# Statistické metody a zpracování dat

## VI. Analýza rozptylu (ANOVA)

***Petr Dobrovolný***

## K čemu to je (příklad)

Studenti se připravovali na test ze statistiky třemi různými metodami.

Existuje na hladině významnosti $\alpha$=0,05 rozdíl mezi metodami přípravy?

| Metoda učení | | |
|---|---|---|
| A | B | C |
| 89 | 104 | 86 |
| 101 | 120 | 98 |
| 87 | 98 | 100 |
| 87 | 110 | 96 |

**Faktor**

| Výběr | Počet | Součet | Průměr | Rozptyl |
|---|---|---|---|---|
| A | 4 | 364 | 91 | 45,3 |
| B | 4 | 432 | 108 | 88 |
| C | 4 | 380 | 95 | 38,7 |

**ANOVA**

| Zdroj variability | SS | Rozdíl | MS | F | Hodnota P | F krit |
|---|---|---|---|---|---|---|
| Mezi výběry | 632 | 2 | 316 | 5,511628 | 0,027364 | 4,256492 |
| Všechny výběry | 516 | 9 | 57,33333 | | | |
| Celkem | 1148 | 11 | | | | |

Existuje rozdíl

## K čemu to je?

- Porovnávání libovolného počtu průměrů (více než dvou).
- Jeden či více tzv. **faktorů** dělí vyšetřované znaky do skupin.
- Testujeme, zda existuje významný rozdíl v průměrech skupin

**Příklady:**

- Vliv průmyslové lokality na koncentraci přízemního ozónu v ovzduší. Pro čtyři lokality jsme získali několik vzorků měření koncentrace přízemního ozónu. Máme zjistit, zda má lokalita významný vliv na koncentraci ozónu. Existuje lokalita, která se významně liší od ostatních?
- Existuje významný rozdíl v názoru různých skupin obyvatelstva na problém polohy brněnského nádraží?

## Obecný problém, který řeší ANOVA

| | *Skupina1* | *Skupina2* | . | . | *Skupina_m* |
|---|---|---|---|---|---|
| *Měření1* | $x_{11}$ | $x_{12}$ | . | . | $x_{1m}$ |
| *Měření2* | $x_{21}$ | $x_{11}$ | . | . | |
| *Měření3* | . | . | $x_{ij}$ | . | |
| . | . | . | . | . | |
| *Měření_n* | $x_{n1}$ | | | . | |
| *Počet* | $n_1$ | $n_2$ | $n_3$ | . | $n_m$ |
| *Průměr* | $\bar{x}_1$ | $\bar{x}_2$ | | . | $\bar{x}_m$ |
| *Sm.odch.* | $s_1$ | $s_2$ | | . | $s_m$ |

Máme *m* nezávislých náhodných výběrů ($m>2$, $j=1,2,\dots m$) vyšetřované proměnné *x*. Rozsahy výběrů $n_j$ nemusí být stejné. V každém výběru je znám průměr $\bar{x}_j$ a rozptyl $s_j^2$.

Výběry vzniknou obvykle tak, že základní soubor rozdělíme podle určitého znaku (FAKTORU) do *m* skupin a v každé z nich pak vybereme $n_i$ prvků.

Prvek $x_{ij}$ označuje i-té pozorování v *j*-tém výběru

## Základní druhy ANOVA

- ANOVA při jednoduchém třídění (**jednofaktorová**) – sledujeme efekt jednoho faktoru na závisle proměnnou
- ANOVA **vícefaktorová** – při dvojnásobném třídění, …
- ANOVA při **vyváženém** třídění (stejný počet prvků ve skupinách) a při **nevyváženém** třídění
- ANOVA s **opakováním** měření
- **Neparametrická** ANOVA

## Dva důvody, proč nemůžeme analýzu provést postupným testováním jednotlivých dvojic (např. t-testem):

(poznámka)

1) Museli bychom provádět **velký počet testování** (pro *m* skupin $m.(m-1)/2$ testů

2) Opakovaným porovnáváním významnosti bychom neoprávněně **zvyšovali pravděpodobnost chyby prvního druhu.**

U každého testu je řekněme 5% možnost chybného pozitivního výsledku (tedy chyby prvního druhu - hladina významnosti $\alpha$ = 0,05) pokud neexistuje žádný rozdíl.

Máme-li tři skupiny a provedeme všechny tři testy, pravděpodobnost, že dostaneme nejméně jeden chybný pozitivní výsledek (chybu prvního druhu) je větší než 5 %.

S rostoucím počtem provedených testů roste pravděpodobnost, že alespoň jeden výsledek bude statisticky významný, přestože ve skutečnosti platí nulová hypotéza.

Abychom se tomuto problému vyhnuli, použijeme k testování hypotézy metodu analýzy rozptylu a testů, které řeší tzv. mnohonásobná porovnávání (viz. dále).

## Obecný model analýzy rozptylu

ANOVA je založena na **předpokladu**, že každý z *m* výběrů pochází z populace s normálním rozdělením se stejnou směrodatnou odchylkou.

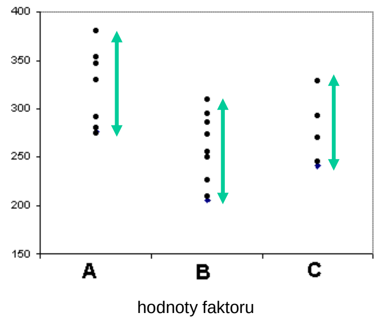
Zajímá nás, zda střední hodnoty (průměry) skupin jsou všechny shodné, nebo zda se navzájem liší.

$$x_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

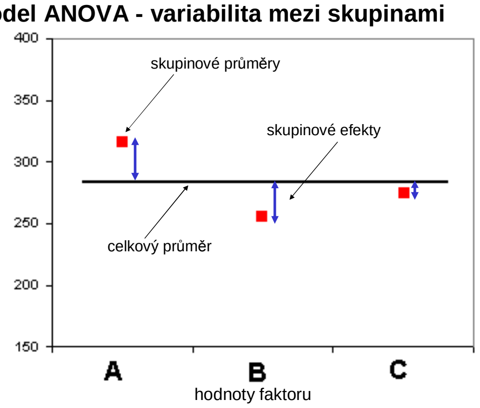
$x_{ij}$ je *i*-té pozorování z *j*-té skupiny.

Každé pozorované *x* je funkcí nějaké celkové průměrné hodnoty $\mu$, **skupinového efektu** $\alpha_i$ a blíže nespecifikované náhodné chyby $\varepsilon_{ij}$ .

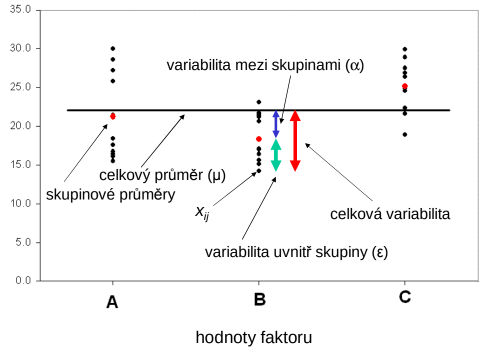
## Model ANOVA - variabilita uvnitř skupin




## Model ANOVA - variabilita mezi skupinami




## Zdroje variability v modelu ANOVA




## Obecný model analýzy rozptylu

Z předchozího plyne, že střední hodnota *j*-té skupiny je rovna:

$$\mu_j = \mu + \alpha_j$$

V analýze rozptylu chceme zjistit, zda jsou skupinové efekty důležité, tj. zda existuje nějaký rozdíl mezi průměry jednotlivých skupin.

**Nulová hypotéza $H_0$:** všechny výběry pocházejí z jednoho základního souboru s normálním rozložením (jinými slovy – faktor neovlivňuje závisle proměnnou)

$H_0$: $\mu_1 = \mu_2 = \ldots = \mu_i = \ldots. = \mu_m = \mu$

nebo:

$H_0$: $\alpha_1 = \alpha_2 = \ldots = \alpha_i = \ldots. = \alpha_m = 0$

**Cílem ANOVA je zjistit, zda se jednotlivé dílčí průměry $\mu_m$ mezi sebou a tedy i od celkového průměru $\mu$ liší pouze v mezích náhodného kolísání.**

## Obecný výpočet ANOVA

Podstatou výpočtů při ANOVA je rozdělení celkového rozptylu ($S_T$) závisle proměnné do dvou částí, na **variabilitu uvnitř skupin** ($S_e$) a **variabilitu mezi skupinami** ($S_A$)

$$S_T = S_A + S_e$$

**Variabilita uvnitř skupin** popisuje, jak se každá hodnota ve skupině liší od skupinového průměru.

**Variabilita mezi skupinami** je funkcí, která ukazuje, jak se navzájem liší skupinové průměry. Zahrnuje porovnání všech k skupinových průměrů s tzv. celkovým průměrem.

Pokud neexistuje žádný rozdíl mezi skupinovými průměry, pak variabilita mezi skupinami i variabilita v rámci skupiny popisují stejný jev - stejný populační rozptyl.

Toto porovnání variability v rámci skupiny a mezi skupinami se provádí pomocí **F testu**.

## Obecný výpočet ANOVA

Zkoumáme, že vypočtené průměry $\bar{x}_j$ se liší jen v mezích náhodného kolísání $\bar{x}$

Odchylku konkrétního měření $x_{ij}$ od celkového průměru lze zapsat:

$$x_{ij} - \bar{x} = (x_{ij} - \bar{x}_j) + (\bar{x}_j - \bar{x})$$

odhad parametru $\alpha_j$ - tedy efekt kategorie j

Umocníme a sečteme obě strany rovnice pro všechna měření:

$$S_T = \sum_j \sum_i (x_{ij} - \bar{x})^2 = \sum_j n_j (\bar{x}_j - \bar{x})^2 + \sum_j \sum_i (x_{ij} - \bar{x}_j)^2 = S_A + S_e$$

## Obecný výpočet ANOVA

Jednotlivé složky celkového rozptylu mají tento význam:

$S_T$ – celkový součet čtverců odchylek všech měření od celkového průměru

$S_A$ - vážený součet druhých mocnin rozdílů každého skupinového průměru a celkového průměru

$S_e$ - součet druhých mocnin rozdílů hodnot a příslušného skupinového průměru

Každé složce rozptylu přísluší jistý počet stupňů volnosti $\nu$ :

$\nu_T$ pro $S_T$ – počet pozorování – 1: *(n-1)*

$\nu_A$ pro $S_A$ - počet skupin – 1: *(m-1)*

$\nu_e$ pro $S_e$ – počet pozorování – počet skupin: *(n – m)*

## Obecný výpočet ANOVA

Charakteristiky

$$MS_A = \frac{S_A}{\nu_A} \qquad MS_e = \frac{S_e}{\nu_e}$$

představují součty čtverců dělené odpovídajícím počtem stupňů volnosti. Tyto veličiny jsou **mírou variability pro jednotlivé zdroje rozptylu** a ve statistických programech jsou označovány anglicky jako Mean Square (průměrné čtverce).

Testovací kritérium se potom vypočte jako podíl míry variability mezi skupinami a míry variability uvnitř skupin podle následujícího vztahu:

$$F = \frac{MS(mezi\_skupinami)}{MS(uvnitř\_skupin)} = \frac{S_A / \nu_A}{S_e / \nu_e}$$

## Typická tabulka výstupu z ANOVA

| Zdroj variability | *S* | *st.v.* | *MS* | *F* |
|---|---|---|---|---|
| faktor *A* | $S_A$ | $m - 1$ | $MS_A = \frac{S_A}{m-1}$ | $F = \frac{MS_A}{MS_e}$ |
| reziduální | $S_e$ | $n - m$ | $MS_e = \frac{S_e}{n-m}$ | |
| Celková variabilita | $S_T$ | $n - 1$ | | |

Výstupy ze statistického programu ještě nabízejí ***p hodnotu*** příslušející vypočtené hodnotě testovacího kritéria

## Interpretace testovacího kritéria

• V případě platnosti $H_0$ (všechny populační průměry shodné) bude čitatel *F* statistiky (zhruba) stejný jako jmenovatel (tzv. reziduální rozptyl)

• Pak by tedy hodnota *F* statistiky byla přibližně rovna jedné. Ve statistických tabulkách zjistíme, zda hodnota *F* je významně větší než 1

• To by ukazovalo, že *MS* mezi skupinami je významně větší než *MS* uvnitř skupin, a tedy že se průměry skupin liší.

• (Pokud by *F* statistika byla menší než 1, pak to znamená, že variabilita mezi skupinami může být dokonce menší než uvnitř skupin, a tedy tím spíše není důvod zamítat nulovou hypotézu.)

• K výpočtu příslušných kritických hodnot i dosažených hladin významnosti lze využít i různé statistické programy.

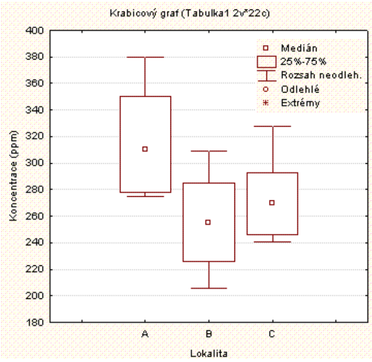
## Příklad ANOVA při jednoduchém třídění

Zjistěte, zda se na hladině významnosti $\alpha$=0,05 liší se koncentrace znečišťující látky (ppm) v ovzduší měřené na třech lokalitách?

| Lokalita | A | B | C |
|---|---|---|---|
| 1 | 276 | 206 | 241 |
| 2 | 280 | 210 | 246 |
| 3 | 275 | 226 | 270 |
| 4 | 291 | 249 | 293 |
| 5 | 347 | 255 | 328 |
| 6 | 354 | 273 | |
| 7 | 380 | 285 | |
| 8 | 330 | 295 | |
| 9 | | 309 | |
| *n* | 8 | 9 | 5 |
| $\bar{x}$ | 316,6 | 256,4 | 275,6 |
| *s* | 41,2 | 37,1 | 35,9 |

## Příklad

Vizuální analýza jednotlivých skupin za pomoci vhodného grafu a porovnání úrovně a variability skupin.




## Příklad

***Výpočet v EXCELU:***

***Nástroje – Analýza dat – ANOVA jeden faktor***

| A | B | C |
|---|---|---|
| 276 | 206 | 241 |
| 280 | 210 | 246 |
| 275 | 226 | 270 |
| 291 | 249 | 293 |
| 347 | 255 | 328 |
| 354 | 273 | |
| 380 | 285 | |
| 330 | 295 | |
| | 309 | |

Anova: jeden faktor

Faktor

| *Výběr* | *Počet* | *Součet* | *Průměr* | *Rozptyl* |
|---|---|---|---|---|
| A | 8 | 2533 | 316,625 | 1699,411 |
| B | 9 | 2308 | 256,4444 | 1378,028 |
| C | 5 | 1378 | 275,6 | 1288,3 |

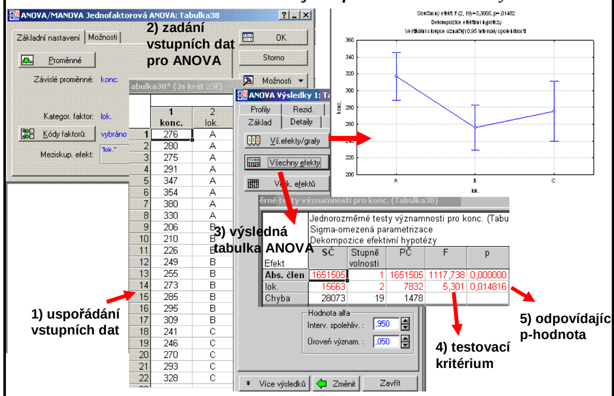
ANOVA

| *Zdroj variability* | *SS* | *Rozdíl* | *MS* | *F* | *Hodnota P* | *F krit* |
|---|---|---|---|---|---|---|
| Mezi výběry | 15663,48 | 2 | 7831,738 | 5,300518 | 0,014816 | 3,52189 |
| Všechny výběry | 28073,3 | 19 | 1477,542 | | | |
| | | | | | | |
| Celkem | 43736,77 | 21 | | | | |

Protože *p = 0,0148*, což je méně než *α = 0,05*, můžeme zamítnout nulovou hypotézu a učinit závěr, že průměrná koncentrace znečišťující látky není ve všech třech skupinách stejná.

## Příklad ANOVA v programu Statistica – část I.

***Statistika – ANOVA – jednofaktorová ANOVA – Rychlé nastavení***




## Dva problémy výsledu ANOVA:

1) Zda jsou výsledky ANOVA vůbec použitelné - musíme ověřit, že náš model splňuje **předpoklady**

2) Výsledek ANOVA nám neříká, které průměry se navzájem liší.

Můžeme se podívat na skupinové průměry a zjistit, že určitá skupina má vyšší průměr než ostatní skupiny.

V tuto chvíli ale nemůžeme říci, že tento průměr je významně vyšší.

Musíme data analyzovat dále použitím metod **mnohonásobného porovnávání**, abychom zjistili, které průměry se navzájem významně liší.

## Předpoklady ANOVA

Aby byly výsledky analýzy rozptylu správné, musí být splněny následující předpoklady:

a) Všechna měření musí být vzájemně nezávislá uvnitř skupin i mezi skupinami

b) Vyšetřovaný znak, jehož průměry chceme porovnávat musí mít normální rozdělení

c) Rozptyly jednotlivých výběrů se mezi sebou statisticky neliší (což ověřujeme testy (Bartlettův test nebo tzv. Hartleyův test ($F_{max}$ test) - pokud mají všechny výběry stejný rozsah.)

## Ad c) předpoklad rovnosti rozptylů

Zkoumáme, zda je splněno:

$$\frac{\max s_j}{\min s_j} \leq 3$$

Hodnoty $s_j$ jsou směrodatné odchylky měření v jednotlivých skupinách

## Ad b) předpoklad normálního rozdělení

Ověřování lze provádět graficky analýzou tzv. **reziduálních** (zbytkových) hodnot

Hodnoty pozorovaných veličin můžeme vyjádřit takto:

$$x_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

$\varepsilon_{ij}$ jsou náhodné navzájem nezávislé chybové složky (**rezidua**)

- Model platí pro základní soubor
- Skutečné parametry však můžeme pouze odhadovat z výběrových souborů.
- V následujícím příkladu index *o* u symbolu parametru znamená, že se jedná o odhad.

## Ověřování normality

$\alpha_{oj}$ - odhady skupinových efektů - tedy toho, jak se každý průměr liší od celkového průměru.

Předpovídaná hodnota pro pozorování z j-té skupiny je průměr j-té skupiny:

$$\mu_{oi} = \mu_o + \alpha_{oi}$$

**Příklad:**
$\mu_o$ – celkový průměr = 282,7
$\alpha_{o1}$ = průměr první skupiny - celkový průměr = 316,6 – 282,7 = 33,9
$\alpha_{o2}$ = průměr druhé skupiny - celkový průměr = 256,4 – 282,7 = -26,3
$\alpha_{o3}$ = průměr třetí skupiny - celkový průměr = -7,1

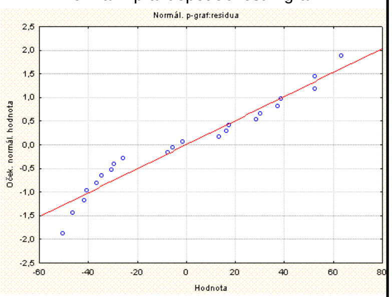
Naším modelem ANOVA jsme tedy vypočetli, že například průměrná hodnota koncentrace měřené látky se v první skupině rovná 282,7 + 33,9 = 316,6.

## Ověřování normality

Rezidua (zbytkové hodnoty) pro každé pozorování spočteme jako rozdíl mezi pozorovanou hodnotou a předpovídanou hodnotou:

| LOK. | MĚŘENO | MODEL | REZIDUUM |
|---|---|---|---|
| A | 276 | 316,6 | -40,6 |
| A | 280 | 316,6 | -36,6 |
| A | 275 | 316,6 | -41,6 |
| A | 291 | 316,6 | -25,6 |
| A | 347 | 316,6 | 30,4 |
| A | 354 | 316,6 | 37,4 |
| A | 380 | 316,6 | 63,4 |
| A | 330 | 316,6 | 13,4 |
| B | 206 | 256,4 | -50,4 |
| B | 210 | 256,4 | -46,4 |
| B | 226 | 256,4 | -30,4 |
| B | 249 | 256,4 | -7,4 |
| B | 255 | 256,4 | -1,4 |
| B | 273 | 256,4 | 16,6 |
| B | 285 | 256,4 | 28,6 |
| B | 295 | 256,4 | 38,6 |
| B | 309 | 256,4 | 52,6 |
| C | 241 | 275,6 | -34,6 |
| C | 246 | 275,6 | -29,6 |
| C | 270 | 275,6 | -5,6 |
| C | 293 | 275,6 | 17,4 |
| C | 328 | 275,6 | 52,4 |

Normální pravděpodobnostní graf




***Statistika- Základní statistiky/tabulky – Popisné statistiky – Prav. & bod. grafy***

## Ověřování předpokladu normality

- Vytvoříme nejprve graf předpovídaných hodnot vs. pozorovaných hodnot.
- Mají-li rezidua normální rozdělení, měl by tzv. normální pravděpodobnostní graf vytvořit přímku.
- Přítomnost jakýchkoli velkých odchylek by mohla znamenat doporučení transformace dat před provedením analýzy nebo nutnost provedení neparametrické verze testu.
- Jak je patrné z normálního grafu, v našem případě je sestavený model ANOVA vyhovující.

## Mnohonásobná porovnávání

- Analýza rozptylu nám pouze říká, že průměry nejsou stejné. Je třeba provést další analýzu, abychom zjistili, jak se liší.
- Jednou z možností je porovnat každou dvojici průměrů, nebo dvojice, které nás zajímají.
- Mnohonásobné testování významnosti dává vysokou pravděpodobnost, že bude nalezen významný rozdíl pouze náhodou.
- Například: test má 5% možnost chybného pozitivního výsledku (hladina významnosti $\alpha$).
- To znamená, že při opakovaném testování bychom chybně zamítli nulovou hypotézu v 5 % případů **– tedy např. při padesáti testech uděláme při $\alpha = 0,05$ 2-3 chyby .**
- Kdybychom měli čtyři skupiny a porovnali je navzájem tak, že bychom provedli všech šest testů, potom by pravděpodobnost, že dostaneme nejméně jednou chybný pozitivní výsledek **(chyba prvního druhu), byla mnohem větší než 5 %.**

## Mnohonásobná porovnávání

Tato situace se označuje jako problém mnohonásobného porovnávání a pro jeho řešení existuje několik metod (např. Bonferroniho, Tukeyova, Newman-Keulsova, Duncanova, Fisherovo LSD (nejmenší významný rozdíl - Least Significant Difference) a Scheffého).

Úkolem každé metody je udržet danou hladinu pravděpodobnosti chyby prvního druhu (5 %) a v podstatě ji rozdělit mezi všechna porovnání.

## Mnohonásobná porovnávání

**Bonferroniho metoda:** Pro ta porovnání, která nás zajímají, provedeme modifikované t-testy s upravenou hladinou významnosti.

Tu získáme tak, že hladinu $\alpha$ jednoduše vydělíme celkových počtem porovnání, která chceme provést.

Tato hodnota pak bude naší hladinou významnosti pro každý t-test.

Řekněme, že pro náš příklad chceme provést všechna možná porovnání - pro tři skupiny existují tři.

Naše hladina významnosti pro každé porovnání nebude tedy 5 %, ale (5/3) % = 1,67 %.

Nulová a alternativní hypotéza jsou stejné jako pro obyčejný t test.

## Mnohonásobná porovnávání

Testová statistika t-testu se v tomto případě počítá následujícím způsobem:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{S_e}{\nu_e}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Od běžného t-testu se liší ve jmenovateli – na místo rozptylu jen ze dvou skupin(které porovnáváme) použijeme sdruženou verzi rozptylu ze **všech** skupin, včetně těch, které nepoužíváme při porovnávání.

Za platnosti nulové hypotézy má testová charakteristika t rozdělení s $\nu_e$ stupni volnosti.

Upravená hladina významnosti při třech skupinách (viz. výše) se rovná 1,67%.

Je-li tedy vypočtená hladina významnosti (p hodnota) menší než 0,0167, potom zamítáme nulovou hypotézu o rovnosti průměrů dvou testovaných skupin.

## Výsledky mnohonásobných porovnávání

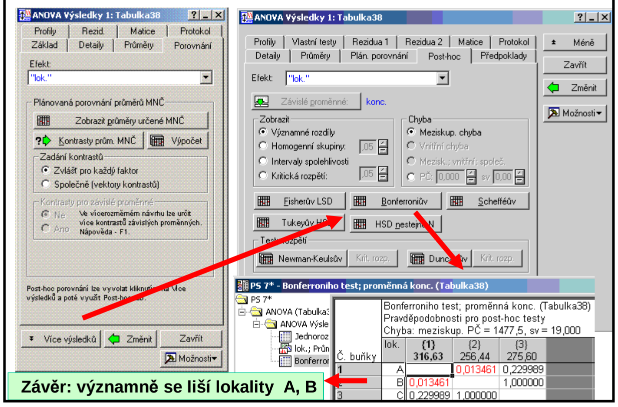
**Příklad:** srovnání jednotlivých skupin:

| | | |
|---|---|---|
| první – druhá | $t = 3{,}22$ | $p < 0{,}0167$ |
| první – třetí | $t = 1{,}87$ | $p > 0{,}0167$ |
| druhá – třetí | $t = -0{,}90$ | $p > 0{,}0167$ |

Výsledky ANOVA nám ukazují, že existuje významný rozdíl mezi průměry skupin 1 a 2.

## Příklad ANOVA v programu Statistica – část II. pokračování

*1) Porovnání – 2) Více výsledků – 3) Bonferroniův*




Závěr: významně se liší lokality A, B

## Jednofaktorová ANOVA – základní interpretace výsledků v programu Statistica

**Příklad:** Zjistěte, zda se významně liší hodnoty maximálních měsíčních nárazů větru naměřené v letech 1921 až 1923 na stanici Praha -Karlov

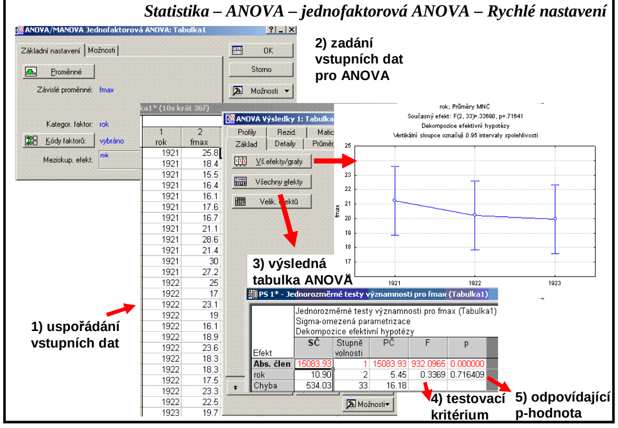
**Příklad řešený v EXCELu:**

| | D | E | F | G | H | I | J | K | L | M | N | O |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | | FMAX21 | FMAX22 | FMAX23 | | Anova: jeden faktor | | | | | | |
| 2 | | 25.8 | 25.0 | 19.7 | | | | | | | | |
| 3 | | 18.4 | 17.0 | 21.2 | | Faktor | | | | | | |
| 4 | | 15.5 | 23.1 | 13.7 | | *Výběr* | *Počet* | *Součet* | *Průměr* | *Rozptyl* | | |
| 5 | | 16.4 | 19.0 | 16.4 | | FMAX21 | 12 | 254.8 | 21.23333 | 28.36242 | | |
| 6 | | 16.1 | 16.1 | 21.7 | | FMAX22 | 12 | 242.6 | 20.21667 | 9.32697 | | |
| 7 | | 17.6 | 18.9 | 22.3 | | FMAX23 | 12 | 239.5 | 19.95833 | 10.85902 | | |
| 8 | | 16.7 | 23.6 | 19.4 | | | | | | | | |
| 9 | | 21.1 | 18.3 | 22.3 | | | | | | | | |
| 10 | | 28.6 | 18.3 | 17.2 | | ANOVA | | | | | | |
| 11 | | 21.4 | 17.5 | 24.7 | | *Zdroj variabil.* | *SS* | *Rozdíl* | *MS* | *F* | *Hodnota P* | *F krit* |
| 12 | | 30.0 | 23.3 | 17.2 | | Mezi výběr | 10.90389 | 2 | 5.451944 | 0.336897 | 0.716409 | 3.284918 |
| 13 | | 27.2 | 22.5 | 23.7 | | Všechny v | 534.0325 | 33 | 16.1828 | | | |
| 14 | | 21.2 | 20.2 | 20.0 | | | | | | | | |
| 15 | | 5.098911 | 2.923991 | 3.155011 | | Celkem | 544.9364 | 35 | | | | |
| 16 | | 254.8 | 242.6 | 239.5 | | | | | | | | |

Hodnoty maximálních měsíčních nárazů větru pro $\alpha$=0,05 se neliší

## Příklad ANOVA v programu Statistica

*Statistika – ANOVA – jednofaktorová ANOVA – Rychlé nastavení*



## Neparametrická Analýza rozptylu (Kruskalův –Wallisův test)

- měření nejsou normálně rozdělena, jsou měřena na ordinální škále, …
- využívá ne vlastních měřených hodnot, ale jejich pořadí (rank), které získáme jejich setříděním.

**Nulová hypotéza $H_o$:** Měření ve skupinách mají stejné mediány

$$H_0 : \tilde{\mu}_1 = \tilde{\mu}_2 = \ldots = \tilde{\mu}_m$$

**Alternativní hypotéza $H_1$:** Alespoň pro jednu dvojici *i,j* platí:

$$\tilde{\mu}_i \neq \tilde{\mu}_j$$

## Kruskalův –Wallisův test – obecný postup

- Uspořádáme všech *n* měření podle velikosti.
- Nahradíme hodnoty měření jejich pořadími
- Vypočítáme hodnoty $SR_j$ – tj. součet pořadí měření ze skupiny j
- Vypočítáme testovací charakteristiku H jako míru rozdílnosti mediánu pořadí ve skupinách

$$H = \left[ \frac{12}{n(n+1)} \sum_j \frac{(SR_j)^2}{n_j} \right] - 3(n+1)$$

- Pokud platí $H_0$, potom pro velká $n_j$ má testovací statistika H přibližně $\chi 2$ rozdělení
- Na zvolené hladině významnosti α zamítáme $H_0$, pokud testovací statistika H je větší než kritická hodnota $\chi 2$ rozdělení o *m-1* stupních volnosti.
- A nebo: vypočtenému H příslušející *p hodnota* je menší než hladina významnosti α.

## Kruskalův – Wallisův test příklad

Zjistěte, zda existuje významný rozdíl v názorech lidí žijících v různých lokalitách na ohrožení životního prostředí?

Tři skupiny respondentů po 10 členech.

- Skupina A – lidé pracující v chemickém závodě a bydlící v jeho okolí
- Skupina B – lidé pracující mimo lokalitu a bydlící v sousedství chemického závodu
- Skupina C – lidé, kteří nepracují v chemické továrně, ani nebydlí v jejím okolí

Výsledky dotazníku jsou v dispozici ve formě skore.

## Kruskalův – Wallisův test - příklad

Vstupní data:

| Skupina A | | Skupina B | | Skupina C | |
|---|---|---|---|---|---|
| skore | pořadí | skore | pořadí | skore | pořadí |
| 2 | 7 | 3 | 15 | 3 | 15 |
| 3 | 15 | 4 | 24 | 4 | 24 |
| 1 | 2 | 4 | 24 | 2 | 7 |
| 2 | 7 | 5 | 29,5 | 1 | 2 |
| 2 | 7 | 3 | 15 | 3 | 15 |
| 1 | 2 | 4 | 24 | 4 | 24 |
| 3 | 15 | 2 | 7 | 2 | 7 |
| 4 | 24 | 4 | 24 | 3 | 15 |
| 3 | 15 | 5 | 29,5 | 3 | 15 |
| 2 | 7 | 4 | 24 | 4 | 24 |

$\sum R_A = 101$ $\sum R_B = 216$ $\sum R_C = 148$

## Kruskalův –Wallisův test

Výpočet testovacího kritéria

$$H = \left[ \frac{12}{n(n+1)} \cdot \left( \frac{(\Sigma R_A)^2}{n_A} + \frac{(\Sigma R_B)^2}{n_B} + \frac{(\Sigma R_C)^2}{n_C} \right) \right] - 3(n+1)$$

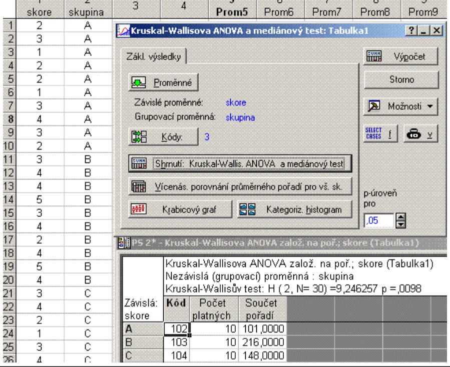
$$H = \left[ \frac{12}{30 \cdot 31} \cdot \left( \frac{101^2}{10} + \frac{216^2}{10} + \frac{148^2}{10} \right) \right] - 3 \cdot 31 = 8{,}627$$

V tabulkách najdeme kritickou hodnotu $\chi 2$ rozdělení pro $\alpha = 0{,}05$ a pro $v = m - 1$, tedy 2 stupně volnosti: 5,991

**Závěr:** Odmítáme nulovou hypotézu. V názorech lidí žijících v různých lokalitách na ohrožení životního prostředí je statisticky významný rozdíl na hladině významnosti $\alpha = 0{,}05$.

## Kruskalův –Wallisův test - Statistica

***Statistika – Neparametrická statistika – Porovnání více nezávislých vzorků (skupiny)***



## Analýza rozptylu při dvojném třídění

Zkoumáme vliv dvou faktorů (např. A, B) na závisle proměnnou

a – počet úrovní faktoru A

b – počet úrovní faktoru B

$n_{ij}$ – počet objektů odpovídajících i-té úrovni faktoru A a j-té úrovni faktoru B

Často jsou všechny četnosti $n_{ij}$ stejné: $n_{ij} = c$ (tzv. vyvážené třídění)

| | Chlapci (hladina B1) | Dívky (hladina B2) |
|---|---|---|
| Metoda výuky 1 (hladina A1) | 89<br>101 | 87<br>87 |
| Metoda výuky 2 (hladina A2) | 120<br>110 | 98<br>104 |
| Metoda výuky 3 (hladina A3) | 100<br>98 | 86<br>96 |

## Model ANOVA při dvojném třídění

$$x_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}$$

$\mu$ - společná část průměru závisle proměnné

$\alpha_i$ - efekt faktoru A na úrovni *i (i=1, …, a)*

$\beta_j$ - efekt faktoru B na úrovni *j (j=1, …, b)*

$\gamma_{ij}$ - **interakce** mezi faktorem A na úrovni *i* a faktorem B na úrovni *j*

$\varepsilon_{ijk}$ – náhodná chyba s nulovou střední hodnotou, normálním rozdělením a stejným rozptylem pro všechna *i, j.*

Pro každou kombinaci faktorů měříme *c* objektů (k=1,2,…c), c>1

## Model ANOVA při dvojném třídění

**Zkoumáme tři páry hypotéz:**

H01: $\alpha 1 = \alpha 2 = \ldots = \alpha a = 0$
H11: Ne všechny efekty *αi* jsou nulové
H02: $\beta 1 = \beta 2 = \ldots = \beta b = 0$
H12: Ne všechny efekty *βj* jsou nulové
H03: Mezi faktory A B není žádná interakce (všechna $\gamma_{ij}=0$)
H13: Některé interakce jsou nenulové

Testovací statistika F opět vychází z rozkladu čtverců odchylek měření od společného průměru $\bar{x}$

Symbolicky:

$$S_T = S_A + S_B + S_I + S_e$$

$S_A$, $S_B$ – efekty faktorů
$S_I$ – interakce
$S_e$ – variabilita uvnitř skupin

### Tabulka výstupu z ANOVA při dvojném třídění

| Zdroj variability | *S* | *st.v.* | *MS* | *F* | $H_0$ |
|---|---|---|---|---|---|
| faktor *A* | $S_A$ | $a-1$ | $MS_A$ | $MS_A/MS_e$ | $H_{01}$ |
| faktor *B* | $S_B$ | $b-1$ | $MS_B$ | $MS_B/MS_e$ | $H_{02}$ |
| interakce | $S_I$ | $(a-1)(b-1)$ | $MS_I$ | $MS_I/MS_e$ | $H_{03}$ |
| reziduální | $S_e$ | $ab(c-1)$ | $MS_e$ | | |
| Celkový rozptyl | $S_T$ | $abc-1$ | | | |

**INTERAKCE:**

Značí, že faktory nepůsobí izolovaně - jinými slovy nejsou nezávislé.

Faktory produkují větší (menší) efekt, než který bychom zjistili, kdybychom posuzovali každý faktor zvlášť.

Významné interakce způsobují, že jednotlivé faktory nevysvětlují veškerou variabilitu

Hypotézu o existenci (H03) či neexistenci (H13) interakcí zkoumáme jako první.

### Příklad – výsledky ANOVA při dvojném třídění

| Zdroj variability | *S* | *st.v.* | *MS* | *F* | Kritická hodnota |
|---|---|---|---|---|---|
| faktor *A* | 632 | 2 | 316 | 9,88 | $F_{0.05}(2,6) = 5,14$ |
| faktor *B* | 300 | 1 | 300 | 9,36 | $F_{0.05}(1,6) = 5,99$ |
| interakce *A × B* | 24 | 2 | 12 | 0,38 | $F_{0.05}(2,6) = 5,14$ |
| reziduální | 192 | 6 | 32 | | |
| Celkový $S_T$ | 1148 | 11 | | | |