

# Pokročilé metody analýzy dat v neurovědách



RNDr. Eva Janoušová  
doc. RNDr. Ladislav Dušek, Dr.

Jaro 2016

# 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):

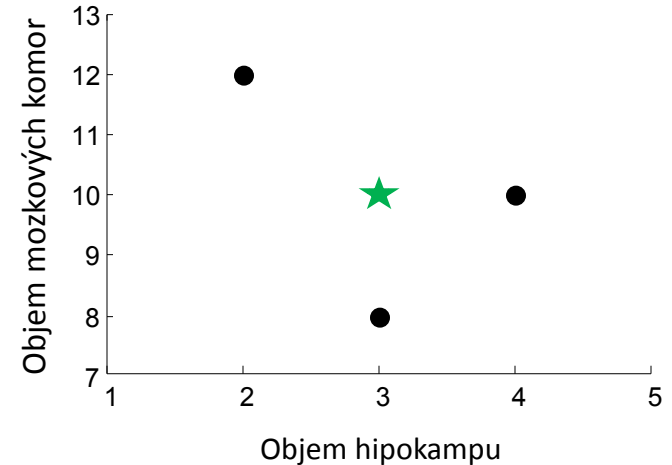
$$\bar{\mathbf{x}} = \left[ \frac{1}{n} \sum_{i=1}^n x_{i1} \quad \frac{1}{n} \sum_{i=1}^n x_{i2} \right]$$

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

$$\mathbf{S} = \begin{bmatrix} s_{11} & s_{12} \\ s_{21} & s_{22} \end{bmatrix}, \text{ kde } s_{11} = \frac{1}{n-1} \sum_{i=1}^n (x_{i1} - \bar{x}_1)^2$$

# Vícerozměrný průměr a kovarianční matice

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



Vícerozměrný průměr:

$$\bar{\mathbf{x}} = \left[ \frac{1}{n} \sum_{i=1}^n x_{i1} \quad \frac{1}{n} \sum_{i=1}^n x_{i2} \right] = \left[ \frac{1}{3} (2 + 4 + 3) \quad \frac{1}{3} (12 + 10 + 8) \right] = [3 \quad 10]$$

Kovarianční matice:  $\mathbf{S} = \begin{bmatrix} s_{11} & s_{12} \\ s_{21} & s_{22} \end{bmatrix}$ , kde:

$$s_{11} = \frac{1}{n-1} \sum_{i=1}^n (x_{i1} - \bar{x}_1)^2 = \frac{1}{3-1} ((2-3)^2 + (4-3)^2 + (3-3)^2) = \frac{1}{2} (1 + 1 + 0) = 1$$

$$s_{22} = \frac{1}{n-1} \sum_{i=1}^n (x_{i2} - \bar{x}_2)^2 = \frac{1}{3-1} ((12-10)^2 + (10-10)^2 + (8-10)^2) = 4$$

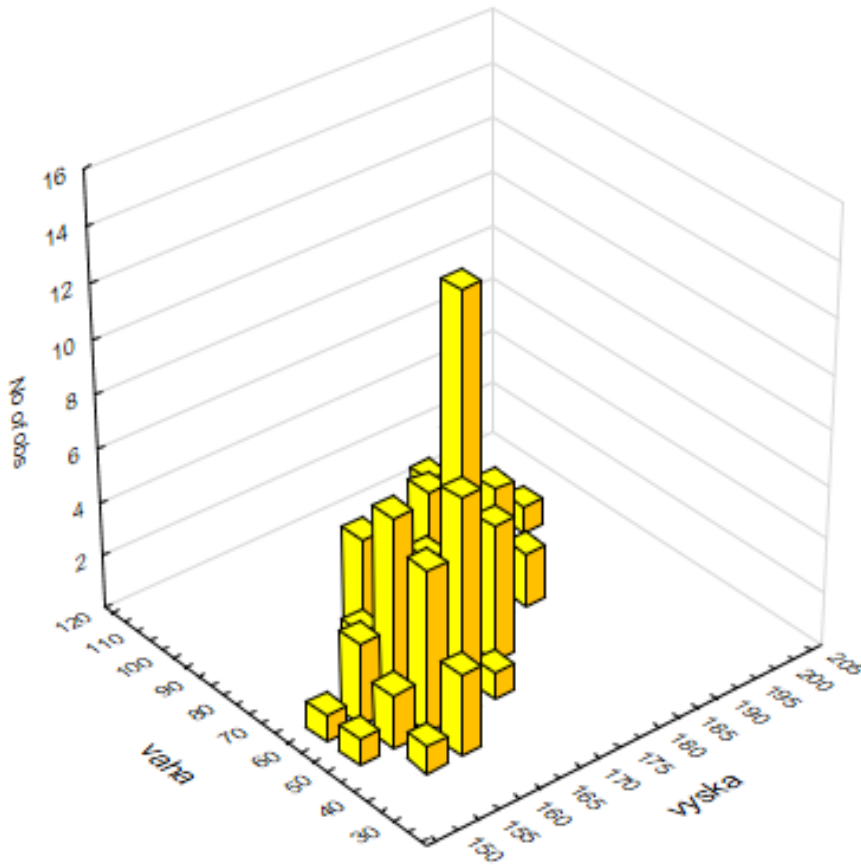
$$s_{21} = s_{12} = \frac{1}{n-1} \sum_{i=1}^n (x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2) = \frac{1}{3-1} ((2-3)(12-10) + (4-3)(10-10) + (3-3)(8-10)) = -1 \quad \rightarrow \mathbf{S} = \begin{bmatrix} 1 & -1 \\ -1 & 4 \end{bmatrix}$$



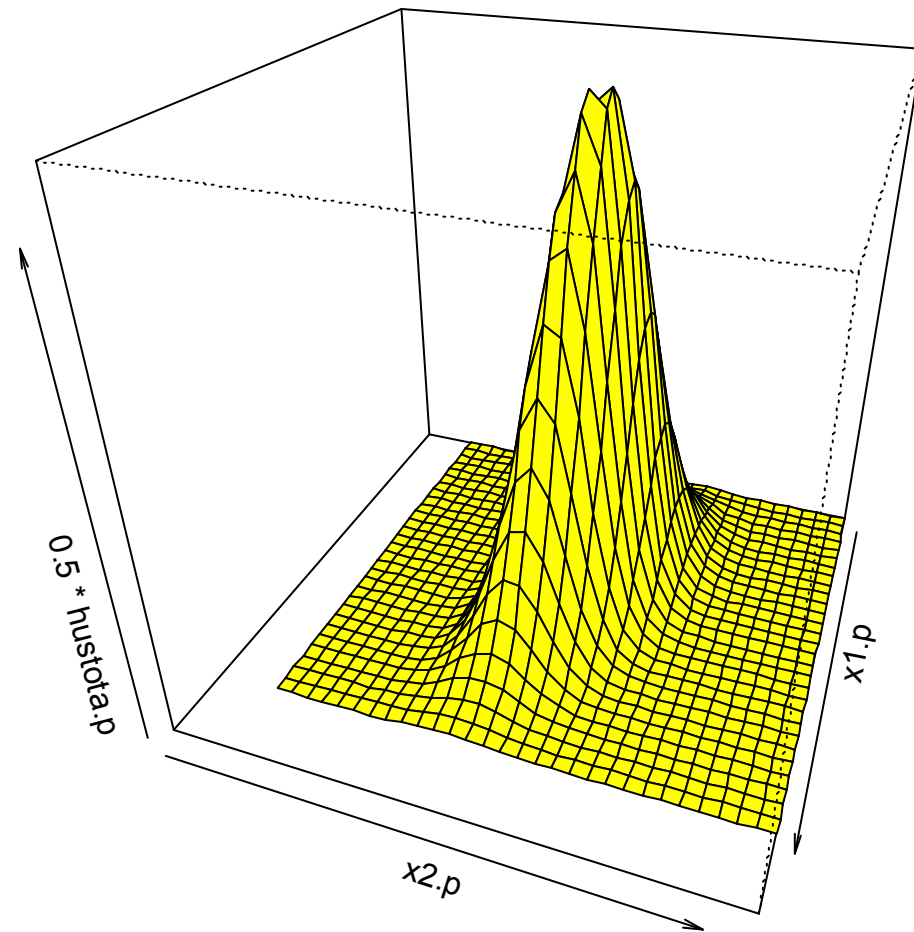
# Vícerozměrné normální rozdělení

# Motivace

Dvourozměrný histogram



Hustota dvourozměrného normálního rozdělení



# Vícerozměrné normální rozdělení

**Hustota jednozměrného normálního rozdělení:**

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

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

**Hustota vícerozměrného normálního rozdělení:**

$$f(x_1, \dots, x_k) = \frac{1}{\sqrt{(2\pi)^k |\Sigma|}} \cdot \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu})\right)$$

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

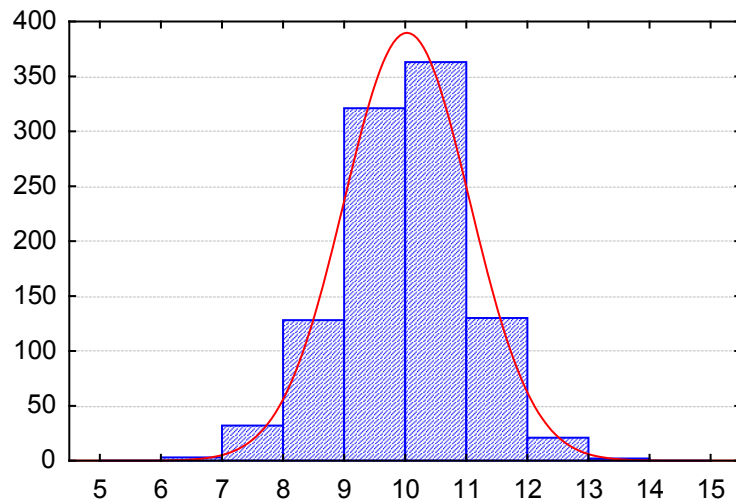
**Hustota dvourozměrného normálního rozdělení:**

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

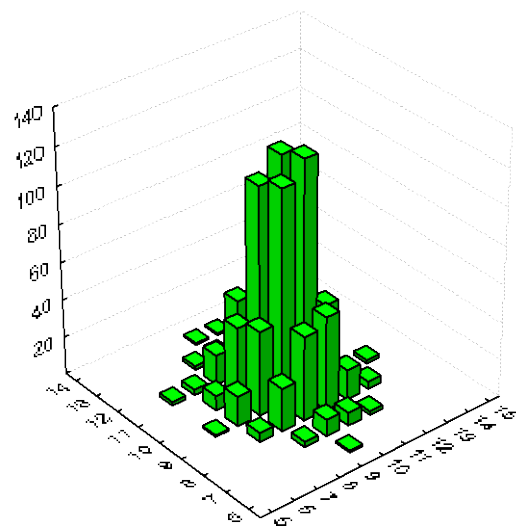
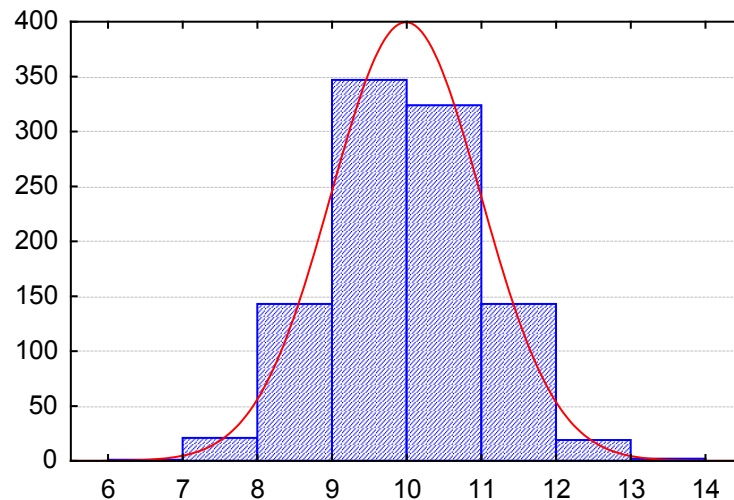
$$\boldsymbol{\mu} = \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix}$$

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

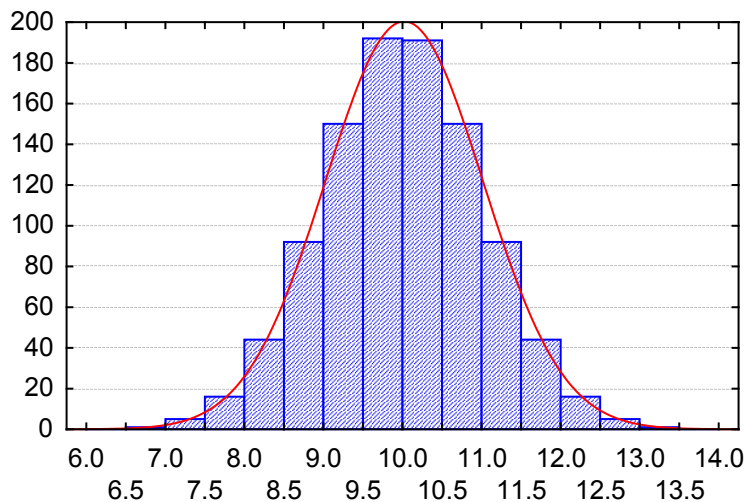
# Je normalita v jednorozměrném prostoru jedinou podmínkou vícerozměrné normality?



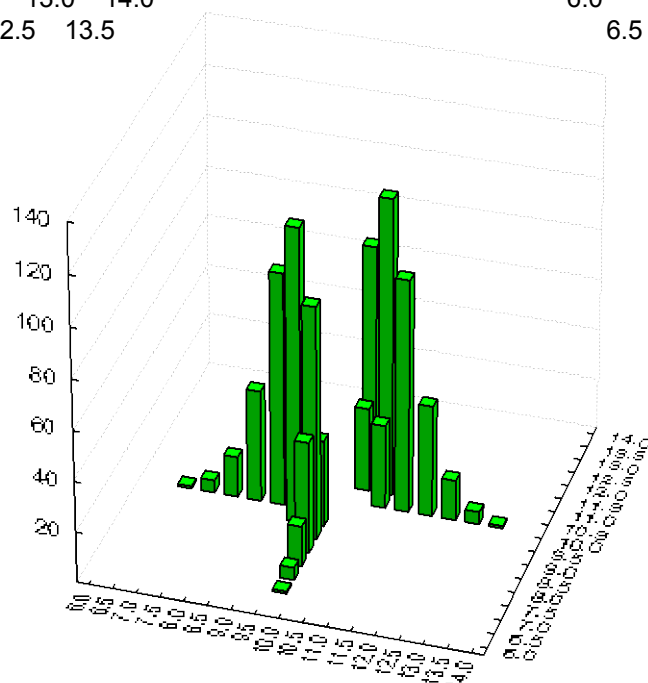
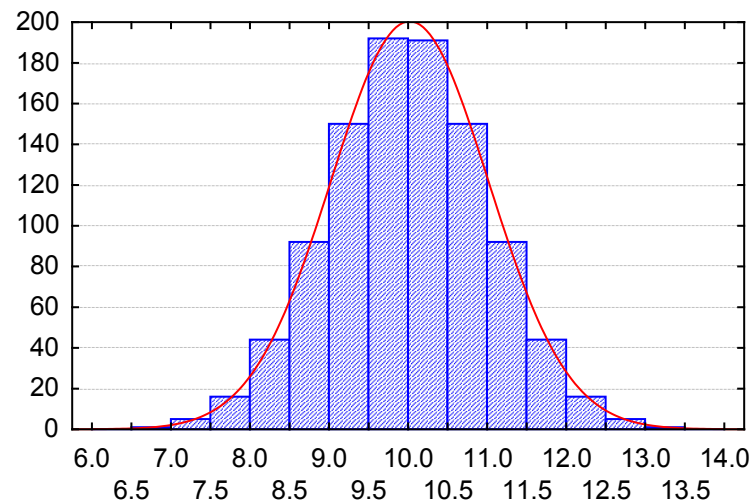
+



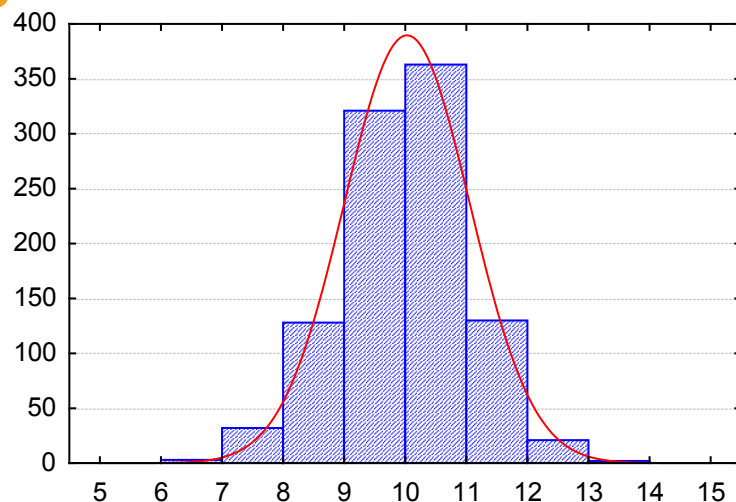
# Je normalita v jednorozměrném prostoru jedinou podmínkou vícerozměrné normality?



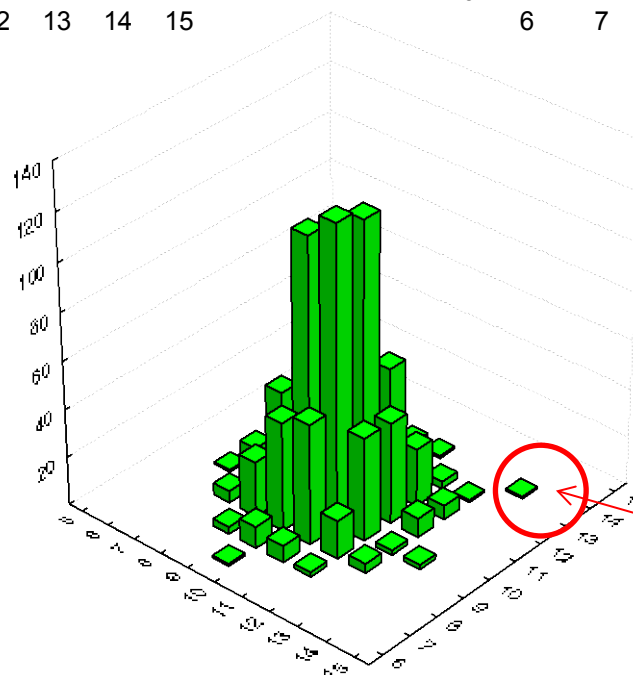
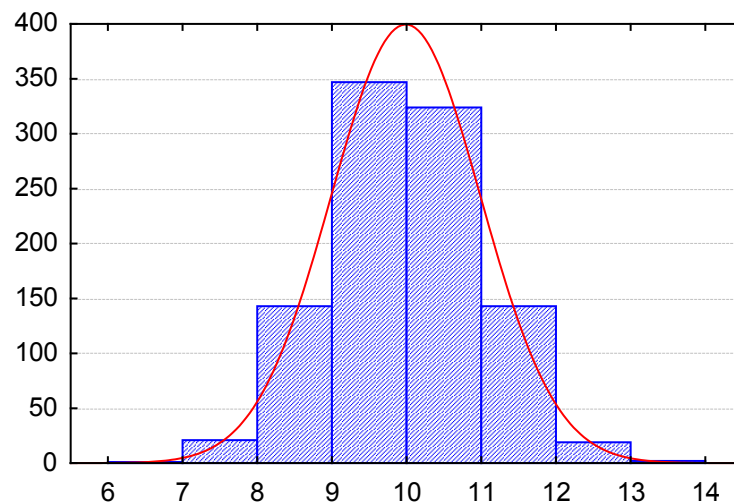
+



# Je normalita v jednorozměrném prostoru jedinou podmínkou vícerozměrné normality?



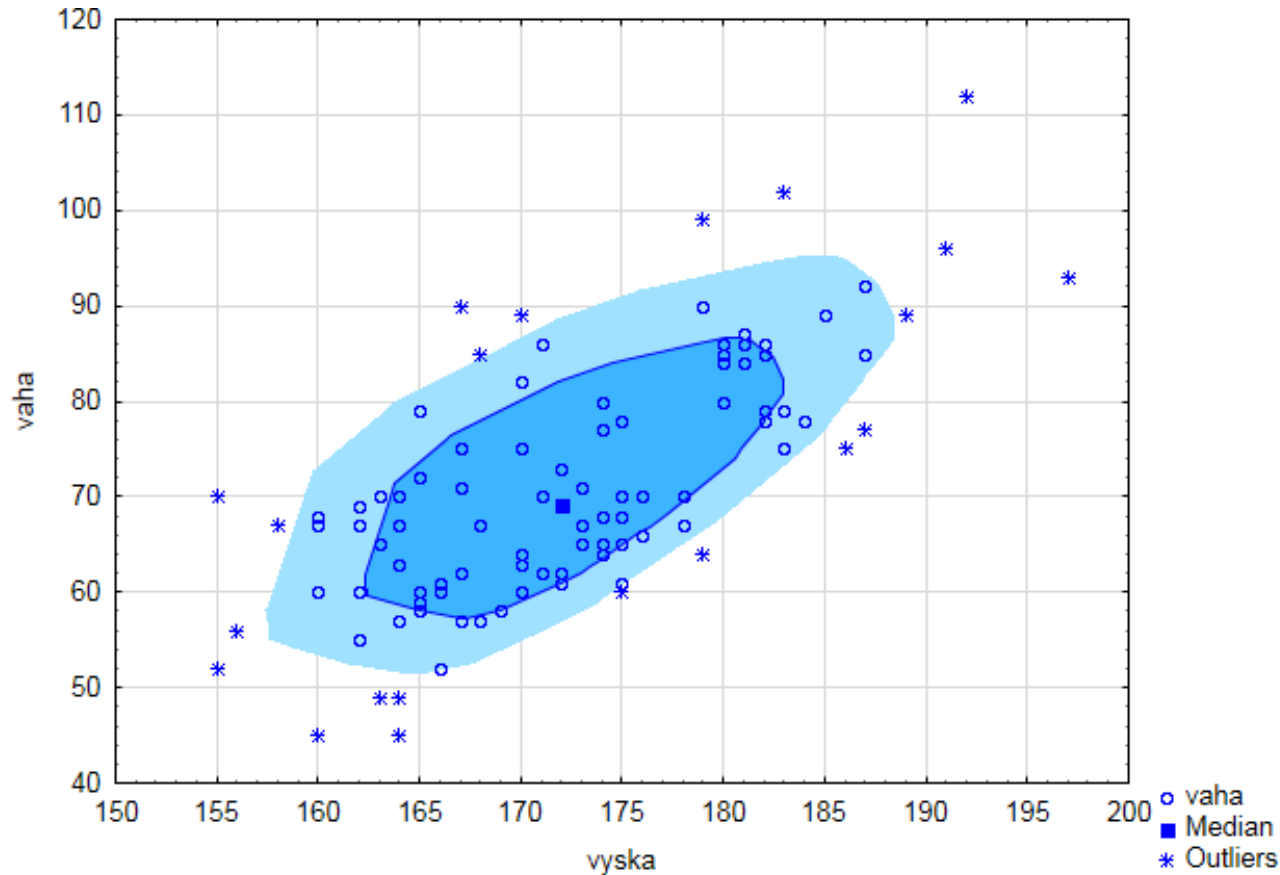
+



Vícerozměrný  
outlier

# Ověření dvourozměrné normality

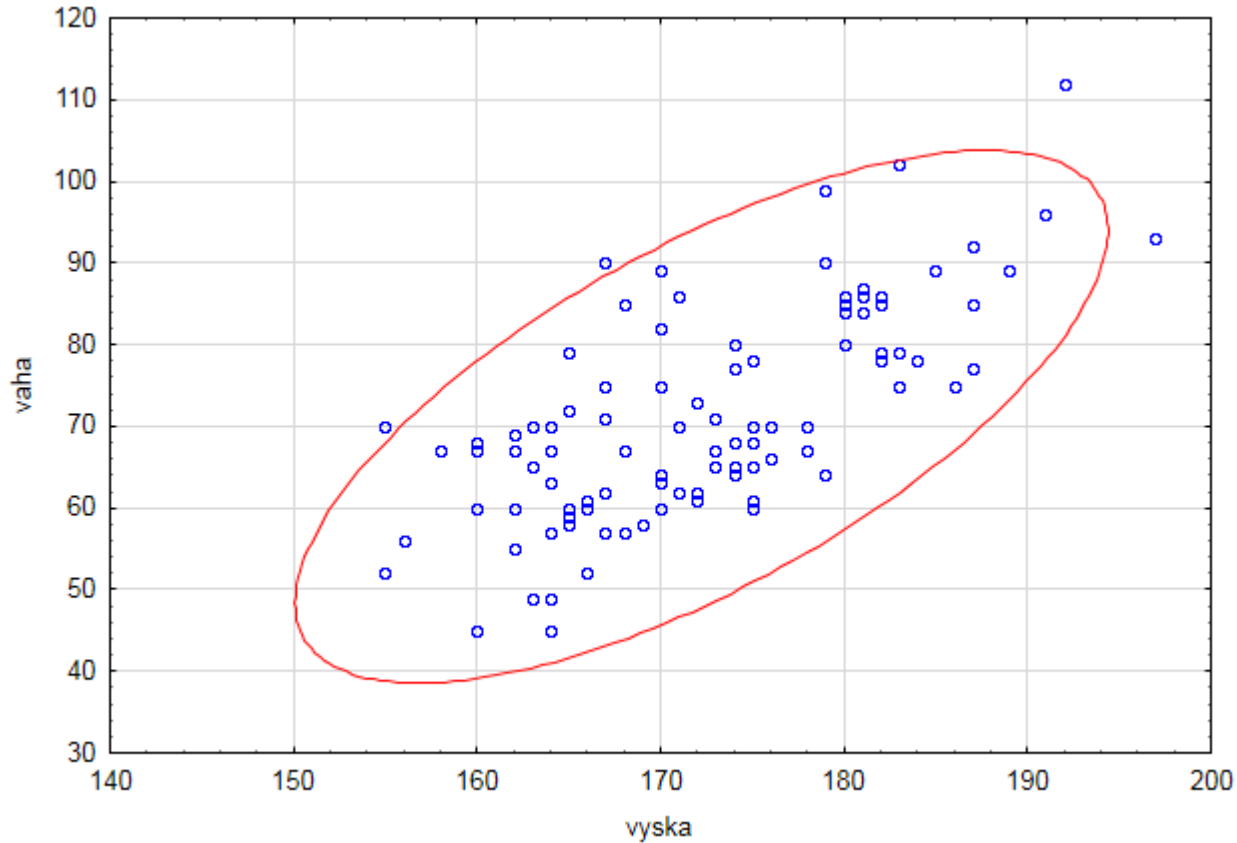
Bagplot = „bivariate boxplot“ (tzn. „dvourozměrný krabicový graf“)



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

# Ověření dvourozměrné normality

Vykreslení regulační elipsy („control“ ellipse):



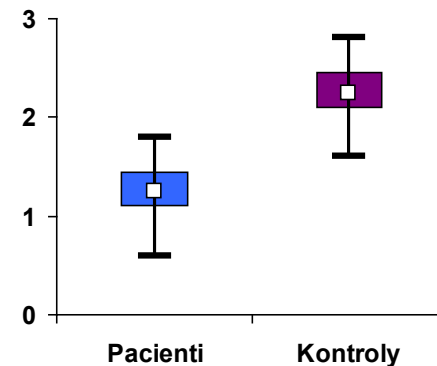
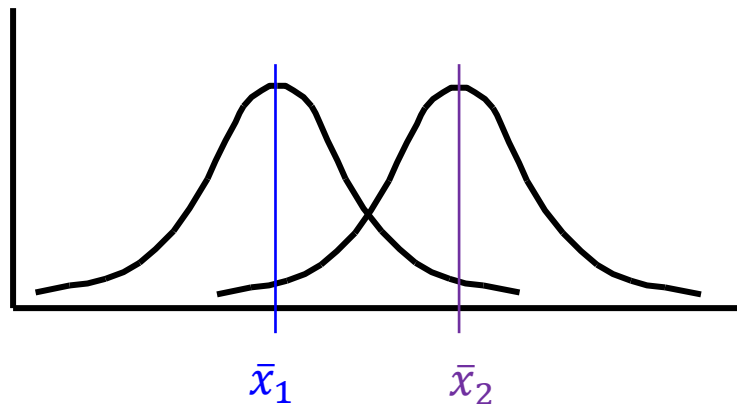
v softwaru Statistica: Graphs – Scatterplots – na záložce Advanced zvolit Elipse Normal



# Vícerozměrný t-test

# Jednorozměrný dvouvýběrový t-test

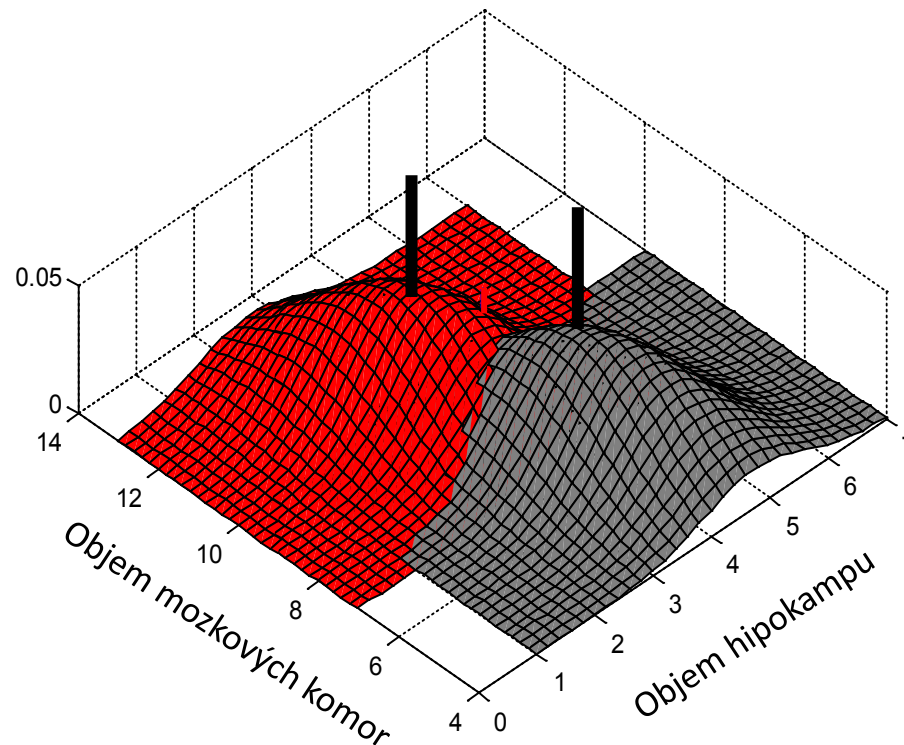
- Srovnáváme dvě skupiny dat, které jsou na sobě nezávislé – mezi objekty neexistuje vazba.
- Příklady: srovnání objemu hipokampu u mužů a u žen, srovnání kognitivního výkonu podle dvou kategorií věku,...



- 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:  $T = \frac{(\bar{x}_D - \bar{x}_H) - c}{s_* \sqrt{\frac{1}{n_D} + \frac{1}{n_H}}}$ , kde  $T \sim t(n_D + n_H - 2)$  ← Studentovo rozdělení
- $s_*^2$  je vážený rozptyl vypočtený jako  $s_*^2 = \frac{(n_D - 1)s_D^2 + (n_H - 1)s_H^2}{(n_D - 1) + (n_H - 1)}$
- $c$  je konstanta, o kterou se rozdíl průměrů má lišit (většinou  $c = 0$ )
- nulová hypotéza zamítnuta, pokud  $|T| > t_{1-\alpha/2}(n_D + n_H - 2)$

## Je ekvivalentní testu:

$$T^2 = \left( \frac{(\bar{x}_D - \bar{x}_H) - c}{s_* \sqrt{\frac{1}{n_D} + \frac{1}{n_H}}} \right)^2 = (\bar{x}_D - \bar{x}_H - c) \left[ s_*^2 \left( \frac{1}{n_D} + \frac{1}{n_H} \right) \right]^{-1} (\bar{x}_D - \bar{x}_H - c), \text{ kde } T^2 \sim F(1, n_D + n_H - 2)$$
← F rozdělení

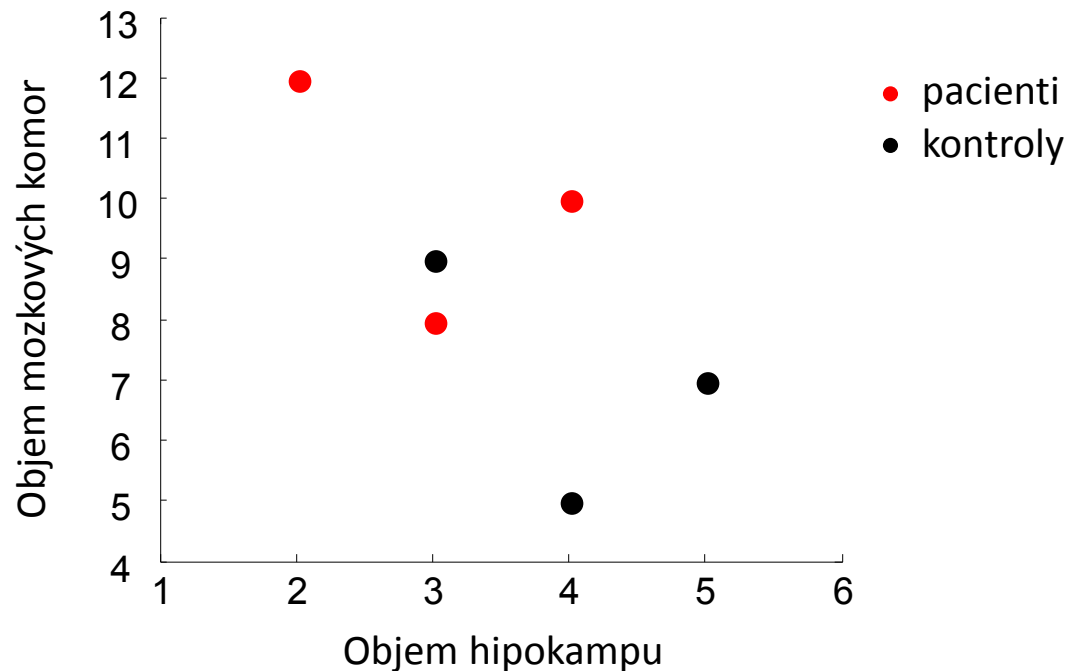
## Vícerozměrný t-test:

- Hotellingova  $T^2$  testová statistika:  $T^2 = (\bar{\mathbf{x}}_D - \bar{\mathbf{x}}_H - \mathbf{c})^T \left[ \mathbf{S}_* \left( \frac{1}{n_D} + \frac{1}{n_H} \right) \right]^{-1} (\bar{\mathbf{x}}_D - \bar{\mathbf{x}}_H - \mathbf{c})$
- kde  $\mathbf{S}_*$  je vážená kovarianční matice:  $\mathbf{S}_* = \frac{(n_D - 1)\mathbf{S}_D + (n_H - 1)\mathbf{S}_H}{(n_D - 1) + (n_H - 1)}$
- $T^2 \sim T^2(p, n - p - 1)$ ; pro malé  $n_D$  a  $n_H$  je lepší použít:  $F = \frac{n - p - 1}{p} \frac{T^2}{n - 2}$ , kde  $n = n_D + n_H$  ← Hotellingovo rozdělení
- nulová hypotéza zamítnuta, když  $F > F_{1-\alpha}(p, n - p - 1)$  ← F rozdělení

# Úkol 1

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

$$\mathbf{X}_D = \begin{bmatrix} 2 & 12 \\ 4 & 10 \\ 3 & 8 \end{bmatrix}, \mathbf{X}_H = \begin{bmatrix} 5 & 7 \\ 3 & 9 \\ 4 & 5 \end{bmatrix}$$



# Úkol 1 - řešení

Vícerozměrné průměry:

$$\bar{\mathbf{x}}_D = \left[ \frac{1}{n_D} \sum_{i=1}^{n_D} x_{i1} \quad \frac{1}{n_D} \sum_{i=1}^{n_D} x_{i2} \right] = [3 \quad 10]$$

$$\bar{\mathbf{x}}_H = \left[ \frac{1}{n_H} \sum_{i=1}^{n_H} x_{i1} \quad \frac{1}{n_H} \sum_{i=1}^{n_H} x_{i2} \right] = [4 \quad 7]$$

Výběrové kovarianční matice:

$$\mathbf{S}_D = \begin{bmatrix} s_{11}^D & s_{12}^D \\ s_{21}^D & s_{22}^D \end{bmatrix} = \begin{bmatrix} 1 & -1 \\ -1 & 4 \end{bmatrix}$$

$$\mathbf{S}_H = \begin{bmatrix} s_{11}^H & s_{12}^H \\ s_{21}^H & s_{22}^H \end{bmatrix} = \begin{bmatrix} 1 & -1 \\ -1 & 4 \end{bmatrix}$$

Vážená kovarianční matice:

$$\mathbf{S}_* = \begin{bmatrix} 1 & -1 \\ -1 & 4 \end{bmatrix}$$

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

$$T^2 = (\bar{\mathbf{x}}_D - \bar{\mathbf{x}}_H - \mathbf{c})^T \left[ \mathbf{S}_* \left( \frac{1}{n_D} + \frac{1}{n_H} \right) \right]^{-1} (\bar{\mathbf{x}}_D - \bar{\mathbf{x}}_H - \mathbf{c})$$

$$F = \frac{n - p - 1}{p} \frac{T^2}{n - 2}$$

# 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)
```

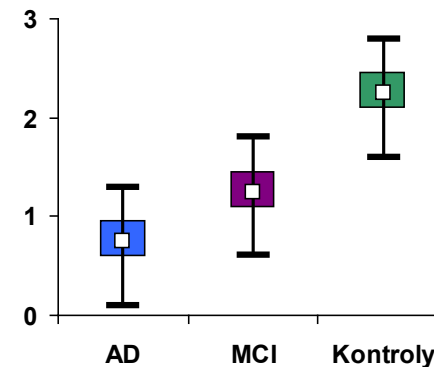
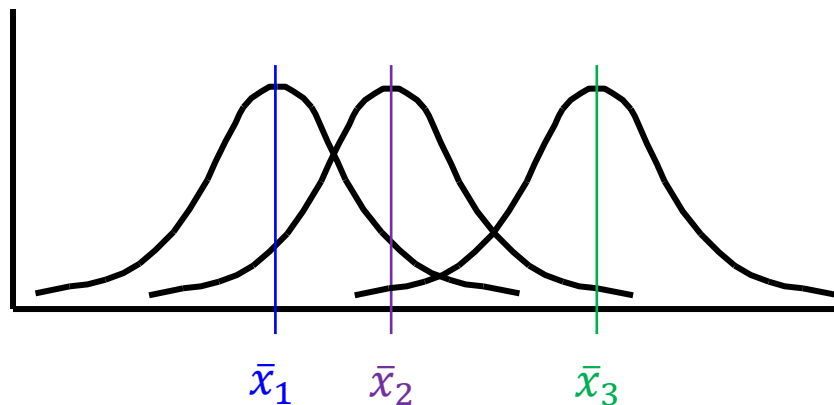
```
HotellingsT2(X, Y)
```

# Analýza rozptylu pro vícerozměrná data



# Analýza rozptylu (ANOVA) jednoduchého třídění

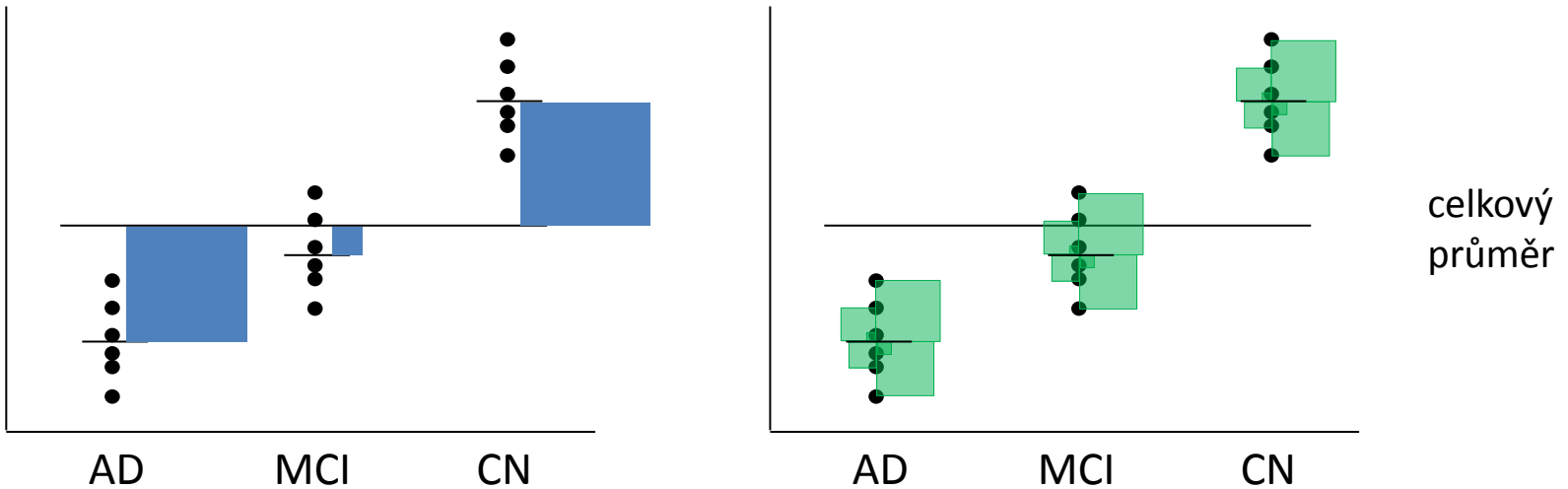
- Srovnáváme tři a více skupin dat, které jsou na sobě nezávislé (mezi objekty neexistuje vazba).
- Příklady: srovnání objemu hipokampu u pacientů s AD, pacientů s MCI a kontrol; srovnání kognitivního výkonu podle čtyř kategorií věku.



- 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 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

- Uvažujeme dvě vysvětlující proměnné zároveň.
- Zápis modelu:

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

- 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:

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

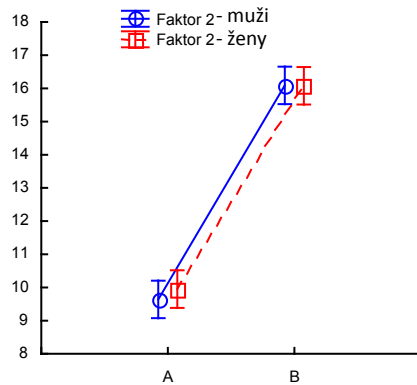
↑ Populační průměr (pointing to  $\mu$ )  
↑  $i$ -tý efekt faktoru A (pointing to  $\alpha_i$ )  
↑  $j$ -tý efekt faktoru B (pointing to  $\beta_j$ )  
↑ Interakce (pointing to  $\gamma_{ij}$ )  
← Reziduum (pointing to  $e_{ij}$ )

- Nulové hypotézy pak máme tři:

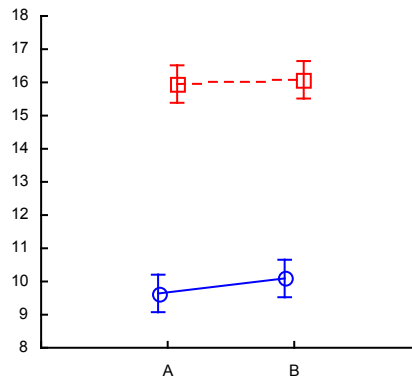
$$H_{01} : \gamma_{11} = \gamma_{12} = \dots = \gamma_{kr} \quad H_{02} : \alpha_1 = \alpha_2 = \dots = \alpha_k \quad H_{03} : \beta_1 = \beta_2 = \dots = \beta_r$$

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	$S_T$	$df_T = n - 1$			

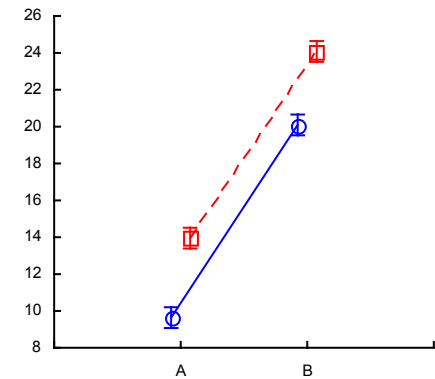
# Hlavní efekty a interakce



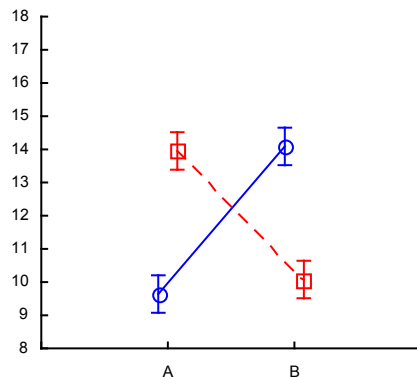
	SS	D.f.	MS	F	p
<b>Faktor 1</b>	<b>1978</b>	<b>1</b>	<b>1978</b>	<b>482.2</b>	<b>0.000</b>
Faktor 2	1	1	1	0.3	0.602
F1*F2	1	1	1	0.3	0.570
Error	804	196	4		



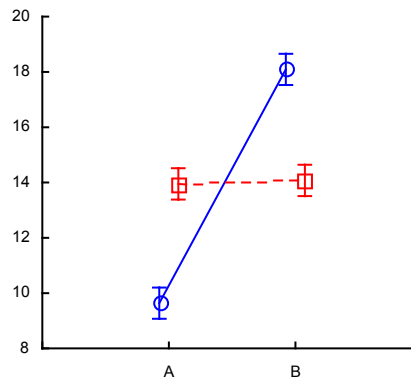
	SS	D.f.	MS	F	p
Faktor 1	4	1	4	1.0	0.314
<b>Faktor 2</b>	<b>1891</b>	<b>1</b>	<b>1891</b>	<b>461.1</b>	<b>0.000</b>
F1*F2	1	1	1	0.3	0.570
Error	804	196	4		



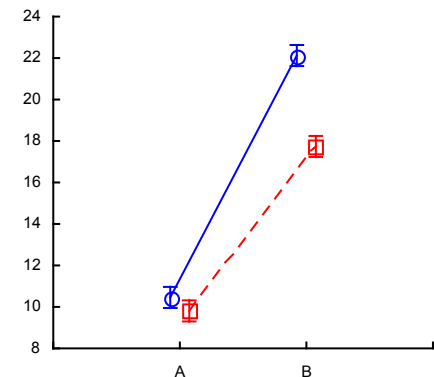
	SS	D.f.	MS	F	p
<b>Faktor 1</b>	<b>5293</b>	<b>1</b>	<b>5293</b>	<b>1290.7</b>	<b>0.000</b>
<b>Faktor 2</b>	<b>861</b>	<b>1</b>	<b>861</b>	<b>209.9</b>	<b>0.000</b>
F1*F2	1	1	1	0.3	0.570
Error	804	196	4		



	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		



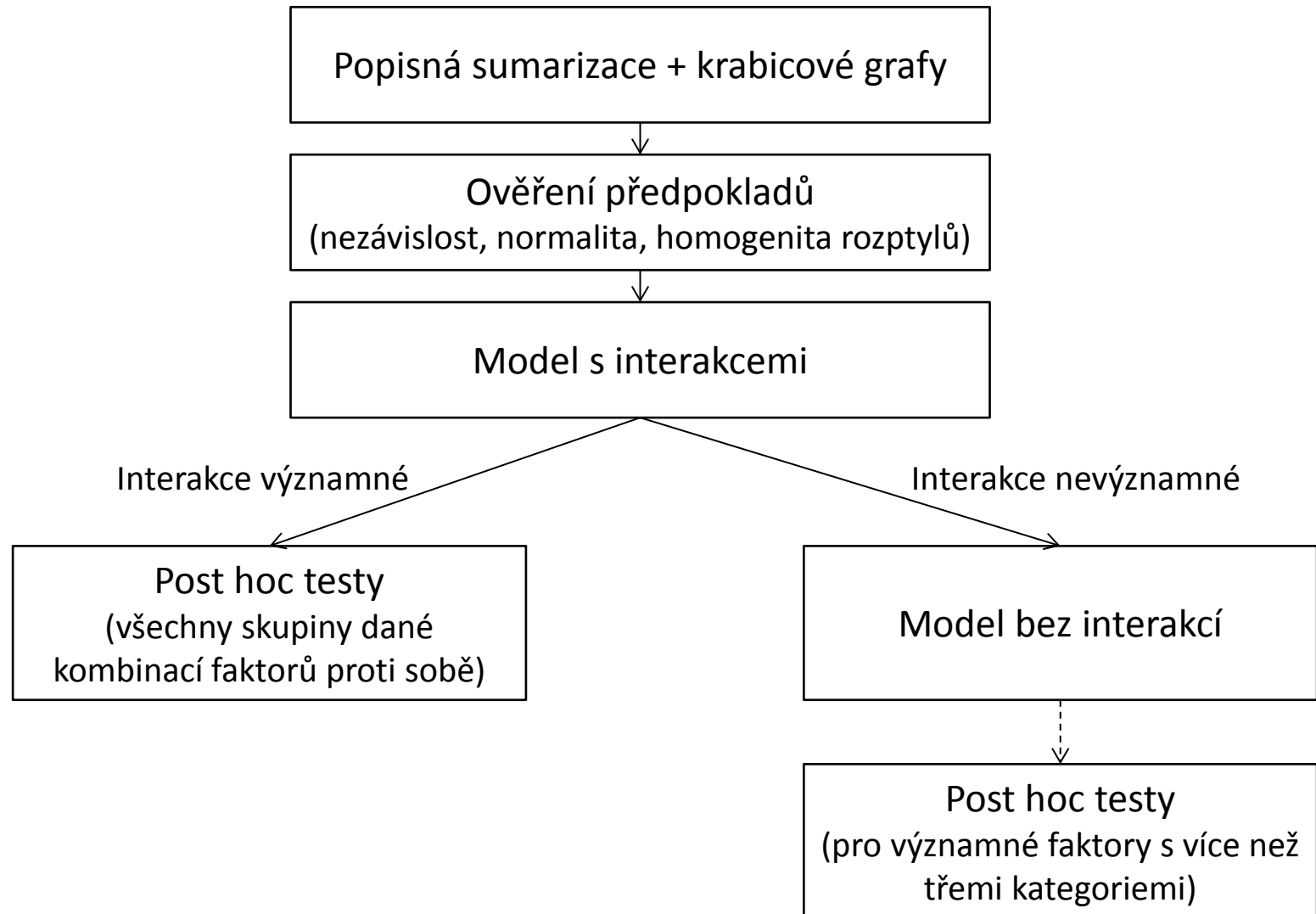
	SS	D.f.	MS	F	p
<b>Faktor 1</b>	<b>920</b>	<b>1</b>	<b>920</b>	<b>224.3</b>	<b>0.000</b>
Faktor 2	1	1	1	0.3	0.602
<b>F1*F2</b>	<b>867</b>	<b>1</b>	<b>867</b>	<b>211.3</b>	<b>0.000</b>
Error	804	196	4		



	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 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 2 – 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 2 – 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 2 – 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 2 – 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 – 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 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 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 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 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)

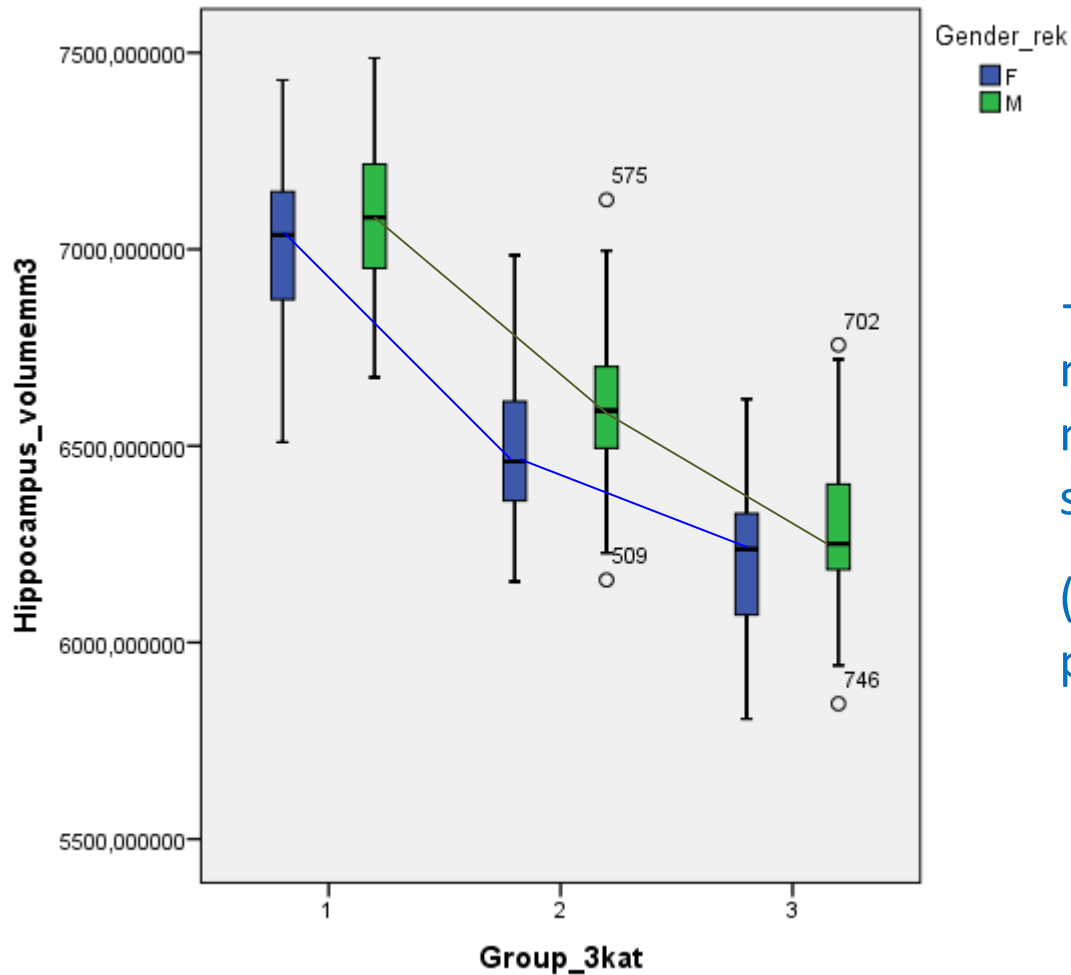
2...MCI (mírná kognitivní porucha)

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

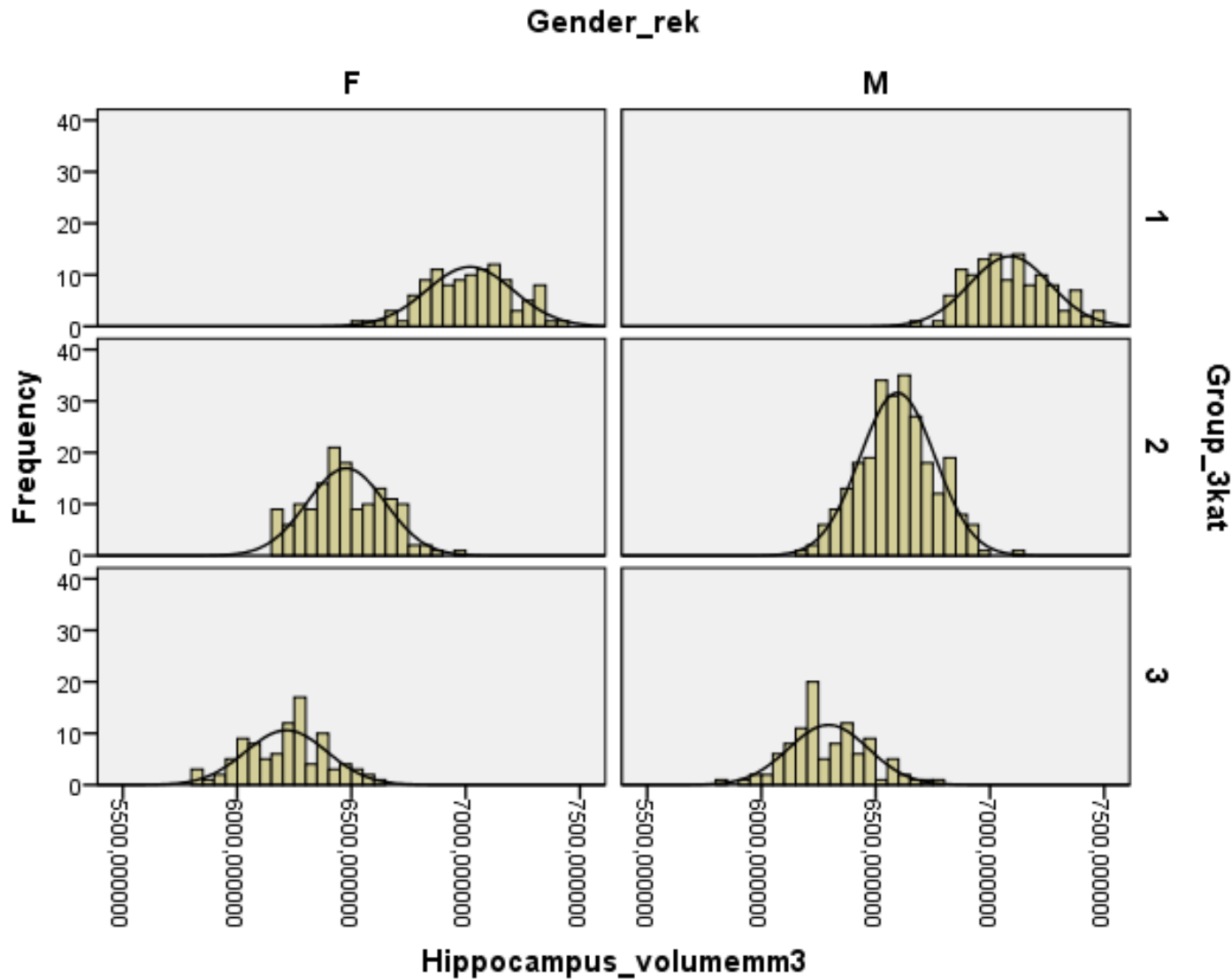
# Úkol 3 – krabicový graf



→ 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

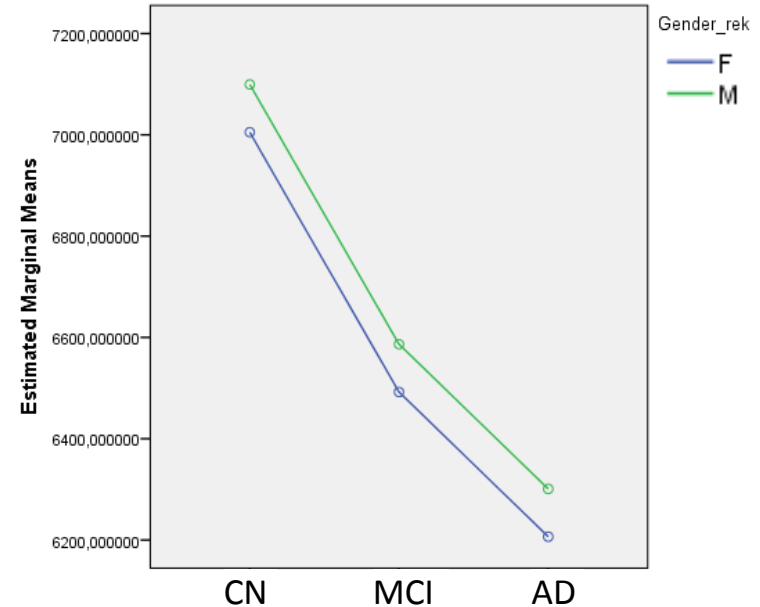
Hippocampus\_volumemm3

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

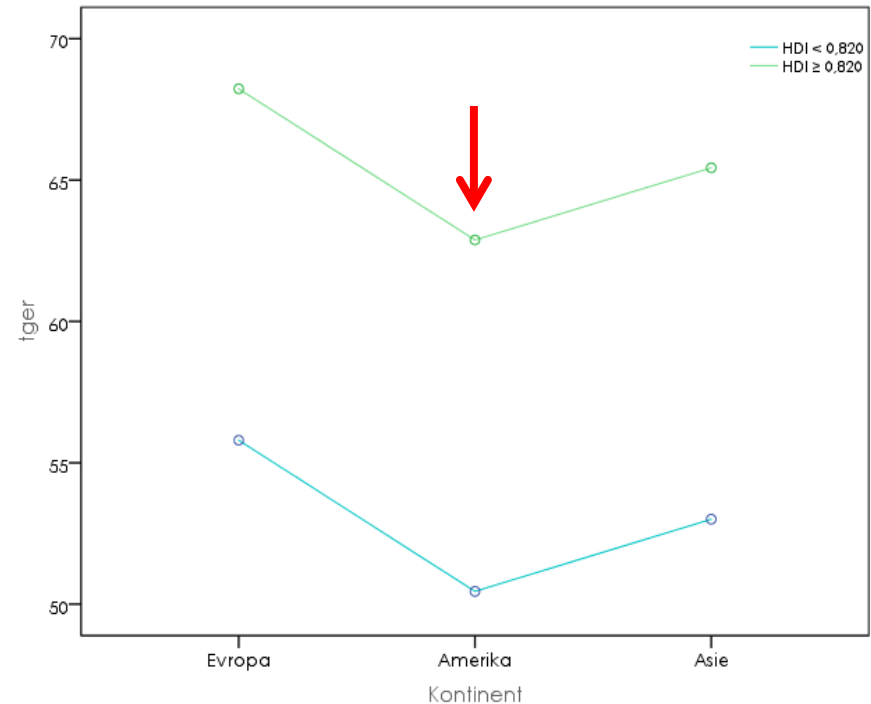
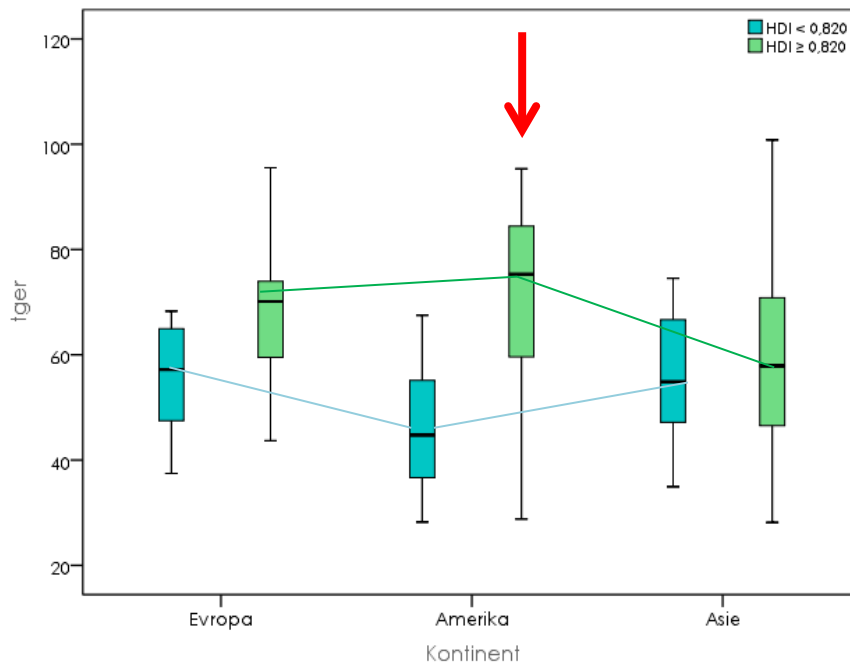


- statisticky významný vliv pohlaví i typu onemocnění na objem hipokampu, přičemž mezi pohlavím a typem onemocnění nenastává interakce
- u mužů statisticky významně vyšší objem hipokampu než u žen
- statisticky významný rozdíl v objemu hipokampu u všech 3 skupin subjektů podle typu onemocnění, přičemž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



# Upozornění I

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



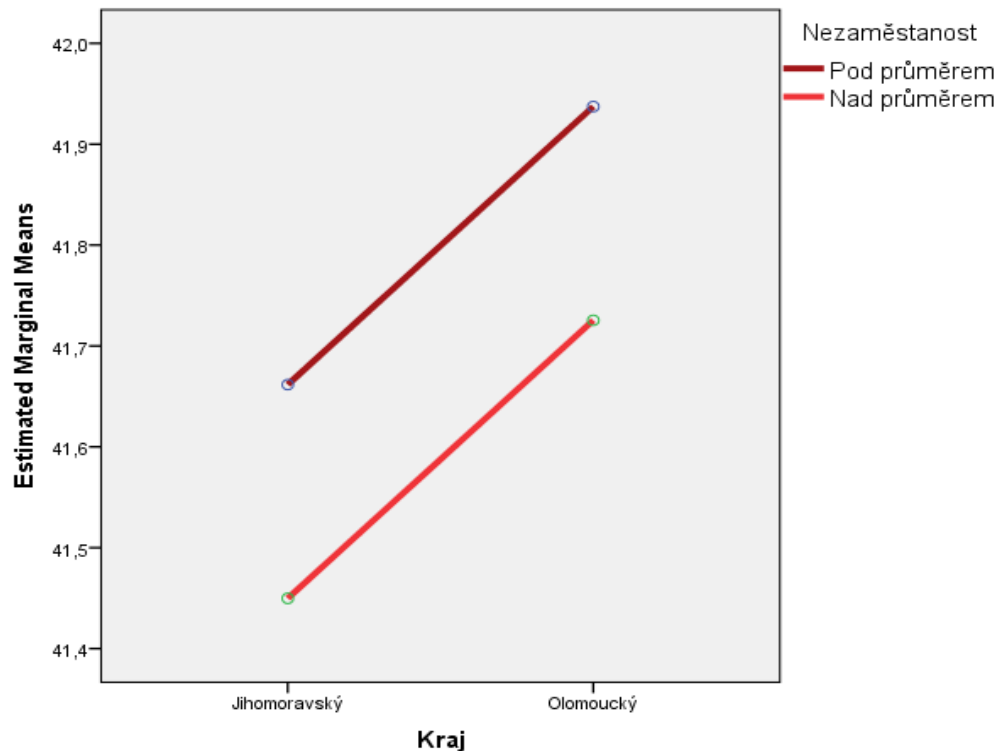
- znamená to, že tam zřejmě není splněn předpoklad normality
- pokud rozdíl není statisticky významný, není zpravidla potřeba to řešit
- pokud by ten rozdíl vyšel statisticky významně, je to problém!
- poznámka: je dobré mít měřítko na ose y stejné u obou grafů

# 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!)



# Transformace a jiné úpravy vícerozměrných dat

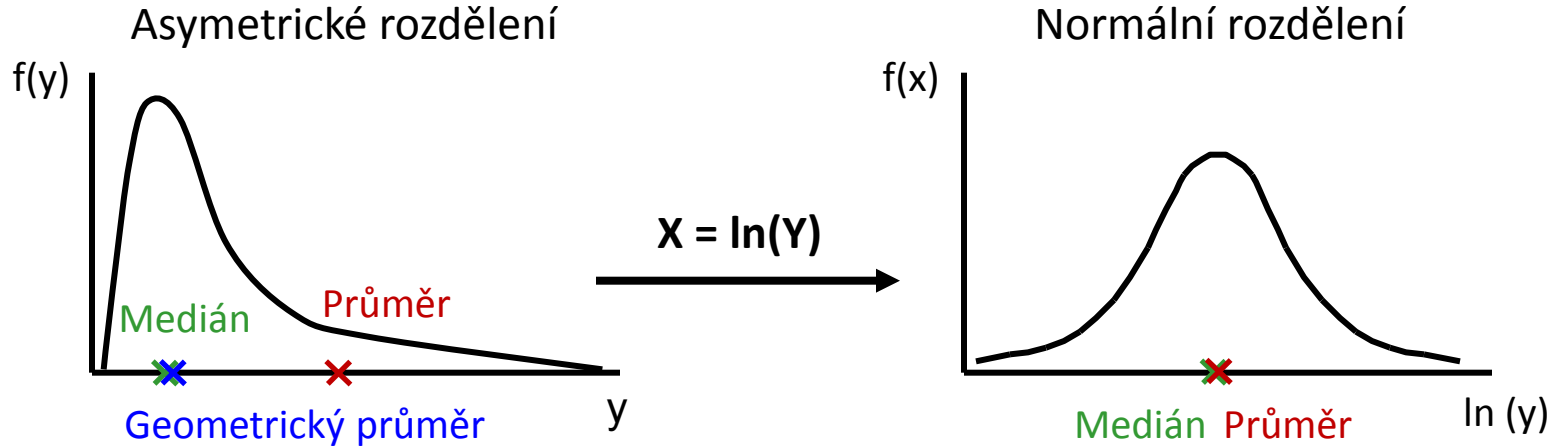
# Typy transformací a jiných úprav vícerozm. dat

---

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

# Normalizace dat

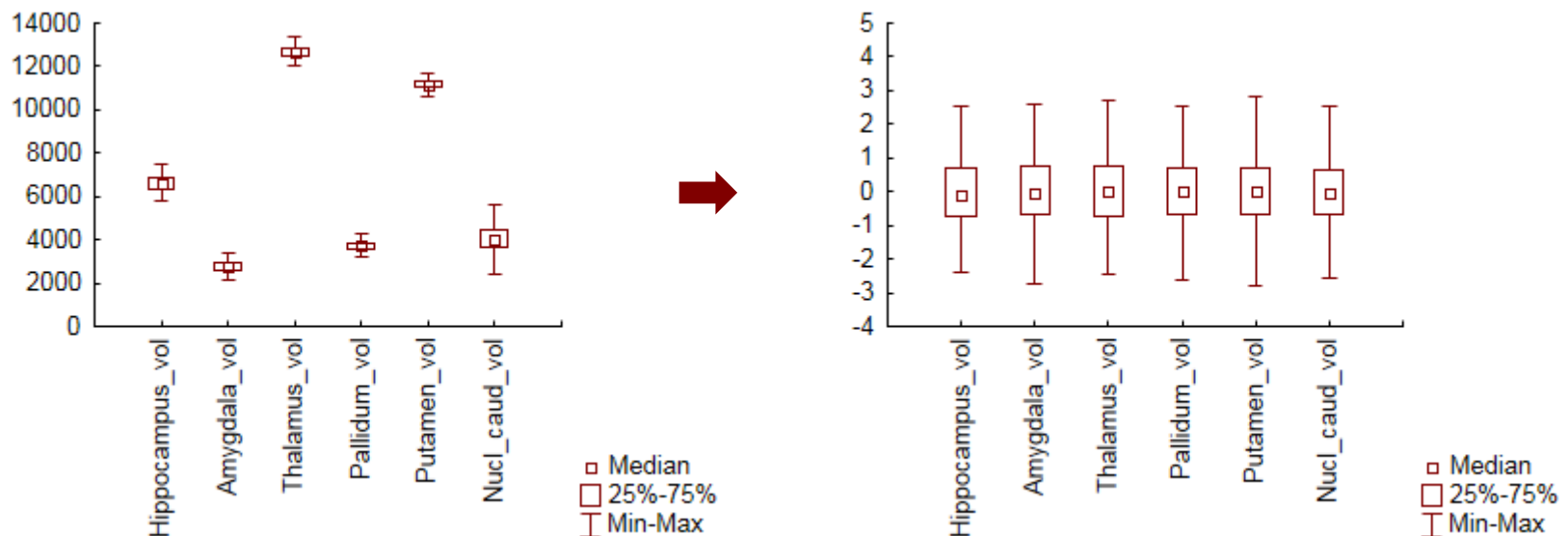
- převod na normální rozdělení (normalita je předpokladem řady statistických testů).
- např. **logaritmická transformace**:  $X = \ln(Y)$  nebo  $X = \ln(Y+1)$ , pokud data obsahují hodnotu 0



- 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**

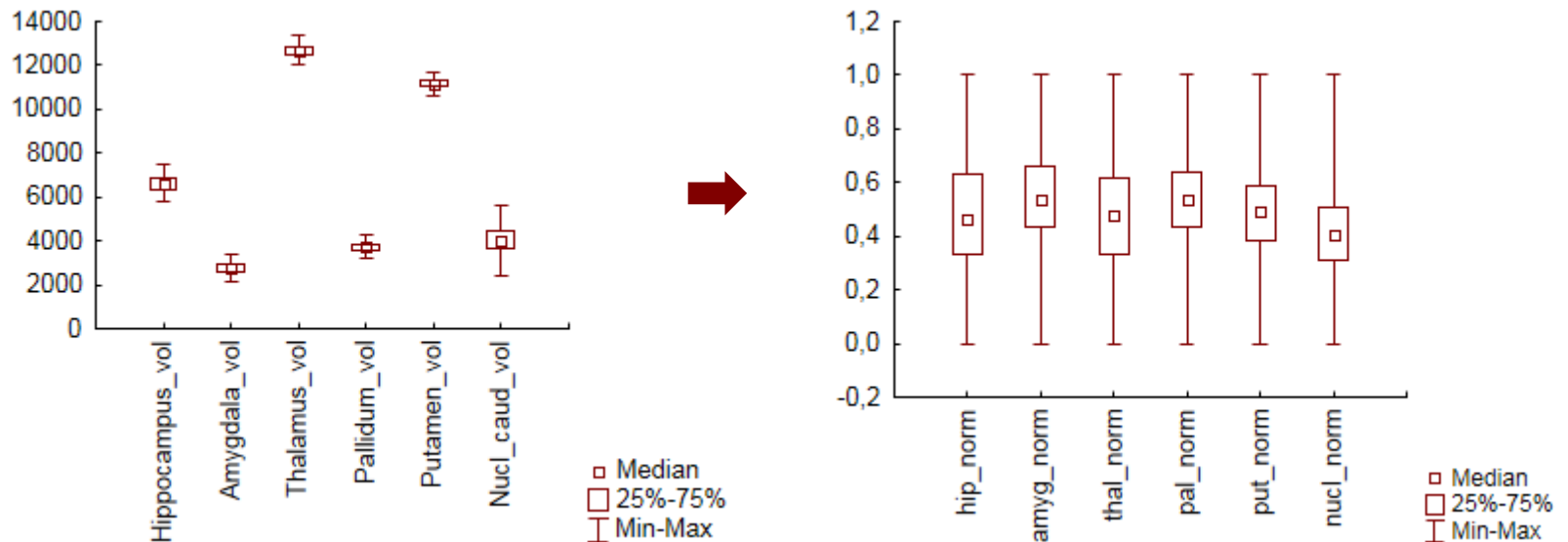
# Standardizace dat

- důvod: převod proměnných na stejné měřítko
- standardizace:  $z_i = \frac{x_i - \bar{x}}{s}$  (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!!!**



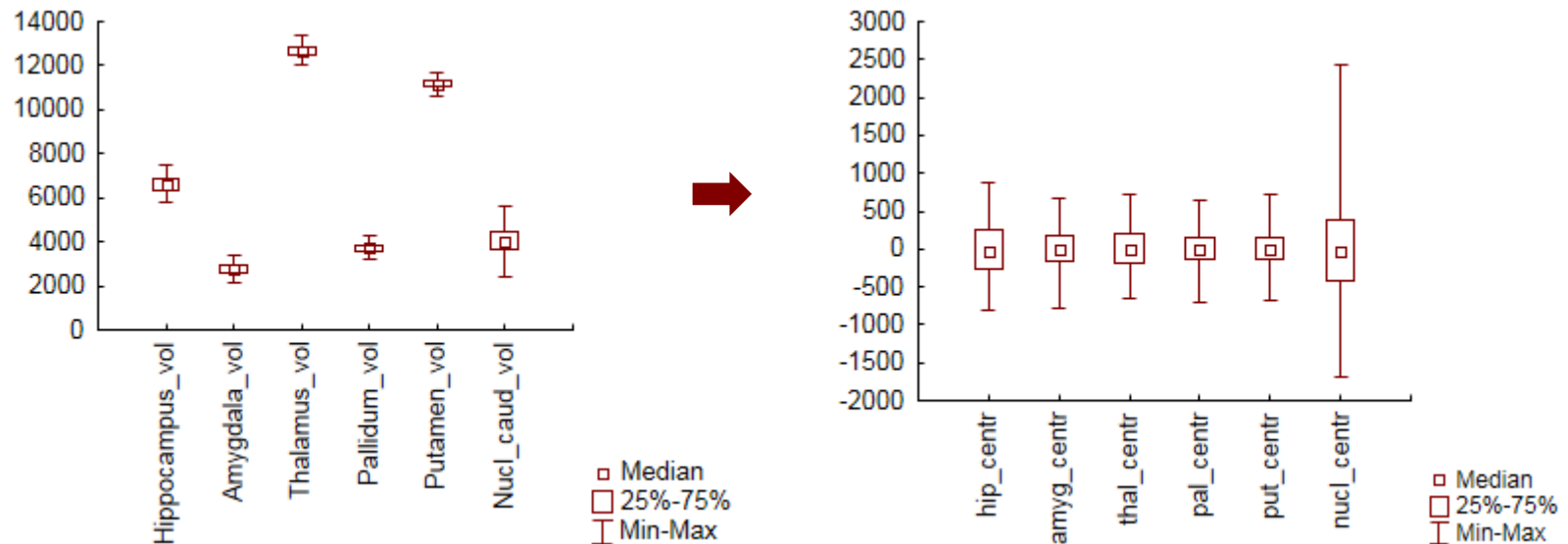
# Min-max normalizace

- důvod: převod proměnných na stejné měřítko
- oproti standardizaci vhodná i na proměnné nemající normální rozdělení či obsahující odlehlé hodnoty
- min-max normalizace:  $y_i = \frac{x_i - \min(x)}{\max(x) - \min(x)}$
- 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:  $z_i = x_i - \bar{x}$

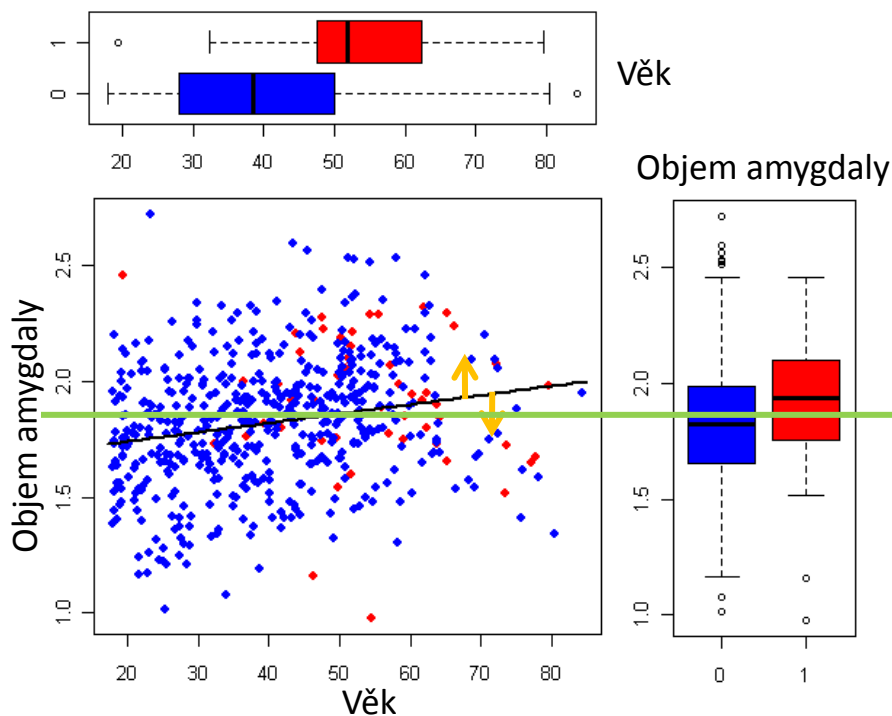




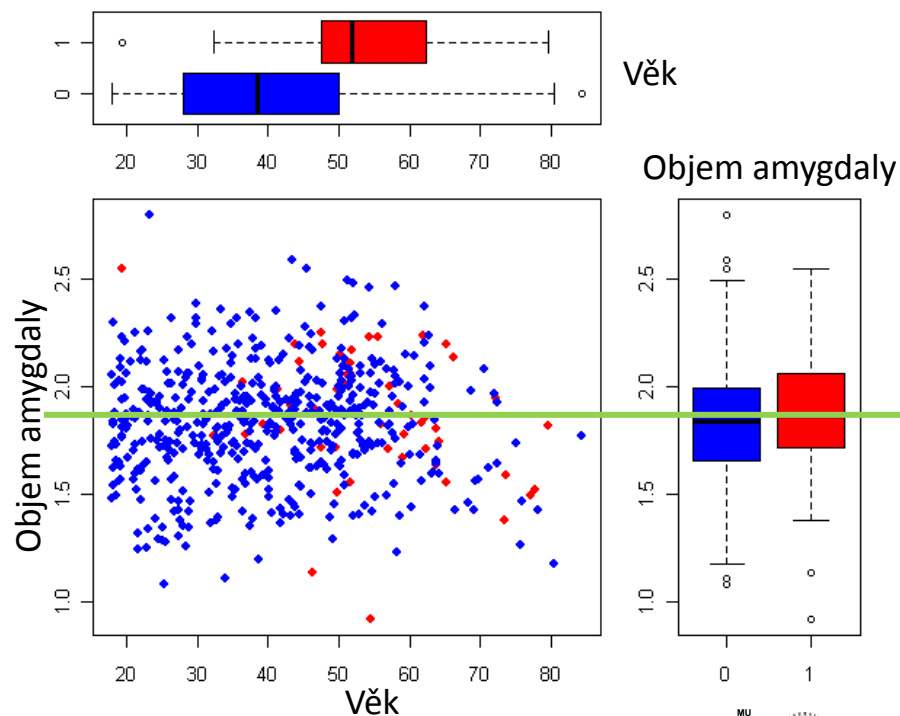
# Odstranění vlivu kovariát (tzv. adjustace)

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“

