Univerzita Palackého v Olomouci

Analýza rozptylu

Chemometrie I (ACH/CHEX1)

1

Univerzita Palackého v Olomouci

Úvod

- Analýza rozptylu (aj. Analysis of Variance – **ANOVA**) je pokročilá statistická metoda popsaná počátkem 20. stol. R. A. Fisherem.
- ANOVA je statistická metoda používaná k porovnání průměrů několika (více než dvou) základních souborů.
  - Název metody je matoucí. Metoda se nepoužívá k porovnání rozptylů ale průměrů.
- Název analýza rozptylu vyplývá ze skutečnosti, že určení, zda jsou či nejsou mezi průměry několika základních souborů rozdíly, se zakládá na analýze různých forem rozptylu spojeného s náhodnými výběry.
- Používá se buď jako samostatná technika, nebo jako postup, umožňující analýzu zdrojů variability v lineární regresi. Příklady použití:
  - k porovnání středních hodnot více než 2 souborů, např. porovnání shody více než dvou metod stanovení analytu, účinku více než 2 léčiv na dané onemocnění, účinku více než 2 hnojiv na výnos, …
  - určení vlivu způsobu přípravy vzorků (několika způsoby),
  - zpracování mezilaboratorních porovnávacích zkoušek (MPZ).

2

Univerzita Palackého v Olomouci

Úvod

- PODSTATA metody: rozklad celkového rozptylu na rozptyl vyvolaný vlivem jednotlivých faktorů (objasněná složka rozptylu) a náhodný rozptyl (neobjasněná složka rozptylu).
- Předmětem statistického testování je statistická významnost poměru mezi rozptylem způsobeným faktorem/faktory a náhodným rozptylem.
- Máme-li 1 faktor = jednofaktorová ANOVA, v případě 2 faktorů = dvoufaktorová ANOVA.
- Předpoklady pro použití ANOVA:
  - data pocházejí z normálního rozdělení, náhodné chyby ε_{ij} jsou náhodné veličiny s nulovou střední hodnotou,
  - rozptyly sloupců dat (v tabulce ANOVA) jsou stejné,
  - každý sloupec je náhodným výběrem ze svého základního souboru.
- Ověření předpokladu normality v QC Expertu:
  - Q-Q graf Jackknife reziduí (odchylek od celkového průměru) – v případě normálního rozdělení vznikne v grafu lineární závislost s nulovým úsekem a jednotkovou směrnicí.

3

Univerzita Palackého v Olomouci

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).
- Z tabulky ANOVA lze vytvořit model ve tvaru: $x_{ij} = \mu + \alpha_i + \varepsilon_{ij}$

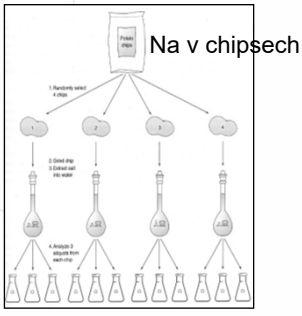
  μ … celkový aritmetický průměr všech hodnot v matici $\bar{\bar{x}}$,

  α_i … efekt i-té úrovně faktoru A popsaný sloupcovým průměrem $\bar{x}_i$,

  ε_{ij} … náhodná chyba.
- Jednofaktorová ANOVA porovnává střední hodnoty (průměry) dvou či více úrovní faktoru ***A*** čili sloupců v tabulce ANOVA za účelem určit, zda alespoň jedna sloupcový průměr ($\bar{x}_i$) se liší od ostatních.
- H_0: „Všechny střední hodnoty jsou stejné“ (H_0: $\alpha_i = 0$).
- H_1: „Alespoň jedna střední hodnota se odlišuje od ostatních“ (H_1: $\alpha_i \neq 0$).

A_1	A_2	$A_.$	A_k
x_{11}	x_{21}	$x_{..}$	x_{k1}
x_{12}	x_{22}	$x_{..}$	x_{k2}
x_{13}	x_{23}	x_{ij}	x_{k3}
x_{14}	x_{24}		x_{k4}
x_{15}			x_{k5}

Tabulka ANOVA



4

Univerzita Palackého v Olomouci

Jednofaktorová ANOVA

- Odvození modelu ANOVA:

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

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

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

sumací přes řádky (i) a sloupce (j) získáme

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

což lze zjednodušit na

$$S_0 = S_A + S_R + 2 \cdot 0 = S_A + S_R$$

- Vzorce pro výpočet jednofaktorvé ANOVA na kalkulačce:

S_0 „představuje“ celkový rozptyl

$$S_0 = \sum_i \sum_j (x_{ij}^2) - \frac{T^2}{N}$$

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 „představuje“ reziduální (zbytkový, neobjasněný) rozptyl, tj. uvnitř úrovní faktoru A

$$S_R = S_0 - S_A$$

T … součet všech hodnot v tabulce ANOVA

T_i … sloupcový součet

5

Univerzita Palackého v Olomouci

Jednofaktorová ANOVA

- Testační statistika pro faktor *A*:

$$F_A = \frac{\frac{S_A}{(k-1)}}{\frac{S_R}{(N-k)}} = \frac{S_A \cdot (N-k)}{S_R \cdot (k-1)}$$

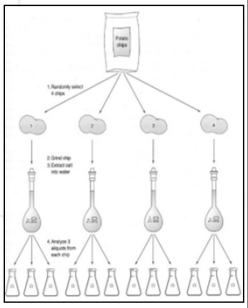
Při platnosti H_0 je $F_A < F_{krit(\alpha,\ k-1,\ N-k)}$, pokud je $F_A < F_{krit(\alpha,\ k-1,\ N-k)}$, je nutné H_0 na hladině významnosti α zamítnout a průměry jsou rozdílné.

	Chips 1	Chips 2	Chips 3	Chips 4
	0,324	0,455	0,420	0,447
	0,311	0,467	0,463	0,377
	0,352	0,448	0,424	0,398
$\bar{x}_i$	*0,329*	*0,457*	*0,436*	*0,407*
T_i	*0,987*	*1,370*	*1,307*	*1,222*

% Na v chipsech

k = 4, n_i = 3
N = 12

T = 0,987 + 1,370 + 1,307 + 1,222 = 4,886



6

Univerzita Palackého v Olomouci

Párové porovnání v jednofaktorové ANOVA

- Když ANOVA určí, že faktor ***A*** je statisticky významný, je pomocí párového porovnání (multiple comparison procedure – MCP) možné nalézt ty úrovně faktoru ***A***, které se liší od ostatních. Je tedy možné určit, který sloupcový(é) průměr(y) ($\bar{x}_i$) se liší od ostatních.
- Při MCP se každý $\bar{x}_i$ porovnáván postupně se všemi ostatními.
  - Např. pro k = 3 jde o tato porovnání $\bar{x}_1$ s $\bar{x}_2$, $\bar{x}_1$ s $\bar{x}_3$ a $\bar{x}_2$ s $\bar{x}_3$.
- Existuje více metod MCP, QC.Expert používá **Scheffeho porovnání**.
  - Testační kritérium má následující podobu pro případ, kdy se $\bar{x}_i$ a $\bar{x}_j$ od sebe liší:

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

Princip: testování významnosti rozdílů jednotlivých sloupcových průměrů, př. pro k = 3 $\bar{x}_1 - \bar{x}_2$, $\bar{x}_1 - \bar{x}_3$ a $\bar{x}_2 - \bar{x}_3$. Je-li rozdíl statistický významný (tj. odlišný od 0), tak IS tohoto rozdílu nebude obsahovat 0. Je-li rozdíl statisticky nevýznamný (tj. rovný 0), tak IS rozdílu bude obsahovat 0.

7

Univerzita Palackého v Olomouci

Dvoufaktorová ANOVA

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

	B_1	B_2	$B_.$	B_m
A_1	.	.	.	.
A_2	.	.	.	.
$A_.$	.	.	x_{ij}	.
A_k	.	.	.	.

ANOVA bez opakování (2P)

	B_1	B_2	$B_.$	B_m
A_1	…	…	…	…
A_2	…	…	…	…
$A_.$	…	…	…	…
A_k	…	…	…	…

Vyvážená ANOVA (2B)

	B_1	B_2	$B_.$	B_m
A_1	..	…	….	….
A_2	…	…	….	..
$A_.$	..	.	..	…
A_k	….	..	…	.

Nevyvážená ANOVA (2U)

- Pokud se kromě řádkových α_i a sloupcových β_j efektů uplatňuje i vliv různých kombinací sloupcových a řádkových efektů, vyskytuje se také interakční člen τ_{ij}.

8

Univerzita Palackého v Olomouci

Dvoufaktorová ANOVA

- Podrobně (pro výpočty na kalkulačce) se budeme zabývat pouze ANOVA 2P.
- Model ANOVA 2P bez interakce: $x_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$

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

  $S_0 = S_A + S_B + S_R$
- 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)

$$S_0 = \sum_i \sum_j (x_{ij}^2) - \frac{T^2}{N} \qquad S_A = \frac{1}{m}\sum_{i=1}^{k}(Z_i^2) - \frac{T^2}{N} \qquad S_B = \frac{1}{k}\sum_{i=1}^{m}(T_j^2) - \frac{T^2}{N}$$

- **S_A „představuje“ rozptyl mezi jednotlivými úrovněmi faktoru *A*,**
- **S_B „představuje“ rozptyl mezi jednotlivými úrovněmi faktoru *B*,**
- **S_0 „představuje“ celkový rozptyl, S_R „představuje“ reziduální rozptyl**
- **$N = k \cdot m$,**
- **Z_i … řádkový součet (součet *i-té* úrovně f. *A*), T_j … sloupcový součet (součet *j-té* úrovně f. *B*).**

9

Univerzita Palackého v Olomouci

Dvoufaktorová ANOVA

- Testační statistiky pro ANOVA 2P bez interakce:

$$F_A = \frac{\frac{S_A}{k-1}}{\frac{S_R}{(k-1)\cdot(m-1)}} \qquad F_B = \frac{\frac{S_B}{m-1}}{\frac{S_R}{(k-1)\cdot(m-1)}}$$

- Za předpokladu platnosti H_0 je $F_A < F_{krit,\alpha}$ s (k-1) a (k-1)(m-1) stupni volnosti a $F_B < F_{krit,\alpha}$ s (m-1) a (k-1)(m-1) stupni volnosti. V opačných případech ($F_i > F_{krit}$) přijímáme H_1.

- INTERAKCE FAKTORŮ
  - Jak již bylo uvedeno, kromě řádkových α_i a sloupcových β_j efektů se může uplatnit vliv různých kombinací sloupcových a řádkových efektů, vyskytuje se také interakční člen τ_{ij}.
  - Model ANOVA pak je: $x_{ij} = \mu + \alpha_i + \beta_j + \tau_{ij} + \varepsilon_{ij}$.
  - Obvykle se užívá ***Tukeyův model interakce*** – $\tau_{ij} = C \cdot \alpha_i \cdot \beta_j$, kde C je konstanta. To vede k rovnici pro celkový rozptyl $S_0 = S_A + S_B + S_I + S_R$. Model s interakcí používá QC-Expert v modulu dvoufaktorové ANOVA a T_i v něm „představuje“ rozptyl příslušející interakci.

10

Univerzita Palackého v Olomouci

Neparametrické testy v ANOVA

- **KRUSKALŮV-WALLISŮV TEST**
  - Je neparametrickou alternativou pro jednofaktorovou ANOVA.
  - Jedná se o rozšíření Wilcoxonova pořadového testu pro porovnání mediánů více než dvou náhodných výběrů.
  - Předpoklady pro použití:
    - rozdělení souborů musí být stejné,
    - rozptyly souborů musí být stejné,
    - Všechny sloupce představují náhodné výběry svých základních souborů.
- **FRIEDMANŮV TEST**
  - Je neparametrickou variantou dvoufaktorové analýzy rozptylu (varianty 2P), faktor A má *k* úrovní a faktor B má *m* úrovní.
  - Postup je obdobný s Kruskalovým-Wallisovým testem.
  - V rámci ACH/CHEX1 nebudeme řešit příklady na tento test.

11

Univerzita Palackého v Olomouci

Kruskalův-Wallisův test

- Formulace hypotéz: H_0: „mediány všech úrovní faktoru jsou stejné“
  H_1: „alespoň jeden medián se liší od ostatních“
- POSTUP:
  1. Všechny hodnoty v tabulce ANOVA seřadíme podle velikosti a přiřadíme jím pořadová čísla (včetně průměrných pořadí pro stejné hodnoty).
  2. Pro každou úroveň faktoru A (výběrový soubor) vypočítáme sumu pořadí R_1, R_2, ..., R_k (*k* je počet úrovní faktoru).
  3. Určíme celkový rozsah výběru $N = n_1 + n_2 + ... + n_k$, tj. součet hodnot všech úrovní faktoru A.
  4. Vypočteme testovací charakteristiku χ^2_{Kru} 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.
  6. Je-li $\chi^2_{Kru} < \chi^2_{krit}$, přijímáme H_0. Pokud je $\chi^2_{Kru} > \chi^2_{krit}$, zamítáme H_0 a přijímáme H_1.
- Test je vhodné používat, pokud je N > asi 15!

12