PRAVDĚPODOBNOST A STATISTIKA, PŘEDNÁŠKA 18. 5. 2020

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

Testy dobré shody χ^2

$X_1, \dots, X_n$ ... náhodný výběr

Testujeme hypotézu
H_0: $X_i \sim F$ (tj. náhodný výběr pochází z rozdělení s distribuční funkcí F) × H_1: $X_i \not\sim F$

Postup:

1. Stanovíme H_0, tj. stanovíme rozdělení s nimž testujeme shodu.
2. Zvolíme třídní intervaly $(t_0, t_1\rangle, (t_1, t_2\rangle, \dots, (t_{r-1}, t_r\rangle$ a sestavíme tabulku třídního rozdělení četností (n_i označíme četnost i-tého třídního intervalu $(t_{i-1}, t_i\rangle$)
3. Odhadneme neznámé parametry $\theta_1, \dots, \theta_k$ (např. metodou maximální věrohodnosti)
4. Vypočteme „teoretické“ pravděpodobnosti $p_i = P(t_{i-1} < X \leq t_i) = F(t_i) - F(t_{i-1}), i = 1, \dots, r$ (dělení na třídní intervaly by mělo být takové, aby pro teoretické četnosti intervalů platilo $np_i \geq 5$)
5. Vypočteme hodnotu testové statistiky $\chi^2 = \sum_{i=1}^{r} {(n_i - np_i)^2}/{np_i}$ a rozhodneme o výsledku testu. H_0, tj. shodu se zvoleným rozdělením zamítneme na hladině α, jestliže $\chi^2 \geq \chi^2_{1-\alpha}(r-k-1)$.

ANOVA - analýza rozptylu

Připomeňme, že dvouvýberový t-test se používá k testování shody středních hodnot dvou nezávislých náhodných výběru pocházejících z normálního rozdělení (se stejným neznámým rozptylem σ^2). Přirozeným zobecněním je pak situace, kdy chceme otestovat shodu středních hodnot více než dvou nezávislých náhodných výběrů z normálního rozdělení.

<u>Analýza rozptylu - jednoduché třídění</u>
Na prvcích statistického souboru sledujeme dva znaky X a Y. Znak X je tzv. faktor (třídící znak), který může nabývat k různých hodnot (úrovní). Máme tak k dispozici k nezávislých náhodných výběrů

$$Y_{11}, \dots, Y_{1n_1} \sim N(\mu_1, \sigma^2),$$
$$Y_{21}, \dots, Y_{2n_2} \sim N(\mu_2, \sigma^2),$$
$$\vdots$$
$$Y_{k1}, \dots, Y_{kn_k} \sim N(\mu_k, \sigma^2).$$

a testujeme hypotézu:
H_0: $\mu_1 = \dots = \mu_k$ × H_1: $\exists i \neq j \ \ \mu_i \neq \mu_j$ (alespoň dvě střední hodnoty se liší)

Postup:

1. Označíme

$n = \sum_{i=1}^{k} n_i$ ... celkový rozsah náhodného výběru

$\bar{Y} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} Y_{ij}$ ... celkový výběrový průměr

$Y_{i\cdot} = \sum_{j=1}^{n_i} Y_{ij}$ ... součet skupiny i

$\bar{Y}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij}$ ... průměr skupiny i (tzv. podmíněný výběrový průměr)

$S_t = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij} - \bar{Y}\right)^2$ ... tzv. celkový součet čtverců

$S_m = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{Y}_i - \bar{Y})^2 = \sum_{i=1}^{k} n_i(\bar{Y}_i - \bar{Y})^2$ ... tzv. meziskupinový součet čtverců

$S_e = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij} - \bar{Y}_i\right)^2$ ... tzv. reziduální součet čtverců

a sestavíme následující tabulku

Úroveň faktoru X	Hodnoty Y_{ij}	n_i	$Y_{i\cdot}$	$\bar{Y}_i$	$\sum_j Y_{ij}^2$
1	Y_{11} , … , Y_{1n_1}	n_1	$Y_{1\cdot}$	$\bar{Y}_1 = {Y_{1\cdot}}/{n_1}$	$Y_{1\cdot}^2$
2	Y_{21} , … , Y_{2n_2}	n_2	$Y_{2\cdot}$	$\bar{Y}_2 = {Y_{2\cdot}}/{n_2}$	$Y_{2\cdot}^2$
⋮	⋮	⋮	⋮	⋮	⋮
k	Y_{k1} , … , Y_{kn_k}	n_k	$Y_{k\cdot}$	$\bar{Y}_k = {Y_{k\cdot}}/{n_k}$	$Y_{k\cdot}^2$
$\sum$	- - - - -	n	$\sum_i\sum_j Y_{ij}$	- - - - -	$\sum_i\sum_j Y_{ij}^2$

součty čtverců obvykle počítáme dle níže uvedených vztahů

$$S_t = \sum_i\sum_j Y_{ij}^2 - {\left(\sum_i\sum_j Y_{ij}\right)^2}\Big/{n}, \qquad S_m = \sum_i {Y_{i\cdot}^2}/{n_i} - {\left(\sum_i\sum_j Y_{ij}\right)^2}\Big/{n}, \qquad S_e = S_t - S_m$$

2. Sestavíme tabulku analýzy rozptylu pro jednoduché třídění, vypočteme testovou statistiku F

Zdoj měnlivosti	**Součty čtverců**	**Stupně volnosti**	**Průměrné čtverce**	**Testová statistika**	**Min. hladina významnosti**
Faktor X	S_m	$k-1$	${S_m}/{(k-1)}$	$F = \frac{S_m/(k-1)}{S_e/(n-k)}$	α^*
Reziduum	S_e	$n-k$	${S_e}/{(n-k)}$		
Celkový	S_t	$n-1$	- - - - -		

α^* je minimální hladina významnosti, při které hypotézu H_0, tj. neúčinnost faktoru X zamítáme (H_0 tedy zamítáme i na každé hladině α, pro kterou $\alpha \geq \alpha^*$)

3. Hypotézu H_0 o shodě středních hodnot zamítáme na hladině α, jestliže $F \geq F_{1-\alpha}(k-1, n-k)$.

Poznámky

- Platí

$$S_t = S_m + S_e$$

Zjednodušeně lze myšlenku analýzy rozptylu u jednoduchého třídění zformulovat následovně: celkový výběrový rozptyl (reprezentovaný celkovým součtem čtverců S_t) lze rozložit na dvě složky - složku vyjadřující meziskupinovou variabilitu odpovídající faktoru X (reprezentuje meziskupinový součet čtverců S_m) a složku odpovídající vnitroskupinové variabilitě („nevysvětlitelná“ faktrorem X; reprezentuje reziduální součet čtverců S_e). Nyní lze vyvodit, že v případě shody středních hodnot bude meziskupinová variabilita v porovnání s vnitroskupinovou „malá“.

- Pokud dojde k zamítnutí H_0, je třeba rozhodnout, které střední hodnoty se liší, což provedeme pomocí tzv. mnohonásobného porovnávání.
  Sheffeho metoda - na hladině α zamítáme rovnost středních hodnot μ_r, μ_s $(r \neq s)$, jestliže

$$(\bar{Y}_r - \bar{Y}_s)^2 \geq \frac{n_r + n_s}{n_r n_s} \cdot \frac{k-1}{n-k} \cdot S_e \cdot F_{1-\alpha}(k-1, n-k)$$

- $s^2 = {S_e}/{(n-k)}$ je nejlepší nestranný odhad neznámého rozptylu σ^2