MA012 Statistika II

2. Dvoufaktorová analýza rozptylu (Two-Way ANOVA)

Ondřej Pokora (pokora@math.muni.cz)

Ústav matematiky a statistiky, Přírodovědecká fakulta, Masarykova univerzita, Brno

(podzim 2015)

Motivační příklad

Příklad 1

Zkoumají se výnosy sena v tunách na hektar v závislosti na typu půdy a na způsobu hnojení. Každá kombinace byla realizována čtyřikrát, nezávisle na sobě.

[t/ha]	způsob hnojení (B):		
typ půdy (A)	bez hnojení	chlévská mrva	vápenaté hnojivo
normální	2,8; 3,2; 3,0; 3,0	3,7; 3,6; 3,9; 3,6	3,4; 3,8; 3,7; 3,6
kyselá	3,1; 2,7; 3,0; 2,9	3,4; 3,4; 3,0; 3,8	4,2; 4,0; 4,1; 3,9

Na hladině významnosti 0,05 testujte hypotézy

- Typ půdy nemá vliv na výnosy.
- Způsob hnojení nemá vliv na výnosy.
- Typ půdy a způsob hnojení jsou nezávislé, tj. neinteragují.

Motivace

V některých reálných situacích se celkový soubor sledovaných náhodných veličin rozpadá na dílčí výběry takovým způsobem, že přihlížíme ke dvěma faktorům (třídicím znakům).

Podobně jako v jednofaktorové analýze rozptylu (One-Way ANOVA), se zajímáme o statistické posouzení toho, zda lze některým z faktorů, či oběma faktory, vysvětlit variabilitu pozorovaných hodnot.

Princip dvoufaktorové analýzy ropztylu (Two-Way ANOVA) je analogický jednofaktorové variantě. Konstruuje se řetězec submodelů a postupně se porovnávají rozptyly, které jednotlivým modelům odpovídají.

Předpoklady

- Uvažujeme dva faktory, A a B.
- Faktor A má $a \geq 2$ úrovní, faktor B má $b \geq 2$ úrovní.
- Pro každou kombinaci úrovní obou faktorů, tzn. pro $(A = i,\ B = j)$, máme n_{ij} výsledků $\left(Y_{ij1}, \ldots, Y_{ijn_{ij}}\right)$, které tvoří náhodný výběr z rozložení $N(\mu_{ij}, \sigma^2)$, $i = 1, \ldots, a$, $j = 1, \ldots b$.
- V označení Y_{ijk} první index označuje skupinu podle úrovně faktoru A, druhý index označuje skupinu podle úrovně faktoru B, třetí index značí pořadí měření v dané skupině.
- Jednotlivé náhodné výběry jsou stochasticky nezávislé, tedy

$$Y_{ijk} = \mu_{ij} + \varepsilon_{ijk},$$

kde ε_{ijk} jsou stochasticky nezávislé náhodné veličiny s rozložením $N(0, \sigma^2)$, kde $i = 1, \ldots, a$, $j = 1, \ldots, b$ a $k = 1, \ldots, n_{ij}$. Rozptyl náhodných chyb σ^2 přitom není známý.

Dvojné třídění

		faktor B		
		1	$\cdots$	b
faktor A	1	$(Y_{111}, \ldots, Y_{11n_{11}})$	$\cdots$	$(Y_{1b1}, \ldots, Y_{1bn_{1b}})$
	$\vdots$	$\vdots$	$(Y_{ij1}, \ldots, Y_{ijn_{ij}})$	$\vdots$
	a	$(Y_{a11}, \ldots, Y_{a1n_{a1}})$	$\cdots$	$(Y_{ab1}, \ldots, Y_{abn_{ab}})$

Příklad 1

[t/ha]	způsob hnojení (B):		
typ půdy (A)	**B**ez hnojení	chlévská **M**rva	**V**ápenaté hnojivo
Normální	2,8; 3,2; 3,0; 3,0	3,7; 3,6; 3,9; 3,6	3,4; 3,8; 3,7; 3,6
Kyselá	3,1; 2,7; 3,0; 2,9	3,4; 3,4; 3,0; 3,8	4,2; 4,0; 4,1; 3,9

Součty a rozsahy výběrů

$$\text{součty ve skupinách: } Y_{ij\cdot} = \sum_{k=1}^{n_{ij}} Y_{ijk}$$

$$\text{součty v řádcích: } Y_{i\cdot\cdot} = \sum_{j=1}^{b} \sum_{k=1}^{n_{ij}} Y_{ijk}, \quad \text{součty ve sloupcích: } Y_{\cdot j\cdot} = \sum_{i=1}^{a} \sum_{k=1}^{n_{ij}} Y_{ijk}$$

$$\text{celkový součet: } Y_{\cdot\cdot\cdot} = \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n_{ij}} Y_{ijk}$$

$$\text{počet měření pro } A = i\text{: } n_{i\cdot} = \sum_{j=1}^{b} n_{ij}, \quad \text{počet měření pro } B = i\text{: } n_{\cdot j} = \sum_{j=1}^{b} n_{ij},$$

$$\text{celkový rozsah souboru: } n = \sum_{i=1}^{a} \sum_{j=1}^{b} n_{ij} = \sum_{i=1}^{a} n_{i\cdot} = \sum_{j=1}^{b} n_{\cdot j}$$

Průměry výběrů

$$\text{průměry ve skupinách: } \overline{Y}_{ij\cdot} = \frac{1}{n_{ij}} \sum_{k=1}^{n_{ij}} Y_{ijk}$$

$$\text{v řádcích: } \overline{Y}_{i\cdot\cdot} = \frac{1}{n_{i\cdot}} \sum_{j=1}^{b} \sum_{k=1}^{n_{ij}} Y_{ijk}, \quad \text{ve sloupcích: } \overline{Y}_{\cdot j\cdot} = \frac{1}{n_{\cdot j}} \sum_{i=1}^{a} \sum_{k=1}^{n_{ij}} Y_{ijk}$$

$$\text{celkový průměr: } \overline{Y}_{\cdot\cdot\cdot} = \frac{1}{n} \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n_{ij}} Y_{ijk}$$

Počty a průměry ve skupinách

		faktor B					
		1	$\cdots$	j	$\cdots$	b	v řádcích
faktor A	1	n_{11}	$\cdots$	n_{1j}	$\cdots$	n_{1b}	$n_{1\cdot}$
	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
	i	n_{i1}	$\cdots$	n_{ij}	$\cdots$	n_{ib}	$n_{i\cdot}$
	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
	a	n_{a1}	$\cdots$	n_{aj}	$\cdots$	n_{ab}	$n_{a\cdot}$
ve sloupcích		$n_{\cdot 1}$	$\cdots$	$n_{\cdot j}$	$\cdots$	$n_{\cdot b}$	n

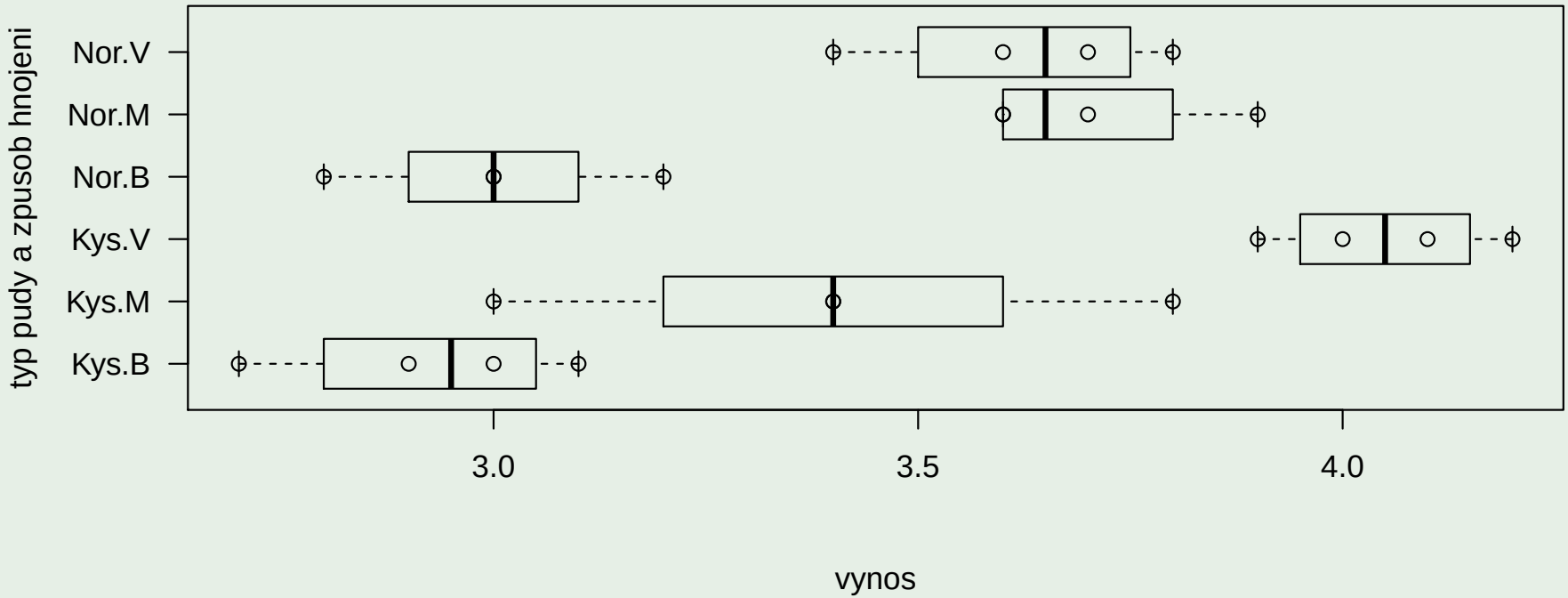
		faktor B					
		1	$\cdots$	j	$\cdots$	b	v řádcích
faktor A	1	$\overline{Y}_{11}$	$\cdots$	$\overline{Y}_{1j}$	$\cdots$	$\overline{Y}_{1b}$	$\overline{Y}_{1\cdot}$
	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
	i	$\overline{Y}_{i1}$	$\cdots$	$\overline{Y}_{ij}$	$\cdots$	$\overline{Y}_{ib}$	$\overline{Y}_{i\cdot}$
	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
	a	$\overline{Y}_{a1}$	$\cdots$	$\overline{Y}_{aj}$	$\cdots$	$\overline{Y}_{ab}$	$\overline{Y}_{a\cdot}$
ve sloupcích		$\overline{Y}_{\cdot 1}$	$\cdots$	$\overline{Y}_{\cdot j}$	$\cdots$	$\overline{Y}_{\cdot b}$	$\overline{Y}_{\cdots}$

Průměry a boxploty

Příklad 1

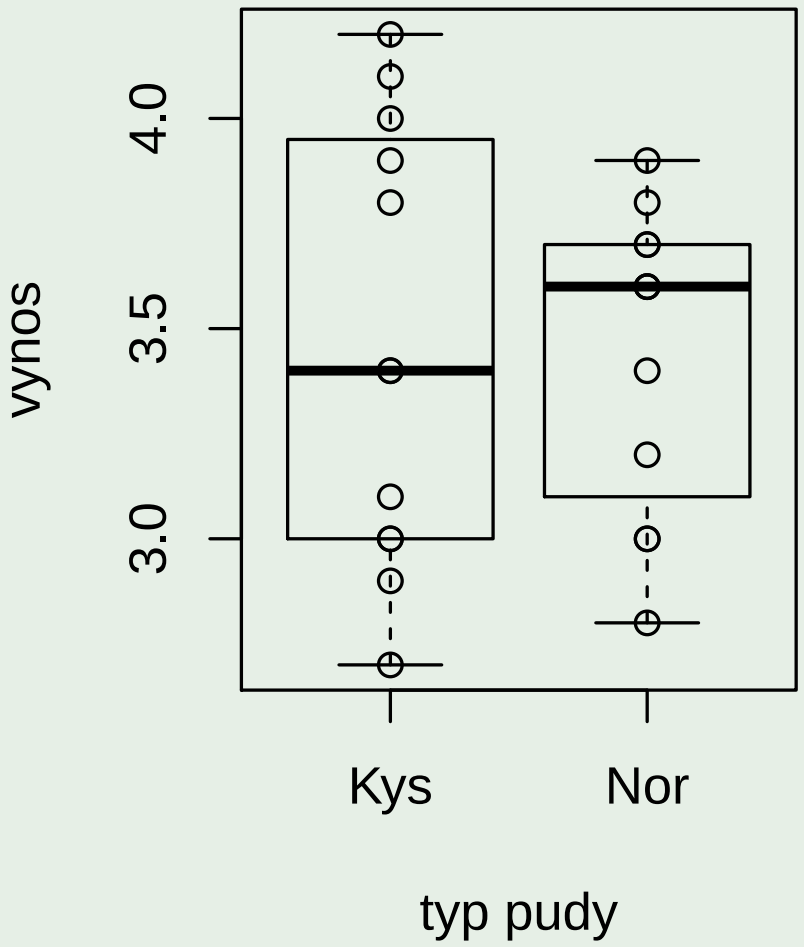
$\overline{Y}_{ij}$	způsob hnojení (B):			průměry
typ půdy (A)	**B**ez hnojení	chlévská **M**rva	**V**ápenaté hnojivo	$\overline{Y}_{i\cdot}$
Normální	3,00	3,70	3,63	3,44
Kyselá	2,93	3,40	4,05	3,46
průměry $\overline{Y}_{\cdot j}$	2,96	3,55	3,84	$\overline{Y}_{\cdot\cdot} = 3,45$

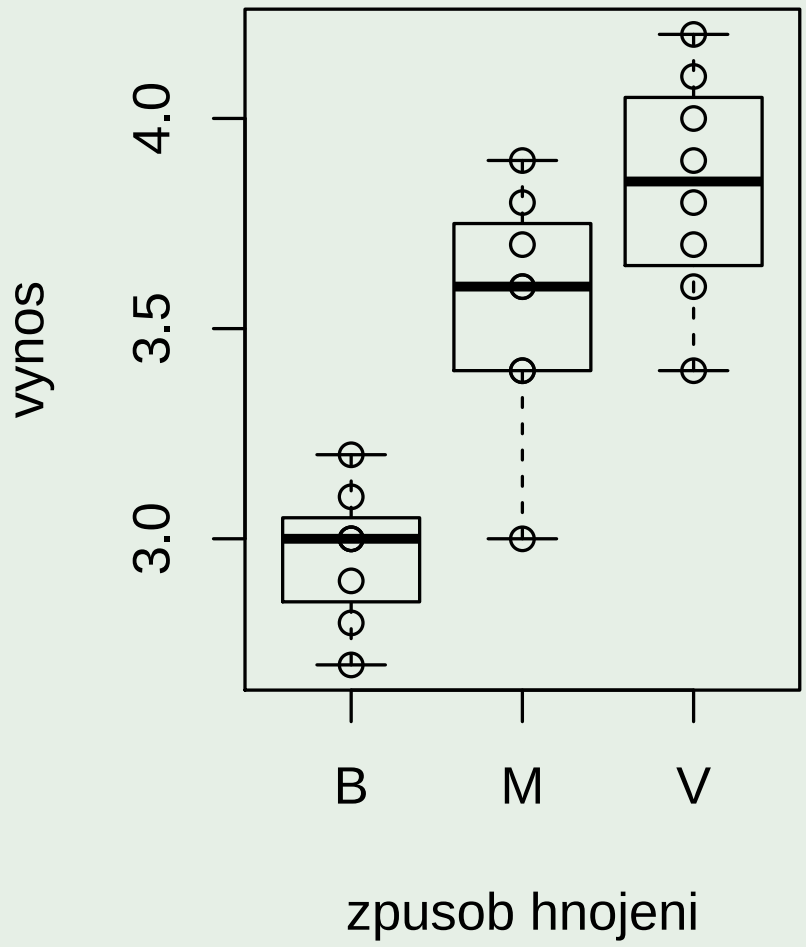
$$n_{ij} = p = 4, \qquad n_{i\cdot} = 12, \qquad n_{\cdot j} = 8, \qquad n = 24$$



Boxploty pro řádky a sloupce

Příklad 1





Hypotézy

Stanovíme hypotézy, které na hladině významnosti α chceme testovat.

H_{A0}: všechny střední hodnoty v řádcích jsou stejné, tzn. faktor A nemá vliv

oproti alternativní hypotéze

H_{A1}: některá(é) dvojice středních hodnot v řádcích se liší, tzn. faktor A má vliv

H_{B0}: všechny střední hodnoty ve sloupcích jsou stejné, tzn. faktor B nemá vliv

oproti alternativní hypotéze

H_{B1}: některá(é) dvojice středních hodnot ve sloupcích se liší, tzn. faktor B má vliv

Základní model M

Definice 1 (model M)

Náhodné veličiny Y_{ijk} se řídí modelem M, pokud

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}$$

pro $i = 1, \ldots, a$, $i = j, \ldots, b$ a $k = 1, \ldots, n_{ij}$, přičemž ε_{ijk} jsou stochasticky nezávislé náhodné veličiny s rozložením $\varepsilon_{ijk} \sim N(0, \sigma^2)$.

Interpretace neznámých parametrů:

- μ je společná část střední hodnoty sledované veličiny
- $\alpha_1, \ldots, \alpha_a$ jsou efekty faktoru A, odchylky od μ způsobené vlivem A
- $\beta_1, \ldots, \beta_b$ jsou efekty faktoru B, odchylky od μ způsobené vlivem B
- σ^2 je rozptyl náhodných chyb

M jako lineární regresní model

$$\underbrace{\begin{pmatrix} Y_{111} \\ \vdots \\ Y_{abn_{ab}} \end{pmatrix}}_{\boldsymbol{Y}} = \underbrace{\left(\begin{array}{cccccccc} \mathbf{1}_{n_{11}} & \mathbf{1}_{n_{11}} & \mathbf{0}_{n_{11}} & \cdots & \mathbf{0}_{n_{11}} & \mathbf{1}_{n_{11}} & \mathbf{0}_{n_{11}} & \cdots & \mathbf{0}_{n_{11}} \\ \mathbf{1}_{n_{12}} & \mathbf{1}_{n_{12}} & \mathbf{0}_{n_{12}} & \cdots & \mathbf{0}_{n_{12}} & \mathbf{0}_{n_{12}} & \mathbf{1}_{n_{12}} & \cdots & \mathbf{0}_{n_{12}} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{1}_{n_{1b}} & \mathbf{1}_{n_{1b}} & \mathbf{0}_{n_{1b}} & \cdots & \mathbf{0}_{n_{1b}} & \mathbf{0}_{n_{1b}} & \mathbf{0}_{n_{1b}} & \cdots & \mathbf{1}_{n_{1b}} \\ \hline \mathbf{1}_{n_{21}} & \mathbf{0}_{n_{21}} & \mathbf{1}_{n_{21}} & \cdots & \mathbf{0}_{n_{21}} & \mathbf{1}_{n_{21}} & \mathbf{0}_{n_{21}} & \cdots & \mathbf{0}_{n_{21}} \\ \mathbf{1}_{n_{22}} & \mathbf{0}_{n_{22}} & \mathbf{1}_{n_{22}} & \cdots & \mathbf{0}_{n_{22}} & \mathbf{0}_{n_{22}} & \mathbf{1}_{n_{22}} & \cdots & \mathbf{0}_{n_{22}} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{1}_{n_{2b}} & \mathbf{0}_{n_{2b}} & \mathbf{1}_{n_{2b}} & \cdots & \mathbf{0}_{n_{2b}} & \mathbf{0}_{n_{2b}} & \mathbf{0}_{n_{2b}} & \cdots & \mathbf{1}_{n_{2b}} \\ \hline \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \hline \mathbf{1}_{n_{a1}} & \mathbf{0}_{n_{a1}} & \mathbf{0}_{n_{a1}} & \cdots & \mathbf{1}_{n_{a1}} & \mathbf{1}_{n_{a1}} & \mathbf{0}_{n_{a1}} & \cdots & \mathbf{0}_{n_{a1}} \\ \mathbf{1}_{n_{a2}} & \mathbf{0}_{n_{a2}} & \mathbf{0}_{n_{a2}} & \cdots & \mathbf{1}_{n_{a2}} & \mathbf{0}_{n_{a2}} & \mathbf{1}_{n_{a2}} & \cdots & \mathbf{0}_{n_{a2}} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{1}_{n_{ab}} & \mathbf{0}_{n_{ab}} & \mathbf{0}_{n_{ab}} & \cdots & \mathbf{1}_{n_{ab}} & \mathbf{0}_{n_{ab}} & \mathbf{0}_{n_{ab}} & \cdots & \mathbf{1}_{n_{ab}} \end{array}\right)}_{X} \underbrace{\left(\begin{array}{c} \mu \\ \hline \alpha_1 \\ \vdots \\ \alpha_a \\ \hline \beta_1 \\ \vdots \\ \beta_b \end{array}\right)}_{\boldsymbol{\theta}}$$

$$\boldsymbol{Y} = X\boldsymbol{\theta}$$

Model M

Jaké rozměry a jakou hodnost má matice plánu $\boldsymbol{X}$?

Model M

Jaké rozměry a jakou hodnost má matice plánu $\boldsymbol{X}$?

$$\boldsymbol{X} : n \times (1 + a + b), \qquad h(\boldsymbol{X}) = a + b - 1$$

Matice plánu tedy není plné hodnosti.

Odhad vektoru parametrů $\boldsymbol{\theta}$ se proto počítá pomocí pseudoinverzní matice

$$\boldsymbol{\theta} = (\boldsymbol{X}'\boldsymbol{X})^{-}\boldsymbol{X}'\boldsymbol{Y},$$

anebo se přidávají tzv. reparametrizační rovnice (proč dvě a co vyjadřují?)

$$\sum_{i=1}^{a} \alpha_i = 0, \qquad \sum_{j=1}^{b} \beta_j = 0.$$

Odhady parametrů v modelu M

Věta 2 (odhady parametrů modelu M)

Odhady parametrů modelu

$$M \ : \quad Y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}$$

metodou nejmenších čtverců jsou rovny

$$\mu = \overline{Y}_{\cdots},$$

$$\alpha_i = \overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdots},$$

$$\beta_j = \overline{Y}_{\cdot j\cdot} - \overline{Y}_{\cdots}.$$

Odhad střední hodnoty měření ve skupině $(A = i,\ B = j)$ *modelu* M *je rovný*

$$\widehat{\mu}_{ijk} = \overline{Y}_{i\cdot\cdot} + \overline{Y}_{\cdot j\cdot} - \overline{Y}_{\cdots}.$$

Tvrzení se dokazuje přímým výpočtem maticové rovnice pro odhad parametrů v lineárním regresním modelu.
Porovnejte s odvozením v přednášce k jednofaktorové analýze rozptylu.

Model M_B

Při zkoumání vlivu faktoru B testujeme v modelu M hypotézu

$$H_{B0}: \ \beta_1 = \cdots = \beta_b = 0,$$
$$H_{B1}: \ \exists\, i \in \{1, \ldots, b\}: \ \beta_i \neq 0.$$

Definice 3 (model M_B)

Náhodné veličiny Y_{ijk} se řídí modelem M_B, pokud

$$Y_{ijk} = \mu + \alpha_i \qquad + \varepsilon_{ijk}$$

pro $i = 1, \ldots, a$, $i = j, \ldots, b$ a $k = 1, \ldots, n_{ij}$, přičemž ε_{ijk} jsou stochasticky nezávislé náhodné veličiny s rozložením $\varepsilon_{ijk} \sim N(0, \sigma^2)$.

Interpretace neznámých parametrů:

- μ je společná část střední hodnoty sledované veličiny
- $\alpha_1, \ldots, \alpha_a$ jsou efekty faktoru A, odchylky od μ způsobené vlivem A
- σ^2 je rozptyl náhodných chyb

M_B jako lineární regresní model

$$\underbrace{\begin{pmatrix} Y_{111} \\ \vdots \\ Y_{abn_{ab}} \end{pmatrix}}_{\boldsymbol{Y}} = \boldsymbol{X}_B \underbrace{\left(\begin{array}{c} \mu \\ \hline \alpha_1 \\ \vdots \\ \alpha_a \end{array}\right)}_{\boldsymbol{\theta}_B} \qquad \text{tj.} \quad \boldsymbol{Y} = \boldsymbol{X}_B \boldsymbol{\theta}_B$$

Jak vypadá matice plánu $\boldsymbol{X}_B$ odpovídající modelu M_B?
Jaké má rozměry a hodnost?

M_B jako lineární regresní model

$$\underbrace{\begin{pmatrix} Y_{111} \\ \vdots \\ Y_{abn_{ab}} \end{pmatrix}}_{\boldsymbol{Y}} = \boldsymbol{X}_B \underbrace{\left(\begin{array}{c} \mu \\ \hline \alpha_1 \\ \vdots \\ \alpha_a \end{array}\right)}_{\boldsymbol{\theta}_B} \qquad \text{tj.} \quad \boldsymbol{Y} = \boldsymbol{X}_B \boldsymbol{\theta}_B$$

Jak vypadá matice plánu $\boldsymbol{X}_B$ odpovídající modelu M_B?
Jaké má rozměry a hodnost?

$\boldsymbol{X}_B$ získáme vynecháním posledních b sloupců z matice $\boldsymbol{X}$.

$$\boldsymbol{X}_B : n \times (1+a), \qquad h(\boldsymbol{X}_B) = a$$

Matice plánu tedy opět není plné hodnosti.

Odhad vektoru parametrů $\boldsymbol{\theta}_B$ se proto počítá pomocí pseudoinverzní matice

$$\boldsymbol{\theta}_B = (\boldsymbol{X}_B' \boldsymbol{X}_B)^- \boldsymbol{X}_B' \boldsymbol{Y},$$

anebo se přidá reparametrizační rovnice $\sum_{i=1}^{a} \alpha_i = 0$.

Odhady parametrů v modelu M_B

Řešením maticové rovnice pro linerání regresní model M_B lze spočítat:

Věta 4 (odhady parametrů modelu M_B)

Odhady parametrů modelu

$$M_B: \quad Y_{ijk} = \mu + \alpha_i \qquad + \varepsilon_{ijk}$$

metodou nejmenších čtverců jsou rovny

$$\mu = \overline{Y}_{\cdot\cdot\cdot},$$

$$\alpha_i = \overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdot\cdot\cdot}.$$

Odhad střední hodnoty měření ve skupině $(A = i,\ B = j)$ *modelu* M_B *je rovný*

$$\widehat{\mu}_{B,ijk} = \overline{Y}_{i\cdot\cdot}.$$

Model M_A

Při zkoumání vlivu faktoru A testujeme v modelu M hypotézu

$$H_{A0}: \ \alpha_1 = \cdots = \alpha_a = 0,$$
$$H_{A1}: \ \exists\, i \in \{1, \ldots, a\}: \ \alpha_i \neq 0.$$

Definice 5 (model M_A)

Náhodné veličiny Y_{ijk} se řídí modelem M_A, pokud

$$Y_{ijk} = \mu \qquad + \beta_j + \varepsilon_{ijk}$$

pro $i = 1, \ldots, a$, $i = j, \ldots, b$ a $k = 1, \ldots, n_{ij}$, přičemž ε_{ijk} jsou stochasticky nezávislé náhodné veličiny s rozložením $\varepsilon_{ijk} \sim N(0, \sigma^2)$.

Interpretace neznámých parametrů:

- μ je společná část střední hodnoty sledované veličiny
- $\beta_1, \ldots, \beta_b$ jsou efekty faktoru B, odchylky od μ způsobené vlivem B
- σ^2 je rozptyl náhodných chyb

M_A jako lineární regresní model, odhady parametrů

Jak vypadá matice plánu $\boldsymbol{X}_A$ pro model M_A? Jaké má rozměry a hodnost?

M_A jako lineární regresní model, odhady parametrů

Jak vypadá matice plánu $\boldsymbol{X}_A$ pro model M_A? Jaké má rozměry a hodnost?

$\boldsymbol{X}_A$ získáme vynecháním 2. až $(a+1)$-ního sloupce z matice $\boldsymbol{X}$.

$$\boldsymbol{X}_A : n \times (1+b), \qquad h(\boldsymbol{X}_A) = b$$

Věta 6 (odhady parametrů modelu M_A)

Odhady parametrů modelu

$$M_A : \quad Y_{ijk} = \mu \qquad + \beta_j + \varepsilon_{ijk}$$

metodou nejmenších čtverců jsou rovny

$$\mu = \overline{Y}_{\cdot\cdot\cdot},$$

$$\beta_j = \overline{Y}_{\cdot j \cdot} - \overline{Y}_{\cdot\cdot\cdot}.$$

Odhad střední hodnoty měření ve skupině $(A = i,\ B = j)$ modelu M_A je rovný

$$\widehat{\mu}_{A,ijk} = \overline{Y}_{\cdot j \cdot}.$$

Minimální model M_0

Při zkoumání vlivu obou faktorů vyjdeme testujeme v modelu M_B hypotézu

$$H_{A0}: \ \alpha_1 = \cdots = \alpha_a = 0,$$
$$H_{A1}: \ \exists\, i \in \{1, \ldots, a\}: \ \alpha_i \neq 0.$$

Definice 7 (model M_0)

Náhodné veličiny Y_{ijk} se řídí modelem M_0, pokud

$$Y_{ijk} = \mu \qquad\qquad + \varepsilon_{ijk}$$

pro $i = 1, \ldots, a$, $i = j, \ldots, b$ a $k = 1, \ldots, n_{ij}$, přičemž ε_{ijk} jsou stochasticky nezávislé náhodné veličiny s rozložením $\varepsilon_{ijk} \sim N(0, \sigma^2)$.

Interpretace neznámých parametrů:

- μ je střední hodnota sledované veličiny bez ohledu na kategorizaci dle faktorů
- σ^2 je rozptyl náhodných chyb

M_0 jako lineární regresní model, odhad μ

$$\boldsymbol{Y} = \boldsymbol{X}_0 \, \mu$$

Jak vypadá matice plánu $\boldsymbol{X}_0$ pro model M_0? Jaké má rozměry a hodnost?

M_0 jako lineární regresní model, odhad μ

$$\boldsymbol{Y} = \boldsymbol{X}_0\,\mu$$

Jak vypadá matice plánu $\boldsymbol{X}_0$ pro model M_0? Jaké má rozměry a hodnost?

$$\boldsymbol{X}_0 = \mathbf{1}_n, \qquad \boldsymbol{X}_A : n \times 1, \qquad h(\boldsymbol{X}_A) = 1$$

Matice je v tomto modelu plné hodnosti a odhad parametru μ počítáme klasicky:

Věta 8 (odhady parametrů modelu M_0)

V modelu

$$M_0: \quad Y_{ijk} = \mu \qquad\qquad + \varepsilon_{ijk}$$

je odhad parametru μ metodou nejmenších čtverců rovný

$$\widehat{\mu}_{A,ijk} = (\boldsymbol{X}_0'\boldsymbol{X}_0)^{-1}\boldsymbol{X}_0'\boldsymbol{Y} = \overline{Y}_{\cdots}.$$

Shrnutí: modely a submodely

$$
\begin{aligned}
M &: \quad Y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk} \\
M_B &: \quad Y_{ijk} = \mu + \alpha_i + \varepsilon_{ijk} \\
M_A &: \quad Y_{ijk} = \mu + \beta_j + \varepsilon_{ijk} \\
M_0 &: \quad Y_{ijk} = \mu + \varepsilon_{ijk}
\end{aligned}
$$

Jaké řetězce submodelů lze vytvořit?

Shrnutí: modely a submodely

$$
\begin{aligned}
M &: \quad Y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk} \\
M_B &: \quad Y_{ijk} = \mu + \alpha_i + \varepsilon_{ijk} \\
M_A &: \quad Y_{ijk} = \mu + \beta_j + \varepsilon_{ijk} \\
M_0 &: \quad Y_{ijk} = \mu + \varepsilon_{ijk}
\end{aligned}
$$

Jaké řetězce submodelů lze vytvořit?

$$M \xrightarrow{H_{B0}} M_B \xrightarrow{H_{A0}} M_0, \qquad M \xrightarrow{H_{A0}} M_A \xrightarrow{H_{B0}} M_0.$$

Pro postupné testování hypotéz v analýze rozptylu se volí vždy jeden z řetězců submodelů, v praxi obvykle $M \longrightarrow M_B \longrightarrow M_0$.

Praxe

V praxi volí většina statistických softwarů (mj. i *R*) odlišnou stavbu modelů s

$$\alpha_1 = 0, \qquad \beta_1 = 0,$$

kdy se neuvažují reparametrizační rovnice. Tato odlišná parametrizace nemá vliv na vlastní výpočet, pouze parametry modelů mají odlišnou interpretaci.

První úrovně faktorů A a B jsou tedy stanoveny jako referenční. Parametr μ zde interpretujeme jako střední hodnotu kategorie $(A = 1,\ B = 1)$, a efekty $\alpha_2, \ldots, \alpha_a$, resp. $\beta_2, \ldots, \beta_b$, vyjadřují odchylky vlivem faktoru A, resp. B, od této střední hodnoty.

Definice 9 (vyvážené třídění)

O vyváženém třídění hovoříme, pokud je počet ve všech kategoriích stejný,

$$n_{ij} = p, \qquad i = 1, \ldots, a,\ j = 1, \ldots, b.$$

Při vyváženém třídění dostáváme $n = a\,b\,p$, $n_{i\cdot} = b\,p$, $n_{\cdot j} = a\,p$.

Součty čtverců

Skupinový součet čtverců S_B charakterizuje variabilitu mezi jednotlivými náhodnými výběry ve skupinách faktoru B, Jde tedy o součet čtverců rozdílů odhadů mezi modely M a M_B:

$$S_B = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}}\left(\widehat{\mu}_{ijk} - \widehat{\mu}_{B,ijk}\right)^2 = \sum_{i=1}^{a}\sum_{j=1}^{b} n_{ij}\left(Y_{\cdot j\cdot} - \overline{Y}_{\cdots}\right)^2 = a\,p\sum_{j=1}^{b} Y_{\cdot j\cdot}^2 - n\,\overline{Y}_{\cdots}$$

Podobně S_A charakterizuje variabilitu mezi jednotlivými náhodnými výběry ve skupinách faktoru A. Musíme však dodržet pořadí submodelů v řetězci, počítáme proto součet čtverců rozdílů odhadů mezi modely M_B a M_0:

$$S_A = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}}\left(\widehat{\mu}_{B,ijk} - \widehat{\mu}_{0,ijk}\right)^2 = \sum_{i=1}^{a}\sum_{j=1}^{b} n_{ij}\left(Y_{i\cdot\cdot} - \overline{Y}_{\cdots}\right)^2 = b\,p\sum_{j=1}^{b} Y_{i\cdot\cdot}^2 - n\,\overline{Y}_{\cdots}$$

Celkový součet čtverců charakterizuje variabilitu jednotlivých pozorování kolem celkového průměru, bez ohledu na faktory. Jde tedy o součet čtverců rozdílů odhadů mezi pozorovanými veličinami a modelem M_0:

$$S_T = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}}\left(Y_{ijk} - \widehat{\mu}_{0,ijk}\right)^2 = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}}\left(Y_{ijk} - \overline{Y}_{\cdots}\right)^2 = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}} Y_{ijk}^2 - n\,\overline{Y}_{\cdots}$$

Součty čtverců

Reziduální součet čtverců charakterizuje varibilitu nevysvětlenou modelem M:

$$S_e = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}} \left(Y_{ijk} - \widehat{\mu}_{ijk}\right)^2 = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}} \left(Y_{ijk} - \overline{Y}_{i\bullet\bullet} - \overline{Y}_{\bullet j\bullet} + \overline{Y}_{\bullet\bullet\bullet}\right)^2$$

Věta 10

Platí

$$S_e = S_T - S_A - S_B$$

Analýza rozptylu dvojného třídění je založena na porovnání podílů roztpylů S_B/S_e a S_A/S_e vážených odpovídajícími stupni volnosti.

ANOVA tabulka dvojného třídění bez interakcí

Předcházející pojmy se shrnují v tzv. **tabulce analýzy rozptylu dvojného třídění**:

Zdroj variability	*Součet čtverců* SS	*Stupně volnosti* df	*Podíl* $MS = \frac{SS}{df}$	$F = \frac{MS}{s^2}$	*p-hodnota*
řádky (A)	S_A	$df_A = a - 1$	$MS_A = \frac{S_A}{df_A}$	$F_A = \frac{MS_A}{MS_e}$	$P(F \geq F_A)$
sloupce (B)	S_B	$df_B = b - 1$	$MS_B = \frac{S_B}{df_B}$	$F_B = \frac{MS_B}{MS_e}$	$P(F \geq F_B)$
reziduální	S_e	$df_e = n - a - b + 1$	$MS_e = \frac{S_e}{df_e}$	–	–
celkový	S_T	$df_T = n - 1$	–	–	–

Testovací statistiky

Věta 11

Rozdíl mezi modely M a M_B ověřujeme pomocí testové statistiky

$$F_B = \frac{S_B / df_B}{S_e / df_e},$$

která má za platnosti H_{B0} rozdělení pravděpodobnosti

$$F_B \sim F(df_B, df_e) = F(b-1, n-a-b+1).$$

Rozdíl mezi modely M_B a M_0 ověřujeme pomocí testové statistiky

$$F_A = \frac{S_A / df_A}{S_e / df_e},$$

která má za platnosti H_{A0} rozdělení pravděpodobnosti

$$F_A \sim F(df_A, df_e) = F(a-1, n-a-b+1).$$

Výsledek testování

Věta 12

Porovnáváme modely M a M_B. Pokud

$$F_B = \frac{S_B/df_B}{S_e/df_e} \geq F_{1-\alpha}(b-1, n-a-b+1),$$

zamítneme na hladině významnosti α nulovou hypotézu H_{B0}, tzn. statisticky prokážeme vliv faktoru B (sloupce).
Pokračujeme porovnáním modelů M_B a M_0. Pokud

$$F_A = \frac{S_A/df_A}{S_e/df_e} \geq F_{1-\alpha}(a-1, n-a-b+1),$$

zamítneme navíc na hladině významnosti α nulovou hypotézu H_{A0}, tzn. statisticky prokážeme také vliv faktoru A (řádky).

Postup testování ve dvojném třídění

1. $M \longrightarrow M_B$

Pomocí F_B testujeme rozdíly mezi sloupci (faktor B) a přitom přihlížíme k eventuálním řádkovým efektům.

2. $M_B \longrightarrow M_0$

Pomocí F_A testujeme rozdíly mezi řádky (faktor A), **nebereme** však v úvahu případný vliv sloupcových efektů.

Postup testování ve dvojném třídění

1. $M \longrightarrow M_B$

Pomocí F_B testujeme rozdíly mezi sloupci (faktor B) a přitom přihlížíme k eventuálním řádkovým efektům.

2. $M_B \longrightarrow M_0$

Pomocí F_A testujeme rozdíly mezi řádky (faktor A), **nebereme** však v úvahu případný vliv sloupcových efektů.

jiný postup: $M \longrightarrow M_A \longrightarrow M_0$

Analýzu rozptylu můžeme ale provést také v řetězci $M \longrightarrow M_A \longrightarrow M_0$, čímž dokážeme testovat rozdíly mezi řádky (faktor A) při přihlédnutí k eventuálním sloupcovým efektům.

V případě vyváženého designu vyjdou analýzy obou řetězců stejně. Pro nevyvážený design mohou analýzy obou řetězců vyjít odlišné.

Mnohonásobné porovnávání

Zjistíme-li významný rozdíl mezi řádky (faktor A), zjišťujeme dále, které řádky (úrovně faktoru A) se od sebe signifikantně liší. K ověření máme opět Scheffého metodu a Tukeyovu metodu, z nichž v praxi vybíráme tu, která je citlivější.

Využíváme přitom následující tvrzení:

Věta 13

Za platnosti M jsou veličiny $(\overline{Y}_{1\bullet\bullet}, \ldots, \overline{Y}_{a\bullet\bullet})$ stochasticky nezávislé a platí

$$\overline{Y}_{i\bullet\bullet} \sim N\left(\mu + \alpha_i + \frac{1}{b}\sum_{i=1}^{b}\beta_j,\ \frac{\sigma^2}{bp}\right), \qquad i = 1, \ldots, a.$$

Zjistíme-li významný rozdíl mezi sloupci (faktor B), analogicky zjišťujeme, které sloupce (úrovně faktoru B) se od sebe signifikantně liší.

Porovnání řádků (faktor A)

Porovnávání řádků $(A = u)$ a $(A = v)$ odpovídá porovnávání efektů α_u a α_v.

Věta 14 (Scheffého metoda)

Hypotézu o rovnosti $\alpha_u = \alpha_v$ zamítáme, pokud

$$|\overline{Y}_{u\bullet\bullet} - \overline{Y}_{v\bullet\bullet}| > \sqrt{\frac{2(a-1)S_e}{b\,p\,(n-a-b+1)}\,F_{1-\alpha}(a-1, n-a-b+1)}.$$

Věta 15 (Tukeyova metoda)

Hypotézu o rovnosti $\alpha_u = \alpha_v$ zamítáme, pokud

$$|\overline{Y}_{u\bullet\bullet} - \overline{Y}_{v\bullet\bullet}| > \sqrt{\frac{S_e}{b\,p(n-a-b+1)}}\,q_{1-\alpha}(a, n-a-b+1).$$

Porovnání sloupců (faktor B)

Porovnávání řádků $(B = u)$ a $(B = v)$ odpovídá porovnávání efektů β_u a β_v.

Věta 16 (Scheffého metoda)

Hypotézu o rovnosti $\beta_u = \beta_v$ zamítáme, pokud

$$|\overline{Y}_{.u.} - \overline{Y}_{.v.}| > \sqrt{\frac{2(b-1)S_e}{a\,p\,(n-a-b+1)}\,F_{1-\alpha}(b-1, n-a-b+1)}.$$

Věta 17 (Tukeyova metoda)

Hypotézu o rovnosti $\beta_u = \beta_v$ zamítáme, pokud

$$|\overline{Y}_{.u.} - \overline{Y}_{.v.}| > \sqrt{\frac{S_e}{a\,p(n-a-b+1)}\,q_{1-\alpha}(b, n-a-b+1)}.$$

Dvojné třídění s interakcemi (vyvážené)

U dvojného třídění se často stává, že se řádkové a sloupcové efekty jen prostě nesčítají, jak to předpokládá náš model M.
V takových situacích uvažujeme následující komplexnější model.

Definice 18 (model M_*)

Náhodné veličiny Y_{ijk} se řídí modelem M_*, pokud

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \lambda_{ij} + \varepsilon_{ijk}$$

pro $i = 1, \ldots, a$, $i = j, \ldots, b$ a $k = 1, \ldots, p$, přičemž ε_{ijk} jsou stochasticky nezávislé náhodné veličiny s rozložením $\varepsilon_{ijk} \sim N(0, \sigma^2)$.

Interpretace neznámých parametrů:

- μ je společná část střední hodnoty sledované veličiny
- $\alpha_1, \ldots, \alpha_a$ jsou efekty faktoru A,
- $\beta_1, \ldots, \beta_b$ jsou efekty faktoru B,
- λ_{ij} $(i = 1, \ldots, a,\ j = 1, \ldots, b)$ jsou tzv. **interakce**,
- σ^2 je rozptyl náhodných chyb

Model M_*

Zapíšeme M_* jako lineární regresní model:

$$\boldsymbol{Y} = \boldsymbol{X}_* \underbrace{(\mu, \alpha_1, \ldots, \alpha_a, \beta_1, \ldots, \beta_b, \lambda_{11}, \ldots, \lambda_{ab})'}_{\text{vektor parametrů } \boldsymbol{\theta}_*}$$

Jaké rozměry a jakou hodnost má matice plánu $\boldsymbol{X}_*$?

Model M_*

Zapíšeme M_* jako lineární regresní model:

$$\boldsymbol{Y} = \boldsymbol{X}_* \underbrace{(\mu, \alpha_1, \ldots, \alpha_a, \beta_1, \ldots, \beta_b, \lambda_{11}, \ldots, \lambda_{ab})'}_{\text{vektor parametrů } \boldsymbol{\theta}_*}$$

Jaké rozměry a jakou hodnost má matice plánu $\boldsymbol{X}_*$?

$$\boldsymbol{X}_* : n \times (1 + a + b + ab), \qquad h(\boldsymbol{X}_*) = ab$$

Matice plánu není plné hodnosti.
Model se obvykle řeší přidáním reparametrizačních rovnic

$$\sum_{i=1}^{a} \alpha_i = 0, \quad \sum_{j=1}^{b} \beta_j = 0, \quad \sum_{i=1}^{a} \lambda_{ij} = 0, \quad \sum_{j=1}^{b} \lambda_{ij} = 0.$$

Odhady parametrů v modelu M_*

Věta 19 (odhady parametrů modelu M_*)

Odhady parametrů modelu

$$M_*: \quad \mu + \alpha_i + \beta_j + \lambda_{ij} + \varepsilon_{ijk}$$

metodou nejmenších čtverců jsou rovny

$$\mu = \overline{Y}_{\cdots},$$

$$\alpha_i = \overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdots},$$

$$\beta_j = \overline{Y}_{\cdot j\cdot} - \overline{Y}_{\cdots},$$

$$\lambda_{ij} = \overline{Y}_{ij\cdot} - \overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdot j\cdot} + \overline{Y}_{\cdots}.$$

Odhad střední hodnoty měření ve skupině $(A = i,\ B = j)$ modelu M_* je rovný

$$\widehat{\mu}_{*ijk} = \overline{Y}_{ij\cdot}.$$

Porovnejte s odhady pro model M.

Součty čtverců

Věta 20

Při vyváženém dvojném třídění s interakcemi platí:

$$S_B = a\,p \sum_{j=1}^{b} Y_{\cdot j \cdot}^2 - n\,\overline{Y}_{\cdots},$$

$$S_A = b\,p \sum_{j=1}^{b} Y_{i \cdot \cdot}^2 - n\,\overline{Y}_{\cdots},$$

$$S_T = \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n_{ij}} Y_{ijk}^2 - n\,\overline{Y}_{\cdots},$$

$$S_e = \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n_{ij}} Y_{ijk}^2 - p \sum_{i=1}^{a} \sum_{j=1}^{b} Y_{ij\cdot}^2,$$

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

Porovnejte se součty čtverců pro model M.

ANOVA tabulka dvojného třídění s interakcemi

Tabulka analýzy rozptylu dvojného třídění s interakcemi:

Zdroj variability	*Součet čtverců* SS	*Stupně volnosti* df	*Podíl* $MS = \frac{SS}{df}$	$F = \frac{MS}{s^2}$	*p-hodnota*
řádky (A)	S_A	$df_A = a - 1$	$MS_A = \frac{S_A}{df_A}$	$F_A = \frac{MS_A}{MS_e}$	$P(F \geq F_A)$
sloupce (B)	S_B	$df_B = b - 1$	$MS_B = \frac{S_B}{df_B}$	$F_B = \frac{MS_B}{MS_e}$	$P(F \geq F_B)$
interakce	S_{AB}	$df_{AB} = (a-1)(b-1)$	$MS_{AB} = \frac{S_{AB}}{df_{AB}}$	$F_{AB} = \frac{MS_{AB}}{MS_e}$	$P(F \geq F_{AB})$
reziduální	S_e	$df_e = n - a\,b$	$MS_e = \frac{S_e}{df_e}$	–	–
celkový	S_T	$df_T = n - 1$	–	–	–

Testování v modelu s interakcemi

Testování ve dvojném třídění s interakcemi probíhá obvykle v řetězci submodelů

$$M_* \longrightarrow M \longrightarrow M_B \longrightarrow M_0.$$

Při vyváženém třídění lze řetězec submodelů libovolně měnit bez změny výsledku.

Věta 21

Porovnáváme modely M_* a M. Pokud

$$F_{AB} = \frac{S_{AB}/df_{AB}}{S_e/df_e} \geq F_{1-\alpha}\left((a-1)(b-1), n-ab\right),$$

zamítneme na hladině významnosti α nulovou hypotézu

$$H_{*0} : \lambda_{ij} = 0, \qquad i = 1,\ldots,a,\ j = 1,\ldots,b,$$

tzn. statisticky prokážeme vliv interakcí řádků a sloupců.

Porovnávání dvojic řádků a dvojic sloupců se dále provádí podle Vět 14–17, v nichž se upraví počet stupňů volnosti df_e tak, že $(n-a-b+1)$ nahradíme $(n-ab)$.

Shrnutí

- k čemu se používá jednoduché a dvojné třídění (s interakcemi)
- testy ověření podmínek – normalita dat, homogenita rozptylů
- lineární regresní modely, význam parametrů, řetězec submodelů
- hypotézy a jejich souvislost s testováním submodelů
- testovací statistiky v analýze rozptylu založené na součtech čtverců
- interpretace výsledků v ANOVA tabulce
- výpočet efektů a interakcí a jejich interpretace
- výpočet středních hodnot ve skupinách
- metody mnohonásobného porovnávání

ANOVA v R

Funkce pro analýzu rozptylu:

```
model <- aov (formule, data)
```

Zápis `formule`:

```
# jednoduche trideni
Y ~ faktorA
# dvojne trideni bez interakci
Y ~ faktorA + faktorB
# dvojne trideni s interakcemi
Y ~ faktorA + faktorB + faktorA:faktorB
# dvojne trideni s interakcemi
Y ~ faktorA * faktorB
```

Porovnání regresních modelů:

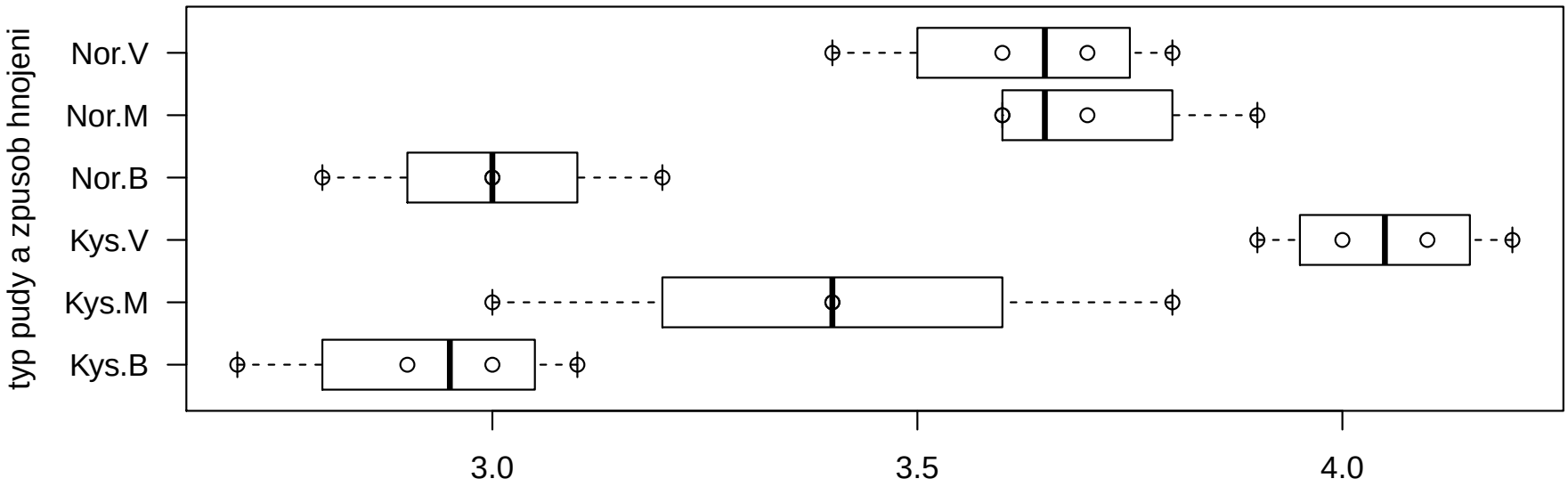
```
model1 <- lm (formule1, data)
model2 <- lm (formule2, data)
model <- anova (model1, model2)
```

Příklad

Příklad 1

Zkoumají se výnosy sena v tunách na hektar v závislosti na typu půdy a na způsobu hnojení. Každá kombinace byla realizována čtyřikrát, nezávisle na sobě.

[t/ha]	*způsob hnojení (B):*		
typ půdy (A)	**B**ez hnojení	chlévská **M**rva	**V**ápenaté hnojivo
Normální	2,8; 3,2; 3,0; 3,0	3,7; 3,6; 3,9; 3,6	3,4; 3,8; 3,7; 3,6
Kyselá	3,1; 2,7; 3,0; 2,9	3,4; 3,4; 3,0; 3,8	4,2; 4,0; 4,1; 3,9



ANOVA tabulka dvojného třídění

```
             Df  Sum Sq  Mean Sq  F value    Pr(>F)
puda          1   0.002   0.0017    0.027     0.871
hnojeni       2   3.182   1.5912   25.752 2.93e-06 **
Residuals    20   1.236   0.0618
```

```
Tables of effects
 puda
puda
      Kys        Nor
 0.008333 -0.008333
 hnojeni
hnojeni
      B        M        V
-0.4875   0.1000   0.3875
```

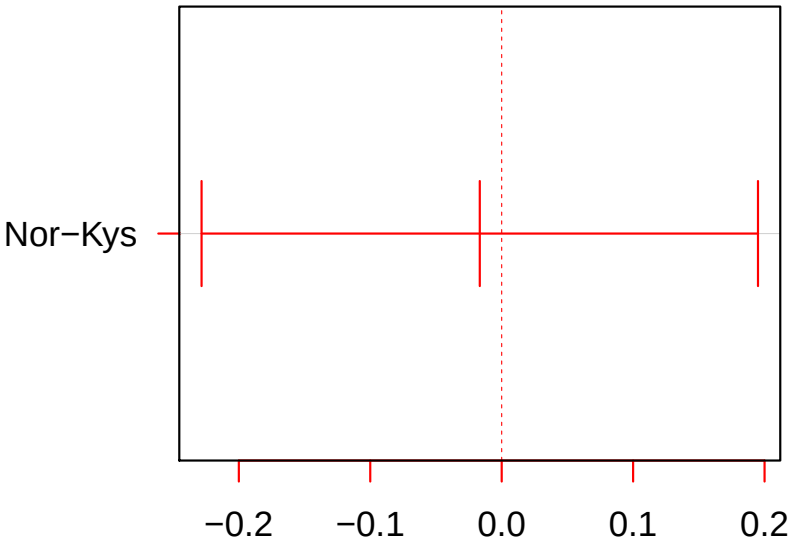
```
Tables of means
Grand mean
      3.45
 puda
puda
  Kys   Nor
3.458 3.442
 hnojeni
hnojeni
    B     M     V
2.963 3.550 3.838
```

```
(Intercept)      pudaNor     hnojeniM     hnojeniV
 2.97083333 -0.01666667   0.58750000   0.87500000
```

Scheffé (typ půdy)

```
  trt     means M
1 Kys 3.458333 a
2 Nor 3.441667 a
```

95% family−wise confidence level

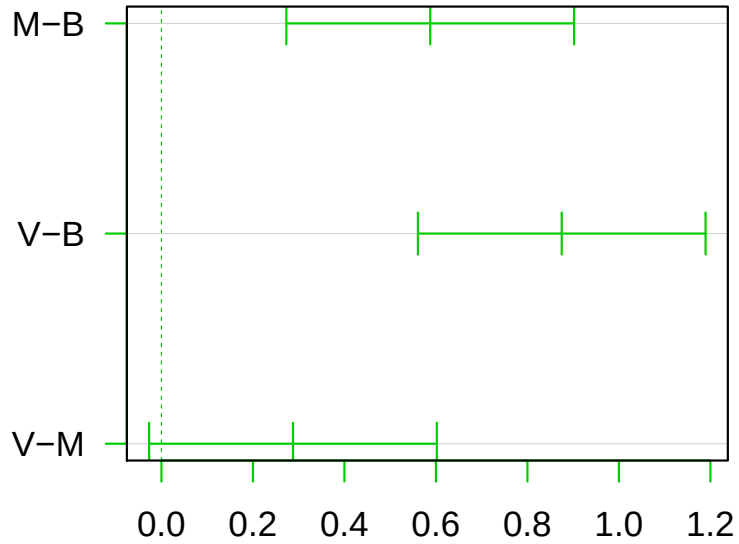


Differences in mean levels of puda

Scheffé (způsob hnojení)

```
  trt   means M
1   V 3.8375 a
2   M 3.5500 a
3   B 2.9625 b
```

95% family−wise confidence level



Differences in mean levels of hnojeni

ANOVA tabulka dvojného třídění s interakcemi

```
             Df  Sum Sq  Mean Sq F value    Pr(>F)
puda          1   0.002   0.0017   0.044   0.83658
hnojeni       2   3.182   1.5912  41.814 1.72e-07 ***
puda:hnojeni  2   0.551   0.2754   7.237   0.00494 **
Residuals    18   0.685   0.0381
```

parametry (nahoře efekty, dole interakce)

```
(Intercept)      pudaNor     hnojeniM     hnojeniV
      2.925        0.075        0.475        1.125
pudaNor:hnojeniM pudaNor:hnojeniV
           0.225               -0.500
```

```
Tables of effects
 puda
puda
      Kys        Nor
 0.008333 -0.008333
 hnojeni
hnojeni
      B        M        V
-0.4875   0.1000   0.3875
 puda:hnojeni
     hnojeni
puda   B         M         V
Kys  -0.04583 -0.15833  0.20417
Nor   0.04583  0.15833 -0.20417
```

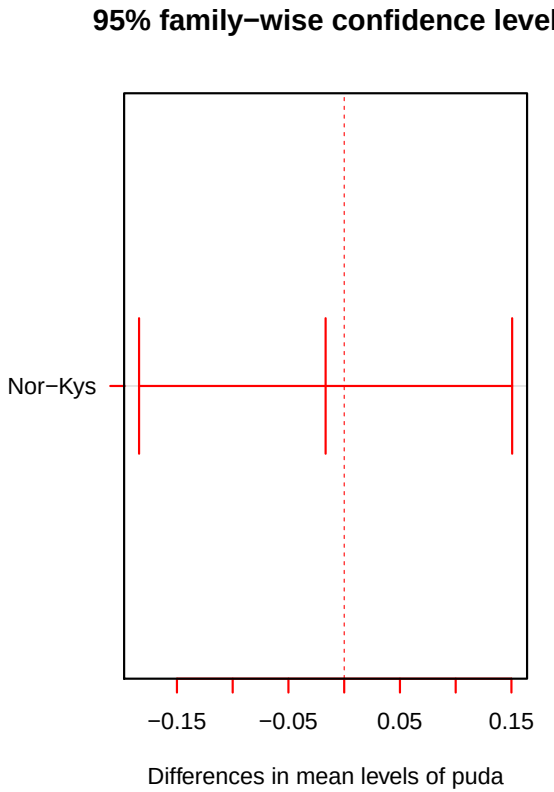
```
Tables of means
Grand mean
3.45
 puda
puda
   Kys   Nor
3.458 3.442
 hnojeni
hnojeni
    B     M     V
2.963 3.550 3.838
 puda:hnojeni
     hnojeni
puda   B     M     V
  Kys 2.925 3.400 4.050
  Nor 3.000 3.700 3.625
```

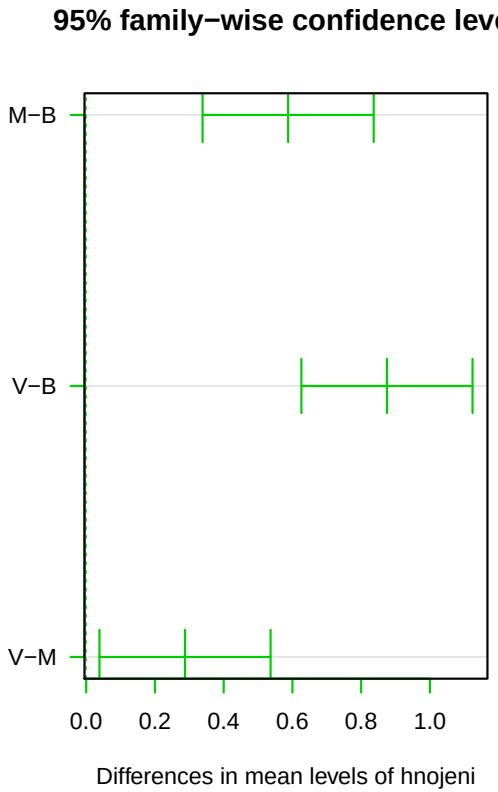
Scheffé (typ půdy)

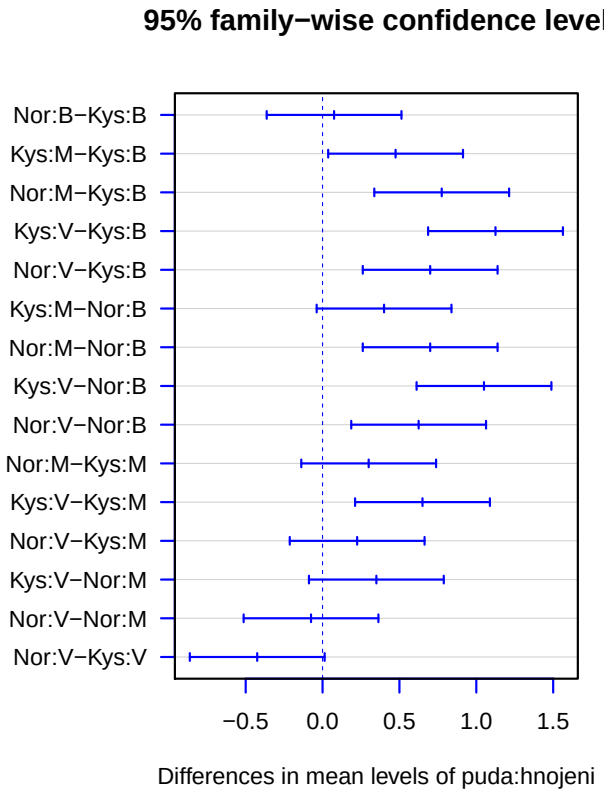
	trt	means	M
1	Kys	3.458333	a
2	Nor	3.441667	a

Scheffé (způsob hnojení)

	trt	means	M
1	V	3.8375	a
2	M	3.5500	b
3	B	2.9625	c







ANOVA tabulka jednoduchého třídění

```
             Df  Sum Sq  Mean Sq  F value    Pr(>F)
group         5   3.735   0.7470    19.63 1.02e-06 ***
Residuals    18   0.685   0.0381
```

parametry

```
(Intercept)
      2.925
groupKys.M  groupKys.V  groupNor.B  groupNor.M  groupNor.V
     0.475       1.125       0.075       0.775       0.700
```

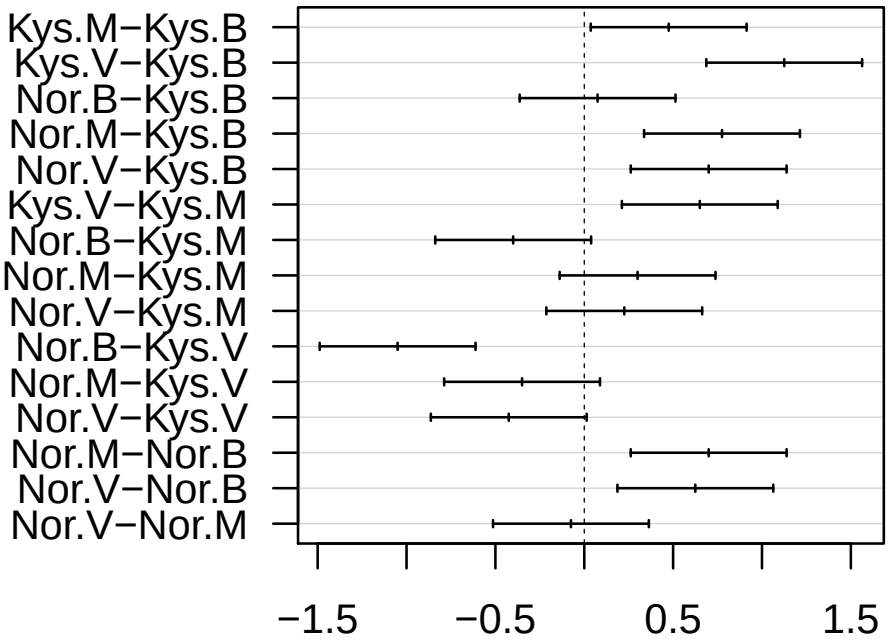
```
Tables of means
Grand mean
3.45
 group
group
Kys.B Kys.M Kys.V Nor.B Nor.M Nor.V
2.925 3.400 4.050 3.000 3.700 3.625
```

```
Tables of effects
 group
group
 Kys.B  Kys.M  Kys.V  Nor.B  Nor.M  Nor.V
-0.525 -0.050  0.600 -0.450  0.250  0.175
```

95% family−wise confidence level

Scheffé (půda.hnojení)

```
    trt means  M
1 Kys.V 4.050  a
2 Nor.M 3.700 ab
3 Nor.V 3.625 ab
4 Kys.M 3.400 bc
5 Nor.B 3.000  c
6 Kys.B 2.925  c
```



Differences in mean levels of group