M5VM05 Statistické modelování

8. Analýza rozptylu

Jan Koláček (kolacek@math.muni.cz)

Ústav matematiky a statistiky, Přírodovědecká fakulta, Masarykova univerzita, Brno

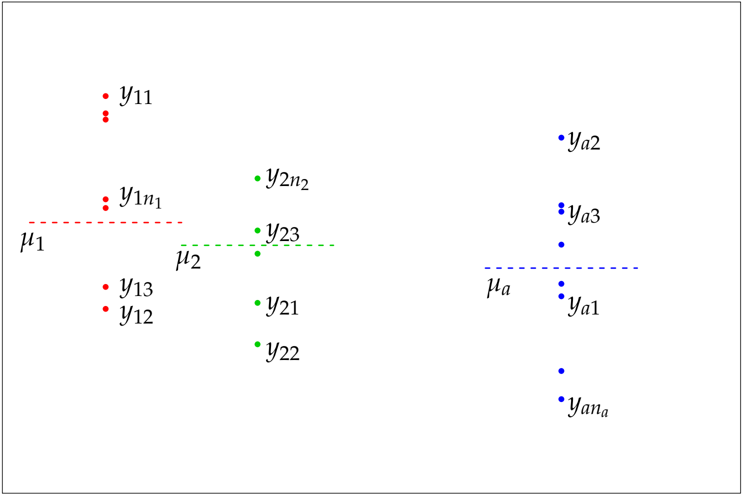
Motivace

Zajímáme se o problém, zda lze určitým faktorem (tj. nominální náhodnou veličinou A) vysvětlit variabilitu pozorovaných hodnot náhodné veličiny Y, která je intervalového či poměrového typu. Např. zkoumáme, zda metoda výuky určitého předmětu (faktor A) ovlivňuje počet bodů dosažených studenty v závěrečném testu (náhodná veličina Y).

Obecný popis

Předpokládáme, že faktor A má $a \geq 3$ úrovní a i-té úrovni odpovídá n_i výsledků $Y_{i1}, \dots, Y_{in_i}$, které tvoří náhodný výběr z rozložení $N(\mu_i, \sigma^2)$, $i = 1, \dots, a$ a jednotlivé náhodné výběry jsou stochasticky nezávislé, tedy $Y_{ij} = \mu_i + \varepsilon_{ij}$, kde ε_{ij} jsou stochasticky nezávislé náhodné veličiny s rozložením $N(0, \sigma^2)$, kde $i = 1, \dots, a$ a $j = 1, \dots, n_i$.

Graficky



Úroveň: 1 2 … a

Obecný popis

Na hladině významnosti α testujeme nulovou hypotézu, která tvrdí, že všechny střední hodnoty jsou stejné oproti alternativní hypotéze, která tvrdí, že alespoň jedna dvojice středních hodnot se liší. Jedná se tedy o zobecnění dvouvýběrového t-testu a na první pohled se zdá, že stačí utvořit $r(r-1)/2$ dvojic náhodných výběrů a na každou dvojici aplikovat dvouvýběrový t-test. Tento postup však nelze použít, neboť nezaručuje splnění podmínky, že pravděpodobnost chyby 1. druhu je α. Proto ve 30. letech 20. století vytvořil R. A. Fisher metodu ANOVA[1] (analýza rozptylu, v popsané situaci analýza rozptylu jednoduchého třídění), která uvedenou podmínku splňuje.

[1] Z anglického ANalysis Of VAriance

Obecný popis

Pokud na hladině významnosti α zamítneme nulovou hypotézu, zajímá nás, které dvojice středních hodnot se od sebe liší. K řešení tohoto problému slouží metoda mnohonásobného porovnávání, např. Scheffého nebo Tukeyova metoda.

Označení

Výsledky pokusu popíšeme pomocí spojité náhodné veličiny Y a to tak, že sledujeme výsledky tohoto pokusu při všech úrovních faktoru A. Zjištěné hodnoty $\mathbf{Y} = (Y_1, \dots, Y_n)'$ roztřídíme do $\boxed{a}$ skupin podle úrovní do následující tabulky:

Úroveň faktoru	*Počet pozorování*	*Naměřené hodnoty*	*Součet úrovně*	*Průměr úrovně*	*Rozdělení úrovně*
1.	n_1	$\mathbf{Y}_1 = (Y_{11}, \dots, Y_{1n_1})'$	$Y_{1.} = \sum_{i=1}^{n_1} Y_{1i}$	$\overline{Y}_{1.} = \frac{1}{n_1} Y_{1.}$	$Y_{1i} \sim \mathcal{L}(\mu_1, \sigma^2)$
2.	n_2	$\mathbf{Y}_2 = (Y_{21}, \dots, Y_{2n_2})'$	$Y_{2.} = \sum_{i=1}^{n_2} Y_{2i}$	$\overline{Y}_{2.} = \frac{1}{n_2} Y_{1.}$	$Y_{2i} \sim \mathcal{L}(\mu_2, \sigma^2)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
a-tá	n_a	$\mathbf{Y}_a = (Y_{a1}, \dots, Y_{an_a})'$	$Y_{a.} = \sum_{i=1}^{n_a} Y_{ai}$	$\overline{Y}_{a.} = \frac{1}{n_a} Y_{a.}$	$Y_{ai} \sim \mathcal{L}(\mu_a, \sigma^2)$
Součet	n		$Y_{..} = \sum_{j=1}^{a} \sum_{i=1}^{n_j} Y_{ji}$	$\overline{Y}_{..} = \frac{1}{n} Y_{..}$	

Základní model

Definice 1 (model M)

Náhodné veličiny Y_{ij} se řídí modelem M:

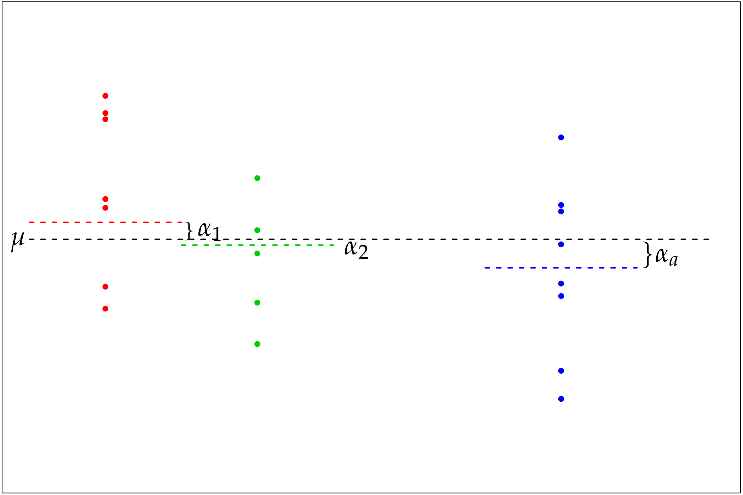
$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij},$$

pro $i = 1, \ldots, a$ a $j = 1, \ldots, n_i$, přičemž ε_{ij} jsou stochasticky nezávislé náhodné veličiny s rozložením $N(0, \sigma^2)$, μ je společná část střední hodnoty proměnné veličiny, α_i je efekt faktoru A na úrovni i.

Při zkoumání vlivu jednoho faktoru A testujeme hypotézu

$$\boxed{H_0} : \alpha_1 = \cdots = \alpha_a = 0 \quad \text{proti alternativě} \quad \boxed{H_1} : \exists\, i : \alpha_i \neq 0$$

Graficky – model M



Úroveň: 1 2 … a

Minimální submodel

Pokud platí nulová hypotéza H_0, dostáváme následující minimální submodel.

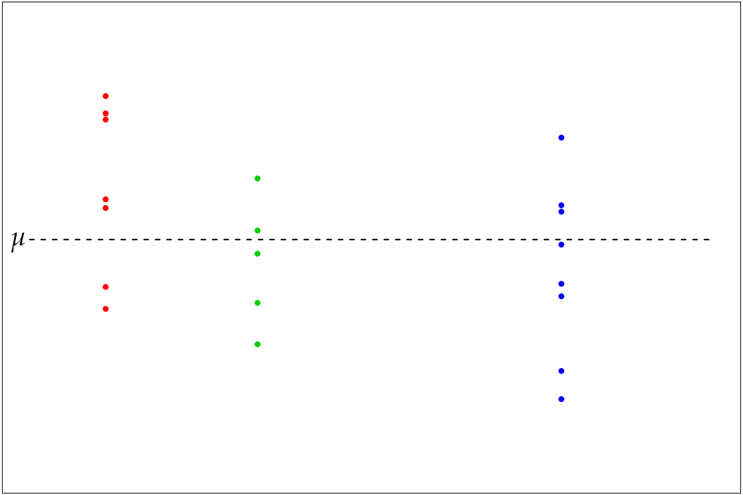
Definice 2 (model M_0)

Náhodné veličiny Y_{ij} se řídí modelem M_0:

$$Y_{ij} = \mu + \varepsilon_{ij},$$

pro $i = 1, \ldots, a$ a $j = 1, \ldots, n_i$, přičemž ε_{ij} jsou stochasticky nezávislé náhodné veličiny s rozložením $N(0, \sigma^2)$.

Graficky – submodel M_0



Úroveň: 1 2 $\dots$ a

Odvození

Základní model M:

Matice plánu je $$\mathbf{X} = \begin{pmatrix} \mathbf{1}_{n_1} & \mathbf{1}_{n_1} & \mathbf{0} & \cdots & \cdots & \mathbf{0} \\ \mathbf{1}_{n_2} & \mathbf{0} & \mathbf{1}_{n_2} & \ddots & & \vdots \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ \mathbf{1}_{n_{a-1}} & \vdots & & \ddots & \mathbf{1}_{n_{a-1}} & \mathbf{0} \\ \mathbf{1}_{n_a} & \mathbf{0} & \cdots & \cdots & \mathbf{0} & \mathbf{1}_{n_a} \end{pmatrix} \quad \text{a} \quad \boldsymbol{\beta} = \begin{pmatrix} \mu \\ \alpha_1 \\ \vdots \\ \vdots \\ \alpha_a \end{pmatrix},$$

kde vektor $\mathbf{1}_k$ značí sloupcový vektor složený z k jedniček. Matice $\mathbf{X}$ má $(a+1)$ sloupců a není plné hodnosti. **Proč?**

Odvození

Systém normálních rovnic $\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{Y}$:

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} n & n_1 & n_2 & \cdots & \cdots & n_a \\ n_1 & n_1 & 0 & \cdots & \cdots & 0 \\ n_2 & 0 & n_2 & \ddots & & \vdots \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ n_{a-1} & \vdots & & \ddots & n_{a-1} & 0 \\ n_a & 0 & \cdots & \cdots & 0 & n_a \end{pmatrix}, \qquad \mathbf{X}'\mathbf{Y} = \begin{pmatrix} \mathbf{1}'_{n_1} & \mathbf{1}'_{n_2} & \cdots & \mathbf{1}'_{n_{a-1}} & \mathbf{1}'_{n_a} \\ \mathbf{1}'_{n_1} & \mathbf{0} & \cdots & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{1}'_{n_2} & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \mathbf{1}'_{n_{a-1}} & \mathbf{0} \\ \mathbf{0} & \cdots & \cdots & \mathbf{0} & \mathbf{1}'_{n_a} \end{pmatrix} \begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \\ \vdots \\ \vdots \\ \mathbf{Y}_{a-1} \\ \mathbf{Y}_a \end{pmatrix} = \begin{pmatrix} Y_{\cdot\cdot} \\ Y_{1\cdot} \\ \vdots \\ \vdots \\ Y_{a-1\cdot} \\ Y_{a\cdot} \end{pmatrix}.$$

Jednou z pseudoinverzních matic k matici $\mathbf{X}'\mathbf{X}$ je matice

$$(\mathbf{X}'\mathbf{X})^{-} = \begin{pmatrix} 0 & 0 & 0 & \cdots & \cdots & 0 \\ 0 & \frac{1}{n_1} & 0 & \cdots & \cdots & 0 \\ 0 & 0 & \frac{1}{n_2} & \ddots & & \vdots \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \vdots & & \ddots & \frac{1}{n_{a-1}} & 0 \\ 0 & 0 & \cdots & \cdots & 0 & \frac{1}{n_a} \end{pmatrix} \quad \Rightarrow \quad \mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}' = \begin{pmatrix} \frac{1}{n_1}\mathbf{E}_{n_1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \mathbf{0} \\ \mathbf{0} & \cdots & \mathbf{0} & \frac{1}{n_a}\mathbf{E}_{n_a} \end{pmatrix},$$

kde $\mathbf{E}_k = \mathbf{1}_k\mathbf{1}'_k$ je matice typu $(k \times k)$ samých jedniček.

Odvození

Odtud

$$\widehat{\mathbf{Y}} = \begin{pmatrix} (\widehat{\mu}+\widehat{\alpha}_1)\cdot \mathbf{1}_{n_1} \\ \vdots \\ \vdots \\ (\widehat{\mu}+\widehat{\alpha}_a)\cdot \mathbf{1}_{n_a} \end{pmatrix} = \begin{pmatrix} \widehat{\mathbf{Y}}_1 \\ \vdots \\ \vdots \\ \widehat{\mathbf{Y}}_a \end{pmatrix} = \mathbf{HY} = \begin{pmatrix} \frac{1}{n_1}\mathbf{E}_{n_1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \mathbf{0} \\ \mathbf{0} & \cdots & \mathbf{0} & \frac{1}{n_a}\mathbf{E}_{n_a} \end{pmatrix} \begin{pmatrix} \mathbf{Y}_1 \\ \vdots \\ \vdots \\ \mathbf{Y}_a \end{pmatrix} = \begin{pmatrix} \overline{Y}_{1\bullet}\,\mathbf{1}_{n_1} \\ \vdots \\ \vdots \\ \overline{Y}_{a\bullet}\,\mathbf{1}_{n_a} \end{pmatrix}$$

takže odhad střední hodnoty je tvaru

$$\widehat{\mu}+\widehat{\alpha}_j = \overline{Y}_{j\bullet}.$$

Přidáním dodatečné podmínky $\sum_{j=1}^{a} n_j\alpha_j = 0$, dostaneme odhad společné střední hodnoty $\widehat{\mu} = \overline{Y}_{\bullet\bullet}$ a pro $j = 1,\dots,a$ odhad příspěvku j-té skupiny $\widehat{\alpha}_j = \overline{Y}_{j\bullet} - \overline{Y}_{\bullet\bullet}$

Odvození

Pokud platí nulová hypotéza H_0, tj. submodel M_0:

$$\mathbf{Y} = \mathbf{X}_0\ \beta_0 + \boldsymbol{\varepsilon},$$

kde $\mathbf{X}_0 = \mathbf{1}_n$, $\mathbf{X}_0'\mathbf{X}_0 = \mathbf{1}_n'\mathbf{1}_n = n$, $\mathbf{X}_0'\mathbf{Y} = \mathbf{1}_n'\mathbf{Y} = Y_{\bullet\bullet}$
a

$$\hat{\beta}_0 = (\mathbf{X}_0'\mathbf{X}_0)^{-1}\mathbf{X}_0'\mathbf{Y} = \frac{1}{n}Y_{\bullet\bullet} = \overline{Y}_{\bullet\bullet}$$

Pak $\mathbf{H}_0 = \mathbf{X}_0(\mathbf{X}_0'\mathbf{X}_0)^{-1}\mathbf{X}_0' = \frac{1}{n}\mathbf{1}_n\mathbf{1}_n' = \frac{1}{n}\mathbf{E}_n$
a

$$\widehat{\mu}_0 = \widehat{\mathbf{Y}}_0 = \mathbf{H}_0\mathbf{Y} = \frac{1}{n}\mathbf{E}_n\mathbf{Y} = \overline{Y}_{\bullet\bullet}\mathbf{1}_n.$$

Odvození

Součty kvadrátů odchylek

$$\boxed{S_e} = \|\widehat{\boldsymbol{\varepsilon}}\|^2 = (\mathbf{Y} - \widehat{\boldsymbol{\mu}})'(\mathbf{Y} - \widehat{\boldsymbol{\mu}}) = (\mathbf{Y} - \widehat{\mathbf{Y}})'(\mathbf{Y} - \widehat{\mathbf{Y}})$$

$$= \sum_{j=1}^{a} (\mathbf{Y}_j - \overline{Y}_{j.} \mathbf{1}_{n_j})'(\mathbf{Y}_j - \overline{Y}_{j.} \mathbf{1}_{n_j}) = \sum_{j=1}^{a} \sum_{i=1}^{n_j} (Y_{ji} - \overline{Y}_{j.})^2 \qquad \textit{reziduální}$$

$$S_{e_0} = \boxed{S_T} = \|\widehat{\boldsymbol{\varepsilon}}_0\|^2 = (\mathbf{Y} - \widehat{\boldsymbol{\mu}}_0)'(\mathbf{Y} - \widehat{\boldsymbol{\mu}}_0) = \sum_{j=1}^{a} (\mathbf{Y}_j - \overline{Y}_{..} \mathbf{1}_{n_j})'(\mathbf{Y}_j - \overline{Y}_{..} \mathbf{1}_{n_j}) = \sum_{j=1}^{a} \sum_{i=1}^{n_j} (Y_{ji} - \overline{Y}_{..})^2 \qquad \textit{celkový}$$

$$S_{\Delta_0} = \boxed{S_A} = \|\boldsymbol{\Delta}_0\|^2 = (\widehat{\boldsymbol{\mu}} - \widehat{\boldsymbol{\mu}}_0)'(\widehat{\boldsymbol{\mu}} - \widehat{\boldsymbol{\mu}}_0) = \sum_{j=1}^{a} (\overline{Y}_{j.} \mathbf{1}_{n_j} - \overline{Y}_{..} \mathbf{1}_{n_j})'(\overline{Y}_{j.} \mathbf{1}_{n_j} - \overline{Y}_{..} \mathbf{1}_{n_j})$$

$$= \sum_{j=1}^{a} (\overline{Y}_{j.} - \overline{Y}_{..})^2 \mathbf{1}'_{n_j} \mathbf{1}_{n_j} = \sum_{j=1}^{a} n_j (\overline{Y}_{j.} - \overline{Y}_{..})^2 \qquad \textit{mezi třídami}$$

$$= S_{e_0} - S_e \qquad \Rightarrow \qquad \boxed{S_T = S_A + S_e}$$

takže pokud platí model $\boxed{M_0}$, pak statistika

$$F_A = \frac{(S_{e_0} - S_e)/(a-1)}{S_e/(n-a)} \quad \sim \quad F(a-1, n-a).$$

Shrnutí

Definice 3

- **Celkový součet čtverců** (charakterizuje variabilitu jednotlivých pozorování kolem celkového průměru), počet stupňů volnosti $df_T = n - 1$:

$$S_T = \sum_{i=1}^{a} \sum_{j=1}^{n_i} \left(Y_{ij} - \overline{Y}_{\bullet\bullet}\right)^2$$

- **Skupinový součet čtverců** (charakterizuje variabilitu mezi jednotlivými náhodnými výběry), počet stupňů volnosti $df_A = a - 1$:

$$S_A = \sum_{j=1}^{a} n_j \left(\overline{Y}_{j\bullet} - \overline{Y}_{\bullet\bullet}\right)^2$$

- **Reziduální součet čtverců** (charakterizuje variabilitu uvnitř jednotlivých výběrů), počet stupňů volnosti $df_e = n - a$:

$$S_e = \sum_{i=1}^{a} \sum_{j=1}^{n_i} \left(Y_{ij} - \overline{Y}_{j\bullet}\right)^2.$$

Shrnutí

Věta 4

Lze dokázat, že

$$S_T = S_A + S_E.$$

Věta 5

Rozdíl mezi modely M a M_0 ověřujeme pomocí testové statistiky

$$F_A = \frac{S_A/df_A}{S_e/df_e},$$

která se řídí rozložením $F(a-1, n-a)$, je-li model M_0 správný. Hypotézu o nevýznamnosti faktoru A tedy zamítáme na hladině významnosti α, když platí:

$$F_A \geq F_{1-\alpha}(a-1, n-a).$$

Shrnutí

Předcházející pojmy se shrnují v **tabulce analýzy rozptylu**

Zdroj variability	*Součet čtverců* SS	*Stupně volnosti* df	*Podíl* $MS = \frac{SS}{df}$	$F = \frac{MS}{s^2}$
Třídy	S_A	$df_a = a - 1$	$MS_A = \frac{S_A}{df_a}$	$F_A = \frac{MS_A}{MS_e}$
Reziduální	S_e	$df_e = n - a$	$MS_e = \frac{S_e}{df_e}$	–
Celkový	S_T	$df_T = n - 1$	–	–

Test shody rozptylů

Věta 6 (Levenův test)

Položme $Z_{ij} = |Y_{ij} - \overline{Y}_{i\bullet}|$. *Označme:*

- $\overline{Z}_{i\bullet} = \frac{1}{n_i} \sum_{j=1}^{n_i} Z_{ij}$
- $\overline{Z}_{\bullet\bullet} = \frac{1}{n} \sum_{i=1}^{a} \sum_{j=1}^{n_i} Z_{ij}$
- $S_{Ze} = \sum_{i=1}^{a} \sum_{j=1}^{n_i} \left(Z_{ij} - \overline{Z}_{i\bullet}\right)^2$
- $S_{ZA} = \sum_{i=1}^{a} n_i \left(\overline{Z}_{i\bullet} - \overline{Z}_{\bullet\bullet}\right)^2$

Platí-li hypotéza o shodě rozptylů, pak statistika

$$F_Z = \frac{S_{ZA}/(a-1)}{S_{Ze}/(n-a)} \sim F(a-1, n-a).$$

Test shody rozptylů

Věta 7 (Bartlettův test)

Platí-li hypotéza o shodě rozptylů, pak statistika

$$B = \frac{1}{C}\left[(n-a)\ln S_*^2 - \sum_{j=1}^{a}(n_j-1)\ln S_j^2\right] \approx \chi^2(a-1),$$

kde

$$C = 1 + \frac{1}{3(a-1)}\left(\sum_{j=1}^{a}\frac{1}{n_j-1} - \frac{1}{n-a}\right), \quad S_*^2 = \frac{S_e}{n-a}.$$

H_0 *zamítáme na asymptotické hladině významnosti* α*, když* $B \geq \chi^2_{1-\alpha}(a-1, n-a)$.

Metody mnohonásobného porovnávání

Zamítneme-li na hladině významnosti α hypotézu o shodě středních hodnot, chceme zjistit, které dvojice středních hodnot se liší na dané hladině významnosti α.

Všechny výběry mají týž rozsah $\boxed{p}$ $\Rightarrow$ Tukeyova metoda
Všechny výběry nemají stejný rozsah $\Rightarrow$ Schefféo metoda.

Metody mnohonásobného porovnávání

Věta 8 (Tukeyova metoda)

Rovnost středních hodnot μ_k a μ_l zamítneme na hladině významnosti α, když:

$$\left|\overline{Y}_{k\cdot} - \overline{Y}_{l\cdot}\right| \geq q_{1-\alpha}(a, n-a)\frac{S_*}{\sqrt{p}},$$

kde $q_{1-\alpha}(a, n-a)$ jsou kvantily studentizovaného rozpětí, které najdeme ve statistických tabulkách.

Věta 9 (Scheffého metoda)

Rovnost středních hodnot μ_k a μ_l zamítneme na hladině významnosti α, když:

$$\left|\overline{Y}_{k\cdot} - \overline{Y}_{l\cdot}\right| \geq S_*\sqrt{(a-1)\left(\frac{1}{n_k} + \frac{1}{n_l}\right)F_{1-\alpha}(a-1, n-a)}.$$

Význam předpokladů v analýze rozptylu

- Nezávislost jednotlivých náhodných výběrů – velmi důležitý předpoklad, musí být splněn, jinak dostaneme nesmyslné výsledky.
- Normalita – ANOVA není příliš citlivá na porušení normality, zvlášť pokud mají všechny výběry rozsah nad 20 (důsledek centrální limitní věty). Při výraznějším porušení se doporučuje Kruskalův – Wallisův test.
- Shoda rozptylů – mírné porušení nevadí, při větším se doporučuje Kruskalův – Wallisův test. Test shody rozptylů má smysl provádět až po ověření předpokladu normality.

Kruskalův – Wallisův test

Kruskalův – Wallisův test je neparametrická obdoba analýzy rozptylu jednoduchého třídění.

Formulace problému

Nechť je dáno a nezávislých náhodných výběrů o rozsazích $n_1, \ldots, n_a$. Předpokládáme, že tyto výběry pocházejí ze spojitých rozložení. Označme $n = n_1 + \ldots + n_a$. Chceme testovat hypotézu, že všechny tyto výběry pocházejí z téhož rozložení.

Kruskalův – Wallisův test

Věta 10 (Kruskalův – Wallisův test)

Všech n hodnot seřadíme do rostoucí posloupnosti a určíme pořadí každé hodnoty. Označme T_j součet pořadí těch hodnot, které patří do j-tého výběru, $j = 1, \ldots, a$ (kontrola: musí platit $T_1 + \ldots + T_a = n(n+1)/2$).
Testová statistika má tvar:

$$Q = \frac{12}{n(n+1)} \sum_{j=1}^{a} \frac{T_j^2}{n_j} - 3(n+1). \tag{1}$$

Platí-li H_0, má statistika Q asymptoticky rozložení $\chi^2(a-1)$, rostou-li rozsahy výběrů nade všechny meze. H_0 tedy zamítneme na asymptotické hladině významnosti α, když $Q \geq \chi^2_{1-\alpha}(a-1)$.

Příklad

Příklad 1

U čtyř odrůd brambor (označených symboly A, B, C, D) se zjišťovala celková hmotnost brambor vyrostlých vždy z jednoho trsu. Výsledky uvádí tabulka:

odrůda	hmotnost (v kg)				
A	0,9	0,8	0,6	0,9	
B	1,3	1,0	1,3		
C	1,3	1,5	1,6	1,1	1,5
D	1,1	1,2	1,0		

Na hladině významnosti 0,05 testujte hypotézu, že střední hodnota hmotnosti trsu brambor nezávisí na odrůdě. Zamítnete-li nulovou hypotézu, zjistěte, které dvojice odrůd se liší na hladině významnosti 0,05.

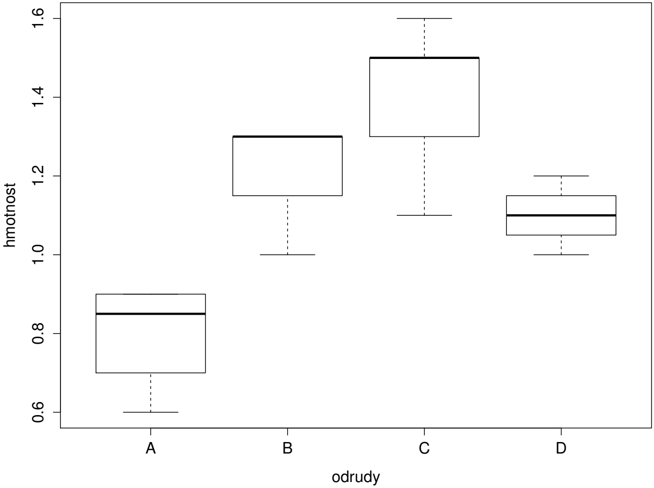
Řešení

Řešení. Data považujeme za realizace čtyř nezávislých náhodných výběrů ze čtyř normálních rozložení se stejným rozptylem. Testujeme hypotézu, že všechny čtyři střední hodnoty jsou stejné.
Výpočtem získáme: $\bar{y}_{1\bullet} = 0,8$, $\bar{y}_{2\bullet} = 1,2$, $\bar{y}_{3\bullet} = 1,4$, $\bar{y}_{4\bullet} = 1,1$, $\bar{y}_{\bullet\bullet} = 1,14$, $S_e = 0,3$, $S_A = 0,816$, $S_T = 1,116$, $F_A = 9,97$. Ze statistických tabulek získáme $F_{0,95}(3,11) = 3,59$. Protože testová statistika se realizuje v kritickém oboru, zamítáme nulovou hypotézu na hladině významnosti 0,05.
Výsledky zapíšeme do tabulky ANOVA:

Zdroj variability	Součet čtverců	Stupně volnosti	Podíl	F_A
třídy	$S_A = 0,816$	3	$S_A/3 = 0,272$	$\frac{S_A/3}{S_E/11} = 9,97$
reziduální	$S_E = 0,3$	11	$S_E/11 = 0,02727$	—
celkový	$S_T = 1,116$	14	—	—

Řešení

Grafické posouzení



Řešení

Nyní pomocí Schefféhо metody zjistíme, které dvojice odrůd se liší na hladině významnosti 0,05.

Srovnávané odrůdy	Rozdíly $\lvert m_k - m_l \rvert$	Pravá strana vzorce
A, B	0,4	0,41
A, C	0,67	0,36
A, D	0,3	0,41
B, C	0,2	0,40
B, D	0,1	0,44
C, D	0,3	0,40

Na hladině významnosti 0,05 se liší odrůdy A a C.

Více nezávislých náhodných výběrů z alternativních rozložení

Test homogenity binomických rozložení

Nechť $Y_{j1}, \ldots, Y_{jn_j} \sim A(\theta_j)$, $j = 1, 2, \ldots, a$ jsou nezávislé náhodné výběry z alternativního rozložení. Testujeme hypotézu H_0: $\theta_1 = \cdots = \theta_a$ proti alternativní hypotéze H_1: „alespoň jedna dvojice parametrů je různá".

Věta 11

Statistika

$$Q = \frac{1}{\overline{Y}_{\bullet\bullet}(1 - \overline{Y}_{\bullet\bullet})} \sum_{j=1}^{a} n_j \left(\overline{Y}_{j\bullet} - \overline{Y}_{\bullet\bullet}\right)^2,$$

má v případě platnosti nulové hypotézy asymptoticky rozložení $\chi^2(a-1)$. H_0 tedy zamítáme na asymptotické hladině významnosti α, když $Q \geq \chi^2_{1-\alpha}(a-1)$.

Více nezávislých náhodných výběrů z alternativních rozložení

Poznámka 12

Test lze použít, pokud $n_j \overline{y}_{\bullet\bullet} > 5$ *pro všechna* $j = 1, \ldots, a$.

Poznámka 13

Statistiku Q *lze snadno upravit do Brandtova – Snedecorova výpočetního tvaru*

$$Q = \frac{1}{\overline{Y}_{\bullet\bullet}(1 - \overline{Y}_{\bullet\bullet})} \sum_{j=1}^{a} n_j \overline{Y}_{j\bullet}^2 - n \frac{\overline{Y}_{\bullet\bullet}}{1 - \overline{Y}_{\bullet\bullet}}. \tag{2}$$

Více nezávislých náhodných výběrů z alternativních rozložení

Test homogenity binomických rozložení založený na arkussinusové transformaci

Není-li splněna podmínka $n_j\bar{y}_{..} > 5$ pro všechna $j = 1, \ldots, a$, doporučuje se následující postup:

Věta 14

Označme

- $A_j = \arcsin\sqrt{\overline{Y}_{j.}}$
- $B = \frac{1}{n}\sum\limits_{j=1}^{a} n_j A_j$.

Pak statistika

$$Q = 4\sum_{j=1}^{a} n_j \left(A_j - B\right)^2 \approx \chi^2(a-1).$$

H_0 *tedy zamítáme na asymptotické hladině významnosti* α, *když* $Q \geq \chi^2_{1-\alpha}(a-1)$.

Mnohonásobné porovnávání

Zamítneme-li nulovou hypotézu na asymptotické hladině významnosti α, chceme zjistit, které dvojice parametrů θ_k a θ_l se liší.

Věta 15

Platí-li nerovnost

$$|A_k - A_l| \geq \sqrt{\frac{1}{8}\left(\frac{1}{n_k} + \frac{1}{n_l}\right)} \cdot q_{1-\alpha}(a, \infty),$$

pak na hladině významnosti α zamítáme hypotézu o shodě parametrů θ_k a θ_l.

Poznámka 16

Hodnoty $q_{1-\alpha}(a, \infty)$ jsou kvantily studentizovaného rozpětí.

Příklad

Příklad 2

Na gymnázium bylo přijato 142 studentů. Ti byli náhodně rozděleni do tříd A, B, C, D. V každé třídě byla matematika vyučována jinou metodou. Na konci školního roku psali všichni studenti stejnou písemnou práci a byl zaznamenán počet těch studentů, kteří vyřešili všechny zadané úkoly.

Třída	A	B	C	D
Počet studentů	35	36	37	34
Počet úspěšných studentů	5	8	17	15

Na asymptotické hladině významnosti 0,05 testujte hypotézu, že rozdíly v podílech studentů v jednotlivých třídách, kteří správně vyřešili všechny zadané úlohy, jsou způsobeny pouze náhodnými vlivy.

Řešení

Řešení. Máme čtyři nazávislé náhodné výběry, j-tý pochází z rozložení $A(\theta_j)$, $j = 1, 2, 3, 4$. Testujeme hypotézu H_0: $\theta_1 = \theta_2 = \theta_3 = \theta_4$. Ze zadání a výpočtem zjistíme: $n_1 = 35$, $n_2 = 36$, $n_3 = 37$, $n_4 = 34$, $\overline{y}_{1\bullet} = 5/35$, $\overline{y}_{2\bullet} = 8/36$, $\overline{y}_{3\bullet} = 17/37$, $\overline{y}_{4\bullet} = 15/34$, $\overline{y}_{\bullet\bullet} = 45/142$, $Q = 12{,}288$, $\chi^2_{0{,}95}(3) = 7{,}81$.
Protože testové kritérium se realizuje v kritickém oboru, H_0 zamítáme na asymptotické hladině významnosti 0,05.
Spočteme arkussinusové transformace výběrových průměrů. Vyjde: $A_1 = 0{,}3876$, $A_2 = 0{,}4909$, $A_3 = 0{,}7448$, $A_4 = 0{,}7264$.
Nyní metodou mnohonásobného porovnávání zjistíme, které dvojice parametrů se od sebe liší na hladině významnosti 0,05.

Řešení

Srovnávané třídy	Rozdíly $\lvert A_k - A_l \rvert$	Pravá strana vzorce
A, B	0,1033	0,30
A, C	0,3572	0,30
A, D	0,3388	0,31
B, C	0,2539	0,30
B, D	0,2356	0,31
C, D	0,0184	0,30

Na hladině významnosti 0,05 se liší třídy A, C a A, D.

Využití ANOVA v lineárním regresním modelu

Analýzy rozptylu lze využít v momentě, kdy chceme zjednodušit zvolený model a vypustit z modelu některé vysvětlující proměnné. Tj. uvažujeme nový podmodel, jehož matice plánu vznikne z původní matice vypuštěním některých sloupců.
Naším úkolem je testovat, zda zvolený podmodel je vhodný k dostatečnému popisu závislosti v datech.
Bez újmy na obecnosti předpokládejme, že matice, které určují model a podmodel se liší právě posledními sloupci matice $\mathbf{X}$, takže $\mathbf{X} = (\mathbf{X}_0, \mathbf{X}_1)$.
Mějme náhodný vektor $\mathbf{Y} = (Y_1, \ldots, Y_n)'$ a předpokládejme, že platí model M a je dán submodel M_0, přičemž

$$\boxed{M} \quad \mathbf{Y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n) \quad \mathbf{X} \text{ je typu } n \times k, \quad h(\mathbf{X}) = r, \quad \boldsymbol{\beta} \text{ je typu } k \times 1$$

$$\boxed{M_0} \quad \mathbf{Y} \sim N_n(\mathbf{X}_0\boldsymbol{\beta}_0, \sigma^2\mathbf{I}_n) \quad \mathbf{X}_0 \text{ je typu } n \times k_0, \quad h(\mathbf{X}_0) = r_0, \quad \boldsymbol{\beta}_0 \text{ je typu } k_0 \times 1$$

$$n \geq k \geq r \geq r_0$$

Model M_0 je podmodelem M pokud $\mathbf{X}_0 = \mathbf{X}\mathbf{K}$, kde matice $\mathbf{K} = \binom{\mathbf{I}_{k_0}}{\mathbf{0}}$ je typu $k \times k_0$.

Využití ANOVA v lineárním regresním modelu

Položme

$$\widehat{\boldsymbol{\mu}} = \mathbf{H}\mathbf{Y} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{Y}, \;\; \widehat{\boldsymbol{\mu}}_0 = \mathbf{H}_0\mathbf{Y} = \mathbf{X}_0(\mathbf{X}_0'\mathbf{X}_0)^{-}\mathbf{X}_0'\mathbf{Y},$$

pak

$$\begin{aligned} S_e &= (\mathbf{Y}-\widehat{\boldsymbol{\mu}})'(\mathbf{Y}-\widehat{\boldsymbol{\mu}}) & \qquad S_{e_0} &= (\mathbf{Y}-\widehat{\boldsymbol{\mu}}_0)'(\mathbf{Y}-\widehat{\boldsymbol{\mu}}_0) \\ S_{\Delta_0} &= (\widehat{\boldsymbol{\mu}}-\widehat{\boldsymbol{\mu}}_0)'(\widehat{\boldsymbol{\mu}}-\widehat{\boldsymbol{\mu}}_0) & \qquad S_e &= S_{e_0} - S_{\Delta_0} \end{aligned}$$

Pokud platí model $\boxed{M_0}$, pak statistika

$$F_0 = \frac{(S_{e_0} - S_e)/(r - r_0)}{S_e/(n-r)} \quad \sim \quad F(r - r_0, n - r).$$

Příklad

Příklad 3

Pro data uvedená v následující tabulce

x	1	2	3	4	5	6	7	8	9	10
y	58,42	37,34	49,64	59,85	24,37	59,29	47,12	75,29	140,49	147,23

uvažujte modely

$$
\begin{aligned}
M_1 : &\quad y = \beta_0 + \beta_1 x \\
M_2 : &\quad y = \beta_0 + \beta_1 x + \beta_2 x^2 \\
M_3 : &\quad y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3.
\end{aligned}
$$

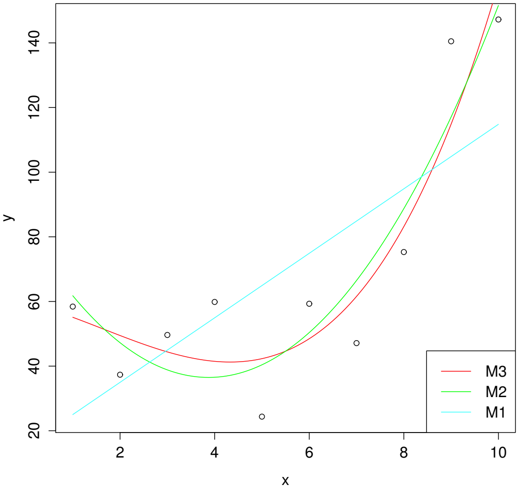
Pomocí analýzy rozptylu porovnejte tyto modely.

Řešení

Řešení. Vycházíme z modelu M_3 a testujeme vhodnost podmodelu M_2. Hodnota statistiky F_0 je v tomto případě $0,6469$, p-hodnota testu je $0,4519$. To znamená, že vynecháním kubického členu se model významně nezhorší. Nadále budeme tedy uvažovat model M_2 a testovat vhodnost podmodelu M_1. Hodnota statistiky F_0 je v tomto případě $15,586$, p-hodnota testu je $0,0055$. To znamená, že vynecháním kvadratického členu se model již významně zhorší. Nejvhodnějším modelem pro popis závislosti je tedy M_2.

Řešení

Graficky



Úlohy k procvičení

Příklad 1

Jsou známy měsíční tržby (v tisících Kč) tří prodavačů za dobu půl roku.

1. prodavač	*12*	*10*	*9*	*10*	*11*	*9*
2. prodavač	*10*	*12*	*11*	*12*	*14*	*13*
3. prodavač	*19*	*18*	*16*	*16*	*17*	*15*

Na hladině významnosti 0,05 testujte hypotézu, že střední hodnoty tržeb všech tří prodavačů jsou stejné. Pokud zamítneme nulovou hypotézu, zjistěte, tržby kterých dvou prodavačů se liší na hladině významnosti 0,05.

[Na hladině významnosti 0,05 se liší tržby prodavačů 1, 3 a 2, 3.]

Úlohy k procvičení

Příklad 2

Naprogramujte funkci „`anovabinom.R`“, která pro vstupní vektory `nj` (počet pozorování ve skupinách) a `pj` (počet „úspěchů“ ve skupinách) provede analýzu rozptylu pro binomická data. V případě zamítnutí nulové hypotézy vypíše indexy skupin, které se od sebe významně liší.

Příklad 3

104 náhodně vybraných matek bylo dotázáno, zda jejich kojenec dostává dudlík. Zjišťoval se též nejvyšší stupeň dosaženého vzdělání matky.

Vzdělání matky	*Počet matek*	*Počet dětí s dudlíkem*
základní	*39*	*27*
středoškolské	*47*	*34*
vysokoškolské	*18*	*15*

Na asymptotické hladině významnosti 0,05 testujte hypotézu, že podíly dětí s dudlíkem nezávisí na vzdělání matky.

Úlohy k procvičení

Příklad 4

Je dáno pět nezávislých náhodných výběrů o rozsazích 5, 7, 6, 8, 5, přičemž i-tý výběr pochází z rozložení $\mathrm{N}(\mu_i, \sigma^2)$, $i = 1, \ldots, 5$. Byl vypočten celkový součet čtverců $S_T = 15$ a reziduální součet čtverců $S_e = 3$. Na hladině významnosti 0,05 testujte hypotézu o shodě středních hodnot.

[$n = 31$, $a = 5$, $S_A = 12$, $f_A = 26$, $F_{0{,}95}(4, 26) = 2{,}7426$ Protože $f_A \geq F_{0{,}95}(4, 26)$, H_0 zamítáme na hladině významnosti 0,05.]

Příklad 5

V proměnné „`LakeHuron`"[a] jsou uloženy roční údaje o hloubce jezera Huron (ve stopách) v letech 1875 – 1972. Data proložte polynomem 8. stupně. Pomocí analýzy rozptylu zkoumejte možnosti zmenšení stupně regresního polynomu.

[a] datový soubor implementovaný v jazyce R

[Možno jít na stupeň 7.]

Úlohy k procvičení

Příklad 6

U 126 podniků řepařské oblasti v České Republice byl sledován hektarový výnos cukrovky ve vztahu ke spotřebě průmyslových hnojiv.
Data jsou uložena v souboru „`cukrovka.Rdata`“ ve 4 sloupcích:

1. *dolní hranice spotřeby K_2O (kg/ha)*
2. *horní hranice spotřeby K_2O (kg/ha)*
3. *četnosti*
4. *průměrné výnosy cukrovky (q/ha)*

a) *odhadněte parametry regresní funkce tvaru*

$$y = \beta_0 + \beta_1 x$$
$$y = \beta_0 + \beta_1 x + \beta_2 x^2$$

Poznámka: Za hodnoty nezávisle proměnné volte střed intervalu.

b) *Porovnejte vhodnost použitých regresních modelů pomocí analýzy rozptylu.*

[Kvadratický model je významný.]