

ZBIGNIEW CIESIELSKI, IMPAN SOPOT
RYSZARD ZIELIŃSKI, IMPAN WARSZAWA

WYGŁADZANIE DYSTRYBUANTY EMPIRYCZNEJ

XXXVII KONFERENCJA ZASTOSOWAŃ MATEMATYKI

KZM 2008

9-16 WRZEŚNIA 2008, ZAKOPANE-KOŚCIELISKO, SIWARNA

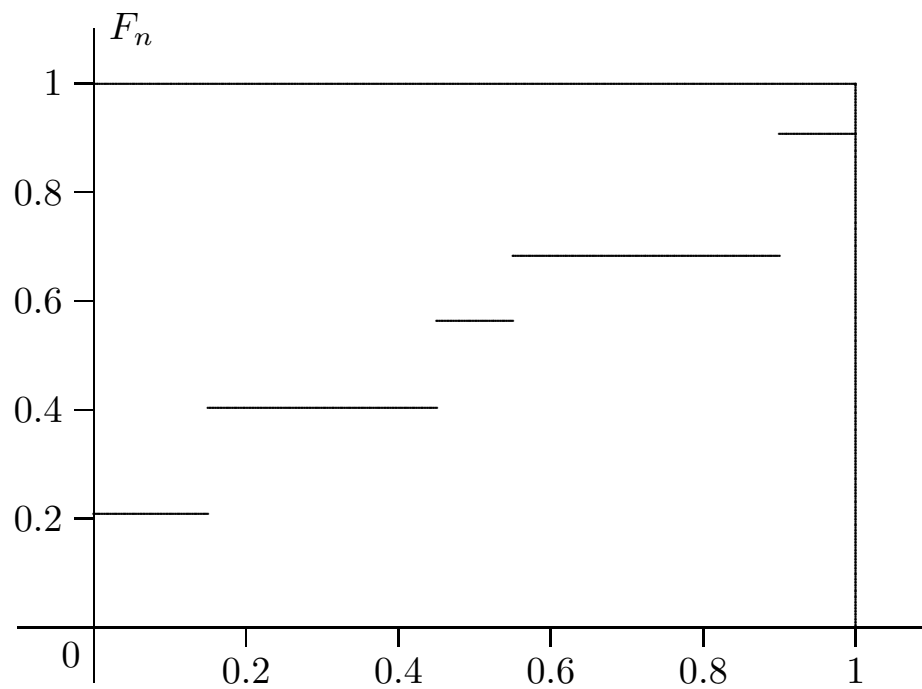
PROBLEM

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

n – LICZBA NATURALNA

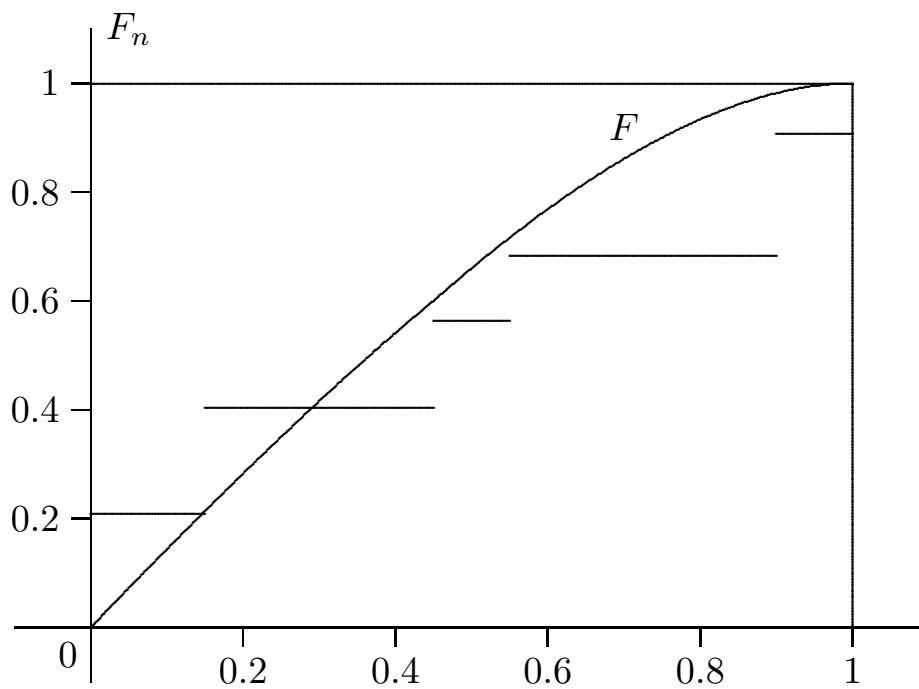
X_1, X_2, \dots, X_n – PRÓBA Z PEWNEGO (NIEZNANEGO) ROZKŁADU $F \in \mathcal{F}$
(iid rv $\sim F$)

NA PODSTAWIE X_1, X_2, \dots, X_n OSZACOWAĆ F



F_n - dystrybuanta empiryczna

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

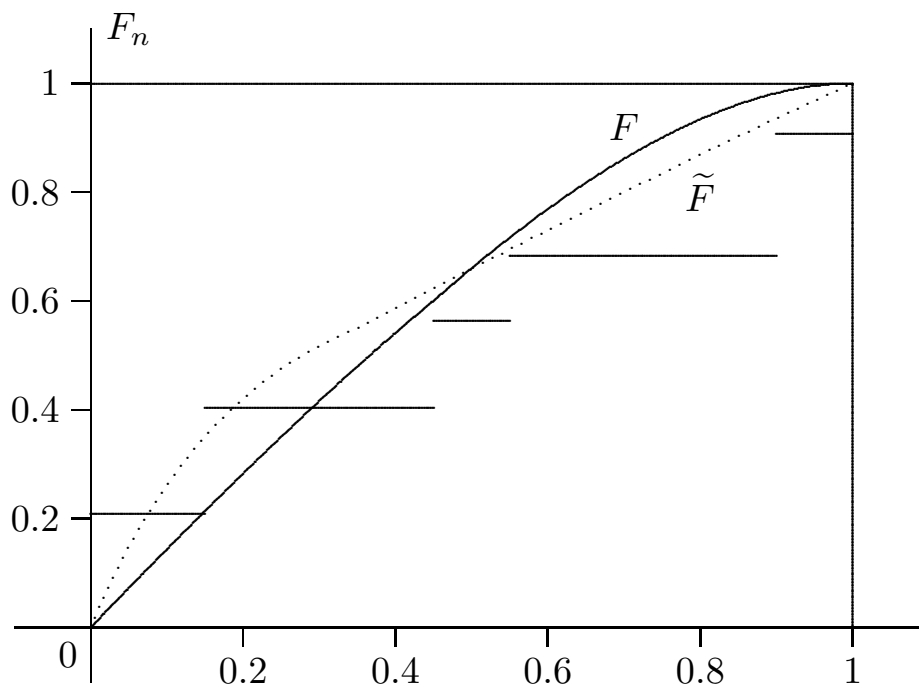


F - dystrybuanta teoretyczna

DVORETZKY-KIEFER-WOLFOWITZ

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

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



\tilde{F} — "typowa" wygładzona dystrybuanta empiryczna

TWIERDZENIE NEGATYWNE

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

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

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?

1. ESTYMATORY JĄDROWE

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

Twierdzenie negatywne (Zieliński 2007): *Jeżeli*

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

– $(h_n, n = 1, 2, \dots)$ 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

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

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, \dots, X_n - próba z rozkładu $F \in \mathcal{F}$

$X_{1:n} \leq X_{2:n} \leq \dots \leq X_{n:n}$ - statystyka pozycyjna z tej próby

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

Estymator jądrowy

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

gdzie

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

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

Nierówność DKW:

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

2. ESTYMATORY WIELOMIANOWE

Wielomiany podstawowe na $[0, 1]$:

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

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 $m+1$

TWIERDZENIE NEGATYWNE:

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

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

TWIERDZENIE POZYTYWNE:

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

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

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

2. ESTYMATORY WIELOMIANOWE (c.d.)

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

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

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

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

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

Własności estymatora $\Phi_{m,n}(x)$:

1. $\Phi_{m,n}(x)$ jest dystrybuantą na $[X_{0:n}, X_{n+1:n}]$
2. $\Phi_{m,n}(X_{i:n}) = \frac{i}{n} - \frac{1}{2n} = F_n(X_{i:n}) - \frac{1}{2n}, \quad i = 1, 2, \dots, n$
3. $\Phi_{m,n}(x) \in C^m(\mathbf{R}^1)$
4. $D^k \Phi_{m,n}(X_{i:n}) = 0$ for $k = 1, 2, \dots, m$ and $i = 1, 2, \dots, n$
5. $\sup_{x \in \mathbf{R}^1} |\Phi_{m,n}(x) - F(x)| \leq \sup_{x \in \mathbf{R}^1} |F_n(x) - F(x)| + \frac{1}{2n}$

Nierówność DKW

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

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

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

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

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

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

$B^{(r)} \in C^{(r-2)}(R)$ (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 $r \geq 1$, $1 \leq k \leq r$, $r - k = 2\nu$, 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 .

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 $\omega(h)$, $h \in \mathbb{R}^+$, będzie modulem ciągłości, tzn. funkcją ciągłą, ograniczoną, niemalejącą, $\omega(0) = 0$. Dla danego modułu ciągłości ω , zdefiniujemy dwie Hölderowskie klasy dystrybuant:

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

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

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 WYŚLANA DO PUBLIKACJI