Zbigniew Ciesielski, IMPAN Sopot
Ryszard Zieliński, IMPAN Warszawa

WYGŁADZANIE DYSTRYBUANTY EMPIRYCZNEJ

XXXVII Konferencja Zastosowań Matematyki
KZM 2008
9-16 września 2008, Zakopane-Kościelisko, Siwarna

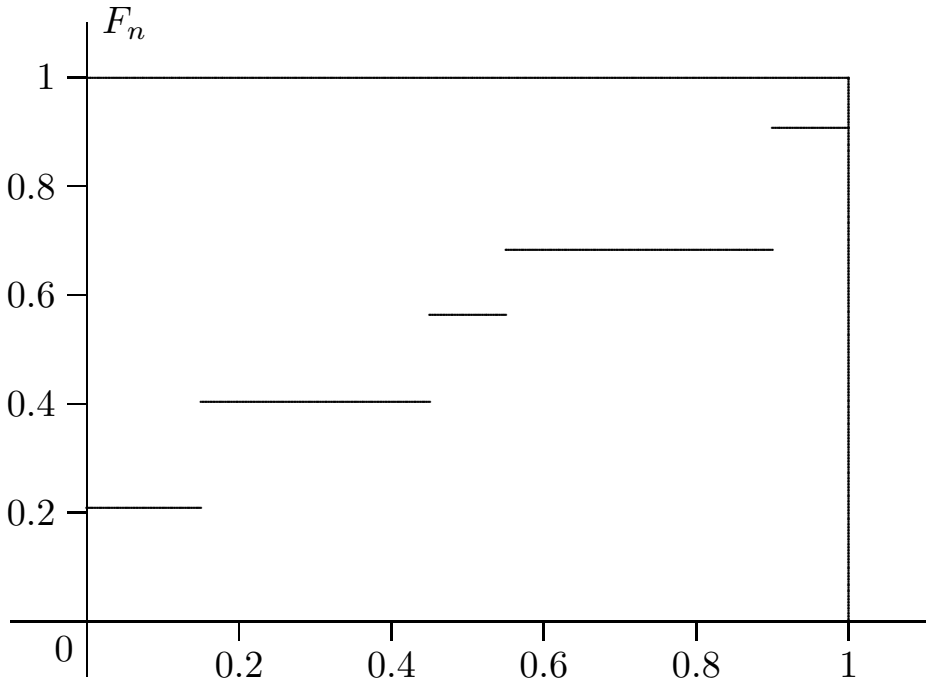
PROBLEM

$\mathcal{F}$ – RODZINA CIĄGŁYCH I ŚCIŚLE ROSNĄCYCH DYSTRYBUANT NA $[0,1]$
(W TYM KOMUNIKACIE CZASAMI TEŻ NA $\mathbf{R}^1$)

n – LICZBA NATURALNA

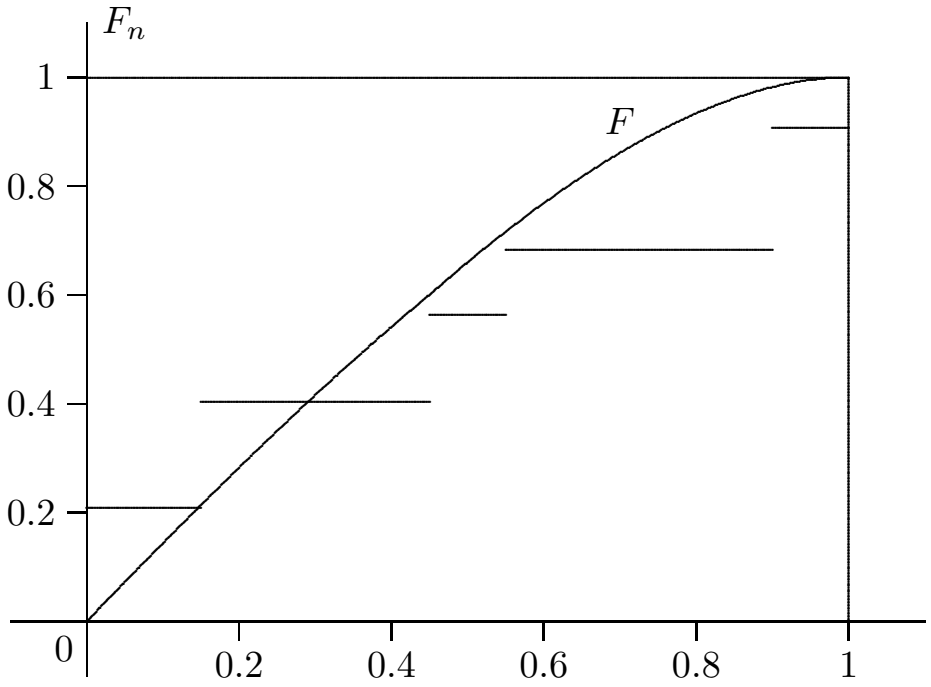
$X_1, X_2, \ldots, X_n$ – PRÓBA Z PEWNEGO (NIEZNANEGO) ROZKŁADU $F \in \mathcal{F}$
(iid rv $\sim F$)

NA PODSTAWIE $X_1, X_2, \ldots, X_n$ OSZACOWAĆ F



F_n - dystrybuanta empiryczna

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

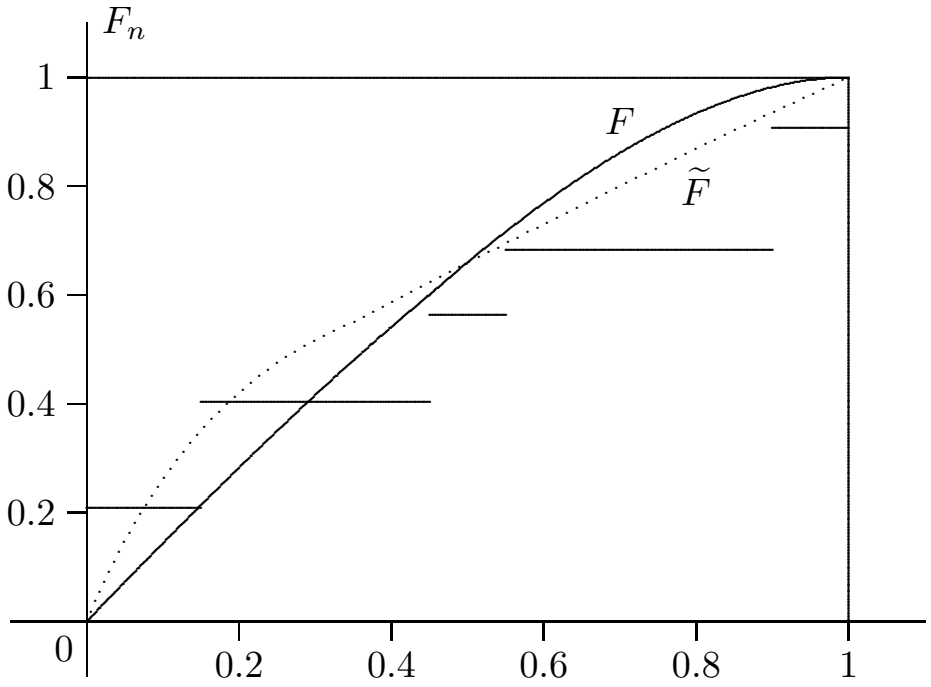


F - dystrybuanta teoretyczna

DVORETZKY-KIEFER-WOLFOWITZ

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

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



$\widetilde{F}$ — "typowa" wygładzona dystrybuanta empiryczna

TWIERDZENIE NEGATYWNE

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

$$P_F\{\sup_{x \in \mathbf{R}^1} |\widetilde{F}_n(x) - F(x)| \geq \varepsilon\} \geq \eta$$

PROBLEMY:

1. Jak wygładzać, żeby nierówność DKW zachodziła w całej klasie $\mathcal{F}$ dystrybuant ciągłych i ściśle rosnących?

2. Jeżeli zdecydujemy się na jakieś wygładzanie, to w jakiej podklasie rodziny rozkładów $\mathcal{F}$ nierówność DKW zachodzi?

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

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*

$$P_F\{\sup_{x\in\mathbf{R}^1} |\widetilde{F}_n(x) - F(x)| \geq \varepsilon\} \geq \eta.$$

KOMENTARZ:

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

1. ESTYMATORY JĄDROWE (c.d.) - twierdzenie pozytywne (Zieliński 2007)

$X_1, X_2, \ldots, X_n$ - 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 $(-1/2, 1/2)$.

Nierówność DKW:

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

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)

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

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

TWIERDZENIE NEGATYWNE:

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

$$P_F\{\sup_{x \in \mathbf{R}^1} |T_m F_n(x) - F(x)| \ge \varepsilon\} \ge \eta$$

TWIERDZENIE POZYTYWNE:

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

$$P_F\{\sup_{x \in \mathbf{R}^1} |T_m F_n(x) - F(x)| \ge \varepsilon\} \le 2e^{-2n\left(\varepsilon - m^{-1/4} M\right)^2}$$

$$\mathcal{F}_M = \{F \in \mathcal{F} : \int_0^1 |D^2 F(x)|^2 dx \le M\}$$

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

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

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

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

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

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. $\Phi_{m,n}(x) \in C^m(\mathbf{R}^1)$

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

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

Nierówność DKW

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

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 $U(-1/2, 1/2)$

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 $r \geq 1$, $1 \leq k \leq r$, $r - k = 2\nu$, $\nu -$ 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.

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:

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

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

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 $F \in H_{\omega,1}^{(k,r)}$, lub odpowiednio dla każdej dystrybuanty $F \in H_{\omega,2}^{(k,r)}$,

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

Wystarczy wyznaczyć h oraz N takie, że

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

Wyniki tu zaprezentowane pochodzą z pracy

Zbigniew Ciesielski i Ryszard Zieliński

Polynomial and spline estimators of the distribution function with prescribed accuracy

Praca wysłana do publikacji