Analýza rozptylu

Statistika II

Jiří Neubauer

Katedra ekonometrie 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 třídit 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 faktorů X. Označíme-li střední hodnoty veličiny Y v jednotlivých skupinách $\mu_1, \mu_2, \ldots, \mu_k$, testujeme hypotézu

$$H : \mu_1 = \mu_2 = \cdots = \mu_k \quad \text{proti alternativě} \quad A : \text{non } H,$$

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

Východiskem jsou naměřené hodnoty proměnné Y roztříděné do k skupin podle úrovní – variant faktoru X,

Jednofaktorová ANOVA

Každý řádek korelační tabulky obsahuje rozdělení četností hodnot znaku Y za podmínky, že znak X nabyl určité obměny, tj. obsahuje podmíněné rozdělení četností hodnot znaku Y, které lze popsat pomocí tzv. **podmíněných charakteristik**

- podmíněný průměr v i-té skupině $\overline{y_i} = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}$
- podmíněný rozptyl v i-té skupině $s^2_{n,i}(y) = \frac{1}{n_i} \sum_{j=1}^{n_i} (y_{ij} - \overline{y_i})^2$
- rozptyl podmíněných průměrů $s^2_n(\overline{y_i}) = \frac{1}{n} \sum_{i=1}^{k} (\overline{y_i} - \overline{y})^2 n_i$
- průměr podmíněných rozptylů $\overline{s^2_{n,i}(y)} = \frac{1}{n} \sum_{i=1}^{k} s^2_{n,i}(y) n_i$
- celkový průměr $\overline{y} = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} y_{ij}$
- celkový rozptyl $s^2_n(y) = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \overline{y})^2$

Jednofaktorová ANOVA

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

x_i	y_j							n_i	$\overline{y_i}$	$s^2_{n,i}(y)$	$S_i(y)$
3	4	7	5	5	5			5	5,200	0,960	4,800
4	9	5	7	6	8	7	8	7	7,143	1,551	10,857
5	9	8	9	10	7	7		6	8,333	1,222	7,333
6	10	8	10	10	10	9		6	9,500	0,583	3,500
7	9	7	8	9	10	9		6	8,667	0,889	5,333
8	8	7	7	8	6	10		6	7,667	1,556	9,333
9	5	4	6	7	6	8		6	6,000	1,667	10,000
$\sum$								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.
Pro celkový rozptyl platí

$$s_n^2(y) = s_n^2(\overline{y_i}) + \overline{s_{n,i}^2(y)}$$

neboli

$$S_c(y) = S_m(y) + S_v(y),$$

kde

- $S_m(y) = n \cdot s_n^2(\overline{y_i})$ je součet čtverců, který představuje **meziskupinovou** – vysvětlenou **variabilitu** proměnné Y,
- $S_v(y) = \sum_{i=1}^{k} S_i(y) = \sum_{i=1}^{k} n_i \cdot s_{n,i}^2(y)$ je součet čtverců, který představuje **vnitroskupinovou** – nevysvětlenou, chybovou, reziduální **variabilitu** proměnné Y,
- $S_c(y) = n \cdot s_n^2(y)$ je součet čtverců, který představuje **celkovou variabilitu** proměnné Y.

Jednofaktorová ANOVA

Z analyzovaného datového souboru dostáváme:

- $S_m(y) = n \cdot s_n^2(\overline{y_i}) = 77{,}248$
- $S_v(y) = \sum_{i=1}^{k} S_i(y) = \sum_{i=1}^{k} n_i \cdot s_{n,i}^2(y) = 51{,}157$
- $S_c(y) = n \cdot s_n^2(y) = 128{,}405$

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 σ^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).

Definice

Koeficient determinace p^2_{yx} je definován vztahem

$$p^2_{yx} = \frac{s^2_n(\overline{y_i})}{s^2_n(y)} = \frac{S_m(y)}{S_c(y)}.$$

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

Jednofaktorová ANOVA

Test o shodě podmíněných středních hodnot:
$H : \mu_1 = \mu_2 = \cdots = \mu_k$
$A : \mu_i \neq \mu_j$ pro nějaké $i, j = 1, \ldots, k, i \neq j$
Testové kritérium je statistika

$$F = \frac{\frac{S_m(y)}{k-1}}{\frac{S_v(y)}{n-k}} = \frac{(n-k) \cdot S_m(y)}{(k-1) \cdot S_v(y)},$$

které 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) $k = 7$, $S_m(y) = 77{,}248$, $S_v(y) = 51{,}157$ a $S_c(y) = 128,405$.

Koeficient determinace má hodnotu

$$p_{yx}^2 = \frac{S_m(y)}{S_c(y)} = 0{,}602.$$

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

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

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 – bartltett.test, Fligner-Killeenův test – fligner.test)

Jednofaktorová ANOVA v Excelu

	A	B	C	D	E	F	G	H
1	stáří stromu	**3**	**4**	**5**	**6**	**7**	**8**	**9**
2		4	9	9	10	9	8	5
3		7	5	8	8	7	7	4
4	sklizeň	5	7	9	10	8	7	6
5		5	6	10	10	9	8	7
6		5	8	7	10	10	6	6
7			7	7	9	9	10	8
8			8					

Anova: jeden faktor

Faktor

Výběr	*Počet*	*Součet*	*Průměr*	*Rozptyl*
3	5	26	5,2	1,2
4	7	50	7,142857	1,809524
5	6	50	8,333333	1,466667
6	6	57	9,5	0,7
7	6	52	8,666667	1,066667
8	6	46	7,666667	1,866667
9	6	36	6	2

ANOVA

Zdroj variability	*SS*	*Rozdíl*	*MS*	*F*	*Hodnota P*	*F krit*
Mezi výběry	77,247619	6	12,8746	8,808371	7,1E-06	2,371781
Všechny výběry	51,157143	35	1,461633			
Celkem	128,40476	41				

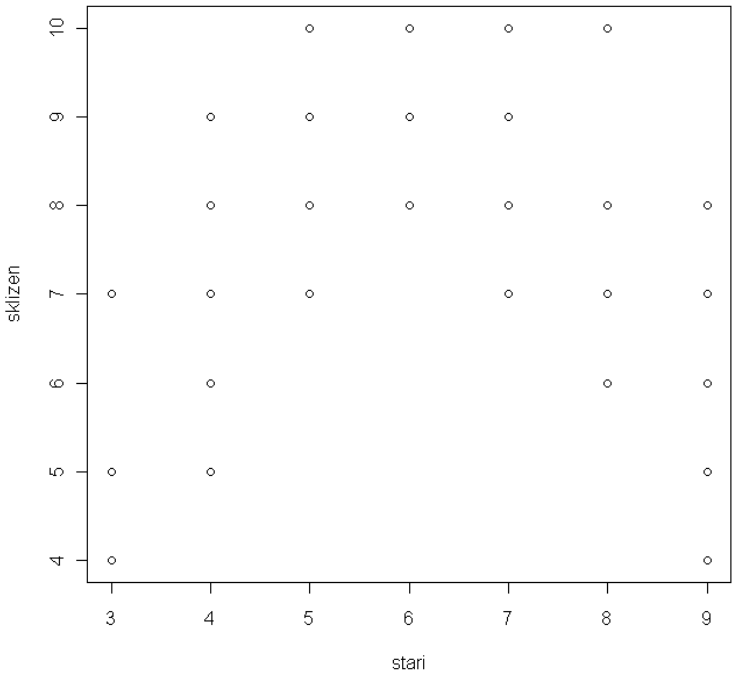
Jednofaktorová ANOVA v R

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

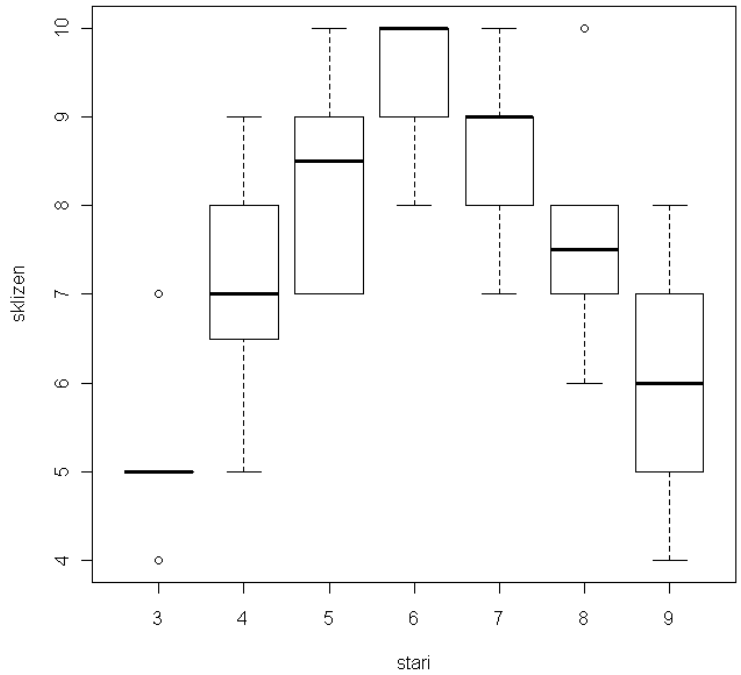
```
> data<-read.table("anova_sklizen.txt",header=T)
> attach(data)
> names(data)
> stari<-factor(stari)
> summary(aov(sklizen ~ stari))
```

	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

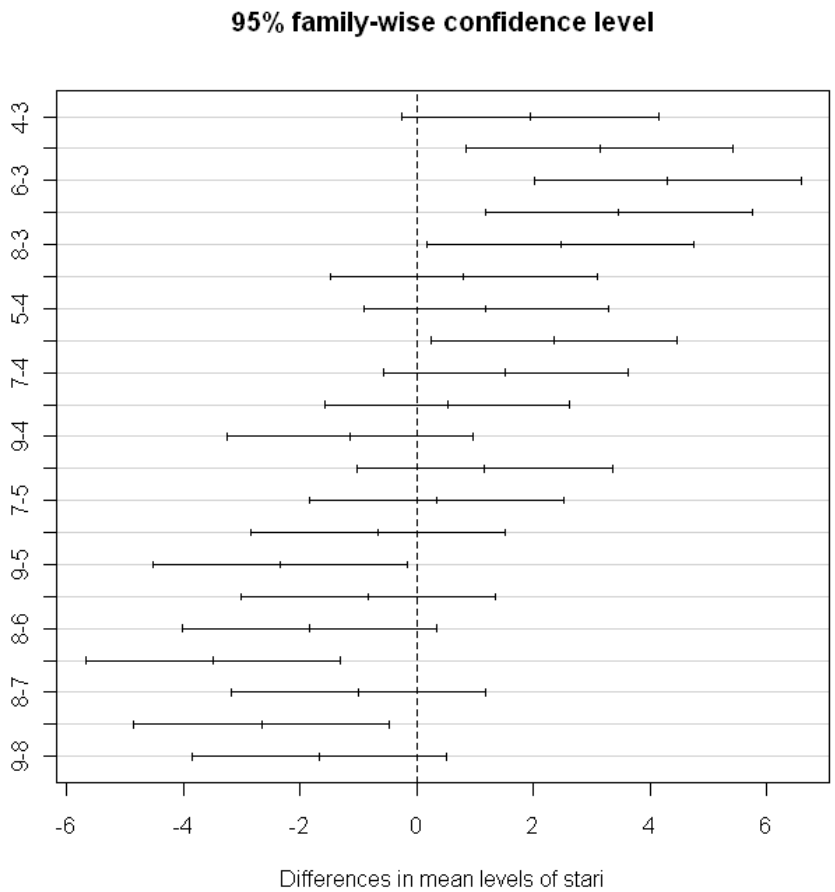
Jedná se v podstatě o řadu dvouvýběrových t-testů, u nichž je upravena hladina významnosti

```
> 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.

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

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

Dvoufaktorová ANOVA

Označme:

$\overline{y}_{ij.} = \frac{1}{r}\sum_{k=1}^{r} y_{ijk}$,
$\overline{y}_{i..} = \frac{1}{br}\sum_{j=1}^{b}\sum_{k=1}^{r} y_{ijk}$,
$\overline{y}_{.j.} = \frac{1}{ar}\sum_{i=1}^{a}\sum_{k=1}^{r} y_{ijk}$,
$\overline{y}_{...} = \frac{1}{abr}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{r} y_{ijk}$

Pro celkovou variabilitu lze psát

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

kde $S_c = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{r}(y_{ijk} - \overline{y}_{...})^2$,
$S_A = br\sum_{i=1}^{a}(\overline{y}_{i..} - \overline{y}_{...})^2$,
$S_B = ar\sum_{j=1}^{b}(\overline{y}_{.j.} - \overline{y}_{...})^2$,
$S_{AB} = r\sum_{i=1}^{a}\sum_{j=1}^{b}(\overline{y}_{ij.} - \overline{y}_{i..} - \overline{y}_{.j.} + \overline{y}_{...})^2$,
$S_e = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{r}(y_{ijk} - \overline{y}_{ij.})^2 = S_c - S_A - S_B - S_{AB}$.

Dvoufaktorová ANOVA

Při dvoufaktorové analýze rozptylu 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

Pro data z příkladu dostaneme ($a = 3$, $b = 2$, $r = 2$)
$S_c = \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{r} (y_{ijk} - \overline{y}_{...})^2 = 3{,}650,$
$S_A = br \sum_{i=1}^{a} (\overline{y}_{i..} - \overline{y}_{...})^2 = 1{,}067,$
$S_B = ar \sum_{j=1}^{b} (\overline{y}_{.j.} - \overline{y}_{...})^2 = 1{,}401,$
$S_{AB} = r \sum_{i=1}^{a} \sum_{j=1}^{b} (\overline{y}_{ij.} - \overline{y}_{i..} - \overline{y}_{.j.} + \overline{y}_{...})^2 = 1{,}006,$
$S_e = \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{r} (y_{ijk} - \overline{y}_{ij.})^2 = S_c - S_A - S_B - S_{AB} = 0{,}181.$

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

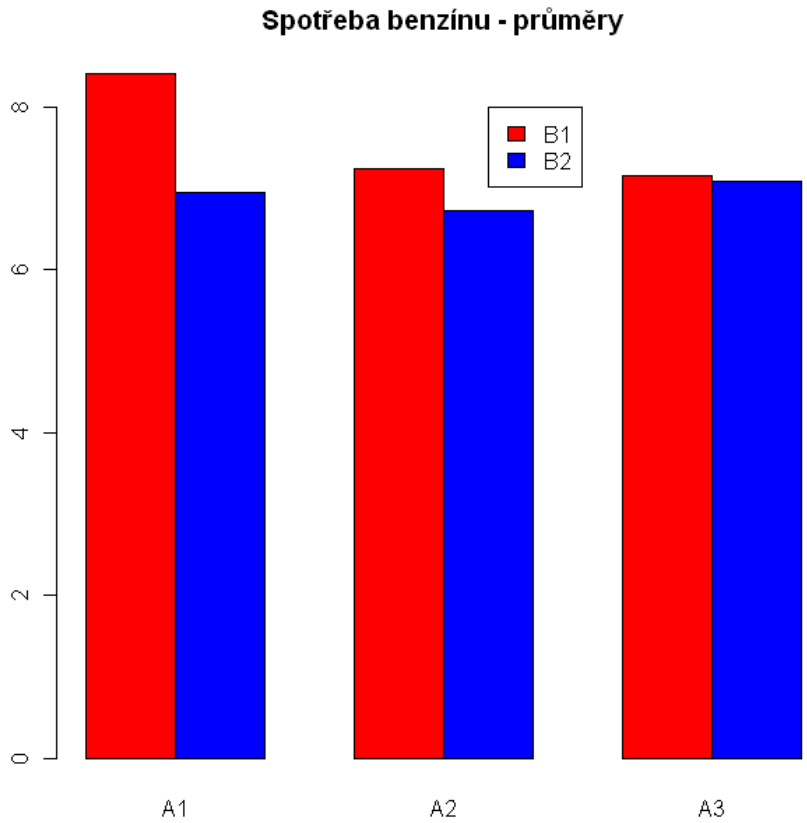
	A	B	C	D	E	F	G	H	I	J	K	L
1		A1	A2	A3		Anova: dva faktory s opakováním						
2	B1	8,58	7,13	7,02								
3		8,22	7,35	7,28		Faktor	A1	A2	A3	Celkem		
4	B2	7,06	6,61	7,04		*B1*						
5		6,82	6,84	7,11		Počet	2	2	2	6		
6						Součet	16,8	14,48	14,3	45,58		
7						Průměr	8,4	7,24	7,15	7,596667		
8						Rozptyl	0,0648	0,0242	0,0338	0,413387		
9												
10						*B2*						
11						Počet	2	2	2	6		
12						Součet	13,88	13,45	14,15	41,48		
13						Průměr	6,94	6,725	7,075	6,913333		
14						Rozptyl	0,0288	0,02645	0,00245	0,036467		
15												
16						*Celkem*						
17						Počet	4	4	4			
18						Součet	30,68	27,93	28,45			
19						Průměr	7,67	6,9825	7,1125			
20						Rozptyl	0,741733	0,105292	0,013958			
21												
22												
23						ANOVA						
24						*Zdroj variability*	*SS*	*Rozdíl*	*MS*	*F*	*Hodnota P*	*F krit*
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(typ,aditivum),mean)
       A1      A2      A3
 B1  8.40   7.240   7.150
 B2  6.94   6.725   7.075
> tapply(spotreba,list(typ,aditivum),var)
         A1        A2        A3
 B1  0.0648   0.02420   0.03380
 B2  0.0288   0.02645   0.00245
```

Dvoufaktorová ANOVA v R



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

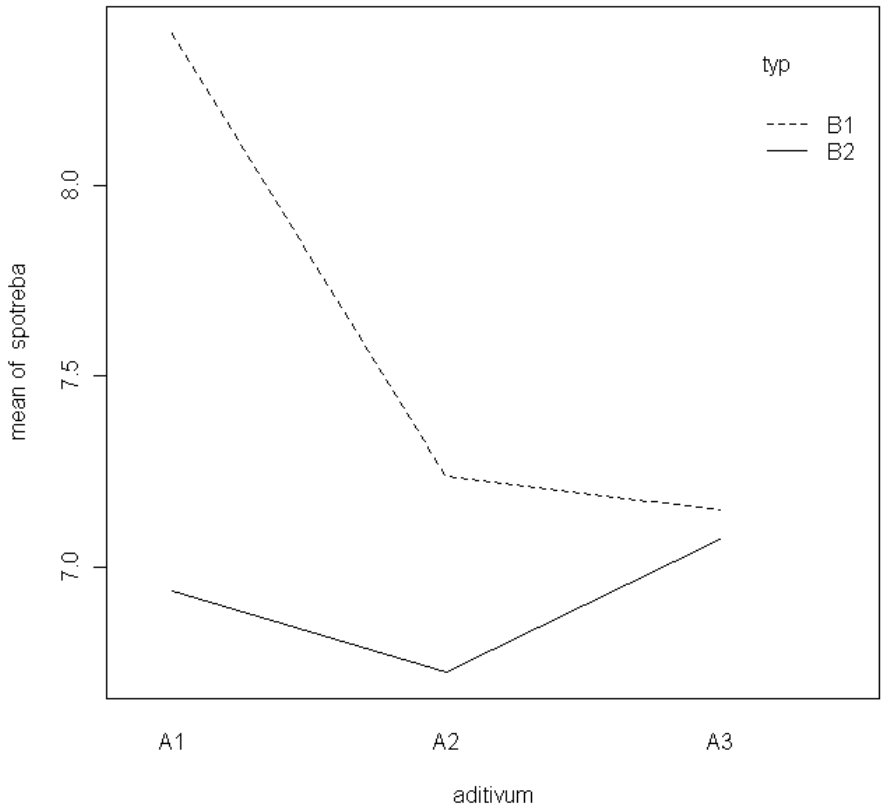
Dvoufaktorová ANOVA v R

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

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
typ	1	1.40083	1.40083	46.565	0.0004861	***
aditivum	2	1.06715	0.53358	17.737	0.0030280	**
typ:aditivum	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