MA012 Statistika II

1. Analýza rozptylu (ANOVA) a lineární regresní model

Ondřej Pokora (pokora@math.muni.cz)

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

(podzim 2015)

Motivační příklad

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.

Motivace

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

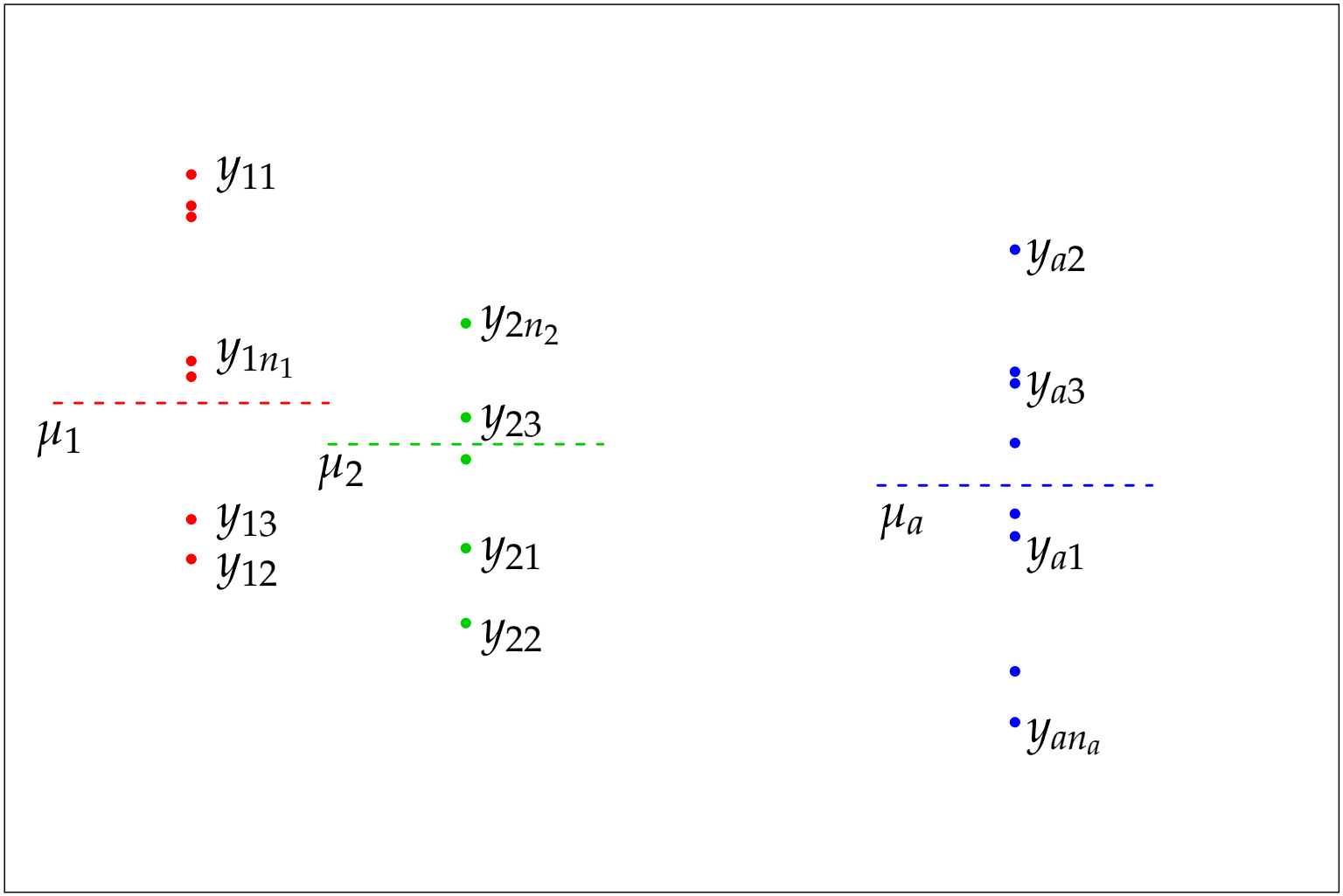
Obecný popis a předpoklady

- Faktor A má $a \geq 3$ úrovní.
- i-té úrovni odpovídá n_i výsledků $Y_{i1}, \ldots, Y_{in_i}$, které tvoří náhodný výběr z rozložení $N(\mu_i, \sigma^2)$, $i = 1, \ldots, a$.
- První index označuje skupinu podle úrovně faktoru, druhý index značí pořadí měření v dané skupině.
- 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 $i = 1, \ldots, a$ a $j = 1, \ldots, n_i$.

Graficky



Úroveň: 1 2 ... a

Obecný popis

Na hladině významnosti α testujeme nulovou hypotézu

H_0: všechny střední hodnoty jsou stejné,

oproti alternativní hypotéze

H_1: 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 (**ANalysis Of VAriance**),

která uvedenou podmínku splňuje.

Obecný popis

Pokud na hladině významnosti α zamítneme nulovou hypotézu H_0, zajímá nás, které dvojice středních hodnot, tedy kategorie podle úrovní faktoru A, se odlišují.

K řešení tohoto problému slouží metody tzv. **mnohonásobného porovnávání**, konkrétně 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 $\mathbf{Y} = (Y_1, \ldots, Y_n)'$ roztřídíme do a skupin podle úrovní do následující tabulky:

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

Všimněte si způsobu indexace výběrových průměrů pomocí tzv. tečkové notace.

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.

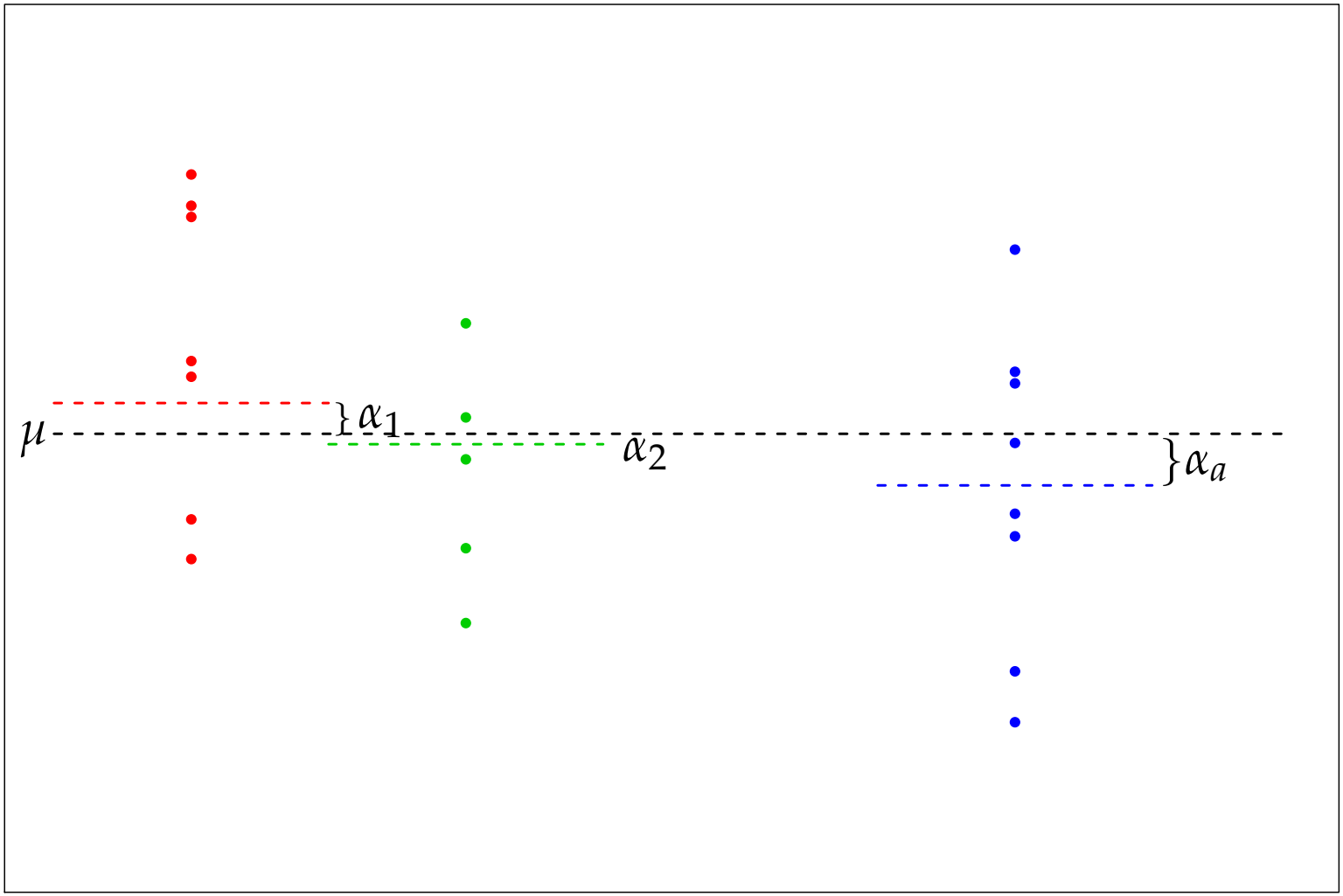
Při zkoumání vlivu jednoho faktoru A testujeme hypotézu

$$H_0 : \alpha_1 = \cdots = \alpha_a = 0$$

proti alternativní hypotéze

$$H_1 : \exists\, i \in \{1, \ldots, a\} :\ \alpha_i \neq 0.$$

Graficky – model M



Minimální (nulový) submodel

Pokud platí nulová hypotéza H_0, dostáváme následující minimální submodel.

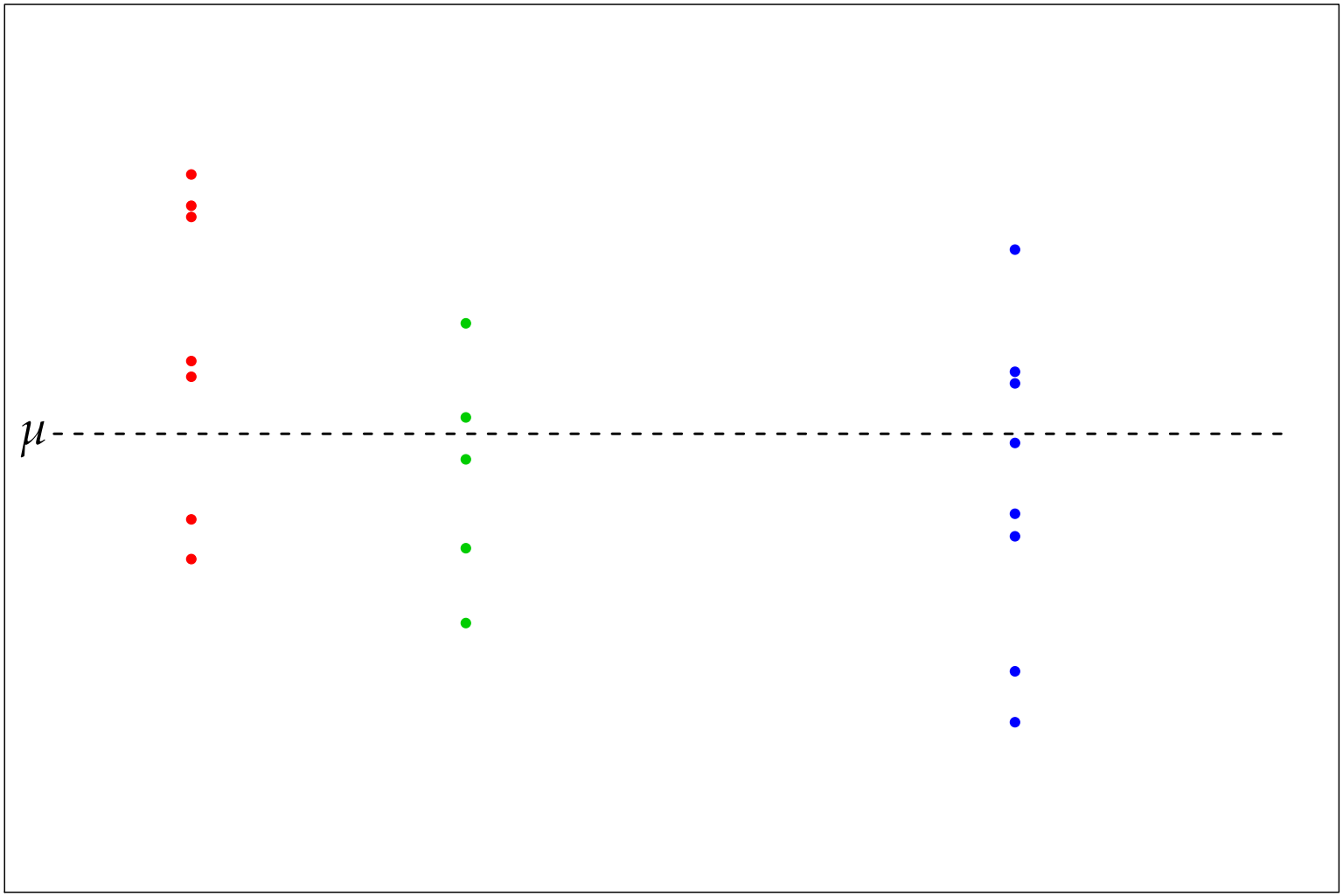
Definice 2 (model M_0)

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



Odvození

Základní model M:

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

kde vektor $\mathbf{1}_k$ značí sloupcový vektor složený z k jedniček.

Jaké rozměry má matice $\mathbf{X}$ a vektor $\boldsymbol{\beta}$? Matice $\mathbf{X}$ 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

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

takže odhad střední hodnoty je tvaru

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

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

Odvození – odhady parametrů modelu

Pokud platí nulová hypotéza H_0, tj. submodel M_0:

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

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

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

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

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

tedy

$$\widehat{\mu}_0 = \overline{Y}_{\cdot\cdot}.$$

Odvození – součty čtverců

Součty kvadrátů odchylek

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

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

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

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

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

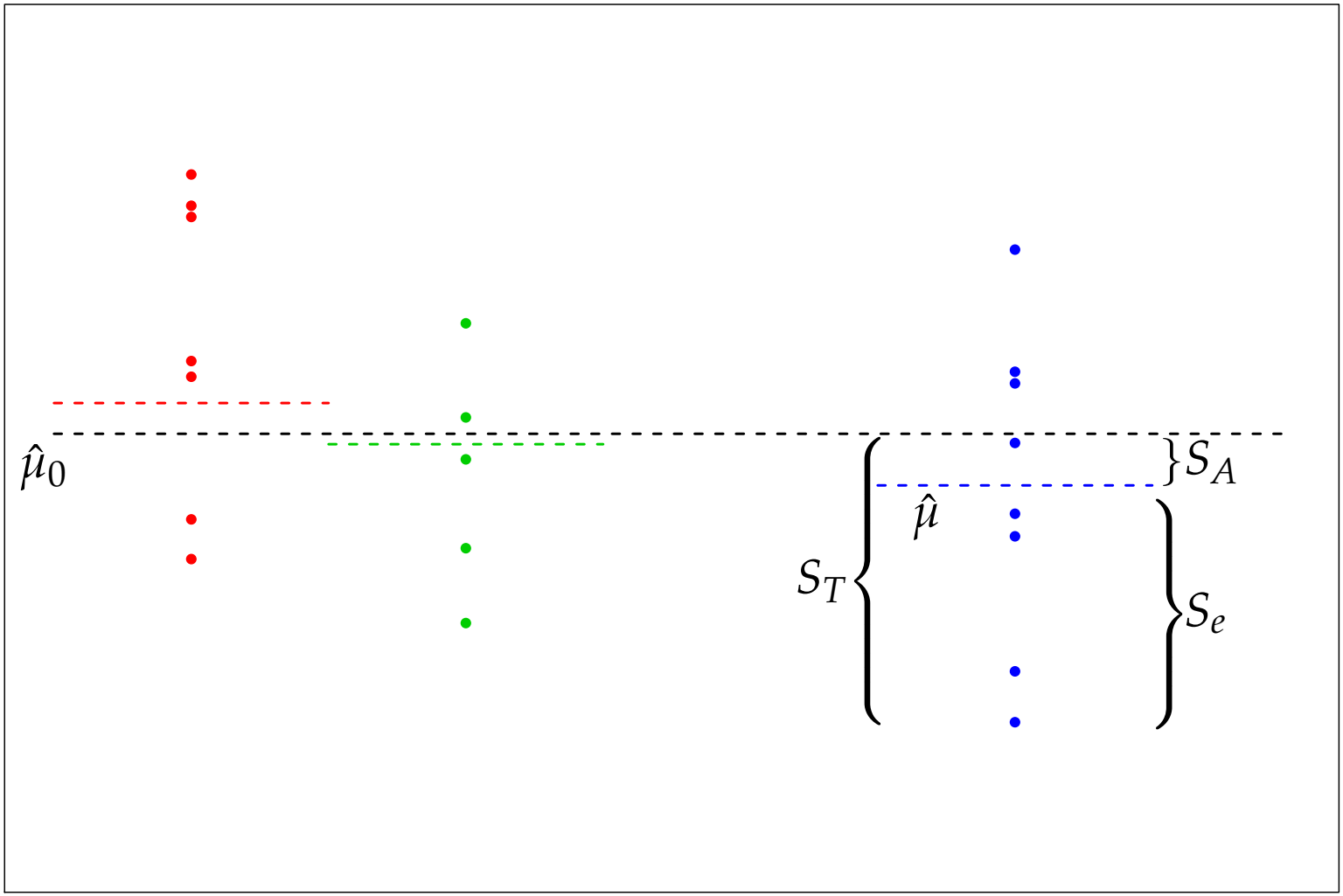
Všimneme si, že platí

$$S_T = S_A + S_e.$$

Pokud tedy platí model M_0, pak statistika

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

Graficky



Shrnutí

Definice 3

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

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

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

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

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

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

Shrnutí

Věta 4

Platí

$$S_T = S_A + S_E.$$

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:

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

Tabulka analýzy rozptylu

Předcházející pojmy se shrnují v tzv. **tabulce analýzy rozptylu**:

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

SS = *sum of squares*, MS = *mean square error*, df = *degrees of freedom*.

Test shody rozptylů

Věta 6 (Levenův test)

- *označmeOZ*$_{ij} = |Y_{ij} - \overline{Y}^*_{i\cdot}|$, *kde* $\overline{Y}^*_{i\cdot}$ *je výběrový průměr* $\overline{Y}_{i\cdot}$, *výběrový medián, příp. 10% ořezaný průměr (trimmed, truncated mean).*
- $\overline{Z}_{i\cdot} = \frac{1}{n_i} \sum_{j=1}^{n_i} Z_{ij}$
- $\overline{Z}_{\cdot\cdot} = \frac{1}{n} \sum_{i=1}^{a} \sum_{j=1}^{n_i} Z_{ij}$
- $S_{Ze} = \sum_{i=1}^{a} \sum_{j=1}^{n_i} \left(Z_{ij} - \overline{Z}_{i\cdot}\right)^2$
- $S_{ZA} = \sum_{i=1}^{a} n_i \left(\overline{Z}_{i\cdot} - \overline{Z}_{\cdot\cdot}\right)^2$

Platí-li hypotéza o shodě rozptylů, pak statistika

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

Praxe – jiný model

V praxi volí většina statistických softwarů (mj. i *R*) mírně odlišnou stavbu modelu.

Definice 7

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 $N(0, \sigma^2)$.

První úroveň faktoru A je přitom stanovena jako referenční, tedy $\alpha_1 = 0$.

Testujeme hypotézu

$$H_0 : \alpha_2^* = \cdots = \alpha_a^* = 0$$

proti alternativní hypotéze

$$H_1 : \exists\, i \in \{2, \ldots, a\} :\ \alpha_i^* \neq 0.$$

Praxe – jiný model

- Matice plánu modelu M^* je již plné hodnosti. Proč?
- Jakou interpretaci mají v modelu M^* parametry μ^* a α_i^*, $i = 2, \ldots, a$?

ANOVA v *R*:

- Jednou z možností je zkonstruovat lineární regresní model pomocí funkce `lm` a na výsledek aplikovat funkci `anova` .
- Obecnější přístup umožňuje přímo funkce `aov` .

Test shody rozptylů

Věta 8 (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

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

a S_j^2 je výběrový rozptyl v j-té kategorii.

Hypotézu o shodě rozptylů zamítáme na asymptotické hladině významnosti α, pokud $B \geq \chi^2_{1-\alpha}(a-1)$.

Metody mnohonásobného porovnávání

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

- Všechny výběry mají stejný rozsah p, resp. v praxi přibližně stejný $\Rightarrow$ Tukeyova metoda
- Výběry nemají stejný rozsah $\Rightarrow$ Scheffého metoda.

Metody mnohonásobného porovnávání

Věta 9 (Tukeyova metoda)

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

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

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

Věta 10 (Scheffého metoda)

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

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

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.

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 11 (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.

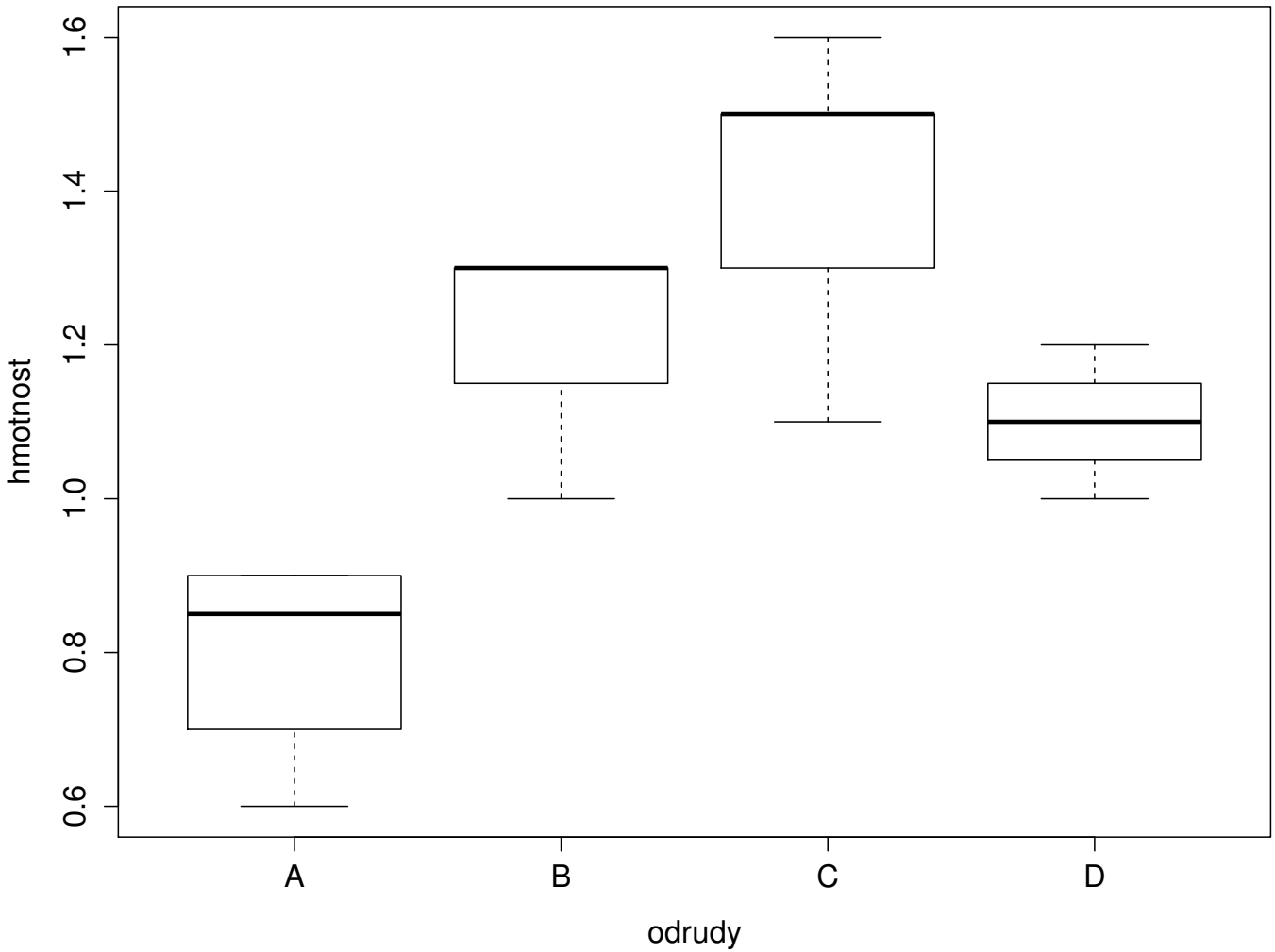
Řešení

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: $\overline{y}_{1\cdot} = 0,8$, $\overline{y}_{2\cdot} = 1,2$, $\overline{y}_{3\cdot} = 1,4$, $\overline{y}_{4\cdot} = 1,1$, $\overline{y}_{\cdot\cdot} = 1,14$, $S_e = 0,3$, $S_A = 0,816$, $S_T = 1,116$, $F_A = 9,97$. 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 Scheffého metody zjistíme, které dvojice odrůd se liší na hladině významnosti 0,05.

Srovnávané odrůdy	Rozdíly $\lvert\overline{Y}_{k\bullet} - \overline{Y}_{l\bullet}\rvert$	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.

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

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

$\boxed{M_0}$ $\quad \mathbf{Y} \sim N_n(\mathbf{X}_0\boldsymbol{\beta}_0, \sigma^2\mathbf{I}_n) \quad \mathbf{X}_0$ je typu $n \times k_0$, $\quad h(\mathbf{X}_0) = r_0$, $\quad \boldsymbol{\beta}_0$ 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 $\mathbf{K} = \begin{pmatrix} \mathbf{I}_{k_0} \\ \mathbf{0} \end{pmatrix}$ je typu $k \times k_0$.

Využití ANOVA v lineárním regresním modelu

Položme

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

pak

$$S_e = (\mathbf{Y} - \widehat{\boldsymbol{\mu}})'(\mathbf{Y} - \widehat{\boldsymbol{\mu}}) \qquad S_{e_0} = (\mathbf{Y} - \widehat{\boldsymbol{\mu}}_0)'(\mathbf{Y} - \widehat{\boldsymbol{\mu}}_0)$$

$$S_{\Delta_0} = (\widehat{\boldsymbol{\mu}} - \widehat{\boldsymbol{\mu}}_0)'(\widehat{\boldsymbol{\mu}} - \widehat{\boldsymbol{\mu}}_0) \qquad S_e = S_{e_0} - S_{\Delta_0}$$

Pokud platí model M_0, pak statistika

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

Využití ANOVA v lineárním regresním modelu

Připomeňme si ještě index (koeficient) determinace R^2 z lineárního regresního modelu a celkový F-test (průkaznosti) modelu. Obě veličiny, které v regresní analýze používáme k hodnocení kvality modelu, mají matematické pozadí právě v analýze rozptylu. Zvolený lineární regresní model M porovnáváme s minimálním (nulovým) modelem M_0.

- Celkový F-test je vlastně jen ANOVA testem modelu M vůči M_0.
- Index determinace R^2 je pak definován pomocí součtů čtverců

$$R^2 = \frac{S_A}{S_T} = 1 - \frac{S_e}{S_T}.$$

  Odtud plyne i jeho interpretace: $R^2 \in [0;1]$ značí, jak velkou část celkové variability dat se navrženým modelem M podaří vysvětlit.
- Korigovaný koficient determinace $\overline{R}^2$ obdržíme, pokud součty čtverců nahradíme nestrannými odhady středních kvadratických chyb

$$\overline{R}^2 = 1 - \frac{S_r/df_e}{S_T/df_T} = 1 - \frac{n-1}{n-a}(1-R^2).$$

Příklad

Příklad 2

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