

Neparametrické a robustní metody

neparametrické metody - nepokládají lineární řádní předpoklady o rozdělení

robustní metody - jsou málo citlivé na odchylky od předpokládaného rozdělení

Mezi neparametrické metody patří např. Spearmanův test nerovnosti, který je založen nikoli na konkrétních hodnotách výběru, ale na jejich pořadí (patří mezi tzv. porádkové testy). Nezpracovává řádní předpoklady o rozdělení dat.

Dále sem patří testy založené na empirických distribučních funkcích

Je často nemusím testovat hypotézu o shodnosti středních hodnot a.

rozptylu po dvou výběrů z normálního rozdělení, ale mohou testovat

hypotézu, zda tyto výběry mají stejnou distribuční funkci

(a tím pádem budou mít i stejný střední kvantil).

Empirická distribučná fca. sa definuje predpisom

$$F_m(x) = \frac{1}{n} \sum_{i=1}^m u(x - X_i),$$

kde $u(x) = 1$ pre $x > 0$, $u(x) = 0$ pre $x \leq 0$.

Je teda $F_m(x) = \frac{1}{n}$ (počet X_i menších než x) $\doteq P(X < x) = F(x)$.

Test je založený na veličine (test Kolmogorov - Smirnovova štatistika.)

$$D = \max_{-\infty < x < \infty} |F_m(x) - G_m(x)|,$$

ktvá by mala byť neplatosť hypotézy mala - existuje pre ni
tabelované kritické hodnoty.

Neparametrickými testy je i jednovýbovú testy dobre shody o tom, či
výber pochádza z rozdelenia s distr. fca' F . Pak

$$D = \sup_{-\infty < x < \infty} |F_m(x) - F(x)|.$$

Jiným typom testu dobre shody je χ^2 -testy dobre shody založené
na čínoch - skráň je test shody s normálnym rozdelením.

Robustní metody

Nejčastěji se řeší situace, kdy výber nemá přesně normální rozdělení, např. pak, že jsou v něm odlehilá pozorování. Da' se to popsat sčítá, např. že skuteční rozdělení má hustotu f , kde

$$F = (1 - \varepsilon)\Phi + \varepsilon H$$

(Φ je hustota normálního rozdělení, H nějaká jiná, $\varepsilon > 0$ malé).

Tomu se říká ε -kontaminovaný normální model, ε značí malý odchylky od ideálního modelu a tím implikuje, že tam jsou měřené má odlehilé hodnoty libovolné velikosti, ale bude jich málo.

Jiný model je Holmegerov - Gmninorova skoli' normálního rozdělení

$$\sup_x |\Phi(x) - F(x)| \leq \delta,$$

kde $0 \leq \delta \leq 1$, klidě říká, že měřené narušit model u velkého počtu pozorování, ale málo.

Evidují různé typy robustnosti a různé typy robustních odhadů.

Robustní odhady parametrů polohy (např. střední hodnoty):

Arithmetický průměr $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ počítaný z celého výběru je dost citlivý na extrémně vylučující hodnoty - šlápl' jedna labuť.

Robustní L-odhad je založen na pořádkových statistikách

$X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$ a má obecně tvar

$$L_n = \sum_{i=1}^n c_i X_{(i)},$$

kde e_i jsou známé konstanty, tzv. váhy, pro které $\sum_{i=1}^m e_i = 1$.

Např. pro $e_1 = \dots = e_m = \frac{1}{m}$ dostáváme obvyklý aritmetický průměr,

$e_1 = e_m = 0$ znamená, že vyměníme a nejmenší a největší hodnota se
škrtně odd.

d -exkluzivní průměr
$$L_{m,d} = \frac{1}{m - 2\lfloor md \rfloor} \sum_{i=\lfloor md \rfloor + 1}^{m - \lfloor md \rfloor} x_{(i)},$$

kde $0 \leq d < \frac{1}{2}$ je stupeň exkluzivnosti.

Pro $d = 0$ dostáváme obyčejný průměr, pro $d = \frac{1}{2}$ obyčejný medián.

R-odhad obvykle využívají výsledky z korelačních testů,

z publikované literatury, zejména pátá kapitola, protože jsou vypracovány
mnohokrát a čas vypracování má jistě výhodu, protože dostáváme spolehlivý
výsledek výše kvalitou výsledků.

Příkladem R-odhadu je Hodgesova-Kermannova odhad

$$R = \text{med}_{1 \leq i, j \leq n} \left\{ \frac{x_i + x_j}{2} \right\}.$$

Existují ještě M-odhady, které jsou robustním maximálně invariantním
odhady, P-odhady, W-odhady, D-odhady, A-odhady, d-odhady,
Bayesovské robustní odhady, S-odhady, SP-odhady a další.

Odhady parametru měřítka (obdobu rozptylu)

L -odhady mají obecně tvar $\sum_{i=1}^m c_i X_{(i)}$, kde měly c_i se součítat, aby $\sum c_i = 0$.

Mají pro $c_1 = -1, c_2 = \dots = c_{m-1} = 0, c_m = 1$ dostáváme tzv. rozpětí $X_{(n)} - X_{(1)}$.

Typické jsou merikvantilové odhady, mají $\tilde{X}_{0,95} - \tilde{X}_{0,25}$,

kde 25% pozorování je menších než $\tilde{X}_{0,25}$, 45% je menších než $\tilde{X}_{0,75}$ (a tedy 25% je větší).

(V tomto uspořádání střední median je $\tilde{X}_{0,5}$).

Robustní odhady o lineárním modelu

Máme model $Y = X\beta + \varepsilon$, kde chyby jsou nezávislé a stejně rozdělené podle nějaké distribuční fce F .

Regressní median je odhad minimalizující výraz

$$\sum_{i=1}^m |Y_i - \sum_{j=1}^p x_{ij}\beta_j| = \sum_{i=1}^m |x_i|$$

(měly se nazývat LAD odhad).

Robustní (neparametrické) testy hypotéz

Jednovýběrový Wilcoxonův test:

Předp.: $X_1, \dots, X_n$ je náhodný výběr ze spojitého rozdělení
s hustotou f , která je symetrická kolem bodu a , tj.

$$f(a+x) = f(a-x).$$

pak medián $\tilde{X} = a$ (a také $EX = a$, pokud existuje)

hypotéza: $H_0: \tilde{X} = x_0$, alternativa $H_1: \tilde{X} \neq x_0$

položíme $Y_i = X_i - x_0$

přesuneme Y_i do neklesající posloupnosti podle absolutní hodnoty

$$|Y|_{(1)} \leq |Y|_{(2)} \leq \dots \leq |Y|_{(n)}$$

R_i = pořadí relací $|Y|_{(i)}$

$$S^+ = \sum_{Y_i \geq 0} R_i, \quad S^- = \sum_{Y_i < 0} R_i$$

H_0 se zamítá, když $\min(S^+, S^-) \leq w_n(\alpha)$ (tj. když S^+ a S^- se
hodně liší)

$w_n(\alpha)$ je kritická hodnota W. testu, je tabulkovaná. %
Příklad:

Dvouvýběrový W. test:

$H_0: F = G$ proti $H_1: F \neq G$, speciálně $H_1': G(x) = F(x - \Delta)$, $\Delta \neq 0$
(alternativa posunutí)

kde F je spojita distrib. funkce náh. $X_1, \dots, X_m$

G

-1-

$Y_1, \dots, Y_m$

náh. pro meravíci

Příklad (bez motivace):

hodnoty X_i : 53, 48, 45, 55, 63, 51, 66, 56, 50, 58

hypotéza: $\tilde{X} = 60$

$Y_i = X_i - 60$: -7, -12, -15, -5, 3, -9, 6, -4, -10, -2

(je vidět, že převažují záporné hodnoty, takže hypotéza se nejspíš bude zamítnout)

pořadí podle absolutních hodnot:

-2, 3, -4, -5, 6, -7, -9, -10, -12, -15

$$S^+ = \sum_{Y_i \geq 0} R_i = 2 + 5 = 7$$

$$S^- = \sum_{Y_i < 0} R_i = 1 + 3 + 4 + 6 + 7 + 8 + 9 + 10 = 48$$

$$\min(S^+, S^-) = 7$$

krit. hodnota $w_{10}(0,05) = 8$

H_0 se zamítá

dvouvýběrový W. test:

postup: všech $m+n$ veličin se ^{uspořádaný} uspořádá podle velikosti (sdružený výběr)

vypočte se $\bar{T}_1 =$ součet pořadí hodnot $X_1, \dots, X_m$
 $\bar{T}_2$ - " - $Y_1, \dots, Y_n$

$$\text{platí } \bar{T}_1 + \bar{T}_2 = \frac{1}{2} (m+n)(m+n+1)$$

$$\text{za platnosti hypotézy: } E\bar{T}_1 = \frac{1}{2} m(m+n+1), \quad E\bar{T}_2 = \frac{1}{2} n(m+n+1)$$

$$\text{def: } U_1 = mn + \frac{1}{2} m(m+1) - \bar{T}_1, \quad U_2 = mn + \frac{1}{2} n(n+1) - \bar{T}_2$$

$$EU_1 = \frac{1}{2} mn = EU_2, \quad U_1 + U_2 = mn$$

H_0 se zamítá, když U_1 a U_2 se hodně liší, tj. když $\min(U_1, U_2) \leq \text{krit. hod.}$
(tabelovaná)

Znaménkový test: (jednovýběrový)

Předp, že $X_1, \dots, X_n$ je výběr ze spojitého rozdělení s mediánem $\tilde{x}$

$$H_0: \tilde{x} = x_0, \quad A: \tilde{x} \neq x_0$$

opět se utvoří rozdíly $Y_i = X_i - x_0$, nulové hodnoty Y_i se vynášejí

označíme Y = počet Y_i s kladným znaménkem

$$\text{za platnosti } H_0: P(Y_i > 0) = P(Y_i < 0) = \frac{1}{2}$$

tedy Y má binomické rozd. s parametry $n, \frac{1}{2}$

H_0 se zamítá, když Y je hodně malé nebo naopak hodně velké

používají se kritické hodnoty $k_1, k_2 = n - k_1$ odvozené z binomického

rozdělení tak, že

$$P(Y \leq k_1) \leq \frac{\alpha}{2}, \quad P(Y \geq k_2) \leq \frac{\alpha}{2}$$

V našem příkladě: $Y = 2$

z tabulek pro $\alpha = 0,05$ máme $k_1 = 1, k_2 = 9$

H_0 tedy nezamítáme

Ríká se, že tento test má poměrně malou sílu (pst chyby 2. druhu

je dost velká), proto je třeba, aby počet pozorování n

byl dost velký (jak?)

Používá se hlavně v případech, kdy rozdělení X_i je výrazně nesymetrické.

Poznámka: Pokud se mezi hodnotami Y_i vyskytne hodně nul (více než $1/4$), tak se tato pozorování vynášejí.