# Osnova přednášky Analýza rozptylu jednoduchého třídění

**Motivace**

**Označení**

**Součty čtverců**

**Testování hypotézy o shodě středních hodnot**

**Testování hypotézy o shodě rozptylů**

**Post – hoc metody mnohonásobného porovnávání**

**Doporučený postup při provádění analýzy rozptylu**

**Příklad**

**ANOVA jako speciální případ obecného lineárního modelu**

**Kódování úrovní faktoru pomocí indikátorových proměnných**

**Odhady parametrů modelu**

# Analýza rozptylu jednoduchého třídění (jednofaktorová ANOVA)

**Motivace:** ANOVA je statistická metoda, která slouží k porovnání úrovně sledované náhodné veličiny Y intervalového či poměrového typu v několika populacích. Tyto populace jsou vymezeny variantami třídicího faktoru A, což je veličina nominálního nebo ordinálního typu. Počet variant faktoru A bývá poměrně malý – do 10.
Jednotlivým variantám faktoru A se říká úrovně.
Prostřednictvím náhodných výběrů z jednotlivých populací zkoumáme, zda faktor A má vliv na variabilitu hodnoty veličiny Y.

Příklady použití ANOVY:

| sledovaná veličina Y | třídicí faktor A |
|---|---|
| hmotnostní přírůstek selat | druh krmiva |
| čas dopravy do zaměstnání | druh dopravy |
| koncentrace ozónu v ovzduší | lokalita |
| počet bodů v závěrečném testu | výuková metoda |
| směnový výkon dělníků | druh osvětlení pracoviště |

Na hladině významnosti $\alpha$ testujeme nulovou hypotézu, která tvrdí, že všechny střední hodnoty jsou stejné, tj.

$H_0$: $\mu_1 = \ldots = \mu_r$

proti alternativní hypotéze

$H_1$: aspoň jedna dvojice středních hodnot se liší.

Jedná se tedy o zobecnění dvouvýběrového t-testu a na první pohled se zdá, že stačí utvořit $\binom{r}{2}$ dvojic náhodných výběrů a na každou dvojici aplikovat dvouvýběrový t-test.

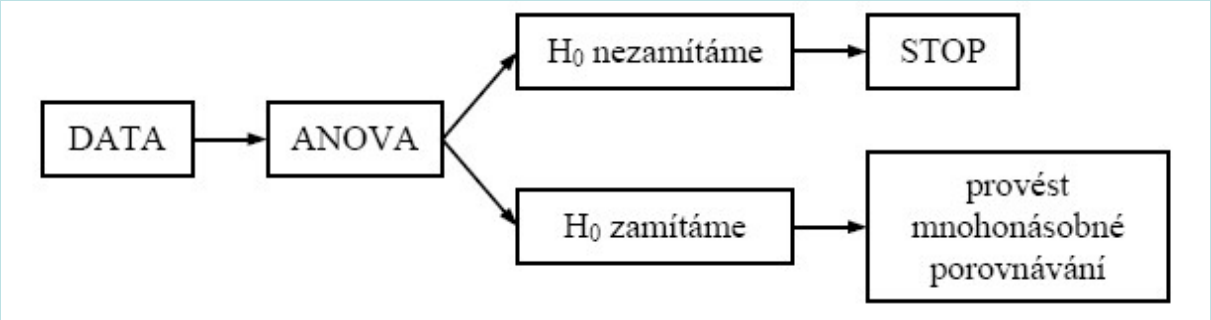
Hypotézu o shodě všech středních hodnot bychom pak zamítli, pokud aspoň v jednom případě z $\binom{r}{2}$ porovnávání se prokáže odlišnost středních hodnot. Odtud je vidět, že

k neoprávněnému zamítnutí nulové hypotézy (tj. k chybě 1. druhu) může dojít s pravděpodobností větší než $\alpha$. Tato pravděpodobnost je shora omezena číslem $1-(1-\alpha)^r$. Např. pro $\alpha = 0{,}05$ a $r = 3$ je tato pravděpodobnost 0,1426, pro $r = 4$ je 0,1855 a pro $r = 5$ dokonce 0,2262.

Proto ve 30. letech 20. století vytvořil R. A. Fisher metodu ANOVA (analýza rozptylu, v popsané situaci konkrétně analýza rozptylu jednoduchého třídění), která uvedenou podmínku splňuje.

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

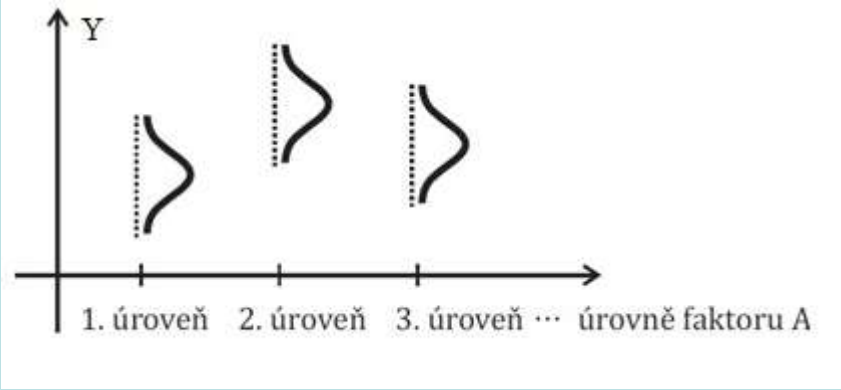
Ilustrace:



Předpokládáme, že faktor A má $r \geq 3$ úrovní a přitom i-té úrovni odpovídá $n_i \geq 2$ pozorování $Y_{i1},\ldots,Y_{in_i}$, které tvoří náhodný výběr z rozložení $N(\mu_i, \sigma^2)$, $i = 1, ..., r$ a jednotlivé náhodné výběry jsou stochasticky nezávislé.
j-té pozorování v i-tém výběru lze zapsat ve tvaru $Y_{ij} = \mu_i + \varepsilon_{ij}$, kde $\varepsilon_{ij}$ jsou stochasticky nezávislé náhodné veličiny s rozložením $N(0, \sigma^2)$, $i = 1, \ldots, r$, $j = 1, \ldots, n_i$.
Výsledky lze zapsat do tabulky

| faktor A | výsledky |
|---|---|
| úroveň 1 | $Y_{11},\ldots,Y_{1n_1}$ |
| úroveň 2 | $Y_{21},\ldots,Y_{2n_2}$ |
| … | … |
| úroveň r | $Y_{r1},\ldots,Y_{rn_r}$ |

Ilustrace:



**Označení:**

$n = \sum_{i=1}^{r} n_i$ … celkový rozsah všech r výběrů

$Y_{i.} = \sum_{j=1}^{n_i} Y_{ij}$ … součet hodnot v i-tém výběru

$M_{i.} = \frac{1}{n_i} Y_{i.}$ … výběrový průměr v i-tém výběru

$S_i^2 = \frac{1}{n_i - 1} \sum_{i=1}^{r} \left(Y_{ij} - M_{i.}\right)^2$ … výběrový rozptyl i-tého výběru

$S_*^2 = \frac{\sum_{i=1}^{r} (n_i - 1) S_i^2}{n - r}$ … vážený průměr výběrových rozptylů

$Y_{..} = \sum_{i=1}^{r} \sum_{j=1}^{n_i} Y_{ij}$ … součet hodnot všech výběrů

$M_{..} = \frac{1}{n} Y_{..}$ … celkový průměr všech r výběrů

**Součty čtverců:**

$S_T = \sum_{i=1}^{r}\sum_{j=1}^{n_i}\left(Y_{ij} - M_{..}\right)^2$ … celkový součet čtverců (charakterizuje variabilitu jednotlivých pozorování kolem celkového průměru), počet stupňů volnosti $f_T = n - 1$,

$S_A = \sum_{i=1}^{r} n_i \left(M_{i.} - M_{..}\right)^2$ … skupinový součet čtverců (charakterizuje variabilitu mezi jednotlivými náhodnými výběry), počet stupňů volnosti $f_A = r - 1$.

Podíl $\frac{S_A}{f_A} = \frac{S_A}{r-1}$ se nazývá průměrný skupinový čtverec nebo též rozptyl vysvětlený faktorem A či meziskupinový rozptyl.

$S_E = \sum_{i=1}^{r}\sum_{j=1}^{n_i}\left(Y_{ij} - M_{i.}\right)^2 = (n-r)S_*^{\,2}$ … reziduální součet čtverců (charakterizuje variabilitu uvnitř jednotlivých výběrů), počet stupňů volnosti $f_E = n - r$.

Podíl $\frac{S_E}{f_E} = \frac{S_E}{n-r}$ se nazývá průměrný reziduální čtverec nebo též rozptyl nevysvětlený faktorem A či vnitroskupinový rozptyl.

Lze dokázat, že $S_T = S_A + S_E$.
Celková variabilita sledované veličiny Y se rozkládá na variabilitu mezi výběry a variabilitu uvnitř výběrů.

Za splnění podmínky homoskedasticity (tzn., že všech r náhodných výběrů pochází z rozložení se stejným rozptylem $\sigma^2$) je průměrný reziduální čtverec $\frac{S_E}{n-r}$ nestranným odhadem neznámého rozptylu $\sigma^2$.

Za platnosti hypotézy o shodě středních hodnot je průměrný skupinový čtverec $\frac{S_A}{r-1}$ také nestranným odhadem neznámého rozptylu $\sigma^2$.

**Testování hypotézy o shodě středních hodnot**

Náhodné veličiny $Y_{ij}$ se řídí modelem

$M_0$: $Y_{ij} = \mu_i + \varepsilon_{ij} = \mu + \alpha_i + \varepsilon_{ij}$

pro $i = 1, \ldots, r$, $j = 1, \ldots, n_i$, přičemž

$\varepsilon_{ij}$ jsou stochasticky nezávislé náhodné veličiny s rozložením $N(0, \sigma^2)$,

$\mu$ je společná část střední hodnoty závisle proměnné veličiny,

$\alpha_i$ je efekt faktoru A na úrovni i.

Parametry $\mu$, $\alpha_i$ neznáme.

Požadujeme, aby platila tzv. reparametrizační rovnice: $\sum_{i=1}^{r} n_i \alpha_i = 0$.

(Pokud je třídění vyvážené, tj. všechny výběry mají stejný rozsah: $n_1 = n_2 = \ldots = n_r$,

pak lze použít zjednodušenou podmínku $\sum_{i=1}^{r} \alpha_i = 0$.)

Pokud by nezáleželo na faktoru A, platila by hypotéza $\alpha_1 = \ldots = \alpha_r = 0$ a dostali bychom model

$M_1$: $Y_{ij} = \mu + \varepsilon_{ij}$.

Během analýzy rozptylu tedy zkoumáme, zda výběrové průměry $M_{1.}, \ldots, M_{r.}$ se od sebe liší pouze v mezích náhodného kolísání kolem celkového průměru $M_{..}$ nebo zda se projevuje vliv faktoru A. Pokud převažuje vliv vnitroskupinové variability nad meziskupinovou variabilitou, vliv faktoru A je nevýznamný. V opačném případě je významný a hypotéza o shodě středních hodnot bude zamítnuta.

Rozdíl mezi modely $M_0$ a $M_1$ ověřujeme pomocí testové statistiky $F_A = \frac{S_A / f_A}{S_E / f_E}$, která se řídí rozložením $F(f_A, f_E)$, je-li model $M_1$ správný. Hypotézu o nevýznamnosti faktoru A tedy zamítneme na hladině významnosti $\alpha$, když platí: $F_A \geq F_{1-\alpha}(f_A, f_E)$.

Vidíme, že test hypotézy o shodě r středních hodnot byl převeden na ekvivalentní test hypotézy o podílu dvou rozptylů (meziskupinového a vnitroskupinového). Proto uvedená metoda nese název analýza rozptylu.

Výsledky výpočtů zapisujeme do tabulky analýzy rozptylu jednoduchého třídění.

| Zdroj variability | součet čtverců | stupně volnosti | podíl | $F_A$ |
|---|---|---|---|---|
| skupiny | $S_A$ | $f_A = r - 1$ | $S_A/f_A$ | $\frac{S_A/f_A}{S_E/f_E}$ |
| reziduální | $S_E$ | $f_E = n - r$ | $S_E/f_E$ | - |
| celkový | $S_T$ | $f_T = n - 1$ | - | - |

Sílu závislosti náhodné veličiny Y na faktoru A můžeme měřit pomocí poměru determinace: $P^2 = \frac{S_A}{S_T}$. Nabývá hodnot z intervalu $\langle 0,1 \rangle$.

**Testování hypotézy o shodě rozptylů**

Před provedením analýzy rozptylu je zapotřebí ověřit předpoklad o shodě rozptylů v daných r výběrech.

a) **Levenův test**: Položme $Z_{ij} = \left|Y_{ij} - M_{i.}\right|$. Označíme

$$M_{Zi} = \frac{1}{n_i}\sum_{j=1}^{n_i} Z_{ij}\,,\quad M_Z = \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{n_i} Z_{ij}\,,\quad S_{ZE} = \sum_{i=1}^{r}\sum_{j=1}^{n_i}\left(Z_{ij} - M_{Zi}\right)^2\,,\quad S_{ZA} = \sum_{i=1}^{r} n_i\left(M_{Zi} - M_Z\right)^2$$

Platí-li hypotéza o shodě rozptylů, pak statistika

$$F_{ZA} = \frac{S_{ZA}/(r-1)}{S_{ZE}/(n-r)} \approx F(r-1,\, n-r).$$

Hypotézu o shodě rozptylů tedy zamítáme na asymptotické hladině významnosti $\alpha$, když $F_{ZA} \geq F_{1-\alpha}(r-1,\, n-r)$.

(Levenův test je vlastně založen na analýze rozptylu absolutních hodnot centrovaných pozorování. Vzhledem k tomu, že náhodné veličiny $X_{ij} - M_i$ nejsou stochasticky nezávislé a absolutní hodnoty těchto veličin nemají normální rozložení, je Levenův test pouze aproximativní.)

b) **Brownův – Forsytheův test** je modifikací Levenova testu. Modifikace spočívá v tom, že místo výběrového průměru i-tého výběru se při výpočtu veličiny $z_{ij}$ používá medián i-tého výběru.

c) **Bartlettův test:** Platí-li hypotéza o shodě rozptylů a rozsahy všech výběrů jsou větší než 6, pak statistika

$$B = \frac{1}{C}\left[(n-r)\ln S_*^{\,2} - \sum_{i=1}^{r}(n_i - 1)\ln S_i^{\,2}\right]$$ se asymptoticky řídí rozložením $\chi^2(r-1)$. Přitom konstanta $$C = 1 + \frac{1}{3(r-1)}\left(\sum_{i=1}^{r}\frac{1}{n_i - 1} - \frac{1}{n-r}\right)$$ a $S_*^{\,2}$ je vážený průměr výběrových rozptylů.

$H_0$ zamítáme na asymptotické hladině významnosti $\alpha$, když statistika B se realizuje v kritickém oboru $W = \langle \chi^2_{1-\alpha}(r-1), \infty)$.

**Poznámka k testům homogenity rozptylů:** Ze simulačních studií vyplývá, že pravděpodobnost chyby 1. druhu (tj. pravděpodobnost neoprávněného zamítnutí pravdivé nulové hypotézy) je u Bartlettova testu blízká obvykle volené hladině významnosti 0,05 pouze pro výběry z normálního rozložení. Pro větší počty výběrů z výrazně nenormálních rozložení (např. výběry z exponenciálního rozložení) výrazně stoupá pravděpodobnost chyby 1. druhu. Naopak Brownův – Forsytheův test udrží nízkou pravděpodobnost chyby 1. druhu i pro velký počet výběrů pocházejících z nenormálních rozložení.

**Post – hoc metody mnohonásobného porovnávání**

Zamítneme-li na hladině významnosti $\alpha$ hypotézu o shodě středních hodnot, chceme zjistit, které dvojice středních hodnot se liší na dané hladině významnosti $\alpha$, tj. na hladině významnosti $\alpha$ testujeme $H_0$: $\mu_l = \mu_k$ proti $H_1$: $\mu_l \neq \mu_k$ pro všechna l, k = 1, .., r, $l \neq k$.

a) Mají-li všechny výběry týž rozsah p (říkáme, že třídění je vyvážené), použijeme **Tukeyovu metodu**.

Testová statistika má tvar $\dfrac{|M_{k.} - M_{l.}|}{\dfrac{S_*}{\sqrt{p}}}$. Rovnost středních hodnot $\mu_k$ a $\mu_l$ zamítneme na hladině významnosti $\alpha$, když $\dfrac{|M_{k.} - M_{l.}|}{\dfrac{S_*}{\sqrt{p}}} \geq q_{1-\alpha}(r, n-r)$, kde hodnoty $q_{1-\alpha}(r, n-r)$ jsou kvantily studentizovaného rozpětí a najdeme je ve statistických tabulkách.

(Studentizované rozpětí je náhodná veličina $Q = \dfrac{Y_{(n)} - Y_{(1)}}{s}$.)

Existuje modifikace Tukeyovy metody pro nestejné rozsahy výběrů, nazývá se **Tukeyova HSD metoda**.

V tomto případě má testová statistika tvar $\frac{\left|M_{k.}-M_{l.}\right|}{S_*\sqrt{\frac{1}{2}\left(\frac{1}{n_k}+\frac{1}{n_l}\right)}}$. Rovnost středních hodnot $\mu_k$ a $\mu_l$ zamítneme na hladině významnosti $\alpha$, když $\frac{\left|M_{k.}-M_{l.}\right|}{S_*\sqrt{\frac{1}{2}\left(\frac{1}{n_k}+\frac{1}{n_l}\right)}} \geq q_{1-\alpha}(r, n-r)$.

b) Nemají-li všechny výběry stejný rozsah, použijeme **Scheffého metodu**: rovnost středních hodnot $\mu_k$ a $\mu_l$ zamítneme na hladině významnosti $\alpha$, když

$$|M_{k.} - M_{l.}| \geq S_* \sqrt{(r-1)\left(\frac{1}{n_k} + \frac{1}{n_l}\right) F_{1-\alpha}(r-1, n-r)}\,.$$

Výhodou Scheffého testu je, že k jeho provedení nepotřebujeme speciální statistické tabulky s hodnotami kvantilů studentizovaného rozpětí, ale stačí běžné statistické tabulky s kvantily Fisherova – Snedecorova rozložení.

V případě vyváženého třídění, kdy lze aplikovat Tukeyovu i Scheffého metodu, použijeme tu, která je citlivější. Tukeyova metoda tedy bude výhodnější, když $q_{1-\alpha}^2(r, n-r) < 2(r-1)F_{1-\alpha}(r-1, n-r)$.

Metody mnohonásobného porovnávání mají obecně menší sílu než ANOVA. Může nastat situace, kdy při zamítnutí $H_0$ nenajdeme metodami mnohonásobného porovnávání významný rozdíl u žádné dvojice středních hodnot. K tomu dochází zvláště tehdy, když p-hodnota pro ANOVU je jen o málo nižší než zvolená hladina významnosti. Pak slabší test patřící do skupiny metod mnohonásobného porovnávání nemusí odhalit žádný rozdíl.

**Doporučený postup při provádění analýzy rozptylu:**

a) **Ověření normality daných r náhodných výběrů** (grafické metody - NP plot, Q-Q plot, histogram, testy hypotéz o normálním rozložení - Lilieforsova varianta Kolmogorovova – Smirnovova testu, Shapirův – Wilkův test, Andersonův – Darlingův test).
Doporučuje se kombinace obou způsobů. Závěry učiníme až na základě posouzení obou výsledků.

**Upozornění:** Při malých rozsazích výběrů se nedoporučuje zkoumat normalitu jednotlivých náhodných výběrů, ale normalitu reziduí $Y_{ij} – M_{i.}$.

Obecně lze říci, že analýza rozptylu není příliš citlivá na porušení předpokladu normality, zvláště při větších rozsazích výběrů (nad 20), což je důsledek působení centrální limitní věty. Mírné porušení normality tedy není na závadu, při větším porušení použijeme např. Kruskalův – Wallisův test jako neparametrickou obdobu analýzy rozptylu jednoduchého třídění.

b) Po ověření normality se testuje **homogenita rozptylů**, tj. předpoklad, že všechny náhodné výběry pocházejí z normálních rozložení s týmž rozptylem. Graficky ověřujeme shodu rozptylů pomocí krabicových diagramů, kdy sledujeme, zda je šířka krabic stejná. Numericky testujeme homogenitu rozptylů pomocí Levenova testu, Brownova – Forsytheova testu (oba jsou implementovány ve STATISTICE, Brownův – Forsytheův test v MINITABu) či Bartlettova testu (je k dispozici v MINITABu). Lze rovněž vytvořit graf závislosti reziduí $Y_{ij} – M_{i.}$ na variantách faktoru A. Měl by to být náhodný mrak bodů.
Při vyváženém třídění se nemusí zkoumat homogenita rozptylů.
Slabé porušení homogenity rozptylů nevadí, při větším se doporučuje použít v ANOVĚ Welchovu aproximaci nebo z neparametrických metod mediánový test.

c) Pokud jsou splněny předpoklady normality a homogenity rozptylů, můžeme přistoupit k **testování shody středních hodnot**. Předtím je samozřejmě vhodné vypočítat průměry a směrodatné odchylky či rozptyly v jednotlivých skupinách.

d) Dojde-li na zvolené hladině významnosti k zamítnutí hypotézy o shodě středních hodnot, zajímá nás, které dvojice středních hodnot se od sebe liší. K řešení tohoto problému slouží **post-hoc metody mnohonásobného porovnávání**, např. Scheffého nebo Tukeyova metoda.

**Příklad:** V rámci psychologického výzkumu bylo náhodně vybráno devět dvanáctiletých dětí a to tak, že tři děti měly matku se základním vzděláním, tři se středoškolským a tři s vysokoškolským. Všechny děti byly podrobeny témuž testu. Počty bodů, které děti v testu získaly, jsou uvedeny v tabulce:

| Vzdělání matky | Počet bodů | | |
|---|---|---|---|
| Základní (ZŠ) | 20 | 23 | 22 |
| Středoškolské (SŠ) | 24 | 26 | 25 |
| Vysokoškolské (VŠ) | 26 | 27 | 27 |

Na hladině významnosti 0,05 testujte hypotézu, že střední hodnota počtu bodů v testu nezávisí na vzdělání matky. Zamítnete-li nulovou hypotézu, zjistěte, které dvojice úrovní vzdělání se liší na hladině významnosti 0,05. Vypočtěte též poměr determinace.

**Řešení:** Data považujeme za realizace tří nezávislých náhodných výběrů ze tří normálních rozložení se stejným rozptylem. Testujeme hypotézu, že všechny tři střední hodnoty jsou stejné.

Výběrové průměry v jednotlivých výběrech: $M_{1.} = 21{,}67$, $M_{2.} = 25$, $M_{3.} = 26{,}67$,

celkový průměr: $M_{..} = 24{,}44$,

výběrové rozptyly: $S_1^2 = 2{,}33$, $S_2^2 = 1$, $S_3^2 = 0{,}33$,

vážený průměr výběrových rozptylů: $S_*^2 = \dfrac{\sum_{i=1}^{r}(n_i - 1)S_i^2}{n - r} = \dfrac{2 \cdot 2{,}33 + 2 \cdot 1 + 2 \cdot 0{,}33}{9 - 3} = 1{,}22$,

reziduální součet čtverců: $S_E = (n - r)S_*^2 = 6 \cdot 1{,}22 = 7{,}33$,

skupinový součet čtverců:

$$S_A = \sum_{i=1}^{r} n_i (M_{i.} - M_{..})^2 = 3 \cdot (21{,}67 - 24{,}44)^2 + 3 \cdot (25 - 24{,}44)^2 + 3 \cdot (26{,}67 - 24{,}44)^2 = 38{,}89$$

celkový součet čtverců: $S_T = S_A + S_E = 38{,}89 + 7{,}33 = 46{,}22$,

testová statistika $F_A = \dfrac{S_A / f_A}{S_E / f_E} = \dfrac{38{,}89/2}{7{,}33/6} = 15{,}9091$,

kritický obor $W = \langle F_{0{,}95}(2,6), \infty) = \langle 5{,}1433, \infty)$. Protože testová statistika se realizuje v kritickém oboru, $H_0$ zamítáme na hladině významnosti 0,05.

Vypočteme poměr determinace: $P^2 = \frac{S_A}{S_T} = \frac{38,89}{46,22} = 0,8414$.

Výsledky zapíšeme do tabulky ANOVA:

| Zdroj variability | Součet čtverců | Stupně volnosti | podíl | $F_A$ |
|---|---|---|---|---|
| skupiny | $S_A = 38,89$ | 2 | $S_A/2 = 19,44$ | $\frac{S_A/(r-1)}{S_E/(n-r)} = 15,9091$ |
| reziduální | $S_E = 7,33$ | 6 | $S_E/6 = 1,22$ | - |
| celkový | $S_T = 46,22$ | 8 | - | - |

Nyní pomocí Tukeyovy metody zjistíme, které dvojice úrovní vzdělání se liší na hladině významnosti 0,05: rovnost středních hodnot $\mu_k$ a $\mu_l$ zamítneme na hladině významnosti $\alpha$, když $\dfrac{|M_{k.} - M_{l.}|}{\dfrac{S_*}{\sqrt{p}}} \geq q_{1-\alpha}(r, n-r)$ neboli $|M_{k.} - M_{l.}| \geq \dfrac{S_*}{\sqrt{p}} q_{1-\alpha}(r, n-r)$.

V našem případě $S_* = \sqrt{1{,}22} = 1{,}105$, $p = 3$, $q_{0,95}(3,6) = 4{,}34$, tedy

$$\frac{S_*}{\sqrt{p}} q_{1-\alpha}(r, n-r) = \frac{1{,}105}{\sqrt{3}} 4{,}34 = 2{,}77$$

| Srovnávané dvojice | Rozdíly $\lvert M_{k.} - M_{l.} \rvert$ | Pravá strana vzorce |
|---|---|---|
| (ZŠ, SŠ) | $\lvert M_{1.} - M_{2.} \rvert = \lvert 21{,}67 - 25 \rvert = 3{,}33$ | 2,77 |
| (ZŠ, VŠ) | $\lvert M_{1.} - M_{3.} \rvert = \lvert 21{,}67 - 26{,}67 \rvert = 5$ | 2,77 |
| (SŠ, VŠ) | $\lvert M_{2.} - M_{3.} \rvert = \lvert 25 - 26{,}67 \rvert = 1{,}67$ | 2,77 |

Na hladině významnosti 0,05 se liší dvojice (ZŠ, SŠ) a (ZŠ, VŠ).

## Řešení pomocí systému R

Načteme data:

```
Y<-c(20, 23, 22, 24, 26, 25, 26, 27, 27)
ID<-c(1,1,1,2,2,2,3,3,3)
ID<-factor(ID,labels=c('ZS','SS','VS'))
```

Vypočteme průměry v jednotlivých skupinách:

```
tapply(Y,ID,mean)
ZS       SS       VS
21.66667 25.00000 26.66667
```

Vypočteme směrodatné odchylky v jednotlivých skupinách:

```
tapply(Y,ID,sd)
       ZS        SS        VS
1.5275252 1.0000000 0.5773503
```

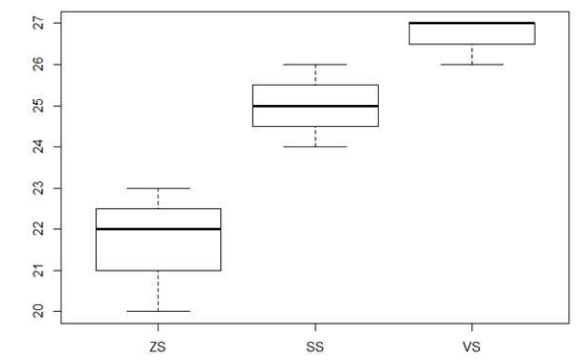
Testujeme hypotézu o shodě středních hodnot:

```
vystup<-aov(Y~ID)
summary(vystup)
            Df Sum Sq Mean Sq F value  Pr(>F)
ID           2  38.89  19.444   15.91 0.00399 **
Residuals    6   7.33   1.222
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Vidíme, že p-hodnota testu o shodě středních hodnot je 0,00399, což je menší než 0,05, tedy na hladině významnosti 0,05 nulovou hypotézu zamítáme. S rizikem omylu nejvýše 5 % jsme prokázali, že výsledky testu se v daných třech skupinách dětí roztříděných podle vzdělání matky se liší.

Vykreslíme krabicové diagramy:

```
boxplot(Y~ID)
```




Provedeme Tukeyovu metodu mnohonásobného porovnávání.

Nejprve načteme knihovnu DescTools: `library(DescTools)`

Použijeme funkci PostHocTest: `PostHocTest(vystup,method=c('hsd'))`

```
Posthoc multiple comparisons of means : Tukey HSD
    95% family-wise confidence level

$ID
          diff     lwr.ci   upr.ci   pval
SS-ZS 3.333333  0.5636912 6.102975 0.0237 *
VS-ZS 5.000000  2.2303579 7.769642 0.0035 **
VS-SS 1.666667 -1.1029755 4.436309 0.2339

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Na hladině významnosti 0,05 se liší dvojice (SŠ, ZŠ) a (VŠ, ZŠ).

## ANOVA jako speciální případ obecného lineárního modelu

Na analýzu rozptylu lze pohlížet jako na speciální případ obecného lineárního modelu. To nám umožní nejenom odhalit případný statisticky významný vliv faktoru A na variabilitu hodnot veličiny Y, ale také interpretovat odhady parametrů modelu a predikovat hodnoty veličiny Y pomocí úrovní faktoru A.

### Kódování úrovní faktoru pomocí indikátorových proměnných

Abychom mohli použít tento přístup, musíme si ukázat, jak se kódují jednotlivé úrovně faktoru A. Jednotlivé úrovně (je jich $r \geq 3$) vlastně představují klasifikaci objektů do skupin. Příslušnost objektů ke skupinám se vyjadřuje pomocí umělých proměnných, tzv. indikátorů. Používá se několik typů kódování. Způsob kódování vysvětlíme na příkladu faktoru se třemi úrovněmi.
Zkoumanou veličinou Y je počet bodů, které dítě získalo v testu a faktorem A je nejvyšší dosažené vzdělání matky (ZŠ, SŠ, VŠ).

**a) Kódování přeparametrizovaného modelu**

Zavedeme r závislých indikátorů $Z_1, \ldots, Z_r$ tak, že každý z nich vyjadřuje vždy jednu úroveň faktoru A hodnotou 1 a všechny ostatní hodnotou 0.

V našem případě zavedeme tři indikátory $Z_1, Z_2, Z_3$ takto:

$$Z_1 = \begin{cases} 1 \text{ pro ZŠ} \\ 0 \text{ jinak} \end{cases}, \quad Z_2 = \begin{cases} 1 \text{ pro SŠ} \\ 0 \text{ jinak} \end{cases}, \quad Z_3 = \begin{cases} 1 \text{ pro VŠ} \\ 0 \text{ jinak} \end{cases}.$$

Vyjádřeno tabulkou:

| Úroveň faktoru | indikátory | | |
|---|---|---|---|
| | $Z_1$ | $Z_2$ | $Z_3$ |
| ZŠ | 1 | 0 | 0 |
| SŠ | 0 | 1 | 0 |
| VŠ | 0 | 0 | 1 |

Součet v každém sloupci tabulky je 1.

Každý indikátor je možno vyjádřit jako lineární kombinaci ostatních indikátorů. Tato vlastnost je pro mnohé statistické postupy nežádoucí, proto budeme uvažovat o jeden indikátor méně. Vynechaná úroveň faktoru bude sloužit jako referenční. Referenční úroveň volíme tak, aby to bylo výhodné z interpretačního hlediska.

## b) Kódování typu dummy

Zavedeme r-1 nezávislých indikátorů $Z_1, \ldots, Z_{r-1}$, které jsou definovány takto:
$Z_1 = 1$ pro 1. úroveň faktoru A, $Z_1 = 0$ jinak,
$Z_2 = 1$ pro 2. úroveň faktoru A, $Z_2 = 0$ jinak,
……..
$Z_{r-1} = 1$ pro (r-1). úroveň faktoru A, $Z_{r-1} = 0$ jinak.
Pro r-tou úroveň faktoru A nabývají všechny indikátory typu dummy $Z_1, \ldots, Z_{r-1}$ hodnoty 0 a tím indikují její výskyt.
V našem případě máme dva indikátory:

$$Z_1 = \begin{cases} 1 \text{ pro ZŠ} \\ 0 \text{ jinak} \end{cases}, \quad Z_2 = \begin{cases} 1 \text{ pro SŠ} \\ 0 \text{ jinak} \end{cases}$$

. Vynechaná úroveň VŠ je referenční.
Vyjádřeno tabulkou:

| Úroveň faktoru | indikátory | |
|---|---|---|
| | $Z_1$ | $Z_2$ |
| ZŠ | 1 | 0 |
| SŠ | 0 | 1 |
| VŠ | 0 | 0 |

Součet v každém sloupci tabulky je 1. Při interpretaci výsledků analýz s indikátory typu dummy konfrontujeme jednotlivé úrovně faktoru A s referenční úrovní.

## c) Kódování typu effect

Zavedeme r-1 nezávislých indikátorů $Z_1, \ldots, Z_{r-1}$, které jsou definovány takto:
$Z_1 = 1$ pro 1. úroveň faktoru A, $Z_1 = -1$ pro r-tou úroveň faktoru A, $Z_1 = 0$ jinak,
$Z_2 = 1$ pro 2. úroveň faktoru A, $Z_2 = -1$ pro r-tou úroveň faktoru A, $Z_2 = 0$ jinak,
……
$Z_{r-1} = 1$ pro (r-1). úroveň faktoru A, $Z_{r-1} = -1$ pro r-tou úroveň faktoru A, $Z_{r-1} = 0$ jinak,
Pro r-tou úroveň faktoru A nabývají všechny indikátory typu effect $Z_1, \ldots, Z_{r-1}$ hodnoty -1 a tím indikují její výskyt.
V našem případě máme dva indikátory:

$$Z_1 = \begin{cases} 1 \text{ pro ZŠ} \\ -1 \text{ pro VŠ} \\ 0 \text{ jinak} \end{cases}, \quad Z_2 = \begin{cases} 1 \text{ pro SŠ} \\ -1 \text{ pro VŠ} \\ 0 \text{ jinak} \end{cases}$$

. Vynechaná úroveň VŠ je referenční.

Vyjádřeno tabulkou:

| Úroveň faktoru | indikátory | |
|---|---|---|
| | $Z_1$ | $Z_2$ |
| ZŠ | 1 | 0 |
| SŠ | 0 | 1 |
| VŠ | -1 | -1 |

Součet v každém sloupci tabulky je 0. Hovoříme o sigma omezené parametrizaci. Při interpretaci výsledků analýz s indikátory typu effect konfrontujeme jednotlivé úrovně faktoru A s celkovým průměrem veličiny Y.

## Odhady parametrů modelu

Model $Y_{ij} = \mu_i + \varepsilon_{ij} = \mu + \alpha_i + \varepsilon_{ij}$

lze vyjádřit jako

$Y_{ij} = \beta_0 + \beta_1 Z_{1j} + \beta_2 Z_{2j} + \varepsilon_{ij}$.

**Ad a) Odhady při kódování pomocí indikátorů typu dummy:**

$\hat{\beta}_0 = M_{3.}$, $\hat{\beta}_1 = M_{1.} - M_{3.}$, $\hat{\beta}_2 = M_{2.} - M_{3.}$

V našem případě:

$\hat{\beta}_0 = M_{3.} = 26{,}67$, $\hat{\beta}_1 = M_{1.} - M_{3.} = 21{,}67 - 26{,}67 = -5$, $\hat{\beta}_2 = M_{2.} - M_{3.} = 25 - 26{,}67 = -1{,}67$

Interpretace $\hat{\beta}_1$: Bude-li mít matka ZŠ vzdělání, tak výsledek dítěte v testu bude v průměru horší o 5 bodů oproti potomkovi matky s VŠ vzděláním.

Interpretace $\hat{\beta}_2$: Bude-li mít matka SŠ vzdělání, tak výsledek dítěte v testu bude v průměru horší o 1,67 bodů oproti potomkovi matky s VŠ vzděláním.

**Ad b) Odhady při kódování pomocí indikátorů typu effect:**

$\hat{\beta}_0 = M_{..}$, $\hat{\beta}_1 = \hat{\alpha}_1 = M_{1.} - M_{..}$, $\hat{\beta}_2 = \hat{\alpha}_2 = M_{2.} - M_{..}$. Odhad efektu 3. úrovně faktoru A získáme jako $\hat{\alpha}_3 = M_{3.} - M_{..}$.

V našem případě:

$\hat{\beta}_0 = M_{..} = 24{,}44$, $\hat{\beta}_1 = M_{1.} - M_{..} = 21{,}67 - 24{,}44 = -2{,}77$, $\hat{\beta}_2 = M_{2.} - M_{..} = 25 - 24{,}44 = -0{,}56$,

$\hat{\alpha}_3 = M_{3.} - M_{..} = 26{,}67 - 24{,}44 = 2{,}23$

Interpretace $\hat{\beta}_1 = \hat{\alpha}_1$: Bude-li mít matka ZŠ vzdělání, tak výsledek dítěte v testu bude v průměru horší o 2,77 bodů oproti průměrnému výsledku všech dětí.

Interpretace $\hat{\beta}_2 = \hat{\alpha}_2$: Bude-li mít matka SŠ vzdělání, tak výsledek dítěte v testu bude v průměru horší o 0,56 bodů oproti průměrnému výsledku všech dětí.

Interpretace $\hat{\alpha}_3$: Bude-li mít matka VŠ vzdělání, tak výsledek dítěte v testu bude v průměru lepší o 2,23 bodů oproti průměrnému výsledku všech dětí.

**Upozornění:** Kromě uvedených bodových odhadů parametrů $\beta_j$, j = 0, 1, …, r-1 lze získat také 100 (1-α)% intervaly spolehlivosti pro tyto parametry a lze vypočítat predikované hodnoty veličiny Y. Predikovaná hodnota Y v i-té skupině se nahradí skupinovým průměrem $M_{i.}$.

V našem případě uvedeme meze 95% intervalů spolehlivosti pro $\beta_0$, $\beta_1$, $\beta_2$:

a) při kódování pomocí indikátorů typu dummy:

$25{,}10 < \beta_0 < 28{,}23$; $-7{,}21 < \beta_1 < -2{,}79$; $-3{,}88 < \beta_2 < 0{,}54$

b) při kódování pomocí indikátorů typu effect:

$23{,}54 < \beta_0 < 25{,}35$; $-4{,}05 < \beta_1 < -1{,}50$; $-0{,}72 < \beta_2 < 1{,}83$.

Dále uvedeme predikované hodnoty počtu bodů v testu:

$$\hat{Y}_{11} = \hat{Y}_{12} = \hat{Y}_{13} = M_{1.} = 21{,}67$$

$$\hat{Y}_{21} = \hat{Y}_{22} = \hat{Y}_{23} = M_{2.} = 25$$

$$\hat{Y}_{31} = \hat{Y}_{32} = \hat{Y}_{33} = M_{3.} = 26{,}67$$