3.1 Analýza rozptylu pro vícerozměrná data

V této podkapitole si představíme rozšíření jednorozměrné analýzy rozptylu (ANOVA) [http://portal.matematickabiologie.cz/index.php?pg=aplikovana-analyza-klinickych-a-biologickych-dat--biostatistika-pro-matematickou-biologii--analyza-rozptylu-anova] pro vícerozměrná data. Pokud zkoumáme vliv jediného faktoru (kategoriální proměnné) na jednu či více vysvětlovaných spojitých proměnných, mluvíme o analýze rozptylu jednoduchého třídění (neboli jednofaktorové analýze rozptylu). Při větším počtu faktorů se jedná o analýzu rozptylu dvojného, trojného apod. třídění (tedy o vícefaktorovou analýzu rozptylu), přičemž se faktory mohou ovlivňovat (model s interakcí) či se ovlivňovat nemusejí (model bez interakce). V případě, že je vysvětlovaná proměnná pouze jedna, hovoříme o jednorozměrné analýze rozptylu (ANOVA), zatímco při zkoumání vlivu jednoho či více faktorů na více vysvětlovaných proměnných mluvíme o vícerozměrné analýze rozptylu (MANOVA). Pro větší názornost si uveďme několik příkladů různých typů úloh:

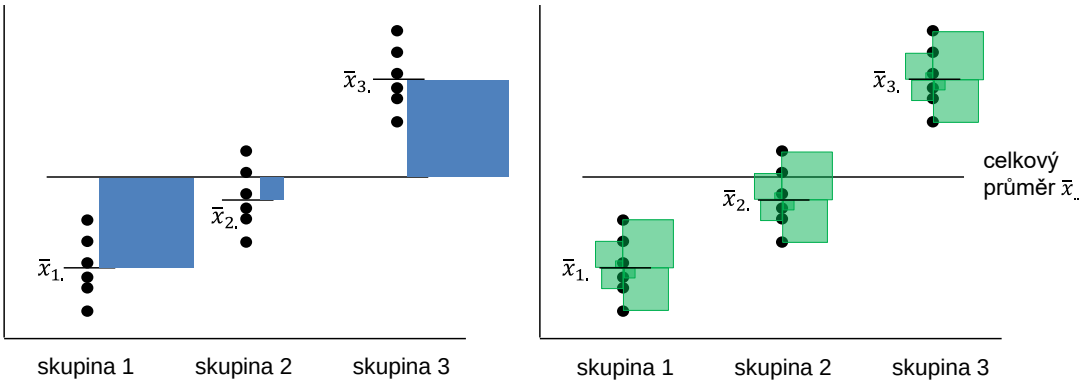
- zkoumáme dlouhodobý vliv třech typů léků na hodnoty systolického tlaku u skupiny osob – jedná se o jednorozměrnou analýzu rozptylu jednoduchého třídění;
- zkoumáme dlouhodobý vliv třech typů léků na hodnoty systolického tlaku u skupiny osob, přičemž chceme zkoumat i vliv pohlaví, předpokládáme však, že ženy i muži reagují na jednotlivé léky obdobně (tzn. např. ženy s léky A a C budou mít nižší tlak než ženy s lékem B a muži s léky A a C budou mít také nižší tlak než muži s lékem B apod.) – jedná se o jednorozměrnou analýzu rozptylu dvojného třídění bez interakce;
- zkoumáme dlouhodobý vliv třech typů léků na hodnoty systolického tlaku u skupiny osob, přičemž chceme zkoumat i vliv pohlaví, a předpokládáme, že ženy a muži budou reagovat na léky různě (tzn. např. ženy s léky A a C budou mít nižší tlak než ženy s lékem B, zatímco muži s léky A a C budou mít vyšší tlak než muži s lékem B apod.) – jedná se o jednorozměrnou analýzu rozptylu dvojného třídění s interakcí;
- zkoumáme dlouhodobý vliv třech typů léků na hodnoty systolického a diastolického tlaku u skupiny osob – jedná se o vícerozměrnou analýzu rozptylu jednoduchého třídění;
- zkoumáme dlouhodobý vliv třech typů léků a vliv pohlaví na hodnoty systolického a diastolického tlaku u skupiny osob – jedná se o vícerozměrnou analýzu rozptylu dvojného třídění.

Začněme nejprve stručným popisem jednorozměrné analýzy rozptylu jednoduchého třídění, kdy srovnáváme tři a více skupin dat, které jsou na sobě nezávislé. Předpokladem je normalita dat ve všech skupinách a shodnost (homogenita) rozptylů všech srovnávaných skupin. Principem je srovnání variability mezi výběry S_A s variabilitou uvnitř výběrů S_e (Obr. 3), které můžeme vypočítat jako

$$S_A = \sum_{i=1}^{a} n_i (\bar{x}_{i.} - \bar{x}_{..})^2, \tag{4}$$

$$S_e = \sum_{i=1}^{a} \sum_{j=1}^{n_i} \left(x_{ij} - \bar{x}_{i.}\right)^2, \tag{5}$$

tedy S_A je součet čtverců rozdílů výběrových průměrů jednotlivých skupin $\bar{x}_{i.}$ od celkového průměru $\bar{x}_{..}$ a S_e je součet čtverců rozdílů pozorovaných hodnot x_{ij} od příslušných skupinových průměrů, přičemž a je počet skupin faktoru A a n_i je počet subjektů v *i*-té skupině.



Obr. 3. Ilustrace výpočtu variability mezi výběry (vlevo) a variability uvnitř výběrů (vpravo).

Můžeme vypočítat také celkový součet čtverců

$$S_T = \sum_{i=1}^{a} \sum_{j=1}^{n_i} \left(x_{ij} - \bar{x}_{..}\right)^2. \tag{6}$$

Výsledky zapíšeme do tzv. tabulky analýzy rozptylu (Tabulka 1), kde *n* je celkový počet subjektů a p_A je výsledná p-hodnota. Pokud $F > F_{1-\alpha}(a-1, n-a)$, zamítáme nulovou hypotézu o shodě středních hodnot jednotlivých skupin subjektů $H_0: \mu_1 = \mu_2 = \cdots = \mu_a$.

Tabulka 1. Tabulka jednorozměrné analýzy rozptylu jednoduchého třídění.

Variabilita	Součet čtverců	Počet stupňů volnosti	Průměrný čtverec	F statistika	p-hodnota
Mezi skupinami	S_A	$df_A = a - 1$	$MS_A = S_A/df_A$	$F = \frac{S_A/df_A}{S_e/df_e}$	p_A
Uvnitř skupin (reziduální var.)	S_e	$df_e = n - a$	$MS_e = S_e/df_e$		
Celková	S_T	$df_T = n - 1$			

Využijeme nyní skutečnosti, že model analýzy rozptylu je speciálním případem obecného lineárního modelu, můžeme tedy jednorozměrnou analýzu rozptylu jednoduchého třídění zapsat jako lineární model následujícím způsobem:

$$X_{ij} = \mu_i + e_{ij} = \mu + \alpha_i + e_{ij}, \tag{7}$$

kde μ je celkový průměr, α_i je *i*-tý efekt faktoru A a e_{ij} je reziduum. Nulovou hypotézu lze pak vyjádřit jako $H_0: \alpha_1 = \alpha_2 = \cdots = \alpha_a$. Rozšířením tohoto modelu můžeme definovat další modely analýzy rozptylu pro více faktorů, hodnocení interakcí, opakovaná měření na jednom subjektu apod.

3.1.1 Jednorozměrná analýza rozptylu dvojného třídění

Jednorozměrná analýza rozptylu dvojného třídění umožňuje srovnání hodnot jedné vysvětlované proměnné podle dvou faktorů (A a B). Předpokladem je normalita dat ve všech $a \cdot b$ skupinách (a je počet skupin faktoru A a b je počet skupin faktoru B) a homogenita rozptylů všech srovnávaných skupin. Model analýzy rozptylu dvojného třídění bez interakcí (tzn. za předpokladu, že se faktory neovlivňují) zapíšeme

$$X_{ij} = \mu + \alpha_i + \beta_j + e_{ij}, \tag{8}$$

kde μ je celkový průměr, α_i je *i*-tý efekt faktoru A, β_j je *j*-tý efekt faktoru B a e_{ij} je reziduum. Nulové hypotézy jsou pak dvě, a to $H_{01}: \alpha_1 = \alpha_2 = \cdots = \alpha_a$ a $H_{02}: \beta_1 = \beta_2 = \cdots = \beta_b$. Výsledky můžeme opět zapsat pomocí tabulky analýzy rozptylu (Tabulka 2), kde součet čtverců pro faktor A (S_A), součet čtverců pro faktor B (S_B), celkový součet čtverců (S_T) a reziduální součet čtverců (S_e) při vyváženém třídění (tedy pro každou skupinu máme stejný počet *c* pozorování) spočítáme jako

$$S_A = bc \sum_{i=1}^{a} (\bar{x}_{i..} - \bar{x}_{...})^2, \tag{9}$$

$$S_B = ac \sum_{j=1}^{b} \left(\bar{x}_{.j.} - \bar{x}_{...}\right)^2, \tag{10}$$

$$S_T = \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{c} \left(x_{ijk} - \bar{x}_{...}\right)^2, \tag{11}$$

$$S_e = S_T - S_A - S_B, \tag{12}$$

kde $\bar{x}_{i..}$ jsou výběrové průměry jednotlivých skupin podle faktoru A, $\bar{x}_{.j.}$ jsou výběrové průměry jednotlivých skupin podle faktoru B, $\bar{x}_{...}$ je celkový průměr a x_{ijk} jsou pozorované hodnoty. Pokud $F_A > F_{1-\alpha}(a-1, n-a-b+1)$, zamítáme nulovou hypotézu o nevýznamnosti faktoru A. Obdobně, pokud $F_B > F_{1-\alpha}(b-1, n-a-b+1)$, zamítáme nulovou hypotézu o nevýznamnosti faktoru B. V případě nevyváženého třídění je situace komplikovanější a vzorce složitější, postupuje se však analogicky.

Tabulka 2. Tabulka jednorozměrné analýzy rozptylu dvojného třídění bez interakce.

Variabilita	Součet čtverců	Počet stupňů volnosti	Průměrný čtverec	F statistika	p-hodnota
Faktor A	S_A	$df_A = a - 1$	$MS_A = S_A/df_A$	$F_A = \frac{S_A/df_A}{S_e/df_e}$	p_A
Faktor B	S_B	$df_B = b - 1$	$MS_B = S_B/df_B$	$F_B = \frac{S_B/df_B}{S_e/df_e}$	p_B
Reziduální	S_e	$df_e = n - a - b + 1$	$MS_e = S_e/df_e$		
Celková	S_T	$df_T = n - 1$			

V případě interakce mezi faktory A a B, tedy pokud se faktory A a B navzájem ovlivňují, mluvíme o analýze rozptylu dvojného třídění s interakcemi, jejíž model lze zapsat

$$X_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk}, \tag{13}$$

kde γ_{ij} odpovídá interakci mezi faktorem A a B. Nulové hypotézy v tomto případě máme tři, a to $H_{01}: \alpha_1 = \alpha_2 = \cdots = \alpha_a$, $H_{02}: \beta_1 = \beta_2 = \cdots = \beta_b$ a $H_{03}: \gamma_{11} = \gamma_{12} = \cdots = \gamma_{ab}$. V tabulce analýzy rozptylu (Tabulka 3) přibude oproti Tabulce 2 další řádek odpovídající interakci. Při vyváženém třídění lze součet čtverců pro faktor A spočítat podle vzorce (9), součet čtverců pro faktor B podle vzorce (10) a celkový součet čtverců S_T podle vzorce (11). Součet čtverců pro interakce vypočteme jako

$$S_{AB} = c \sum_{i=1}^{a} \sum_{j=1}^{b} \left(\bar{x}_{ij.} - \bar{x}_{i..} - \bar{x}_{.j.} + \bar{x}_{...}\right)^2, \tag{14}$$

kde $\bar{x}_{ij.}$ jsou výběrové průměry jednotlivých skupin podle kombinace faktorů A a B. Reziduální součet čtverců (S_e) pak spočítáme pomocí

$$S_e = S_T - S_A - S_B - S_{AB}. \tag{15}$$

Pokud $F_A > F_{1-\alpha}(a-1, n-ab)$, zamítáme nulovou hypotézu o nevýznamnosti faktoru A, a pokud $F_B > F_{1-\alpha}(b-1, n-ab)$, zamítáme nulovou hypotézu o nevýznamnosti faktoru B. V případě, že $F_{AB} > F_{1-\alpha}\big((a-1)\cdot(b-1), n-ab\big)$, zamítáme nulovou hypotézu o nevýznamnosti interakce faktorů A a B.

Tabulka 3. Tabulka jednorozměrné analýzy rozptylu dvojného třídění s interakcí.

Variabilita	Součet čtverců	Počet stupňů volnosti	Průměrný čtverec	F statistika	p-hodnota
Faktor A	S_A	$df_A = a-1$	$MS_A = S_A/df_A$	$F_A = \frac{S_A/df_A}{S_e/df_e}$	p_A
Faktor B	S_B	$df_B = b-1$	$MS_B = S_B/df_B$	$F_B = \frac{S_B/df_B}{S_e/df_e}$	p_B
Interakce AxB	S_{AB}	$df_{AB} = (a-1)\cdot(b-1)$	$MS_{AB} = S_{AB}/df_{AB}$	$F_{AB} = \frac{S_{AB}/df_{AB}}{S_e/df_e}$	p_{AB}
Reziduální	S_e	$df_e = n-ab$	$MS_e = S_e/df_e$		
Celková	S_T	$df_T = n-1$			

V případě analýzy rozptylu trojného či dalších vícenásobných třídění by byl postup analogický, tedy by přibývaly další řádky do tabulky analýzy rozptylu, přičemž výpočet součtů čtverců pro další faktory a jejich interakce bychom počítali obdobným způsobem jako v případě analýzy rozptylu dvojného třídění.

3.1.2 Příklad

Zjistěte, zda má vliv pohlaví a typ léku na počet nežádoucích účinků u pacientů s leukémií, přičemž neuvažujeme interakci mezi oběma faktory. Data jsou zaznamenána v následující tabulce:

ID	Pohlaví	Typ léku	Počet nežádoucích účinků
P1	M	lék X	1
P2	M	lék Y	1
P3	M	lék Z	6
P4	Z	lék X	3
P5	Z	lék Y	4
P6	Z	lék Z	9

Řešení:

Pro větší názornost si data překódujeme tak, že ve druhém sloupečku M=1, Z=2 a ve třetím sloupečku lék X = 1, lék Y = 2 a lék Z = 3. Získáme tedy tabulku

ID	Pohlaví	Typ léku	Počet nežádoucích účinků
P1	1	1	1
P2	1	2	1
P3	1	3	6
P4	2	1	3
P5	2	2	4
P6	2	3	9

Z tabulky vyplývá, že počet kategorií faktoru A (pohlaví) je $a = 2$, počet kategorií faktoru B (typ léku) je $b = 3$, počet pozorování jednotlivých kombinací $c = 1$ a celkový počet pacientů $n = 6$.

Nejprve vypočteme jednotlivé výběrové průměry: $\bar{x}_{1..} = \frac{(1+1+6)}{3} = 8/3$; $\bar{x}_{2..} = \frac{(3+4+9)}{3} = 16/3$; $\bar{x}_{.1.} = \frac{(1+3)}{2} = 2$; $\bar{x}_{.2.} = \frac{(1+4)}{2} = 2{,}5$; $\bar{x}_{.3} = \frac{(6+9)}{2} = 7{,}5$. Dále vypočteme celkový průměr $\bar{x}_{...} = \frac{(1+1+6+3+4+9)}{6} = \frac{24}{6} = 4$.

Součet čtverců pro faktor A (pohlaví) vypočteme jako $S_A = bc\sum_{i=1}^{a}(\bar{x}_{i..} - \bar{x}_{...})^2 = 3 \cdot ((8/3 - 4)^2 + (16/3 - 4)^2) = 32/3 = 10{,}67$ a počet stupňů volnosti je $f_A = a - 1 = 1$.

Součet čtverců pro faktor B (typ léku) vypočteme jako $S_B = ac\sum_{j=1}^{b}\left(\bar{x}_{.j.} - \bar{x}_{...}\right)^2 = 2 \cdot ((2 - 4)^2 + (2{,}5 - 4)^2 + (7{,}5 - 4)^2) = 37$ a počet stupňů volnosti je $f_B = b - 1 = 2$.

Celkový součet čtverců je $S_T = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}\left(x_{ijk} - \bar{x}_{...}\right) = (1 - 4)^2 + (1 - 4)^2 + \cdots + (9 - 4)^2 = 48$ s počtem stupňů volnosti $f_T = n - 1 = 5$.

Reziduální součet čtverců pak spočítáme jako $S_E = S_T - S_A - S_B = 0{,}33$ a počet stupňů volnosti jako $f_E = n - a - b + 1 = 2$.

Výsledky zapíšeme do tabulky jednorozměrné analýzy rozptylu dvojného třídění bez interakcí:

Variabilita	Součet čtverců	Počet stupňů volnosti	Průměrný čtverec	F statistika
Faktor A	$S_A = 10{,}67$	$f_A = 1$	10,67	63,99
Faktor B	$S_B = 37$	$f_B = 2$	18,5	110,98
Reziduální	$S_E = 0{,}33$	$f_E = 2$	0,16	-
Celková	$S_T = 48$	$f_T = 5$	-	-

Protože $F_A = 63{,}99 > F_{0,95}(1,2) = 18{,}1$, zamítáme nulovou hypotézu o nevýznamnosti faktoru A, tedy pohlaví má vliv na počet nežádoucích účinků, přičemž ze vstupní tabulky vidíme, že ženy měly více nežádoucích účinků než muži. Protože $F_B = 110{,}98 > F_{0,95}(2,2) = 19$, zamítáme nulovou hypotézu o nevýznamnosti faktoru B, tedy typ léku má vliv na počet nežádoucích účinků, přičemž ze vstupní tabulky je patrné, že u léku Z bylo nejvíce nežádoucích účinků.

Poznámka: Tento příklad je pouze ilustrativní, v praxi je potřebné, aby u jednotlivých kombinací faktorů A a B bylo mnohem více pacientů než pouze jeden.

3.2 Literatura

1. Everitt, B., Horthorn, T. An Introduction to Applied Multivariate Analysis with R. Springer, New York. (2011)
2. Hebák, P., Hustopecký, J., Jarošová, E., Pecáková, I. Vícerozměrné statistické metody (1). Informatorium, Praha. (2007)
3. Johnson, R.A., Wichern, D.W. Applied Multivariate Statistical Analysis. Prentice Hall, Upper Sadlle River, N.J. (2007)