Použití jednoduchého třídění

a) prosté porovnání

Příklad: Y_{ij} jsou denní tržby v obchodě

řádky (1 - 12) odpovídají měsícům v roce

$H_0: \mu_1 = \dots = \mu_{12}$ znamená, že průměrné tržby se v jednotlivých měsících výrazně neliší

b) testování vlivu nějakého faktoru

Pro tento účel se předpokládá, že $\mu_i = \mu + \alpha_i$, kde μ je nějaká konstantní složka pro všechny výběry stejná a α_i je vliv i-tého faktoru.

Testuje se hypotéza $H_0: \alpha_1 = \dots = \alpha_I$.

Je to model rozvinutý předchozímu a výsledné vzorce jsou stejné, jenom se trochu liší jejich odvození.

Příklad: Y_{ij} je prodané množství nějakého zboží (třeba za týden)

řádky odpovídají různým způsobům reklamy

(bez reklamy, reklama v tisku, v televizi, ...)

μ odpovídá množství, které se tak jako tak prodá

(lidi potřebují nějaké potraviny nebo oblečení ať už je na něj reklama nebo ne)

α_i je nárůst prodeje po reklamě i-tého druhu

H_0 znamená, že způsob reklamy nemá na prodané množství vliv

Dvojné třídění

Testují se vlivy 2 faktorů (např. reklamy a způsobu balení), které mohou být na sobě nezávislé nebo taky ne.

a) dvojné třídění

Předpokládá se, že pozorujeme veličiny Y_{ijp}, kde

$i = 1, \dots, I$ je počet úrovní faktoru A

$j = 1, \dots, J$ — " — B

$p = 1, \dots, n_{ij}$ je počet pozorování při i-té úrovni faktoru A a j-té — " — B

často $n_{ij} = P$ pro v. i,j (tzv. vyvážené třídění)

střední hodnota $\mu_{ij} = \mu + \alpha_i + \beta_j$

přičemž $\sum_i \alpha_i = 0$, $\sum_j \beta_j = 0$ (tohle jsou tzv. reparametrizační rovnice, které jsou tam kvůli tomu, aby model neměl nadbytečné parametry - neměl by pak plnou hodnost)

Hypotézy: $H_A: \alpha_1 = \dots = \alpha_I = 0$

Za platnosti H_A se model redukuje na model jednoduchého třídění $\mu_{ij} = \mu + \beta_j$, v němž se pak testuje hypotéza

$H_B: \beta_1 = \dots = \beta_J = 0$. Testujeme tedy posloupnost modelů: $\mu_{ij} = \mu + \alpha_i + \beta_j \rightarrow \mu_{ij} = \mu + \beta_j \rightarrow \mu_{ij} = \mu$

Nebo v opačném pořadí:

$$\mu_{ij} = \mu + \alpha_i + \beta_j \rightarrow \mu_{ij} = \mu + \alpha_i \rightarrow \mu_{ij} = \mu$$

Oba způsoby dávají stejné výsledky jen za předpokladu, že $n_{ij} = P$ pro v. i,j.

Tabulka analýzy rozptylu vypadá následovně.

variabilita	součet čtverců S	počet st. volnosti f	podíl S/f	F
řádková	S_A	f_A	S_A/f_A	F_A
sloupcová	S_B	f_B	S_B/f_B	F_B
reziduální	S_e	f_e	S_e/f_e	–
celková	S_T	f_t	–	–

Odvození je zdlouhavé, uvedeme jenom výsledné vzorce.

$$S_A = \frac{1}{JP} \sum_i \left(\sum_j \sum_p Y_{ijp} \right)^2 - \frac{1}{n} \left(\sum_i \sum_j \sum_p Y_{ijp} \right)^2 \qquad f_A = I - 1$$

$$S_B = \frac{1}{IP} \sum_j \left(\sum_i \sum_p Y_{ijp} \right)^2 - \frac{1}{n} \left(\sum_i \sum_j \sum_p Y_{ijp} \right)^2 \qquad f_B = J - 1$$

$$S_T = \sum_i \sum_j \sum_p Y_{ijp}^2 - \frac{1}{n} \left(\sum_i \sum_j \sum_p Y_{ijp} \right)^2 \qquad f_T = n - 1$$

$$S_e = S_T - S_A - S_B \qquad f_e = n - I - J + 1$$

$$F_A = \frac{(n - I - J + 1)\, S_A}{(I-1)\, S_e} \sim F_{I-1,\, n-I-J+1}$$

$$F_B = \frac{(n - I - J + 1)\, S_B}{(J-1)\, S_e} \sim F_{J-1,\, n-I-J+1}$$

Při zamítnutí některé hypotézy se dále testují elementární kontrasty.

Rovnost $\alpha_i = \alpha_t$ se zamítne, když

$$\left| \frac{1}{JP} \sum_j \sum_p Y_{ijp} - \frac{1}{JP} \sum_j \sum_p Y_{tjp} \right| \geq \sqrt{\frac{s^2}{JP}}\; q_{I,\, n-I-J+1}(\alpha)$$

kde $s^2 = \frac{S_e}{f_e}$, q je kritická hodnota stud. rozpětí.

Rovnost $\beta_j = \beta_t$ se zamítne, když

$$\left| \frac{1}{IP} \sum_i \sum_p Y_{ijp} - \frac{1}{IP} \sum_i \sum_p Y_{itp} \right| \geq \sqrt{\frac{s^2}{IP}}\; q_{J,\, n-I-J+1}(\alpha).$$

b) dvojné třídění s interakcemi

Používá se v případech, kdy se dva sledované faktory mohou navzájem ovlivňovat – např. testujeme, zda na výši výnosu nějaké zemědělské plodiny má vliv druh půdy a druh použitého hnojiva, ale musíme počítat i s tím, že některé hnojivo je vhodné jen pro některý typ půdy a pro jiný ne, takže se zde projeví i v kombinaci.

Obecně $\mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}$, kde γ_{ij} jsou <u>interakce</u>.

Přibývá navíc ještě jedna hypotéza $H_{AB}: \gamma_{ij} = 0$ pro vš. i, j.

Dostáváme posloupnost modelů např.

$$\mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij} \rightarrow \mu_{ij} = \mu + \alpha_i + \beta_j \rightarrow \mu_{ij} = \mu + \alpha_i \rightarrow \mu_{ij} = \mu$$

nebo v jiném pořadí.

V případě vyváženého třídění, kde $P \geq 2$, na pořadí nezáleží.

Při výpočtu testových statistik se předpokládá, že

$$\sum_i \alpha_i = 0, \quad \sum_j \beta_j = 0, \quad \sum_i \pi_{ij} = 0 \ \forall j, \quad \sum_j \pi_{ij} = 0 \ \forall i.$$

V tabulce analýzy rozptylu bude:

$$S_A = \frac{1}{JP} \sum_i \Big(\sum_j \sum_r Y_{ijr} \Big)^2 - \frac{1}{n} \Big(\sum_i \sum_j \sum_r Y_{ijr} \Big)^2, \quad f_A = I - 1$$

$$S_B = \frac{1}{IP} \sum_j \Big(\sum_i \sum_r Y_{ijr} \Big)^2 - \frac{1}{n} \Big(\sum_i \sum_j \sum_r Y_{ijr} \Big)^2, \quad f_B = J - 1$$

$$S_e = \sum_i \sum_j \sum_r Y_{ijr}^2 - \frac{1}{P} \sum_i \sum_j \Big(\sum_r Y_{ijr} \Big)^2, \quad f_e = n - IJ$$

$$S_T = \sum_i \sum_j \sum_r Y_{ijr}^2 - \frac{1}{n} \Big(\sum_i \sum_j \sum_r Y_{ijr} \Big)^2, \quad f_T = n - 1$$

$$S_{AB} = S_T - S_A - S_B - S_e, \quad f_{AB} = (I-1)(J-1).$$

$$F_A = \frac{S_A / f_A}{S_e / f_e}, \quad F_B = \frac{S_B / f_B}{S_e / f_e}, \quad F_{AB} = \frac{S_{AB} / f_{AB}}{S_e / f_e}.$$

V případě zamítnutí hypotéz H_A nebo H_B se dále testují elementární kontrasty.

Další zobecnění – <u>trojné třídění</u> atd.

Analýza rozptylu pro závislé výběry

(obdoba párového t-testu)

Máme výběry:

$Y_{11}, \dots, Y_{1r}$ $\qquad r = n_1 = n_2 = \dots = n_I$

$Y_{21}, \dots, Y_{2r}$ $\qquad$ výběry musí být stejně dlouhé

$\vdots$ $\qquad$ předpokládáme normální rozdělení

$Y_{I1}, \dots, Y_{Ir}$

Představujeme si, že na data mají vliv 2 efekty:

klasický pevný efekt (účinek léku, vliv reklamy, použitý algoritmus) – ten nás zajímá, jeho vliv testujeme

náhodný efekt – v tomto případě něco, co způsobuje závislost mezi výběry (příbuzenské vztahy mezi sledovanými pacienty, podobná data při testování algoritmů)

tento efekt chceme nějak odstínit

představujeme si, že je to další náhodná chyba, tj. náhodná veličina s nulovou stř. hodnotou a nenulovým rozptylem

Střední hodnota má potom tvar $\mu + \alpha_i + b_j + e_{ij}$

α_i ... vliv pevného efektu

b_j ... vliv náhodného efektu

$i = 1, \dots, I$ je úroveň pevného efektu (číslo výběru)

$j = 1, \dots, r$ $\quad$ –"– $\quad$ náhodného –"– $\quad$ (číslo jednotky ve výběru)

Testujeme hypotézu $H_0: \alpha_1 = \dots = \alpha_I = 0$

proti alternativě $\alpha_i \neq 0$ pro některé i

18 b

Označíme $S_T = \sum_{i=1}^{I}\sum_{j=1}^{r}(y_{ij} - \bar{y})^2$ kde $\bar{y} = \frac{1}{n}\sum_{i=1}^{I}\sum_{j=1}^{r} y_{ij}$

$$n = \sum_{i=1}^{I} n_i = rI$$

$$S_A = \sum_{i=1}^{I} r(\bar{y}_{i\cdot} - \bar{y})^2 \quad \text{kde} \quad \bar{y}_{i\cdot} = \frac{1}{r}\sum_{j=1}^{r} y_{ij}$$

$$S_B = \sum_{j=1}^{r} I(\bar{y}_{\cdot j} - \bar{y})^2 \quad \text{kde} \quad \bar{y}_{\cdot j} = \frac{1}{I}\sum_{i=1}^{I} y_{ij}$$

$$S_e = S_T - S_A - S_B$$

S_T popisuje celkovou variabilitu

S_A –"– variabilitu mezi výběry

S_B –"– variabilitu způsobenou náhodným efektem

S_e –"– zbytkovou (reziduální variabilitu)

počty stupňů volnosti: $f_T = n-1$

$$f_A = I-1$$

$$f_B = r-1$$

$$f_e = f_T - f_A - f_B$$

testová statistika pro vliv pevného efektu: $F_A = \dfrac{\frac{S_A}{f_A}}{\frac{S_e}{f_e}}$

má rozdělení F s f_A a f_e stupni volnosti

H_0 se zamítá, když $F \geq$ krit. hodnota

Je to vlastně předchozí dvojné třídění s jedním pozorováním v každé podtřídě, tj. $P=1$