

# Vícerozměrné metody - cvičení



RNDr. Eva Koriťáková, Ph.D.

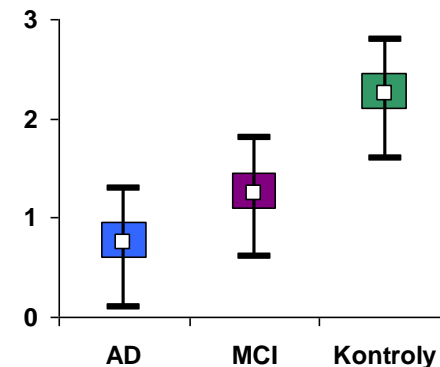
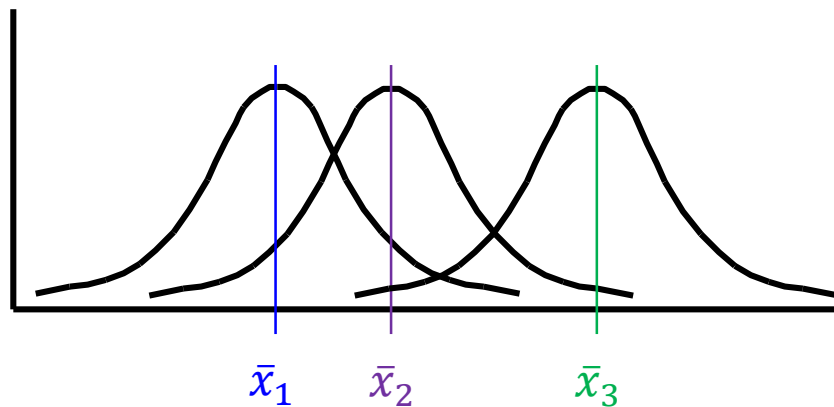
Podzim 2017

# Cvičení 3

## Analýza rozptylu pro vícerozměrná data

# Analýza rozptylu (ANOVA) jednoduchého třídění

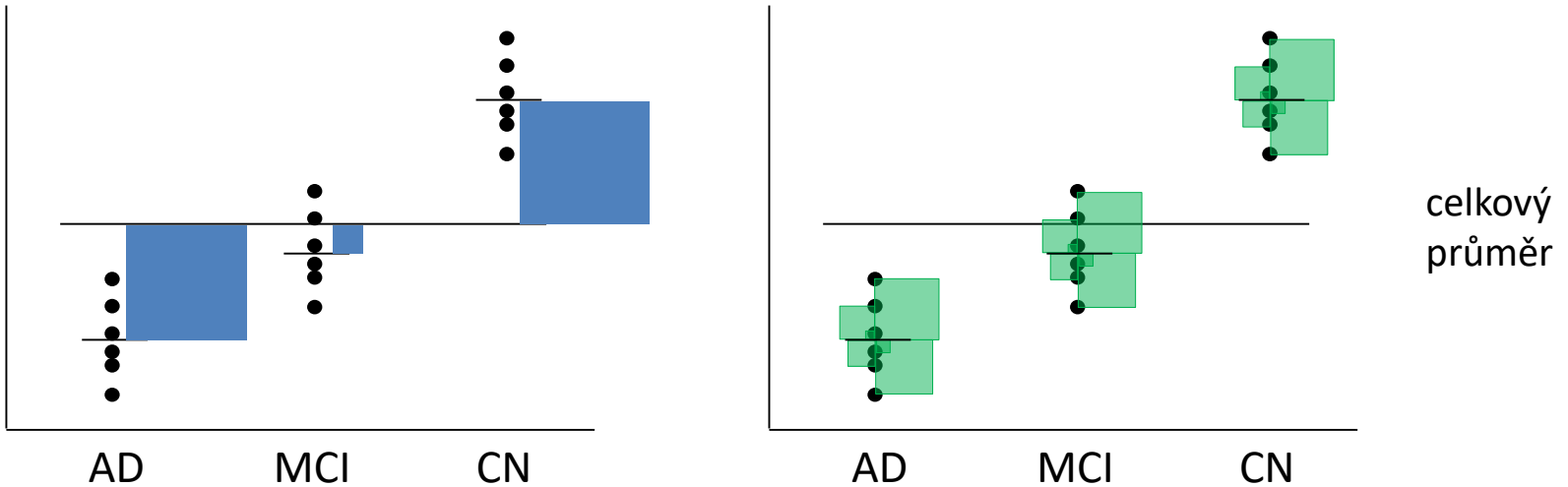
- Srovnáváme tři a více skupin dat, které jsou na sobě nezávislé (mezi objekty neexistuje vazba).
- Příklady: srovnání objemu hipokampu u pacientů s AD, pacientů s MCI a kontrol; srovnání kognitivního výkonu podle čtyř kategorií věku.



- Předpoklady: **normalita dat ve VŠECH skupinách, shodnost (homogenita) rozptylů VŠECH srovnávaných skupin**, nezávislost jednotlivých pozorování.
- Testová statistika: 
$$F = \frac{S_A / df_A}{S_e / df_e}$$

# Analýza rozptylu (ANOVA) – princip

- Srovnání variability (rozptylu) mezi výběry s variabilitou uvnitř výběrů.



- Tabulka analýzy rozptylu jednoduchého třídění (One-Way ANOVA):

Variabilita	Součet čtverců	Počet stupňů volnosti	Průměrný čtverec	F statistika	p-hodnota
Mezi skupinami	$S_A$	$df_A = a - 1$	$MS_A = S_A / df_A$	$F = \frac{S_A / df_A}{S_e / df_e}$	$p$
Uvnitř skupin (reziduální var.)	$S_e$	$df_e = n - a$	$MS_e = S_e / df_e$		
Celkem	$S_T$	$df_T = n - 1$			

# Analýza rozptylu jako lineární model

- Analýza rozptylu pro jednu vysvětlující proměnnou (jednoduché třídění) lze zapsat jako lineární model:

$$Y_{ij} = \mu_i + e_{ij} = \mu + \alpha_i + e_{ij}$$

Populační průměr       $\alpha_i$        $e_{ij}$

Reziduum  
 $i$ -tý efekt faktoru A

- Nulovou hypotézu pak lze vyjádřit jako:  $H_0 : \alpha_1 = \alpha_2 = \dots = \alpha_k$
- Rozšířením tohoto zápisu můžeme definovat další modely ANOVA:** více faktorů, hodnocení interakcí, opakovaná měření na jednom subjektu.

# Analýza rozptylu pro vícerozměrná data

- podle počtu vysvětlovaných proměnných:
  - 1 vysvětlovaná proměnná – jednorozměrná analýza rozptylu (ANOVA)
  - 2 a více vysvětlovaných proměnných – vícerozměrná analýza rozptylu (MANOVA)
- podle počtu faktorů:
  - 1 faktor – ANOVA jednoduchého třídění (jednofaktorová ANOVA)
  - 2 faktory – ANOVA dvojného třídění (dvoufaktorová ANOVA)
  - ...
- podle toho, zda se faktory ovlivňují či nikoliv:
  - faktory se mohou ovlivňovat – model s interakcí
  - faktory se neovlivňují – model bez interakce

# Analýza rozptylu pro vícerozměrná data - příklady

**Počet proměnných:** jednorozměrná x vícerozměrná analýza rozptylu

**Počet faktorů:** jednoduché x dvojné x trojné, ... třídění

**Faktory se ovlivňují či neovlivňují:** s interakcí x bez interakce

- zkoumáme dlouhodobý vliv třech typů léků na hodnoty systolického tlaku u stovky osob – **jednorozměrná analýza rozptylu jednoduchého třídění**
- zkoumáme dlouhodobý vliv třech typů léků na hodnoty systolického tlaku u stovky osob, přičemž chceme zkoumat i vliv pohlaví, předpokládáme však, že ženy i muži reagují na jednotlivé léky obdobně (tzn. např. ženy s léky A a C budou mít nižší tlak než ženy s lékem B a muži s léky A a C budou mít také nižší tlak než muži s lékem B apod.) – **jednorozměrná analýza rozptylu dvojného třídění bez interakce**
- zkoumáme dlouhodobý vliv třech typů léků na hodnoty systolického tlaku u stovky osob, přičemž chceme zkoumat i vliv pohlaví, a předpokládáme, že ženy a muži budou reagovat na léky různě (tzn. např. ženy s léky A a C budou mít nižší tlak než ženy s lékem B, zatímco muži s léky A a B budou mít vyšší tlak než muži s lékem C apod.) – **jednorozměrná analýza rozptylu dvojného třídění s interakcí**
- zkoumáme dlouhodobý vliv třech typů léků na hodnoty systolického a diastolického tlaku u stovky osob – **vícerozměrná analýza rozptylu jednoduchého třídění**
- zkoumáme dlouhodobý vliv třech typů léků a vliv pohlaví na hodnoty systolického a diastolického tlaku u stovky osob – **vícerozměrná analýza rozptylu dvojného třídění**

# Analýza rozptylu dvojného třídění (bez interakce)

- Uvažujeme dvě vysvětlující proměnné zároveň.
- Zápis modelu:

$$Y_{ij} = \mu + \alpha_i + \beta_j + e_{ij}$$

Diagrammatic explanation of the model components:
 

- $\mu$ : Populační průměr (Population mean)
- $\alpha_i$ :  $i$ -tý efekt faktoru A
- $\beta_j$ :  $j$ -tý efekt faktoru B
- $e_{ij}$ : Reziduum (Residual)

- Nulové hypotézy pak máme dvě:  $H_{01} : \alpha_1 = \alpha_2 = \dots = \alpha_k$  ,  $H_{02} : \beta_1 = \beta_2 = \dots = \beta_r$

Variabilita	Součet čtverců	Počet stupňů volnosti	Průměrný čtverec	F statistika	p-hodnota
Faktor A	$S_A$	$df_A = a - 1$	$MS_A = S_A / df_A$	$F_A$	$p$
Faktor B	$S_B$	$df_B = b - 1$	$MS_B = S_B / df_B$	$F_B$	$p$
Rezidua	$S_e$	$df_e = n - a - b + 1$	$MS_e = S_e / df_e$		
Celkem	$S_T$	$df_T = n - 1$			



# Analýza rozptylu dvojného třídění s interakcí

- Uvažujeme dvě vysvětlující proměnné a zároveň i jejich společné působení.

- Zápis modelu:

$$Y_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ij}$$

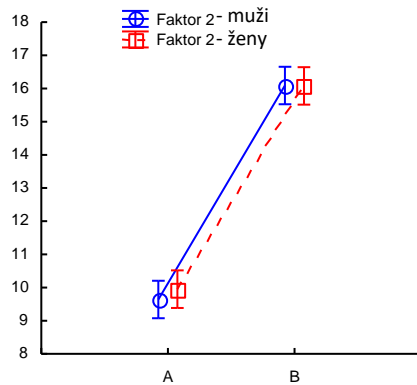
↑ Populační průměr (pointing to  $\mu$ )  
↑  $i$ -tý efekt faktoru A (pointing to  $\alpha_i$ )  
↑  $j$ -tý efekt faktoru B (pointing to  $\beta_j$ )  
↑ Interakce (pointing to  $\gamma_{ij}$ )  
← Reziduum (pointing to  $e_{ij}$ )

- Nulové hypotézy pak máme tři:

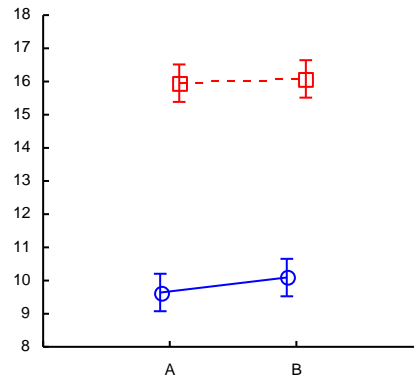
$$H_{01} : \gamma_{11} = \gamma_{12} = \dots = \gamma_{kr} \quad H_{02} : \alpha_1 = \alpha_2 = \dots = \alpha_k \quad H_{03} : \beta_1 = \beta_2 = \dots = \beta_r$$

Variabilita	Součet čtverců	Počet stupňů volnosti	Průměrný čtverec	F statistika	$p$ -hodnota
Faktor A	$S_A$	$df_A = a - 1$	$MS_A = S_A / df_A$	$F_A$	$p$
Faktor B	$S_B$	$df_B = b - 1$	$MS_B = S_B / df_B$	$F_B$	$p$
Interakce A×B	$S_{AB}$	$df_{AB} = (a-1)(b-1)$	$MS_{AB} = S_{AB} / df_{AB}$	$F_{AB}$	$p$
Rezidua	$S_e$	$df_e = n - ab$	$MS_e = S_e / df_e$		
Celkem	$S_T$	$df_T = n - 1$			

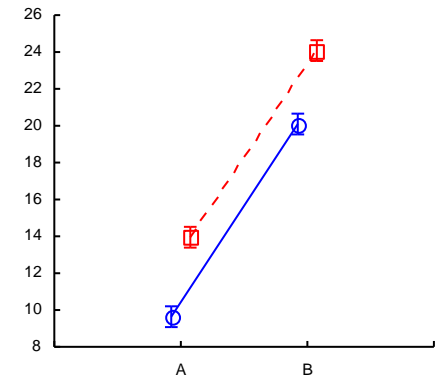
# Hlavní efekty a interakce



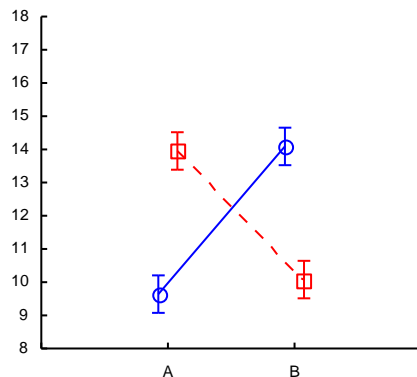
	SS	D.f.	MS	F	p
<b>Faktor 1</b>	<b>1978</b>	<b>1</b>	<b>1978</b>	<b>482.2</b>	<b>0.000</b>
Faktor 2	1	1	1	0.3	0.602
F1*F2	1	1	1	0.3	0.570
Error	804	196	4		



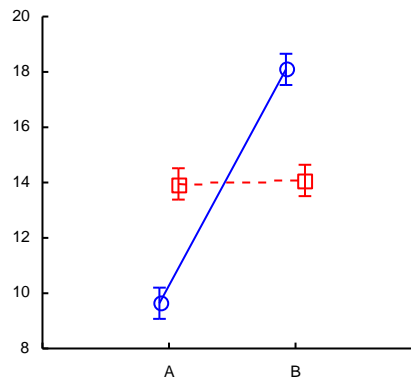
	SS	D.f.	MS	F	p
<b>Faktor 1</b>	<b>4</b>	<b>1</b>	<b>4</b>	<b>1.0</b>	<b>0.314</b>
<b>Faktor 2</b>	<b>1891</b>	<b>1</b>	<b>1891</b>	<b>461.1</b>	<b>0.000</b>
F1*F2	1	1	1	0.3	0.570
Error	804	196	4		



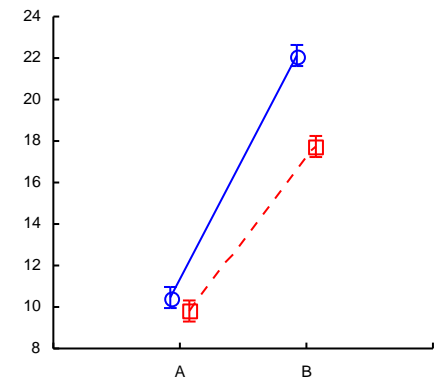
	SS	D.f.	MS	F	p
<b>Faktor 1</b>	<b>5293</b>	<b>1</b>	<b>5293</b>	<b>1290.7</b>	<b>0.000</b>
<b>Faktor 2</b>	<b>861</b>	<b>1</b>	<b>861</b>	<b>209.9</b>	<b>0.000</b>
F1*F2	1	1	1	0.3	0.570
Error	804	196	4		



	SS	D.f.	MS	F	p
<b>Faktor 1</b>	<b>4</b>	<b>1</b>	<b>4</b>	<b>1.0</b>	<b>0.314</b>
Faktor 2	1	1	1	0.3	0.602
<b>F1*F2</b>	<b>867</b>	<b>1</b>	<b>867</b>	<b>211.3</b>	<b>0.000</b>
Error	804	196	4		

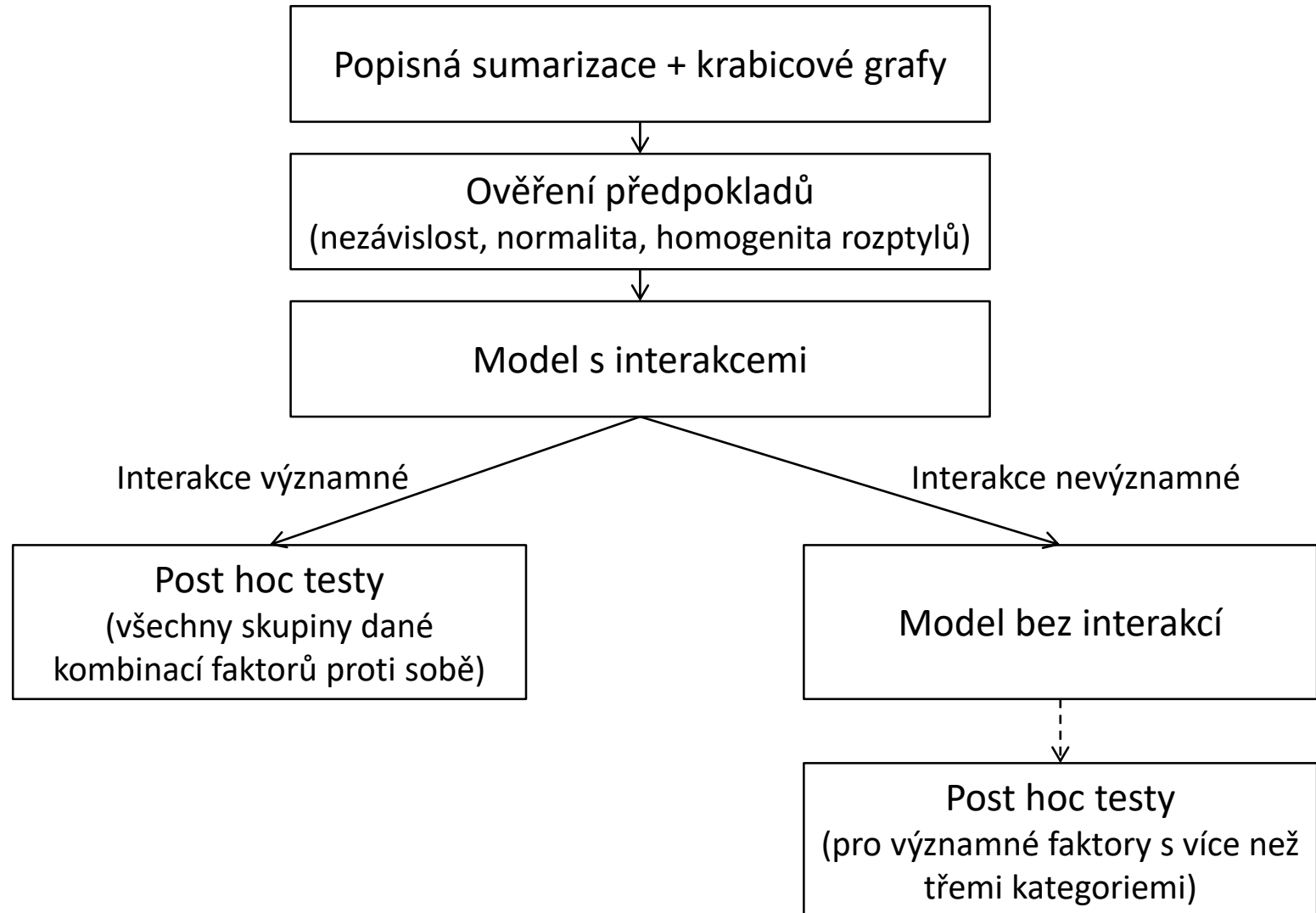


	SS	D.f.	MS	F	p
<b>Faktor 1</b>	<b>920</b>	<b>1</b>	<b>920</b>	<b>224.3</b>	<b>0.000</b>
Faktor 2	1	1	1	0.3	0.602
<b>F1*F2</b>	<b>867</b>	<b>1</b>	<b>867</b>	<b>211.3</b>	<b>0.000</b>
Error	804	196	4		



	SS	D.f.	MS	F	p
<b>Faktor 1</b>	<b>4799</b>	<b>1</b>	<b>4799</b>	<b>1443.4</b>	<b>0.000</b>
<b>Faktor 2</b>	<b>316</b>	<b>1</b>	<b>316</b>	<b>95.0</b>	<b>0.000</b>
<b>F1*F2</b>	<b>175</b>	<b>1</b>	<b>175</b>	<b>52.5</b>	<b>0.000</b>
Error	652	196	3		

# Analýza rozptylu pro vícerozměrná data - postup



# Úkol 1

Zjistěte, zda má vliv pohlaví a typ léku na počet nežádoucích účinků u pacientů s leukémií (neuvažujeme možnou interakci).

ID	Pohlaví	Typ léku	Počet nežádoucích účinků
P1	M	lék X	1
P2	M	lék Y	1
P3	M	lék Z	6
P4	Z	lék X	3
P5	Z	lék Y	4
P6	Z	lék Z	9

# Úkol 1 - řešení

Zjistěte, zda má vliv pohlaví a typ léku na počet nežádoucích účinků u pacientů s leukémií (neuvažujeme možnou interakci).

Překódování:

Pohlaví	Typ léku	Počet nežádoucích účinků
1	1	1
1	2	1
1	3	6
2	1	3
2	2	4
2	3	9

Legenda:

Pohlaví: 1=M  
2=Z

Typ léku: 1=lék X  
2=lék Y  
3=lék Z

# Úkol 1 - řešení

Pohlaví	Typ léku	Počet než. účinků
1	1	1
1	2	1
1	3	6
2	1	3
2	2	4
2	3	9

# Úkol 1 - řešení

Pohlaví	Typ léku	Počet než. účinků	
1	1	$X_{1..} = 8$ $M_{1..} = 8/3$	1
1	2		1
1	3		6
2	1	$X_{2..} = 16$ $M_{2..} = 16/3$	3
2	2		4
2	3		9

$$a = 2; \quad b = 3; \quad c = 1; \quad n = 6;$$

$$X_{.1.} = 4; \quad M_{.1.} = 4/2 = 2$$

$$X_{.2.} = 5; \quad M_{.2.} = 5/2 = 2,5$$

$$X_{.3.} = 15; \quad M_{.3.} = 15/2 = 7,5$$

$$X_{...} = 24; \quad M_{...} = 24/6 = 4$$

**Součet čtverců pro faktor A (pohlaví):**

počet stupňů volnosti:  $f_A = a - 1 = 1$

$$S_A = bc \sum_{i=1}^a (M_{i..} - M_{...})^2 = 3 \cdot ((8/3 - 4)^2 + (16/3 - 4)^2) = 32/3 = 10,67$$

**Součet čtverců pro faktor B (typ léku):**

počet stupňů volnosti:  $f_B = b - 1 = 2$

$$S_B = ac \sum_{j=1}^b (M_{.j.} - M_{...})^2 = 2 \cdot ((2 - 4)^2 + (2,5 - 4)^2 + (7,5 - 4)^2) = 37$$

**Celkový součet čtverců :**

počet stupňů volnosti:  $f_T = n - 1 = 5$

$$S_T = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^c (X_{ijk} - M_{...})^2 = (1 - 4)^2 + (1 - 4)^2 + \dots + (9 - 4)^2 = 48$$

**Reziduální součet čtverců :**

$$S_E = S_T - S_A - S_B = 0,33$$

počet stupňů volnosti:  $f_E = n - a - b + 1 = 2$

# Úkol 1 - řešení

Tabulka analýzy rozptylu dvojného třídění:

Zdroj variability	Součet čtverců	Stupně volnosti	Podíl S/f	$F = \frac{S/f}{S_E/f_E}$
Faktor A (pohlaví)	$S_A = 10,67$	$f_A = 1$	10,67	63,99
Faktor B (typ léku)	$S_B = 37$	$f_B = 2$	18,5	110,98
Reziduální	$S_E = 0,33$	$f_E = 2$	0,16	-
Celkový	$S_T = 48$	$f_T = 5$	-	-

Srovnání s kvantily:

$F_A = 63,99 > F_{0,95}(1,2) = 18,1 \rightarrow$  pohlaví má vliv na počet nežádoucích účinků

$F_B = 110,98 > F_{0,95}(2,2) = 19 \rightarrow$  typ léku má vliv na počet nežádoucích účinků



# Úkol 1 – řešení v softwaru SPSS

Zjistěte, zda má vliv pohlaví a typ léku na počet nežádoucích účinků u pacientů s leukémií.

Pohlaví	Typ léku	Počet než. účinků
M	lék X	1
M	lék Y	1
M	lék Z	6
Z	lék X	3
Z	lék Y	4
Z	lék Z	9

**V softwaru SPSS:** Analyze – General Linear Model – Univariate – Dependent Variable: spojitá proměnná, Fixed Factor(s): kategoriální proměnné →

- Model – zatrhneme Custom – vybereme Typ:Main effects – do Model přetáhneme A, B (*pokud bychom chtěli model s interakcemi, necháme zatržené Full factorial*) – odškrtneme Include intercept in model – Continue
- Post Hoc – Post hoc Tests for: zvolit kategoriální proměnnou – zatrhneme Tukey's-b – Continue
- Plots: zvolit proměnné do Horizontal Axis a Separate Lines – Add – Continue
- Options... – Homogeneity tests – Continue

*Vykreslení krabicových grafů podle obou proměnných:* Graphs – Legacy Dialogs – Boxplot... – Clustered – Define – zvolit Variable, Category Axis a Define Clusters by – OK

# Úkol 1 – řešení v softwaru R

Zjistěte, zda má vliv pohlaví a typ léku na počet nežádoucích účinků u pacientů s leukémií.

## V softwaru R:

```
data <- data.frame(pohl=c(1,1,1,2,2,2),lek=c(1,2,3,1,2,3),pocet=c(1,1,6,3,4,9))
data
```

```
model_bez_interakce <- aov(data$pocet ~ (as.factor(data$pochl)+as.factor(data$lek)))
summary(model_bez_interakce)
TukeyHSD(model_bez_interakce) # post-hoc test
```

```
# 2. způsob: anova(lm(data$pocet ~ (as.factor(data$pochl)+as.factor(data$lek))))
```

```
model_s_interakci <- aov(data$pocet ~ (as.factor(data$pochl)*as.factor(data$lek)))
summary(model_s_interakci)
```

```
boxplot(data$pocet ~(as.factor(data$pochl)*as.factor(data$lek)))
```

```
library("car") # instalace balíku car pomocí: install.packages("car")
leveneTest(data$pocet ~ (as.factor(data$pochl)*as.factor(data$lek)),center=mean)
```

# Úkol 1 – řešení v softwaru STATISTICA

Zjistěte, zda má vliv pohlaví a typ léku na počet nežádoucích účinků u pacientů s leukémií.

Pohlaví	Typ léku	Počet uzdrav. pacientů
M	lék X	1
M	lék Y	1
M	lék Z	6
Z	lék X	3
Z	lék Y	4
Z	lék Z	9

**V softwaru STATISTICA:** Statistics – ANOVA – Main effects ANOVA – Quick specs dialog – OK – Variables – Dependent variable list: X, Categorical predictors (factors): A, B – OK – All effects.

*Post hoc testy:* More results – Post hoc – zvolit Effect – Tukey HSD (nebo Scheffé)

*Levenův test:* More results – Assumptions – zvolit proměnnou – Levene's test (ANOVA)

*Vykreslení krabicových grafů podle obou proměnných:* Graphs – 2D Graphs – Box Plots... – zvolit spojitou proměnnou jako Dependent variable, zvolit jednu kategoriální proměnnou jako Grouping variable – na listu Categorized u X-Categories zatrhnout On a Layout změnit na Overlaid – pokud chceme spojit mediány či průměry, na záložce Advanced zatrhnout Connect middle points – OK

*Pokud bychom uvažovali model s interakcemi, zvolíme Factorial ANOVA (namísto Main effects A.)*

# Úkol 2

Zjistěte, zda má vliv pohlaví a typ onemocnění na objem hipokampu.

Ukázka datového souboru:

ID	Group_3kat	Gender_rek	Hippocampus_volume (mm3)
101	1	M	6996.1
102	1	F	7187.3
103	1	M	7030.2
331	2	M	6891.6
332	2	M	6332.9
334	2	F	6303.7
737	3	M	6170.8
739	3	F	5984.1
740	3	F	6052.4

Legenda k proměnné Group\_3kat:

1...CN (kontroly)

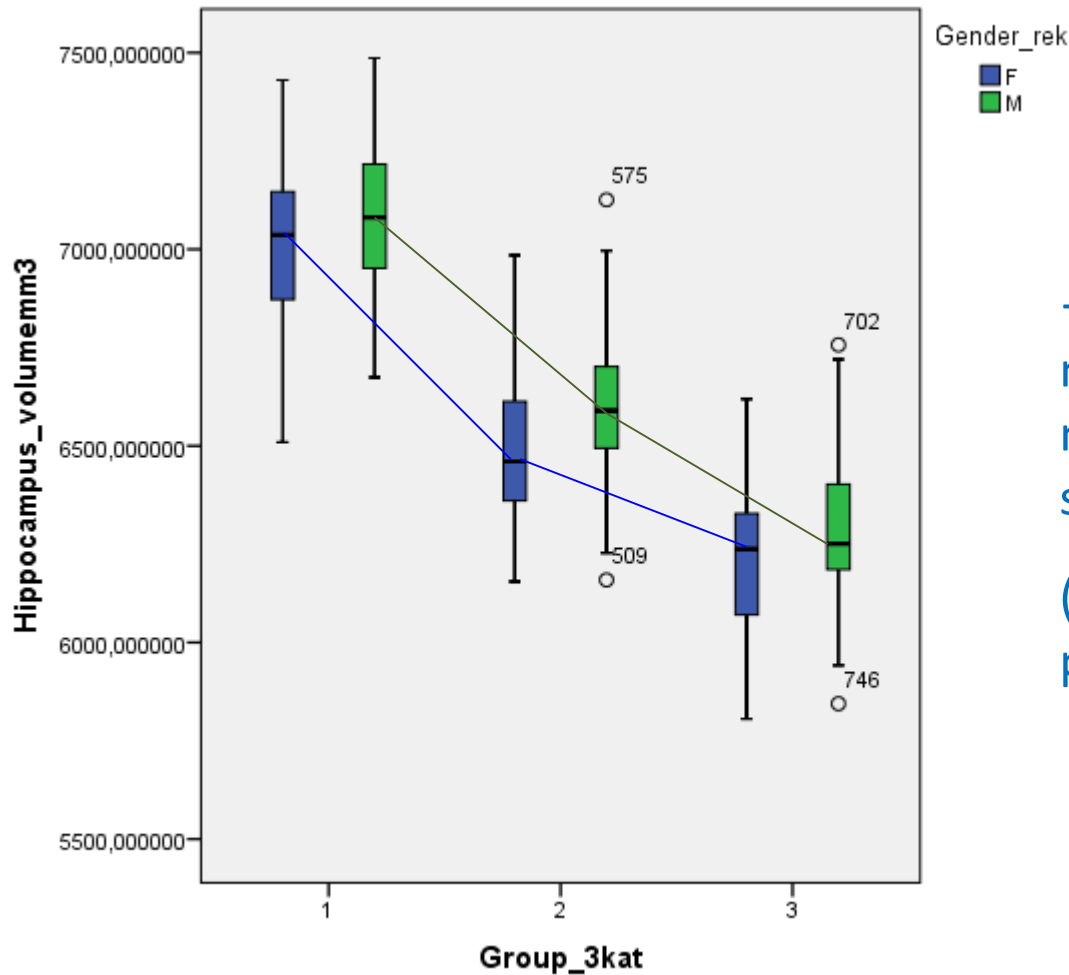
2...MCI (mírná kognitivní porucha)

3...AD (Alzheimerova choroba)

# Úkol 2 – popisná sumarizace dat

Skupina	Pohlaví	N	Průměr	SD	Medián	Minimum	Maximum
CN	F	110	7018.3	190.1	7036.1	6509.6	7430.1
	M	120	7087.3	176.0	7081.1	6674.4	7486.6
	Celkem	230	7054.3	185.7	7048.6	6509.6	7486.6
MCI	F	146	6476.7	171.8	6460.4	6155.1	6984.8
	M	260	6595.2	164.1	6589.5	6159.1	7125.6
	Celkem	406	6552.6	176.2	6555.0	6155.1	7125.6
AD	F	95	6215.0	178.8	6237.8	5805.2	6619.0
	M	102	6293.0	174.8	6250.8	5844.3	6756.9
	Celkem	197	6255.4	180.6	6248.0	5805.2	6756.9
Celkem	F	351	6575.6	364.8	6498.2	5805.2	7430.1
	M	482	6653.8	323.9	6610.0	5844.3	7486.6
	Celkem	833	6620.9	343.7	6580.9	5805.2	7486.6

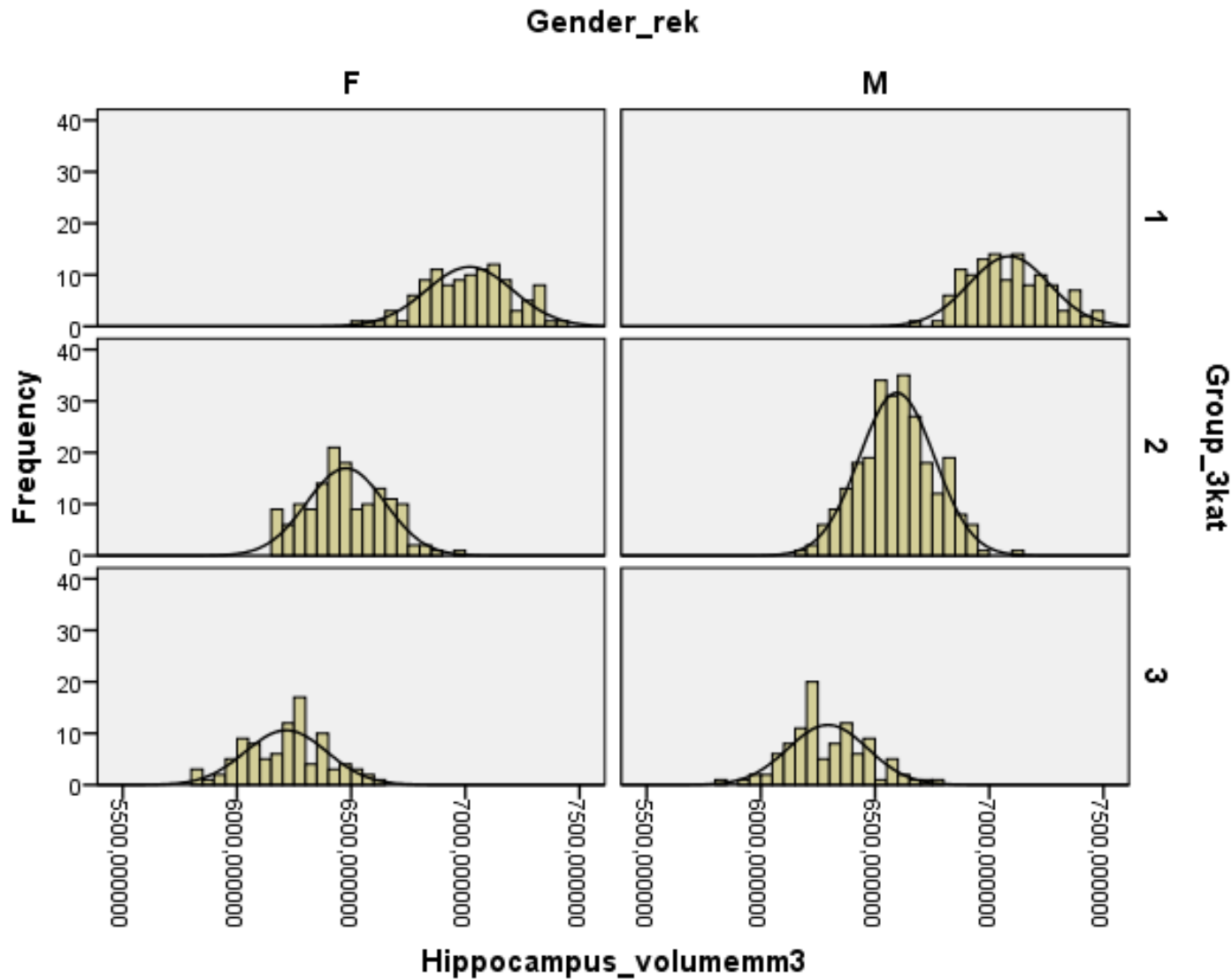
# Úkol 2 – krabicový graf



→ interakci sice očekávat  
nebudeme, přesto si ale  
model s interakcí raději  
spočítáme

(nejdřív ale musíme ověřit  
předpoklady)

# Úkol 2 – ověření normality



# Úkol 2 – homogenita rozptylů a nezávislost

## Homogenita rozptylů:

### Levene's Test of Equality of Error Variances<sup>a</sup>

Dependent Variable: Hippocampus\_volumem

F	df1	df2	Sig.
,962	5	827	,440

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

a. Design: Group\_3kat + Gender\_rek +  
Group\_3kat \* Gender\_rek

$p=0,440 > 0,05 \rightarrow$  nezamítáme homogenitu rozptylů

## Nezávislost:

Protože žádný subjekt nebyl současně ve více skupinách, nezávislost můžeme předpokládat.



# Úkol 2 – model s interakcí

## Tests of Between-Subjects Effects

Dependent Variable: Hippocampus\_volumemm3

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Model	3,659E+10 <sup>a</sup>	6	6098069036	201956,010	,000
Group_3kat	71984656,14	2	35992328,07	1191,995	,000
Gender_rek	1455184,169	1	1455184,169	48,193	,000
Group_3kat * Gender_rek	104654,379	2	52327,189	1,733	,177
Error	24971294,93	827	30195,036		
Total	36613385510	833			

a. R Squared = .999 (Adjusted R Squared = .999)

→ není statisticky významná interakce, proto spočítáme model bez interakce

# Úkol 2 – model bez interakce

## Tests of Between-Subjects Effects

Dependent Variable: Hippocampus\_volumemm3

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Model	3,659E+10 <sup>a</sup>	4	9147077390	302398,408	,000
Group_3kat	71962303,15	2	35981151,58	1189,521	,000
Gender_rek	1781192,205	1	1781192,205	58,885	,000
Error	25075949,31	829	30248,431		
Total	36613385510	833			

a. R Squared = .999 (Adjusted R Squared = .999)

- statisticky významný vliv pohlaví i typu onemocnění na objem hipokampu
- protože typ onemocnění má více než 2 kategorie, musíme provést post-hoc test, abychom zjistili, mezi kterými kategoriemi je statisticky významný rozdíl

# Úkol 2 – interpretace

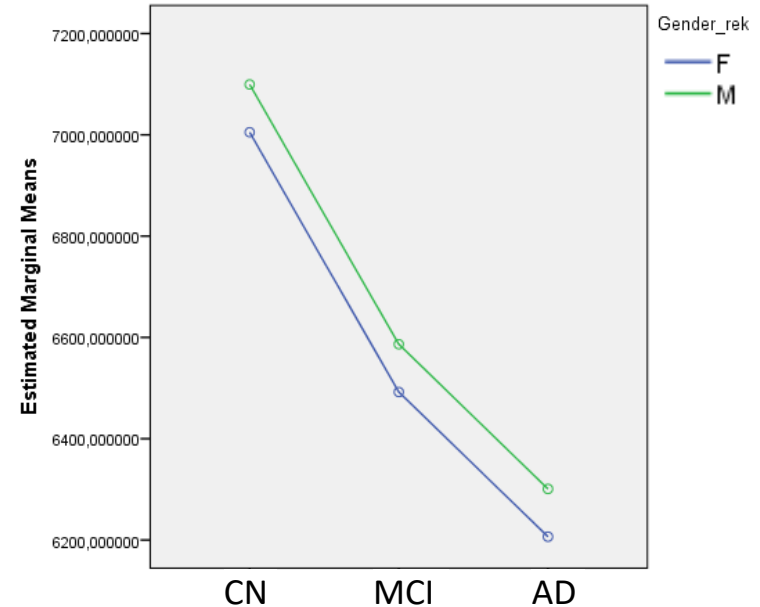
Hippocampus\_volumemm3

Tukey B<sup>a,b,c</sup>

Group 3kat	N	Subset		
		1	2	3
3	197	6255,381784		
2	406		6552,613882	
1	230			7054,334947

Means for groups in homogeneous subsets are displayed.

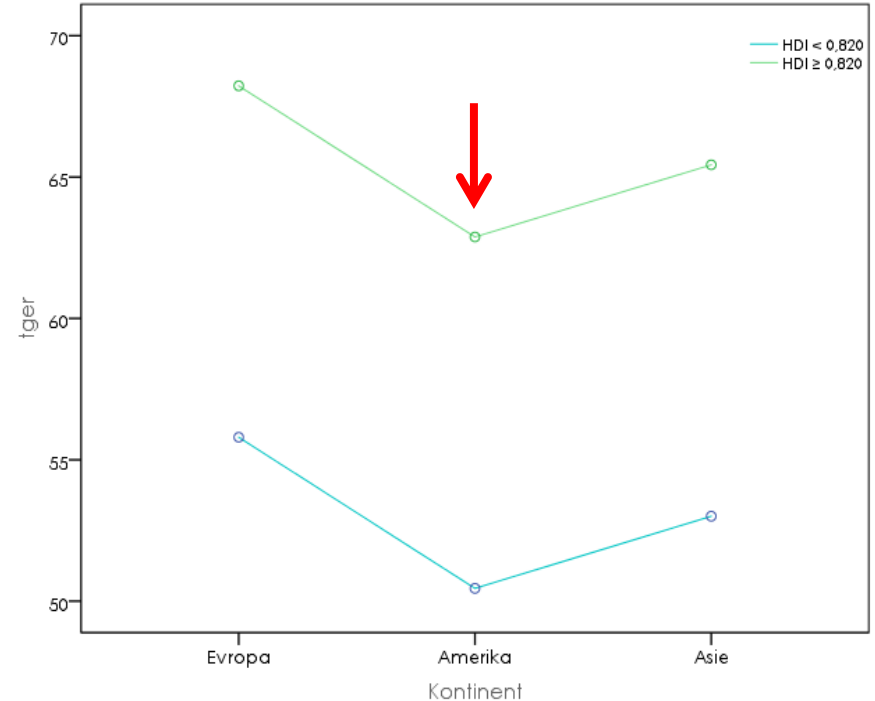
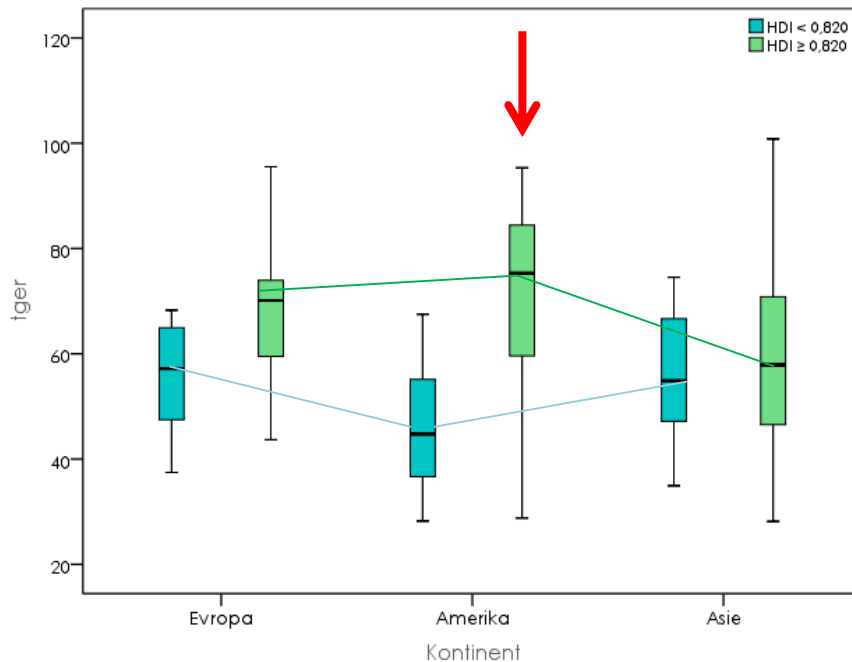
Estimated Marginal Means of Hippocampus\_volumemm3



- statisticky významný vliv pohlaví i typu onemocnění na objem hipokampu, přičemž mezi pohlavím a typem onemocnění nenastává interakce
- u mužů statisticky významně vyšší objem hipokampu než u žen
- statisticky významný rozdíl v objemu hipokampu u všech 3 skupin subjektů podle typu onemocnění, přičemž u pacientů s AD je objem nejmenší a u CN největší

# Upozornění I

Pozor, pokud mediány ukazují úplně jiný „trend“ než průměry!



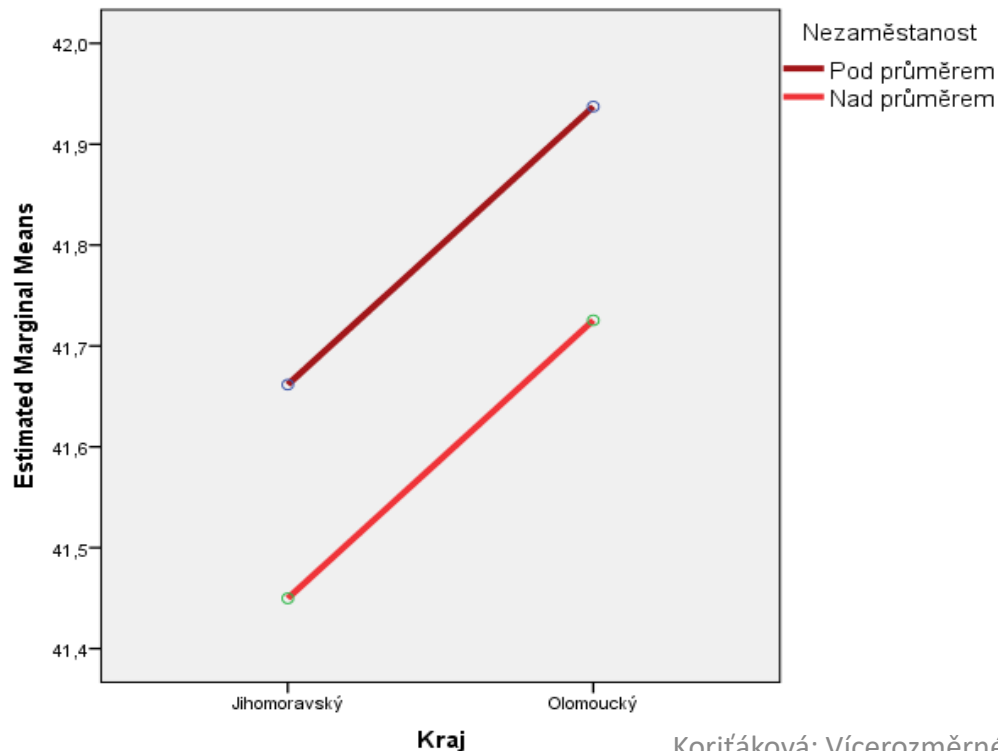
- znamená to, že tam zřejmě není splněn předpoklad normality
- pokud rozdíl není statisticky významný, není zpravidla potřeba to řešit
- pokud by ten rozdíl vyšel statisticky významně, je to problém!
- poznámka: je dobré mít měřítko na ose y stejné u obou grafů

# Upozornění II

## Pozor na interpretaci!

Na první pohled z grafu vypadá, že tam je vliv kraje i nezaměstnanosti, že to nevychází statisticky významně může být:

- malým počtem subjektů ve skupině
- ale i velikostí efektu! (tady efekty malé, průměry ve všech čtyřech skupinách se podle posledního grafu pohybují jen od cca 41,4 do 42!)



# Úkol 3

---

Vyberte (případně vymyslete) si vícerozměrný datový soubor, zvolte statistický software podle svého uvážení a spočítejte analýzu rozptylu. Nezapomeňte ověřit předpoklady! Udělejte i popisnou sumarizaci dat.

Datový soubor a wordovský dokument s přehledně popsány výstupy vložte do odevzdáárny v ISu.