Analýza rozptylu

Analýza rozptylu umožňuje ověřit významnost rozdílu mezi výběrovými průměry většího počtu náhodných výběrů, umožňuje posoudit vliv různých faktorů.

Podle počtu analyzovaných faktorů rozlišujeme jednofaktorovou, dvoufaktorovou a vícefaktorovou analýzu rozptylu.

Analýza rozptylu se často označuje akronymem ANOVA „ANalysis Of VAriance“.

Analýza rozptylu

Např. zjišťujeme vliv vzdělání (první nezávislý faktor *A*) a pohlaví (druhý nezávislý faktor *B*) na příjem (závislý kvantitativní faktor *Y*).
Nezávislé faktory jsou zpravidla kvalitativní (pohlaví, vzdělání) ale mohou být i kvantitativní (věk).

Cílem ANOVA je prokázat, že hodnoty znaků *A,B* - nezávislých faktorů, ovlivňují hodnoty kvantitativního znaku *Y* - závislého faktoru.

ANOVA je lepší alternativou pro *t*-test v případě, že porovnáváme víc než dva průměry.

Jednofaktorová ANOVA

Předpokládáme, že faktor *A* je pouze jeden a má *k* úrovní (hodnot x_i), s účinkem na znak *Y*, který lze vyjádřit vztahem:

$$\mu_i = \mu + \alpha_i$$

kde μ_i je průměr znaku *Y* v *i*-té úrovni,

μ je celkový průměr znaku *Y*,

α_i je vliv faktoru *A* na znak *Y* v *i*-té úrovni.

Předpokládáme, že hodnoty α_i pocházejí z normálně rozdělené populace s nulovou střední hodnotou a konstantním rozptylem.

Nulová hypotéza:

H_0: $\alpha_1 = \alpha_2 = ... = \alpha_k = 0$ resp. $\mu_1 = \mu_2 = ... = \mu_k$

Jednofaktorová ANOVA

Součet čtverců odchylek od celkového průměru μ:

$$S_c = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \mu)^2$$

lze rozložit na dvě složky:

$$S_c = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left((y_{ij} - \mu_i) + (\mu_i - \mu) \right)^2 = \underbrace{\sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \mu_i)^2}_{S_R} + \underbrace{\sum_{i=1}^{k} n_i (\mu_i - \mu)^2}_{S_A} = S_R + S_A$$

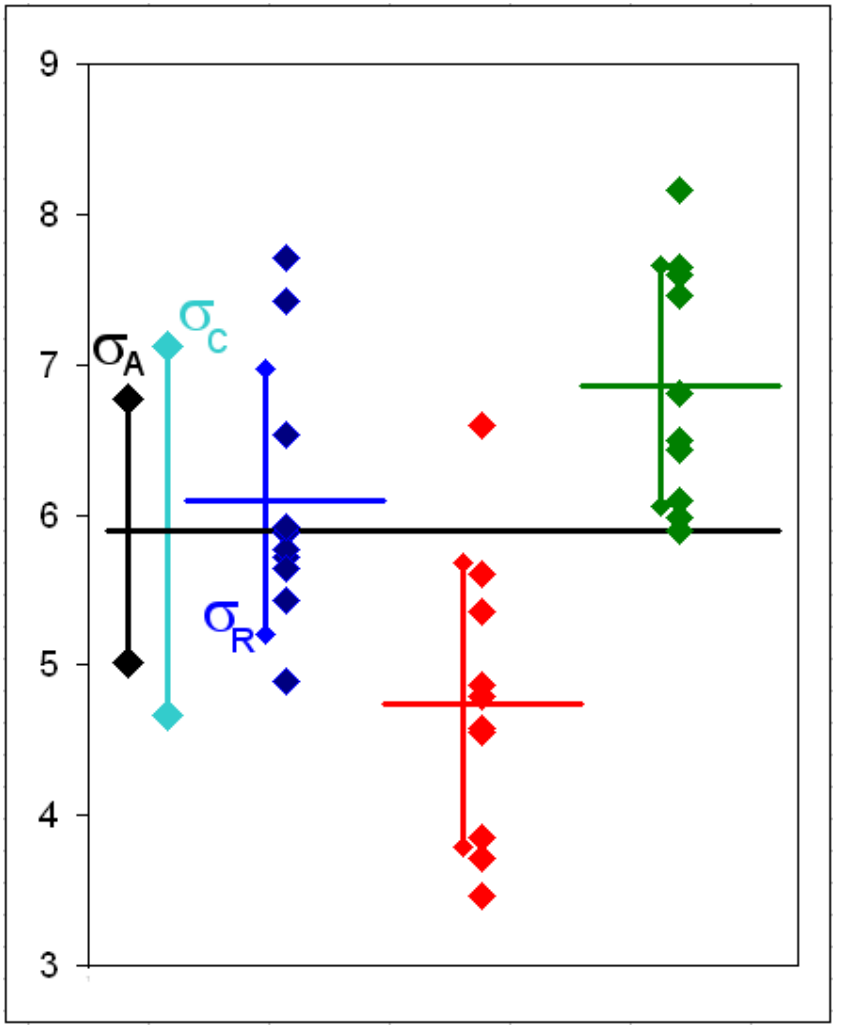
kde S_R je součet čtverců odchylek uvnitř jednotlivých úrovní a S_A je součet čtverců odchylek mezi úrovněmi.

Testuje se, zda je S_A významné ve srovnání s S_R.

Jednofaktorová ANOVA

S_c je s.č.o. od celkového průměru;
S_R je s.č.o. uvnitř jednotlivých úrovní;
S_A je s.č.o. mezi úrovněmi.
$S_c = S_R + S_A$
Testovací kritérium:

$$F = \frac{S_A(n-k)}{S_R(k-1)}$$



kde k je počet úrovní a
n je celkový počet měření.
Platí-li nulová hypotéza, má
F statistika ***Fisherovo rozdělení***
$F(k\text{-}1, n\text{-}k)$ s k-1 a n-k stupni volnosti.
Je-li $F > F_\alpha(k\text{-}1, n\text{-}k)$, můžeme nulovou hypotézu na hladině α zamítnout.

Kritické hodnoty lze spočítat v Excelu: F.INV.RT(α;k-1;n-k).

ANOVA v Excelu

Některé varianty ANOVA lze vypočítat v Excelu.

Instalace:

V menu: **Soubor** → **Možnosti** → **Doplňky**,
dole na kartě **Spravovat:** vybrat **Doplňky aplikace Excel**,
zmáčknout tlačítko **Přejít**,
zaškrtnout **Analytické nástroje** a zmáčknout tlačítko **OK**.

Spuštění:
V menu: **Data** → **Analýza dat**
Podle potřeby vybrat **Anova: Jeden faktor**
Anova: Dva faktory s opakováním
Anova: Dva faktory bez opakování

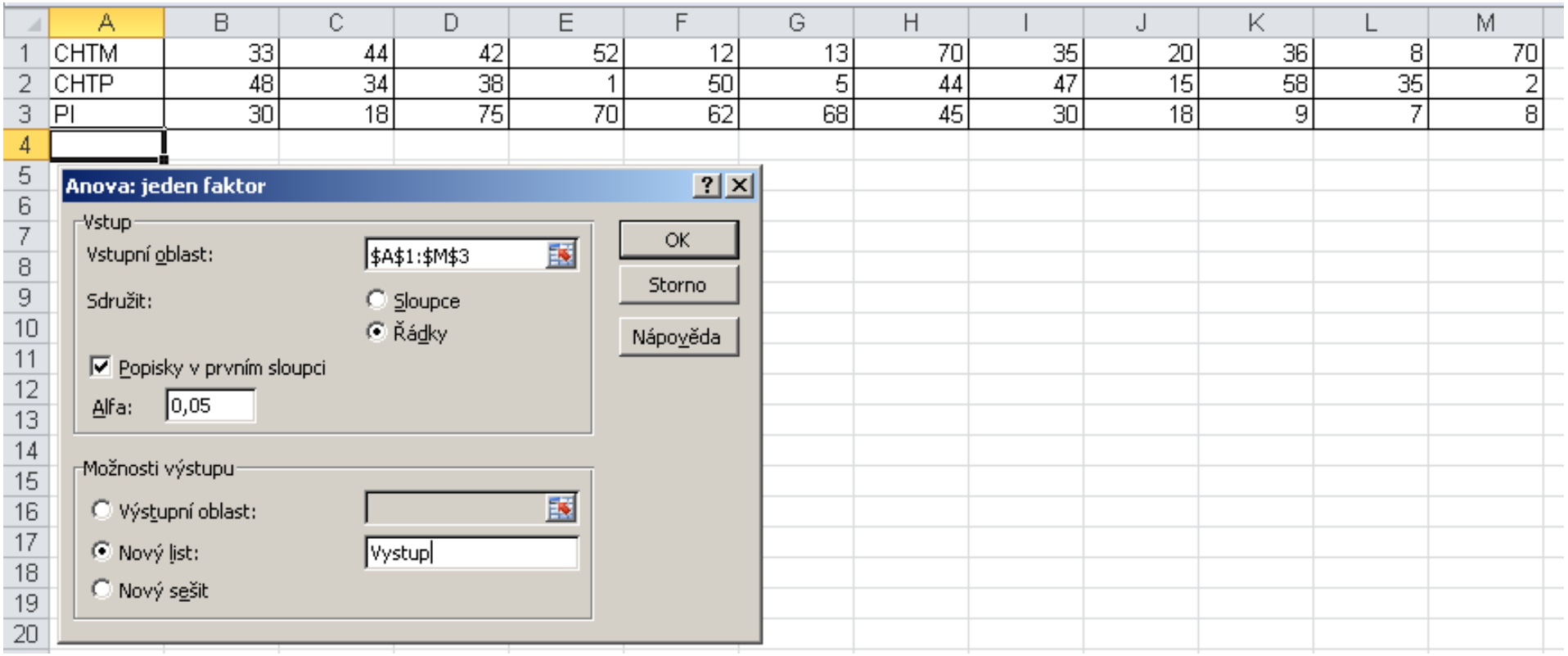
Jednofaktorová ANOVA v Excelu

Po písemce z Fyziky II bylo vybráno podle abecedy po 12 studentech studijních programů CHTM, CHTP a PI.
Body těchto studentů byly zapsány do tabulky:

CHTM	33	44	42	52	12	13	70	35	20	36	8	70
CHTP	48	34	38	1	50	5	44	47	15	58	35	2
PI	30	18	75	70	62	68	45	30	18	9	7	8

Jednofaktorová ANOVA v Excelu

Zadání parametrů:



Jednofaktorová ANOVA v Excelu

Výstup:

	A	B	C	D	E	F	G
1	Anova: jeden faktor						
2							
3	Faktor						
4	*Výběr*	*Počet*	*Součet*	*Průměr*	*Rozptyl*		
5	CHTM	12	435	36,25	438,3864		
6	CHTP	12	377	31,41667	413,5379		
7	PI	12	440	36,66667	684,2424		
8							
9							
10	ANOVA						
11	*Zdroj variability*	*SS*	*Rozdíl*	*MS*	*F*	*Hodnota P*	*F krit*
12	Mezi výběry	204,3889	2	102,1944	0,199577	0,820059	3,284918
13	Všechny výběry	16897,83	33	512,0556			
14							
15	Celkem	17102,22	35				

$$F = \frac{S_A (n-k)}{S_R (k-1)}$$

S_A $k - 1$ S_R $n - k$ S_C

Jednofaktorová ANOVA

Zamítneme-li nulovou hypotézu, víme, že některé se liší od ostatních. Které to jsou?

Scheffého metoda vícenásobného porovnání:

Je-li

$$|\mu_i - \mu_j| \geq \sqrt{\frac{k-1}{n-k} S_R F_\alpha(k-1, n-k) \left[\frac{1}{n_i} + \frac{1}{n_j}\right]}$$

lze nulovou hypotézu $\mu_i = \mu_j$ zamítnout.

Dvoufaktorová ANOVA

Posuzujeme vliv dvou faktorů *A* a *B* na různých úrovních. Kombinace faktorů tvoří mřížkovou strukturu. Mřížka se skládá z cel. (*i*,*j*) -tá cela odpovídá kombinaci úrovně A_i faktoru *A* a B_j faktoru *B*.

	B_1	B_2	B_3
A_1			
A_2	cela (2,1)		
A_3			

Je-li v každé cele jedna hodnota, mluvíme o ANOVA bez opakování.
Je-li v některé cele více než jedna hodnota, mluvíme o ANOVA s opakováním. Budeme se zabývat pouze případem, kdy je v každé cele stejný počet hodnot *p* (tzv. vyvážená třídění).

Dvoufaktorová ANOVA

Předpokládáme, že existují dva faktory *A* a *B*, které mají *k*, resp. *m* úrovní, s účinkem na znak Y, který lze vyjádřit vztahem:

$$\mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}$$

kde μ_{ij} je průměr znaku *Y* v (*i,j*)-té cele,

μ je celkový průměr znaku *Y*,

α_i je vliv faktoru *A* na znak *Y* v *i*-té úrovni,

β_j je vliv faktoru *B* na znak *Y* v *j*-té úrovni,

$\gamma_{i,j}$ charakterizuje interakci mezi faktory.

Nulová hypotéza pro všechny skupiny (úrovně faktoru *A*):

H_0: $\alpha_1 = \alpha_2 = ... = \alpha_k = 0$

Nulová hypotésa pro všechny bloky (úrovně faktoru *B*):

H_0': $\beta_1 = \beta_2 = ... = \beta_m = 0$

Dvoufaktorová ANOVA

Označme:

$\mu_{i.}$ - průměr v *i*-té skupině

$\mu_{.j}$ - průměr v *j*-tém bloku

μ - celkový průměr

$$S_c = \sum_{i=1}^{k} \sum_{j=1}^{m} \sum_{r=1}^{p} (y_{ijr} - \mu)^2 - \text{celkový součet čtverců}$$

$$S_A = mp \sum_{i=1}^{k} (\mu_{i.} - \mu)^2 - \text{meziskupinový součet čtverců}$$

$$S_B = kp \sum_{j=1}^{m} (\mu_{.j} - \mu)^2 - \text{meziblokový součet čtverců}$$

$$S_R = \sum_{i=1}^{k} \sum_{j=1}^{m} \sum_{r=1}^{p} (y_{ijr} - \mu_{i.} - \mu_{.j} + \mu)^2 - \text{vnitroskupinový-blokový s. č.}$$

$$S_c = S_A + S_B + S_{AB} + S_R$$

Dvoufaktorová ANOVA s opakováním

Pro ověření nulové hypotézy H_0 použijeme statistiku

$$F_A = \frac{(n-k-m+1)}{(k-1)} \frac{S_A}{S_R}$$

která má při platnosti nulové hypotézy *Fisherovo rozdělení* $F(k\text{-}1, n\text{-}k\text{-}m\text{+}1)$. Kritickou hodnotu vypočítáme v Excelu pomocí funkce =F.INV.RT(α;*k*-1;*n*-*k*-*m*+1).

Analogicky pro ověření hypotézy H_0' použijeme statistiku

$$F_B = \frac{(n-k-m+1)}{(m-1)} \frac{S_A}{S_R}$$

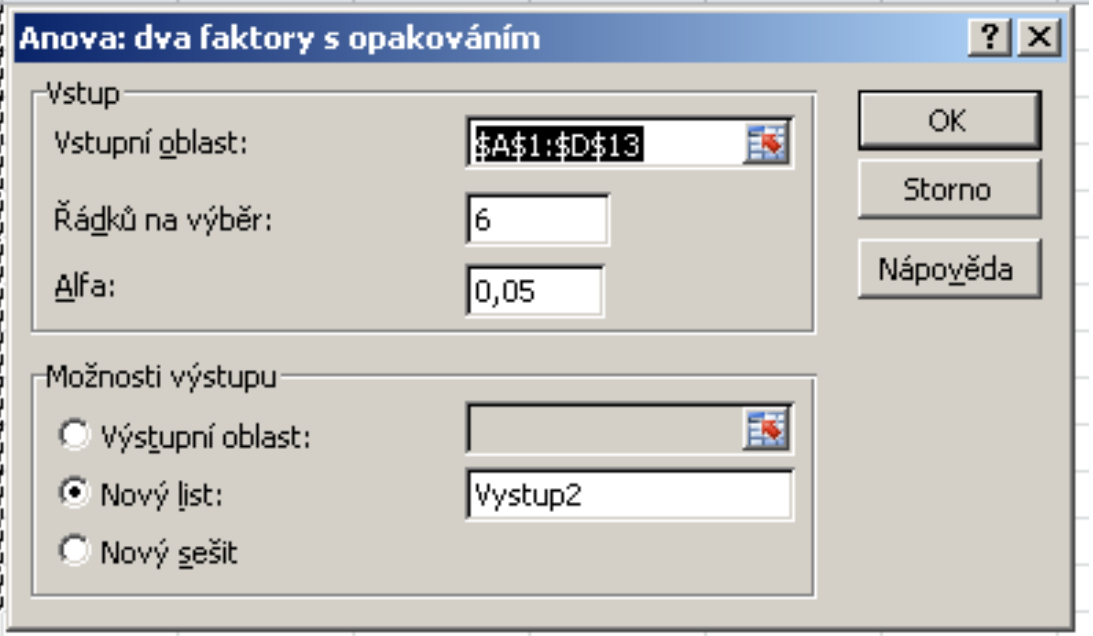
která má při platnosti nulové hypotézy *Fisherovo rozdělení* $F(m\text{-}1, n\text{-}k\text{-}m\text{+}1)$.

V obou případech nulovou hypotézu zamítneme, je-li F_A resp. F_B větší než příslušná hodnota Fisherova rozdělení.

Dvoufaktorová ANOVA

Po písemce z Fyziky II bylo vybráno podle abecedy po 12 studentech studijních programů CHTM, CHTP a PI, vždy 6 studentů a 6 studentek. Máme 2 faktory (program a pohlaví) a 6 hodnot v každé cele (s opakováním).

	A	B	C	D
1	pohlaví\obor	CHTM	CHTP	PI
2	student	44	48	30
3		42	34	18
4		52	35	75
5		70	2	70
6		35	41	62
7		20	33	68
8	studentka	33	38	30
9		12	1	42
10		13	50	18
11		22	5	27
12		64	44	34
13		35	47	30
14				



Poznámky:
1) Vstupní oblast musí obsahovat i záhlaví tabulky.
2) V každé cele musí být stejný počet hodnot.

Dvoufaktorová ANOVA

	A	B	C	D	E	F	G
1	Anova: dva faktory s opakováním						
2							
3	Faktor	CHTM	CHTP	PI	Celkem		
4	*student*						
5	Počet	6	6	6	18		
6	Součet	263	193	323	779		
7	Průměr	43,83333	32,16667	53,83333	43,27778		
8	Rozptyl	280,1667	250,1667	565,7667	405,3889		
9							
10	*studentka*						
11	Počet	6	6	6	18		
12	Součet	179	185	181	545		
13	Průměr	29,83333	30,83333	30,16667	30,27778		
14	Rozptyl	373,3667	482,1667	62,56667	270,2124		
15							
16	*Celkem*						
17	Počet	12	12	12			
18	Součet	442	378	504			
19	Průměr	36,83333	31,5	42			
20	Rozptyl	350,5152	333,3636	438,3636			
21							
22							
23	ANOVA						
24	*Zdroj variability*	*SS*	*Rozdíl*	*MS*	*F*	*Hodnota P*	*F krit*
25	Výběr	1521	1	1521	4,530831	0,041616	4,170877
26	Sloupce	661,5556	2	330,7778	0,985337	0,385072	3,31583
27	Interakce	752,6667	2	376,3333	1,121041	0,339213	3,31583
28	Dohromady	10071	30	335,7			
29							
30	Celkem	13006,22	35				

S_A S_B S_C S_R S_I

Výběr - meziskupinový SS (faktor A)
Sloupce - meziblokový SS (faktor B)
Interakce - SS pro interakci nezi faktory A, B
Dohromady -vnitroskupinový SS
Celkem - celkový SS

Dvoufaktorová ANOVA

Zjistili jsme, že není rozdíl mezi obory, ale je rozdíl mezi pohlavími. Pokusme se ověřit t-testem rozdíl mezi pohlavími:

pohlaví	průměr	rozptyl	směr. odch.
studenti	43.28	405.39	20.13
studentky	30.28	270.21	16.44

$$t = \frac{|\hat{\mu}_1 - \hat{\mu}_2|}{\sqrt{\frac{\hat{\sigma}_1^2}{N_1} + \frac{\hat{\sigma}_2^2}{N_2}}}$$

$t = 2.21$ $t_{krit}(0.05) = 2.02$ $\alpha = 0.041$

Je rozdíl mezi studenty CHTP a PI?

program	průměr	rozptyl	směr. odch.
CHTP	32.17	250	15.82
PI	53.83	565	23.79

$t = 1.86$ $t_{krit}(0.05) = 2.23$ $\alpha = 0.093$

Dvoufaktorová ANOVA s opakováním

Zamítneme-li nulovou hypotézu, víme, že některé se liší od ostatních. Které to jsou?

Scheffého metoda vícenásobného porovnání:

Je-li

$$|\mu_i - \mu_t| \geq \sqrt{\frac{2(k-1)}{mp(n-km)} S_R F_\alpha(k-1, n-km)}$$

lze nulovou hypotézu $\mu_i = \mu_t$ zamítnout.

Je-li

$$|\mu_j - \mu_t| \geq \sqrt{\frac{2(m-1)}{kp(n-km)} S_R F_\alpha(m-1, n-km)}$$

lze nulovou hypotézu $\mu_j = \mu_t$ zamítnout.