Cvičení z biostatistiky 08

Poslední úprava dokumentu: 10. dubna 2024.

Analýza rozptylu jednoduchého třídění

Rozcvička

Představte si, že pomocí jednovýběrového t-testu testujeme hypotézu $H_0 : \mu = 3$ proti alternativě $H_1 : \mu \neq 3$. Test jsme provedli na základě 36 pozorování a vyšla nám hodnota testové statistiky $t = 1.8$. Na základě statistických tabulek (viz 7. cvičení) rozhodněte, zda H_0 zamítnout, nebo nikoli. Hladinu testu α uvažujte 5 %.

Je známo, že testová statistika Z má za platnosti H_0 normované normální rozdělení. Pro naše data nám vyšla hodnota $Z = 2.1$. Zamítneme na hladině 5 % nulovou hypotézu ve prospěch oboustranné alternativy?

1 Úvod

Připomeňte si problematiku jednoduchého třídění buď z přednášky, nebo z textu Analýza rozptylu (dále jako *ANV*), který naleznete na mých stránkách.

Budeme pokračovat v analýze datové tabulky Kojeni, s níž jsme poprvé pracovali na 4. cvičení. V pracovním listu 06 lze nalézt podrobnější popis jednotlivých proměnných.

Spusťte si RStudio a načtěte dříve uložená data Kojeni. Zajistěte si přímý přístup k proměnným:

```
attach(Kojeni)
```

2 Závislost hmotnosti dítěte (ve 24. týdnu) na vzdělání matky

Pokusíme se zjistit, zda hmotnost dítěte ve 24. týdnu souvisí se vzděláním matky.

1) Spočtěte základní popisné statistiky pro hmotnost dítěte ve 24. týdnu (hmotnost) v závislosti na vzdělání matky (Vzdelani) a vhodně je graficky znázorněte.

```
tapply(hmotnost, Vzdelani, summary)
boxplot(hmotnost~Vzdelani)
```

2) Znázorněte graficky průměrné hmotnosti v závislosti na vzdělání matky spolu s 95% intervaly spolehlivosti pro střední hmotnosti.

```
library(RcmdrMisc)
plotMeans(hmotnost, Vzdelani, error.bars = "conf.int", level = 0.95)
```

Pravděpodobnostní model a formulace hypotéz

Označme jako Y_1 hmotnost náhodně vybraného[1] dítěte matky se základoškolským vzděláním, jako Y_2 hmotnost náhodně vybraného dítěte matky s maturitou, a jako Y_3 hmotnost náhodně vybraného

[1] míněno vždy z celé populace

dítěte matky s vysokoškolským vzděláním. V případě, že lze předpokládat, že $Y_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$, $Y_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$, $Y_3 \sim \mathcal{N}(\mu_3, \sigma_3^2)$, a platí-li navíc $\sigma_1 = \sigma_2 = \sigma_3$, umíme pomocí analýzy rozptylu testovat $\mathrm{H}_0 : \mu_1 = \mu_2 = \mu_3$ proti $\mathrm{H}_1 : \mu_1 \neq \mu_2 \ \vee \ \mu_1 \neq \mu_3 \ \vee \ \mu_2 \neq \mu_3$ (kde $\vee$ znamená „nebo“)[2].

Ověřování předpokladů

3) Normalitu rozdělení veličin Y_1, Y_2 a Y_3 budeme zkoumat až na konci.

4) Zjistěme si alespoň, zda je smysluplné předpokládat shodnou variabilitu hmotnosti dětí v jednotlivých skupinách matek dle jejich vzdělání. K tomu lze použít např. Leveneův test z knihovny `car`

```
install.packages("car")          # instalace knihovny (trvá dlouho)
library(car)                     # otevření knihovny
leveneTest(hmotnost ~ Vzdelani)  # Leveneův test
```

nebo z knihovny `lawstat`

```
install.packages("lawstat")      # instalace knihovny (rychlejší)
library(lawstat)                 # otevření knihovny
levene.test(hmotnost, Vzdelani)  # Leveneův test
```

✧ Z výstupu nás zajímá pouze p-hodnota, která v tomto případě činí 0.34. Jelikož je p-hodnota větší než zvolená hladina $\alpha = 0.05$, nezamítáme shodu populačních rozptylů.

Obdobně lze použít též Bartlettův test, který je však mnohem citlivější vůči případné nenormalitě.

```
bartlett.test(hmotnost ~ Vzdelani)
```

✧ I zde nás případně zajímá pouze p-hodnota, která vyšla 0.6276. Jelikož je to hodnota vyšší než předpokládaná hladina $\alpha = 0.05$, nezamítáme shodu populačních rozptylů hmotností v jednotlivých skupinách a můžeme ji tedy předpokládat.

V praxi samozřejmě stačí použít pouze jeden z těchto testů - buď Leveneův, nebo Bartlettův. Vyšlo nám, že shodu populačních rozptylů hmotností v jednotlivých skupinách lze předpokládat. Označme si hodnotu tohoto společného rozptylu jako σ^2 (tj. předpokládáme $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 =: \sigma^2$).

Test hypotézy

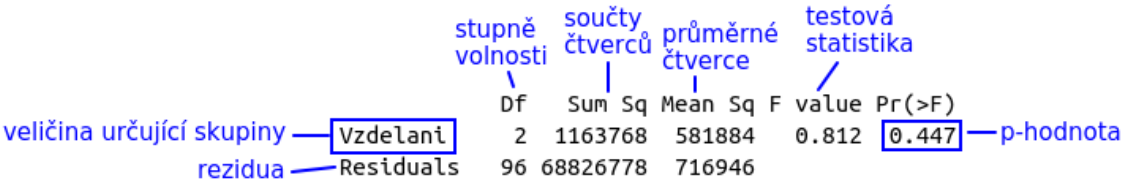
5) Nyní pomocí analýzy rozptylu otestujme

$$H_0 : \mu_1 = \mu_2 = \mu_3 \quad \text{proti}$$
$$H_1 : \mu_1 \neq \mu_2 \ \vee \ \mu_1 \ \neq \ \mu_3 \ \vee \ \mu_2 \neq \mu_3.$$

Hladinu testu budeme uvažovat 5 %, tj. $\alpha = 0.05$.

```
mod01 <- aov(hmotnost ~ Vzdelani)   # vytvoření modelu
summary(mod01)                      # tabulka analýzy rozptylu
```

[2] Uvědomte si, že takto definovaná H_1 je opravdu negací H_0



✧ vysvětlení anglických zkratek:

$$\texttt{Df} = \text{degrees of freedom}$$
$$\texttt{Sum Sq} = \text{sum of squares}$$
$$\texttt{Mean Sq} = \text{mean squares}$$
$$\texttt{F value} = \text{value of F-statistics}$$

✧ Porovnejte si tabulku z výstupu s tabulkou z textu *ANV* na str. 4. Vidíme, že:

$$\text{Součet čtverců } S_A = 1163768$$
$$\text{Reziduální součet čtverců } S_e = 68826778$$
$$\text{Reziduální rozptyl } s^2 = 716946 \quad \text{(v textu } ANV \text{ vzorec (2))}$$
$$\text{Hodnota testové statistiky } F = 0.812 \quad \text{(v textu } ANV \text{ vzorec (3))}$$
$$\text{p-hodnota } = 0.447$$

Celková variabilita S_T v tabulce uvedena není, neboť je součtem S_A a S_e.

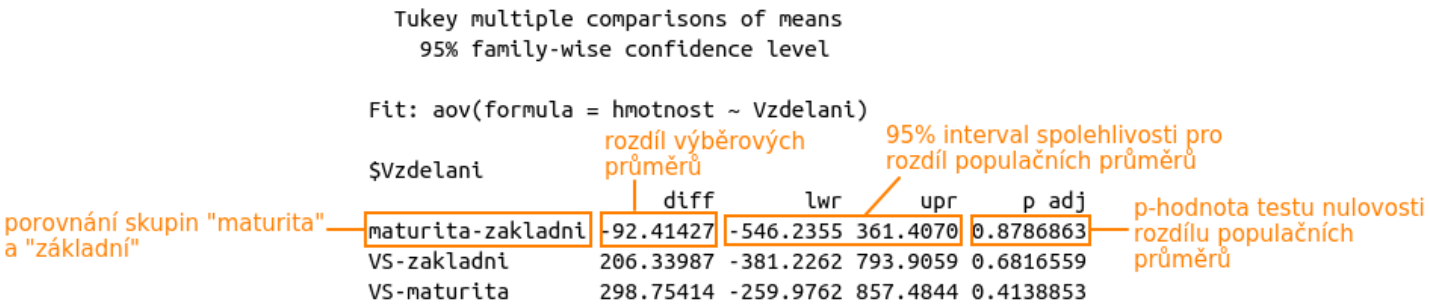
✧ Jaký je závěr (zamítáme/nezamítáme H_0 na požadované hladině)?
P-hodnota vyšla 0.447, což je víc než zvolená hladina 0.05, tudíž nulovou hypotézu nezamítáme.

✧ Jak byste formulovali svůj závěr bez použití výrazů „zamítáme/nezamítáme H_0“?
Data neprokázala, že by se hmotnost dětí ve 24. týdnu lišila v závislosti na vzdělání matky.

✧ Jaký je odhad společného rozptylu hmotností σ^2?
Tím odhadem je reziduální rozptyl $s^2 = \frac{S_e}{f_e} = 716946$.

6) V případě, kdy je nulová hypotéza zamítnuta, potřebujeme zjistit, u kterých skupin se populační průměry liší. K tomu slouží mnohonásobné porovnávání, které se většinou provádí Tukeyho metodou. V našem příkladu jsme rovnost populačních průměrů hmotností v jednotlivých skupinách nezamítli, takže mnohonásobné porovnávání zde nemá velký smysl. Pojd'me se na něj ale podívat aspoň ze cvičných důvodů.

```
TukeyHSD(mod01)                    # mnohonásobné porovnání
```



✧ Z dřívějška máme označeno, že μ_1 je populační průměr hmotností dětí u matek se ZŠ. Označme si dále $\bar{y}_1$ výběrový průměr hmotností dětí od matek se ZŠ v našich datech. Analogicky pak budeme mít μ_2 = populační průměr hmotností dětí u matek se SŠ a $\bar{y}_2$ příslušný výběrový průměr. A nakonec μ_3 je populační průměr hmotností dětí u matek s VŠ a $\bar{y}_3$ příslušný výběrový průměr. Z výstupu výše vidíme, že:

➔ $\bar{y}_2 - \bar{y}_1 = -92.41427$

➔ 95% interval spolehlivosti pro $\mu_2 - \mu_1$ je $(-546.24, 361.41)$

➔ p-hodnota testu hypotézy $H_0 : \mu_2 = \mu_1$ je 0.8787

Analogicky pro ostatní dvojice. Všechny intervaly spolehlivosti obsahují nulu, tedy u porovnání všech dvojic má nula šanci být skutečným rozdílem příslušných populačních průměrů. Stejně tak všechny p-hodnoty jsou větší než 0.05.

✧ Zkratka `p adj` by se dala přečíst jako „p-value **adjusted** for multiple comparisons“ a upozorňuje na fakt, že p-hodnoty berou v potaz, kolik průměrů porovnáváme. Zohledňují to i intervaly spolehlivosti. Spolehlivost 95 % se vztahuje na všechny intervaly současně!

✧ Kdyby měla některá dvojice skupin populační průměry odlišné, poznaly bychom to tak, že příslušný interval spolehlivosti by neobsahoval 0 a p-hodnota by byla menší než 0.05.

✧ Intervaly spolehlivosti lze přehledně znázornit v obrázku, kde je ihned vidět, které z nich neobsahují nulu, a tudíž které z nich způsobily zamítnutí nulové hypotézy:

```
plot(TukeyHSD(mod01))
```

Zde všechny intervaly nulu obsahují, což je v souladu s tím, že nulová hypotéza $H_0 : \mu_1 = \mu_2 = \mu_3$ nebyla zamítnuta.

7) Normální rozdělení lze nyní ověřit „najednou“ pomocí *standardizovaných reziduí* z odhadnutého „modelu“ (viz text *ANV*, str. 4, dole):

```
shapiro.test(rstandard(mod01))
```

P-hodnota vyšla 0.7055, což je více než hladina 0.05, a tudíž normální rozdělení hmotnosti v jednotlivých skupinách nezamítáme.

8) ANOVA je v podstatě dvouvýběrový t-test rozšířený na srovnání více jak dvou skupin. Obdobně jako u t-testu lze vynechat předpoklad shodnosti rozptylů v jednotlivých skupinách a použít příkaz:

```
oneway.test(hmotnost ~ Vzdelani)
```

✧ Výstup této funkce je bohužel trochu chudší než u funkce `aov` a neobsahuje kompletní tabulku analýzy rozptylu.

```
        One-way analysis of means (not assuming equal variances)

data:  hmotnost and Vzdelani
F = 0.98266, num df = 2.000, denom df = 48.235, p-value = 0.3817
```

hodnota testové statistiky F — první stupně volnosti (odpovídají čitateli, proto "numerator degrees of freedom") — druhé stupně volnosti (odpovídají jmenovateli, proto "denominator degrees of freedom") — p-hodnota

✧ P-hodnota testu hypotézy, že populační průměry jsou ve všech skupinách stejné, vyšla 0.3817, což je podobná hodnota jako u ANOVY s předpokladem shodných rozptylů. Náš závěr by byl tedy stejný.

9) Pojďme si cvičně zkusit rozhodnout o platnosti H_0 pouze na základě testové statistiky a kvantilu F-rozdělení nalezeného v tabulkách (viz text *ANV*, vzorec (4)). Tabulky jsou k dispozici např. na mých stránkách k 7. cvičení.

- hodnota testové statistiky (u ANOVy pro shodné rozptyly) vyšla $F = 0.812$
- stupně volnosti jsou $f_A = 2$ a $f_e = 96$
- Nyní potřebujeme najít hodnotu 95% kvantilu rozdělení F_{f_A,f_e}, tj. potřebujeme $qF_{2,96}(0.95)$, který bude představovat kritickou hodnotu pro zamítnutí. V tabulkách na mých stránkách jsou kvantily $qF_{k,n}(1-\alpha)$ značené jakožto kritické hodnoty $F_{k,n}(\alpha)$.
- Bohužel v těchto tabulkách není možné zvolit druhý stupeň volnosti $n = 96$. Vezmeme tedy nejbližší ($n = 100$) a dostáváme $F_{2,100}(0.05) = 3.09$. Skutečný kvantil bychom mohli zjistit v R jakožto

  ```
  qf(0.95,2,96)
  ```

  a dozvěděli bychom se, že je to 3.091191.
- Tak jako tak hodnota testové statistiky nepřekročila kritickou hodnotu v podobě kvantilu, a tudíž bychom dospěli ke stejnému závěru - nezamítnout H_0.

3 Závislost výšky otce na vzdělání matky

Pokuste se zjistit, zda výška otce závisí na vzdělání matky. Jestliže zjistíte významnou závislost, jak se projevuje? Hladinu testování zvolte 5 %.

Označme si opět jako Y_1 výšku otce náhodně vybraného dítěte matky se základoškolským vzděláním, jako Y_2 výšku otce náhodně vybraného dítěte matky s maturitou, a jako Y_3 výšku otce náhodně vybraného dítěte matky s vysokoškolským vzděláním. Budeme předpokládat, že $Y_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$, $Y_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$, $Y_3 \sim \mathcal{N}(\mu_3, \sigma_3^2)$, a že platí-li navíc $\sigma_1 = \sigma_2 = \sigma_3$. Tyto předpoklady samozřejmě pro naše data ověříme. Pomocí analýzy rozptylu pak otestujeme hypotézu $\mathrm{H}_0 : \mu_1 = \mu_2 = \mu_3$ proti $\mathrm{H}_1 : \mu_1 \neq \mu_2 \ \vee \ \mu_1 \neq \mu_3 \ \vee \ \mu_2 \neq \mu_3$ (kde $\vee$ znamená „nebo“).

1) Základní popisné statistiky.

```
tapply(vyska.o, Vzdelani, summary)
```

2) Popisné obrázky.

```
plot(vyska.o ~ Vzdelani, ylab="Vyska otce (cm)")     # krabicové grafy
plot_of_means(vyska.o, Vzdelani, "ci", prob=0.95)    # graf průměrů
```

3) Leveneův test homoskedasticity.

```
leveneTest(vyska.o ~ Vzdelani)   # nebo:
levene.test(vyska.o, Vzdelani)
```

✧ P-hodnota je 0.8092, tudíž shodu populačních rozptylů nezamítáme. Označme si tuto společnou hodnotu populačního rozptylu jako σ^2 (předpokládáme tedy $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 =: \sigma^2$).

4) ANOVA (při předpokladu shodných rozptylů).

```
mod02 <- aov(vyska.o ~ Vzdelani)   # vytvoření modelu
summary(mod02)                     # tabulka analýzy rozptylu
```

✧ P-hodnota vyšla 0.0137, což je menší než zvolená hladina $\alpha = 0.05$. Zamítáme tedy nulovou hypotézu. Na 5% hladině jsme prokázali, že populační průměry výšky otců se liší pro různá vzdělání matky. Mezi některými skupinami je statisticky významný rozdíl.

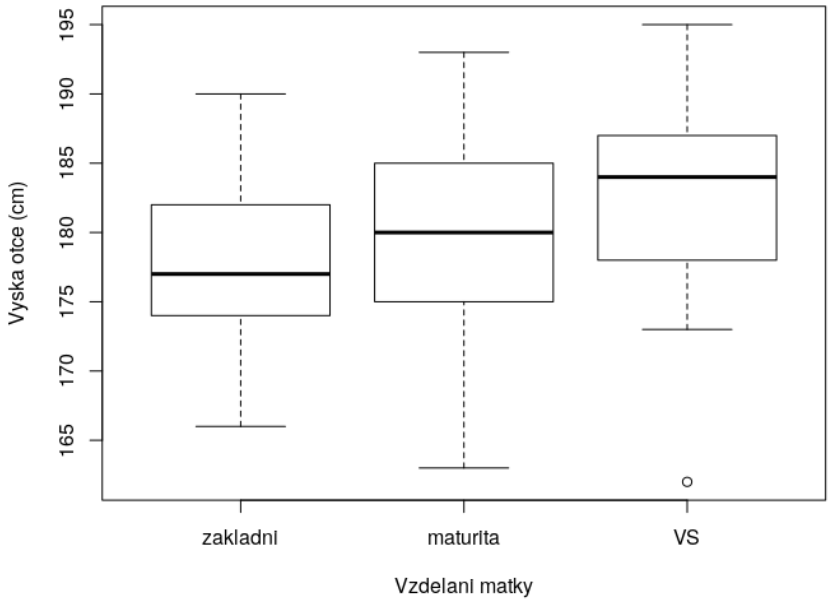
✧ Odhaden společného rozptylu výšky otců σ^2 je reziduální rozptyl $s^2 = \frac{S_e}{f_e} = 42.97$.

5) Dále by nás zajímalo, které dvě skupiny se liší natolik, že došlo k zamítnutí. Mnohonásobné porovnání provedeme opět pomocí Tukeyho metody.

```
TukeyHSD(mod02)
plot(TukeyHSD(mod02))
```

✧ Z p-hodnot (a také z intervalů spolehlivosti) vidíme, že významně se liší skupiny „VŠ“ a „základní“. Příslušný interval spolehlivosti pro $\mu_3 - \mu_1$ totiž neobsahuje nulu a také p-hodnota testu $H_0 : \mu_3 = \mu_1$ je menší než 0.05.

✧ Už z pohledu na krabicové grafy je patrné, že skupina „VŠ“ a „základní“ mají mezi sebou největší rozdíl ve výškách otců. Dvojice „VŠ“-„maturita“ nebo „maturita“-„základní“ se samozřejmě taky trochu liší, ale ne natolik, aby to vyšlo statisticky významné.



6) Test normality se standardizovanými rezidui z ANOVy.

```
shapiro.test(rstandard(mod02))
```

✧ Je to dobré, p-hodnota vyšla 0.346, tedy větší než hladina testu 0.05, takže normalitu rozdělení výšek otců můžeme ve všech třech skupinách předpokládat. Použití standardní analýzy rozptylu tedy bylo oprávněné.

4 Závislost věku otce na vzdělání matky

Pokuste se zjistit, zda věk otce závisí na vzdělání matky. Hladinu testu uvažujme opět 5 %.

Nejprve si opět zavedeme náš pravděpodobnostní model, abychom slovní zadání převedli do řeči matematiky.

Označme si opět jako Y_1 věk otce náhodně vybraného dítěte matky se základoškolským vzděláním, jako Y_2 věk otce náhodně vybraného dítěte matky s maturitou, a jako Y_3 věk otce náhodně vybraného dítěte matky s vysokoškolským vzděláním. Chceme otestovat hypotézu $\mathrm{H}_0 : \mu_1 = \mu_2 = \mu_3$

proti $H_1 : \mu_1 \neq \mu_2 \ \vee \ \mu_1 \neq \mu_3 \ \vee \ \mu_2 \neq \mu_3$ (kde $\vee$ znamená „nebo“). Pokud bychom mohli předpokládat, že $Y_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$, $Y_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$, $Y_3 \sim \mathcal{N}(\mu_3, \sigma_3^2)$, mohli bychom použít analýzu rozptylu.

1) Základní popisné statistiky.

```
tapply(vek.o, Vzdelani, summary)
```

2) Popisné obrázky.

```
plot(vek.o ~ Vzdelani, ylab="Vek otce")
plot_of_means(vek.o, Vzdelani, "ci", prob=0.95)  # graf průměrů s 95% intervaly
                                                 # spolehlivosti
```

3) Leveneův test homoskedasticity.

```
leveneTest(vek.o ~ Vzdelani)
```

✧ P-hodnota vyšla 0.1077, tedy nezamítáme shodu rozptylů věku otců v jednotlivých skupinách.

4) ANOVA (při předpokladu shodných rozptylů).

```
mod03 <- aov(vek.o ~ Vzdelani)
summary(mod03)
```

5) Test normality se standardizovanými rezidui z ANOVy.

```
shapiro.test(rstandard(mod03))
```

P-hodnota Shapiro.Wilkova testu vyšla 0.001746. Zjišťujeme tedy, že rozdělení reziduí v ANOVA modelu je významně nenormální. V tom případě byla celá předchozí analýza zbytečná a můžeme ji zahodit. Budeme se muset uchýlit k nějakému neparametrickému (pořadovému) testu.

V případě dvouvýběrového t-testu jsme alternativně používali neparametrický Wilcoxonův (Mannův-Whitneyův) test. V případě srovnání více jak dvou skupin je neparametrickou obdobou ANOVy Kruskalův-Wallisův test.

Otestujme tedy pomocí Kruskalova-Wallisova testu významnost závislosti věku otce na vzdělání matky. Musíme si ale přeformulovat tvar hypotéz, aby odpovídal tomuto testu:

H_0 : rozdělení věku otců je ve všech skupinách vzdělání matky shodné (tj. jsou shodné i střední hodnoty)
H_1 : rozdělení věku otců se liší pro různé skupiny vzdělání matky

✧ v R provedeme Kruskalův-Wallisův test pomocí příkazu

```
kruskal.test(vek.o ~ Vzdelani)
```

```
        Kruskal-Wallis rank sum test

data:  vek.o by Vzdelani
Kruskal-Wallis chi-squared = 11.888, df = 2, p-value = 0.002621
```

p-hodnota

hodnota testové statistiky (v textu ANV vzorec (7))

stupně volnosti (počet skupin mínus 1)

✧ Jaký je závěr?
Vidíme, že p-hodnota je menší než uvažovaná hladina 0.05. Na hladině 5 % jsme tedy prokázali, že rozdělení věku otců se liší pro různá vzdělání matky.

✧ Zvládli byste o platnosti H_0 rozhodnout opět pouze na základě statistických tabulek (tj. bez použití p-hodnoty)?