ANALÝZA ROZPTYLU
(Analysis of Variance ANOVA)

Používá se buď jako samostatná technika, nebo jako postup, umožňující analýzu zdrojů variability v lineární regresi. Př. použití:

- k porovnání středních hodnot (průměrů) více než 2 souborů,
- určení vlivu způsobu přípravy vzorků (několika způsobů),
- zpracování mezilaboratorních porovnávacích zkoušek (MPZ).

Porovnání shody více než dvou metod stanovení analytu, účinek více než dvou léčiv na dané onemocnění, porovnání účinku více než dvou hnojiv na výnos, atd.

PODSTATA: rozklad celkového rozptylu na rozptyl vyvolaný vlivem jednotlivých faktorů (známé zdroje variability) a složku náhodnou (neobjasněnou), o níž se předpokládá, že je náhodná.
Předmětem testování je statistická významnost poměru mezi rozptylem způsobeným faktorem (MS_A) náhodným rozptylem (MS_R). Pokud máme 1 faktor, mluvíme o jednofaktorové *ANOVě*, máme-li 2 faktory, jde o dvoufaktorovou *ANOVu*, apod.

Základní předpoklady pro (jednofaktorovou) analýzu rozptylu:

- data pocházejí z normálního rozdělení,
- náhodné chyby ε_{ij} jsou náhodné veličiny s $N(0, \sigma^2)$,
- rozptyly sloupců dat (úrovní faktoru) jsou stejné (homoskedasticita).

Jednofaktorová ANOVA

Formulace modelu: sleduje se faktor A na k úrovních $A_1, \ldots, A_k$; na každé úrovni je provedeno n_i měření (celkový počet měření označujeme N). Model ANOVA má tvar:

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

jednotlivé parametry se odhadují pomocí odpovídajících aritmetických průměrů, a to následovně:

μ … celkový aritmetický průměr všech hodnot v matici $\bar{\bar{x}}$,

α_i … efekt i-té úrovně faktoru A, $\boldsymbol{\alpha_i = \mu_i - \mu}$, kde μ_i je sloupcový průměr $\bar{x}_i$.

Jednofaktorová ANOVA porovnává střední hodnoty (průměry) dvou či více úrovní faktoru A čili sloupců v matici dat za účelem určit, zda alespoň jedna sloupcová střední hodnota se liší od ostatních. Statistická významnost je testována F-testem tak, že H_0 říká „*Všechny střední hodnoty jsou stejné*“ a H_1 „*Alespoň jedna střední hodnota se odlišuje od ostatních*“.

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

$$(x_{ij} - \bar{\bar{x}})^2 = \left[(\bar{x}_i - \bar{\bar{x}}) + (x_{ij} - \bar{x}_i)\right]^2$$

sumací přes i a j získáme následující rovnici, ve které je poslední člen = 0

$$\sum_i \sum_j (x_{ij} - \bar{\bar{x}})^2 = \sum_i \sum_j (\bar{x}_i - \bar{\bar{x}})^2 + \sum_i \sum_j (x_{ij} - \bar{x}_i)^2 + 2\sum_i \sum_j (\bar{x}_i - \bar{\bar{x}})(x_{ij} - \bar{x}_i),$$

$$S_0 = S_A + S_R$$

S_0 … součet čtverců odchylek od celkového průměru:

$$S_0 = \sum_i \sum_j (x_{ij}^2) - \frac{T^2}{N} = S_A + S_R \text{ , kde}$$

S_A představuje rozptyl mezi jednotlivými úrovněmi faktoru A:

$$S_A = \sum_{i=1}^{k} \left(\frac{T_i^2}{n_i}\right) - \frac{T^2}{N}$$

S_R je reziduální (zbytkový) rozptyl uvnitř jednotlivých úrovní a vypočte se jako rozdíl $S_0 - S_A$. Odhadem rozptylu chyb σ^2_ε je průměrný reziduální čtverec MS_R:

$$MS_R = \frac{S_R}{N - k}$$

T … součet všech hodnot v matici
T_i … sloupcové součty

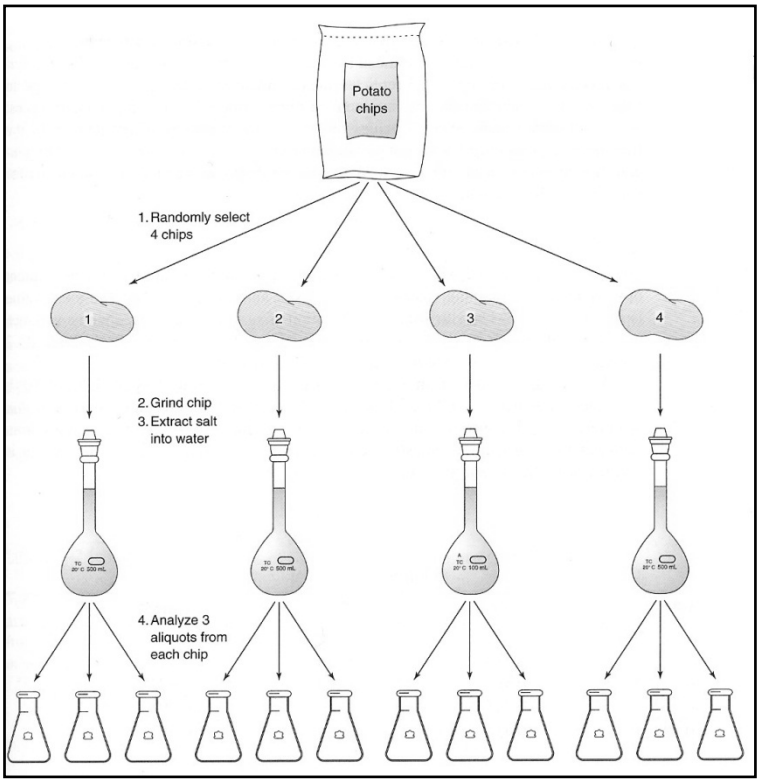
Formulace hypotéz: H_0: $\alpha_i = 0$; H_1: $\alpha_i \neq 0$

Testační statistika F_A (pro faktor A):

$$F_A = \frac{MS_A}{MS_R} = \frac{S_A / (k-1)}{S_R / (N-k)}$$

Při platnosti H_0 má F_A statistika F-rozdělení s (k-1, N-k) stupni volnosti. Vyjde-li F_A větší než kvantil $F(krit)_{(1-\alpha,\ k-1,\ N-k)}$, je nutné H_0 na hladině významnosti α zamítnout a vliv úrovní faktoru α_i je nenulový.

Ilustrační příklad ANOVA– stanovení obsahu NaCl v chipsech:



Tabulka výsledků (% Na) v chipsech:

	Lupínek 1	Lupínek 2	Lupínek 3	Lupínek 4
1. stanovení	0,324	0,455	0,420	0,447
2. stanovení	0,311	0,467	0,463	0,377
3. stanovení	0,352	0,448	0,424	0,398
Aritm. průměr	***0,329***	***0,457***	***0,436***	***0,407***
Směrodatná odchylka	***0,021***	***0,0096***	***0,0238***	***0,0359***
Sloupcový součet	***0,987***	***1,37***	***1,307***	***1,222***

N = 12; počet úrovní faktoru k = 4; n = 3

GRAFY (V QC EXPERTU)

- GRAF ANOVA – zobrazuje polohu měřených dat v jednotlivých úrovních. Lze vizuálně posoudit rozdíly a rozptyl.
- KRABICOVÝ GRAF – zobrazí se pro každou úroveň faktoru; k identifikaci OB.

Základní předpoklady (normalitu) lze ověřit:

- *Testem normality*
- *Q-Q graf Jackknife reziduí (odchylek od celkového průměru)* – v případě normálního rozdělení vznikne v grafu lineární závislost s nulovým úsekem a jednotkovou směrnicí.

VÍCENÁSOBNÉ POROVNÁVÁNÍ
(MULTIPLE COMPARSION PROCEDURE – MCP)

- Když ANOVA určí, že faktor A je statisticky významný, je možné nalézt úrovně faktoru A, které se významně liší od ostatních.

SCHEFFEHO POROVNÁNÍ

Vyšetřuje všechna možná porovnání k sloupcových průměrů. Princip spočívá v testování významnosti rozdílů jednotlivých sloupcových průměrů.

Např.: $\boxed{\begin{array}{l}\overline{X}_1 - \overline{X}_2 \cong 0 \\ \overline{X}_2 - \overline{X}_3 \cong 0 \\ \ldots\end{array}}$ a sledujeme zda IS jednotlivých rozdílů obsahují 0.

Testační kriterium má následující podobu:

$$\frac{\left|\overline{X}_i - \overline{X}_j\right|}{\sqrt{s^2\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}} \geq \sqrt{(k-1)\,F_{krit(k-1,\,N-k)}}$$

Dvoufaktorová ANOVA

Provádí se experimenty na různých úrovních dvou faktorů A a B. Kombinace úrovní faktoru tvoří mřížkovou strukturu jejímž elementem je tzv. cela. Platí že cela [ij] odpovídá *i-té* úrovní faktoru A a *j-té* úrovni faktoru B. V každé cele je obecně n_{ij} opakování.

- Pokud je v každé cele jen 1 opakování = ANOVA bez opakování (2P).
- Pokud je v každé cele více než jedno opakování, ale ve všech celách stejný počet = vyvážená dvoufaktorová ANOVA (2B).
- Pokud je v každé cele více než jedno opakování, a počet se v celách liší = nevyvážená dvoufaktorová ANOVA (2U).

Podrobněji se budeme zabývat pouze ANOVou 2P.

Tabulka pro ANOVA 2P:

	B_1	B_2	…	B_m
A_1				
A_2		A_2B_2		
…				
A_k				

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

α_i … vliv *i-té* úrovně faktoru A
β_j … vliv *j-té* úrovně faktoru B

$$S_0 = S_A + S_B + S_R$$

$$S_0 = \sum_i \sum_j (x_{ij}^2) - \frac{T^2}{N} \quad \text{, kde N = k·m}$$

$$S_A = \frac{1}{m}\sum_{i=1}^{k} (Z_i^2) - \frac{T^2}{N} \qquad S_B = \frac{1}{k}\sum_{i=1}^{m} (T_i^2) - \frac{T^2}{N}$$

S_A představuje rozptyl mezi jednotlivými úrovněmi faktoru A, S_B pak mezi úrovněmi faktoru B. Význam S_0 a S_R je stejný jako u jednofaktorové ANOVy.

Z_i … součet hodnot v *i-té* úrovni faktoru A (řádkový součet)
T_j … součet hodnot v *j-té* úrovni faktoru B (sloupcový součet)

Formulace hypotéz:
H_0: $\alpha_i = 0$ a $\beta_j = 0$ (efekty úrovní faktorů A a B jsou nevýznamné)
H_1: $\alpha_i \neq 0$ a $\beta_j \neq 0$ (efekty úrovní faktorů A a B jsou významné)

Testovací kritéria:

$$F_A = \frac{MS_A}{MS_R} = \frac{S_A / k-1}{S_R / (k-1)(m-1)} \qquad F_B = \frac{MS_B}{MS_R} = \frac{S_B / m-1}{S_R / (k-1)(m-1)}$$

Za předpokladu platnosti H_0 má testační charakteristika F_A Fisher-Snedecorovo rozdělení s (k-1) a (k-1)(m-1) stupni volnosti a testační charakteristika F_B s (m-1) a (k-1)(m-1) stupni volnosti.

INTERAKCE FAKTORŮ

Rozptyl může být kromě efektu faktorů A a B ovlivněn i interakčním členem τ_{ij}, který je důsledkem různých kombinací řáskových a sloupcových efektů. Tzn., že efekty faktorů A a B nejsou ve svém vlivu na každý výsledek x_{ij} nezávislé.

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

Obvykle se užívá ***Tukeyův model interakce*** $\tau_{ij} = C \cdot \alpha_i \cdot \beta_j$, kde C je konstanta určovaná jako směrnice přímky v grafu závislosti reziduí na $\alpha_i \cdot \beta_j / \mu$.

Formulace hypotéz: H_0: $\tau_{ij} = 0$; H_1: $\tau_{ij} \neq 0$

NEPARAMETRICKÉ TESTY V ANOVA

KRUSKAL-WALLISŮV TEST

- Tento test je rozšířením Wilcoxonova testu pro porovnání mediánů více než dvou náhodných výběrů.
- Je alternativou pro jednofaktorovou ANOVA.
- Předpoklady pro použití:
  - rozdělení souborů (úrovní faktoru) musí být stejné,
  - rozptyly souborů (úrovní faktoru) musí být stejné.

Formulace hypotéz: H_0: „*mediány všech úrovní faktoru jsou stejné*“
H_1: „*alespoň jeden medián se liší od ostatních*“

POSTUP:

1. Všechny hodnoty v matici seřadíme od nejmenší do největší a přiřadíme jím pořadová čísla (včetně průměrných pořadí pro stejné hodnoty).
2. Pro každý výběrový soubor (úroveň faktoru) vypočítáme sumu pořadí R_1, R_2, ..., R_k (k je počet výběrových souborů – úrovní faktoru).
3. Určíme celkový rozsah výběru $N = n_1 + n_2 + ... + n_k$, kde n_i, … označuje počet hodnot pro každou úroveň faktoru.
4. Vypočteme testovací charakteristiku χ^2_{exp} pomocí následujícího vztahu:

$$\chi^2_{Kru} = \frac{12}{N^2 + N}\sum_{i=1}^{k}\left(\frac{R_i^2}{n_i}\right) - 3(N+1)$$

5. Porovnáváme s kritickou hodnotou $\chi^2_{krit}(0,95)$ s k - 1 stupni volnosti. Srovnání s hodnotou χ^2 rozdělení je možné použít, pokud je N > asi 15!

FRIEDMANŮV TEST

- Je neparametrickým testem pro dvoufaktorovou analýzu rozptylu (2P), faktor A má k úrovní a faktor B má m úrovní.
- **POSTUP:** totožný s Kruskal-Wallisovým testem, R_j jsou sumy pořadí sloupců. *Matice dat by měla být použita tak, aby rozptyl v řádcích byl menší než ve sloupcích (lze řešit záměnou faktorů – otočením matice).*

$$\chi^2_{Fri} = \frac{12}{km(k+1)}\sum_{j=1}^{m}(R_j^2) - 3k(m+1)$$

Porovnáváme s kritickou hodnotou $\chi^2_{krit}(0,95)$ s m - 1 stupni volnosti, pokud je k.m > 15.