

# Analýza rozptylu

## Ekonometrie

Jiří Neubauer

Katedra kvantitativních metod FVL UO Brno  
kancelář 69a, tel. 973 442029  
email: Jiri.Neubauer@unob.cz

# Analýza rozptylu – ANOVA

Analýza rozptylu je nástroj pro zkoumání vztahu mezi vysvětlovanými a vysvětlujícími proměnnými. Vysvětlované proměnné jsou vždy kvantitativní, u vysvětlujících proměnných (označují se jako **faktory**) na typu nezáleží. Faktory nabývají pouze malého počtu obměn (úrovní), podle nichž lze hodnoty vysvětlovaných proměnných třídít do skupin.

- **jednofaktorová ANOVA** – vliv jednoho faktoru na vysvětlovanou proměnnou
- **vícefaktorová ANOVA** – vliv více faktorů (dvojné, trojné třídění, atd.)
- **vícerozměrná analýza rozptylu – MANOVA** – vliv jednoho či více faktorů na několik vysvětlovaných proměnných současně

## Jednofaktorová ANOVA

Prokázat závislost vysvětlované proměnné  $Y$  (kvantitativní proměnná) na vysvětlujících proměnných (faktorech), znamená prokázat rozdílné úrovně proměnné  $Y$  v jednotlivých podsouborech – skupinách, vzniklých tříděním podle faktoru  $A$ . Předpokládejme, že máme  $k$  nezávislých náhodných výběrů z normálních rozdělení  $N(\mu_1, \sigma^2), N(\mu_2, \sigma^2), \dots, N(\mu_k, \sigma^2)$ , kde  $\mu_1, \mu_2, \dots, \mu_k$  jsou střední hodnoty veličiny  $Y$  v jednotlivých skupinách. Testujeme hypotézu

$$H : \mu_1 = \mu_2 = \dots = \mu_k$$

proti alternativě

$$A : \mu_i \neq \mu_j, \text{ pro alespoň jednu dvojici } i \neq j, i = 1, 2, \dots, k, j = 1, 2, \dots, k,$$

kteřá znamená, že alespoň některá rovnost mezi středními hodnotami neplatí.

Východiskem jsou naměřené hodnoty proměnné  $Y$  rozříděné do  $k$  skupin podle úrovní (variant) faktoru  $A$ . Náhodný výběr z rozdělení  $N(\mu_i, \sigma^2)$  označíme  $Y_{i1}, Y_{i2}, \dots, Y_{i, n_i}$ , kde  $n_i$  je jeho rozsah. Všechny výběry potom mají rozsah  $n = n_1 + n_2 + \dots + n_k = \sum_{i=1}^k n_i$ .

# Jednofaktorová ANOVA

V jednotlivých skupinách určíme tzv. **podmíněné charakteristiky**

- podmíněný průměr v  $i$ -té skupině  $\bar{Y}_{i.} = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}$
- podmíněný rozptyl v  $i$ -té skupině  $S_{i.}^2 = \frac{1}{n_i-1} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i.})^2$

Označme

- celkový průměr  $\bar{Y}_{..} = \frac{1}{n} \sum_{i=1}^k \sum_{j=1}^{n_i} Y_{ij}$
- celkový rozptyl  $S_{..}^2 = \frac{1}{n-1} \sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{..})^2$

Dále definujeme

- rozptyl podmíněných průměrů  $S_{\bar{Y}_{i.}}^2 = \frac{1}{n-1} \sum_{i=1}^k (\bar{Y}_{i.} - \bar{Y}_{..})^2 n_i$
- průměr podmíněných rozptylů  $\bar{S}_{i.}^2 = \frac{1}{n} \sum_{i=1}^k S_{i.}^2 n_i$

# Jednofaktorová ANOVA

**Příklad:** U 42 zákrsku jabloní bylo zaznamenáno stáří stromu v letech (faktor  $A$ ) a roční sklizeň (znak  $Y$ ).

$x_i$	$y_j$	$n_i$	$\bar{Y}_i$	$S_{i.}^2$	$(n_i - 1)S_{i.}^2$
3	4 7 5 5 5	5	5,200	1,200	4,800
4	9 5 7 6 8 7 8	7	7,143	1,810	10,857
5	9 8 9 10 7 7	6	8,333	1,467	7,333
6	10 8 10 10 10 9	6	9,500	0,700	3,500
7	9 7 8 9 10 9	6	8,667	1,067	5,333
8	8 7 7 8 6 10	6	7,667	1,867	9,333
9	5 4 6 7 6 8	6	6,000	2,000	10,000
$\Sigma$		42			51,157

## Jednofaktorová ANOVA

Podstatou analýzy rozptylu je rozklad celkového rozptylu na složku objasněnou – známý zdroj variability, a na složku neobjasněnou – reziduální, chybovou, o níž se předpokládá, že je náhodná. Zřejmě platí

$$Y_{ij} - \bar{Y}_{..} = (Y_{ij} - \bar{Y}_{i.}) + (\bar{Y}_{i.} - \bar{Y}_{..}).$$

Umocněním a sumací přes  $i$  a  $j$  dostáváme

$$\sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{..})^2 = \sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i.})^2 + 2 \sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i.})(\bar{Y}_{i.} - \bar{Y}_{..}) + \sum_{i=1}^k \sum_{j=1}^{n_i} (\bar{Y}_{i.} - \bar{Y}_{..})^2.$$

Lze ukázat, že člen  $2 \sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i.})(\bar{Y}_{i.} - \bar{Y}_{..})$  je roven nule. Potom platí

$$\underbrace{\sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{..})^2}_{S_c} = \underbrace{\sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i.})^2}_{S_e} + \underbrace{\sum_{i=1}^k \sum_{j=1}^{n_i} (\bar{Y}_{i.} - \bar{Y}_{..})^2}_{S_A}.$$

- $S_c$  značí **celkovou variabilitu**,  $S_c = \sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{..})^2 = (n - 1)S_{..}^2$ ,
- $S_A$  je **variabilita vysvětlená faktorem A**,  
 $S_A = \sum_{i=1}^k \sum_{j=1}^{n_i} (\bar{Y}_{i.} - \bar{Y}_{..})^2 = \sum_{i=1}^k n_i (\bar{Y}_{i.} - \bar{Y}_{..})^2 = (n - 1)S_{\bar{Y}_{i.}}^2$ ,
- $S_e$  je **variabilita reziduální (vnitřní)**,  $S_e = \sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i.})^2 = \sum_{i=1}^k (n_i - 1)S_{i.}^2$ .

# Jednofaktorová ANOVA

Z analyzovaného datového souboru dostáváme:

- $S_A = 77,248$
- $S_e = 51,157$
- $S_c = 128,405$

## Jednofaktorová ANOVA

Důležitým předpokladem použití analýzy rozptylu je, že každý z  $k$  nezávislých výběrů (což odpovídá k řádkům v korelační tabulce) proměnné  $Y$  pochází z normálního rozdělení  $N(\mu_i, \sigma^2)$  se stejným rozptylem  $\sigma^2$ .

- Předpoklad normality lze ověřit testy normality, avšak s přihlédnutím k rozsahům výběrů se v praxi se od toho často upouští a posuzuje se pouze, zda se ve skupinách hodnot proměnné  $Y$ , zjištěných na jednotlivých úrovních faktoru  $X$ , nevyskytují výslovně extrémní hodnoty a zda se hodnoty blízké podmíněným průměrům vyskytují častěji než hodnoty, jejichž vzdálenost od podmíněných průměrů je větší.
- K ověření hypotézy o stejných rozptylech  $k$  normálních rozdělení lze použít **Bartlettův test** (je velmi citlivý na porušení předpokladu normality), lze použít i jiné testy, např. **Hartleyův** nebo **Cochranův test** (předp. se stejné četnosti ve třídách) případně **Fligner-Killeenův test**.



## Jednofaktorová ANOVA

Jestliže k nezávislých výběrů pochází z normálních rozdělení se stejnými rozptyly, lze kolísání – variabilitu podmíněných průměrů interpretovat jako závislost proměnné  $Y$  na faktoru  $X$ , zatím co kolísání hodnot proměnné  $Y$  uvnitř jednotlivých skupin budeme vnímat jako závislosti proměnné  $Y$  na dalších činitelích (v analýze nesledovaných).

**Koeficient determinace  $p^2$**  je definován vztahem

$$p^2 = \frac{S_A}{S_c}.$$

- $p^2 \in \langle 0, 1 \rangle$ ,
- udává, jaké % rozptylu závisle proměnné  $Y$  lze vysvětlit vlivem faktoru  $A$ ,
- neshoda mezi středními hodnotami  $\mu_i, i = 1, \dots, k$  se považuje za tím silnější, čím více se  $p^2$  blíží k 1 a naopak

# Jednofaktorová ANOVA

Test o shodě podmíněných středních hodnot:

$$H : \mu_1 = \mu_2 = \dots = \mu_k$$

$$A : \mu_i \neq \mu_j \text{ pro nějaké } i, j = 1, \dots, k, i \neq j$$

Testové kritérium je statistika

$$F = \frac{\frac{S_A}{k-1}}{\frac{S_e}{n-k}} = \frac{(n-k) \cdot S_A}{(k-1) \cdot S_e},$$

kteří má při platnosti hypotézy  $H$  Fisherovo-Snedecorovo rozdělení  $F(k-1, n-k)$ .

Kritický obor je dán  $W_\alpha : F \geq F_{1-\alpha}(k-1, n-k)$ .

## Jednofaktorová ANOVA

Z analyzovaných dat (velikost sklizně v závislosti na stáří stromu) jsme získali následující údaje:  $n = 42$ , počet skupin (hodnot faktorů)  $k = 7$ ,  $S_A = 77,248$ ,  $S_e = 51,157$  a  $S_c = 128,405$ .

Koeficient determinace má hodnotu

$$p^2 = \frac{S_A}{S_c} = 0,602.$$

Budeme testovat hypotézu (na hladině významnosti 0,05)  $H : \mu_1 = \mu_2 = \dots = \mu_7$   
 $A : \mu_i \neq \mu_j$  pro nějaké  $i, j = 1, \dots, 7, i \neq j$

$$F = \frac{\frac{S_A}{k-1}}{\frac{S_e}{n-k}} = \frac{(n-k) \cdot S_A}{(k-1) \cdot S_e} = \frac{(42-7) \cdot 77,248}{(7-1) \cdot 51,157} = 8,808.$$

Kritický obor je  $W_\alpha : 8,808 \geq F_{0,95}(6, 35) = 2,372$ , na hladině významnosti 0,05 zamítáme nulovou hypotézu o rovnosti středních hodnot. S pravděpodobností 95 % můžeme tvrdit, že stáří stromu ovlivňuje velikost sklizně.

# Jednofaktorová ANOVA

Možnosti výpočtu:

- Excel – Analýza dat – Anova: jeden faktor
- R – funkce aov, (Bartlettův test – bartlett.test, Fligner-Killeenův test – fligner.test)

## Jednofaktorová ANOVA v Excelu

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	stáří stromu	3	4	5	6	7	8	9		Anova: jeden faktor						
2		4	9	9	10	9	8	5								
3		7	5	8	8	7	7	4		Faktor						
4	sklizeň	5	7	9	10	8	7	6		<i>Výběr</i>	<i>Počet</i>	<i>Součet</i>	<i>Průměr</i>	<i>Rozptyl</i>		
5		5	6	10	10	9	8	7		3	5	26	5,2	1,2		
6		5	8	7	10	10	6	6		4	7	50	7,142857	1,809524		
7		7	7	9	9	10	8			5	6	50	8,333333	1,466667		
8		8								6	6	57	9,5	0,7		
9										7	6	52	8,666667	1,066667		
10										8	6	46	7,666667	1,866667		
11										9	6	36	6	2		
12																
13																
14										ANOVA						
15										<i>Zdroj variability</i>	<i>SS</i>	<i>Rozdíl</i>	<i>MS</i>	<i>F</i>	<i>Hodnota P</i>	<i>F krit</i>
16										Mezi výběry	77,247619	6	12,8746	8,808371	7,1E-06	2,371781
17										Všechny výběry	51,157143	35	1,461633			
18										Celkem	128,40476	41				
19																
20																

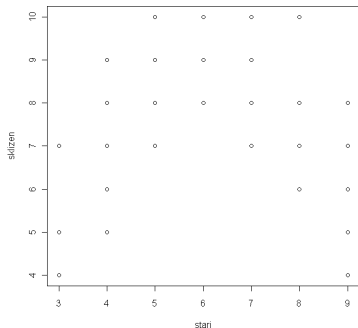
## Jednofaktorová ANOVA v R

Datový soubor `anova_sklizen.txt` obsahuje 2 sloupce se záhlavím "stari" a "sklizen"

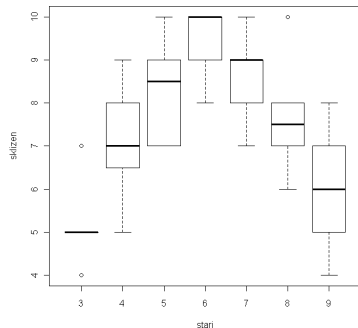
```
> data<-read.table("anova_sklizen.txt",header=T)
> attach(data)
> names(data)
> stari<-factor(stari)
> summary(aov(sklizen ~ stari))
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
stari	6	77.248	12.875	8.8084	7.104e-06	***
Residuals	35	51.157	1.462			

## Jednofaktorová ANOVA v R



Obrázek: Bodový diagram



Obrázek: Krabicové diagramy – boxplot

## Jednofaktorová ANOVA v R

Ověření předpokladu homoskedasticity (stejné rozptyly ve všech skupinách) je možné provést pomocí Bartlettova nebo Fligner-Killeenůvova testu.

```
> bartlett.test(sklizen~stari)
```

Bartlett test of homogeneity of variances

data: sklizen by stari

Bartlett's K-squared = 1.8159, df = 6, p-value = 0.9358

```
> fligner.test(sklizen~stari)
```

Fligner-Killeen test of homogeneity of variances

data: sklizen by stari

Fligner-Killeen:med chi-squared = 2.9335, df = 6, p-value = 0.8171

Předpoklad homoskedasticity je přijatelný.



## Mnohonásobné porovnávání v R – Tukeyho metoda

Jedná se v podstatě o řadu dvouvýběrových  $t$ -testů, u nichž je upravena hladina významnosti

```
> TukeyHSD(aov(sklizen~stari))
> plot(TukeyHSD(aov(sklizen~stari)))
```

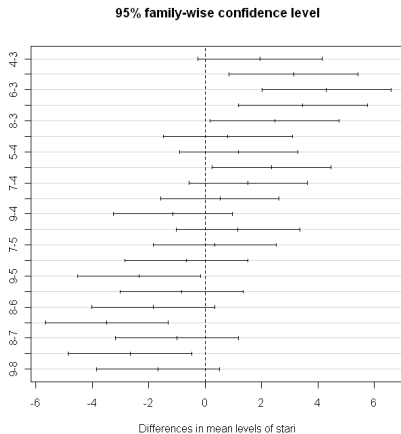
Tukey multiple comparisons of means

95% family-wise confidence level

Fit: aov(formula = sklizen ~ stari)

	diff	lwr	upr	p adj
4-3	1.9428571	-0.2700120	4.1557263	0.1169792
5-3	3.1333333	0.8449180	5.4217487	0.0024115
6-3	4.3000000	2.0115847	6.5884153	0.0000220
7-3	3.4666667	1.1782513	5.7550820	0.0006503
8-3	2.4666667	0.1782513	4.7550820	0.0277023
⋮	⋮	⋮	⋮	⋮
9-7	-2.6666667	-4.8485851	-0.4847483	0.0085912
9-8	-1.6666667	-3.8485851	0.5152517	0.2339036

# Mnohonásobné porovnávání v R – Tukeyho metoda



Obrázek: Mnohonásobné porovnávání – Tukeyho metoda

## Dvoufaktorová ANOVA

Často je třeba zkoumat závislost kvantitativní proměnné na více faktorech. omezíme se na případ dvou faktorů.

Možnosti výpočtu:

- Excel – Analýza dat – Anova: dva faktory s opakováním, dva faktory bez opakování
- R – funkce aov

**Příklad:** Cílem experimentu je zkoumat vliv dvou typů benzínu a tří různých aditiv na spotřebu automobilu. Výsledky jsou uvedeny v tabulce.

<i>Aditivum</i>	<i>Typ</i>	
	B1	B2
A1	8,58	7,06
	8,22	6,82
A2	7,13	6,61
	7,35	6,84
A3	7,02	7,04
	7,28	7,11

# Dvoufaktorová ANOVA

- Budeme se zabývat vlivem dvou vysvětlujících proměnných (faktorů  $A$ ,  $B$ ) na proměnnou vysvětlovanou  $Y$ .
- Označme  $a$  počet úrovní faktoru  $A$ , podobně  $b$  bude označovat počet úrovní faktoru  $B$ .
- Předpokládejme, že pro každou dvojici hodnot faktorů máme  $r \geq 2$  pozorování.
- Pro pozorování s  $i$ -tou hodnotou faktoru  $A$  a  $j$ -tou hodnotou faktoru  $B$  platí

$$Y_{ij1}, \dots, Y_{ijr} \sim N(\mu_{ij}, \sigma^2).$$

## Dvoufaktorová ANOVA bez interakcí

Označme:

$$\bar{Y}_{ij.} = \frac{1}{r} \sum_{k=1}^r Y_{ijk},$$

$$\bar{Y}_{i..} = \frac{1}{br} \sum_{j=1}^b \sum_{k=1}^r Y_{ijk},$$

$$\bar{Y}_{.j.} = \frac{1}{ar} \sum_{i=1}^a \sum_{k=1}^r Y_{ijk},$$

$$\bar{Y}_{...} = \frac{1}{abr} \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^r Y_{ijk}$$

Pro celkovou variabilitu lze psát

$$S_c = S_A + S_B + S_e.$$

- $S_c = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^r (Y_{ijk} - \bar{Y}_{...})^2$  – celková variabilita
- $S_A = br \sum_{i=1}^a (\bar{Y}_{i..} - \bar{Y}_{...})^2$  – variabilita vysvětlenou faktorem A
- $S_B = ar \sum_{j=1}^b (\bar{Y}_{.j.} - \bar{Y}_{...})^2$  – variabilita vysvětlenou faktorem B
- $S_e = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^r (Y_{ijk} - \bar{Y}_{ij.})^2$  – variabilita reziduální

## Dvoufaktorová ANOVA bez interakcí

Při dvoufaktorové analýze rozptylu bez interakcí nás bude zajímat vlivu faktorů  $A$  a  $B$  na vysvětlovanou proměnnou  $Y$  viz tabulka.

Zdroj variability	Součet čtverců	Stupně volnosti	Testová statistika
Faktor $A$	$S_A$	$f_A = a - 1$	$F_A = \frac{S_A/f_A}{S_e/f_e}$
Faktor $B$	$S_B$	$f_B = b - 1$	$F_B = \frac{S_B/f_B}{S_e/f_e}$
Reziduální	$S_e$	$f_e = n - a - b + 1$	–
Celkový	$S_c$	$f_c = n - 1$	–

Kritické hodnoty jednotlivých testů jsou kvantily rozdělení  $F$ . Vliv faktoru  $A$  je statisticky významný, je-li  $F_A \geq F_{1-\alpha}(f_A, f_e)$ , podobně vliv faktoru  $B$  je významný, pokud  $F_B \geq F_{1-\alpha}(f_B, f_e)$ .

## Dvoufaktorová ANOVA s interakcemi

Označme:

$$\bar{Y}_{ij.} = \frac{1}{r} \sum_{k=1}^r Y_{ijk},$$

$$\bar{Y}_{i..} = \frac{1}{br} \sum_{j=1}^b \sum_{k=1}^r Y_{ijk},$$

$$\bar{Y}_{.j.} = \frac{1}{ar} \sum_{i=1}^a \sum_{k=1}^r Y_{ijk},$$

$$\bar{Y}_{...} = \frac{1}{abr} \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^r Y_{ijk}$$

Pro celkovou variabilitu lze psát

$$S_c = S_A + S_B + S_{AB} + S_e.$$

- $S_c = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^r (Y_{ijk} - \bar{Y}_{...})^2$  – **celková variabilita**
- $S_A = br \sum_{i=1}^a (\bar{Y}_{i..} - \bar{Y}_{...})^2$  – **variabilita vysvětlenou faktorem A**
- $S_B = ar \sum_{j=1}^b (\bar{Y}_{.j.} - \bar{Y}_{...})^2$  – **variabilita vysvětlenou faktorem B**
- $S_{AB} = r \sum_{i=1}^a \sum_{j=1}^b (\bar{Y}_{ij.} - \bar{Y}_{i..} - \bar{Y}_{.j.} + \bar{Y}_{...})^2$  – **variabilita příslušející interakci**
- $S_e = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^r (Y_{ijk} - \bar{Y}_{ij.})^2 = S_c - S_A - S_B - S_{AB}$  – **variabilita reziduální**

## Dvoufaktorová ANOVA s interakcemi

Při dvoufaktorové analýze rozptylu s interakcemi nás bude zajímat kromě vlivu faktorů  $A$  a  $B$  na vysvětlovanou proměnnou  $Y$ , také vliv interakce obou faktorů – viz tabulka.

Zdroj variability	Součet čtverců	Stupně volnosti	Testová statistika
Faktor $A$	$S_A$	$f_A = a - 1$	$F_A = \frac{S_A/f_A}{S_e/f_e}$
Faktor $B$	$S_B$	$f_B = b - 1$	$F_B = \frac{S_B/f_B}{S_e/f_e}$
Interakce	$S_{AB}$	$f_{AB} = (a - 1)(b - 1)$	$F_{AB} = \frac{S_{AB}/f_{AB}}{S_e/f_e}$
Reziduální	$S_e$	$f_e = n - ab$	–
Celkový	$S_c$	$f_c = n - 1$	–

Kritické hodnoty jednotlivých testů jsou kvantily rozdělení  $F$ . Vliv faktoru  $A$  je statisticky významný, je-li  $F_A \geq F_{1-\alpha}(f_A, f_e)$ , podobně vliv faktoru  $B$  je významný, pokud  $F_B \geq F_{1-\alpha}(f_B, f_e)$ . Mezi faktory  $A$  a  $B$  je významná interakce když  $F_{AB} \geq F_{1-\alpha}(f_{AB}, f_e)$ .



## Dvoufaktorová ANOVA s interakcemi

Pro data z příkladu dostaneme ( $a = 3$ ,  $b = 2$ ,  $r = 2$ )

$$S_c = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^r (y_{ijk} - \bar{y}_{...})^2 = 3,650,$$

$$S_A = br \sum_{i=1}^a (\bar{y}_{i..} - \bar{y}_{...})^2 = 1,067,$$

$$S_B = ar \sum_{j=1}^b (\bar{y}_{.j.} - \bar{y}_{...})^2 = 1,401,$$

$$S_{AB} = r \sum_{i=1}^a \sum_{j=1}^b (\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...})^2 = 1,002,$$

$$S_e = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^r (y_{ijk} - \bar{y}_{ij.})^2 = S_c - S_A - S_B - S_{AB} = 0,181.$$

Testovací statistiky:  $F_A = 17,737$ ,  $F_B = 46,565$ ,  $F_{AB} = 16,647$ .

Kritické hodnoty pro hladinu významnosti 0,05 jsou postupně:  $F_{0,95}(2, 6) = 5,143$ ,  
 $F_{0,95}(1, 6) = 5,987$ ,  $F_{0,95}(2, 6) = 5,143$ .

## Dvoufaktorová ANOVA v Excelu

	A	B	C	D	E	F	G	H	I	J	K	L
1		A1	A2	A3		Anova: dva faktory s opakováním						
2	B1	8,58	7,13	7,02								
3		8,22	7,35	7,28		Faktor	A1	A2	A3	Celkem		
4	B2	7,06	6,61	7,04		<i>B1</i>						
5		6,82	6,84	7,11		Počet	2	2	2	6		
6						Součet	16,8	14,48	14,3	45,58		
7						Průměr	8,4	7,24	7,15	7,596667		
8						Rozptyl	0,0648	0,0242	0,0338	0,413387		
9						<i>B2</i>						
10						Počet	2	2	2	6		
11						Součet	13,88	13,45	14,15	41,48		
12						Průměr	6,94	6,725	7,075	6,913333		
13						Rozptyl	0,0288	0,02645	0,00245	0,036467		
14						<i>Celkem</i>						
15						Počet	4	4	4			
16						Součet	30,68	27,93	28,45			
17						Průměr	7,67	6,9825	7,1125			
18						Rozptyl	0,741733	0,105292	0,013958			
19												
20												
21												
22												
23						ANOVA						
24						<i>Zdroj variability</i>	<i>SS</i>	<i>Rozdíl</i>	<i>MS</i>	<i>F</i>	<i>Hodnota P</i>	<i>F krit</i>
25						Výběr	1,400833	1	1,400833	46,5651	0,000486	5,987378
26						Sloupce	1,06715	2	0,533575	17,73657	0,003028	5,143253
27						Interakce	1,001617	2	0,500808	16,64737	0,00356	5,143253
28						Dohromady	0,1805	6	0,030083			
29												
30						Celkem	3,6501	11				
31												

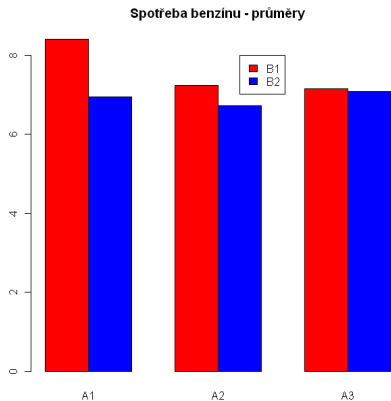
## Dvoufaktorová ANOVA v R

Datový soubor `anova2_spotreba.txt` obsahuje 3 sloupce se záhlavím "typ", "aditivum" a "spotreba"

```
> data<-read.table("anova2_spotreba.txt",header=T)
> attach(data)
> names(data)
> tapply(spotreba,list(aditivum,spotreba),mean)
```

	B1	B2
A1	8.40	6.940
A2	7.24	6.725
A3	7.15	7.075

## Dvoufaktorová ANOVA v R



Obrázek: Průměrná spotřeba benzínu v závislosti na typ benzínu a aditivu

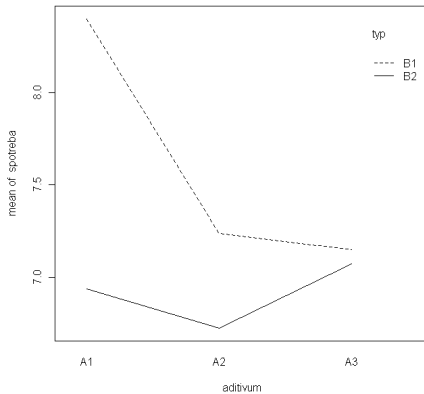
## Dvoufaktorová ANOVA v R

```
> data<-read.table("anova2_spotreba.txt",header=T)
> model <-aov(spotreba ~ aditivum*typ)
> summary(model)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
aditivum	2	1.06715	0.53358	17.737	0.0030280	**
typ	1	1.40083	1.40083	46.565	0.0004861	***
aditivum:typ	2	1.00162	0.50081	16.647	0.0035600	**
Residuals	6	0.18050	0.03008			

Na základě vypočtených  $p$ -hodnot můžeme tvrdit, že vliv typu benzínu i aditiva na spotřebu byl prokázán. Vliv interakce byl také prokázán.

## Dvoufaktorová ANOVA v R



Obrázek: Interakce dvou faktorů