Marie Turčičová

Neparametrické jednovýběrové testy

Mějme data ve tvaru

$$X_1, X_2, \dots, X_n$$

která představují měření nějakého znaku provedené na n jedincích. Průměrnou hodnotu daného znaku v populaci (tj. populační průměr) si označíme jako μ, tedy naše data jsou výběrem z rozdělení se střední hodnotou μ. Chceme testovat hypotézu:

$$\begin{aligned} H_0 : \mu = c & \quad \text{proti alternativě} \\ H_1 : \mu \neq c & \quad (\text{nebo } H_1 : \mu < c, \text{ nebo } H_1 : \mu > c). \end{aligned}$$

Na to je potřeba nějaký jednovýběrový test. Co si ale počít, když se nám pro naše data nepovede ověřit předpoklad normálního rozdělení a nepůjde tedy použít jednovýběrový t-test? V takovém případě se musíme uchýlit k testu, který tento předpoklad nepotřebuje. Těmto testům se většinou říká neparametrické, nebo též pořadové, protože většina z nich pracuje s pořadími. A jaké jsou tedy neparametircké alternativy k jednovýběrovému t-testu?

Jednovýběrový Wilcoxonův test

Předpoklady: $X_1, X_2, \dots, X_n$ je výběr ze **spojitého** a **symetrického** rozdělení.
Většina neparametrických testů již nepracuje s pojmem populačního průměru, ale s populačním mediánem (což je medián zkoumaného znaku ve studované populaci). Tomu odpovídá i tvar testovaných hypotéz:

$$\begin{aligned} &H_0 : \text{populační medián znaku } X \text{ je roven } c \\ &H_1 : \text{populační medián znaku } X \text{ je různý od } c \end{aligned}$$

Jak spočítat testovou statistiku? (Index i v dalším označuje i-tého jedince a nabývá hodnoty $1, \dots, n$).

- z dat vyloučíme případy, kdy je $X_i = c$ (tím dojde ke změně počtu pozorování, výsledný počet si označme n^*)
- veličiny $X_i - c$ seřadíme do neklesající posloupnosti podle absolutní hodnoty a přiřadíme jim pořadí R_i^+
- označme si W = součet pořadí, kdy bylo $X_i > c$
- testová statistika má tvar

$$Z = \frac{W - \mathrm{E}\,W}{\sqrt{\operatorname{var} W}} = \frac{W - n^*(n^*+1)\frac{1}{4}}{\sqrt{n^*(n^*+1)(2n^*+1)\frac{1}{24}}} \tag{1}$$

  která má v případě, že H_0 opravdu platí, a je-li n^* dost velké, normované normální rozdělení $N(0,1)$.

Vyjde-li nám tedy $|Z| \geq qnorm(1 - \frac{\alpha}{2})$, kde $qnorm$ značí příslušný kvantil rozdělení $N(0,1)$, tak zamítáme H_0 ve prospěch H_1.

Uveďme si ještě pro ilustraci malý příklad výpočtu hodnoty W. Předpokládejme, že $c = 2$.

i	1	2	3	4	5	6
x_i	8	5	1	0	-2	2
$x_i - c$	6	3	-1	-2	-4	0
$\lvert x_i - c\rvert$	6	3	1	2	4	-
R_i^+	**5**	**3**	1	2	4	-

Z tabulky vidíme, že $n^* = 5$ (hodnotu $x_6 = c$ jsme vyřadili) a $W = 5 + 3 = 8$.

Znaménkový test

V případě, že rozdělení zkoumaného znaku v populaci nelze pokládat za symetrické, nelze jednovýběrový Wilcoxonův test použít. Místo něho nám poslouží tzv. znaménkový test.
Předpoklady: $X_1, \ldots, X_n$ je výběr ze spojitého rozdělení.
Testovaná hypotéza:

$$H_0 : \text{populační medián znaku } X \text{ je roven } c$$
$$H_1 : \text{populační medián znaku } X \text{ je různý od } c$$

Výpočet testové statistiky:

- vynecháme pozorování, kdy $X_i = c$ (počet zbylých pozorování si označíme n^*)
- spočteme rozdíly $X_1 - c, X_2 - c, \ldots, X_n - c$
- počet rozdílů s kladným znaménkem označíme U

Pro naše cvičná data z tabulky by nám vyšlo $U = 2$ a $n^* = 5$. Testová statistika má tvar

$$Z_2 = \frac{U - \mathrm{E}\,U}{\mathrm{var}\,U} = \frac{U - \frac{n^*}{2}}{\sqrt{\frac{n^*}{4}}} \tag{2}$$

a pokud by H_0 platila, měla by mít veličina Z_2 pro velká n^* přibližné normované normální rozdělení. Tj. asymptoticky za H_0 by mělo platit $Z_2 \sim N(0,1)$.

Je-li c opravdu hodnota populačního mediánu, měla by být v datech přibližně polovina pozorování větších než c a přibližně polovina menších než c. Hodnota U by tedy neměla být ani příliš malá, ani příliš velká. Pokud se stane, že U (a tudíž i Z_2) budou příliš malé nebo velké, zamítneme H_0. Za pomoci té asymptotické normality si kritický obor vymezíme pomocí kvantilů normovaného normálního rozdělení. Tj. nulovou hypotézu zamítneme pokud $|Z_2| \geq qnorm(1 - \frac{\alpha}{2})$.

Pro malá n^* je vhodné použít tzv. Yatesovu korekci, tj. testová statistika má potom tvar

$$Z_3 = \frac{|U - \frac{n^*}{2}| - \frac{1}{2}}{\sqrt{\frac{n^*}{4}}}. \tag{3}$$

Rozhodovací kritérium je pak stejné jako dříve.

Poznámka 1 *U obou zmiňovaných testů je samozřejmě možné volit též jednostranné alternativní hypotézy. Kritickou hodnotou vedoucí k zamítnutí H_0 pak bude kvantil $qnorm(1-\alpha)$, nebo $-qnorm(1-\alpha)$.*

Poznámka 2 *Jak jste si asi všimli, testům postupně ubývaly předpoklady:*

- *t-test potřebuje výběr z normálního rozdělení*
- *Wilcoxonovu testu stačí již pouze spojité a symetrické rozdělení*
- *znaménkový test se spokojí pouze se spojitým rozdělením.*

Spolu s předpoklady ale testům ubývala též síla. Znaménkový test je sice co do předpokladů nejskromnější, ale má také nejmenší sílu. Naopak jednovýběrový t-test je nejsilnější.

Poznámka 3 *Asymptotická normalita testových statistik Z a Z_1 je důsledkem centrální limitní věty.*