Vícefaktorová analýza rozptylu

Mgr. Jiří Mazurek, Ph.D.

Vícefaktorová analýza

- Jde o situaci, kdy se zkoumá, zda kvantitativní znak *Y* je ovlivňován dvěma nebo třemi faktory, opět ne nutně kvantitativními znaky.
- Vícefaktorová analýza rozptylu má svůj experimentální plán. (experimentální plány – později).
- Tento plán může být navržen efektivně tak, aby výsledky analýzy rozptylu byly přesvědčivé a přitom nebylo třeba mít k dispozici příliš mnoho údajů.
- Jak přibývá faktorů, které slouží ke klasifikaci sledovaného znaku *Y*, zvyšuje se tím rychle i požadavek na objem dat.

Dvoufaktorová a vícefaktorová ANOVA

- Techniky testů hypotéz o rozdílech mezi skupinami, kdy rozdíly způsobuje 2 nebo více faktorů
- Příklad otázek, na které odpovídá VF ANOVA:
- (Příklad - Má barva auta, resp. pohlaví respondentů vliv na pravděpodobnost prodeje auta?)

  - Která složka má větší vliv?

  - Je celkový vliv součtem vlivů jednotlivých znaků posuzovaných odděleně?
- Účinky jednotlivých znaků mohou být vzájemně nezávislé (bez interakce) nebo závislé (s interakcí)

Dvojné třídění

- Sleduje-li se vliv dvou faktorů, které mohou ovlivnit hodnotu sledovaného (kvantitativního) znaku *Y*, hovoříme o dvojném třídění.
- Obdobně jako v případě jednoduchého třídění je možné pro různé kombinace těchto dvou faktorů provést náhodné výběry a na jejich základě pak testovat individuální vliv obou faktorů.
- Kromě uvedených dvou faktorů je možno uvažovat jako samostatný faktor také jejich interakci. Podle toho se pak rozlišuje analýza rozptylu dvojné třídění s interakcemi nebo bez interakcí.
- My - bez interakcí.

Trojné třídění

- Analogická tvrzení jako pro dvojné třídění platí také pro případ, kdy pracujeme se třemi „hlavními“ faktory – v tomto případě mluvíme o analýze rozptylu trojné třídění.
- Můžeme zkoumat také jako speciální faktory všechny možné dvoučlenné interakce tří hlavních faktorů a také trojčlennou interakci tvořenou všemi třemi hlavními faktory.

Vyvážené třídění

- Vzhledem k časté náročnosti požadavku na objem dat v případě vícefaktorové analýzy rozptylu se omezujeme na případ, kdy pro danou kombinaci faktorů obsahuje příslušný náhodný výběr **pouze jedno pozorování.**
- Hovoříme pak o analýze rozptylu s jedním pozorováním v každé podskupině. Tento případ také patří mezi případy vyváženého třídění.
- Zatímco u jednofaktorové ANOVA vyvážené třídění není až tak zásadní požadavek, v případech vícefaktorové ANOVA hraje podstatně důležitější roli a doporučujeme jej v praxi dodržovat. Splnění tohoto požadavku obvykle v praxi ani nečiní žádné zvláštní problémy. Pokud tento požadavek splněn není, potom záleží na tom, jak jsou vícefaktorové ANOVA prováděny (mohou být totiž prováděny vícero způsoby) a každý z těchto postupů může dát obecně jiný závěr a mít jinou intepretaci.
- V případě vyváženého třídění toto úskalí nenastává.

Dvojné třídění

- Je-li sledovaný znak ovlivňován dvěma faktory, hovoříme o dvojném třídění.
- I v tomto případě dochází ke vhodnému rozkladu celkové variability znaku na dílčí zdroje variability.
- Rozklad celkového součtu čtverců S se provede analogicky jako v případě jednoduchého třídění s tím rozdílem, že přibyde v rozkladu nový činitel odrážející vliv druhého faktoru.

- Vysvětlení na příkladu

Příklad 1

- 6 řidičů absolvovalo s každým typem benzínu jednu jízdu. Na hladině významnosti 0,05 testujte, je-li průměrná spotřeba paliva (l/100km) závislá na typu použitého benzínu (faktor A) a na řidiči (faktor B).
- **Pozn.: Jako faktor A označujte vždy faktor s úrovněmi v levém sloupci (Benzín) a tedy s řádkovými průměry!**

	Řidič						
Benzín	**1**	**2**	**3**	**4**	**5**	**6**	**Průměr**
Aral	7,5	6,9	7,9	7,3	6,9	7,8	*7,38*
Shell	7,6	7,2	7,5	8	7,3	8,2	*7,63*
Benzina	7,2	8,1	7,8	7,6	7,8	6,9	*7,57*
Slovnaft	7	7,3	7,2	7,5	8,2	7,7	*7,48*
Průměr	*7,33*	*7,38*	*7,6*	*7,6*	*7,55*	*7,65*	*7,5*

Testy hypotéz

- Zkoumáme tedy závislost průměrné spotřeby (znak *Y*) na typu použitého benzínu (faktor A) a na řidiči (faktor B).
- Faktor A nabývá n = 4 úrovní a faktor B nabývá k = 6 úrovní.
- Pro oba faktory testujeme dvě hypotézy:
- Pro faktor A formulujeme hypotézu:
  - H_0: faktor A neúčinkuje
  - H_1: faktor A účinkuje (V tomto případě to značí, že průměrná spotřeba závisí na použitém druhu benzínu)
- Pro faktor B formulujeme hypotézu:
  - H_0: faktor B neúčinkuje
  - H_1: faktor B účinkuje (V tomto případě to značí, že průměrná spotřeba závisí na řidiči, který s vozem jel)

Rozklad variability – součty čtverců

Celkový součet čtverců se rozdělí takto:

$$S = S_A + S_B + S_R$$

Celkový součet čtverců: $S = \sum_{j=1}^{k}\sum_{i=1}^{n}(y_{ij} - \bar{y})^2$

Součet čtverců pro faktor A: $S_A = k\sum_{i=1}\left(\bar{y}_i - \bar{y}\right)^2$

(v sumě jsou řádkové průměry a před sumou počet sloupců)

Součet čtverců pro faktor B:

$$S_B = n\sum_{j=1}^{k}\left(\bar{y}_j - \bar{y}\right)^2$$

(v sumě jsou sloupcové průměry a před sumou počet řádků)

Příklad 1: Součty čtverců

- Celkový součet čtverců:

$$S = \sum_{i=1}^{4}\sum_{j=1}^{6}\left(y_{ij} - \overline{y}\right)^2 = (7{,}5 - 7{,}5)^2 + (6{,}9 - 7{,}5)^2 + \ldots + (8{,}2 - 7{,}5)^2 + (7{,}7 - 7{,}5)^2 = 3{,}79$$

- Součet čtverců pro faktor A:

$$S_A = k\sum_{i=1}^{4}(\overline{y}_{i.} - \overline{y})^2 = 6\left[(7{,}38 - 7{,}5)^2 + \ldots (7{,}48 - 7{,}5)^2\right] = 0{,}21$$

- Součet čtverců pro faktor B:

$$S_B = n\sum_{j=1}^{6}\left(\overline{y}_{.j} - \overline{y}\right)^2 = 4\left[(7{,}33 - 7{,}5)^2 + \ldots (7{,}38 - 7{,}5)^2\right] = 0{,}36$$

- Nakonec spočteme $S_R = S - S_A - S_B$
- S_R= 3,22.

Postup testování – Testové kritérium:

- Testové kritérium pro 1. hypotézu (faktor A):

$$T = \frac{S_A / (n-1)}{S_R / (nk - n - k + 1)}$$

- Testové kritérium pro 2. hypotézu (faktor B):

$$T = \frac{S_B / (k-1)}{S_R / (nk - n - k + 1)}$$

Postup testování: Kritická hodnota a výsledek

- Kritická hodnota:
  - tabulce kritických hodnot Fisherova rozdělení (*F*-rozdělení):

$$\mathrm{K} = F_{n-1,nk-n-k+1}(\alpha)$$

  - Nebo v programu MS Excel: funkce =F.INV.RT.()

- Pokud $T \geq K$, zamítáme nulovou hypotézu. Můžeme tedy v takovém případě říci, že faktor A statisticky významně ovlivňuje sledovaný znak *Y*.
- Je-li naopak $T < K$, přijímáme nulovou hypotézu, jinými slovy, faktor A statisticky významně neovlivňuje sledovaný znak *Y*.

Příklad 1: Test 1

- Testové kritérium pro 1. hypotézu (faktor A):

$$F = \frac{\frac{S_A}{k-1}}{\frac{S_R}{(k-1)(r-1)}} = \frac{\frac{0,21}{3}}{\frac{3,22}{3,5}} = 0,33$$

- V tabulce kritických hodnot *F*-rozdělení najdeme $F_{3,15}(0,05) = 3,29$
- Protože 0,33 < 3,29 , nelze zamítnout H_0, což znamená, že **použitý typ benzínu nemá na průměrnou spotřebu vliv**.

Příklad 1: Test 2

- Testové kritérium pro 1. hypotézu (faktor B):

$$F = \frac{\frac{S_B}{r-1}}{\frac{S_R}{(k-1)(r-1)}} = \frac{\frac{0,36}{5}}{\frac{3,22}{3,5}} = 0,34$$

- V tabulce kritických hodnot *F*-rozdělení najdeme $F_{25,15}(0{,}05) = 2{,}9$
- Protože 0,34 < 2,9 , nelze zamítnout H_0, což znamená, že **volba řidiče nemá na průměrnou spotřebu vliv**.

Příklad 1: Výstup Excel

Anova: dva faktory bez opakování

Faktor	Počet	Součet	Průměr	Rozptyl
Aral	6	44,3	7,383333333	0,185666667
Shell	6	45,8	7,633333333	0,154666667
Benzina	6	45,4	7,566666667	0,194666667
Slovnaft	6	44,9	7,483333333	0,181666667
A	4	29,3	7,325	0,075833333
B	4	29,5	7,375	0,2625
C	4	30,4	7,6	0,1
D	4	30,4	7,6	0,086666667
E	4	30,2	7,55	0,323333333
F	4	30,6	7,65	0,296666667

ANOVA

Zdroj variability	SS	Rozdíl	MS	F	Hodnota P	F krit
Řádky	0,21	3	0,07	0,325581395	0,806868171	3,287382105
Sloupce	0,358333333	5	0,071666667	0,333333333	0,884912733	2,901294536
Chyba	3,225	15	0,215			
Celkem	3,793333333	23				

Příklad 2

- Byla zkoumána kvalita hroznového vína (vyjádřená na stupnici od 1 do 10) v závislosti na dvou faktorech: průměrné době slunečního svitu za den a frekvenci zavlažování. Pro každou kombinaci obou faktorů existuje jedno pozorování, viz tabulka níže. Na hladině významnosti *alfa* = 0,05 rozhodněte o statistické významnosti obou faktorů.
-

Frekvence/svit	4h	5h	6h	7h	8h
denní	3	4	6	8	8
dvoudenní	5	6	6	8	9
2 krát za týden	4	7	7	7	8
1 za týden	2	3	4	4	6

Příklad 2: řešení v Excelu

Anova: dva faktory bez opakování						
Faktor	Počet	Součet	Průměr	Rozptyl		
Řádek 1	5	29	5.8	5.2		
Řádek 2	5	34	6.8	2.7		
Řádek 3	5	33	6.6	2.3		
Řádek 4	5	19	3.8	2.2		
Sloupec 1	4	14	3.5	1.666667		
Sloupec 2	4	20	5	3.333333		
Sloupec 3	4	23	5.75	1.583333		
Sloupec 4	4	27	6.75	3.583333		
Sloupec 5	4	31	7.75	1.583333		
Zdroj variability	SS	Rozdíl	MS	F	Hodnota P	F krit
Řádky	28.15	3	9.383333	**15.85915**	**0.000178**	**3.490295**
Sloupce	42.5	4	10.625	**17.95775**	**5.28E-05**	**3.259167**
Chyba	7.1	12	0.591667			
Celkem	77.75	19				

TROJNÉ TŘÍDĚNÍ (LATINSKÉ ČTVERCE)

- Do analýzy rozptylu patří také speciální případ trojného třídění, tzv. latinské čtverce.
- Latinské čtverce patří mezi klasické metody plánování experimentů (analýza rozptylu rovněž spadá do plánování experimentů).
- Historicky pochází tento pojem z 18. století, kdy L. Euler (1707 – 1783) předložil petrohradské akademii úlohu o 36 důstojnících:
  - Sestavte 36 důstojníků 6 různých hodností ze 6 různých pluků do čtverce tak, aby v každé řadě a v každém sloupci byli důstojníci všech hodností a všech pluků.

Zobecnění problému

- Úkolem je sestavit n^2 objektů do čtverce tak, aby v každé vodorovné řadě i v každé svislé řadě tohoto čtverce byly vždy objekty všech kategorií vlastnosti A a zároveň všech kategorií vlastnosti B (např. v první řadě stojí podporučík z pluku 6, poručík z pluku 5,..., plukovník z pluku 1).
- Takovéto schéma objektů se nazývá latinský čtverec řádu *n*.
- Známý výsledek, který pochází od samotného Eulera, říká, že pro každé přirozené číslo *n* existuje alespoň jeden latinský čtverec řádu *n v* uvedeném slova smyslu.

Tři faktory

- Představme si, že sledujeme vliv tří faktorů na znak *Y*.
- Vzhledem k tomu, že jde o tři faktory, dosti obtížně se nám podaří reprezentovat takový experiment dvojrozměrnou tabulkou.
- My: pro každou kombinaci úrovní sledovaných tří faktorů budeme realizovat jediné pozorování a takový experiment budeme reprezentovat dvojrozměrnou tabulkou, jejíž záhlaví bude obsahovat různé úrovně dvou faktorů a vnitřek tabulky bude obsahovat záznam úrovní třetího faktoru.
- Tyto úrovně třetího faktoru budou přitom vepsány do tabulky tak, aby vznikl latinský čtverec.

Příklad tabulky

- Uvažujeme-li faktory *A*, *B*, *C* a hovoříme o latinském čtverci řádu $n = 3$, můžeme náš experiment zapsat například v podobě tabulky :

a	b	c
b	c	a
c	a	b

- Jedna strana tohoto čtverce představuje tři úrovně faktoru *A*.
- Druhá strana tabulky - sloupce reprezentují tři úrovně faktoru *B*.
- Vnitřek tabulky obsahuje tři úrovně třetího faktoru *C*.
- Návrh takového experimentu čteme tak, že když je faktor A na první úrovni, faktor B je na první a faktor C je rovněž na první úrovni (to je prvek [1,1] tabulky), pak právě pro takovou kombinaci tří faktorů realizujeme jedno pozorování.
- Výhoda: pracujeme s devíti údaji místo 27, přitom je tento návrh zvolen tak, aby výsledná analýza dávala věrohodné výsledky.

Součty čtverců

- Celkový rozklad variability, z něhož se vychází při testování vlivu jednotlivých faktorů, má tvar $S = S_A + S_B + S_C + S_R$.

$$S_A = n\sum_{i=1}^{n}(\overline{y}_{i\bullet\bullet} - \overline{y})^2 \qquad S_B = n\sum_{j=1}^{n}(\overline{y}_{\bullet j\bullet} - \overline{y})^2$$

$$S_C = n\sum_{k=1}^{n}(\overline{y}_{\bullet\bullet k} - \overline{y})^2 \qquad S = \sum_{i}\sum_{j}\sum_{k}(y_{ijk} - \overline{y})^2$$

Test

- U analýzy rozptylu trojné třídění provádíme tři testy. Každý z nich se týká vlivu jednoho ze tří faktorů.
- Testovaná hypotéza H_0: **daný faktor není významný**. Alternativní hypotéza H_1: negace H_0 .
- Testová kritéria pro testování vlivu faktorů *A, B, C* :

Zdroj variability	Součet čtverců	Stupně volnosti	Odhad rozptylu	F testové kritérium
Faktor A	S_A	$df_A=n-1$	$MS_A=S_A / df_A$	$F_A=MS_A / MS_R$
Faktor B	S_B	$df_B=n-1$	$MS_B=S_B / df_B$	$F_B=MS_B / MS_R$
Faktor C	S_C	$df_C=n-1$	$MS_C=S_C / df_C$	$F_C=MS_C / MS_R$
Rezidua	S_R	$df_R=(n-1)(n-2)$	$MS_R=S_R / df_R$	
Celek	S	$df_T=n^2-1$		

Test: výsledek

Je-li $F_A \geq F_{n-1,(n-1)(n-2)}(\alpha)$, zamítáme nulovou hypotézu na hladině významnosti α a tvrdíme, že faktor A je vlivný. Při opačné nerovnosti vlivný není.

Je-li $F_B \geq F_{n-1,(n-1)(n-2)}(\alpha)$, zamítáme nulovou hypotézu na hladině významnosti α a tvrdíme, že faktor B je vlivný. Při opačné nerovnosti vlivný není.

Je-li $F_C \geq F_{n-1,(n-1)(n-2)}(\alpha)$, zamítáme nulovou hypotézu na hladině významnosti α a tvrdíme, že faktor C je vlivný. Při opačné nerovnosti vlivný není.

Příklad

- Uvažujme případ, kdy sledujeme množství emisí výfukových plynů *Y* v závislosti na těchto třech faktorech:
  - Faktor 1 = *typ přísady do benzinu (A, B, C, D)*,
  - Faktor 2 = *řidič vozidla (I, II, III, IV)*,
  - Faktor 3 = *použité vozidlo (1, 2, 3, 4)*.
- Výsledky experimentu:

Řidič\vozidlo:	*1*	*2*	*3*	*4*
I	*A : 21*	*B : 26*	*D : 20*	*C : 25*
II	*D : 23*	*C : 26*	*A : 20*	*B : 27*
III	*B : 15*	*D : 13*	*C : 16*	*A : 16*
IV	*C : 17*	*A : 15*	*B : 20*	*D : 20*

Příklad: příprava

- Testujeme potenciální vliv jednotlivých faktorů na *Y* na pětiprocentní hladině významnosti.

i	$\boxed{\bar{y}_{i\bullet\bullet}}$
1=A	18
2=B	22
3=C	21
4=D	19

j	$\boxed{\bar{y}_{\bullet j\bullet}}$
1=I	23
2=II	24
3=III	15
4=IV	18

k	$\boxed{\bar{y}_{\bullet\bullet k}}$
1	19
2	20
3	19
4	22

Příklad: Součty čtverců

$$S_1 = n\sum_{i=1}^{n}(\overline{y}_{i\bullet\bullet} - \overline{y})^2 = 40.$$

$$S_2 = n\sum_{j=1}^{n}(\overline{y}_{\bullet j\bullet} - \overline{y})^2 = 216.$$

$$S_3 = n\sum_{k=1}^{n}(\overline{y}_{\bullet\bullet k} - \overline{y})^2 = 24.$$

$$S = \sum_{i}\sum_{j}\sum_{k}(y_{ijk} - \overline{y})^2 = 296.$$

$$S_R = S - S_A - S_B - S_C = 16.$$

Příklad: Test

- H_0: **daný faktor není významný**.
- H_1: negace H_0
- Testová kritéria

Faktor 1: $$T = \frac{(40/3)}{(16/6)} = 5.$$

Faktor 2: $$T = \frac{(216/3)}{(16/6)} = 27.$$

Faktor 3: $$T = \frac{(24/3)}{(16/6)} = 3.$$

- Kritická hodnota je ve všech třech případech stejná: K = FINV(0,05,3,6) = 4,757.
- Závěr testů: faktory 1 a 2 jsou statisticky významné, pokud jde o jejich vliv na znak *Y*, třetí faktor, tj. typ použitého vozidla, neovlivňuje znak *Y*.

Děkuji za pozornost.