PRAVDĚPODOBNOST A STATISTIKA, PŘEDNÁŠKA 4. 5. 2020

Doporučená literatura: Likeš J., Machek J.: Matematická statistika. SNTL, Praha 1982.
Anděl J.: Statistické metody. MatfyzPress 2019

Neparametrické testy

V předchozích částech jsme se věnovali testům parametrů normálního rozdělení, které (důsledek centrálních limitních vět) nejsou při větších rozsazích náhodných výběrů příliš citlivé na porušení předpokladu normality dat a jsou proto relativně univerzálněji použitelné. V praxi se však lze setkat s náhodnými výběry menších rozsahů, které pochází z výrazně nenormálních rozdělení a v tomto případě nelze zmíněné testy použít. Vhodnou alternativou je použití tzv. neparamterických testů, které (zjednodušeně řečeno) nepředpokládají konkrétní typ rozdělení a vychází z mnohem obecnějších předpokladů (např. spojitost rozdělení).

Pořadové testy

Důležitou, a díky své jednoduchosti v praxi i poměrně oblíbenou, kategorii tvoří tzv. pořadové testy, v nichž se napozorované údaje nahrazují pořadím.

Pořadí - nechť $X_1, \dots, X_n$ jsou různá reálná čísla, potom symbolem R_i označujeme pořadí X_i, které definujeme jako počet hodnot $X_j, j = 1, \dots, n$ pro která platí $X_j \le X_i$. Jsou-li některá z čísel $X_1, \dots, X_n$ sobě rovna, mluvíme o tzv. shodách, a uvedeným číslům přiřazujeme jejich průměrné pořadí.
Je-li $X_{i1} = X_{i2} = \cdots = X_{ik}$ shoda, potom $R_{i1} = R_{i2} = \cdots = R_{ik} = m + \frac{k+1}{2}$, kde m je počet hodnot ostře menších než X_{i1} a k je počet shodných hodnot. Viz příklad níže.

X_i	6	6	1	0	7	7	7
R_i	3,5	3,5	2	1	6	6	6

Znaménkový test

$X_1, \dots, X_n$ je náhodný výběr ze spojitého rozdělení, testujeme:

$$H_0 \colon x_{0,5} = x_0 \quad \times \quad H_1 \colon x_{0,5} \neq x_0,$$

kde $x_{0,5}$ je neznámý medián a x_0 je daná konstanta.

Postup:

1. Vypočteme rozdíly $X_1 - x_0, \dots, X_n - x_0$.
   Jako testovou statistiku použijeme $Y = |\{X_i - x_0 > 0, i = 1, \dots, n\}|$, tj. počet kladných rozdílů $X_i - x_0$.
2. Rozhodneme o výsledku testu - H_0 zamítneme na hladině α, jestliže
   $$Y \le z_{\alpha/2}(n), \text{ nebo } Y \ge z_{1-\alpha/2}(n),$$
   kde $z_\alpha(n)$ je kritická hodnota znaménkového testu (viz stat. tabulky, resp. stat. software).

Poznámky

- Pokud některé rozdíly $X_i - x_0$ vyjdou rovny nule, tak je obvykle vynecháme a zmenšíme rozsah výběru.
- Znaménkový test používáme často jako párový test, tj. máme $(X_1, Y_1), \dots, (X_n, Y_n)$ a testujeme shodu mediánů, tj. $H_0: x_{0,5} = y_{0,5} \quad \times \quad H_1: x_{0,5} \neq y_{0,5}$.
- Zřejmou nevýhodou znaménkového testu je skutečnost, že nebere v úvahu velikosti $X_i - x_0, i = 1, \dots, n$.
- Je zřejmé, že pokud platí H_0, má testová statistika rozdělení $Bi\left(n, p = {}^1\!/_2\right)$ a tedy H_0 zamítneme, jestliže hodnota testové statistiky Y je „velká“ (tj. blízká n), nebo naopak malá (tj. blízká 0).

  Kritické hodnoty můžeme určit výpočtem, neboť platí
  $$P\big(Y \leq z_\alpha(n)\big) = 2^{-n}\left[\binom{n}{0} + \binom{n}{1} + \cdots + \binom{n}{z_\alpha(n)}\right] \leq \alpha, \text{ resp.}$$
  $$P\big(Y \geq z_{1-\alpha}(n)\big) = 2^{-n}\left[\binom{n}{z_{1-\alpha}(n)} + \binom{n}{z_{1-\alpha}(n)+1} + \cdots + \binom{n}{n}\right] \leq \alpha.$$
  Pro kritické hodnoty navíc zřejmě platí $z_\alpha(n) + z_{1-\alpha}(n) = n$.
- Pro dostatečně velký rozsah n (v praxi alespoň 20) využijeme centrální limitní věty a z nich plynoucí skutečnost, že pokud platí H_0, má veličina $U = \frac{2Y-n}{\sqrt{n}}$ rozdělení $N(0,1)$. Tedy H_0 zamítneme na hladině α, jestliže $|U| \geq u_{1-\alpha/_2}$.

Jednovýběrový Wilcoxonův test

$X_1, \dots, X_n$ je náhodný výběr ze spojitého rozdělení symetrického kolem (neznámého) mediánu $x_{0,5}$. Testujeme:
$$H_0: x_{0,5} = x_0 \quad \times \quad H_1: x_{0,5} \neq x_0,$$
kde x_0 je daná konstanta.

Postup:

1. Analogicky znaménkovému testu vypočteme $Z_i = X_i - x_0$. Určíme pořadí absolutních hodnot $|Z_i|$, které označíme R_i^+.
2. Vypočteme Wilcoxonovu testovou statistiku $S^+ = \sum_{Z_i>0} R_i^+$

   (alternativně $S^- = \sum_{Z_i<0} R_i^+$, zřejmě platí $S^+ + S^- = {}^{n(n+1)}\!/_2$)
3. Rozhodneme o výsledku testu - H_0 zamítneme na hladině α, jestliže
   $$S^+ \leq w_{\alpha/_2}(n), \text{ nebo } S^+ \geq w_{1-\alpha/_2}(n),$$
   kde $w_\alpha(n)$ je kritická hodnota jednovýběrového Wilcoxonova testu (viz statistické tabulky).

Poznámky

- Testová statistika jednovýběrového Wilcoxonova testu zohledňuje velikosti rozdílů $X_i - x_0$ a test je ve svém důsledku „citlivější“ než znaménkový test.
- Pokud některé rozdíly $X_i - x_0$ vyjdou rovny nule, tak je obvykle vynecháme a zmenšíme rozsah výběru.
- Pro dostatečně velká n lze použít testovou statistiku $U = \frac{S^+ - {}^{n(n+1)}\!/_4}{\sqrt{{}^{n(n+1)(2n+1)}\!/_{24}}}$, která má (důsledek centrálních limitních vět) za platnosti H_0 rozdělení $N(0,1)$, tedy H_0 zamítneme na α, jestliže $|U| \geq u_{1-\alpha/_2}$
- Jednovýběrový Wilcoxonův test používáme často jako párový test.

Mann-Whitney-Wilcoxonův test

$X_1, \dots, X_n; Y_1, \dots, Y_m$ jsou dva nezávislé náhodné výběry ze spojitých rozdělení s neznámými distribučními funkcemi $F(x)$, resp. $G(x)$.
Testujeme shodu distribučních funkcí, tj.:

$$H_0: \forall x \in R \quad F(x) = G(x) \quad \times \quad H_1: \exists x \quad F(x) \neq G(x)$$

Postup:

1. Spočteme testovou statistiku $M = \sum_{i=1}^{n} \sum_{j=1}^{m} Z_{i,j}$,
   kde $Z_{i,j} = \begin{cases} 1 & \text{, jestliže } X_i < Y_j \\ 0 & \text{, jestliže } X_i \geq Y_j \end{cases}$
   (tj. M je počet dvojic (X_i, Y_j), kde $X_i < Y_j$)
2. Rozhodneme o výsledku testu - H_0 zamítneme na hladině α, jestliže
   $$M \leq m_{\alpha/2}(n, m), \text{ nebo } M \geq m_{1-\alpha/2}(n, m),$$
   kde $m_\alpha(n, m)$ je kritická hodnota Mann-Whitney-Wilcoxonova testu (viz statistické tabulky).

Poznámky

- Zřejmě platí $m_\alpha(n, m) + m_{1-\alpha}(n, m) = nm$
- Pro dostatečně velká n, m (≥ 8), lze použít přibližný test a H_0 zamítnout na hladině α, pokud
  $\frac{|M - nm/2|}{\sqrt{nm(n+m+1)/12}} \geq u_{1-\alpha/2}$.

Test založený na Spearmanově korelačním koeficientu

$(X_1, Y_1), \dots, (X_n, Y_n)$ je dvourozměrný náhodný výběr ze spojitého rozdělení. Testujeme nezávislost X, Y, tj.:

H_0: složky X, Y jsou nezávislé $\quad \times \quad$ H_1: složky X, Y jsou závislé

Postup:

1. Určíme pořadí R_i hodnot X_i a pořadí Q_i hodnot Y_i.
   (dvojice (X_i, Y_i) se často nejprve uspořádají vzestupně dle X_i, takže $R_i = i$)
2. Vypočteme testovou statistiku $S = \sum_{i=1}^{n} (R_i - Q_i)^2$.
3. Rozhodneme o výsledku testu - H_0 zamítneme na hladině α, jestliže $S \leq s_\alpha(n)$, kde $s_\alpha(n)$ je kritická hodnota testu založeného na Spearmanově korelačním koeficientu (viz statistické tabulky).

Poznámka

- Spearmanův koeficient pořadové korelace
  $$r^{(S)} = \frac{\sum_{i=1}^{n} (R_i - \bar{R})(Q_i - \bar{Q})}{\sqrt{\left[\sum_{i=1}^{n} (R_i - \bar{R})^2\right] \cdot \left[\sum_{i=1}^{n} (Q_i - \bar{Q})^2\right]}} = 1 - \frac{6}{n(n^2 - 1)} \sum_{i=1}^{n} (R_i - Q_i)^2$$
  (v podstatě jde o výběrový korelační koeficient, ve kterém hodnoty nahradíme jejich pořadím)
  Myšlenka - jsou-li pořadí vzájemně si odpovídajících hodnot $R_i, Q_i, i = 1, \dots, n$ podobná (tj. hodnota $r^{(S)}$, resp. S jsou „malé“), svědčí to o závislosti mezi X a Y.