Vlastnosti pořadí a pořádkových statistik

$X_1, \dots, X_n$ je náh. výběr se spojitou distribuční funkcí F a hustotou f.

vektor $(X_{(1)}, X_{(2)}, \dots, X_{(n)})^T$ ... vektor pořádkových statistik

$X_{(1)} \le X_{(2)} \le \dots \le X_{(n)}$ ... $X_{(i)}$ je i-tá pořádková statistika (i-té nejmenší pozorování)

Tvrzení

Hustota náh. veličiny $X_{(k)}$ je rovna $f_{(k)}(x) = n \binom{n-1}{k-1} F^{k-1}(x) \left(1 - F(x)\right)^{n-k} \cdot f(x)$, pro $k = 1, 2, \dots, n$.

Poznámka

Je-li $X_1, \dots, X_n$ náh. výběr z $Rs(0,1)$, pak $f_{(k)}(x) = n \binom{n-1}{k-1} x^{k-1} (1-x)^{n-k}$, $0 \le x \le 1$ a $EX_{(k)} = \frac{k}{n+1}$ a $DX_{(k)} = \frac{k(n-k+1)}{(n+1)^2 (n+2)}$.

R_i je pořadí n.v. X_i v náh. výběru $X_1, \dots, X_n$, jestliže $R_i = \sum_{j=1}^{n} \mathbb{1}\{X_j \le X_i\}$.

$\mathbb{R} = (R_1, \dots, R_n)^T$ je vektor pořadí

Tvrzení

a) $P(\mathbb{R} = \pi) = \frac{1}{n!}$, kde π je libovolná permutace čísel $\{1, \dots, n\}$.

b) $P(R_i = j) = \dfrac{1}{n} \quad \forall i, j = 1, \dots, n$

$$P(R_i = j) = \sum_{r_1, \dots r_n \neq j} \dots \sum P(R_1 = r_1, \dots R_i = j, \dots, R_n = r_n) = \frac{(n-1)!}{n!} = \frac{1}{n}.$$

c) $P(R_i = k, R_j = l) = \dfrac{1}{n(n-1)}, \quad \forall i \neq j, k \neq l.$

d) $ER_i = \dfrac{n+1}{2}$

$$ER_i = 1 \cdot \frac{1}{n} + 2 \cdot \frac{1}{n} + \dots + n \cdot \frac{1}{n} = \frac{1}{n}(1 + 2 + \dots + n) = \frac{1}{n} \cdot \frac{n(n+1)}{2} = \frac{n+1}{2}$$

e) $DR_i = \dfrac{n^2 - 1}{12}$

$$ER_i^2 = 1^2 \cdot \frac{1}{n} + 2^2 \cdot \frac{1}{n} + \dots + n^2 \cdot \frac{1}{n} = \frac{1}{n}(1^2 + 2^2 + \dots + n^2) = \frac{1}{n} \frac{n(n+1)(2n+1)}{6} = \frac{2n^2 + 3n + 1}{6}$$

$$DR_i = ER_i^2 - (ER_i)^2 = \frac{2n^2 + 3n + 1}{6} - \left(\frac{n+1}{2}\right)^2 = \frac{n^2 - 1}{12}.$$

f) $C(R_i, R_j) = -\dfrac{n+1}{12}, \quad i \neq j$

g) R_i je invariantní vůči posunutí, tj. vektor pořadí $(R_1, \dots, R_n)$ pro $(X_1, \dots, X_n)$ je stejný i pro $(X_1 + c, \dots, X_n + c)$, $\forall c \in \mathbb{R}$.

Poznámka

F je spojitá $\Rightarrow X_i = X_j$ s pravd. 0, tedy i $R_i = R_j$ s pravd. 0 (nemohou nastat žádné shody).

v praxi ale shoda $X_i = X_j$ může nastat

- randomizace: má-li měl skupina pozorování stejné pořadí, pak se každému pozorování přiřadí hodnota pořadí náhodně

-1,	5,	3,	3,	2,	1,	-2	X	pořadové testy pak fungují i nadále beze změny
2.	7.	5.-6.		4.	3.	1.	R	
		5.	6.	s pravd. $\frac{1}{2}$				
		6.	5	s pravd. $\frac{1}{2}$				

- metoda průměrných pořadí (midranks): každému shodnému pozorování se přiřadí průměrné pořadí celé skupiny

-1,	5,	3,	3,	2,	1,	-2	X	nutná modifikace testů
2.	7.	5,5.	5,5.	4.	3.	1.	R	

Dvouvýběrové testy (o střední hodnotě)

klasický dvouvýběrový t-test

model: $X_1, \dots, X_m$ náh. výběr z $N(\mu_1, \sigma^2)$
$Y_1, \dots, Y_n$ náh. výběr z $N(\mu_2, \sigma^2)$ } oba výběry vzájemně nezávislé

$H_0: \mu_1 = \mu_2$

$H_A: \mu_1 \neq \mu_2$ (příp. $\mu_1 > \mu_2$, $\mu_1 < \mu_2$)

rozšířený model: $X_1, \dots, X_m$ náh. výběr s distribuční funkcí F $\}$ oba výběry vzájemně nezávislé
$Y_1, \dots, Y_n$ -||- G $\}$

$H_0: F(x) = G(x), \forall x \in \mathbb{R}$

$H_1: G(x) = F(x-\Delta), \forall x \in \mathbb{R}$ pro nějaké $\Delta \neq 0$ (případně $\Delta > 0, \Delta < 0$)

Poznámky

- je-li F dist. fce normálního rozdělení => model dvouvýběrového t-testu
- distribuce F a G se liší jen posunutím.

Konstrukce pořadových testů

Předpokládejme nejprve, že známe distribuční funkci F, respektive hustotu f. Budeme hledat „optimální" pořadový test pro dané rozdělení.

uvažujme sdružený náh. výběr $(X_1, \dots, X_m, Y_1, \dots, Y_n)$ a označme jej
$(Z_1, \dots, Z_m, Z_{m+1} \dots Z_{m+n})$. Dále nechť $N = m+n$.
R_1 R_m R_{m+1} R_{m+n}

označme $R_i = \sum_{j=1}^{N} \mathbb{1}\{Z_j \leq Z_i\}$ pořadí Z_i ve sdruženém výběru $Z_1, \dots, Z_N$.

Definujme testovou statistiku $S_N^f = \sum_{i=m+1}^{N} a_N(R_i, f)$. Test určený testovou statistikou S_N^f je lokálně nejsilnější pořadový test H_0 proti H_1.

$a_N(i, f)$ je funkce pořadí $\{1, 2, \dots, n\}$ závislá na hustotě f (skóry)

$a_N(i, f) = E\varphi(U_{(i)}, f)$, kde $U_{(1)} \le U_{(2)} \le \dots \le U_{(N)}$ je uspořádaný náh. výběr z $Rs(0,1)$ o rozsahu N.

$$\varphi(u, f) = -\frac{f'(F^{-1}(u))}{f(F^{-1}(u))}, \quad 0 < u < 1.$$

$$a_N(i, f) = E\varphi(U_{(i)}, f) = -\int_0^1 \frac{f'(F^{-1}(u))}{f(F^{-1}(u))} \cdot N\binom{N-1}{i-1} u^{i-1}(1-u)^{N-i}\, du$$ ... často obtížné spočítat => použijeme tzv. přibližné skóry

$$a_N^*(i, f) = \varphi(EU_{(i)}, f) = \varphi\left(\frac{i}{N+1}, f\right) = -\frac{f'\left(F^{-1}\left(\frac{i}{N+1}\right)\right)}{f\left(F^{-1}\left(\frac{i}{N+1}\right)\right)}.$$

a definujeme lineární statistiku $S_N^{*f} = \sum_{i=m+1}^{N} a_N^*(R_i, f)$.

<u>Tvrzení</u>

S_N^f a S_N^{*f} jsou asymptoticky ekvivalentní, tj. $\dfrac{E(S_N^f - S_N^{*f})^2}{D S_N^f} \xrightarrow{N \to \infty} 0$.

Příklady

a) $f(x) = \dfrac{e^{-x}}{(1+e^{-x})^2}$, $x \in \mathbb{R}$ … hustota logistického rozdělení

$$\varphi(u,f) = -\frac{f'(F^{-1}(u))}{f(F^{-1}(u))} = 2u - 1 \quad , \; a_N^*(i,f) = \varphi\left(\frac{i}{N+1}, f\right) = \frac{2i}{N+1} - 1$$

$$S_N^{*f} = \sum_{i=m+1}^{N} a_N^*(R_i,f) = \sum_{i=m+1}^{N} \frac{2 \cdot R_i}{N+1} - 1 = \frac{2}{N+1} \sum_{i=m+1}^{N} R_i - n$$

… Wilcoxonův test (je lokálně nejsilnější pořadový test pro náh. výběry z logistického rozdělení lišící se posunutím).

testová stat.

„klasického“ Wilcoxonova testu

b) $f(x) = \dfrac{1}{\sqrt{2\pi}} \cdot e^{-\frac{x^2}{2}}$, $x \in \mathbb{R}$ … hustota normálního rozdělení

$$\varphi(u,f) = \Phi^{-1}(u) \quad , \; a_N^*(i,f) = \Phi^{-1}\left(\frac{i}{N+1}\right)$$

, kde Φ^{-1} je kvantilová funkce $N(0,1)$.

$$S_N^{*f} = \sum_{i=m+1}^{N} \Phi^{-1}\left(\frac{R_i}{N+1}\right)$$

… van der Waerdenův test

c) $f(x) = \frac{1}{2} \cdot e^{-|x|}$, $x \in \mathbb{R}$ … hustota Laplaceova (dvojitě exponenciálního) rozdělení

$$\varphi(u,f) = \operatorname{sign}\left(u - \frac{1}{2}\right) \quad , \; S_N^{*f} = \sum_{i=m+1}^{N} \operatorname{sign}\left(\frac{R_i}{N+1} - \frac{1}{2}\right)$$

… mediánový test

V praxi ale hustotu f neznáme. Proto se snažíme „tipnout", jak by mohla vypadat.

Zvolíme tedy funkci $\varphi: [0,1] \mapsto \mathbb{R}$ neklesající, nekonstantní a integrovatelnou se čtvercem, tj. $\int_0^1 \varphi^2(u)\,du < \infty$.

a definujeme skóry $a_N(i) = \varphi\left(\frac{i}{N+1}\right)$ a testovou statistiku $S_N = \sum_{i=m+1}^{N} a_N(R_i) = \sum_{i=m+1}^{N} \varphi\left(\frac{R_i}{N+1}\right)$.

Potřebujeme rozdělení testové statistiky za platnosti H_0, abychom určili kritický obor.

<u>Přesné rozdělení:</u>

vektor pořadí $(R_1,\dots,R_N)^T$ má diskrétní rovnoměrné rozdělení na množině $\{1,\dots,N\}$ $\Rightarrow$ věta o transformaci $\Rightarrow$ rozdělení statistiky S_N

spočítáme hodnotu testové statistiky pro každou z $N!$ permutací čísel $\{1,\dots,N\}$ $\rightarrow$ pravděpodobnostní funkce S_N (výpočetně náročné).

$$\begin{matrix} Z_1,\dots,Z_m,Z_{m+1},\dots,Z_N \\ R_1,\dots,R_m,R_{m+1},\dots,R_N \end{matrix} \rightarrow S_N = \underbrace{\sum_{i=m+1}^{N} a_N(R_i)}$$

Závisí jen na m pořadí druhého výběru $\Rightarrow$ stačí uvažovat jen m pořadí z čísel $\{1,\dots,N\}$, tj. celkem $\binom{N}{m}$ možností (každá stejně pravděpodobná s pravd. $\frac{1}{\binom{N}{m}}$)

spočítáme hodnotu testové statistiky pro každou z $\binom{N}{m}$ možností, jak z množiny $\{1,\dots,N\}$ vybrat m čísel. $\Rightarrow$ pravd. fce S_N (za platnosti H_0) $\Rightarrow$ kvantily (kvantilová fce) Q_α.

- $H_1: \Delta \neq 0$, $S_N < Q_{\frac{\alpha}{2}}$ nebo $S_N > Q_{1-\frac{\alpha}{2}}$ $\Rightarrow$ zamítáme H_0
- $H_1: \Delta > 0$, $S_N > Q_{1-\alpha}$ $\Rightarrow$ zamítáme H_0
- $H_1: \Delta < 0$, $S_N < Q_{\alpha}$ $\Rightarrow$ zamítáme H_0