Statistické zpracování vodohospodářských dat 2. Metodologie počítačové analýzy rozptylu, ANOVA

Milan Meloun

Klíčová slova
ANOVA, jednofaktorová analýza rozptylu, dvoufaktorová analýza rozptylu, Fisher-Snedecorův F-test, variabilita, homogenita rozptylu,

Souhrn

Analýza rozptylu, ANOVA umožňuje analýzu významnosti zdrojů variability u statistických modelů dat. Podstatou metody je rozklad celkového rozptylu dat na složky objasněné čili známé zdroje variability a složku neobjasněnou čili náhodný šum, následovaný testy hypotéz o významnosti jednotlivých zdrojů variability. Postup lze rozdělit do kroků: 1. Odhad parametrů modelu ANOVA. 2. Testování významnosti modelu. 3. Konstrukce zpřesněných modelů a testy jejich významnosti. 4. Ověření normality, homogenity rozptylů a indikace silně vybočujících pozorování. 5. Interpretace výsledků s ohledem na zadání dat.

1 Úvod

Analýza rozptylu, označovaná ANOVA (z anglického **An**alysis **o**f **V**ariance), se v technické praxi používá buď jako samostatná technika nebo jako postup umožňující analýzu zdrojů variability u statistických modelů. ANOVA jako samostatná technika umožňuje posouzení významnosti zdrojů variability v datech, vlivu přípravy vzorků na výsledek analýzy, vlivu typu přístroje, lidského faktoru a obsluhy na výsledek měření. Podstatou analýzy rozptylu je rozklad celkového rozptylu dat na *složky objasněné*, jež představují známé zdroje variability a *složku neobjasněnou*, náhodnou čili šum. Následně se testují hypotézy o významnosti jednotlivých zdrojů variability. Podle konkrétního uspořádání experimentu existuje řada variant analýzy rozptylu. Přehled základních technik lze nalézt v řadě článků [1,2] a monografií[3-6]. Často se ANOVA vyskytuje v technické praxi v souvislosti s technikami plánovaných experimentů. Omezíme se zde na jednodušší techniky, vhodné k řešení běžných vodohospodářských úloh.

2 Základní pojmy

Historicky se analýza rozptylu začala rozvíjet zejména při vyhodnocování dat v zemědělství. Její terminologie je proto poněkud speciální. Vedle kvalitativních faktorů se vyskytují také *faktory kvantitativní*, jako jsou fyzikální a chemické veličiny. Jednotlivé faktory se vyskytují na jistých *úrovních* Z_1, Z_2, Z_3, jež se označují jako *zpracování*. Tyto úrovně mohou být opět kvalitativní nebo kvantitativní. Zdrojem variability výsledků měření y_{ij} jsou jednotlivé úrovně faktoru. Tomu odpovídá jednoduchý model $y_{ij} = \mu_i + \varepsilon_{ij}$, kde μ_i je skutečná hodnota výsledků analýz a ε_{ij} pak označuje náhodnou chybu. Veličina μ_i se skládá ze složky odpovídající celkovému průměru μ ze všech úrovní faktoru a *efektu i-té* úrovně daného faktoru α_i, tj. $\mu_i = \mu + \alpha_i$, kde μ_i je střední hodnota pro *i*-tou úroveň. Účelem analýzy rozptylu je testování shody jednotlivých úrovní, čili nulové hypotézy H_0: $\mu_1 = \mu_2 = \mu_3$, nebo jinak vyjádřeno významnosti efektů α_i čili nulové hypotézy H_0: $\alpha_1 = \alpha_2 = \alpha_3 = 0$. Pokud jsou předmětem zájmu pouze rozdíly mezi danými úrovněmi, jde o modely s *pevnými efekty*. Pokud jsou jednotlivé úrovně pouze *výběrem* z konečného či nekonečného souboru, jde o modely s *náhodnými efekty*. Výběr mezi pevnými a náhodnými efekty závisí na vlastním záměru analýzy rozptylu a může se podle něho měnit. Je-li sledován pouze jeden faktor, jde o *jednofaktorovou analýzu rozptylu*, čili třídění dle jednoho faktoru. Často se však sleduje i vliv několika faktorů, kdy jde o *vícefaktorovou analýzu rozptylu*. Jako u jednofaktorové analýzy rozptylu, můžeme provést rozklad μ_{ip} na celkovou střední hodnotu, složky α_i odpovídající efektům faktoru Z, složky β_{ip} odpovídající efektům faktoru L a interakce τ_{ip}, $\mu_{ij} = \mu + \alpha_i + \beta_j + \tau_{ij}$. Člen τ_{ip} označuje efekt *interakce* úrovní Z_i a L_j. Používá se v případech, kdy nelze objasnit variabilitu y_{ijk} pouze aditivním působením jednotlivých faktorů. Pro vlastní zpracování modelů analýzy rozptylu je důležité, zda je při všech kombinacích faktorů proveden stejný počet měření čili opakování. Kombinace úrovní jednotlivých faktorů, např. Z_i L_j se pak označuje jako *cela*. Pro stejný počet opakování ve všech celách se *experimenty* označují jako *vyvážené*, zatímco pro nestejný počet opakování jako *nevyvážené*. Postupy analýzy nevyvážených experimentů jsou komplikovanější a navíc může při extrémních rozdílech mezi počty opakování dojít při malých odchylkách od základních předpokladů, např. normality, ke značnému zkreslení výsledků testů [5].

3 Jednofaktorová analýza rozptylu

Při třídění podle jednoho faktoru se zkoumá jeho vliv na výsledek experimentu. Pro případ dvou úrovní jde o porovnání dvou výběrů. Zajímavý bude obecnější případ, kdy daný faktor A má celkem K různých úrovní A_1, ..., A_K. Na každé úrovni A_i je provedeno n_i měření $\{y_{ij}\}$, $j = 1, ..., n_i$. Celkový počet měření je

$$N = \sum_{i=1}^{K} n_i$$

Přehlednější je uspořádání dat v **Tabulce 1.**

Tabulka 1. Uspořádání dat pro jednofaktorovou analýzu rozptylu

	Úroveň faktoru						
	A_1	A_2	...	A_i	...	A_K	Celek
	y_{11}	y_{21}	...	y_{i1}	...	y_{K1}	
	y_{12}	y_{22}	...	y_{i2}	...	y_{K2}	
	...	...	...	...	...	...	
	...	...	...	...	...	...	
Opakování	...	...	...	...	...	...	
měření	y_{1n_1}	y_{2n_2}	....	y_{in_i}		y_{Kn_K}	
Průměry	$\hat{\mu}_1$	$\hat{\mu}_2$	...	$\hat{\mu}_i$	...	$\hat{\mu}_K$	$\hat{\mu}$
Počet	n_1	n_2	...	n_i	...	n_K	N

Sloupcový průměr $\hat{\mu}_i$ představuje součet prvků sloupce pro A_i dělený počtem opakování n_i, $\hat{\mu}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}$. *Celkový průměr* $\hat{\mu}$ je součet všech hodnot dělený celkovým počtem dat $\hat{\mu} = \frac{1}{K}\sum_{i=1}^{K}\hat{\mu}_i$. Pro výpočet odhadů efektů α_i lze pak použít vztah $\hat{\alpha}_i = \hat{\mu}_i - \hat{\mu}$. Při zavedení μ_i vznikne přeurčený model, obsahující o jeden parametr více. Proto se při odhadu efektů α_i používá ještě jedna omezující podmínka $\sum_{i=1}^{K} n_i\alpha_i = 0$. Pro případ vyvážených experimentů lze použít zjednodušenou podmínku $\sum_{i=1}^{K}\alpha_i = 0$. Vlastní analýza rozptylu, tj. rozklad celkového rozptylu, závisí také na tom, zda jde o modely s pevnými nebo náhodnými efekty.

Základním předpokladem statistické analýzy je fakt, že náhodné chyby ε_{ip} jsou nezávislé a náhodné veličiny s normálním rozdělením $N(0, \sigma^2)$. Střední hodnota chyb je rovna nule a rozptyl σ^2 je konstantní. Součet čtverců odchylek od celkového průměru $\hat{\mu}$, definovaný vztahem $S_c = \sum_{i=1}^{K}\sum_{j=1}^{n_i}(y_{ij} - \hat{\mu})^2$ se rozloží s využitím μ_i na dvě složky

$$S_c = \sum_{i=1}^{K}\sum_{j=1}^{n_i}[(y_{ij} - \hat{\mu}_i) + (\hat{\mu}_i - \hat{\mu})]^2 = S_A + S_R,$$

kde S_A představuje součet čtverců odchylek mezi jednotlivými úrovněmi daného faktoru $S_A = \sum_{i=1}^{K} n_i(\hat{\mu}_i - \hat{\mu})^2$ a S_R je reziduální součet čtverců odchylek uvnitř jednotlivých úrovní, $S_R = \sum_{i=1}^{K}\sum_{j=1}^{n_i}(y_{ij} - \hat{\mu}_i)^2$. Jednotlivé součty čtverců resp. složky rozptylu se zapisují do tabulky, která má pro jednofaktorovou analýzu rozptylu s pevnými efekty tvar **Tabulky 2.**

≠ zmenšit zdrojku

Poslední sloupec tabulky obsahuje očekávanou hodnotu průměrného čtverce. Nevychýleným odhadem rozptylu chyb σ_e^2 je průměrný reziduální čtverec $\sigma_e^2 = \frac{S_R}{N - K}$. Cílem je především testování, zda jsou efekty α_i nulové, tedy zda jednotlivé úrovně daného faktoru vedou ke statisticky nevýznamným rozdílům ve výsledcích. Nulová hypotéza H_0: $\alpha_i = 0$, $i = 1, ..., K$, se ověřuje proti alternativní hypotéze H_A: $\alpha_i \neq 0$, $i = 1, ..., K$. Při testování se využívá faktu, že veličina S_A/σ_e^2 má

Tabulka 2. Tabulka analýzy rozptylu pro jednoduché třídění u modelu s pevnými efekty

Součet čtverců	Počet stupňů volnosti	Průměrný čtverec	Očekávaná hodnota
Mezi úrovněmi S_A	$K - 1$	$\frac{S_A}{K-1}$	$\sigma_e^2 + \frac{\sum_{i=1}^{K} n_i \alpha_i^2}{K-1}$
Reziduální S_R	$N - K$	$\frac{S_R}{N-K}$	σ_e^2
Celkový S_c	$N - 1$	-	-

χ^2-rozdělení s $(K - 1)$ stupni volnosti a veličina S_R / σ_e^2 má nezávislé χ^2-rozdělení s $(N - K)$ stupni volnosti. Jejich podíl má pak F-rozdělení s $(K - 1)$ a $(N - K)$ stupni volnosti. Testovací Fisherova statistika F_e má tvar $F_e = \frac{S_A\ (N-K)}{S_R\ (K-1)}$. Při platnosti nulové hypotézy H_0 má F_e statistika Fisherovo F-rozdělení s $(K - 1)$ a $(N - K)$ stupni volnosti. Vyjde-li F_e větší než kvantil Fisherova rozdělení $F_{1-\alpha}(K - 1, N - K)$, je nutné nulovou hypotézu H_0 na hladině významnosti α zamítnout a efekty považovat za nenulové a statisticky významné.

3.1 Technika vícenásobného porovnání

Pokud vyjde vliv jednotlivých efektů jako statisticky významný, jsou rozdíly mezi průměry μ_i, μ_j, $i \neq j$ rovněž významné. Pro hlubší analýzu se používá řady metod, například Scheffého metoda vícenásobného porovnání, pro kterou se zamítá hypotéza H_0: $\mu_i = \mu_j$ pro všechny dvojice (i, j), pro které platí

$$\left|\hat{\mu}_i - \hat{\mu}_j\right| \geq \sqrt{(K-1)\ \hat{\sigma}^2\ F_{1-\alpha}(K-1,\ N-K)\left[\frac{1}{n_i} + \frac{1}{n_j}\right]}$$

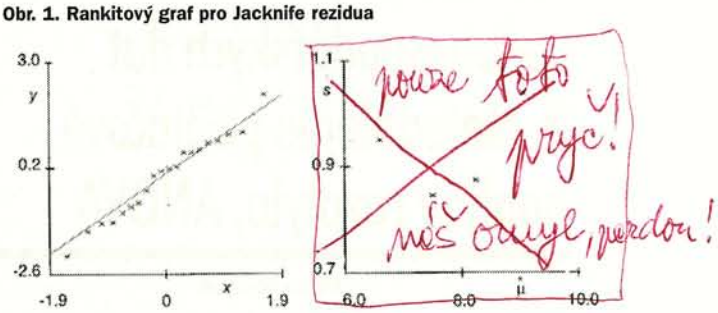
kde $\hat{\sigma}^2$ je reziduální rozptyl $\hat{\sigma}_e^2$.

Tento vztah se používá pro všechny možné dvojice indexů (i, j). V některých případech je třeba testovat pouze zvolený *lineární kontrast* q definovaný vztahem $q = \sum_{i=1}^{K} C_i \mu_i$ se známými konstantami C_i, pro které platí $\sum_{i=1}^{K} C_i = 0$, $\sum_{i=1}^{K} C_i^2 > 0$. Odhadem lineárního kontrastu q je veličina $\hat{q} = \sum_{i=1}^{K} C_i\ \hat{\mu}_i$. Mají-li výsledky měření y_{ij} normální rozdělení $N(\mu_i, \sigma^2)$, lze testovat nulovovu hypotézu H_0: $q = 0$ pomocí statistiky $F_q = \frac{\hat{q}^2}{\hat{\sigma}^2 \sum_{i=1}^{K} \frac{C_i^2}{n_i}}$. Při platnosti nulové hypotézy H_0 má tato testovací statistika F-rozdělení s 1 a $(N - K)$ stupni volnosti. Hypotéza H_0 se zamítá, pokud F_q je větší než kvantil $F_{1-\alpha}(1, N - K)$. Dosavadní postupy analýzy rozptylu jsou správné jen za předpokladu, když jednotlivé hodnoty y_{ij} jsou vzájemně nezávislé, a když chyby ε_{ij} mají normální rozdělení s konstantním rozptylem. V praxi však bývá důležité tyto předpoklady rovněž ověřit.

3.2 Ověření normality chyb

Pro posouzení normality chyb lze použít především rankitové grafy (**obr. 1.**). Výhodné je v těchto grafech užití standardizovaných reziduí $\hat{e}_{Si} = \frac{\hat{e}_{ij}}{\hat{\sigma}\sqrt{1 - \frac{1}{n_i}}}$. V případě platnosti předpokladů klasické analýzy rozptylu mají standardizovaná rezidua přibližně normální rozdělení $N(0, 1)$. Pokud platí podmínka, že $\varepsilon_{ij} \sim N(0, \sigma^2)$, vznikne v rankitovém grafu lineární závislost s nulovým úsekem a jednotkovou směrnicí. V řadě případů je možné zlepšit rozdělení dat ve smyslu přiblížení k normalitě s využitím vhodné transformace. Častým případem je, že data jsou zešikmená směrem k vyšším hodnotám. Pak je vhodné použít např. *posunutou logaritmickou transformaci* $y^* = \ln(y + C)$. Optimální hodnota C se volí tak, aby rezidua byla přibližně symetrická se špičatostí blízkou hodnotě Gaussova rozdělení tj. třem. Pro účely identifikace vybočujících hodnot je však výhodné použít Jackknife reziduí e_{Jij}, která jsou definována vztahem $\hat{e}_{Jij} = \hat{e}_{Sij}\sqrt{\frac{N-K-1}{N-K-\hat{e}_{Sij}^2}}$. Za předpokladu normality vykazují tato rezidua Studentovo rozdělení s $(N - K - 1)$ stupni volnosti. Orientačně platí, že pokud $\hat{e}_{Jij}^2 > 10$, lze danou hodnotu y_{ij} považovat za velmi silně vybočující.

Obr. 1. Rankitový graf pro Jacknife rezidua



3.3 Ověření konstantnosti rozptylu (homoskedasticity)

Předpoklad konstantnosti rozptylu (homoskedasticity) lze ověřit stejnými metodami jako u lineárních regresních modelů. U nevyvážených plánů je třeba uvažovat s nekonstantností rozptylu klasických reziduí způsobem nestejného počtu měření na jednotlivých úrovních. Pokud je k dispozici dostatečný počet opakování při jednotlivých úrovních daného faktoru, lze kromě průměru $\hat{\mu}_i$ počítat také výběrové rozptyly s_i^2. Předpoklad konstantnosti rozptylu lze pak ověřit na základě grafu s_i vs. $\hat{\mu}_i$. Pokud vznikne náhodný shluk bodů, lze považovat předpoklad shody rozptylů u všech úrovní za přijatelný. Jinak je možné použít vhodnou transformaci stabilizující rozptyl.

4 Dvoufaktorová analýza rozptylu

Při dvoufaktorové analýze rozptylu se provádí experimenty na různých úrovních dvou faktorů A a B. Kombinace úrovní faktorů tvoří typickou mřížkovou strukturu, jejímž elementem je tzv. ***cela***. Platí, že (i, j)-tá cela odpovídá kombinaci úrovně A_i faktoru A a B_j faktoru B. Schematicky je mřížková struktura znázorněna v **Tabulce 3:** V každé cele je obecně n_{ij} pozorování. Často se však setkáváme s případem *bez opakování*, kdy v každé cele je pouze jediné pozorování, $n_{ij} = 1$. Pro případ analýzy rozptylu bez opakování dojde ke zjednodušení zápisu $y_{ij} = \mu_{ij} + \varepsilon_{ij}$, kde μ_{ij} lze rozložit tak, že kromě řádkových α_i a sloupcových β_j efektů se zde vyskytuje také interakční člen τ_{ij}. Tento člen je pak důsledkem různých kombinací sloupcových a řádkových efektů.

Tabulka 3. Uspořádání dat pro dvoufaktorovou analýzu rozptylu

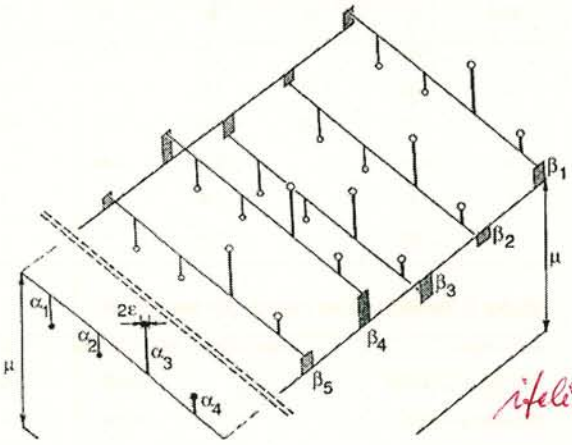
	B_1	B_2	...	B_M
A_1	.	.	...	
A_2	.		...	.
.	.	.	...	.
				cela A_2B_2
	.	.	...	.
			...	.
A_N	.	.	...	.

Nejjednodušším je *Tukeyův model interakce*, vyjádřený tvarem $\tau_{ij} = C\alpha_i\beta_j$, kde C je konstanta. Složitější jsou *řádkově lineární* modely interakcí, vyjádřené tvarem $\tau_{ij} = \gamma_i\beta_j C_R$ nebo *sloupcově lineární* modely interakcí ve tvaru $\tau_{ij} = \alpha_i C_K \delta_j$. Kompletnější je *aditivně-multiplikativní* model interakcí $\tau_{ij} = \gamma_i \delta_j C_W$. Uvedené vztahy obsahují kromě sloupcových a řádkových konstant δ_j a γ_i i obecné konstanty C_R, C_K, C_W. Omezme se zde pouze na nejjednodušší Tukeyův model interakce. Vzhledem ke své speciální definici obsahuje tento model pouze jeden parametr C a proto se označuje jako model *neaditivity s jedním stupněm volnosti*. Použití Tukeyova modelu interakce je výhodné zejména v případech, kdy je v každé cele pouze jedno pozorování, **obr. 2.**

Dvojné třídění je v průmyslové praxi nejužívanější. Umožňuje kvantifikovat vliv dvou faktorů na výsledek chemických analýz. Je účelné rozdělit úlohy dvojného třídění podle počtu opakování v jednotlivých celách. U modelů bez opakování měření obsahuje každá cela právě jednu hodnotu y_{ij}. O chybách

α_i lépe umístit.

Obr. 2. Schematické znázornění modelu dvojného třídění $\mu_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$, kde α_i je řádkový efekt a β_j sloupcový efekt a náhodná chyba ε_{ij} odpovídá poloměru kolečka.



italic

ε_{ij} se předpokládá, že jsou to nezávislé a shodně rozdělené náhodné veličiny s nulovou střední hodnotou a konstantním rozptylem. Při testování se navíc předpokládá, že rozdělení chyb je normální. Pokud v ANOVA modelu provedeme rozklad, je model analýzy rozptylu přeurčený. Proto se definují omezující podmínky $\sum_{i=1}^{N}\alpha_i = 0;\ \sum_{j=1}^{M}\beta_j = 0;\ \sum_{i=1}^{N}\tau_{ij} = 0;\ \sum_{j=1}^{M}\tau_{ij} = 0$

italic

V případě pouze aditivního působení jednotlivých faktorů je $\tau_{ij} = 0$ pro všechna i = 1, ..., N a j = 1, ..., M. Odhady parametrů μ, α_i, β_j lze pak určit ze vztahů $\hat{\mu} = \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij}$, $\alpha_i = \frac{1}{M}\sum_{j=1}^{M} y_{ij} - \hat{\mu}$, $\hat{\beta}_j = \frac{1}{N}\sum_{i=1}^{N} y_{ij} - \hat{\mu}$.

Pro rezidua $\hat{e}_{ij}$ platí $\hat{e}_{ij} = y_{ij} - \hat{\mu} - \hat{\alpha}_i - \hat{\beta}_j$. K určení interakcí můžeme využít skutečnosti, že $\tau_{ij} = E(y_{ij}) - \mu - \alpha_i - \beta_j$ a pro odhad interakcí platí přibližně $\hat{\tau}_{ij} \approx \hat{e}_{ij}$. Lze snadno identifikovat *Tukeyův model* interakce. Platí-li tento model, vyjde na grafu $\hat{e}_{ij}$ vs. $\hat{\alpha}_i \hat{\beta}_j$ lineární trend. Ze směrnice odpovídající regresní přímky se odhadne parametr *C*. Platí pro něj výraz

$$\hat{C} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M}\hat{e}_{ij}\,\hat{\alpha}_i\,\hat{\beta}_j}{\sum_{i=1}^{N}\sum_{j=1}^{M}\hat{\alpha}_i^2\,\hat{\beta}_j^2}$$

Graf $\hat{e}_{ij}$ vs. $\hat{\alpha}_i \hat{\beta}_j / \hat{\mu}$ se označuje jako *graf neaditivity*. Pokud vyjde v tomto grafu nenáhodný trend, znamená to, že je třeba uvažovat interakce.

Tabulka 4. Tabulka analýzy rozptylu pro dvojné třídění s interakcí Tukeyova typu

Součet čtverců pro	Stupně volnosti	Průměrný čtverec	Kritérium *F*
Faktor *A*, $S_A = M\sum_{i=1}^{N}\alpha_i^2$	$N - 1$	$M_A = S_A/(N-1)$	$F_A = M_A/M_{AB}$
Faktor *B*, $S_B = N\sum_{j=1}^{M}\beta_j^2$	$M - 1$	$M_B = S_B/(M-1)$	$F_B = M_B/M_{AB}$
Interakce (Tukey) S_T	1	$M_T = S_T$	$F_T = M_T/M_E$
Reziduální $S_R = S_{AB} - S_T$	$NM - N - M$	$M_E = S_R/(NM-N-M)$	-
Celkový $S_C = \sum_{i=1}^{N}\sum_{j=1}^{M}(\hat{\mu} - y_{ij})^2$	$NM - 1$	-	-

V **tabulce 4.** představuje S_T součet čtverců odchylek odpovídající Tukeyově interakci [3]

$$S_T = \frac{\left(\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij}\,\hat{\alpha}_i\,\hat{\beta}_j\right)^2}{\sum_{i=1}^{N}\sum_{j=1}^{M}\hat{\alpha}_i^2\,\hat{\beta}_j^2}$$

Symbol S_{AB} označuje reziduální součet čtverců pro případ bez interakcí $S_{AB} = \sum_{i=1}^{N}\sum_{j=1}^{M}(y_{ij} - \hat{\mu} - \hat{\alpha}_i - \hat{\beta}_j)^2$ a průměrný čtverec je $M_{AB} = \frac{S_{AB}}{(N-1)(M-1)}$. Hodnota M_{AB} je nevychýleným odhadem rozptylu σ^2. Pomocí *F*-kritéria lze opět provádět statistické testy. Začíná se testováním nulové hypotézy H_0: „Tukeyova interakce je nevýznamná", pro kterou lze použít testovací statistiku F_T z **tabulky 4.** Za předpokladu platnosti nulové hypotézy H_0 má tato testovací statistika *F*-rozdělení s 1 a (N M - N - M) stupni volnosti. Pokud nelze tuto hypotézu zamítnout, testuje se nulová hypotéza H_0: $\alpha_i = 0$, i = 1, ..., *N* (efekty řádků, čili faktoru *A*, jsou nevýznamné) pomocí statistiky F_A nebo nulová hypotéza H_0: $\beta_j = 0$, *j* = 1, ..., *M* (efekty sloupců, čili faktoru *B*, jsou nevýznamné) pomocí statistiky F_B. Obě tyto testovací statistiky jsou uvedeny v **Tabulce 4.** Za předpokladu platnosti hypotézy H_0 má statistika F_A *F*-rozdělení s $(N - 1)$ a $(N - 1)(M - 1)$ stupni volnosti a statistika F_B také *F*-rozdělení s $(M - 1)$ a $(N - 1)(M - 1)$ stupni volnosti. Pokud však vyjde F_T vyšší než odpovídající kvantil *F*-rozdělení, je efekt Tukeyovy interakce významný. V některých případech je výhodné provést eliminaci neaditivity s využitím mocninné transformace

italic

$$y_{ij}^{*} = \begin{cases} \dfrac{(y_{ij} + K)^{\lambda} - 1}{\lambda} & \text{pro } \lambda \neq 0 \\ \ln\,(y_{ij} + K) & \text{pro } \lambda = 0 \end{cases},$$

italic

kde *K* je vhodně volená konstanta zajišťující, aby platilo $(y_{ij} + K) > 0$, a kde λ je parametr transformace.

4.1 Vyvážené modely

Pro vyvážené modely platí, že v každé cele je $n_{ij} = n$ pozorování. Odhadem μ_{ij} jsou aritmetické průměry $\hat{\mu}_{ij} = \frac{1}{n}\sum_{k=1}^{n} y_{ijk}$. Pro odhady ostatních parametrů se použijí vztahy

$$\hat{\mu} = \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}\hat{\mu}_{ij},\quad \hat{\alpha}_i = \frac{1}{M}\sum_{j=1}^{M}\hat{\mu}_{ij} - \hat{\mu},\quad \hat{\beta}_j = \frac{1}{N}\sum_{i=1}^{N}\hat{\mu}_{ij} - \hat{\mu}.$$

Rezidua vyjádříme vztahem $\hat{e}_{ijk} = y_{ijk} - \hat{\mu} - \hat{\alpha}_i - \hat{\beta}_j$. Podobně lze i v tomto případu definovat odhad interakcí $\hat{\tau}_{ij} = \hat{\mu}_{ij} - \hat{\mu} - \hat{\alpha}_i - \beta_j$. Povšimněme si, že tento vztah se liší jen tím, že se místo veličin y_{ij} používá průměrů $\hat{\mu}_{ij}$. Pro ověření Tukeyova modelu interakce neaditivity lze vynášet graf $\hat{\tau}_{ij}$ vs. $\hat{\alpha}_i \hat{\beta}_j$. Náhodný obrazec bodů zde dokazuje aditivní působení obou faktorů. Součty čtverců modelu analýzy rozptylu pro obecný případ interakcí jsou uvedeny v **Tabulce 5.**

Tabulka 5. Tabulka analýzy rozptylu pro dvojné třídění a vyvážený experiment

Součet čtverců pro	Stupně volnosti	Průměrný čtverec	Kritérium *F*
Faktor *A*,			
$S_A = nM\sum_{i=1}^{N}\hat{\alpha}_i^2$	$N - 1$	$M_A = \frac{S_A}{N-1}$	$F_A = \frac{M_A}{M_E}$
Faktor *B*,			
$S_B = nN\sum_{j=1}^{M}\hat{\beta}_j^2$	$M - 1$	$M_B = \frac{S_B}{M-1}$	$F_B = \frac{M_B}{M_E}$
Interakce *AB*,			
$S_{AB} = n\sum_{i=1}^{N}\sum_{j=1}^{M}\hat{\tau}_{ij}^2$	$(N - 1)(M - 1)$	$M_{AB} = \frac{S_{AB}}{(N-1)(M-1)}$	$F_{AB} = \frac{M_{AB}}{M_E}$
Reziduální			
$S_E = \sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{k=1}^{n}(y_{ijk} - \hat{\mu}_{ij})^2$	$MN(n - 1)$	$M_E = \frac{S_E}{MN(n-1)}$	-
Celkový			
$S_C = \sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{k=1}^{n}(y_{ijk} - \hat{\mu})^2$	$MNn - 1$	-	-

Odpovídající průměrné hodnoty (očekávané hodnoty) průměrných čtverců jsou

$$E(M_A) = \sigma^2 + \frac{n M \sum_{i=1}^{N} \alpha_i^2}{(N-1)\,\sigma^2} = \sigma^2 + n M \sigma_A^2$$

$$E(M_B) = \sigma^2 + \frac{n N \sum_{j=1}^{M} \beta_j^2}{(M-1)\,\sigma^2} = \sigma^2 + n N \sigma_B^2$$

$$\text{a } E(M_{AB}) = \sigma^2 + \frac{n \sum_{i=1}^{N} \sum_{j=1}^{M} \tau_{ij}^2}{(N-1)\,(M-1)\,\sigma^2} = \sigma^2 + n \sigma_{AB}^2 .$$

Očekávaná hodnota $E(M_R) = \sigma^2$ ukazuje, že rozptyl M_R je nevychýleným odhadem σ^2 rozptylu chyb. Rozptyly σ_A^2, σ_B^2 a σ_{AB}^2 odpovídají efektům řádků, sloupců a interakcí. Těchto vztahů lze využít i v případech, kdy se hledají odhady rozptylů příslušející faktorům a interakcím. Pak se místo středních hodnot $E(.)$ dosazují přímo průměrné čtverce a místo rozptylu σ^2 reziduální rozptyl $\hat{\sigma}^2$. Důležité je, že průměrné čtverce nejsou přímo odhady odpovídajících rozptylů. Také v případě analýzy rozptylu definované **Tabulkou 5.** se využitím statistik F_{AB}, F_B a F_A testuje, zda je možné považovat sloupcové a řádkové efekty, resp. interakce, za nevýznamné. Pro test nulové hypotézy H_0: $\tau_{ij} = 0$, $i = 1, \ldots, N$ a $j = 1, \ldots, M$, lze použít testační statistiku F_{AB}, která má za předpokladu platnosti hypotézy H_0 F-rozdělení s $\{(N-1)(M-1)\}$ a $\{MN(n-1)\}$ stupni volnosti. Při testování významnosti řádkových efektů faktoru A je H_0: $\alpha_i = 0$, $i = 1, \ldots, N$. Pokud nulová hypotéza platí, má testovací F_A statistika F-rozdělení s $(N-1)$ a $\{MN(n-1)\}$ stupni volnosti. Analogicky při testování významnosti sloupcových efektů faktoru B je H_0: $\beta_j = 0$, $j = 1, \ldots, M$. Pokud nulová hypotéza platí, má testovací F_B statistika F-rozdělení s $(M-1)$ a $\{MN(n-1)\}$ stupni volnosti. Nevychýleným odhadem rozptylu je M_R. Výhodou vyvážených experimentů je fakt, že jednotlivé složky modelů analýzy rozptylu jsou vzájemně nezávislé. Proto je možno sčítat jednotlivé dílčí součty čtverců v **tabulce 5.** a **4.**, což umožňuje současné testování několika hypotéz. V podstatě lze tímto jednoduchým postupem ověřovat platnost různých submodelů analýzy rozptylu od nejjednoduššího $y_{ijk} = \mu + \varepsilon_{ijk}$ přes všechny dílčí (obsahující jen některé z parametrů α, β a τ). Sčítání součtů čtverců se doporučuje i v případech, kdy se vliv některých faktorů či interakcí prokáže jako nevýznamný. Pak se příslušný součet čtverců přičte k reziduálnímu a v modelu se odpovídající členy vypustí.

5 Souhrn: Postup při analýze rozptylu

Obecně lze postup analýzy rozptylu rozdělit do těchto kroků:

1. Provede se odhad parametrů základního modelu ANOVA.
2. Provede se testování jeho významnosti a konstrukce různých submodelů u modelů s pevnými efekty.
3. Provede se ověření předpokladů normality, homogenity rozptylů a přítomnosti silně vybočujících pozorování. Ne vždy se pro tyto účely hodí klasicky definovaná rezidua a využívají se proto i jiné typy reziduí.
4. Provede se interpretace výsledků s ohledem na zadání dat a jejich případné úpravy.

6 Ilustrativní příklady

Předložený postup analýzy rozptylu bude ilustrován na následujících příkladech:

Příklad 1. Testování kvality $AgNO_3$ od různých výrobců

U pěti lahví obsahujících $AgNO_3$, získaných od pěti dodavatelů, byla gravimetrickým stanovením chloridů sledována kvalita užité chemikálie. Z každé láhve byl proto odebrán různý počet vzorků k analýze (**Tabulka 6.**). Účelem je zjistit, zda existují významné rozdíly v kvalitě $AgNO_3$ od pěti dodavatelů.

Z údajů v tabulce lze určit odhady sloupcových průměrů $\hat{\mu}_i$, celkového průměru $\hat{\mu}$ a efektů $\hat{\alpha}_i$. Vyšlo $\hat{\mu} = 5.2715$, $\hat{\mu}_1 = 5.1417$; $\hat{\mu}_2 = 5.0833$; $\hat{\mu}_3 = 5.5483$; $\hat{\mu}_4 = 4.7375$; $\hat{\mu}_5 = 5.8467$; $\hat{\alpha}_1 = -0.1298$; $\hat{\alpha}_2 = -0.1882$; $\hat{\alpha}_3 = 0.277$; $\hat{\alpha}_4 = -0.534$; $\hat{\alpha}_5 = 0.575$. Jednotlivé součty čtverců a složky rozptylu jsou sumarizovány v **tabulce 7.** znázorněny na **obr. 3.** Pro $\alpha = 0.05$ je kvantil $F_{0.95}(4, 17) = 2.96$.

Je třeba ověřit platnost předpokladu normality, a to na základě rankitového grafu pro standardizovaná rezidua $\hat{e}_{Si}$ na **obr. 4.**

Závěr: Protože je F_r vyšší než kvantil $F_{0.95}(4, 17)$, je nutné zamítnout hypotézu H_0 a kvalita $AgNO_3$ od jednotlivých pěti dodavatelů se co do kvality statisticky významně liší. Z rankitového grafu je patrné, že lze přijmout předpoklad normality.

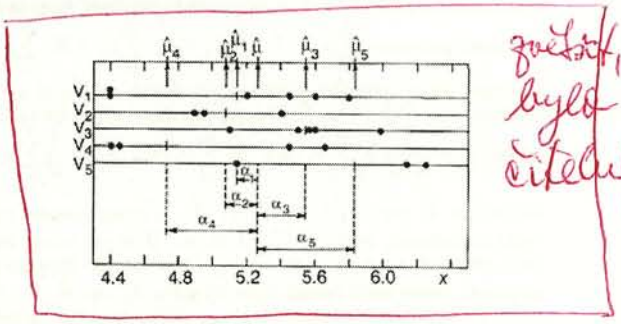
Tabulka 6. Procentuální obsah gravimetricky stanoveného chloru při užití $AgNO_3$ od pěti výrobců

	Výrobce				
Měření	V_1	V_2	V_3	V_4	V_5
1	4.40	4.90	5.55	4.45	5.15
2	4.40	4.95	5.10	5.45	6.25
3	5.20	5.40	5.50	4.65	6.14
4	5.45	-	5.98	4.40	-
5	5.80	-	5.60	-	-
6	5.60	-	5.56	-	-

Tabulka 7. Tabulka analýzy rozptylu pro obsah $AgNO_3$

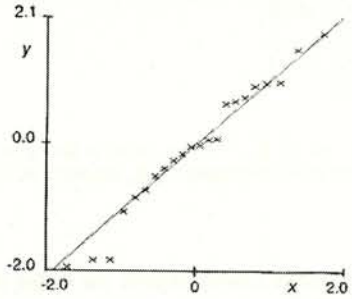
Součet čtverců	Stupně volnosti	Průměrný čtverec	F_r
Mezi výrobci $S_A = 2.8000$	4	0.6999	3.106
Reziduální $S_R = 3.8322$	17	0.2254	-
Celkový $S_c = 6.632$	21	-	-

Obr. 3. Znázornění dat, odhadů sloupcových průměrů μ_i, celkového průměru $\hat{\mu}$ a efektů $\hat{\alpha}_i$.



zvětšit, aby bylo čitelnější

Obr. 4. Rankitový graf pro standardizovaná rezidua.



Příklad 2. Stanovení obsahu vody v rozpouštědlech v různých laboratořích

Byly získány tři vzorky A_1, A_2 a A_3 nového rozpouštědla. U všech těchto vzorků byl ve čtyřech laboratořích B_1, B_2, B_3 a B_4 určen obsah vody. Je třeba rozhodnout, zda existují významné odchylky v obsahu vody v zadaných vzorcích rozpouštědla u výsledků sledovaných laboratoří. Vyčíslí se $\hat{\mu} = 1.2775$; $\hat{\alpha}_1 = -0.1475$; $\hat{\alpha}_2 = 0$; $\hat{\alpha}_3 = 0.1475$; $\hat{\beta}_1 = 0.1358$; $\hat{\beta}_2 = 0.0042$; $\hat{\beta}_3 = -0.0375$ a $\hat{\beta}_4 = -0.0942$.

Součet čtverců odchylek odpovídající Tukeyově interakci je roven $S_T = 0.0156$ a $S_{AB} = 0.02215$ a $M_{AB} = 0.003692$. Výsledky analýzy rozptylu dat **tabulky 8.** jsou uvedeny v **tabulce 9.**

Z tabulek F-rozdělení lze určit kvantily $F_{0.95}(1, 5) = 6.61$, $F_{0.95}(2, 5) = 5.79$ a $F_{0.95}(3, 5) = 5.41$.

Závěr: Z uvedených údajů plyne, že efekt interakce je nevýznamný a lze použít aditivní model analýzy rozptylu, zatímco efekty vzorků a laboratoří významné jsou. Mezi výsledky laboratoří a mezi vzorky rozpouštědla existují nenáhodné rozdíly. Graf neaditivity indikuje, že mocninná transformace by aditivitu zlepšila, i když ne statisticky významně.

Tabulka 8. Obsah vody [%] v rozpouštědle určený v různých laboratořích

Vzorek	Laboratoř			
	B_1	B_2	B_3	B_4
A_1	1.35	1.13	1.06	0.98
A_2	1.40	1.23	1.26	1.22
A_3	1.49	1.46	1.40	1.35

Tabulka 9. Tabulka analýzy rozptylu dat obsahu vody v rozpouštědlech

Součet čtverců pro	Stupně volnosti	Průměrný čtverec	Kritérium F
Vzorky A, S_A=0.174	2	0.087	19.64
Laboratoe B, S_B=0.0862	3	0.0287	6.49
Interakce Tukey, S_T= 0.0156	1	0.0156	3.522
Reziduální S_R= 0.0222	5	0.00443	-
Celkový S_C=0.2980	11	0.02568	-

Poděkování
Článek vznikl za finanční podpory vědeckého záměru Ministerstva školství mládeže a tělovýchovy č. MSM253100002.

7 Doporučená literatura:

[1] Searle S. R.: Biometrics **27**, 1 (1971).
[2] Bartlett M. S., Kendall D. G.: J. Roy. Stat. Soc. **B8**, 128 (1946).
[3] Scheffé H.: The Analysis of Variance. J. Wiley, New York 1959.
[4] Searle S. R.: Linear Models. J. Wiley, New York 1971.
[5] Miller P. G.: Beyond ANOVA, Basics of Applied Statistics. J. Wiley, New York 1986.
[6] Speed T. P.: Annals of Statist. **15**, 885 (1987).
[7] Emerson J. D., Hoaglin D. C., Kempthorne P. I.: J. Amer. Statist. Assoc. **79**, 329 (1984).
[8] Bradu D., Hawkins D. M.: Technometrics **24**, 103 (1982).
[9] Bloomfield P., Steiger W.: Least Absolute Deviations: Theory, Applications and Algorithms. Birkhäuser, Boston, 1983.
[10] Gabriel K. R.: J. R. Stat. Soc. **B40**, 186 (1978).
[11] Cressie N. A. C.: Biometrics **34**, 505 (1978).
[12] Potocký R a kol.: Zbierka úloh z pravdepodobnosti a matematickej štatistiky, ALFA-SNTL Bratislava 1986.
[13] Anderson R. L.: Practical Statistics for Analytical Chemists, van Nostrand Reinhold Comp., New York 1987.
[14] Miller J. C., Miller J. N.: Statistics for Analytical Chemistry, Ellis Horwood Chichester 1984.
[15] Liteanu C., Rica I.: Statistical Theory and Methodology of Trace Analysis, Ellis Horwood Chichester 1980.
[16] Rice J. A.: Mathematical Statistics and Data Analysis, Wadsworth & Brooks, California 1988, str. 397.
[17] Meloun M., Militký J.: Statistické zpracování experimentálních dat, Plus Praha 1994, resp. East Publishing Praha 1998, Academia Praha 2004.
[18] Meloun M., Militký J.: Kompendium statistického zpracování dat, Academia Praha 2002, Academia Praha 2006.
[19] ADSTAT 1.25, 2.0 a verze 3.0, TriloByte Statistical Software Pardubice, 1992, 1993, 1999.

Prof. RNDr. Milan Meloun, DrSc.
Katedra analytické chemie, Chemicko-technologická fakulta,
Univerzita Pardubice,
nám. Čs. Legií 565, 532 10 Pardubice,
Email: milan.meloun@upce.cz

Computer-Assisted Statistical Data Analysis: 1. Analysis of Variance (Meloun, M.)

Key Words
XXXXXXXXXXXXXX

Analysis of variance enables an analysis of various sources on the variability of data to determine which part of variation in a population is due to systematic reasons called factors and which is due to random effects. The profusion of instrumental techniques in analytical chemistry is such that often more than two possible techniques have to be compared. The techniques to be examined may be subject to systematic errors. The choice of a technique is called a controlled factor. Moreover, the results of the analytical determinations are subject to random errors. The ANOVA compares both causes of error, with the purpose of deciding whether or not the controlled factor has a significant effect.

Klíčová slova anglicky:

ANOVA, analysis of variance, one-way ANOVA, two-way ANOVA, Fisher-Snedecor F-test, multiple-comparison procedure, checking for normality, checking for homoscedasticity, two-way balanced models, two-way unbalanced models,