstością rozkładu prawdopodobieństwa sumy r niezależnych zmiennych losowych o jednakowym rozkładzie $U(-1/2, 1/2)$

Wzór dla $B^{(r)}(x) = P_{U(-1/2,1/2)}\{U_1 + \ldots + U_r \leq x\}$:

$$B^{(r)}(x) = \begin{cases} 0, & \text{dla } x < -r/2, \\ \dfrac{1}{(r-1)!} \displaystyle\sum_{i=0}^{[x+r/2]} (-1)^i \binom{r}{i} (x + \frac{r}{2} - i)^{r-1}, & \text{dla } x \geq -r/2 \\ 0, & \text{dla } x > r/2 \end{cases}$$

$$\mathcal{B}^{(r)}(x) = \int_{-\infty}^{x} B^{(r)}(t)dt$$

$$\mathcal{B}^{(r)}(x) = \begin{cases} 0, & \text{dla } x < -r/2, \\ \dfrac{1}{r!} \displaystyle\sum_{i=0}^{[x+r/2]} (-1)^i \binom{r}{i} (x + \frac{r}{2} - i)^{r}, & \text{dla } x \geq -r/2 \\ 1, & \text{dla } x > r/2 \end{cases}$$

Dla danych $r \geq 1$, $h > 0$, $i \in \mathcal{Z}$ oznaczamy

$$B_{h,i}^{(r)}(x) = B^{(r)}\left(\frac{x}{h} - i\right)$$

Dla danych $r \geq 1$, $1 \leq k \leq r$, $r - k = 2\nu$, ν – całkowite, $i \in \mathcal{Z}$ oraz $h > 0$ definiujemy operator (Ciesielski 1988, 1991)

$$T_h^{(k,r)}F(x) = \frac{1}{h}\sum_{i\in\mathcal{Z}}\int_R B_{h,i+\nu}^{(k)}(y)dF(y)\int_{-\infty}^x B_{h,i}^{(r)}(y)dy$$

Ten operator przeprowadza m.in. dystrybuanty (ciągłe lub skokowe) w dystrybuanty, które są splajnami stopnia r. Za estymator dystrybuanty, skonstruowany na podstawie dystrybuanty empirycznej F_n, przyjmujemy wartość tego operatora dla F_n.

Por: Wielomiany podstawowe na $[0, 1]$:

$$N_{i,m}(x) = \binom{m}{i}x^i(1-x)^{m-i}, \quad 0 \leq x \leq 1, \quad i = 0, 1, \ldots, m; \quad m \geq 1$$

Operator

$$T_mF(x) = \sum_{i=0}^m \int_0^1 (m+1)N_{i,m}(y)dF(y)\int_0^x N_{i,m}(z)dz$$

Estymator splajnowy dla wybranych $r \geq 1$, $1 \leq k \leq r$:

$$T_h^{(k,r)} F_n(x) = \sum_{i \in \mathcal{Z}} \left[\frac{1}{n} \sum_{j=1}^{n} B^{(k)} \left(\frac{X_j}{h} - \big(i + \frac{r-k}{2}\big) \right) \right] \mathcal{B}^{(r)} \Big(\frac{x}{h} - i \Big)$$

Dla przypomnienia:

$$T_h^{(k,r)} F(x) = \frac{1}{h} \sum_{i \in \mathcal{Z}} \int_R B_{h,i+\nu}^{(k)}(y) dF(y) \int_{-\infty}^{x} B_{h,i}^{(r)}(y) dy$$

Klasy dystrybuant, dla których możemy dla tego estymatora podać nierówność typu DKW, konstruujemy w następujący sposób.

Definiujemy

$$\omega_1(F,\delta) = \sup_{|t|<\delta} \sup_x |F(x+t) - F(x)|$$

oraz

$$\omega_2(F,\delta) = \sup_{|t|<\delta} \sup_x |F(x+2t) - 2F(x+t) + F(x)|.$$

Niech $\omega(h)$, $h \in R^+$, będzie modułem ciągłości, tzn. funkcją ciągłą, ograniczoną, niemalejącą, $\omega(0) = 0$. Dla danego modułu ciągłości ω, zdefiniujmy dwie Hölderowskie klasy dystrybuant:

$$H^{(k,r)}_{\omega,1} = \{F \in \mathcal{F} : \omega_1\big(F, \frac{r+k}{2}h\big) \leq \omega(h)\}$$

$$H^{(k,r)}_{\omega,2} = \{F \in \mathcal{F} : \big(2(4+(r+k)^2\big)\omega_2(F,h) \leq \omega(h)\}$$

W każdej z tych klas spełniona jest nierówność DKW, tzn. dla każdej z tych klas, dla każdego $\varepsilon > 0$ oraz $\eta > 0$ można wyznaczyć takie $h > 0$ oraz N, że jeżeli $n \geq N$, to dla każdej dystrybuanty $F \in H^{(k,r)}_{\omega,1}$, lub odpowiednio dla każdej dystrybuanty $F \in H^{(k,r)}_{\omega,2}$,

$$P_F\{||T^{(k,r)}_h F_n - F||_\infty > \varepsilon\} < \eta$$

Wystarczy wyznaczyć h oraz N takie, że

$$\omega(h) < \frac{\varepsilon}{2} \quad \text{oraz} \quad 2\exp\left(-\frac{N\varepsilon^2}{2}\right) < \eta.$$