Testy nezávislosti

klasicky: $(X_1,Y_1)^T, (X_2,Y_2)^T, \dots, (X_n,Y_n)^T$ je náhodný výběr z dvourozměrného normálního rozdělení.

označme ρ korelační koeficient mezi X_i a Y_i

$H_0: \rho = 0$ (X_i a Y_i jsou nezávislé)

$H_1: \rho \neq 0$ (příp. $\rho > 0, \rho < 0$)

Pearsonův (výběrový) korelační koeficient
$$r = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2 \cdot \frac{1}{n-1}\sum_{i=1}^{n}(Y_i-\bar{Y})^2}}$$

$$T = \sqrt{n-2}\cdot\frac{r}{\sqrt{1-r^2}} \overset{H_0}{\sim} t(n-2).$$

zobecněný model: $(X_1,Y_1)^T, \dots, (X_n,Y_n)^T$ je náhodný výběr z dvourozměrného spojitého rozdělení s distribuční funkcí $F(x,y)$

označme $F_1(x)$ marginální distribuční funkci a $f_1(x)$ marginální hustotu náh. veličiny X_i,

$F_2(y)$ -||- $f_2(y)$ -||- Y_i.

$H_0: F(x,y) = F_1(x)\cdot F_2(y)$ (X_i a Y_i jsou nezávislé)

$H_1: F(x,y) \neq F_1(x)\cdot F_2(y)$

Konstrukce pořadových testů

Nejprve uvažujme, že známe distr. fce $F_1(x)$ a $F_2(y)$, respektive hustoty $f_1(x)$ a $f_2(y)$.

Označme R_i pořadí X_i mezi $X_1, \dots, X_n$

a S_i pořadí Y_i mezi $Y_1, \dots, Y_n$.

Lokálně nejsilnější pořadový test H_0 proti H_1 je určen testovou statistikou $S_n^{f_1,f_2} = \sum_{i=1}^{n} a_n(R_i, f_1) \cdot a_n(S_i, f_2)$.

Skóry $a_n(i,f)$ můžeme opět definovat přímo jako $E\varphi(U_{(i)}, f)$, nebo použít aproximativní skóry $\varphi(EU_{(i)}, f) = \varphi\left(\frac{i}{n+1}, f\right)$.

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

V praxi ale marginální hustoty f_1 a f_2 neznáme. Zvolíme proto funkce $\varphi, \psi : [0,1] \to \mathbb{R}$ neklesající, nekonstantní a integrovatelné se čtvercem a definujme

$$a_n^{(1)}(i) = \varphi\left(\frac{i}{n+1}\right)$$

$$S_n = \sum_{i=1}^{n} a_n^{(1)}(R_i) \cdot a_n^{(2)}(S_i) = \sum_{i=1}^{n} \varphi\left(\frac{R_i}{n+1}\right) \cdot \psi\left(\frac{S_i}{n+1}\right).$$

$$a_n^{(2)}(i) = \psi\left(\frac{i}{n+1}\right)$$

Jak určíme rozdělení testové statistiky S_n za platnosti H_0?

permutační princip:

- položíme $R_1 = 1, R_2 = 2, \dots, R_n = n$
- pro každou permutaci pořadí $S_1, S_2, \dots, S_n$, tj. čísel $\{1, 2, \dots, n\}$ spočítáme hodnotu S_n (testové statistiky) $\longrightarrow$ pravd. funkce S_n za platnosti H_0

Výpočetně náročné i pro n středmí $\rightarrow$ asymptotická aproximace.

__Poznámka__

Za platnosti H_0 platí:

$$ES_n = \frac{1}{n}\left(\sum_{i=1}^{n} a_n^{(1)}(i)\right)\cdot\left(\sum_{j=1}^{n} a_n^{(2)}(j)\right), \quad DS_n = \frac{1}{n-1}\left(\sum_{i=1}^{n}\left(a_n^{(1)}(i) - \overline{a^{(1)}}\right)^2\right)\cdot\left(\sum_{j=1}^{n}\left(a_n^{(2)}(j) - \overline{a^{(2)}}\right)^2\right).$$

__Věta__

Za platnosti H_0 má testová statistika S_n asymptoticky normální rozdělení $N(ES_n, DS_n)$ při $n \to \infty$.

__Poznámka__

Rozptyl DS_n můžeme aproximovat pomocí $n\left(\int_0^1 (\varphi(u) - \overline{\varphi})^2\, du\right)\left(\int_0^1 (\psi(u) - \overline{\psi})^2\, du\right)$.

Příklady

a) $f_1(x) = f_2(x) = \dfrac{e^{-x}}{(1+e^{-x})^2},\ x \in \mathbb{R}$ hustota logistického rozdělení

$$\varphi(u,f) = 2u - 1$$

$$a_n^{(1)}(i) = a_n^{(2)}(i) = \frac{2i}{n+1} - 1$$

$$S_n^{lok} = \sum_{i=1}^{n} \left(\frac{2R_i}{n+1} - 1\right)\cdot\left(\frac{2S_i}{n+1} - 1\right) = \sum_{i=1}^{n} \left(\frac{4R_i\cdot S_i}{(n+1)^2} - \frac{2R_i}{n+1} - \frac{2S_i}{n+1} + 1\right) = \frac{4}{(n+1)^2}\sum_{i=1}^{n} R_i\cdot S_i - \frac{\cancel{2}}{\cancel{n+1}}\cdot\frac{n\cancel{(n+1)}}{\cancel{2}} - \frac{\cancel{2}}{\cancel{n+1}}\cdot\frac{n\cancel{(n+1)}}{\cancel{2}} + n =$$

$$= \frac{4}{(n+1)^2}\sum_{i=1}^{n} R_i\cdot S_i - n$$ ... souvislost se Spearmanovým korelačním koeficientem

$$r_S = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(R_i-\bar{R})\cdot(S_i-\bar{S})}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(R_i-\bar{R})^2\cdot\frac{1}{n-1}\sum_{i=1}^{n}(S_i-\bar{S})^2}}$$ ... = Pearsonův korelační koeficient z pořadí $(R_1,S_1)^T, (R_2,S_2)^T, \dots, (R_n,S_n)^T$

Na rozdíl od Pearsonova korelačního koeficientu neměří pouze lineární závislost, ale i závislost popsanou pomocí monotónních transformací.

Poznámka

$$\sum_{i=1}^{n} R_i = \sum_{i=1}^{n} S_i = 1 + \dots + n = \frac{n(n+1)}{2} \quad\Rightarrow\quad \bar{R} = \bar{S} = \frac{n+1}{2}$$

$$\sum_{i=1}^{n} R_i^2 = \sum_{i=1}^{n} S_i^2 = 1^2 + \dots + n^2 = \frac{n(n+1)(2n+1)}{6} \quad\Rightarrow\quad \sum_{i=1}^{n}(R_i-\bar{R})^2 = \sum_{i=1}^{n} R_i^2 - n\cdot(\bar{R})^2 = \frac{n(n+1)(2n+1)}{6} - n\cdot\left(\frac{n+1}{2}\right)^2 = \sum_{i=1}^{n}(S_i-\bar{S})^2.$$

$$r_S = \frac{12}{n(n^2-1)} \underline{\sum_{i=1}^{n} R_i \cdot S_i} - 3\frac{n+1}{n-1} \qquad -1 \le r_S \le 1$$

$$r_S = 1 - \frac{6}{n^3-n} \sum_{i=1}^{n} (R_i - S_i)^2 .$$

Za platnosti Ho platí: $E r_S = 0$, $D r_S = \frac{1}{n-1}$ a $r_S \approx N\left(0, \frac{1}{n-1}\right)$.

b) $f_1(x) = f_2(x) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{x^2}{2}}$, $x \in R$ hustota normálního rozdělení

$$\varphi(u, f) = \Phi^{-1}(u)$$

$$a_n^{(1)}(i) = a_n^{(2)}(i) = \Phi^{-1}\left(\frac{i}{n+1}\right)$$

$$S_n = \sum_{i=1}^{n} \Phi^{-1}\left(\frac{R_i}{n+1}\right) \cdot \Phi^{-1}\left(\frac{S_i}{n+1}\right)$$ ... van der Waerdenův test

za platnosti Ho platí při $n \to \infty$ $S_n \approx N\left(0, \frac{1}{n-1}\left\{\sum_{i=1}^{n}\left[\Phi^{-1}\left(\frac{i}{n+1}\right)\right]^2\right\}^2\right)$.

c) Kendallovo τ

$$\tau = \frac{\sum\sum_{i \ne j} \operatorname{sign}(R_i - R_j) \cdot \operatorname{sign}(S_i - S_j)}{n(n-1)} \qquad -1 \le \tau \le 1$$

- je to pořadová statistika, ale není lineární.

$$\tau = \frac{a-b}{\binom{n}{2}}$$

, kde a je počet konkordantních párů (x_i, y_i), tj. $x_i > x_j \wedge y_i > y_j$ nebo $x_i < x_j \wedge y_i < y_j$.

b je počet diskordantních párů (x_i, y_i), tj. $x_i > x_j \wedge y_i < y_j$ nebo $x_i < x_j \wedge y_i > y_j$.

Poznámka

Projekce Kendallova τ do prostoru lineárních pořadových statistik je až na násobek Spearmanův korelační koeficient.

Za platnosti H_0 platí: $E\tau = 0$, $D\tau = \frac{2(2n+5)}{9n(n-1)}$ a $\tau \approx N(0, D\tau)$ pro $n \to \infty$.