Matematická statistika A

2014

Outline

1. Úvod, struktura přednášky

Outline

Outline

Outline

Outline

Outline

Hlavní kapitoly:

I. Opakovaní zakladů pravděpodobnosti a mat.statistiky

II. Testy založené na pořadích

III. Mnohorozměrná statistika

IV. Neparametrická regrese

V. Resampling metody

VI. Základy Bayesovského přístupu k řešení statistických problémů (bude-li čas, popř. jiné téma)

Outline

I. Opakování základních pojmů

Literatura pro úvod

K. Zvára, J.Štěpán: Pravděpodobnost a matematická statistika, Matfyzpress 1997

J. Anděl: Statistické metody, Matfyzpress 1993

R. Bartoszynski, M. Nuewiadomska-Bugaj, Wiley 1996

Pro další části bude literatura postupně doplňována

pravděpodobnost P -míra na $(\Omega, \mathcal{A})$, A –jevy,
$0 \leq P(A) \leq 1, P(\Omega) = 1,\ P(A \cup B) = P(A) + P(B) - P(A \cap B)$

nahodná veličina $X : \Omega \to R^1$

distribuční funkce náhodné veličiny

střední hodnota

rozptyl

binomické rozdělení, normalnií rozdělení, t-rozdělení,

centralni limitní věta a její použití

základy teorie odhadu a testování hypotéz, některé zakladní odhady a testy

Outline

II. Pořadové testy

II. 1. Základní pojmy

pořadí, základní vlastnosti, lineární pořadové statistiky,

II.2 Pořadová statistika pro test hypotézy symetrie

II.3 Pořadové statistiky pro dvouvýběrový problém

obecný test, jednovýběrový Wilcoxonův test a jeho vlastnosti, znaménkový test

obecný test, Wilcoxonův test a jeho vlastnosti, mediánový test, kvartilový test

II.4 Některé další používané pořadové testy

Kruskal -Wallisův test, Friedmanův test, Spearmanův korelační koeficient,
Kolmogorov-Smirnovovy statistiky

II.1 Základní pojmy

Nechť $X_1, \ldots, X_n$ jsou nezávislé náhodné veličiny se spojitou distribuční funkcí F,

R_i– pořadí X_i mezi $X_1, \ldots, X_n$, $\quad i = 1, \ldots, n$
Platí $P(X_i = X_j) = 0, \quad i \neq j$

Příklad

X_i	53	48	45	55	63	51	66	56	50	58
R_i	05	02	01	06	09	04	10	07	03	08

Pořadové testy jsou založeny na $R_1, \ldots, R_n$, jednoduché, mají řadu výhod, ale ztrácíme část informace

Věta 1 Nechť $X_1, \ldots, X_n$ jsou nezávislé náhodné veličiny se spojitou distribuční funkcí F. Pak platí

$$P(R_1 = r_1, \ldots, R_n = r_n) = \frac{1}{n!}, \quad (r_1, \ldots, r_n) \quad permutace \quad 1, \ldots, n,$$

$$P(R_i = j) = \frac{1}{n}, \quad i, j = 1, \ldots, n$$

$$P(R_1 = r_1, R_2 = r_2) = \frac{1}{n(n-1)}, \quad i \neq j \quad 1 \leq r_i, r_j \leq n \quad r_i \neq r_j$$

$$E\, R_i = (n+1)/2, \quad i = 1, \ldots, n$$

$$var\, R_i = \frac{1}{n-1} \sum_{j=1}^{n} j^2 - \frac{1}{n-1}\left(\frac{n+1}{2}\right)^2$$

Důkaz Počet permutací čísel $1, \ldots, n$ je $n!$.

$$P(R_i = j) = \frac{(n-1)!}{n!}$$

$$E\, R_i = \sum_{j=1}^{n} j P(R_i = j) = \frac{1}{n} \frac{n(n+1)}{2} = \frac{n+1}{2}$$

Matematická statistika A

Outline

Úvod, struktura přednášky

I. Opakování základních pojmů

II. Pořadové testy

II.2. Pořadová statistiky pro test hypotézy symetrie

II.3. Dvouvýběrový problém

Rozdělení $R_1, \ldots, R_n$ nezávisí na F!!!

Definujeme *lineární pořadovou statistiku* předpisem

$$S = \sum_{i=1}^{n} c_i a(R_i),$$

kde $c_1, \dots, c_n$ a $a(1), \dots, a(n)$ jsou daná čísla, např. $a(i) = i,\ i^2, \dots$.

Věta 2 Za předpokladu předchozí věty platí

$$E\,S = \frac{1}{n} \sum_{i=1}^{n} c_i \sum_{j=1}^{n} a(j)$$

$$var\ S = \frac{1}{n-1} \sum_{i=1}^{n} (c_i - \overline{c}_n)^2 \sum_{j=1}^{n} (a(j) - \overline{a}_n)^2$$

kde

$$\overline{c}_n = \frac{1}{n} \sum_{i=1}^{n} c_i, \quad \overline{a}_n = \frac{1}{n} \sum_{j=1}^{n} a(j).$$

Důkaz Platí:

$$E\, a^v(R_i) = \frac{1}{n}\sum_{j=1}^{n} a^v(j), \quad v = 1, 2, \dots,$$

$$E\, a(R_1)a(R_2) = \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j=1, j\neq i}^{n} a(j)a(i)$$

□

Matematická statistika A

Outline

Úvod, struktura přednášky

I. Opakování základních pojmů

II. Pořadové testy

II.2. Pořadová statistiky pro test hypotézy symetrie

II.3. Dvouvýběrový problém

Jiná pořadová statistika - pořadová statistika pro hypotézu symetrie:

$$S^+ = \sum_{i=1}^{n} a(R_i^+) sign\, X_i$$

kde R_i^+ je pořadí $|X_i|$ mezi $|X_1|, \dots, |X_n|$ a $sign\, x = 1,\ x > 0; = -1,\ x < 0; = 0\, x = 0$.

Věta 3 Nechť jsou splněny předpoklady věty 1 a nechť $F(x) + F(-x) = 1,\ x \in R^1$. Pak jsou $|X_i|$ a $sign\, X_i$ nezávislé náhodné veličiny.nezávislé Navíc, náhodné vektory $(R_1^+, \dots, R_n^+)$ a $(sign\, X_1, \dots, sign\, X_n)$ jsou nezávislé.

Důkaz Platí pro $x > 0$:

$$P(|X_1| < x, sign\, X_1 = 1) = P(0 < X_1 < x) = F(x) - F(0),$$
$$P(|X_1| < x) = P(-x < X_1 < x) = F(x) - F(-x) = 2F(x) - 1,$$
$$P(sign\, X_1 = 1) = P(X_1 > 0) = 1/2, \quad F(0) = 1/2$$

Tedy

$$P(|X_1| < x, sign\, X_1 = 1) = P(|X_1| < x) P(sign\, X_1 = 1)$$

Podobně pro $P(|X_1| < x, sign\, X_1 = -1)$. □

Věta 4 Nechť jsou splněny předpoklady věty 3. Pak

$$E\, S^+ = 0, \quad var\, S^+ = \frac{1}{n} \sum_{j=1}^{n} a^2(j)$$

Důkaz Platí $E\, S^+ = \sum_{i=1}^{n} \left(E\, sign\, X_i \right) \left(E\, a(R_i^+) \right)$

$$E\, sign\, X_i = 0$$
$$var\, S^+ = ...$$

Outline

II.2. Pořadová statistiky pro test hypotézy symetrie

$X_1, \ldots, X_n$ nezávislé stejně rozdělené náh. veličiny s distr. fcí F

Hypotéza symetrie: formulace

$$H_0 : F(x) + F(-x) = 1, \quad x \in R^1$$

ex.-li hustota f pak

$$H_0 : f(x) = f(-x), \quad x \in R^1 \quad (az \quad Leb. \quad miru \quad 0)$$

v tomto případě je *medián* roven 0

zobecnění: ex. a reálné, že $f(x+a) = f(a-x)$ pro vš. x

alternativní hypotéza: H_0 neplatí

používají se mimo jiné testy založené na S^+

motivace pro H_0 a klasický t-test

dvojice nezávislých pozorovní $(Y_1, Z_1), \ldots, (Y_n, Z_n)$,

(Y_i, Z_i) má distribuční funkci $F(y, z)$

Z_i – odpovídá ošetření (zdravotní stav před léčbou)

Y_i – odpovídá kontrole (zdravotní stav po léčbou)

$H_0 : F(y, z) = F(z, y)$ pro vš. y, z

tedy ošetření nemá vliv, popř. léčba nemá vliv na zdravotní stav

často definujeme: $X_i = Y_i - Z_i,\ i = 1, \ldots$

G – distr. fce náhodné veličiny X_i

pak $H_0 : G(x) + G(-x) = 1$

je-li G distribuční fce $N(\mu, \sigma^2)$, $\mu \in R^1$ and $\sigma^2 > 0$ neznmé

pak $H_0 : \mu = 0$ a aplikujeme tzv. jednovýběrový t-test testová statistika:

$T = \frac{\overline{X}_n}{s_n(x)}\sqrt{n}$

$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad s_n^2(x) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$

zamítáme na hladině α, jestliže

$$|T| \geq t_{n-1,1-\alpha/2}$$

kde $t_{n-1,1-\alpha/2}$ je kvantil t-rozdělení on $n-1$ stupních volnosti

Test je vhodný, pokud je splněn předpoklad normality nebo je n dost velké a jsou konečné momenty (centrální limitní věta).

Wilcoxonův jednovýběrový test

$$a(i) = i, \quad S^+ = \sum_{i=1}^{n} sign\, X_i\, R_i^+$$

ekvivalentní vyjádření

$$S^* = \sum_{i=1, X_i>0}^{n} R_i^+, \qquad S^{**} = \sum_{i=1, X_i<0}^{n} R_i^+$$

Platí:

$$S^+ = S^* - S^{**}$$

$$S^* + S^{**} \sum_{i=1}^{n} i = \frac{n(n+1)}{2}, \quad S^{**} = \frac{n(n+1)}{2} - S^*,$$

$$S^+ = 2S^* - \frac{n(n+1)}{2},$$

$$E\, S^+ = 0, \quad var\, S^+ = \sum_{i=1}^{n} i^2 = n(n+1)(2n+1)/6$$

Lze dokázat:

Věta 5 Nechť $X_1, \dots, X_n$ jsou nezávislé náhodné veličiny se spojitou distribuční funkcí F takovou, že

$$F(x) + F(-x) = 1, \quad pro \quad vs. \quad x \in R.$$

Pak pro $n \to \infty$

$$\sup_x |P(S^+ < x\sqrt{n(n+1)(2n+1)/6}) - \Phi(x)| \to 0$$

kde $\Phi(x)$ je hodnota distribuční funkce normovaného normálního rozdělení $N(0,1)$.

Test:

(i) Statistický software spočítá tzv. p-hodnotu.

(ii) Pro malá n lze použít tabulky, napv r. v Andělovi, jsou tabelovány hodnoty $w_n(\alpha)$ takové, že

$$P(\min(S^*, S^{**}) \leq w_n(\alpha)) = \alpha$$

a hypotézu H_0 zamítáme, jestliže $\min(S^*, S^{**}) \leq w_n(\alpha)$, α je hladina testu

Příklad 10 pokusných osob mělo nezávisle na sobě a bez předchozího tréninku odhadnout, kdy od daného signálu uplyne 1 minuta.

Výsledky (v sekundách):

53, 48, 45, 55, 63, 51, 66, 56, 50, 58

H_0 : F je symetrická kolem 60, tj. $F(x-60)+F(-(x-60))=1$

odeţeme 60 od každého pozorování

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

to vede na hypotézu symetrie kolem 0

spočteme Wilcoxonovu jednovýběrovou statistiku:

$R_1^+ = 6, \quad R_2^+ = 9, \quad R_3^+ = 10, \quad R_4^+ =, \quad R_5^+ = 2, \quad R_6^+ = 7, \quad R_7^+ = 5, \quad R_8^+ = 3, \quad R_9^+ = 8, \quad R_{10}^+ = 1$

$$S^* = 2+5 = 7, \quad S^{**} = \frac{10 \times 11}{2} - S^* = 48, \quad \min(S^*, S^{**}) = 7 < w_{10}(0,05) =$$

Tedy zamítáme H_0 na hladině $\alpha = 0,05$

Použijeme-li limitní věty:

$$|S^+| = 41 > \Phi^{-1}(0,975)\sqrt{387} = 1,96 \times 3,???$$

Tedy i tady zamítáme.

Znaménkový test

$$S^+ = sgn\, X_i, \quad S^* = pocet \quad X_i > 0, \quad S^{**} = pocet \quad X_i < 0$$

Používá se pro test: $H_0 : \ median\ = 0$

S^* má binomické rozdělení $(n, P(X_i > 0))$, jsou-li $X_1, \dots, X_n$ nezávislé stejně rozdělené náhodné veličiny

$H_0 : P(X_i > 0) = 1/2$

H_0 zamítáme, jestliže

$$S^* \leq k_1, \qquad S^{**} \geq n - k_1$$

kde

$$P_{H_0}(S^* \leq k_1)\alpha/2, \qquad P_{H_0}(S^* \geq n - k_1) \leq \alpha/2$$

Je-li n velké, má $2(S^* - n/2)/\sqrt{n}$ má přibližně $N(0,1)$ a tedy H_0 zamítneme, jestliže

$$|2S^* - n|/\sqrt{n} \geq \Phi^{-1}(1 - \alpha/2)$$

Outline

II.3. Dvouvýběrový problém

$X_1, \dots, X_{n_1}$ jsou nezávislé stejně rozdělené náhodné veličiny se spojitou distribuční funkcí F
$Y_1, \dots, Y_{n_2}$ jsou nezávislé stejně rozdělené náhodné veličiny se spojitou distribuční funkcí G
všechny náhodné veličiny nezávislé

$$H_0 : F = G \qquad H_1 : F \neq G$$

$R_1, \dots, R_{n_1+n_2}$ jsou pořadí odpovídající $X_1, \dots, X_{n_1}, Y_1, \dots, Y_{n_2}$
Používají se testové statistiky:

$$S_1 = \sum_{i=1}^{n_1} a(R_i), \qquad S_2 = \sum_{i=n_1+1}^{n_2+n_1} a(R_i)$$

Zřejmě

$$S_1 + S_2 = \sum^{n_1+n_2} a(i)$$

Nejčastěji se používá *Wilcoxonův dvouvýběrový test*:
$a(i) = i,\ i = 1, \dots, n_1 + n_2$
Pak $S_1 + S_2 = \sum_{i=1}^{n_1+n_2} i = \frac{(n_1+n_2)(n_1+n_2+1)}{2}$

$$E\, S_1 = (n_1 + n_2 + 1)n_1/2, \quad var\, S_1 = (n_1 + n_2 + 1)n_1 n_2/12$$

Častěji se používá tvar:

$$U_1 = n_1 n_2 + n_1(n_1 + 1)/2 - S_1, \quad U_2 = n_1 n_2 + n_2(n_2 + 1)/2 - S_2$$

$$U_1 + U_2 = n_1 n_2, \quad E\, U_1 = E\, U_2 = n_1 n_2/2$$

U_1 je též známo jako Mann-Whitney test

Test pro malé n_1, n_2 existují tabulky, běžný software poskytuje p hodnotu, pro n_1, n_2 velká lze použít aproximaci:

$$\sup_x \left| P_{H_0}\left(U_1 - E_{H_0} U_1 \leq x\sqrt{var\, U_1} \right) - \Phi(x) \right| \to 0$$

pro $\min(n_1, n_2) \to \infty$

Příklad Je třeba zjistit, zda jsou dva druhy hnojení ekvivalentní při stejných ostatních podmínkách.

1. skupina (X_i) $5,7 \quad 5,5 \quad 4,3 \quad 5,9 \quad 5,2 \quad 5,6 \quad 5,8 \quad 5,1$

2. skupina (Y_i) $5,0 \quad 4,5 \quad 4,2 \quad 5,4 \quad 4,4$

t-test:

$n_1 = 8, \quad \overline{X} = 5,3875, \quad s_x^2 = 0,2698$

$n_2 = 5, \quad \overline{Y} = 4,7000, \quad s_y^2 = 0,24$

$t = 2,370 > t_{11}(0,975) = 2,160$ -tedy zamítáme na hladině $\alpha = 0,05$, ovšem musí být splněny předpoklady!!!

poř adí v sech pozorování: 11, 9, 2, 13, 7, 10, 12, 6, 5, 4, 2, 8, 3

$S_1 = 70, \quad S_2 = 21, \quad U_1 = 6, \quad U_2 = 34$

zamítáme

$\frac{|U_1 - E\,U_1|}{\sqrt{var\,U_1}} = 2.049 > 1.96 = \Phi^{-1}(0,975)$

Závěrečné poznámky k základním pořadovým testům

(a) výhody pořadových testů: rozdělení testových statistik při H_0 nezávisí na distribuční funkci náhodných veličin, výpočetně jednoduché

(b) nevýhoda: ztrácíme část informace, vhodné jen pro jendoduche situace

(c) existují pořadové testy i pro další situace (testy nezávisloti, pro regresní modely)

(d) existují odhady založené na pořadích

(e) existuje teorie týkající se volby a, známe-li distribuční funkci F

(f) pozor na tzv. shody, jestliže se dvě pozorování shodují.

Kolmogorovovy-Smirnovovy testy

$X_1, \ldots, X_n$ jsou nezávislé stejně rozdělené náhodné veličiny s spojitou distribuční funkcí F

empirická distribuční funkce F_n je definována:

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I\{X_i \leq x\} = \frac{1}{n}\, pocet \quad X_i \leq x, \qquad x \in R$$

Platí:

$$P(\sup_x |F_n(x) - F(x)| \geq \varepsilon) \to 0, \quad n \to \infty$$

pro každé $\varepsilon > 0$

Matematická statistika A

Outline

Úvod, struktura přednášky

I. Opakování základních pojmů

II. Pořadové testy

II.2. Pořadová statistiky pro test hypotézy symetrie

II.3. Dvouvýběrový problém

Lze využít pro konstrukci testů:

(i) $X_1, \ldots, X_n$ jsou nezávislé stejně rozdělené náhodné veličiny se spojitou distribuční funkcí F a testujeme hypotézu:

$$H_0 : F = F_0 \left(F_0 - \ dano\right) \quad H_1 : F \neq F_0$$

Test:

$$\sup_x |F_n(x) - F(x)| \sqrt{n} \geq q_n(1-\alpha)$$

$q_n(1-\alpha)$ určeno tak, aby test měl hladinu α, pro malá n existují tabulky, pro velká se používá limitní rozdělení. Statistický software spočítá p- hodnotu.

Existuje i verze, kde v H_0 je místo F_0 množina disttribučních funkcí, statitický software existuje i pro tuto variantu.

(ii) *Dvouvýběrový Kolmogorovův-Smirnovův test*

$X_1, \ldots, X_{n_1}$ jsou nezávislé stejně rozdělené náhodné veličiny se spojitou distribuční funkcí F
$Y_1, \ldots, Y_{n_2}$ jsou nezávislé stejně rozdělené náhodné veličiny se spojitou distribuční funkcí G
všechny náhodné veličiny nezávislé

$$H_0 : F_1 = F_2, \qquad F_1 \neq F_2$$

Test

$$\sup_x |F_{n_1}(x) - G_{n_2}(x)| \sqrt{n} \geq q_{n_1,n_2}(1-\alpha)$$

$q_{n_1,n_2}(1-\alpha)$ určeno tak, aby test měl hladinu α, pro malá n_1 a n_2 existují tabulky, pro velká se používá limitní rozdělení. Statistický software spočítá p- hodnotu.