11 Neparametrické metody

11.1 Jednovýběrové testy

11.1.1 Znaménkový test

Nechť $X_1, \dots, X_n$ je výběr ze spojitého rozdělení s jediným mediánem $\tilde{x}$. Platí tedy

$$\mathsf{P}(X_i < \tilde{x}) = \mathsf{P}(X_i > \tilde{x}) = \frac{1}{2}, \qquad i = 1, \dots, n.$$

Testujme hypotézu $H_0 : \tilde{x} = x_0$, kde x_0 je dané číslo. Zabývejme se oboustranným testem, kdy alternativou je $H_1 : \tilde{x} \neq x_0$. Utvoří se rozdíly $X_1 - x_0, \dots, X_n - x_0$. Počet rozdílů s kladným znaménkem označme Y. Platí-li H_0, má Y binomické rozdělení $\mathsf{Bi}(n, \frac{1}{2})$. Hypotézu H_0 zamítneme, bude-li Y blízké nule nebo blízké číslu n. Je-li n malé, používají se tabulky kritických hodnot k_1 a k_2 s vlastnostmi

$$\mathsf{P}(Y \leq k_1) \leq \frac{\alpha}{2}, \qquad \mathsf{P}(Y \geq k_2) \leq \frac{\alpha}{2}. \tag{11.1}$$

Přitom k_1 je největší a k_2 nejmenší z čísel, pro něž (11.1) platí. Vzhledem k symetrii rozdělení $\mathsf{Bi}(n, \frac{1}{2})$ je $k_2 = n - k_1$.

Hypotézu H_0 tedy zamítneme, bude-li $Y \leq k_1$ nebo $Y \geq k_2$. Hladina testu je nejvýše rovna α. Obvykle je však značně menší než α, zejména při malých hodnotách n.

Zaveďme náhodné veličiny $\xi_1, \dots, \xi_n$ tak, že $\xi_i = 0$ v případě $X_i - x_0 \leq 0$ a $\xi_i = 1$ v případě $X_i - x_0 > 0$. Tudíž $Y = \xi_1 + \dots + \xi_n$. Platí-li H_0, je $\mathsf{E}\xi_i = \frac{1}{2}$, $\mathsf{var}\ \xi_i = \frac{1}{4}$ a veličina

$$\frac{Y - \dfrac{n}{2}}{\sqrt{n}}$$

má podle centrální limitní věty asymptoticky rozdělení $\mathsf{N}(0, \frac{1}{4})$, tzn. veličina

$$U = \frac{2Y - n}{\sqrt{n}} \tag{11.2}$$

má asymptoticky rozdělení $\mathsf{N}(0, 1)$. Hypotézu H_0 zamítneme, když $|U| \geq u(\frac{\alpha}{2})$. Hladina tohoto testu se s rostoucím n blíží číslu α. V praxi se tohoto postupu

používá, je-li $n \geq 20$. Ekvivalentně se někdy místo toho používá modifikace, při níž se H_0 zamítá, když $U^2 \geq \chi_1^2(\alpha)$.

Další možností je využít *transformací stabilizujících rozptyl* vzhledem k tomu, že často mají lepší normální aproximaci. V případě $\xi \sim \mathsf{Bi}(n,p)$ pro $W = \xi/n$ máme $\mathsf{E}W = p$, $\mathsf{var}\, W = \sigma^2(p) = p(1-p)/n$. Ze vzorce (6.5) při $c = 1/(2\sqrt{n})$ dostaneme $g(p) = \arcsin\sqrt{p}$. Odtud máme $\mathsf{E}g(W) \doteq g(p)$, $\mathsf{var}\, g(W) \doteq 1/(4n)$. Funkci g se říká *arkussinová transformace*. Někdy se používá transformace

$$g_1(p) = \arcsin\sqrt{\left(p + \frac{3}{8n}\right) \Big/ \left(1 + \frac{3}{4n}\right)},$$

která je v jistém smyslu přesnější. Zde máme $\mathsf{E}g_1(W) \doteq g_1(p)$, $\mathsf{var}\, g_1(W) \doteq 1/(4n+2)$. Vrátíme-li se zpět ke znaménkovému testu, v první variantě můžeme použít veličinu

$$U_1 = 2\sqrt{n}\left(\arcsin\sqrt{\frac{Y}{n}} - \arcsin\sqrt{\frac{1}{2}}\right)$$

a H_0 zamítneme, když $|U_1| \geq u(\frac{\alpha}{2})$. Ve druhé variantě vypočítáváme

$$U_2 = \sqrt{4n+2}\left(\arcsin\sqrt{\frac{8Y+3}{8n+6}} - \arcsin\sqrt{\frac{1}{2}}\right)$$

a H_0 zamítáme v případě $|U_2| \geq u(\frac{\alpha}{2})$. Z testů založených na U, U_1 a U_2 lze doporučit první z nich, protože nejlépe zachovává hladinu chyby prvního druhu (viz Zvára 1995). Statistické programy umožňují provést i exaktní test.

Znaménkový test používáme zejména v případě, kdy rozdělení veličin X_i je výrazně zešikmené. Jelikož tento test má poměrně malou sílu (pravděpodobnost chyby druhého druhu je ve srovnání s jinými testy dosti veliká), je žádoucí mít k dispozici větší počet pozorování n.

Jestliže se z technických důvodů stane, že některé rozdíly $X_i - x_0$ jsou rovny nule (což teoreticky má nulovou pravděpodobnost, ale může k tomu dojít třeba vlivem zaokrouhlovacích chyb), pak se tyto hodnoty vynechají a o jejich počet se sníží číslo n. Jinak test proběhne beze změny.

Znaménkový test může být také *jednostranný*. *Mnohorozměrný znaménkový test* uvádí Bennett (1962).

Znaménkový test řadíme mezi neparametrické testy, protože k jeho odvození nebylo nutné pro daný výběr specifikovat přesný typ rozdělení či dokonce jeho parametry.

Příklad 11.1 Deset pokusných osob mělo nezávisle na sobě bez předchozího nácviku odhadnout, kdy od daného signálu uplyne jedna minuta. Byly získány následující výsledky (v sekundách):

$$53, \quad 48, \quad 45, \quad 55, \quad 63, \quad 51, \quad 66, \quad 56, \quad 50, \quad 58.$$

Testujeme-li hypotézu, že medián rozdělení je $x_0 = 60$ sekund, tj. $H_0 : \tilde{x} = 60$, vypočteme veličiny $Y_i = X_i - 60$, které jsou rovny

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

Máme $n = 10$, počet rozdílů s kladným znaménkem je $Y = 2$. Podle tabulky kritických hodnot při $n = 10$ a $\alpha = 0{,}05$ je $k_1 = 1$, $k_2 = 9$. Protože $Y \in (k_1, k_2)$, nelze znaménkovým testem nulovou hypotézu zamítnout. Dá se vypočítat, že v tomto případě skutečná hladina testu není 0,05, ale jen 0,02. Ale test, který by použil sousedních hodnot $k_1 = 2$, $k_2 = 8$, už by měl hladinu 0,11.

Pro ilustraci uveďme, že v našem případě vyjde $U = -1{,}897$, $U_1 = -2{,}035$ a $U_2 = -1{,}919$. Při $\alpha = 0{,}05$ by se absolutní hodnota těchto čísel porovnávala s kritickou hodnotou 1,96. $\diamondsuit$

11.1.2 Jednovýběrový Wilcoxonův test

Nechť $X_1, \ldots, X_n$ je náhodný výběr ze spojitého rozdělení s hustotou f, která je symetrická kolem bodu a a kladná v jeho okolí. Platí tedy $f(a + x) = f(a - x)$. Z toho plyne, že a musí být rovno mediánu $\tilde{x}$. Existuje-li konečná střední hodnota tohoto rozdělení, pak musí také pro každé i platit $\mathsf{E}X_i = a$. Konečnost střední hodnoty se však obecně nepředpokládá. *Jednovýběrový Wilcoxonův test* je určen k testování hypotézy $H_0 : \tilde{x} = x_0$ proti alternativě $H_1 : \tilde{x} \neq x_0$.

Nejprve předpokládejme, že žádná z veličin X_i není rovna x_0. Položme $Y_i = X_i - x_0$. Veličiny Y_i seřadíme do neklesající posloupnosti podle jejich absolutní hodnoty

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

Budiž R_i^+ pořadí veličiny $|Y_i|$. Označme

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

Přitom platí $S^+ + S^- = n(n+1)/2$. Je-li číslo $\min(S^+, S^-)$ menší nebo rovno tabelované kritické hodnotě $w_n(\alpha)$, zamítáme H_0.

Z učiněných předpokladů vyplývá, že při platnosti H_0 jsou $Y_1, \ldots, Y_n$ nezávislé stejně rozdělené náhodné veličiny, jejichž rozdělení je symetrické kolem nuly.

Věta 11.2 *Platí-li H_0, pak vektory* $(\operatorname{sign} Y_1, \ldots, \operatorname{sign} Y_n)'$ *a* $(|Y|_{(1)}, \ldots, |Y|_{(n)})'$ *jsou nezávislé.*

Důkaz. Jelikož veličiny Y_i jsou nezávislé, vektory $(\operatorname{sign} Y_i,\ |Y_i|)'$ jsou také nezávislé. Ze spojitosti a ze symetrie rozdělení vyplývá, že

$$\mathsf{P}(\operatorname{sign} Y_i = 1) = \mathsf{P}(\operatorname{sign} Y_i = -1) = \frac{1}{2}.$$

Dále máme pro libovolné $y > 0$

$$\begin{aligned}\mathsf{P}(\text{sign } Y_i = 1, |Y_i| < y) &= \mathsf{P}(0 < Y_i < y) = \frac{1}{2}\mathsf{P}(-y < Y_i < y) \\ &= \frac{1}{2}\mathsf{P}(|Y_i| < y) = \mathsf{P}(\text{sign } Y_i = 1)\,\mathsf{P}(|Y_i| < y).\end{aligned}$$

Proto veličiny sign Y_i a $|Y_i|$ jsou pro každé i nezávislé. Celkově dostáváme, že vektory $(\text{sign } Y_1, \ldots, \text{sign } Y_n)'$ a $(|Y_1|, \ldots, |Y_n|)'$ jsou nezávislé. Protože vektor $(|Y|_{(1)}, \ldots, |Y|_{(n)})'$ je funkcí vektoru $(|Y_1|, \ldots, |Y_n|)'$, je tím věta dokázána. □

Věta 11.3 *Označme* $S = \sum_{i=1}^{n} R_i^+ \text{ sign } Y_i$. *Pak*

$$S^+ = \frac{1}{2}S \; + \frac{n(n+1)}{4}.$$

Důkaz. Platí $S^+ - S^- = S$, $S^+ + S^- = n(n+1)/2$. Odtud vypočteme S^+. □

Věta 11.4 *Platí-li* H_0, *pak*

$$\mathsf{E}S^+ = \frac{1}{4}n(n+1), \qquad \mathsf{var}\; S^+ = \frac{1}{24}n(n+1)(2n+1).$$

Důkaz. Nejprve si všimneme, že $\mathsf{E}\,\text{sign } Y_i = 0$ pro každé i. Z věty 11.2 dostaneme, že $\mathsf{E}(R_i^+ \,\text{sign}\, Y_i) = (\mathsf{E}R_i^+)(\mathsf{E}\,\text{sign}\, Y_i)$, a tak

$$\mathsf{E}(R_i^+ \text{ sign } Y_i) = 0. \tag{11.3}$$

Odtud

$$\mathsf{E}S = \sum_{i=1}^{n} \mathsf{E}(R_i^+ \text{ sign } Y_i) = 0.$$

Vzhledem k (11.3) platí

$$\begin{aligned}\mathsf{var}(R_i^+ \text{ sign } Y_i) &= \mathsf{E}(R_i^+ \text{ sign } Y_i)^2 = \mathsf{E}(R_i^+)^2\mathsf{E}(\text{sign } Y_i)^2 \\ &= \mathsf{E}(R_i^+)^2 = 1^2\frac{1}{n} + 2^2\frac{1}{n} + \cdots + n^2\frac{1}{n} = \frac{1}{6}(n+1)(2n+1).\end{aligned}$$

Obdobně se dokáže, že platí

$$\mathsf{cov}(R_i^+\text{sign}\, Y_i, R_j^+\text{sign}\, Y_j) = 0 \qquad \text{pro } i \neq j.$$

Nyní se použije věta o rozptylu součtu náhodných veličin. □

Dá se dokázat (viz Hájek a Šidák 1967), že za platnosti H_0 má S^+ asymptoticky normální rozdělení. Test hypotézy H_0 lze tudíž také založit na veličině

$$U = \frac{S^+ - \mathsf{E}S^+}{\sqrt{\mathsf{var}\; S^+}},$$

kde $\mathsf{E}S^+$ a $\mathsf{var}\ S^+$ jsou uvedeny ve větě 11.4. Vyjde-li $|U| \geq u(\frac{\alpha}{2})$, zamítneme H_0 na hladině, která se s rostoucím n blíží číslu α.

Je třeba zdůraznit, že jedním z předpokladů jednovýběrového Wilcoxonova testu je i symetrie hustoty f kolem mediánu. K zamítnutí H_0 může tedy oprávněně dojít i tehdy, je-li medián roven x_0, ale hustota f je výrazně nesymetrická.

Je-li některá z veličin X_i rovna x_0, zpravidla se toto pozorování vynechává.

Příklad 11.5 Použijeme data z příkladu 11.1. Testujme opět hypotézu, že medián rozdělení je roven $x_0 = 60$ sekund. Již jsme vypočetli, že veličiny $Y_i = X_i - 60$ jsou rovny

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

Seřadíme je do neklesající posloupnosti podle jejich absolutních hodnot. Dostaneme

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

Číslo 3 má pořadí rovné dvěma, číslo 6 má pořadí rovné pěti. Proto $S^+ = 2+5 = 7$. Odtud $S^- = 10 \times 11/2 - S^+ = 48$. Rozsah výběru $n = 10$ je poměrně malý, proto užijeme přesných kritických hodnot. V tabulkách kritických hodnot najdeme $w_{10}(0{,}05) = 8$. Protože $\min(S^+, S^-) = 7 \leq w_{10}(0{,}05) = 8$, zamítneme hypotézu, že v lidské populaci polovina osob délku jedné minuty podhodnotí a polovina nadhodnotí. Kdybychom pro ilustraci užili asymptotický postup, měli bychom

$$\mathsf{E}S^+ = 27{,}5, \qquad \mathsf{var}\ S^+ = 96{,}25, \qquad U = -2{,}09.$$

Protože $|U| \geq u(0{,}025) = 1{,}96$, zamítla by se nulová hypotéza i tímto postupem. ♢

11.2 Dvouvýběrové testy

11.2.1 Dvouvýběrový Wilcoxonův test

Nechť $X_1, \ldots, X_m$ je náhodný výběr ze spojitého rozdělení s distribuční funkcí F a nechť $Y_1, \ldots, Y_n$ je na něm nezávislý náhodný výběr ze spojitého rozdělení s distribuční funkcí G. Je třeba testovat hypotézu $H_0 : F = G$ proti alternativě $H_1 : F \neq G$.

Všech $m + n$ veličin $X_1, \ldots, X_m, Y_1, \ldots, Y_n$ (tzv. *sdružený výběr*) uspořádáme vzestupně podle velikosti. Označme T_1 součet pořadí hodnot $X_1, \ldots, X_m$ a T_2 součet pořadí hodnot $Y_1, \ldots, Y_n$. Je jasné, že

$$T_1 + T_2 = \frac{1}{2}(m+n)(m+n+1).$$