Cvičenie 11 –

Test zhody strednej hodnoty a rozptylu

!! Neváhajte použiť google alebo AI na vysvetlenie jednotlivých príkazov a testov (osobitne na zorientovanie sa v dlhých výpisoch výsledkov), prípadne si nechajte poradiť, ak niečo nefunguje podľa očákavaní.

Príklad 1

Psychologický experiment skúma krátkodobú pamäť dvoch skupín účastníkov, líšiacich sa vekom. Testuje sa schopnosť zapamätať si 20 položiek a po 2 minútach rozptýlenia ich správne identifikovať. Skupina A je vek 15-30 rokov, skupina B je 40+, vektory a, b predstavujú počet správne určených položiek u každého z účastníkov.

Testovanou hypotézou bude to, že schopnosti krátkodobej pamäte nie sú závislé od veku, konkrétne že stredné hodnoty vektorov a, b sa nebudú významným spôsobom líšiť. Testovať budeme dvojvýberovým t-testom.

– Zamietnutie tejto hypotézy potvrdí, že odlišné okolnosti (vek) majú vplyv na získané hodnoty vzorky a štatisticky významne menia strednú hodnotu.

– Nezamietnutie hypotézy naopak bude znamenať, že zmenené okolnosti neovplyvňujú charakter vzorky štatisticky významným spôsobom.

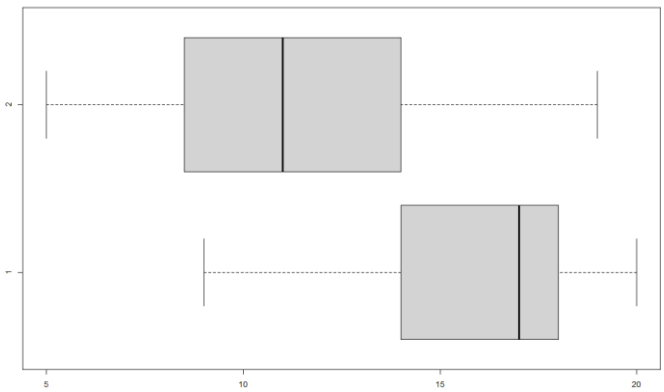
```
a <- c(14, 19, 15, 17, 16, 20, 11, 18, 20, 9, 13, 14, 17, 12, 20, 17, 18)

b <- c(12, 14, 9, 11, 16, 11, 7, 7, 5, 8, 14, 15, 8, 19, 17, 10, 12, 9, 11)
```

Riešenie:

Predbežný rozbor

Na začiatok sa vizuálne zorientujeme v situácii.

```
boxplot(a,b,horizontal = T)
```



Zbežný pohľad naznačuje, že u mladších účastníkov je krátkodobá pamäť lepšia. U starších účastníkov je v priemere horšia, ale „ako u koho“ (väčší rozptyl).

Normálnosť údajov

Overíme si, že údaje vo vektoroch a,b majú normálne rozdelenie (tj. nemáme dôvod na vážne pochybnosti o normálnom rozdelení).

```
shapiro.test(a)
        Shapiro-Wilk normality test        data:  a
        W = 0.94211, p-value = 0.3441

> shapiro.test(b)
        Shapiro-Wilk normality test        data:  b
        W = 0.97375, p-value = 0.8482
```

F-test

Skôr ako začneme testovať zhodu stredných hodnôt, musíme overiť zhodu rozptylov, čo je podmienka korektného použitia t-testu.

H_0: $\sigma^2_a = \sigma^2_b$, tj. $\sigma^2_a/\sigma^2_b = 1$; alfa = 5% // H_1: $\sigma^2_a \neq \sigma^2_b$

Na testovanie zhody disperzií sa používa F-test (George W. Snedecor, Ronald A. Fisher). Test je založený na skutočnosti, že podiel dvoch skúmaných disperzií zo vzoriek veľkostí n, m je náhodná veličina s Fisherovým rozdelením F(n-1, m-1).

https://en.wikipedia.org/wiki/F-distribution
https://en.wikipedia.org/wiki/F-test

V našom prípade $\sigma^2_a/\sigma^2_b \sim F(16, 18)$. Na nezamietnutie hypotézy o zhode disperzií sa musí σ^2_a/σ^2_b nachádzať medzi 0.025 a 0.975 kvantilom rozdelenia F(16,18).

```
var.test(a,b)
        F test to compare two variances,  data:  a and b
        F = 0.77816, num df = 16, denom df = 18, p-value = 0.6191
        alternative hypothesis: true ratio of variances is not equal to 1
        95 percent confidence interval:        0.2947195   2.1142712
        sample estimates: ratio of variances 0.7781629
```

Vysvetlenie:

- $\sigma^2_a/\sigma^2_b = 0.77816$ je “štatistika F”, 16 a 18 sú na-1, nb-1
- interval spoľahlivosti je (0.2947195, 2.1142712), hodnota 0.77816 do neho patrí
  (!! rozdiel oproti doterajším testom, tu je úroveň II, III tá istá)
- p-value je 0.62 >> 0.05, hypotézu H_0 nezamietame.

Nezamietnutie hypotézy H_0 znamená, že rozdiel disperzií σ^2_a , σ^2_b nie je natoľko významný*, aby bránil použitiu t-testu. */ *Prakticky by malo platiť 0.5 < σ^2_a/σ^2_b < 2.*

---

t-test

Dvojvýberový t-test je založený na Studentovom rozdelení. Pre dvojicu vektorov x, y dĺžok n, m má veličina

$$T = \frac{\overline{X} - \overline{Y} - \delta}{\sqrt{(n-1)S_x^2 + (m-1)S_y^2}} \sqrt{\frac{nm(n+m-2)}{n+m}}$$

pri platnosti hypotézy, že rozdiel stredných hodnôt sa rovná 0 alebo všeobecnejšie δ, t-rozdelenie s $n+m-2$ stupňami volnosti.

https://cs.wikipedia.org/wiki/T-test

V našom prípade bude skúmaná hypotéza nasledujúca:

H_0: $\mu_a = \mu_b$, tj. $\mu_a - \mu_b = 0$; alfa = 5% // H_1: $\mu_a \neq \mu_b$

Použijeme systémový nástroj t.test. V argumentoch okrem vektorov a, b uvedieme:

paired=F – nechceme párový test
var.equal = T – overili sme, že disperzie sa môžu považovať za zhodné

```
t.test(a,b,paired=F,var.equal = T)
```

```
Two Sample t-test, data:  a and b
t = 3.8469, df = 34, p-value = 0.0005011
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:  2.154132  6.978995
sample estimates:  mean of x  mean of y
                   15.88235   11.31579
```

Hodnota p-value vychádza výrazne menšia ako alfa = 0.05, preto hypotézu musíme zamietnuť. Túto skutočnosť vidno aj pri pohľade na interval spoľahlivosti (2.154132, 6.978995), v ktorom sa hypotetická 0 nenachádza.

Záver: hypotézu o zhode stredných hodnôt vzoriek a, b zamietame na hladine 5%. Vek účastníkov má teda štatisticky významný vplyv na celkové výsledky, pozorovaná slabšia pamäť u starších účastníkov sa nedá vysvetliť iba ako výsledok náhodných vplyvov.

Príklad 2

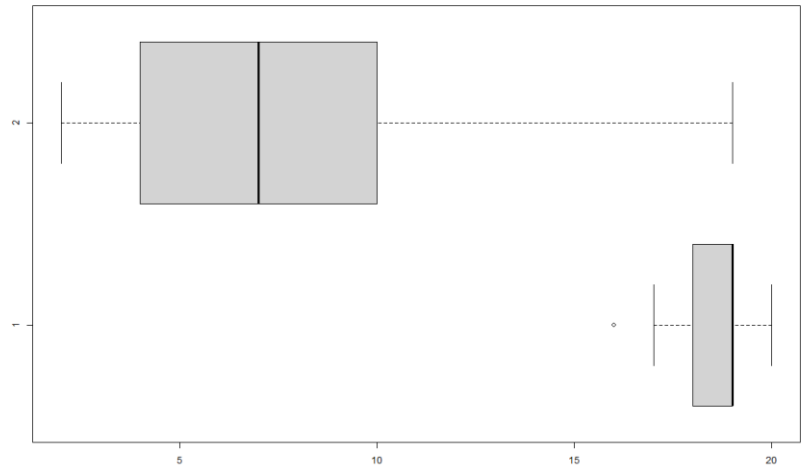
Test krátkodobej pamäte sa zopakoval na ďalších dvoch skupinách účastníkov, ktoré sa okrem veku líšili aj ďalšími okolnosťami (vzdelanie, únava, ...).

Získané vektory sú nasledujúce:

```
p <- c(18, 19, 19, 17, 20, 18, 16, 19, 20, 19)
```

```
q <- c(4, 17, 2, 16, 10, 5, 8, 19, 4, 7, 8, 5, 2)
```

Vizuálna orientácia v situácii – vyzerá to zaujímavo...:



Overíme normálnosť oboch vektorov:

```
shapiro.test(p)
```

W = 0.90321, p-value = 0.2375

```
shapiro.test(q)
```

W = 0.87301, p-value = 0.05739

Ďalej overíme zhodu rozptylov:

```
var.test(p,q)
```

```
F test to compare two variances,      data:  p and q
F = 0.049281, num df = 9, denom df = 12, p-value = 9.486e-05
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:       0.01434321,   0.19062994
sample estimates:     ratio of variances      0.04928105
```

Hodnota p-value je extrémne malá, hypotézu o zhode rozptylov zamietame.
To znamená, že nie sú splnené podmienky pre použitie štandardného t-testu. V takomto prípade je možné použiť Welchov test, pri ktorom sa zhoda rozptylov nevyžaduje.

https://en.wikipedia.org/wiki/Student%27s_t-test

Testujeme tým istým systémovým nástrojom t.test, v ktorom "priznáme" zamietnutú zhodu variancií (var.equal = F):

```
t.test(p,q,paired=F,var.equal = F)
```

```
Welch Two Sample t-test      data:  p and q
t = 6.2777, df = 13.513, p-value = 2.385e-05
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:       6.748845,   13.789617
sample estimates:     mean of x       mean of y
                      18.500000       8.230769
```

Hodnota p-value je takmer nulová. Hypotézu zamietame.
Všetky okolnosti, ktoré odlišujú účastníkov vzoriek p, q, majú štatisticky významný vplyv na odlišnosť stredných hodnôt (nie je to vec náhody).

Príklad 3

Potrebujeme knižnicu (package):

```
library(readxl)
```
– potrebné na čítanie excelovských súborov

Do pracovného adresára si treba stiahnuť súbor data_vyuka.xlsx.
Ten si načítame do tabuľky data, z ktorej následne vynecháme riadky s neúplnými dátami.

```
data <- read_xlsx("data_vyuka.xlsx")
new_data <- na.omit(data)
```

Z hľadiska testovania nás bude zaujímať stĺpec mprij (príjem) a vzdelanie (hodnoty 1,2,3 podľa stupňa), budeme skúmať vplyv vzdelania na príjem respondentov.

```
mprij <- new_data$mprij
vzd <- new_data$vzdelanie
```

Nasleduje overenie podmienok normálnosti a rovnosti disperzií.

Množinu respondentov ankety vieme rozdeliť na tri skupiny podľa stupňa vzdelania. V každej skupine by hodnoty výšky príjmu mali mať normálne rozdelenie. To si overíme klasickým testom, aplikovaným trojmo po skupinách:

```
tapply(mprij,vzd,shapiro.test)
```

```
$`1`    Shapiro-Wilk normality test

data:  X[[i]]
W = 0.9556, p-value = 0.7197

$`2`    Shapiro-Wilk normality test

data:  X[[i]]
W = 0.96792, p-value = 0.6867

$`3`    Shapiro-Wilk normality test

data:  X[[i]]
W = 0.93161, p-value = 0.3975
```

Normálnosť dát nebola spochybnená v žiadnej z troch skupín.

Prejdeme k overeniu zhody rozptylov.

H_0: variancie vo všetkých 3 skupinách sú rovnaké alfa = 5%
H_1: variancia aspoň v jednej skupine sa významne líši od ostatných

Použijeme Bartlettov test. https://en.wikipedia.org/wiki/Bartlett%27s_test

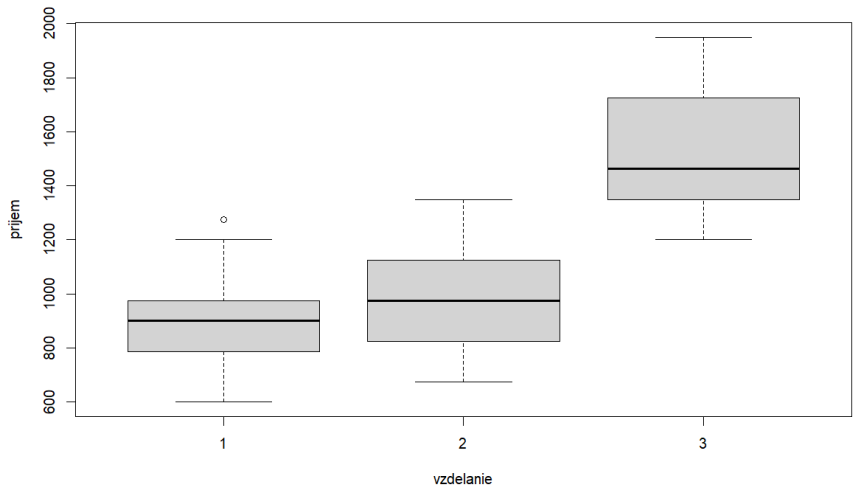
```
bartlett.test(mprij,vzd)
```

Bartlett test of homogeneity of variances data: mprij and vzd
Bartlett's K-squared = 0.89242, df = 2, p-value = 0.6401

Zhoda rozptylov nebola testovaním spochybnená.

Vizualizácia:

```
boxplot(mprij~vzd,xlab="vzdelanie", ylab="prijem")
```



Prejdeme k testovaniu ANOVA (ANalysis Of VAriance).
https://en.wikipedia.org/wiki/Analysis_of_variance

H_0: stredné hodnoty vo všetkých 3 skupinách sú rovnaké; alfa = 5%
H_1: stredná hodnota aspoň v jednej skupine sa významne líši od ostatných

Použijeme príkaz aov, ktorý si vyžaduje faktorizovať vektor vzd, tj. jeho hodnoty pretlmočiť ako úrovne/skupiny 1,2,3.

```
vzdf <- factor(vzd)

anova <- aov(mprij~vzdf)
summary(anova)
```

	Df	*Sum Sq*	*Mean Sq*	*F value*	*Pr(>F)*
vzdf	*2*	*2835237*	*1417618*	*33.72*	*1.84e-09 ******
Residuals	*42*	*1765513*	*42036*		

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

Hodnota p-value v stĺpci Pr(>F) je takmer nulová, hypotézu zamietame. Vzdelanie preto **<u>má</u>** štatisticky významný vplyv na príjem. Toto konštatovanie rozoberieme na drobné.

Veľkosť vplyvu vzdelania na príjem vyjadríme číselne na škále 0 až 1.

```
library(effectsize)                    !! treba nainštalovať
eta_squared(anova)
```

```
Parameter | Eta2 |      95% CI
------------------------------
vzdf      | 0.62 | [0.45, 1.00]
```

Sila vplyvu vzdelania na príjem je 0.62, štatisticky presnejšie povedané, na 95% je v intervale 0.45 až 1.
Rozoberme si to ešte po dvojiciach tried:

```
TukeyHSD(anova)
```

```
Tukey multiple comparisons of means
95% family-wise confidence level
Fit: aov(formula = mprij ~ vzdf)
$vzdf
        diff        lwr        upr        p adj
2-1 90.17857   -90.07478  270.4319   0.4508302
3-1 618.75000  415.39679  822.1032   0.0000000
3-2 528.57143  348.31808  708.8248   0.0000000
```

```
plot(TukeyHSD(anova))
```

Dôležitý je posledný stĺpec (p-value). Hovorí o tom, že štatisticky významné rozdiely sú medzi dvojicami 1-3 a 2-3. Skupiny 1 a 2 nie sú navzájom až tak vzdialené.

Podobnú odpoveď dáva Scheffeho test.

```
library(DescTools)
ScheffeTest(anova)
plot(ScheffeTest(anova))
```

Poznámka:

Ak sa v Bartlettovom teste ukáže významná nezhoda rozptylov, dá sa pokračovať Welchovou modifikáciou testu:

```
anova <- oneway.test(DátovýVektor~FaktorováMnožina, var.equal = F)
print(anova)
```

Príklad 4

V sérii experimentov sa merala utajovaná fyzikálna veličina. Meranie vykonávali traja laboranti (a,b,c) na štyroch prístrojoch (A,B,C,D). Kladieme si otázku (stane sa predmetom testovania), či výsledky meraní sú závislé od toho, kto meria a na čom sa meria.

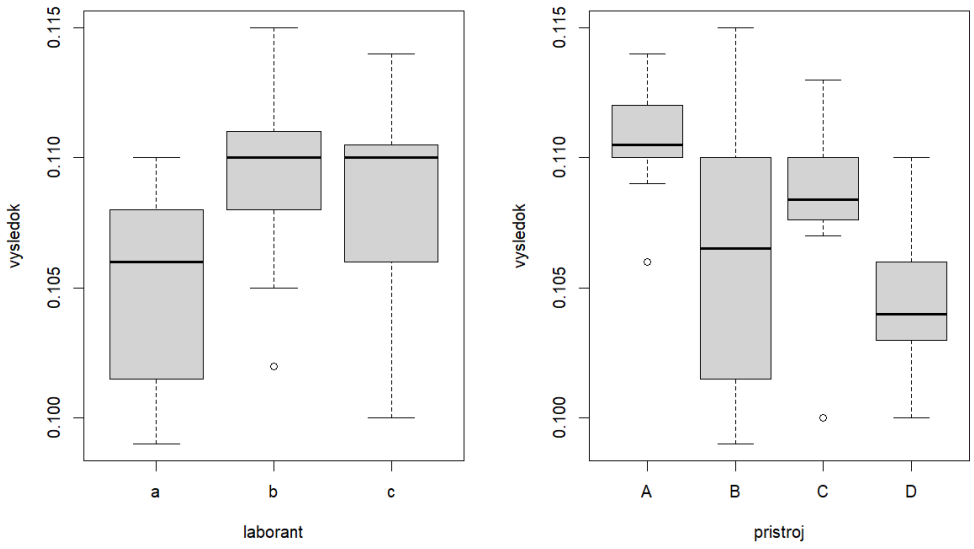
Výsledky meraní sú uložené v súbore anova1 (treba si stiahnuť do pracovného adresára). Načítame stiahnutý súbor a preventívne (kvôli príkazu aov) hneď urobíme faktorizáciu:

```
anova1 <- read_xlsx("anova1.xlsx")

laborant <- factor(anova1$laborant)
pristroj <- factor(anova1$pristroj)
```

Vizuálne sa zorientujeme:

```
par(mfrow=c(1,2))
```
– rozdelí plochu na dve časti, aby sa vedľa seba zmestili 2 grafy

```
boxplot(anova1$deg~anova1$laborant,xlab="laborant",ylab="vysledok")
boxplot(anova1$deg~anova1$pristroj,xlab="pristroj",ylab="vysledok")
```



Podľa obrázkov vidíme, že laborant **a** „vytŕča“ z davu, podobne v prípade pristrojov vidíme výrazné rozdiely vo výsledkoch.

Pred samotným testovaním musíme overiť, či dáta spĺňajú potrebné podmienky.

Budeme testovať normalitu výsledkov v 12 podskupinách:

```
tapply(anova1$deg, list(anova1$laborant, anova1$pristroj), function(x) shapiro.test(x)$p.value )
```

	A	B	C	D
a	0.4632629	0.3689993	0.5130401	0.6368868
b	0.2724532	0.8776501	0.9718771	0.7262263
c	0.5835206	0.6368868	0.4243509	0.6368868

Je to v poriadku.

Overme aspoň zbežne zhodu disperzií.

```
bartlett.test(anova1$deg,anova1$laborant)
```

```
        data:  anova1$deg and anova1$laborant
        Bartlett's K-squared = 0.65843, df = 2, p-value = 0.7195
```

```
> bartlett.test(anova1$deg,anova1$pristroj)
```

```
        data:  anova1$deg and anova1$pristroj
        Bartlett's K-squared = 7.7151, df = 3, p-value = 0.05228
```

Je to dosť na tesno... ale teda ok.

Prejdeme konečne k testu zhody stredných hodnôt.

```
an1 <- aov(anova1$deg~laborant+pristroj)
summary(an1)
```

```
                Df     Sum Sq       Mean Sq     F value  Pr(>F)
laborant         2  0.0001496     7.481e-05       8.227  0.001179 **
pristroj         3  0.0001972     6.574e-05       7.229  0.000672 ***
Residuals   35 0.0003183 9.090e-06
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Hodnoty p-value sú malé, hypotézu o zhode stredných hodnôt naprieč skupinami zamietame. To znamená, že laborant aj prístroj sú štatisticky významné faktory ovplyvňujúce výsledky.

Skúmajme ešte interakcie medzi faktormi laborant a prístroj (vyjadrené znakom násobenia)

```
an2 <- aov(anova1$deg~laborant*pristroj)
> summary(an2)
```

```
                Df     Sum Sq       Mean Sq    F  value  Pr(>F)
laborant         2  1.496e-04     7.481e-05       9.574  0.000642 ***
pristroj         3  1.972e-04     6.574e-05       8.413  0.000355 ***
laborant:pristroj  6  9.168e-05     1.528e-05       1.956  0.105173
Residuals       29 2.266e-04 7.810e-06
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’
```

Hypotézu o neinterakcii oboch faktorov nezamietame. Tj. nemáme presvedčivé argumenty o tom, že by spolu významným spôsobom interagovali – tj. faktor laborant a faktor prístroj sú nezávislé. Preto ďalej s an2 už nebudeme pracovať.

Post-testy

Sila vplyvu faktorov na výsledky:

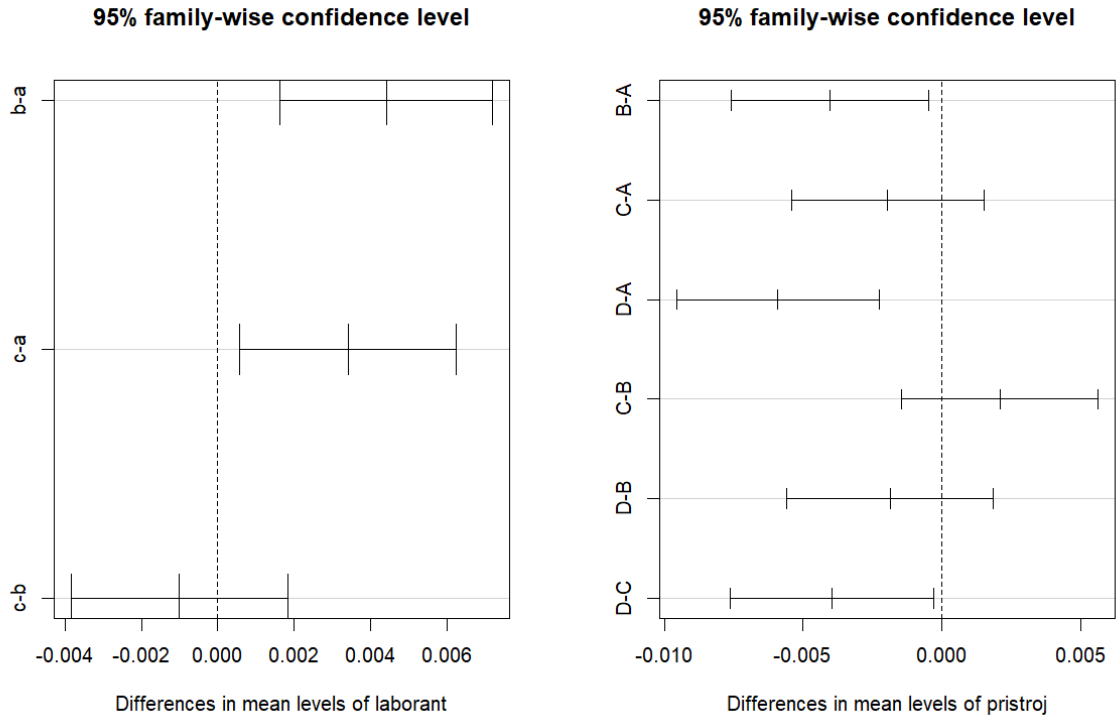
```
eta_squared(an1)
```

```
# Effect Size for ANOVA (Type I)
Parameter | Eta2 (partial) |      95% CI
-----------------------------------------
laborant  |           0.32 | [0.10, 1.00]
pristroj  |           0.38 | [0.14, 1.00]

- One-sided CIs: upper bound fixed at [1.00].
```

TukeyHSD(an1)

```
  Tukey multiple comparisons of means
    95% family-wise confidence level
Fit: aov(formula = anova1$deg ~ laborant + pristroj)

$laborant
         diff         lwr        upr     p adj
b-a  0.004421429  0.0016320534 0.007210804 0.0012560
c-a  0.003410440  0.0005679287 0.006252950 0.0157068
c-b -0.001010989 -0.0038534999 0.001831522 0.6622408

$pristroj
         diff         lwr         upr     p adj
B-A -0.004042517 -0.007595980 -0.0004890554 0.0206029
C-A -0.001953596 -0.005421420  0.0015142275 0.4371571
D-A -0.005909135 -0.009564543 -0.0022537280 0.0006058
C-B  0.002088921 -0.001464541  0.0056423832 0.3998229
D-B -0.001866618 -0.005603367  0.0018701312 0.5400054
D-C -0.003955539 -0.007610946 -0.0003001316 0.0297302
```

plot(TukeyHSD(an1))



Vyhodnotenie určite zvládne každý aj samostatne.

Príklad 5

Testovanie rýchlosti e-shopu. Úlohou je zistiť, ako sa spracovanie užívateľskej požiadavky (merané ako **čas** do dokončenia úlohy v sekundách) mení v závislosti od dvoch kľúčových faktorov:

Faktor A:	Latencia siete	–	Nízka (bežné kancelárske pripojenie, Low Latency) Vysoká (mobilné pripojenie v špičke, High Latency)
Faktor B:	Zložitosť UI	–	Jednoduché UI (intuitívne, moderné rozhranie) Zložité UI (zastaralé, preplnené rozhranie)

Overovanie normálnosti a homogenity variancií ponechávame na samostatnú prácu. Pracujeme ďalej s vedomím, že niekto to už overil a je to v poriadku.

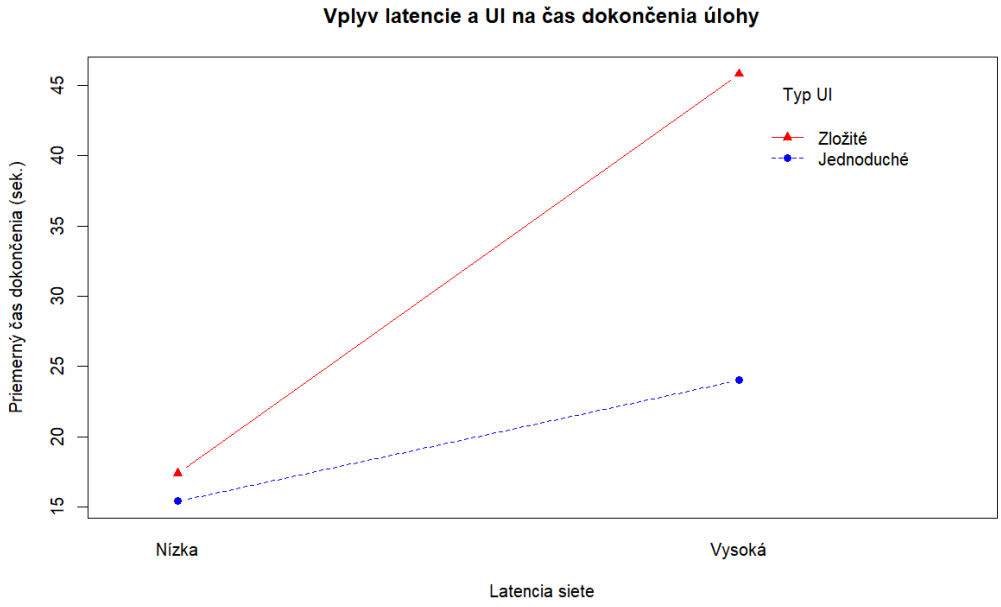
```
anova3 <- read_xlsx("anova_interakcie.xlsx")
an3 <- aov(cas~latencia*uiz,data=anova3)
summary(an3)
```

```
             Df  Sum Sq Mean Sq F value   Pr(>F)
latencia      1    3426    3426  226.94  < 2e-16 ***
uiz           1    1413    1413   93.58 1.50e-11 ***
latencia:uiz  1     983     983   65.09 1.38e-09 ***
Residuals    36     543      15
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Pri latencii a uiz vidno, že tieto faktory významne ovplyvňujú výsledné časy. Nejdeme to skúmať podrobnejšie (robí sa to analogicky ako v predošlom príklade).
Pri interakcii (latencia:uiz) je p-value prakticky nulová, hypotézu o neinterakcii teda zamietame. Oba faktory sa navzájom ovplyvňujú, ukážeme si to vizuálne:

```
par(mfrow=c(1,1))
```
– prepneme “grafickú tabuľu” do pôvodného módu

```
interaction.plot(
    x.factor = anova3$latencia,
    trace.factor = anova3$uiz,
    response = anova3$cas,
    fun = mean,
    type = "b",
    col = c("blue", "red"), # Farby pre lepšiu vizualizáciu
    pch = c(19, 17), # Typy bodov
    trace.label = "Typ UI",
    xlab = "Latencia siete",
    ylab = "Priemerný čas dokončenia (sek.)",
    main = "Vplyv latencie a UI na čas dokončenia úlohy")
```



Čiary nie sú rovnobežné, to znamená že typ UI ovplyvňuje rýchlosť rastu meraného času v závislosti od latencie.