Osnova přednášky Analýza rozptylu dvojného třídění

Motivace

Označení

Dvojné třídění bez interakcí

- Součty čtverců
- Testování hypotézy o nevýznamnosti sloupcového faktoru
- Testování hypotézy o nevýznamnosti řádkového faktoru
- Scheffého a Tukeyova metoda mnohonásobného porovnávání
- Příklad

Dvojné třídění s interakcemi

- Možné problémy v analýze rozptylu dvojného třídění s interakcemi
- Příklad

Analýza rozptylu dvojného třídění

Motivace: Zkoumáme vliv dvou faktorů A a B na závisle proměnnou veličinu Y. Např. zjišťujeme, zda výnosy určité plodiny (náhodná veličina Y) jsou ovlivněny typem půdy (faktor A) a způsobem hnojení (faktor B). Předpokládáme, že faktor A má a úrovní (tj. počet typů půdy) a faktor B má b úrovní (tj. počet způsobů hnojení). Přitom máme n_{ij} pokusů takových, že na i-tém typu půdy byl použit j-tý způsob hnojení. Výsledky (tzn. výnosy dané plodiny) těchto n_{ij} pokusů označíme $Y_{ij1}, Y_{ij2}, ..., Y_{ijn_{ij}}$. Omezíme se na případy, kdy počet pozorování $n_{ij} = c \geq 1$ (jde o tzv. **vyvážené třídění**). Výsledky lze zapsat do tabulky:

		faktor B			
		1	2	...	b
faktor A	1	$Y_{111}, \ldots, Y_{11c}$	$Y_{121}, \ldots, Y_{12c}$	...	$Y_{1b1}, \ldots, Y_{1bc}$
	2	$Y_{211}, \ldots, Y_{21c}$	$Y_{221}, \ldots, Y_{22c}$	...	$Y_{2b1}, \ldots, Y_{2bc}$
	⋮	⋮	⋮	...	⋮
	a	$Y_{a11}, \ldots, Y_{a1c}$	$Y_{a21}, \ldots, Y_{a2c}$	...	$Y_{ab1}, \ldots, Y_{abc}$

Analogicky jako u analýzy rozptylu jednoduchého třídění předpokládáme, že data se řídí normálním rozložením, tj.

$Y_{ij1}, Y_{ij2}, ..., Y_{ijc} \sim N(\mu_{ij}, \sigma^2)$, $i = 1, \ldots, a$, $j = 1, \ldots, b$

a jednotlivé náhodné výběry jsou stochasticky nezávislé, tedy

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

kde ε_{ijk} jsou stochasticky nezávislé náhodné veličiny s rozložením $N(0, \sigma^2)$.

Zajímá nás, zda všechny střední hodnoty μ_{ij} jsou stejné.

Přístup k problému se liší podle toho, zda faktory A, B jsou nezávislé (pak se jedná o **analýzu rozptylu dvojného třídění bez interakcí**) nebo se mohou nějakým způsobem ovlivňovat (jde o **analýzu rozptylu dvojného třídění s interakcemi**).

Označení

$n = abc,$

$$Y_{ij.} = \sum_{k=1}^{c} Y_{ijk},$$

$$M_{ij.} = \frac{1}{c} Y_{ij.},$$

$$Y_{i..} = \sum_{j=1}^{b} \sum_{k=1}^{c} Y_{ijk},$$

$$M_{i..} = \frac{1}{bc} Y_{i..},$$

$$Y_{...} = \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{c} Y_{ijk},$$

$$M_{...} = \frac{1}{n} Y_{...}$$

Analogické označení zavedeme i pro jiné kombinace indexů.

Dvojné třídění bez interakcí

Předpokládáme, že řádkový faktor A a sloupcový faktor B se neovlivňují (např. to znamená, že každý ze čtyř způsobů hnojení působí stejně na každém ze tří druhů půdy).

Náhodné veličiny Y_{ijk} se řídí modelem

M_0: $Y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}$ pro $i = 1, \ldots, a$, $j = 1, \ldots, b$, $k = 1, \ldots, c$, přičemž

ε_{ijk} jsou stochasticky nezávislé náhodné veličiny s rozložením $N(0, \sigma^2)$,

μ je společná část střední hodnoty závisle proměnné veličiny,

α_i je efekt faktoru A na úrovni i,

β_j je efekt faktoru B na úrovni j.

Parametry μ, α_i, β_j neznáme. Požadujeme, aby platily tzv. reparametrizační rovnice:

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

Součty čtverců

Podobně jako v analýze rozptylu jednoduchého třídění se počítají součty čtverců.

$S_T = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}\left(Y_{ijk} - M_{...}\right)^2$ … **celkový součet čtverců**,

počet stupňů volnosti $f_T = n - 1$,

$S_A = bc\sum_{i=1}^{a}\left(M_{i..} - M_{...}\right)^2$ … **součet čtverců pro řádkový faktor A**,

počet stupňů volnosti $f_A = a - 1$,

$S_B = ac\sum_{j=1}^{b}\left(M_{.j.} - M_{...}\right)^2$ … **součet čtverců pro sloupcový faktor B**,

počet stupňů volnosti $f_B = b - 1$,

$S_E = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}\left(Y_{ijk} - M_{ij.}\right)^2$ … **reziduální součet čtverců**,

počet stupňů volnosti $f_E = n - a - b + 1$.

Lze dokázat, že $S_T = S_A + S_B + S_E$:

$$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}\left(Y_{ijk} - M_{...}\right)^2 = bc\sum_{i=1}^{a}\left(M_{i..} - M_{...}\right)^2 + ac\sum_{j=1}^{b}\left(M_{.j.} - M_{...}\right)^2 + \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}\left(Y_{ijk} - M_{ij.}\right)^2$$

Celkový průměr $M_{...}$ je bodovým odhadem střední hodnoty μ,

rozdíl $M_{i..} - M_{...}$ představuje bodový odhad i-té úrovně řádkového faktoru α_i

rozdíl $M_{.j.} - M_{...}$ představuje bodový odhad j-té úrovně sloupcového faktoru β_j.

Odhad $\hat{Y}_{ijk}$ pozorování Y_{ijk} má tedy tvar:

$$\hat{Y}_{ijk} = M_{...} + \left(M_{i..} - M_{...}\right) + \left(M_{.j.} - M_{...}\right).$$

Testování hypotézy o nevýznamnosti sloupcového faktoru

Pokud by nezáleželo na sloupcovém faktoru B, platila by hypotéza $\beta_1 = \ldots = \beta_b = 0$ a dostali bychom model

M_1: $Y_{ijk} = \mu + \alpha_i + \varepsilon_{ijk}$

Platnost uvedené hypotézy ověřujeme pomocí testové statistiky

$F_B = \frac{S_B / f_B}{S_E / f_E}$, která se řídí rozložením F(b-1,n-a-b+1), je-li model M_1 správný.

Hypotézu o nevýznamnosti sloupcového faktoru tedy zamítneme na hladině významnosti α, když platí:

$F_B \geq F_{1-\alpha}(b-1,n-a-b+1)$.

Testování hypotézy o nevýznamnosti řádkového faktoru

Kdyby nezáleželo ani na řádkovém faktoru, platila by hypotéza $\alpha_1 = \ldots = \alpha_a = 0$ a dostali bychom model

M_2: $Y_{ijk} = \mu + \varepsilon_{ijk}$

Rozdíl mezi modely M_1 a M_2 ověřujeme pomocí testové statistiky

$F_A = \frac{S_A / f_A}{S_E / f_E}$, která se řídí rozložením F(a-1,n-a-b+1), je-li model M2 správný.

Hypotézu o nevýznamnosti řádkového faktoru tedy zamítneme na hladině významnosti α, když platí:

$F_A \geq F_{1-\alpha}(a-1,n-a-b+1)$.

Při uvedeném postupu tedy zjišťujeme, zda záleží na sloupcovém efektu B. Pokud ne, platí model M_1 a ptáme se, zda záleží na řádkovém efektu A, tj. zda platí model M_2. Postup lze samozřejmě provést i v jiném pořadí – nejdřív zkoumáme řádkový efekt A (tj. ověřujeme platnost modelu M_1': $Y_{ijk} = \mu + \beta_j + \varepsilon_{ijk}$) a poté sloupcový efekt B. Lze ukázat, že oba řetězce $M_0 \rightarrow M_1 \rightarrow M_2$ a $M_0 \rightarrow M_1' \rightarrow M_2'$ dají stejné výsledky. (To platí pouze za předpokladu, že $n_{ij} = c$ pro všechna i, j.)

Výsledky výpočtů zapisujeme do **tabulky analýzy rozptylu dvojného třídění bez interakcí**.

Zdroj variability	součet čtverců	st. vol.	podíl S/f	$F = \frac{S/f}{S_E/f_E}$
řádkový efekt A	S_A	$f_A = a-1$	S_A/f_A	$F_A = \frac{S_A/f_A}{S_E/f_E}$
sloupcový efekt B	S_B	$f_B = b-1$	S_B/f_B	$F_B = \frac{S_B/f_B}{S_E/f_E}$
reziduální	S_E	$f_E = n-a-b+1$	S_E/f_E	-
celkem	S_T	$f_T = n-1$	-	-

Scheffého a Tukeyova metoda mnohonásobného porovnávání

Zjistíme-li, že existují významné rozdíly mezi řádky, můžeme pomocí Scheffého nebo Tukeyovy metody zjistit, které dvojice řádků se významně liší. Určíme tedy, které rozdíly $\alpha_i - \alpha_t$ jsou nenulové (na dané hladině významnosti).

Podle **Scheffého metody** zamítneme rovnost $\alpha_i = \alpha_t$, když

$$|M_{i..} - M_{t..}| > \sqrt{\frac{2(a-1)}{bc} \cdot \frac{S_E}{n-a-b+1} \cdot F_{1-\alpha}(a-1, n-a-b+1)}$$

a podle **Tukeyovy metody**, když

$$|M_{i..} - M_{t..}| > \sqrt{\frac{1}{bc} \cdot \frac{S_E}{n-a-b+1}} q_{1-\alpha}(a, n-a-b+1),$$ kde $q_{1-\alpha}(a,n-a-b+1)$ najdeme

v tabulkách kvantilů studentizovaného rozpětí.

Jestliže zjistíme významný rozdíl mezi sloupci, určujeme podobně, které dvojice sloupců se mezi sebou liší, tj. které rozdíly $\beta_j - \beta_t$ jsou nenulové.

Podle **Scheffého metody** zamítneme rovnost $\beta_j = \beta_t$, když

$$\left|M_{.j.} - M_{.t.}\right| > \sqrt{\frac{2(b-1)}{ac} \cdot \frac{S_E}{n-a-b+1} \cdot F_{1-\alpha}(b-1, n-a-b+1)}$$

a podle **Tukeyovy metody**, když

$$\left|M_{.j.} - M_{.t.}\right| > \sqrt{\frac{1}{ac} \cdot \frac{S_E}{n-a-b+1}} q_{1-\alpha}(b, n-a-b+1).$$

Příklad:
Byly zaznamenány tržby za prodej určitého zboží během tří stejně dlouhých časových období. Přitom byl sledován jednak vliv balení zboží (řádkový faktor A, úroveň 1 – balení v sáčku, úroveň 2 – balení v krabičce) a jednak vliv druhu reklamy (sloupcový faktor B, úroveň 1 – bez reklamy, úroveň 2 – reklama v novinách, úroveň 3 – reklama v TV a novinách). Výsledky prodeje (tj. hodnota prodaného zboží v miliónech Kč) jsou zaznamenány v tabulce:

		B		
		1-bez reklamy	2-reklama v novinách	3-reklama v TV a novinách
A	1-balení v sáčku	1	1	6
	2-balení v krabičce	3	4	9

Na hladině významnosti 0,05 je třeba posoudit vliv reklamy i vliv balení zboží na jeho prodej.

Řešení:

		B		
		1-bez reklamy	2-reklama v novinách	3-reklama v TV a novinách
A	1-balení v sáčku	1	1	6
	2-balení v krabičce	3	4	9

Data zpracujeme pomocí analýzy rozptylu dvojného třídění bez interakcí. Přitom a = 2, b = 3, c = 1, n = 6. Nejprve provedeme pomocné výpočty:
Součet všech hodnot: $Y_{...} = 24$
Průměr všech hodnot: $M_{...} = 24/6 = 4$
Řádkové součty a průměry:
$Y_{1..} = 8$, $Y_{2..} = 16$, $M_{1..} = 8/3 = 2{,}67$, $M_{2..} = 16/3 = 5{,}33$
Sloupcové součty a průměry:
$Y_{.1.} = 4$, $Y_{.2.} = 5$, $Y_{.3.} = 15$, $M_{.1.} = 4/2 = 2$, $M_{.2.} = 5/2 = 2{,}5$, $M_{.3.} = 15/2 = 7{,}5$.

$a = 2$, $b = 3$, $c = 1$, $n = 6$,
Celkový součet a průměr: $Y_{...} = 24$, $M_{...} = 24/6 = 4$
Řádkové součty a průměry: $Y_{1..} = 8$, $Y_{2..} = 16$, $M_{1..} = 8/3$, $M_{2..} = 16/3$,
Sloupcové součty a průměry: $Y_{.1.} = 4$, $Y_{.2.} = 5$, $Y_{.3.} = 15$, $M_{.1.} = 4/2 = 2$, $M_{.2.} = 5/2$, $M_{.3.} = 15/2$.

Řádkový součet čtverců:

$$S_A = bc\sum_{i=1}^{a}\left(M_{i..} - M_{...}\right)^2 = 3\left[\left(\frac{8}{3}-4\right)^2 + \left(\frac{16}{3}-4\right)^2\right] = \frac{32}{3} = 10,\overline{6},$$

Sloupcový součet čtverců:

$$S_B = ac\sum_{j=1}^{b}\left(M_{.j.} - M_{...}\right)^2 = 2\left[(2-4)^2 + \left(\frac{5}{2}-4\right)^2 + \left(\frac{15}{2}-4\right)^2\right] = 37,$$

Celkový součet čtverců:

$$S_T = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}\left(Y_{ijk} - M_{...}\right)^2 = (1-4)^2 + (1-4)^2 + (6-4)^2 + (3-4)^2 + (4-4)^2 + (9-4)^2 = 48,$$

Reziduální součet čtverců:

$$S_E = S_T - S_A - S_B = 48 - 10,\overline{6} - 37 = 0,\overline{3}.$$

Výsledky zapíšeme do **tabulky analýzy rozptylu dvojného třídění bez interakcí**.

Zdroj variability	součet čtverců	st. vol.	podíl S/f	$F = \frac{S/f}{S_E/f_E}$
způsob balení	$10,\overline{6}$	1	$10,\overline{6}$	63,99
druh reklamy	37	2	18,5	110,98
reziduální	$0,\overline{3}$	2	$0,1\overline{6}$	-
celkem	48	5	-	-

Odpovídající kvantily:
pro řádkový efekt $F_{0,95}(1,2) = 18,1$,
pro sloupcový efekt $F_{0,95}(2,2) = 19$.
Protože $F_A = 63,99 \geq 18,1$, zamítáme na hladině významnosti 0,05 hypotézu, že způsob balení nemá vliv na prodej zboží. Podobně $F_B = 110,98 \geq 19$, tedy na hladině významnosti 0,05 zamítáme hypotézu, že druh reklamy nemá vliv na prodej zboží.

V případě sloupcového faktoru – druh reklamy - lze pomocí Scheffého nebo Tukeyovy metody zjistit, které druhy reklamy se od sebe liší na hladině významnosti 0,05.
Nejprve vypočítáme absolutní hodnoty rozdílů sloupcových průměrů:

$$\left|M_{.1.}-M_{.2.}\right|=\left|2-\frac{5}{2}\right|=0{,}5,\left|M_{.1.}-M_{.3.}\right|=\left|2-\frac{15}{2}\right|=5{,}5,\left|M_{.2.}-M_{.3.}\right|=\left|\frac{5}{2}-\frac{15}{2}\right|=5$$

Pravá strana Scheffého vzorce je:

$$\sqrt{\frac{2(b-1)}{ac}\cdot\frac{S_E}{n-a-b+1}\cdot F_{1-\alpha}(b-1,n-a-b+1)}=\sqrt{\frac{2}{2}\cdot 0{,}1\overline{6}\cdot 19}=2{,}52.$$

Vidíme, že podle Scheffého metody se na hladině významnosti 0,05 liší sloupce 1, 3 (tj. bez reklamy a s reklamou v TV a novinách) a sloupce 2, 3 (tj. s reklamou jen v novinách a reklamou v TV a novinách).
Pravá strana Tukeyova vzorce je:

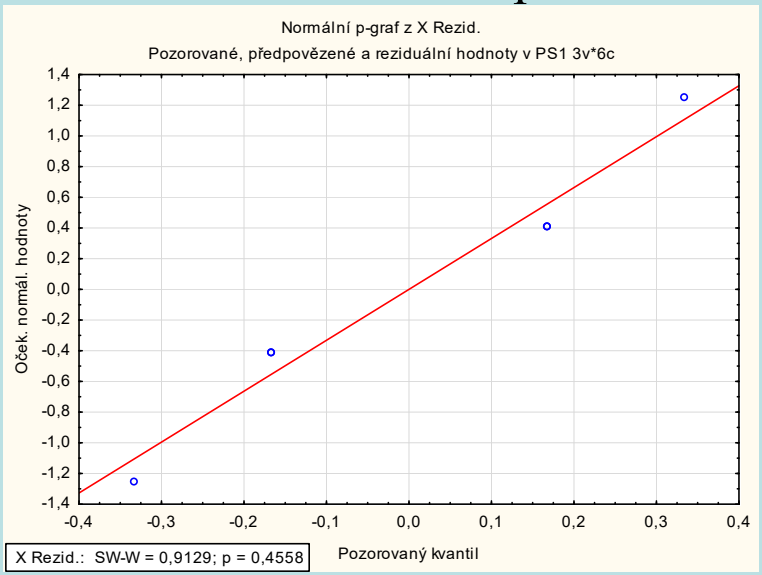
$$\sqrt{\frac{1}{ac}\cdot\frac{S_E}{n-a-b+1}}q_{1-\alpha}(b,n-a-b+1)=\sqrt{\frac{0{,}1\overline{6}}{2}}\cdot q_{0{,}95}(3,2)=\sqrt{\frac{0{,}1\overline{6}}{2}}\cdot 8{,}33=2{,}4.$$

Podle Tukeyovy metody se na hladině významnosti 0,05 také liší sloupce 1, 3 a sloupce 2, 3. Výhodnější je hodnota získaná Tukeyovou metodou, protože je menší.

Podívejme se ještě na počítačové výstupy. Nejprve ověříme předpoklady metody.

Nezávislost: splněno, plyne přímo za způsobu získání dat.

Normalita dat: ověříme pomocí N-P grafu a S-W testu aplikovaného na rezidua:



Na hladině významnosti 0,05 hypotézu o normalitě nezamítáme, p-hodnota S-W testu je 0,4558, což je větší než 0,05.

Homogenita rozptylů: nelze ověřit, všech šest výběrů má rozsah 1.

Výpočet průměrů:

Efekt	Úroveň Faktor	N	X Průměr
Celkem		6	4,000000
A	sacek	3	2,666667
A	krabicka	3	5,333333
B	bez reklamy	2	2,000000
B	reklama v novinach	2	2,500000
B	reklama v TV a novinach	2	7,500000

Tabulka dvoufaktorové ANOVY bez interakcí:

Jednorozm. výsledky pro každou záv. proměnnou (baleni_a_reklama.sta)
Přeparametrizovaný model
Dekompozice typu III

Efekt	Stupně volnosti	X SČ	X PČ	X F	X p
A	1	10,6667	10,66667	64,0000	0,015268
B	2	37,0000	18,50000	111,0000	0,008929
Chyba	2	0,3333	0,16667		
Celkem	6	144,0000			

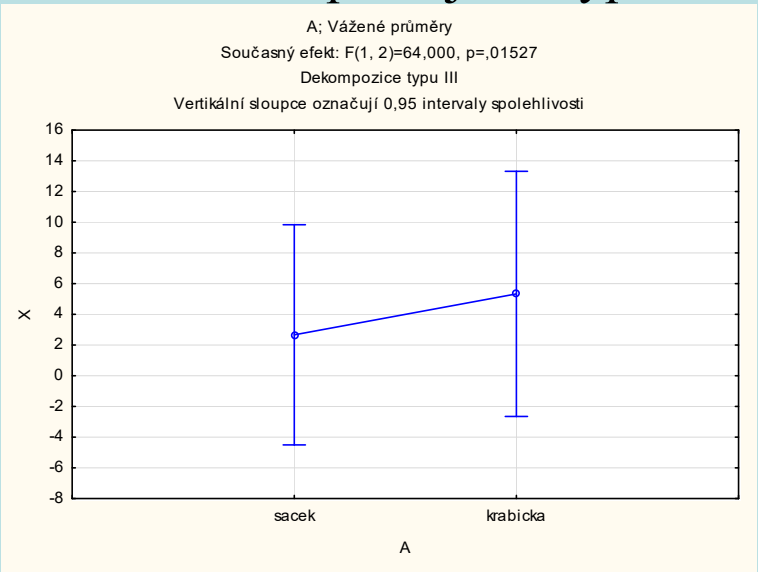
Na hladině významnosti 0,05 zamítáme jak hypotézu o nevýznamnosti typu balení výrobku tak hypotézu o nevýznamnosti druhu reklamy.

Tabulka p-hodnot pro Tukeyovu metodu mnohonásobného porovnávání druhů reklamy:

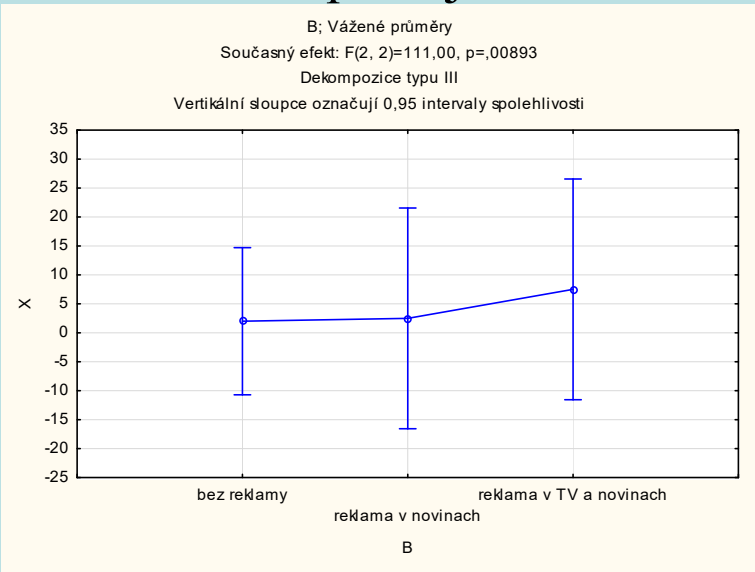
Č. buňky	B	{1} 2,0000	{2} 2,5000	{3} 7,5000
1	bez reklamy		0,548301	0,010156
2	reklama v novinach	0,548301		0,012218
3	reklama v TV a novinach	0,010156	0,012218	

Na hladině významnosti 0,05 se liší dvojice variant (bez reklamy, reklama v TV a novinách) a dvojice (reklama v novinách, reklamy v TV a novinách). Naopak se neliší dvojice (bez reklamy, reklama v novinách).

Graf závislosti prodeje na typu balení:



Graf závislosti prodeje na druhu reklamy:



Řešení pomocí systému R

Načteme data:

```
> Y<-c(1,1,6,3,4,9)
> A<-c(1,1,1,2,2,2)
> B<-c(1,2,3,1,2,3)
> A<-factor(A,labels=c('sacek','krabicka'))
> B<-factor(B,labels=c('bez reklamy','reklama v novinach','reklama v TV a
novinach'))
```

Vypočteme průměry a směrodatné odchylky v jednotlivých skupinách tříděných podle faktoru A:

```
> tapply(Y,A,mean)
   sacek krabicka
2.666667 5.333333

> tapply(Y,A,sd)
   sacek krabicka
2.886751 3.214550 tapply(Y,ID,sd)
```

Vypočteme průměry a směrodatné odchylky v jednotlivých skupinách tříděných podle faktoru B:

```
> tapply(Y,B,mean)
bez reklamy      reklama v novinach reklama v TV a novinach
    2.0                     2.5                     7.5

> tapply(Y,B,sd)
bez reklamy      reklama v novinach reklama v TV a novinach
  1.414214                2.121320                2.121320
```

Celkový průměr:

```
> mean(Y)
[1] 4
```

Testujeme hypotézu o nevýznamnosti faktorů A, B:

```
> vystup<-aov(Y~A+B)
> summary(vystup)
            Df Sum Sq Mean Sq F value  Pr(>F)
A            1  10.67  10.667      64 0.01527 *
B            2  37.00  18.500     111 0.00893 **
Residuals    2   0.33   0.167
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Vidíme, že p-hodnota testu o nevýznamnosti způsobu balení 0,01527, což je menší než 0,05, tedy na hladině významnosti 0,05 nulovou hypotézu zamítáme. Rovněž p-hodnota 0,00893 testu o nevýznamnosti druhu reklamy je menší než 0,05, tedy na hladině významnosti 0,05 nulovou hypotézu zamítáme. S rizikem omylu nejvýše 5 % jsme prokázali, že úroveň prodeje daného druhu zboží závisí na způsobu balení a druhu reklamy.

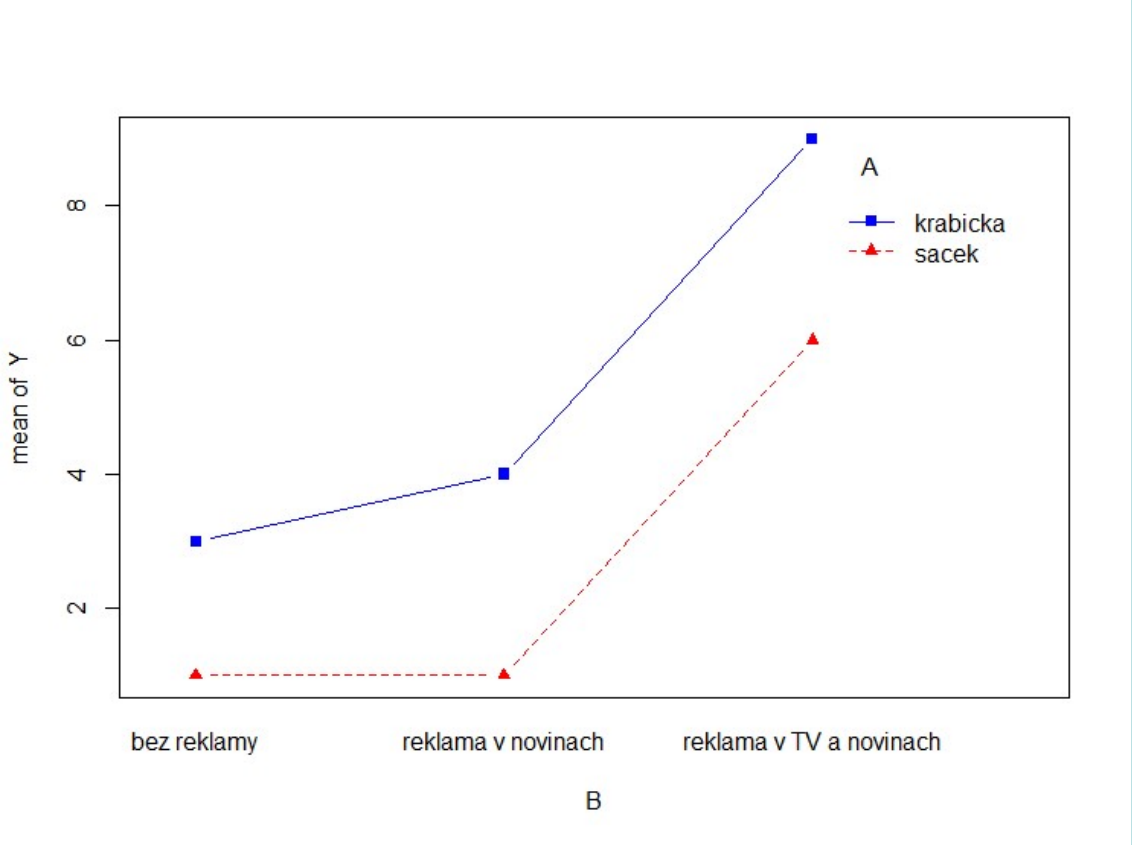
Protože jsme hypotézy o nevýznamnosti faktorů A, B zamítli na hladině významnosti 0,05, přistoupíme k Tukeyově metodě mnohonásobného porovnání:

```
> TukeyHSD(vystup)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Y ~ A + B)

$A
                    diff      lwr      upr    p adj
krabicka-sacek 2.666667 1.233682 4.099651 0.015116

$B
                                                diff      lwr      upr      p adj
reklama v novinach-bez reklamy                   0.5 -1.90489 2.90489 0.5481840
reklama v TV a novinach-bez reklamy              5.5  3.09511 7.90489 0.0099614
reklama v TV a novinach-reklama v novinach       5.0  2.59511 7.40489 0.0120317
```

Vidíme, že p-hodnota pro porovnání úrovně prodeje při použití reklamy v TV a novinách a bez reklamy je 0,0099, tedy na hladině významnosti 0,05 je rozdíl v úrovních prodeje prokázán. Dále, p-hodnota pro porovnání úrovně prodeje při použití reklamy v TV a novinách a reklamy jenom v novinách je 0,012, tedy na hladině významnosti 0,05 je rozdíl v úrovních prodeje prokázán.

Nakonec vykreslíme graf závislosti počtu prodaných kusů zboží na druhu reklamy pro různé způsoby balení:

```
> interaction.plot(B,A,Y,mean,"b",col=c("red","blue"),pch=c(17,15))
```



Dvojné třídění s interakcemi

Nyní předpokládáme, že faktory A a B se mohou ovlivňovat (např. některý způsob hnojení má zcela specifický vliv na určitý typ půdy). Náhodné veličiny Y_{ijk} se řídí modelem

M_0: $Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}$ pro $i = 1, \ldots, a$, $j = 1, \ldots, b$, $k = 1, \ldots, c$, přičemž γ_{ij} je interakce mezi faktorem A na úrovni i a faktorem B na úrovni j. V této situaci předpokládáme, že $c \geq 2$. Parametry μ, α_i, β_j neznáme. Požadujeme, aby platily tzv. reparametrizační rovnice:

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

Nyní můžeme utvořit modely

M_1: $Y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}$

M_2: $Y_{ijk} = \mu + \alpha_i + \varepsilon_{ijk}$

M_3: $Y_{ijk} = \mu + \varepsilon_{ijk}$

(Lze samozřejmě použít i jiný řetězec modelů, kdy postupně klademe rovny nule parametry α_i, β_j, γ_{ij} v jiném pořadí.)

Vypočítáme součty čtverců:

celkový $S_T = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}\left(Y_{ijk} - M_{...}\right)^2$, $f_T = n - 1$,

řádkový $S_A = bc\sum_{i=1}^{a}\left(M_{i..} - M_{...}\right)^2$, $f_A = a - 1$,

sloupcový $S_B = ac\sum_{j=1}^{b}\left(M_{.j.} - M_{...}\right)^2$, $f_B = b - 1$,

reziduální $S_E = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}\left(Y_{ijk} - M_{ij.}\right)^2$, $f_E = n - ab$

a součet čtverců pro interakce $S_{AB} = c\sum_{i=1}^{a}\sum_{j=1}^{b}\left[\left(M_{ij.} - M_{i..}\right) - \left(M_{.j.} - M_{...}\right)\right]^2$, $f_{AB} = (a-1)(b-1)$.

Vliv interakcí je prokázán na hladině významnosti α, když

$$F_{AB} = \frac{S_{AB}/f_{AB}}{S_E/f_E} \geq F_{1-\alpha}\left((a-1)(b-1), n-ab\right).$$

Výsledky zapisujeme do **tabulky analýzy rozptylu dvojného třídění s interakcemi:**

Zdroj variability	součet čtverců	st. vol.	podíl S/f	$F = \frac{S/f}{S_E/f_E}$
řádkový faktor A	S_A	$f_A = a-1$	S_A/f_A	$F_A = \frac{S_A/f_A}{S_E/f_E}$
sloupcový faktor B	S_B	$f_B = b-1$	S_B/f_B	$F_B = \frac{S_B/f_B}{S_E/f_E}$
interakce A,B	S_{AB}	$f_{AB} = (a-1)(b-1)$	S_{AB}/f_{AB}	$F_{AB} = \frac{S_{AB}/f_{AB}}{S_E/f_E}$
reziduální	S_E	$f_E = n-ab$	S_E/f_E	-
celkem	S_T	$f_T = n-1$	-	-

Je třeba si povšimnout, že součet $S_{AB} + S_E$ resp. $f_{AB} + f_E$ dá hodnotu S_E resp. f_E v tabulce bez interakcí.

Možné problémy v analýze rozptylu dvojného třídění s interakcemi

a) Ukáže-li se vliv interakcí nevýznamný, vzniká otázka, zda testovat vliv řádků resp. sloupců pomocí tabulky s interakcemi nebo provést novou analýzu rozptylu, ale tentokrát bez interakcí. Převládá názor, že je zapotřebí dokončit analýzu rozptylu s interakcemi.

b) Pokud interakce vyjdou významné a řádky a sloupce rovněž, zpravidla se nedoporučuje provádět mnohonásobné porovnávání, protože by se mohlo stát, že některá interakce by byla mnohem výraznější než příslušný řádkový resp. sloupcový efekt.

c) Nejsou-li interakce významné a řádky resp. sloupce ano, pak lze provést mnohonásobné porovnávání zcela analogicky jako v případě třídění bez interakcí, avšak je jiný počet stupňů volnosti f_E.

Tabulka odhadů různých parametrů a rozptylů těchto odhadů

parametr	odhad	rozptyl odhadu
μ	$M_{...}$	σ^2/n
$\mu + \alpha_i$	$M_{i..}$	σ^2/bc
$\mu + \beta_j$	$M_{.j.}$	σ^2/ac
$\mu + \alpha_i + \beta_j + \gamma_{ij}$	$M_{ij.}$	σ^2/c
α_i	$M_{i..} - M_{...}$	$\sigma^2(a-1)/n$
β_j	$M_{.j.} - M_{...}$	$\sigma^2(b-1)/n$
γ_{ij}	$(M_{ij.} - M_{i..}) - (M_{.j.} - M_{...})$	$\sigma^2(a-1)(b-1)/n$

Neznámý rozptyl σ^2 nahradíme jeho odhadem, tj. průměrným reziduálním čtvercem

$$s^2 = \frac{S_e}{n - ab}.$$

Příklad:
Byly zkoumány výnosy sena (v q/ha) v závislosti na typu půdy (řádkový faktor A, úroveň 1 – normální půda, úroveň 2 – kyselá půda) a na způsobu hnojení (sloupcový faktor B, úroveň 1 – bez hnojení, úroveň 2 – hnojení chlévskou mrvou, úroveň 3 – hnojení vápenatým hnojivem). Každá kombinace faktorů A a B byla realizována čtyřikrát nezávisle na sobě. Výnosy sena jsou uvedeny v tabulce:

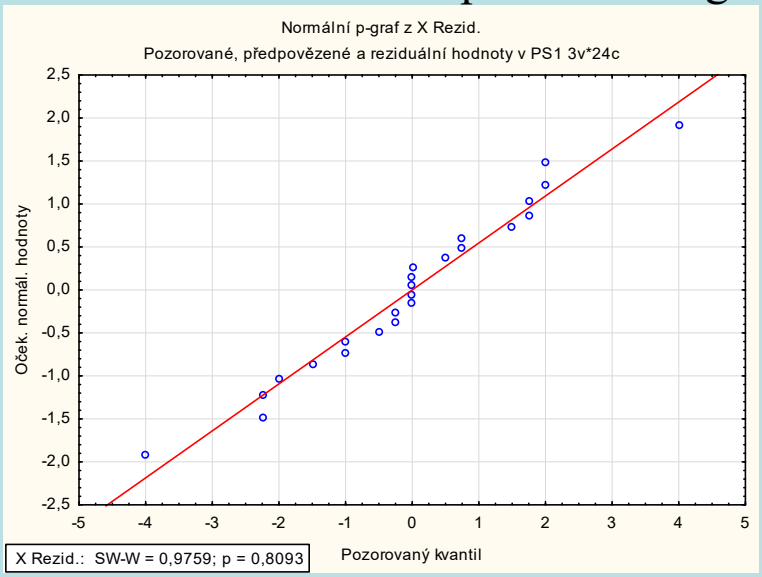
		B		
		1-bez hnojení	2-chlévská mrva	3-vápenaté hnojivo
A	1-normální půda	28 32 30 30	37 36 39 36	34 38 37 36
	2-kyselá půda	31 27 30 29	34 34 30 38	42 40 41 39

Na hladině významnosti 0,05 máme posoudit vliv typu půdy a způsobu hnojení (včetně případných interakcí) na výnosy sena.

Řešení:
Data zpracujeme pomocí analýzy rozptylu dvojného třídění s interakcemi. Přitom a = 2, b = 3, c = 4, n = abc = 24.

Ověření předpokladů:
Nezávislost všech šesti výběrů: splněno, plyne přímo ze způsobu získání dat.

Normalita dat: ověřeno pomocí N-P grafu a S-W testu aplikovaného na rezidua.



Příslušná p-hodnota je 0,8093, tedy na hladině významnosti 0,05 hypotézu o normalitě reziduí nezamítáme.

Homogenita rozptylů: nemusí se zkoumat, jde o vyvážené třídění. Jinak lze ověřit pomocí Levenova testu.

	PČ Efekt	PČ Chyba	F	p
X	0,600000	1,555556	0,385714	0,852058

Levenův test hypotézu o homogenitě rozptylů nezamítá na hladině významnosti 0,05, protože jeho p-hodnota je 0,852.

Průměrné výnosy ve všech šesti skupinách:

Č. buňky	A	B	X Průměr	N
1	normální	bez hnojení	30	4
2	normální	chlévská mrva	37	4
3	normální	vápenaté hnojivo	36,25	4
4	kyselá	bez hnojení	29,25	4
5	kyselá	chlévská mrva	34	4
6	kyselá	vápenaté hnojivo	40,5	4

Tabulka dvoufaktorové ANOVY s interakcemi:

Zdroj variability	součet čtverců	st. vol.	podíl S/f	$F = \frac{S/f}{S_E/f_E}$
typ půdy	0,166	1	0,166	0,04
způsob hnojení	318,25	2	159,125	41,81
interakce	55,084	2	27,542	7,24
reziduální	68,5	18	3,8056	-
celkem	442	23	-	-

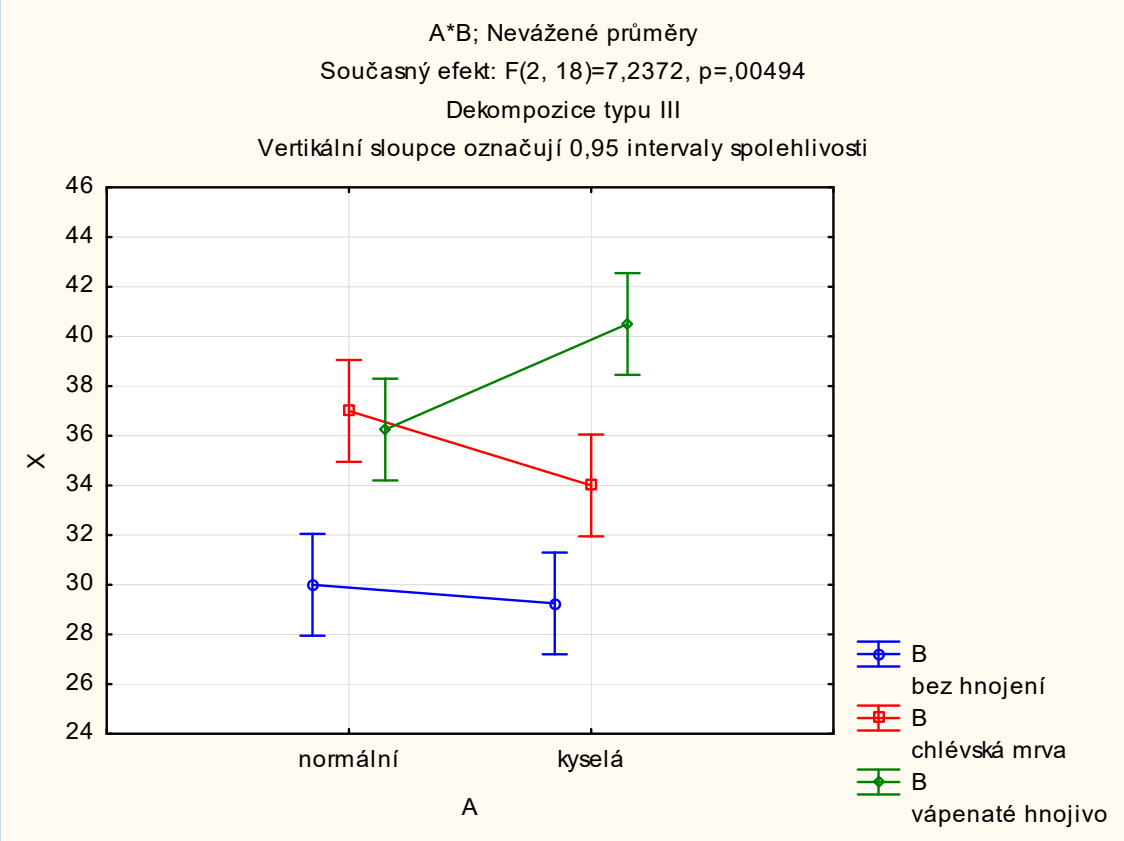
Odpovídající kvantily:
pro řádkový efekt $F_{0,95}(1,18) = 4,41$, pro sloupcový efekt $F_{0,95}(2,18) = 3,55$, pro interakce $F_{0,95}(2,18) = 3,55$.
Protože $F_A = 0,04 < 4,41$, nezamítáme na hladině významnosti 0,05 hypotézu, že typ půdy neovlivňuje výnos sena.
Dále $F_B = 41,81 \geq 3,55$, tedy na hladině významnosti 0,05 se prokázal rozdíl mezi použitými způsoby hnojení.
Jelikož $F_{AB} = 7,24 \geq 3,55$, zamítáme na hladině významnosti 0,05 hypotézu o nevýznamnosti interakcí (tj. aspoň jeden způsob hnojení působí jinak na půdu normální než kyselou).

Počítačový výstup:

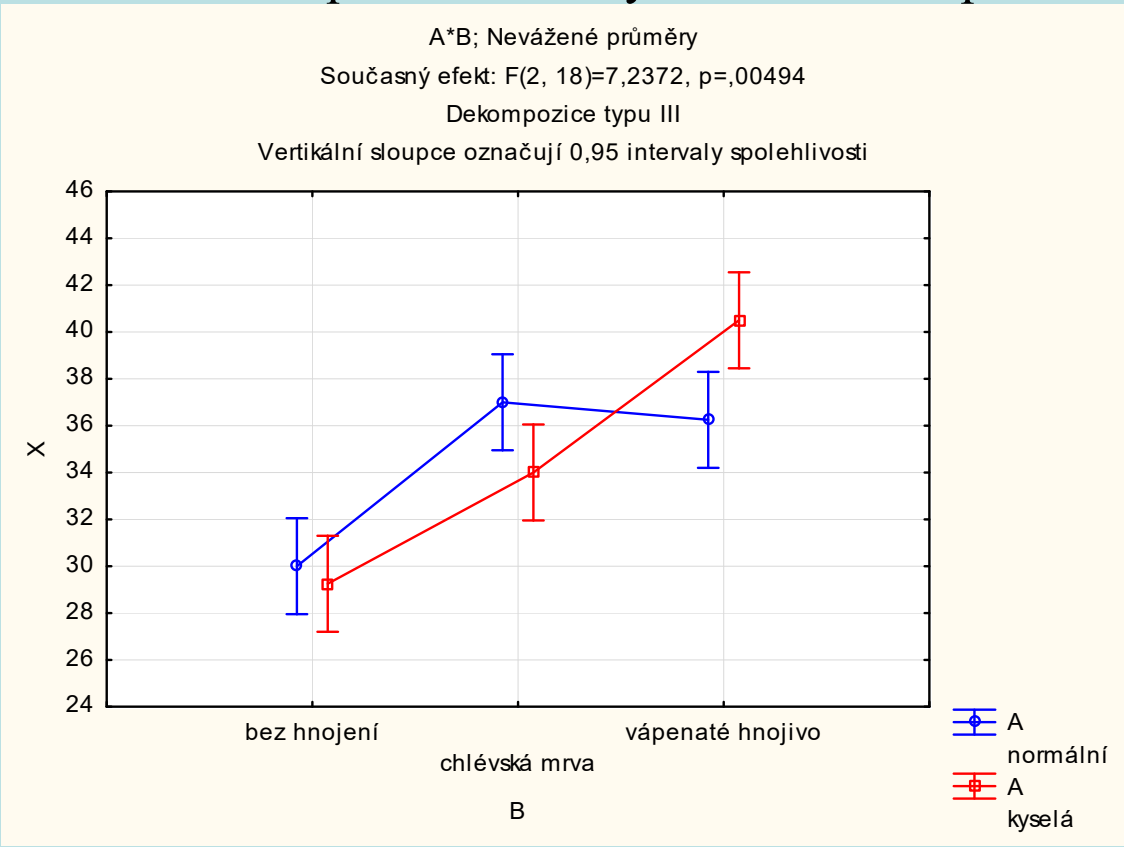
Efekt	Jednorozměrné testy významnosti pro X (seno.sta) Přeparametrizovaný model Dekompozice typu III SČ	Stupně volnosti	PČ	F	p
A	0,1667	1	0,1667	0,04380	0,836585
B	318,2500	2	159,1250	41,81387	0,000000
A*B	55,0833	2	27,5417	7,23723	0,004938
Chyba	68,5000	18	3,8056		

Vidíme, že p-hodnota pro testovou statistiku F_B je velmi blízká 0, tedy na hladině významnosti 0,05 zamítáme hypotézu, že způsob hnojení nemá vliv na výnosy sena. Podobně p-hodnota pro testovou statistiku F_{AB} je 0,004938, což znamená, že na hladině významnosti 0,05 zamítáme hypotézu, že způsob hnojení působí na oba typy půd stejně.

Graf závislosti průměrného výnosu sena na typu půdy:



Graf závislosti průměrného výnosu sena na způsobu hnojení:



V obou grafech se objevuje křížení, které je typické pro případ, kdy působí interakce mezi faktory A, B.

Řešení pomocí systému R

Načteme data:

```
> seno <- read_csv("seno.csv")
```

Zavedeme faktory A, B:

```
> seno$A<-factor(seno$A)
> seno$B<-factor(seno$B)
```

Zjistíme počty pozorování v jednotlivých skupinách:

```
> table(seno$A,seno$B)
          bez hnojeni chlevska mrva vapenate hnojivo
  kysela             4             4                4
  normalni           4             4                4
```

Vypočteme průměry a směr. odchylky výnosu sena ve skupinách tříděných podle faktoru A:

```
> tapply(seno$X,seno$A,mean)
  kysela normalni
34.58333 34.41667
> tapply(seno$X,seno$A,sd)
  kysela normalni
5.230302 3.579191
```

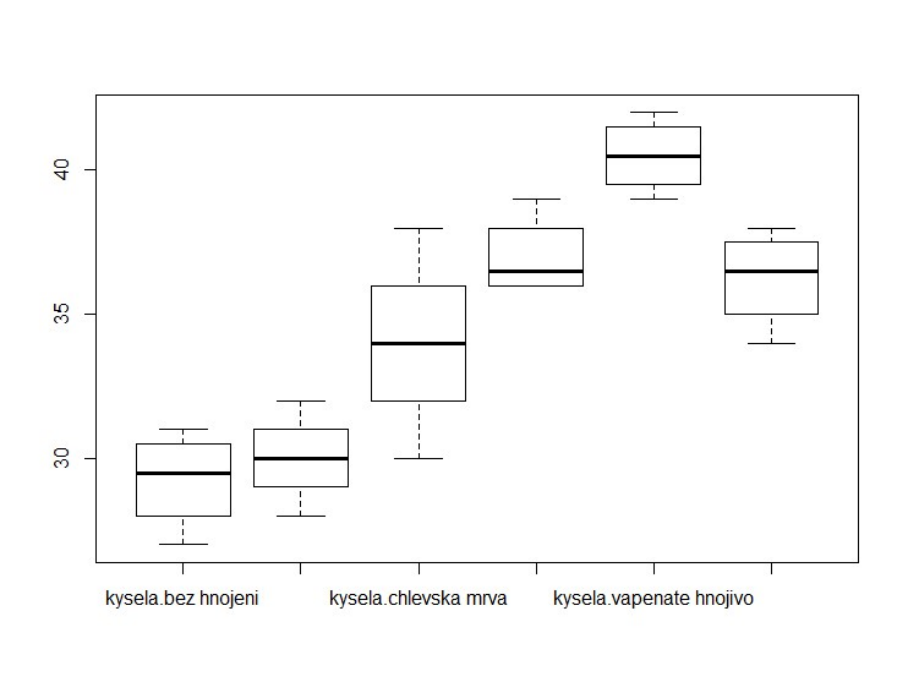
Vypočteme průměry a směr. odchylky výnosu sena ve skupinách tříděných podle faktoru B:

```
> tapply(seno$X,seno$B,mean)
    bez hnojeni    chlevska mrva vapenate hnojivo
          29.625           35.500           38.375
> tapply(seno$X,seno$B,sd)
    bez hnojeni    chlevska mrva vapenate hnojivo
       1.597990         2.828427         2.669270
```

Celkový průměr:

```
> mean(seno$X)
[1] 34.5
```

Vykreslíme krabicové diagramy:



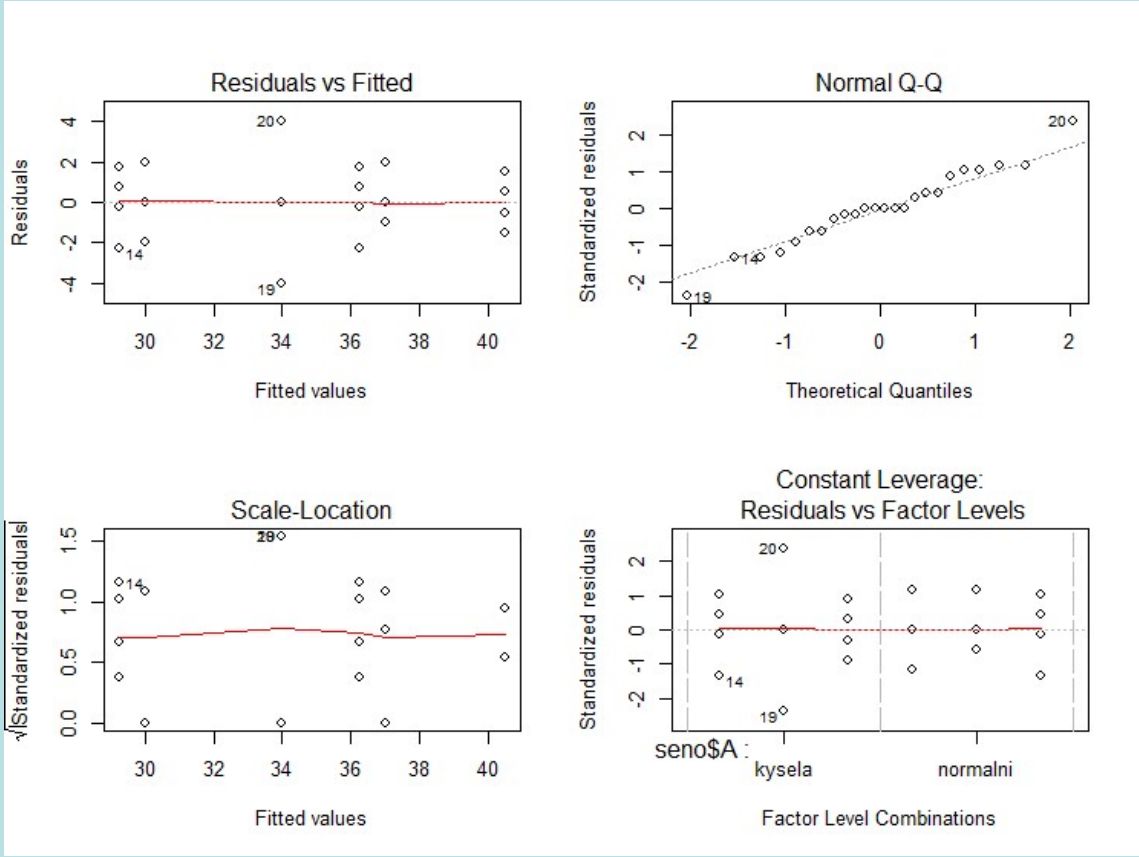
Testujeme hypotézu o nevýznamnosti faktorů A, B a jejich interakcí:

```
> vystup<-aov(seno$X~seno$A*seno$B)
> summary(vystup)
              Df Sum Sq Mean Sq F value   Pr(>F)
seno$A         1    0.2    0.17   0.044  0.83658
seno$B         2  318.2  159.12  41.814 1.72e-07 ***
seno$A:seno$B  2   55.1   27.54   7.237  0.00494 **
Residuals     18   68.5    3.81
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Vidíme, že p-hodnota testu o nevýznamnosti typu půdy je 0,83658, což je větší než 0,05, tedy na hladině významnosti 0,05 nulovou hypotézu nezamítáme. Ovšem p-hodnota $1{,}72*e^{-7}$ testu o nevýznamnosti způsobu hnojení je menší než 0,05, tedy na hladině významnosti 0,05 nulovou hypotézu zamítáme. Také p-hodnota 0,00494 testu o nevýznamnosti interakcí mezi typem půdy a způsobem hnojení je menší než 0,05. S rizikem omylu nejvýše 5 % jsme prokázali, že výnos sena závisí na způsobu hnojení a že různé způsoby hnojení působí jinak na různých typech půdy.

Předpoklady modelu ověříme pomocí diagnostických grafů:

```
> par(mfrow=c(2,2))
> plot(vystup)
```



Protože jsme hypotézy o nevýznamnosti faktoru B a nevýznamnosti interakcí zamítli na hladině významnosti 0,05, přistoupíme k Tukeyově metodě mnohonásobného porovnání:

```
> TukeyHSD(vystup)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = seno$X ~ seno$A * seno$B)

$`seno$A`
                     diff       lwr      upr     p adj
normalni-kysela -0.1666667 -1.839849 1.506516 0.8365845

$`seno$B`
                               diff       lwr       upr     p adj
chlevska mrva-bez hnojeni      5.875 3.3856414  8.364359 0.0000307
vapenate hnojivo-bez hnojeni   8.750 6.2606414 11.239359 0.0000001
vapenate hnojivo-chlevska mrva 2.875 0.3856414  5.364359 0.0223215

$`seno$A:seno$B`
                                                 diff        lwr        upr     p adj
normalni:bez hnojeni-kysela:bez hnojeni          0.75 -3.6338199  5.1338199 0.9934178
kysela:chlevska mrva-kysela:bez hnojeni          4.75  0.3661801  9.1338199 0.0293348
normalni:chlevska mrva-kysela:bez hnojeni        7.75  3.3661801 12.1338199 0.0003078
kysela:vapenate hnojivo-kysela:bez hnojeni       11.25  6.8661801 15.6338199 0.0000025
normalni:vapenate hnojivo-kysela:bez hnojeni      7.00  2.6161801 11.3838199 0.0009516
kysela:chlevska mrva-normalni:bez hnojeni         4.00 -0.3838199  8.3838199 0.0856265
normalni:chlevska mrva-normalni:bez hnojeni       7.00  2.6161801 11.3838199 0.0009516
kysela:vapenate hnojivo-normalni:bez hnojeni     10.50  6.1161801 14.8838199 0.0000065
normalni:vapenate hnojivo-normalni:bez hnojeni    6.25  1.8661801 10.6338199 0.0029984
normalni:chlevska mrva-kysela:chlevska mrva       3.00 -1.3838199  7.3838199 0.2959479
kysela:vapenate hnojivo-kysela:chlevska mrva      6.50  2.1161801 10.8838199 0.0020426
normalni:vapenate hnojivo-kysela:chlevska mrva    2.25 -2.1338199  6.6338199 0.5900609
kysela:vapenate hnojivo-normalni:chlevska mrva    3.50 -0.8838199  7.8838199 0.1649328
normalni:vapenate hnojivo-normalni:chlevska mrva -0.75 -5.1338199  3.6338199 0.9934178
normalni:vapenate hnojivo-kysela:vapenate hnojivo -4.25 -8.6338199  0.1338199 0.0604845
```

Při zkoumání vlivu faktoru B vidíme, že se na hladině významnosti liší všechny tři dvojice způsobu hnojení.
Při zkoumání vlivu interakcí na výnosy sena vidíme, že ze zkoumaných 15 dvojic se na hladině významnosti 0,05 liší 8.

Nakonec vykreslíme graf závislosti průměrného výnosu sena na způsobu hnojení pro různé typy půdy:

```
> interaction.plot(x.factor     = seno$B,
+                  trace.factor = seno$A,
+                  response     = seno$X,
+                  fun = mean,
+                  type="b",
+                  col=c("red","blue"),
+                  pch=c(17, 15),
+                  fixed=TRUE,
+                  leg.bty = "o")
```