

# Pokročilé metody analýzy dat v neurovědách



RNDr. Eva Koritáková, Ph.D.  
doc. RNDr. Ladislav Dušek, Dr.

# Blok 2

## Vícerozměrné statistické testy a rozložení

# Osnova

---

1. Vícerozměrné charakteristiky
2. Vícerozměrné normální rozdělení
3. Vícerozměrný t-test
4. Vícerozměrná analýza rozptylu
5. Transformace a jiné úpravy vícerozměrných dat

# Vícerozměrné charakteristiky

# Vícerozměrná data

## PROMĚNNÉ

OBJEKTY (SUBJEKTY)

ID	Pohlaví	Věk	Váha	MMSE skóre	Objem hipokampu	...
1	muž	84	85,5	29	7030	
2	žena	25	62,0	28	6984	
3						
4						
...						

Poznámka: proměnné označovány i jako znaky, pozorování, diskriminátory, příznakové proměnné či příznaky

Anglicky označení pouze jedním termínem: feature

# Maticový zápis datového souboru

OBJEKTY (SUBJEKTY)	PROMĚNNÉ					
	ID	Pohlaví	Věk	Váha	MMSE skóre	Objem hipokampu ...
1	muž	84	85,5	29	7030	
2	žena	25	62,0	28	6984	
...						



$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

maticový zápis datového souboru  $n$  objektů (subjektů), které jsou popsány  $p$  proměnnými

jeden prvek matice  $x_{ij}$  je hodnota  $j$ -té proměnné u  $i$ -tého objektu (subjektu), přičemž  $j = 1, \dots, p$  a  $i = 1, \dots, n$

# Vícerozměrný průměr a kovarianční matice

- vícerozměrný průměr (např. pro datový soubor se 2 proměnnými):

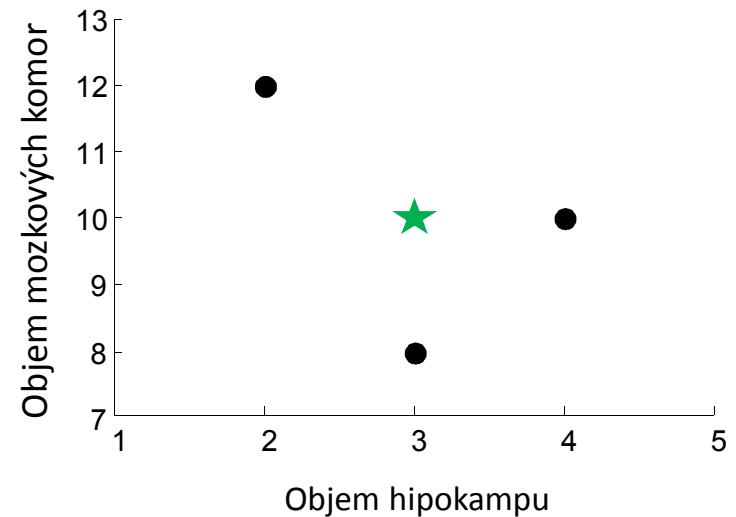
$$\left[ \begin{array}{c} - \\ - \end{array} \right]$$

- výběrová kovarianční matice (např. pro datový soubor se 2 proměnnými):

$$\left[ \begin{array}{c} \\ \end{array} \right], \text{ kde } \text{---} ( \quad )$$

# Vícerozměrný průměr a kovarianční matice

ID	Objem hipokampu	Objem mozkových komor
1	2	12
2	4	10
3	3	8



Vícerozměrný průměr:

$$\left[ \begin{array}{c} - \\ - \end{array} \right] \left[ \begin{array}{c} - ( ) - ( ) \end{array} \right] \left[ \begin{array}{c} - \\ - \end{array} \right]$$

Kovarianční matice:  $\left[ \begin{array}{c} \end{array} \right]$ , kde:

$$\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n} \left( \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \bar{y} - \sum_{i=1}^n \bar{x} y_i + \sum_{i=1}^n \bar{x} \bar{y} \right) = \frac{1}{n} \left( \sum_{i=1}^n x_i y_i - \bar{y} \sum_{i=1}^n x_i - \bar{x} \sum_{i=1}^n y_i + n \bar{x} \bar{y} \right)$$

$$\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x}) = \frac{1}{n} \left( \sum_{i=1}^n x_i^2 - \sum_{i=1}^n x_i \bar{x} - \sum_{i=1}^n \bar{x} x_i + \sum_{i=1}^n \bar{x} \bar{x} \right) = \frac{1}{n} \left( \sum_{i=1}^n x_i^2 - 2 \bar{x} \sum_{i=1}^n x_i + n \bar{x}^2 \right)$$

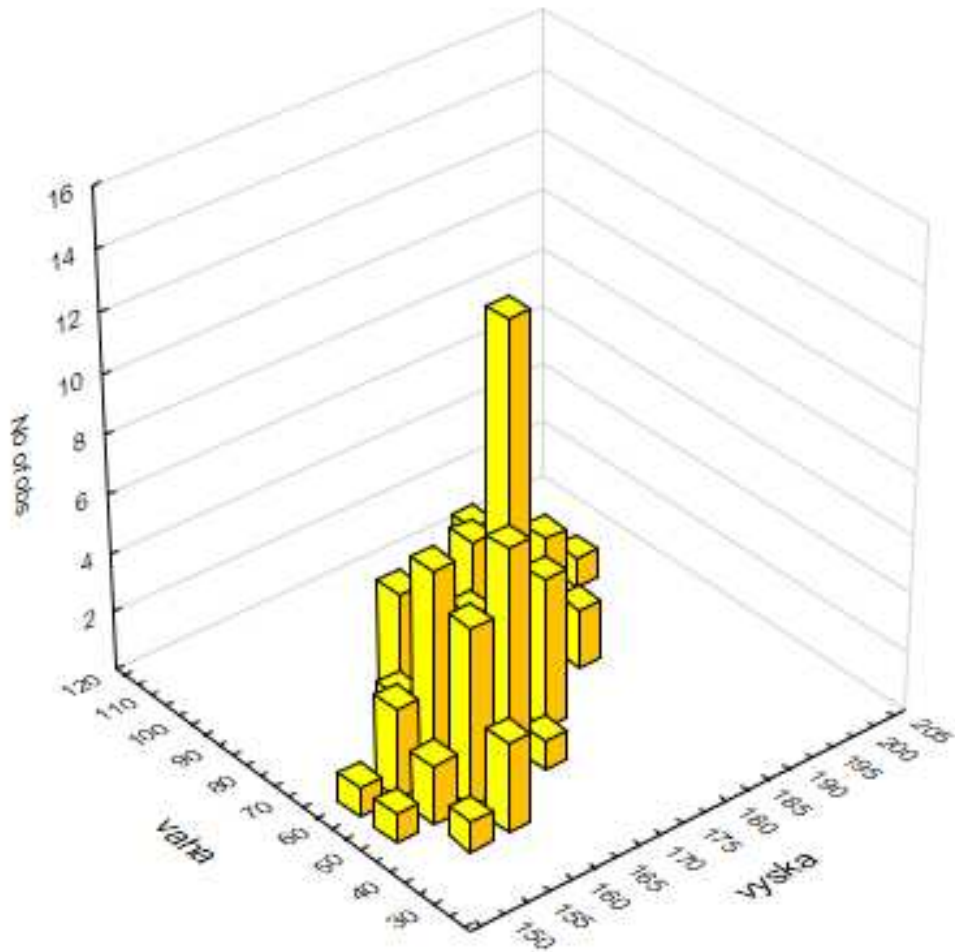
$$\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n} \left( \sum_{i=1}^n x_i y_i - \bar{y} \sum_{i=1}^n x_i - \bar{x} \sum_{i=1}^n y_i + n \bar{x} \bar{y} \right) \rightarrow \left[ \begin{array}{c} \end{array} \right]$$



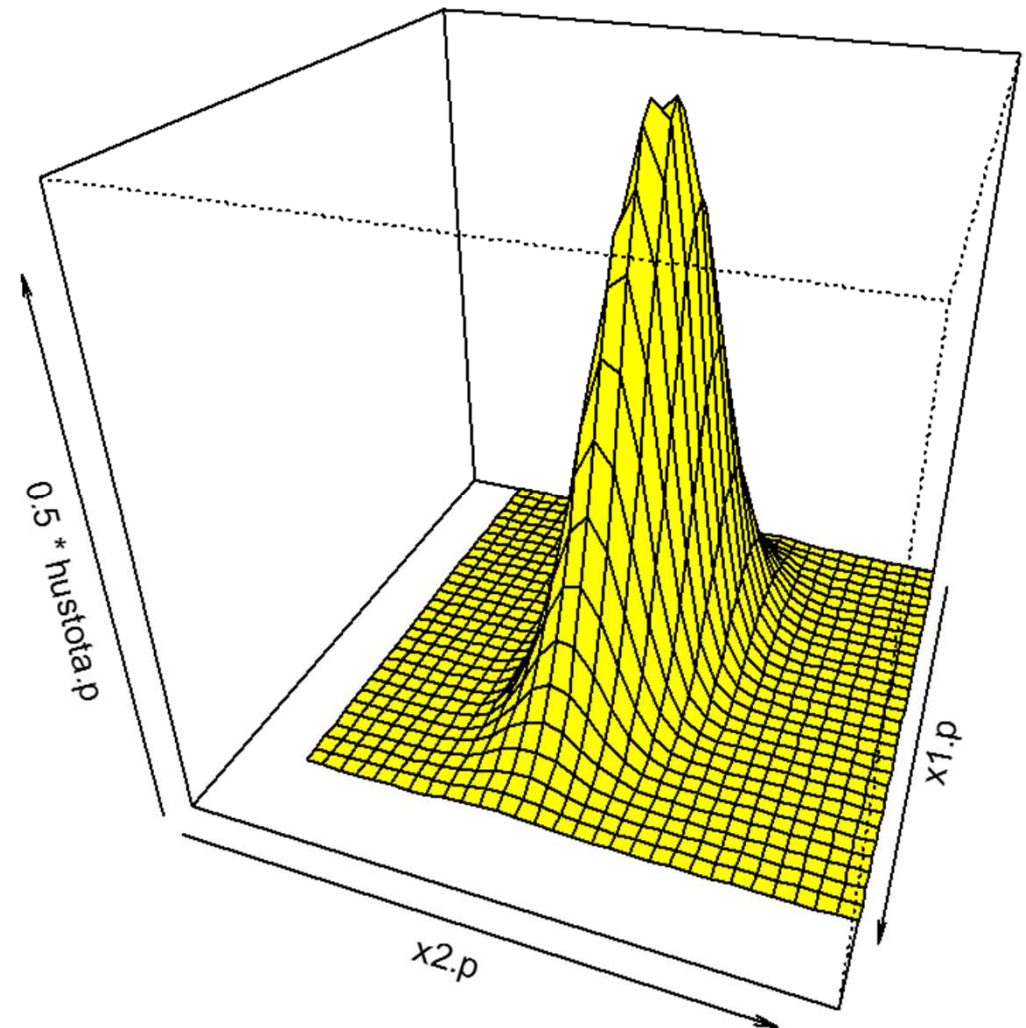
# Vícerozměrné normální rozdělení

# Motivace

Dvourozměrný histogram

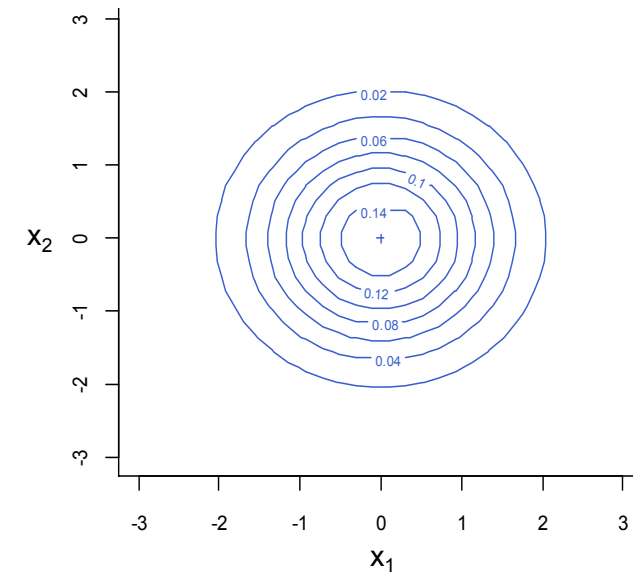
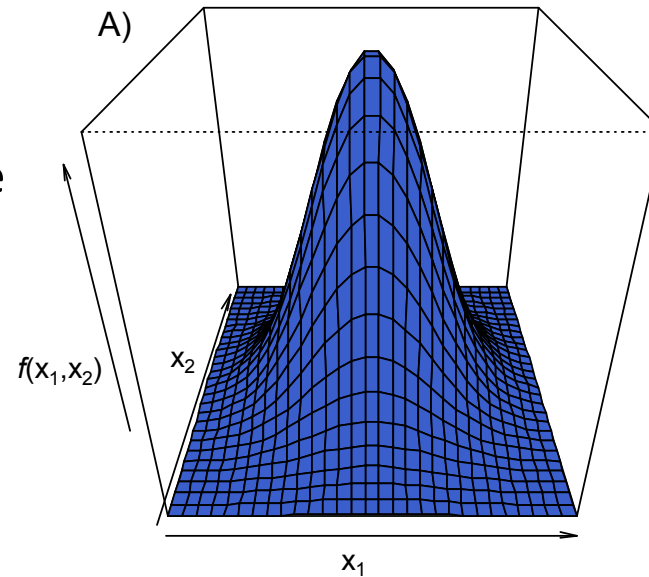


Hustota dvourozměrného normálního rozdělení

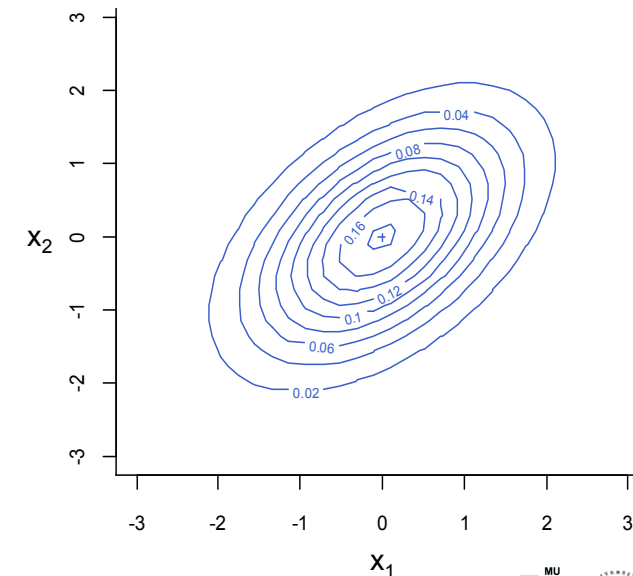
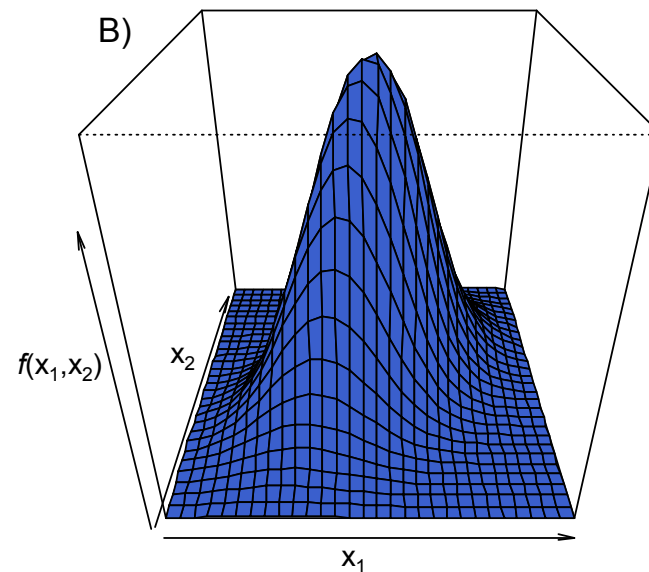


# Hustota u nekorelovaných a korelovaných proměnných

Nekorelované proměnné  
( $\mu_1 = \mu_2 = 0$ ,  $\sigma_1 = \sigma_2 = 1$ ,  
 $\rho = 0$ )



Korelované proměnné  
( $\mu_1 = \mu_2 = 0$ ,  $\sigma_1 = \sigma_2 = 1$ ,  
 $\rho = 0,5$ )



# Vícerozměrné normální rozdělení

**Hustota jednozměrného normálního rozdělení:**

$$f(x) = \frac{1}{\sqrt{\sigma^2}} \exp\left(-\frac{(x - \mu)^2}{\sigma^2}\right)$$

- střední hodnota  $\mu$   $\sigma^2$  – rozptyl

**Hustota vícerozměrného normálního rozdělení:**

$$f(\mathbf{x}) = \frac{1}{\sqrt{|\Sigma|}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})\right)$$

- vektor středních hodnot  $\boldsymbol{\mu}$  - kovarianční matice  $\Sigma$

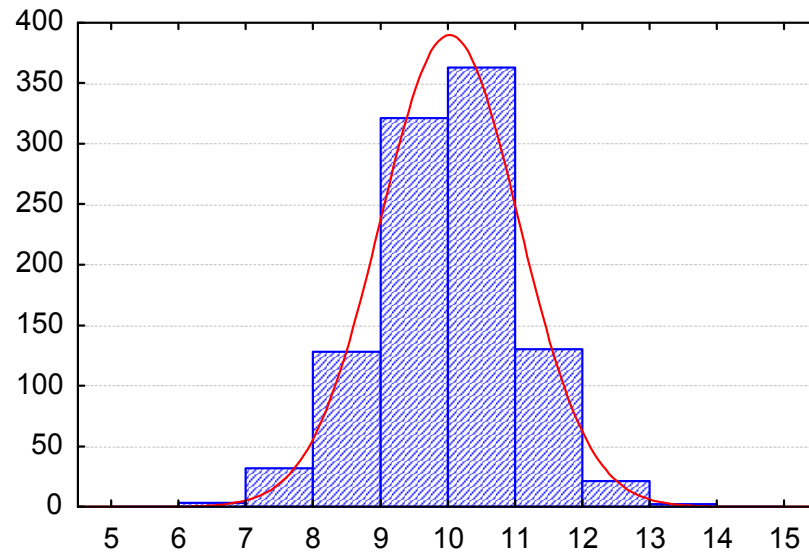
**Hustota dvourozměrného normálního rozdělení:**

$$f(x, y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)} \left[\frac{(x-\mu_x)^2}{\sigma_x^2} + \frac{(y-\mu_y)^2}{\sigma_y^2} - \frac{2\rho(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y}\right]\right),$$

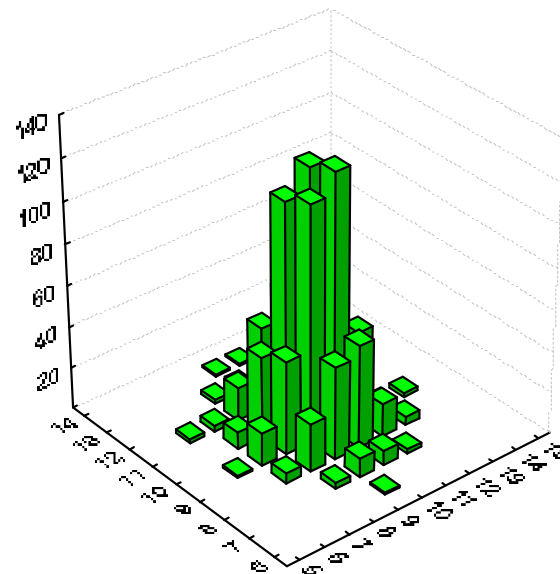
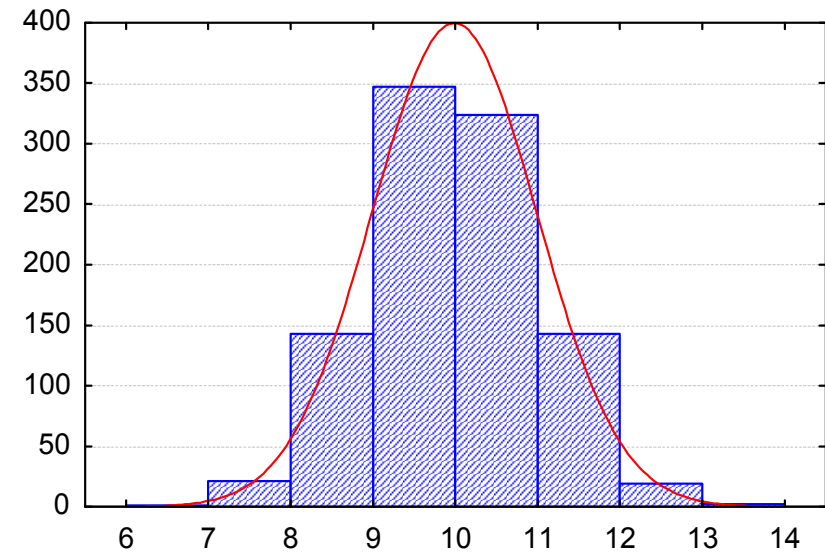
$$\boldsymbol{\mu} = \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix}$$

- korelace mezi X a Y;  
 $\sigma$  – směrodatná odchylka

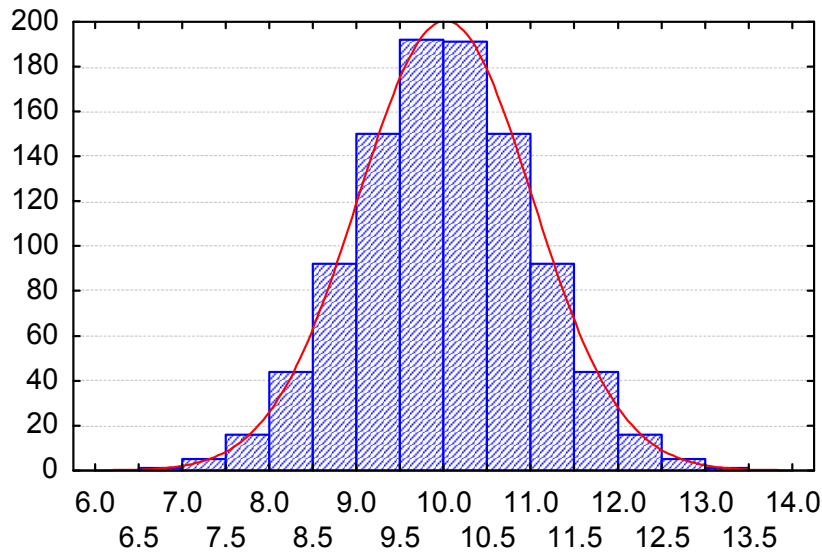
# Je normalita v jednorozměrném prostoru jedinou podmínkou vícerozměrné normality?



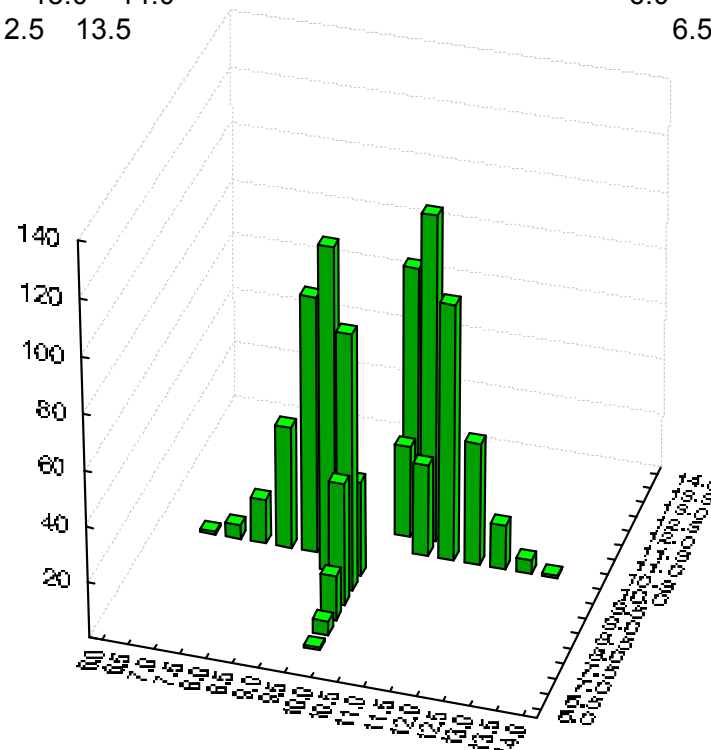
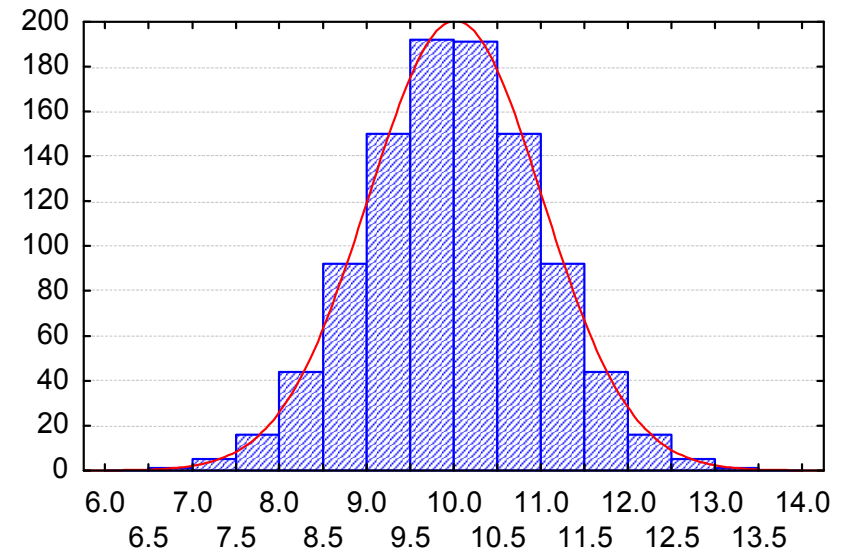
+



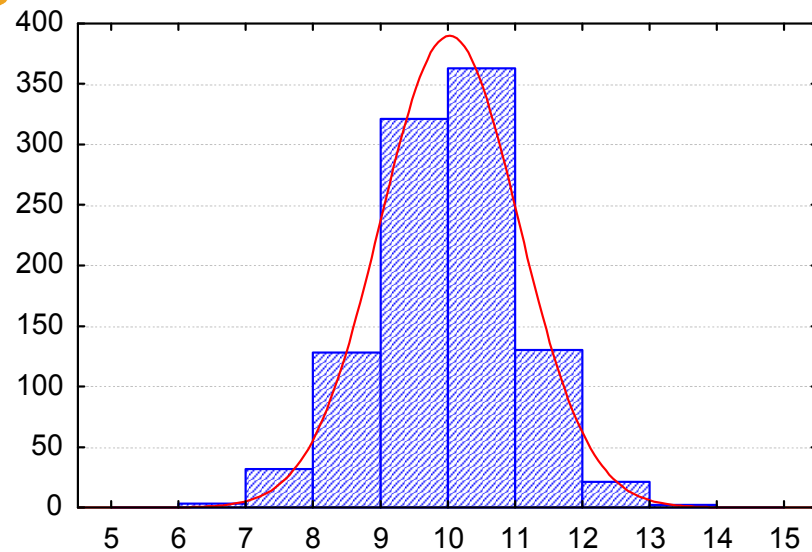
# Je normalita v jednorozměrném prostoru jedinou podmínkou vícerozměrné normality?



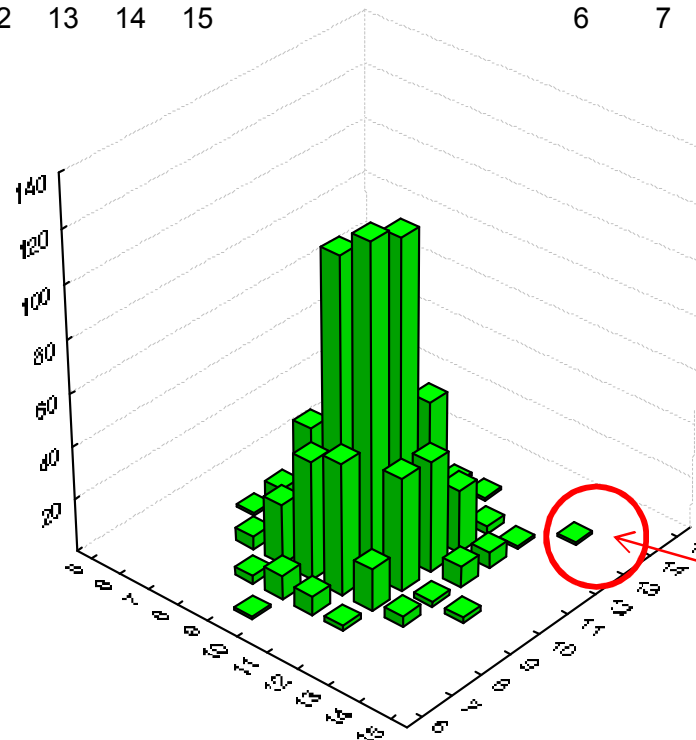
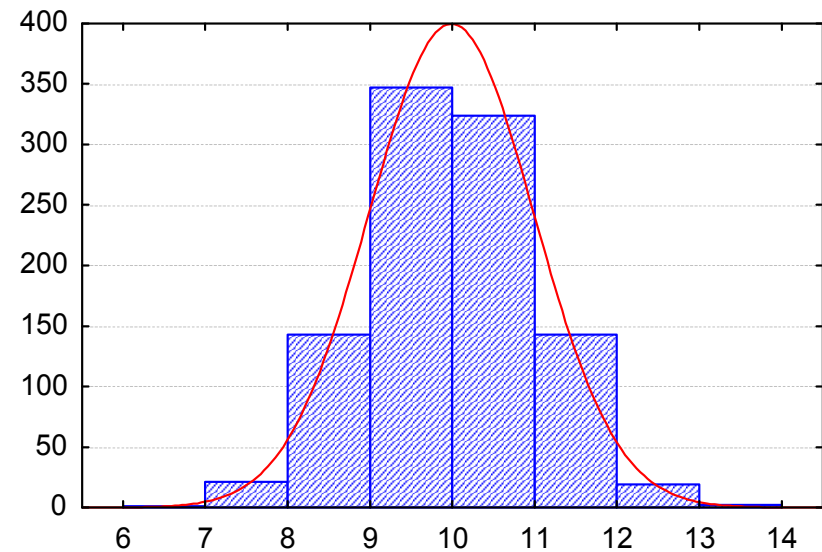
+



# Je normalita v jednorozměrném prostoru jedinou podmínkou vícerozměrné normality?



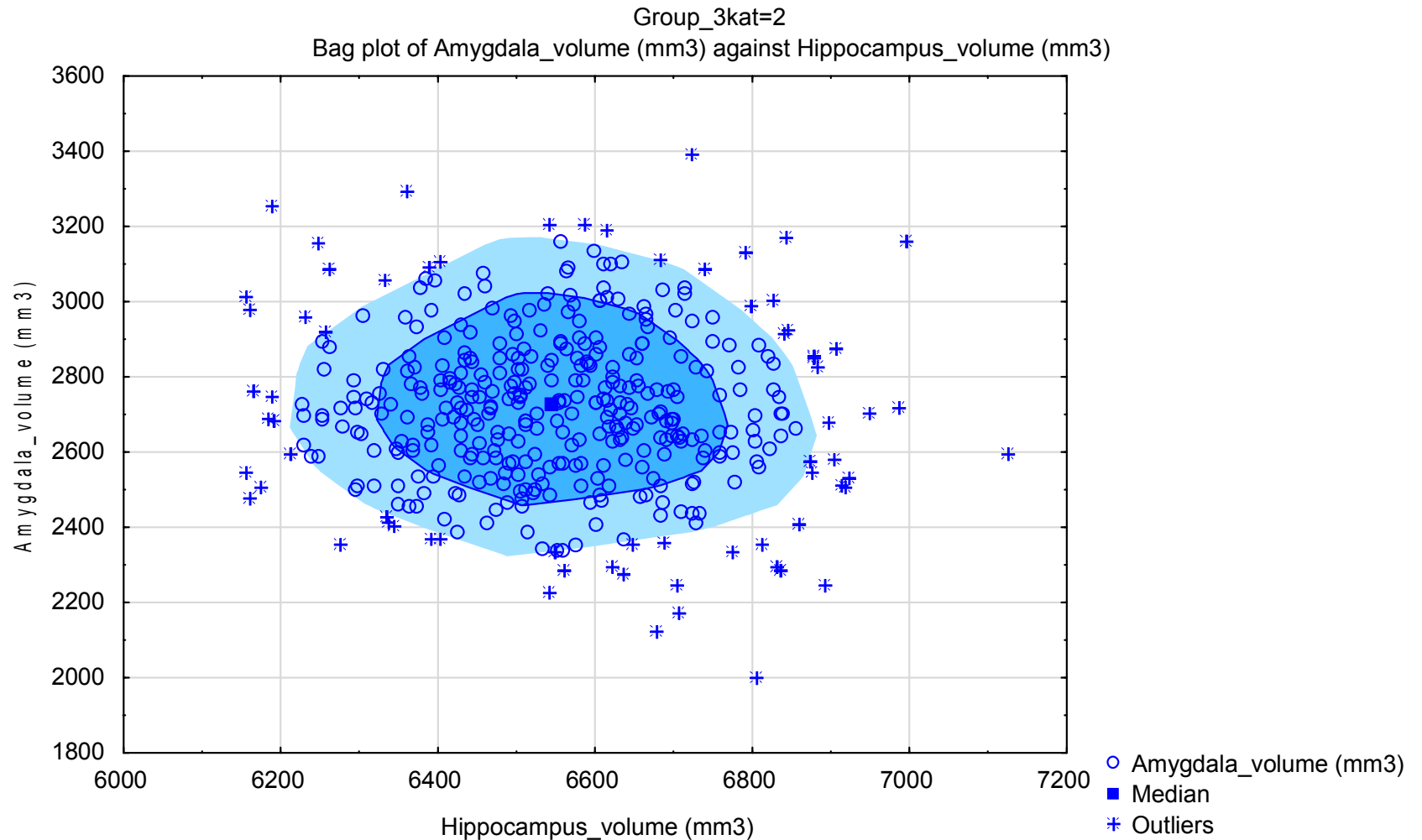
+



Vícerozměrný outlier

# Ověření dvourozměrné normality

Bagplot = „bivariate boxplot“ (tzn. „dvourozměrný krabicový graf“)

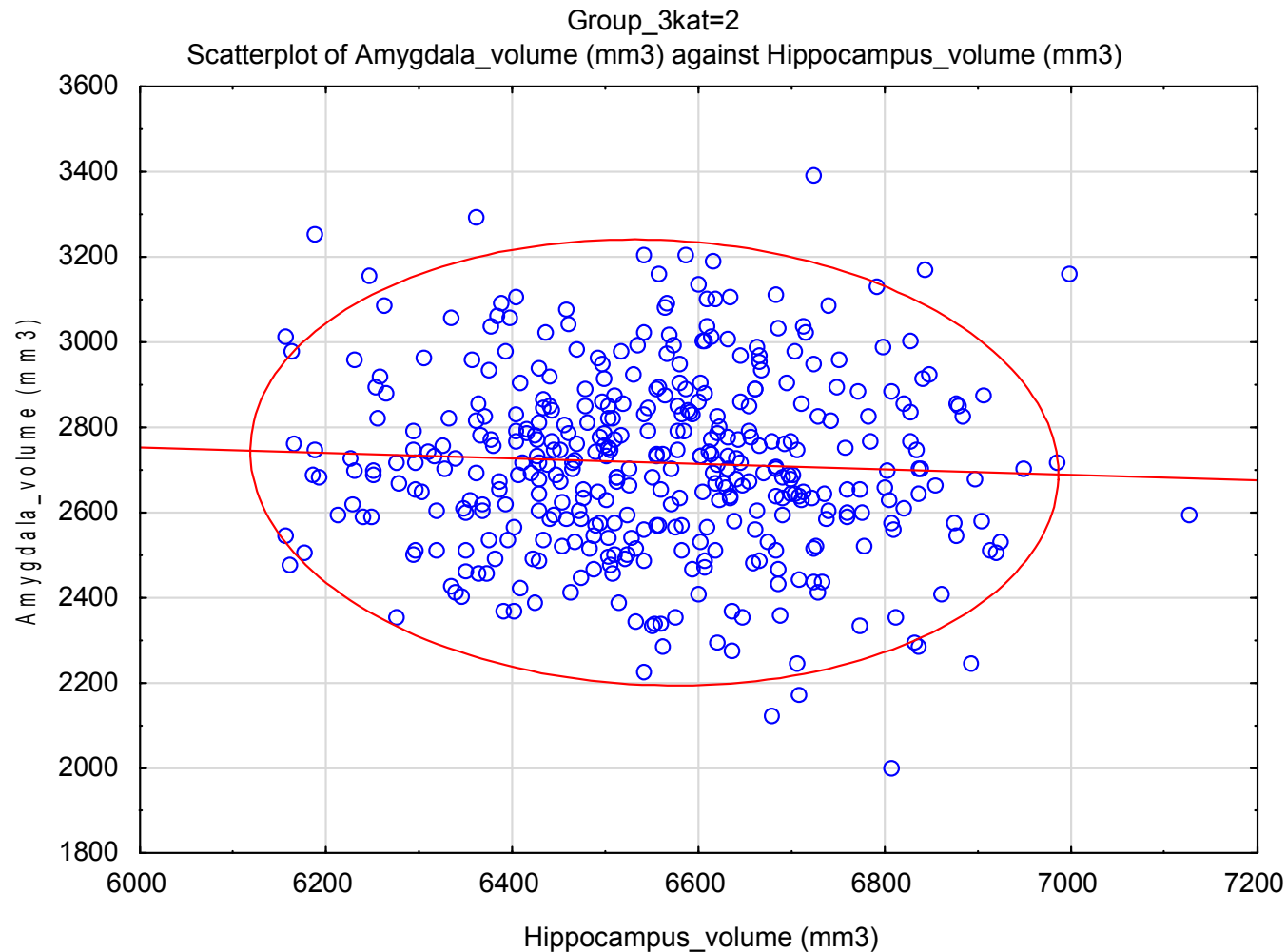


v softwaru Statistica: Graphs – 2D Graphs – Bag Plots



# Ověření dvourozměrné normality

Vykreslení regulační elipsy („control“ ellipse):

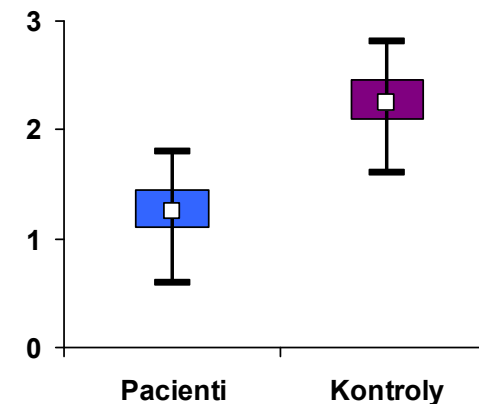
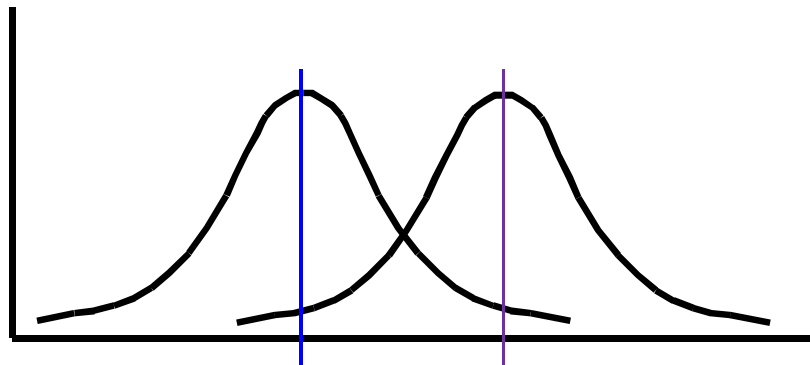


v softwaru Statistica: Graphs – Scatterplots – na záložce Advanced zvolit Ellipse Normal

# Vícerozměrný t-test

# Jednorozměrný dvouvýběrový t-test

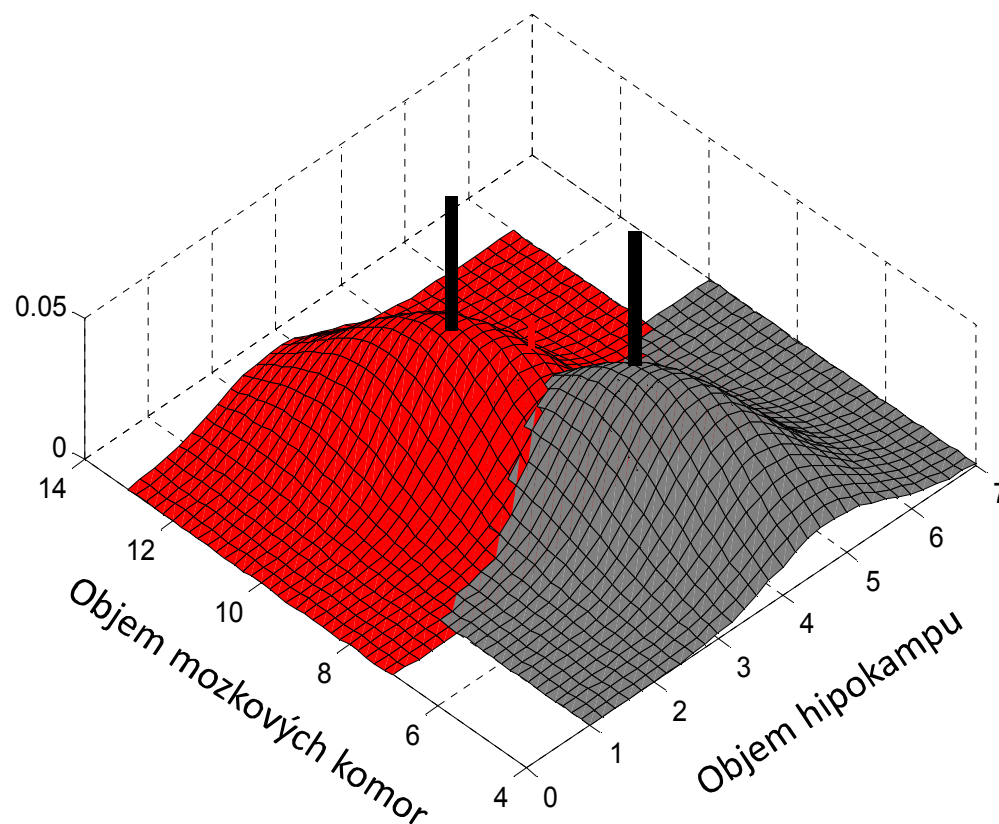
- Srovnáváme dvě skupiny dat, které jsou na sobě nezávislé – mezi objekty neexistuje vazba.
- Příklady: srovnání objem hipokampu u mužů a u žen, srovnání kognitivního výkonu podle dvou kategorií věku,...



- Předpoklad: **normalita dat v OBOU skupinách, shodnost (homogenita) rozptylů** v obou skupinách
- Testová statistika: 
$$t = \frac{\bar{x}_1 - \bar{x}_2 - c}{s_* \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$
, kde  $s_*$  je vážená směrodatná odchylka,  $c$  je konstanta, o kterou se rozdíl průměrů má lišit (většinou rovna 0)

# Vícerozměrný t-test

- Srovnáváme dvě skupiny dat, které jsou na sobě nezávislé – mezi objekty neexistuje vazba.
- Na rozdíl od jednorozměrného dvouvýběrového t-testu jsou dvě skupiny dat popsány více proměnnými.



# Vícerozměrný t-test

## Jednorozměrný dvouvýběrový t-test:

- testová statistika:  $\frac{(\bar{x}_D - \bar{x}_H)}{\sqrt{s_p^2 \left( \frac{1}{n_D} + \frac{1}{n_H} \right)}}$ , kde  $s_p^2$  ← Studentovo rozdělení
- $s_p^2$  je vážený rozptyl vypočtený jako  $\frac{(n_D - 1)s_D^2 + (n_H - 1)s_H^2}{n_D + n_H - 2}$
- $\mu_0$  je konstanta, o kterou se rozdíl průměrů má lišit (většinou 0)
- nulová hypotéza zamítnuta, pokud  $|\bar{x}_D - \bar{x}_H - \mu_0| > t_{\alpha/2, n_D+n_H-2} \cdot \sqrt{s_p^2 \left( \frac{1}{n_D} + \frac{1}{n_H} \right)}$

## Je ekvivalentní testu:

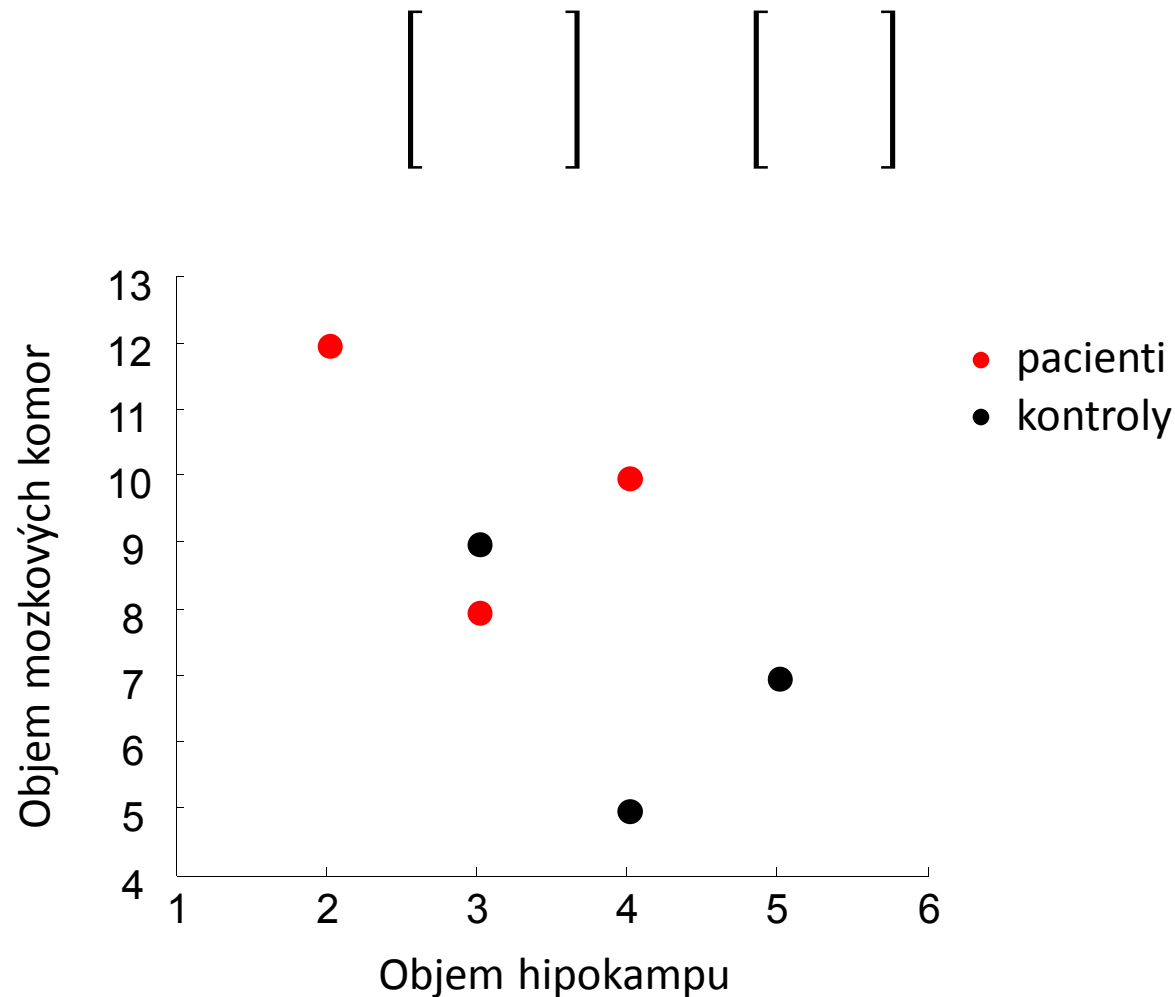
$$\left( \frac{(\bar{x}_D - \bar{x}_H - \mu_0)}{\sqrt{s_p^2 \left( \frac{1}{n_D} + \frac{1}{n_H} \right)}} \right)^2 \sim F(1, n_D + n_H - 2), \text{ kde } T^2 \sim F(1, n_D + n_H - 2) \text{ ← F rozdělení}$$

## Vícerozměrný t-test:

- Hotellingova  $T^2$  testová statistika:  $\frac{(\bar{x}_D - \bar{x}_H - \mu_0)' S_p^{-1} (\bar{x}_D - \bar{x}_H - \mu_0)}{\frac{1}{n_D + n_H} \text{tr}(S_p^{-1} \Sigma)}$
- kde  $S_p^{-1}$  je vážená kovarianční matice:  $\frac{(n_D - 1)S_D^{-1} + (n_H - 1)S_H^{-1}}{n_D + n_H - 2}$
- $T^2 \sim T^2(p, n - p - 1)$  ← Hotellingovo rozdělení; pro malé  $n_D$  a  $n_H$  je lepší použít:  $\frac{T^2}{n} \sim F\left(\frac{p}{2}, \frac{n-p-1}{2}\right)$ , kde  $n = n_D + n_H$
- nulová hypotéza zamítnuta, když  $T^2 > F_{\alpha, p/2, (n-p-1)/2} \cdot \frac{n}{n-1}$  ← F rozdělení

# Úkol 1

- Zjistěte, zda se liší skupina pacientů se schizofrenií od zdravých subjektů na základě parametrů popisujících objem mozkových struktur subjektů.



# Úkol 1 - řešení

Vícerozměrné průměry:

$$\begin{bmatrix} - & - \\ - & - \end{bmatrix} \quad \begin{bmatrix} \\ \end{bmatrix}$$

Výběrové kovarianční matice:

$$\begin{bmatrix} \\ \\ \end{bmatrix} \quad \begin{bmatrix} \\ \\ \end{bmatrix}$$

Vážená kovarianční matice:

$$\begin{bmatrix} \\ \end{bmatrix}$$

Vícerozměrný t-test:

n	6
p	2
$T^2$	3,5
F	1,31
df1= p	2
df2 = n-p-1	3
$\alpha$	0,05
F-crit	9,55
p-hodnota	0,389

$$\left( \quad \right) \begin{bmatrix} - & - \\ - & - \end{bmatrix} \left( \quad \right)$$

# Úkol 1 – řešení v software R

```
library("ICSNP")
```

```
X=matrix(c(2 4 3 12 10 8),3,2)
```

```
Y=matrix(c(5,3,4,7,9,5),3,2)
```

```
HotellingsT2(X, Y)
```

```
Hotelling's two sample T2-test
```

```
data: Xd and Xh
```

```
T.2 = 1.3125, df1 = 2, df2 = 3, p-value = 0.3895
```

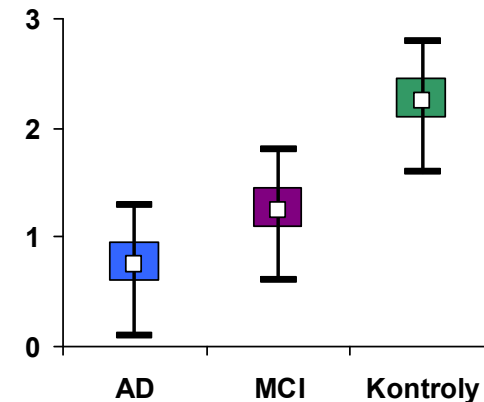
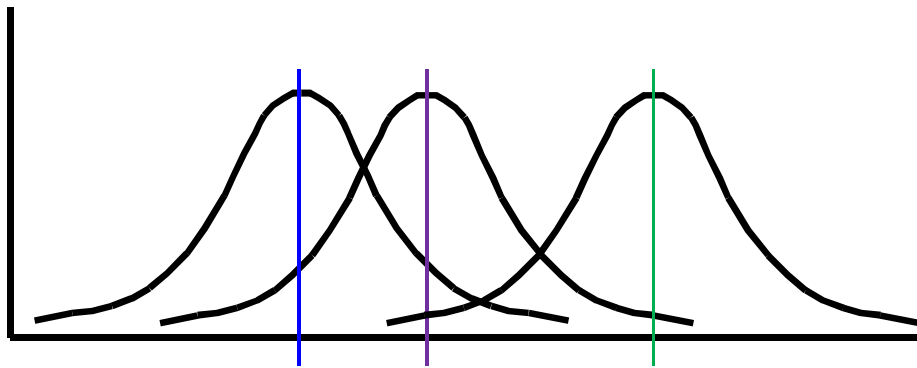
```
alternative hypothesis: true location difference is not equal to c(0,0)
```



# Analýza rozptylu pro vícerozměrná data

# Analýza rozptylu (ANOVA) jednoduchého třídění

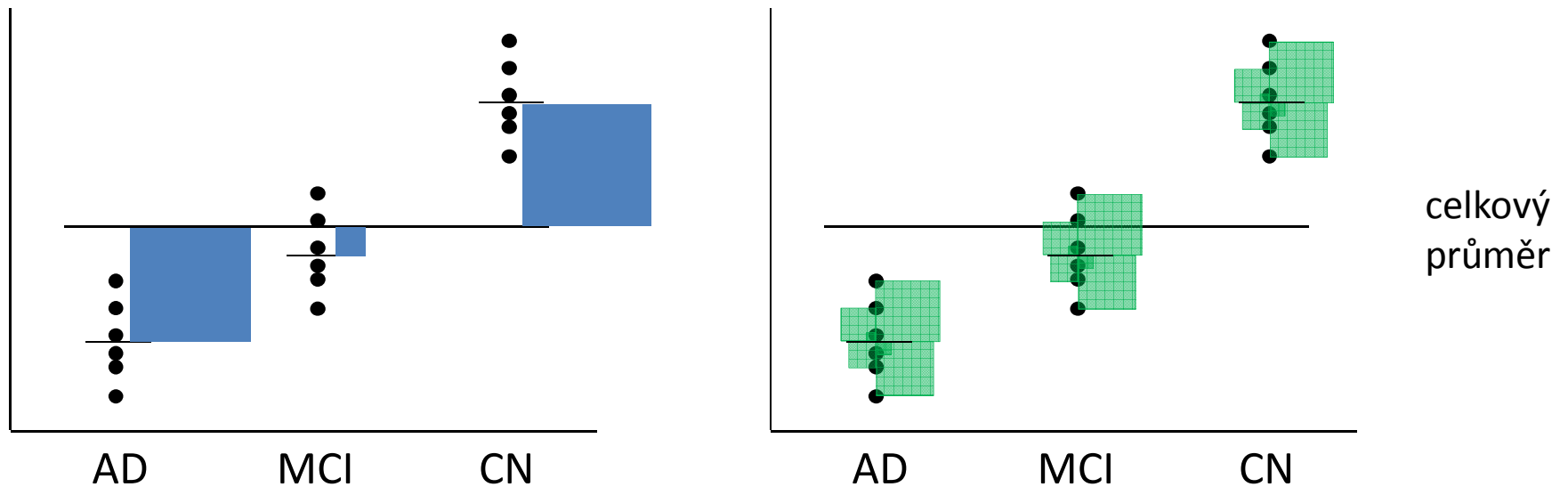
- Srovnáváme tři a více skupin dat, které jsou na sobě nezávislé (mezi objekty neexistuje vazba).
- Příklady: srovnání objemu hipokampu u pacientů s AD, pacientů s MCI a kontrol; srovnání kognitivního výkonu podle čtyř kategorií věku.



- 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	$S_A$	$df_A = k - 1$	$MS_A = S_A / df_A$	$F = \frac{S_A / df_A}{S_e / df_e}$	$p$
Uvnitř skupin (reziduální var.)	$S_e$	$df_e = n - k$	$MS_e = S_e / df_e$		
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 léků na hodnoty systolického tlaku u stovky osob  
– **jednorozměrná analýza rozptylu jednoduchého třídění**
- zkoumáme dlouhodobý vliv třech 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 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 léků na hodnoty systolického a diastolického tlaku u stovky osob – **vícerozměrná analýza rozptylu jednoduchého třídění**
- zkoumáme dlouhodobý vliv třech léků a vliv pohlaví na hodnoty systolického a diastolického tlaku u stovky osob – **vícerozměrná analýza rozptylu dvojného třídění**

# Analýza rozptylu dvojného třídění

- Uvažujeme dvě vysvětlující proměnné zároveň.
- Zápis modelu:

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

Populační průměr  $\mu$  (blue arrow),  $\alpha_i$   $i$ -tý efekt faktoru A (blue arrow),  $\beta_j$   $j$ -tý efekt faktoru B (red arrow),  $e_{ij}$  Reziduum (blue arrow)

- 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:

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

Diagrammatic labels for the equation above:
 

- $\mu$ : Populační průměr (blue arrow)
- $\alpha_i$ :  $i$ -tý efekt faktoru A (blue arrow)
- $\beta_j$ :  $j$ -tý efekt faktoru B (blue arrow)
- $\gamma_{ij}$ : Interakce (red arrow)
- $e_{ij}$ : Reziduum (blue arrow)

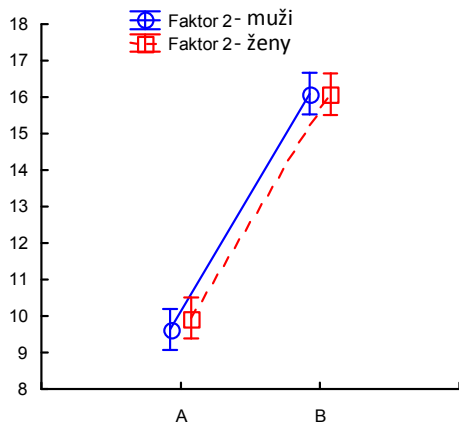
- Nulové hypotézy pak máme tři:

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

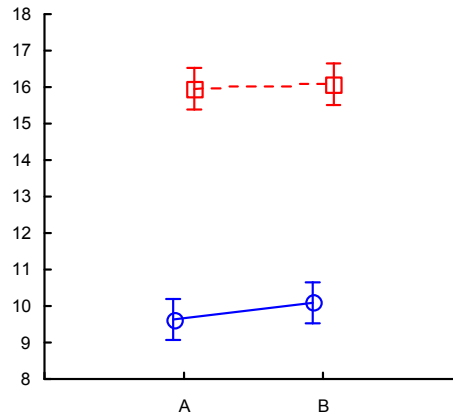
Variabilita	Součet čtverců	Počet stupňů volnosti	Průměrný čtverec	F statistika	p-hodnota
Faktor A	$S_A$	$df_A = a - 1$	$MS_A = S_A / df_A$	$F_A$	$p$
Faktor B	$S_B$	$df_B = b - 1$	$MS_B = S_B / df_B$	$F_B$	$p$
Interakce AxB	$S_{AB}$	$df_{AB} = (a-1)(b-1)$	$MS_{AB} = S_{AB} / df_{AB}$	$F_{AB}$	$p$
Rezidua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



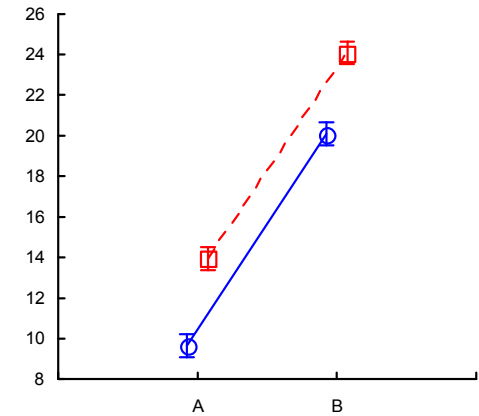
# Hlavní efekty a interakce



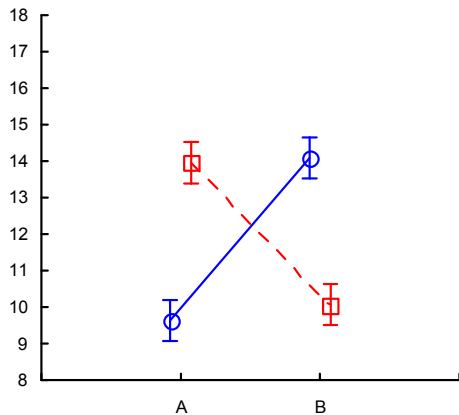
	SS	D.f.	MS	F	p
<b>Faktor 1</b>	<b>1978</b>	<b>1</b>	<b>1978</b>	<b>482.2</b>	<b>0.000</b>
Faktor 2	1	1	1	0.3	0.602
F1*F2	1	1	1	0.3	0.570
Error	804	196	4		



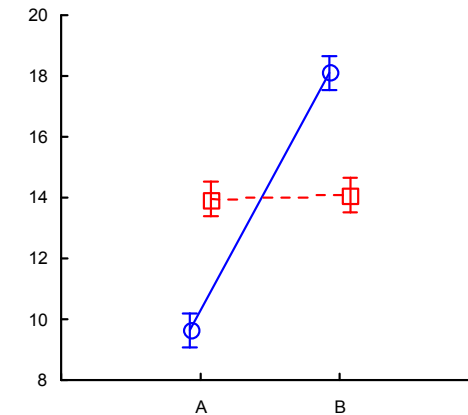
	SS	D.f.	MS	F	p
Faktor 1	4	1	4	1.0	0.314
<b>Faktor 2</b>	<b>1891</b>	<b>1</b>	<b>1891</b>	<b>461.1</b>	<b>0.000</b>
F1*F2	1	1	1	0.3	0.570
Error	804	196	4		



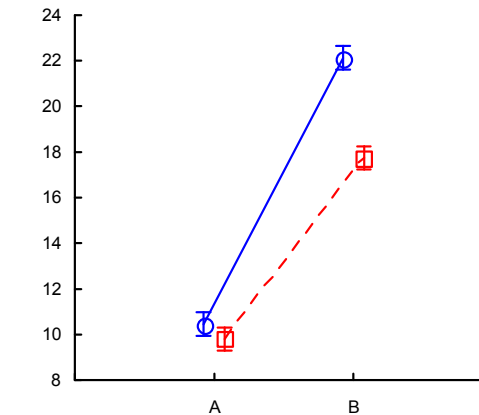
	SS	D.f.	MS	F	p
<b>Faktor 1</b>	<b>5293</b>	<b>1</b>	<b>5293</b>	<b>1290.7</b>	<b>0.000</b>
<b>Faktor 2</b>	<b>861</b>	<b>1</b>	<b>861</b>	<b>209.9</b>	<b>0.000</b>
F1*F2	1	1	1	0.3	0.570
Error	804	196	4		



	SS	D.f.	MS	F	p
Faktor 1	4	1	4	1.0	0.314
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		

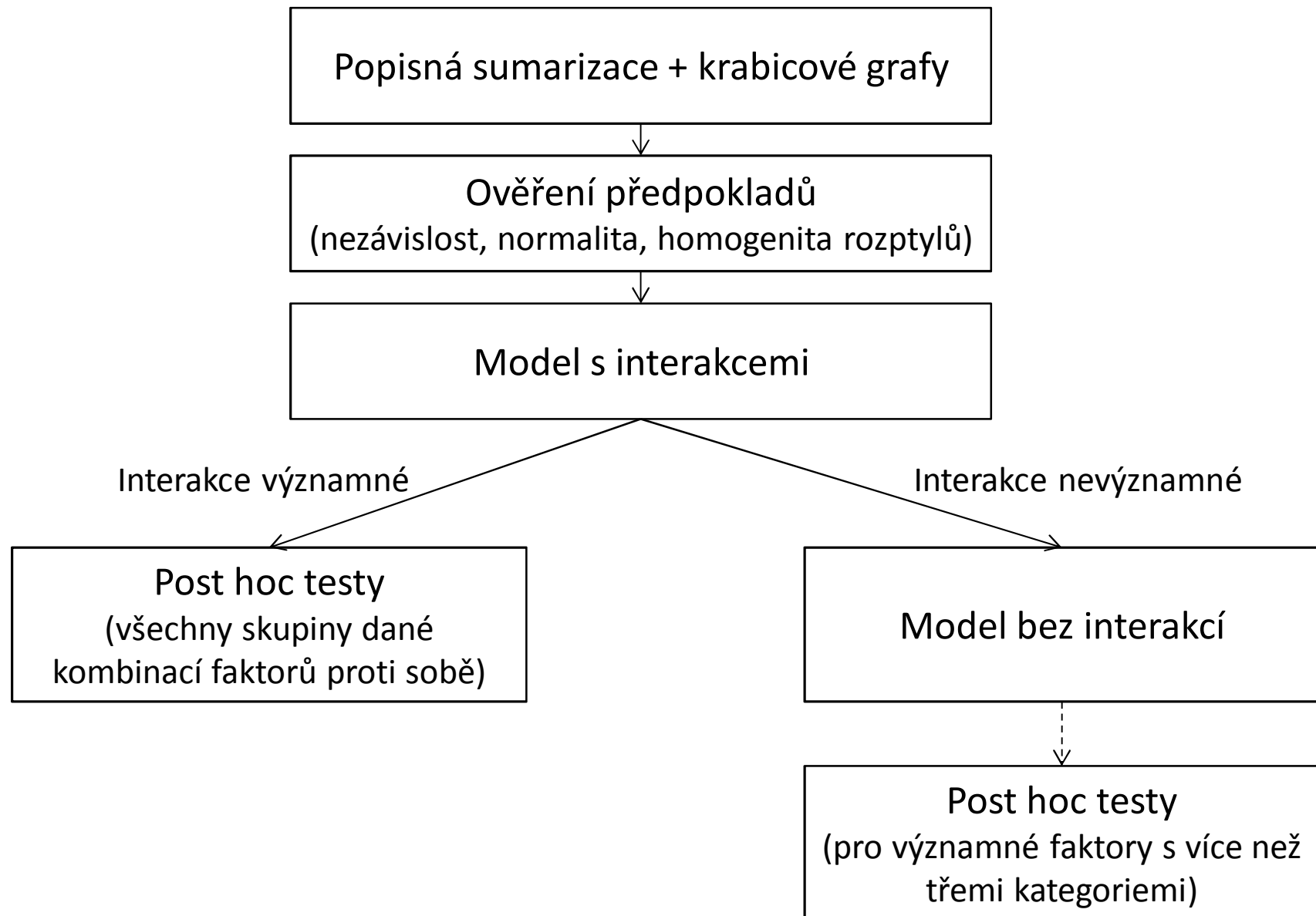


	SS	D.f.	MS	F	p
<b>Faktor 1</b>	<b>920</b>	<b>1</b>	<b>920</b>	<b>224.3</b>	<b>0.000</b>
Faktor 2	1	1	1	0.3	0.602
<b>F1*F2</b>	<b>867</b>	<b>1</b>	<b>867</b>	<b>211.3</b>	<b>0.000</b>
Error	804	196	4		



	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 2

Zjistěte, zda má vliv pohlaví a typ léku na počet nežádoucích účinků u pacientů se schizofrenií (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 2 – řešení

Zjistěte, zda má vliv pohlaví a typ léku na počet nežádoucích účinků u pacientů se schizofrenií (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

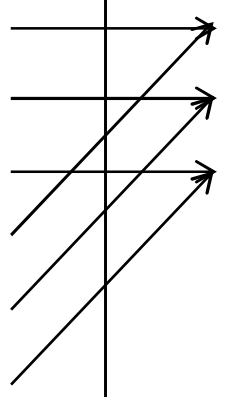
Legenda:

Pohlaví: 1=M  
2=Z

Typ léku: 1=lék X  
2=lék Y  
3=lék Z

# Úkol 2 – 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



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

$$\left( \begin{array}{c} 1 \\ 1 \\ 1 \\ 2 \\ 2 \\ 2 \end{array} \right) \left( \left( \begin{array}{c} 1 \\ 1 \\ 6 \\ 3 \\ 4 \\ 9 \end{array} \right) \left( \begin{array}{c} 1 \\ 1 \\ 6 \\ 3 \\ 4 \\ 9 \end{array} \right) \right)$$

počet stupňů volnosti:

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

$$\left( \begin{array}{c} 1 \\ 1 \\ 1 \\ 2 \\ 2 \\ 2 \end{array} \right) \left( \left( \begin{array}{c} 1 \\ 1 \\ 6 \\ 3 \\ 4 \\ 9 \end{array} \right) \left( \begin{array}{c} 1 \\ 1 \\ 6 \\ 3 \\ 4 \\ 9 \end{array} \right) \left( \begin{array}{c} 1 \\ 1 \\ 6 \\ 3 \\ 4 \\ 9 \end{array} \right) \right)$$

počet stupňů volnosti:

**Celkový součet čtverců :**

$$\left( \begin{array}{c} 1 \\ 1 \\ 1 \\ 2 \\ 2 \\ 2 \end{array} \right) \left( \begin{array}{c} 1 \\ 1 \\ 6 \\ 3 \\ 4 \\ 9 \end{array} \right) \left( \begin{array}{c} 1 \\ 1 \\ 6 \\ 3 \\ 4 \\ 9 \end{array} \right) \left( \begin{array}{c} 1 \\ 1 \\ 6 \\ 3 \\ 4 \\ 9 \end{array} \right)$$

počet stupňů volnosti:

**Reziduální součet čtverců :**

počet stupňů volnosti:

# Úkol 2 – řešení

Tabulka analýzy rozptylu dvojného třídění:

Zdroj variability	Součet čtverců	Stupně volnosti	Podíl	————	p
Faktor A (pohlaví)			10,67	63,99	0,015
Faktor B (typ léku)			18,5	110,98	0,009
Reziduální			0,16	-	-
Celkový			-	-	-

Srovnání s kvantily:

→ pohlaví má vliv na počet nežádoucích účinků

→ typ léku má vliv na počet nežádoucích účinků

# Úkol 2 – řešení v softwaru STATISTICA

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

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 – Unequal N HSD, 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 – řešení v softwaru SPSS

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

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 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 2 – řešení v softwaru R

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

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

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

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

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

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

# Úkol 3

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)

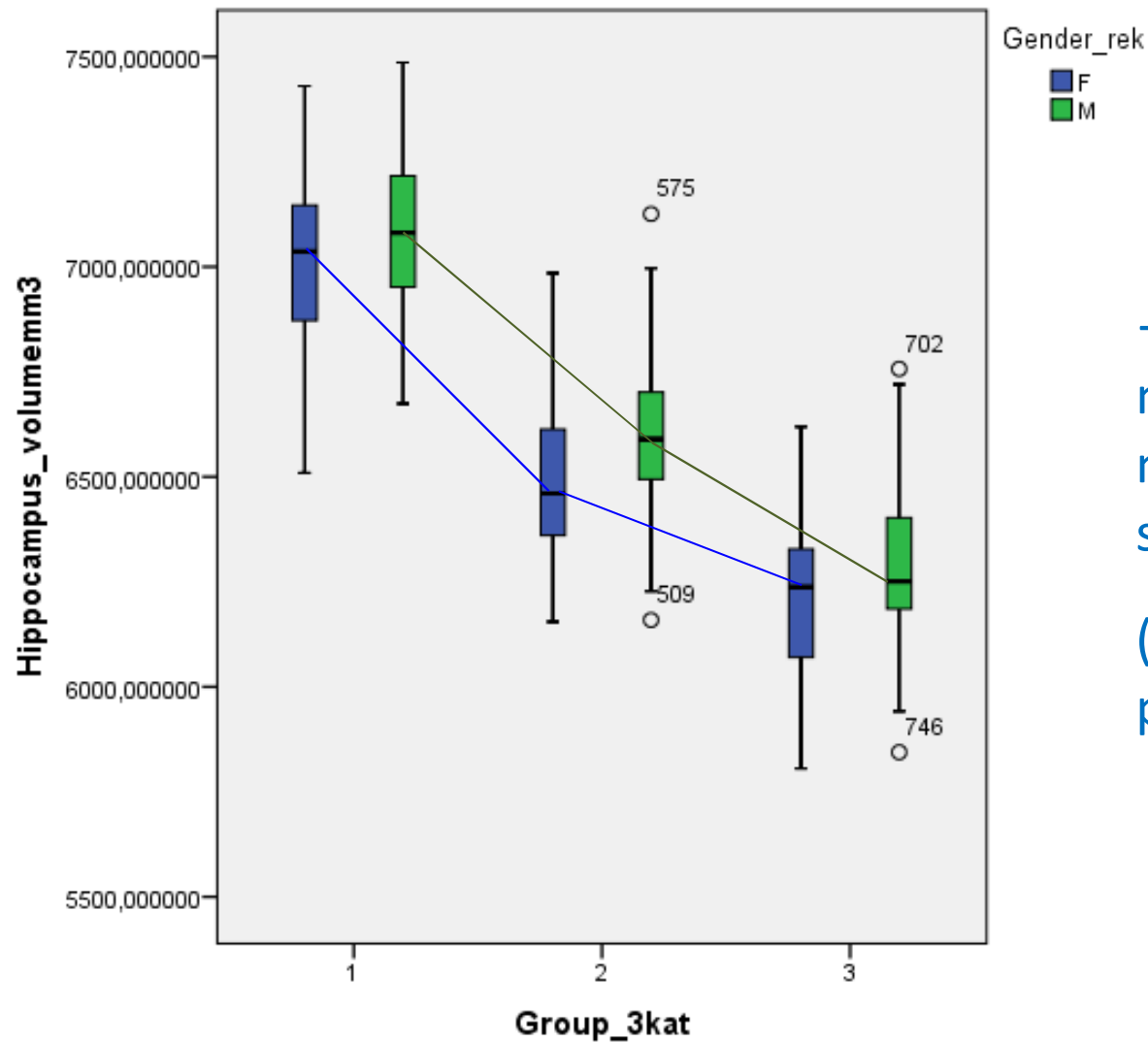
2...MCI (mírná kognitivní porucha)

3...AD (Alzheimerova choroba)

# Úkol 3 – 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

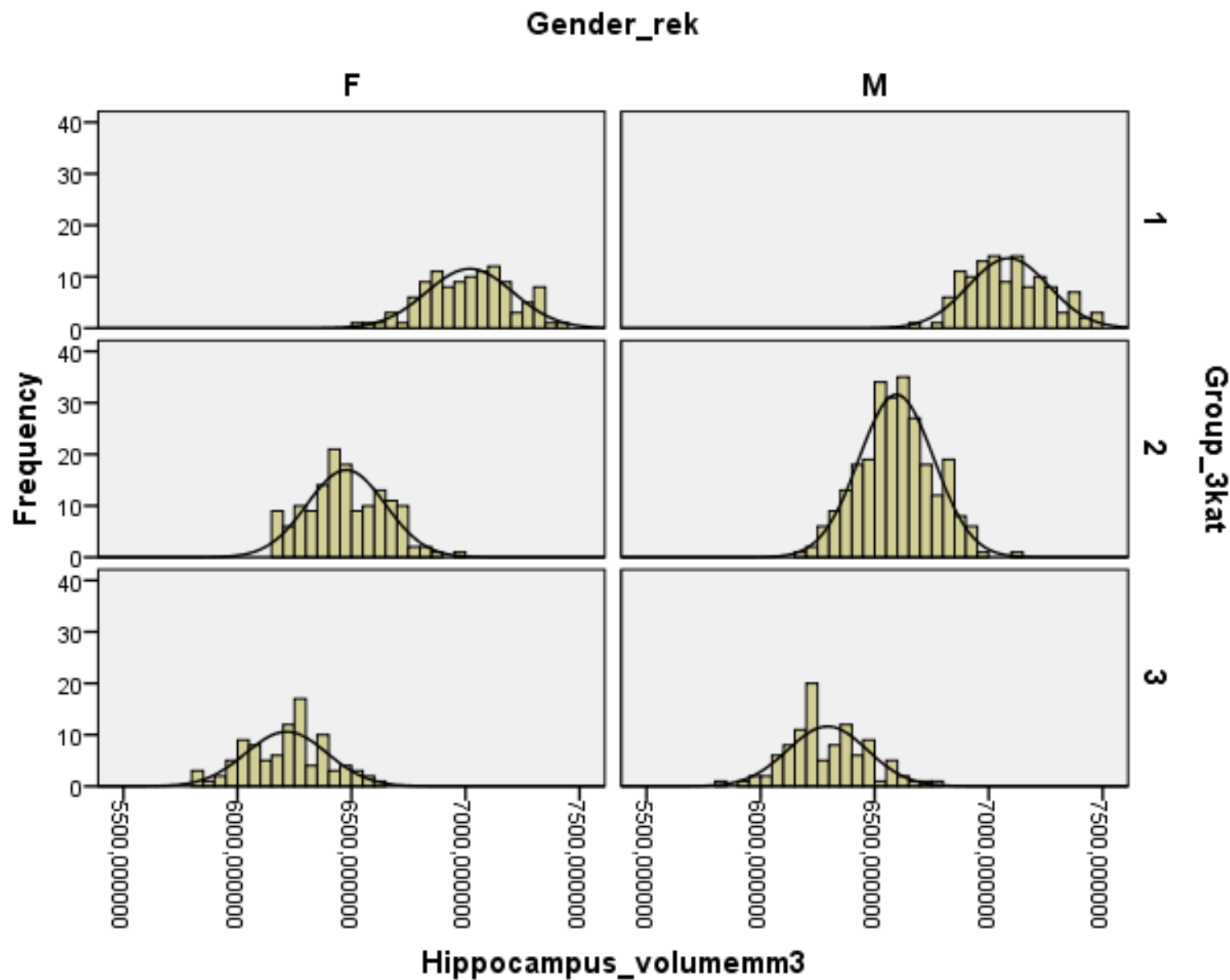
# Úkol 3 – krabicový graf



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

(nejdřív ale musíme ověřit  
předpoklady)

# Úkol 3 – ověření normality



# Úkol 3 – 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 3 – 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 3 – 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 3 – interpretace

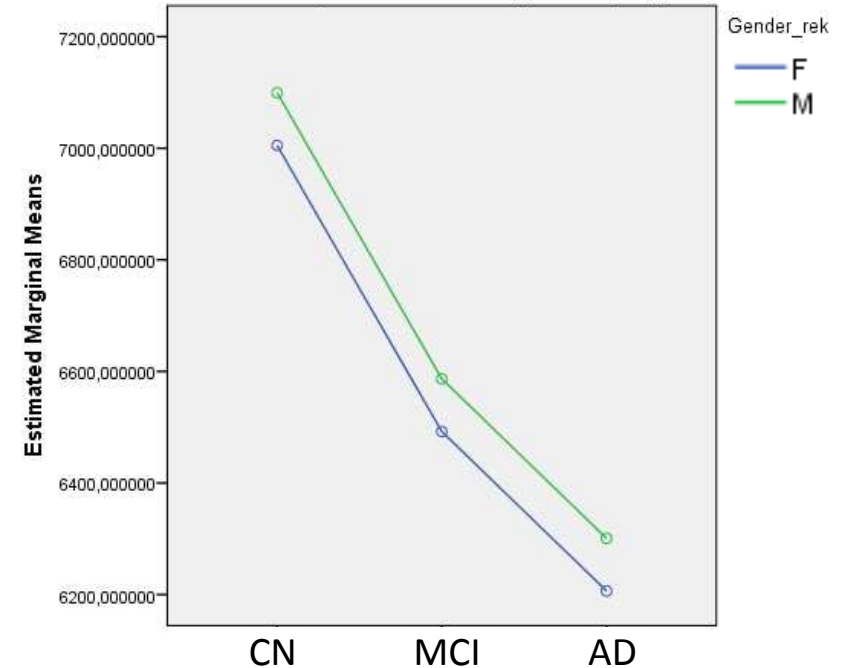
Hippocampus\_volumem3

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.

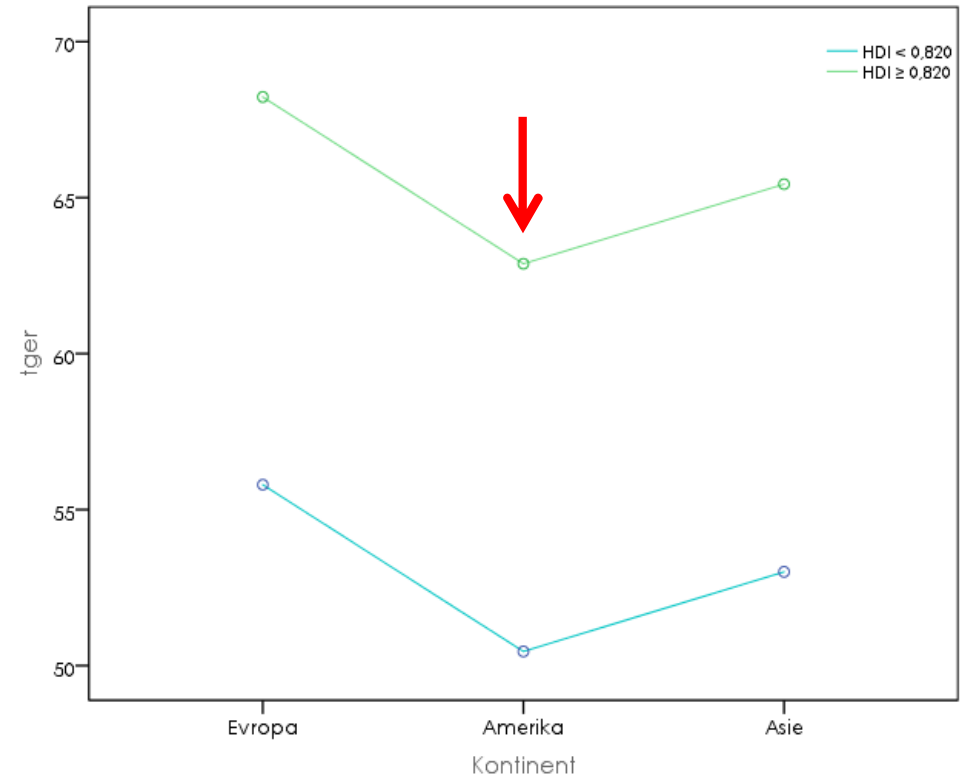
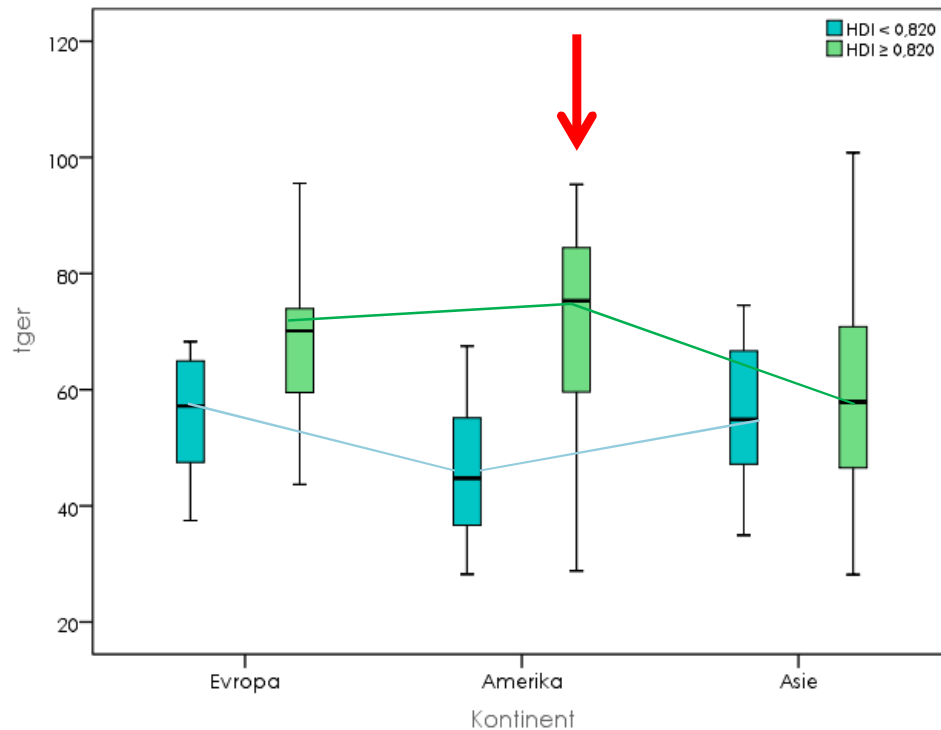
Estimated Marginal Means of Hippocampus\_volumem3



- 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!



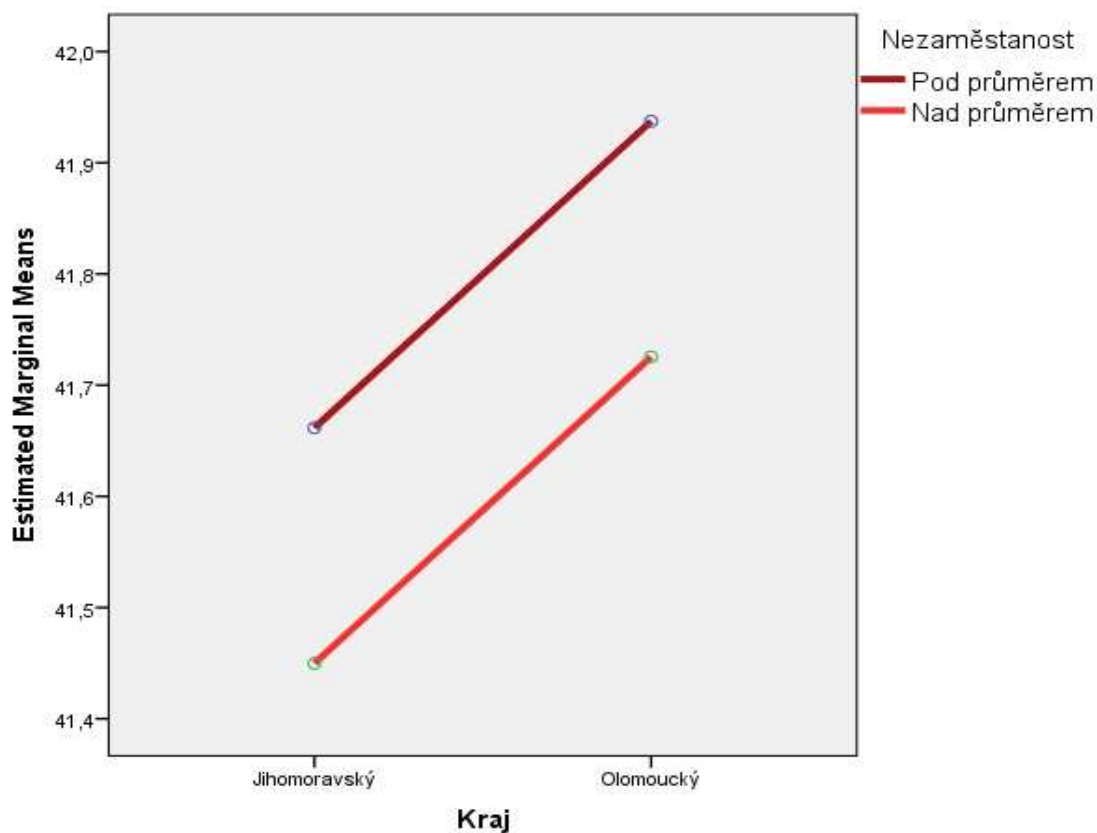
- znamená to, že tam zřejmě není splněn předpoklad normality
- pokud rozdíl není statisticky významný, není zpravidla potřeba to řešit
- pokud by ten rozdíl vyšel statisticky významně, je to problém!
- poznámka: je dobré mít měřítko na ose y stejné u obou grafů

# 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!)



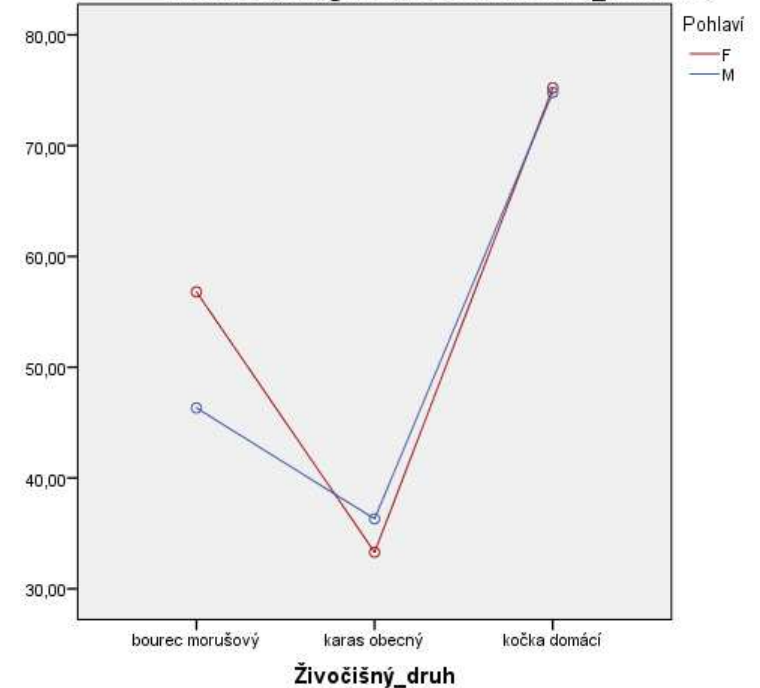
# Doplnění – model s interakcemi

## Tests of Between-Subjects Effects

Dependent Variable: Celková\_bílkovina

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	152178,501 <sup>a</sup>	5	30435,700	4942,124	,000
Intercept	1562235,885	1	1562235,885	253674,570	,000
Živočišný_druh	146815,301	2	73407,651	11919,874	,000
Pohlaví	931,626	1	931,626	151,277	,000
Živočišný_druh * Pohlaví	4431,573	2	2215,787	359,798	,000
Error	3288,599	534	6,158		
Total	1717702,985	540			
Corrected Total	155467,100	539			

Estimated Marginal Means of Celková\_bílkovina



a. R Squared = ,979 (Adjusted R Squared = ,979)

Cell No.		Živočišný_druh	Pohlaví	{1}	{2}	{3}	{4}	{5}	{6}
Unequal N HSD; variable Celková_bílkovina Approximate Probabilities for Post Hoc Tests Error: Between MS = 6.1584, df = 534.00				56.801	46.318	75.211	74.794	33.289	36.308
1	bourec morušový	F			0.000020	0.000020	0.000020	0.000020	0.000020
2	bourec morušový	M		0.000020		0.000020	0.000020	0.000020	0.000020
3	kočka domácí	F		0.000020	0.000020		0.870236	0.000020	0.000020
4	kočka domácí	M		0.000020	0.000020	0.870236		0.000020	0.000020
5	karas obecný	F		0.000020	0.000020	0.000020	0.000020		0.000020
6	karas obecný	M		0.000020	0.000020	0.000020	0.000020	0.000020	

Závěr:

- Nejvyšší koncentrace celkové bílkoviny zjištěny u kočky domácí a nejnižší u karase obecného.
- Vliv pohlaví různý u různých druhů. Největší vliv u bource morušového, přičemž F statisticky významně vyšší koncentrace než u M. Žádný vliv u kočky domácí. U karase obecného významně vyšší koncentrace u M než F.

# Transformace a jiné úpravy vícerozměrných dat

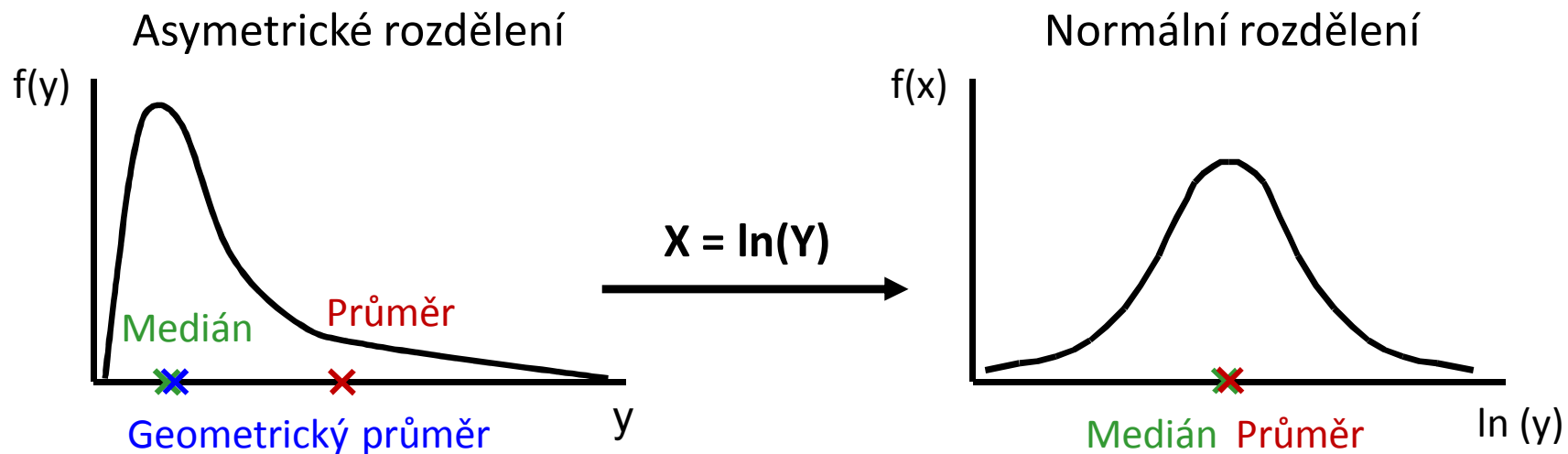
# Typy transformací a jiných úprav vícerozm. dat

---

- normalizace dat (= převod na normální rozdělení)
- standardizace dat
- min-max normalizace
- centrování dat
- odstranění vlivu kovariát na jiné proměnné

# Normalizace dat

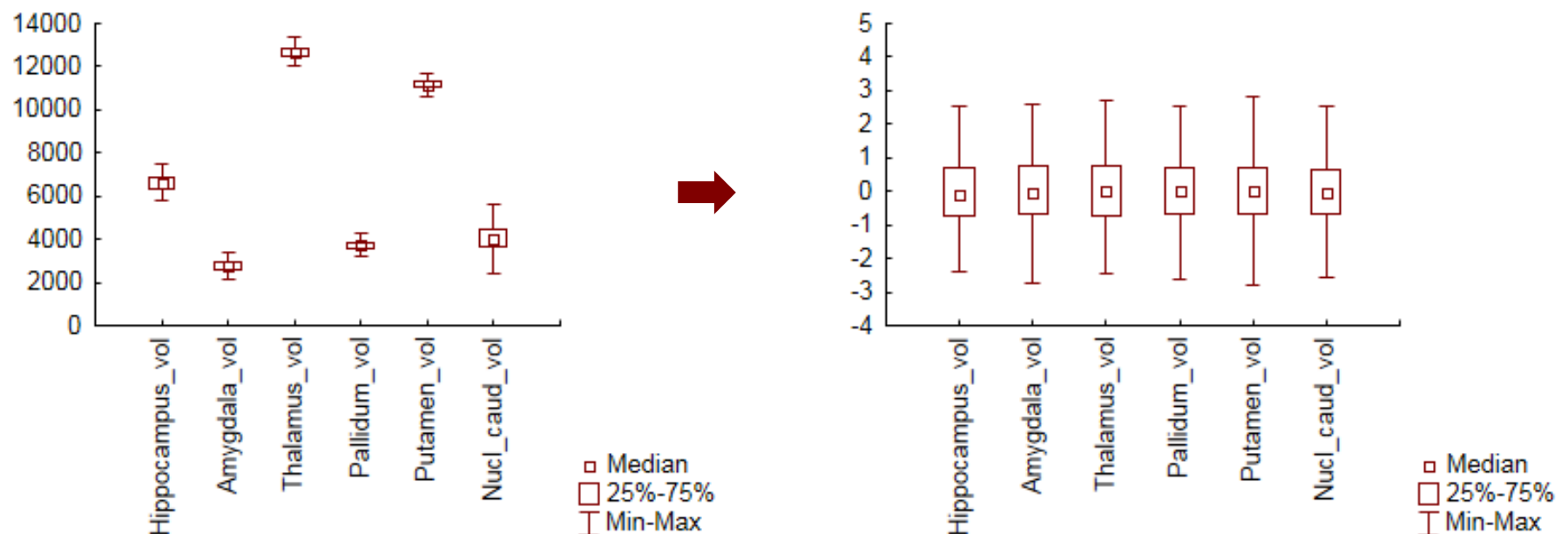
- převod na normální rozdělení (normalita je předpokladem řady statistických testů).
- např. **logaritmická transformace**:  $X = \ln(Y)$  nebo  $X = \ln(Y+1)$ , pokud data obsahují hodnotu 0



- další příklady:
  - **odmocninová transf.** (pro proměnné s Poissonovým rozložením nebo obecně data typu počet jedinců, buněk apod.:  $X = \sqrt{Y}$  nebo  $X = \sqrt{Y + 1}$ )
  - **arcsin transformace** (pro proměnné s binomickým rozložením)
  - **Box-Coxova transformace**

# Standardizace dat

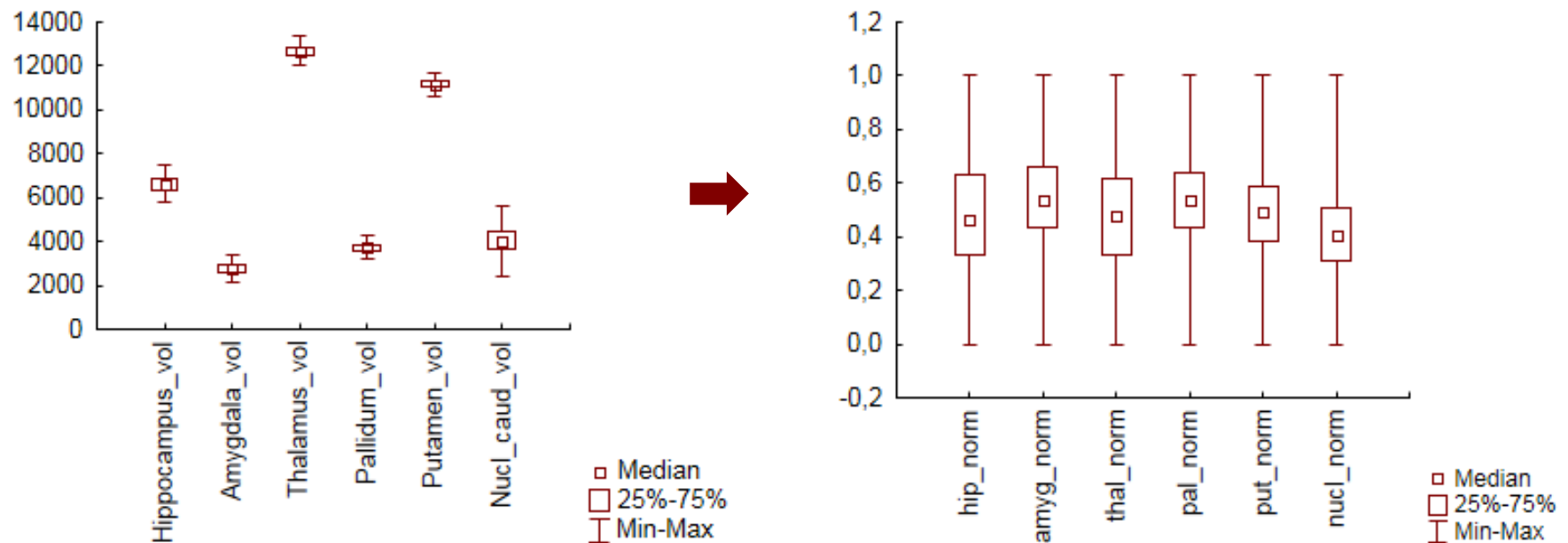
- důvod: převod proměnných na stejné měřítko
- standardizace:  $\frac{x - \mu}{\sigma}$  (tzn. odečtení průměru od jednotlivých hodnot a podělení směrodatnou odchylkou)
- proměnné budou mít rozsah přibližně od -3 do 3
- získáme tím současně i tzv. z-skóre (které vyjadřuje, o kolik směrodatných odchylek se i-tá hodnota odchýlila od průměru)
- **pozor: standardizace je nevhodná v případě, když proměnné nemají normální rozdělení a když se v datech vyskytují odlehlé hodnoty!!!**





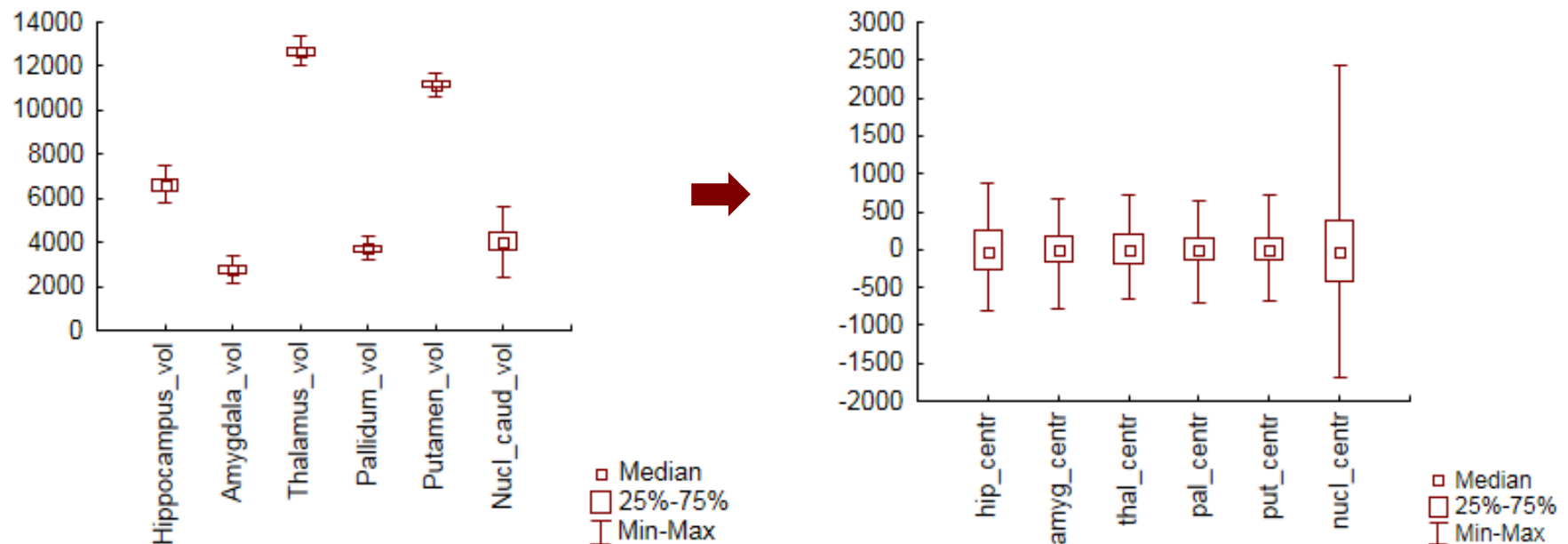
# Min-max normalizace

- důvod: převod proměnných na stejné měřítko
- oproti standardizaci vhodná i na proměnné nemající normální rozdělení či obsahující odlehlé hodnoty
- min-max normalizace: 
$$\frac{() - ()}{() - ()}$$
- rozsah hodnot proměnných po min-max normalizaci je od 0 do 1



# Centrování dat

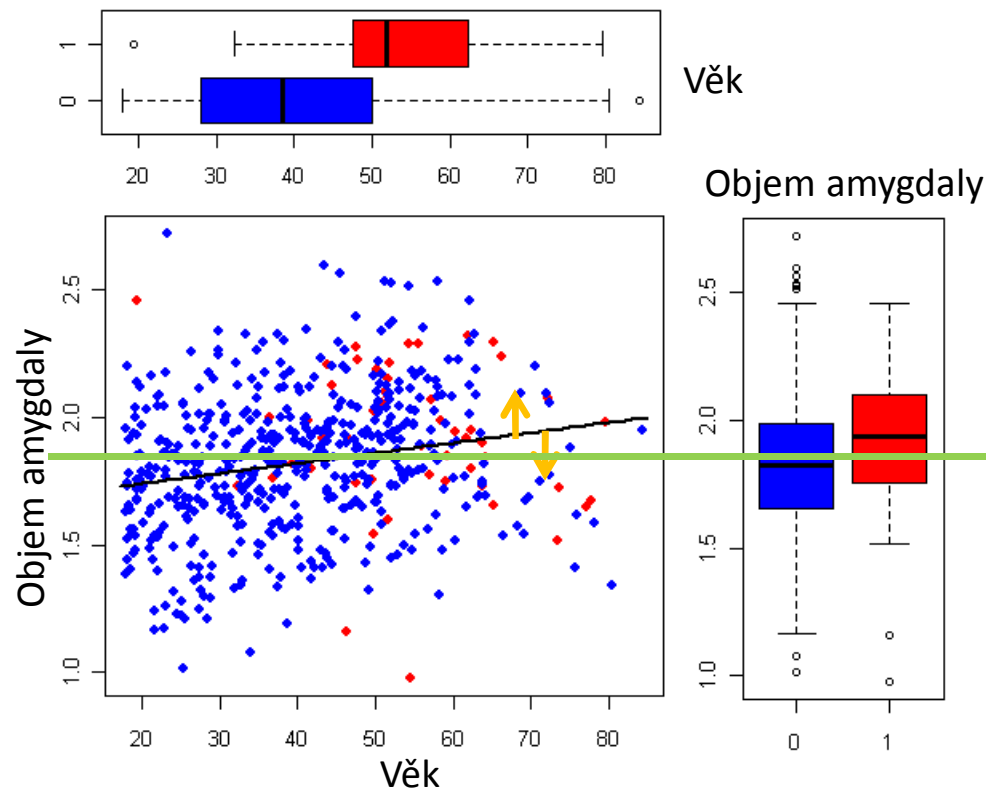
- odečtení průměru od dat – získáme novou proměnnou, která bude mít průměr roven nule
- důvod: centrování je důležitou podmínkou některých pokročilých statistických metod (např. klasifikačních)
- centrování:



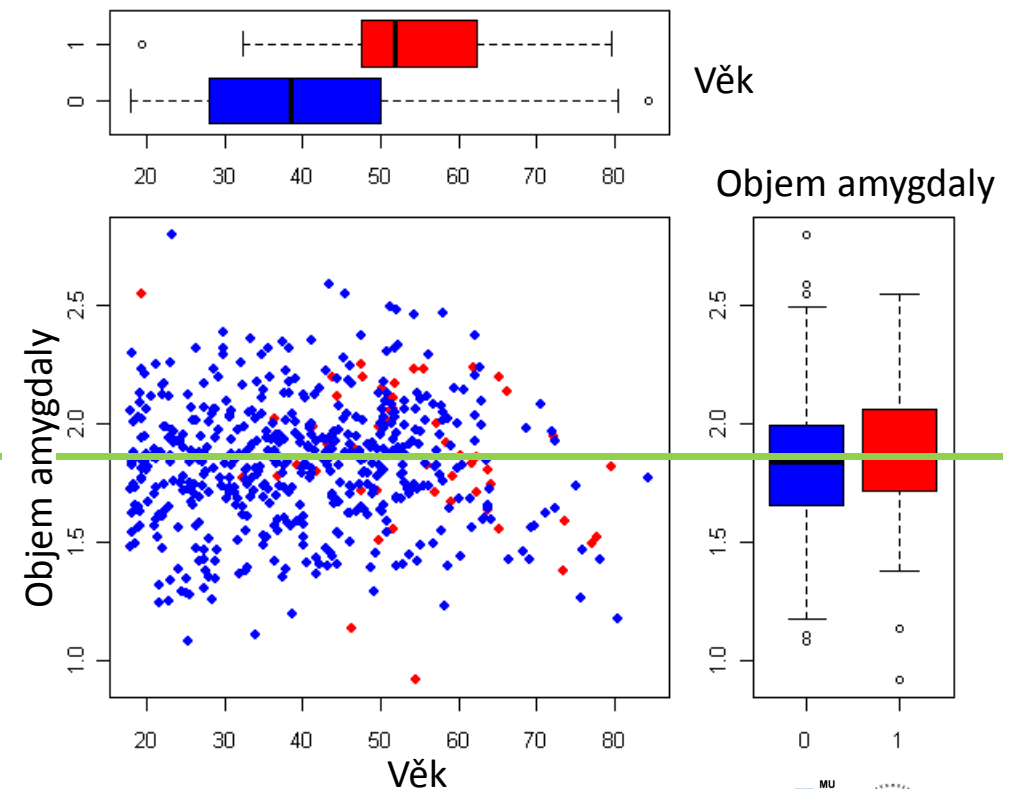
# Odstranění vlivu kovariát (tzv. adjustace)

1. V prvním kroku definujeme regresní model vztahu kovariáty (např. věku) a dané proměnné
2. Pro každého pacienta je vypočteno jeho reziduum od regresní přímky  $\uparrow\downarrow$
3. Reziduum (představující hodnotu parametru po odečtení vlivu věku, jeho průměr je 0) je přičteno k průměrné hodnotě parametru  $\text{—————}$
4. Výsledná adjustovaná hodnota má odečten vliv věku, ale zároveň není změněna číselná hodnota parametru

**Původní data**



**Adjustovaná data**



# Poděkování

Příprava výukových materiálů předmětu „DSAN02 Pokročilé metody analýzy dat v neurovědách“ byla finančně podporována prostředky projektu FRMU č. MUNI/FR/0260/2014 „Pokročilé metody analýzy dat v neurovědách jako nový předmět na LF MU“

