

Pokročilé metody analýzy dat v neurovědách



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

Jaro 2015

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	
...						



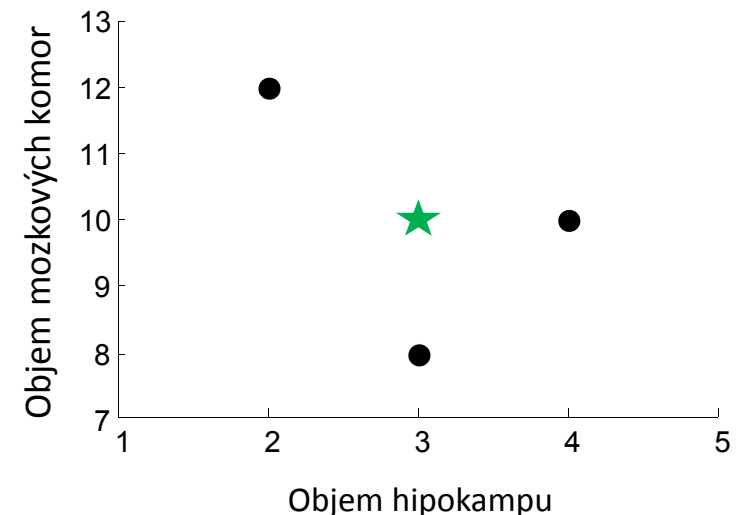
$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

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

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:

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

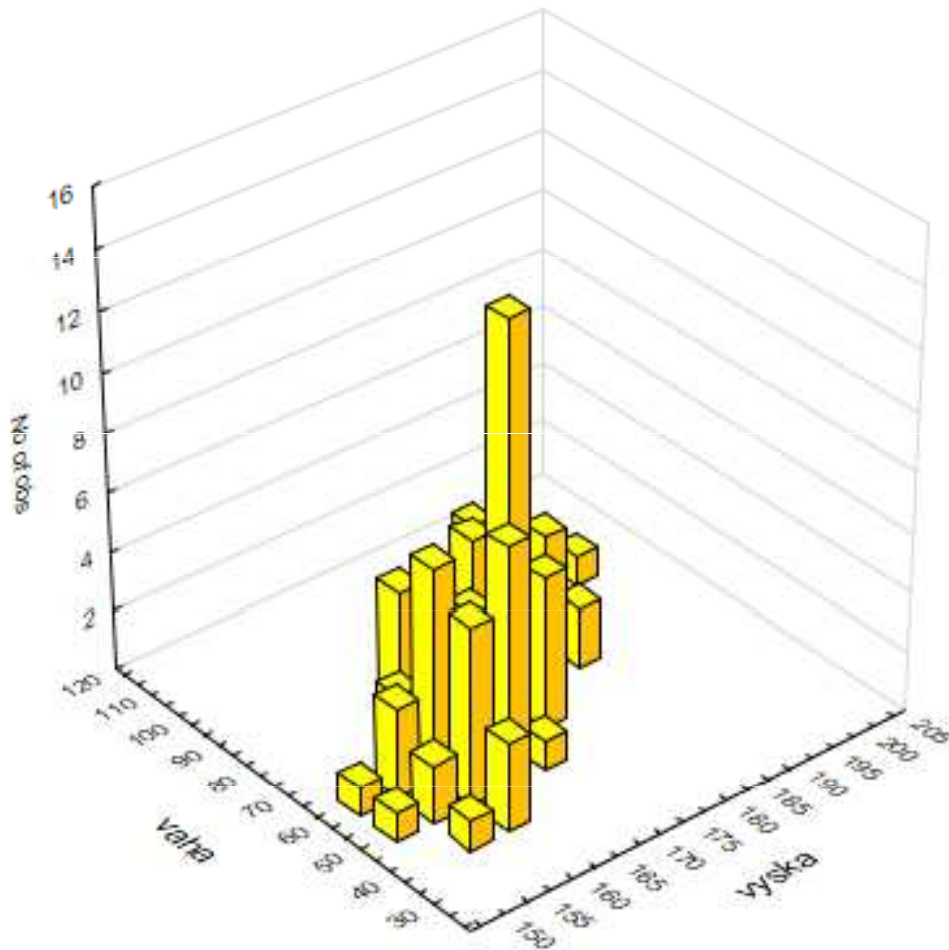
$$s_{22} = \frac{1}{n-1} \sum_{i=1}^n (x_{i2} - \bar{x}_2)^2 = \frac{1}{3-1} ((12-10)^2 + (10-10)^2 + (8-10)^2) = 4$$

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

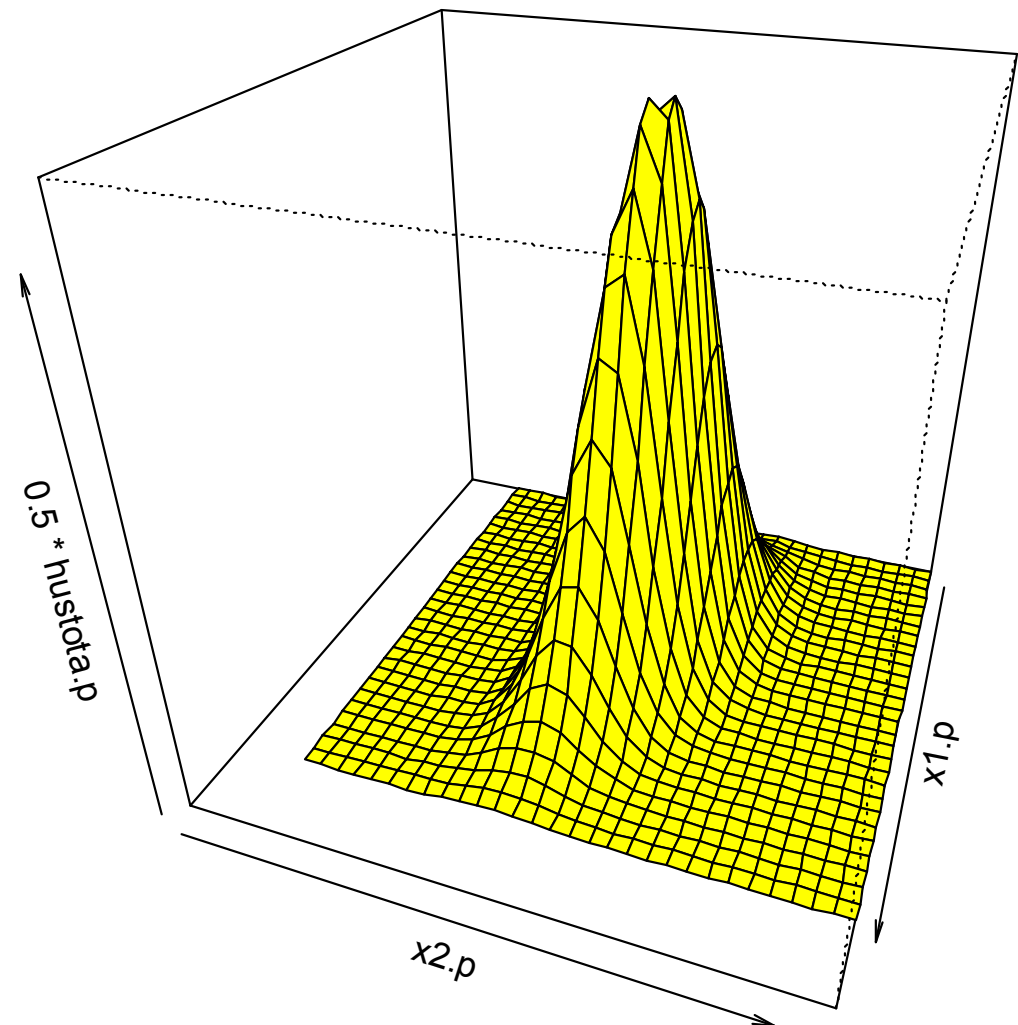
Vícerozměrné normální rozdělení

Motivace

Dvourozměrný histogram



Hustota dvourozměrného normálního rozdělení



Vícerozměrné normální rozdělení

Hustota vícerozměrného normálního rozdělení:

$$f_{\mathbf{x}}(\mathbf{x}_1, \dots, \mathbf{x}_k) = \frac{1}{\sqrt{(2\pi)^k |\Sigma|}} \cdot \exp\left(-\frac{1}{2}(\mathbf{x} - \mu)^T \Sigma^{-1}(\mathbf{x} - \mu)\right)$$

μ - střední hodnota

Σ - kovarianční matice

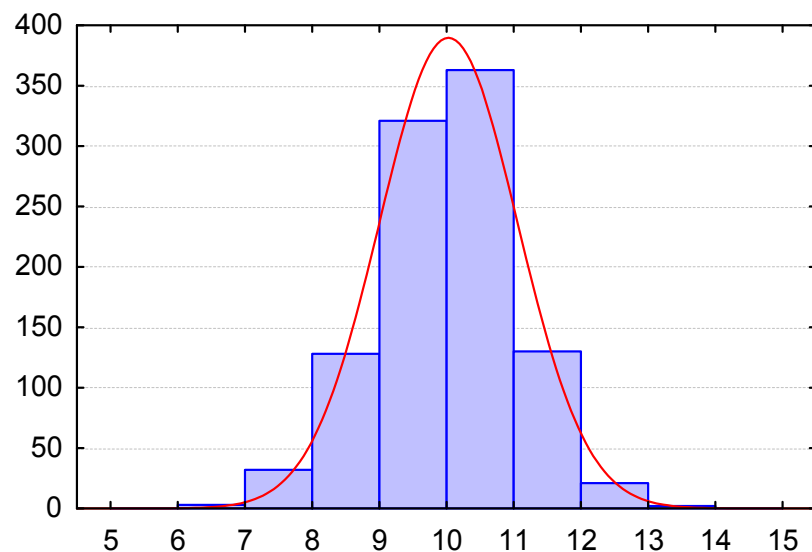
Dvourozměrné normálního rozdělení:

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

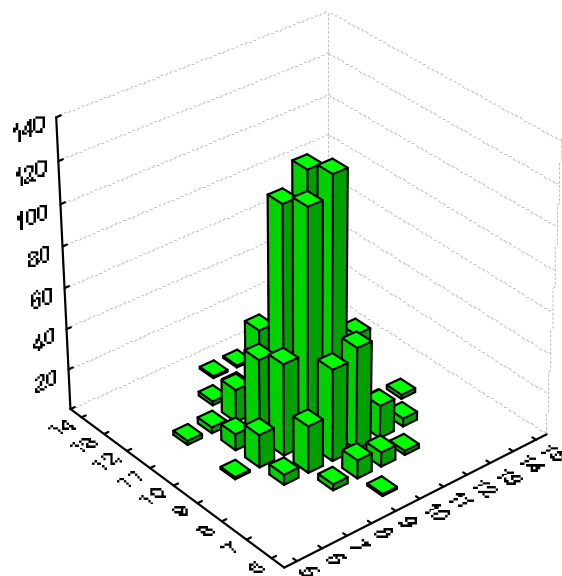
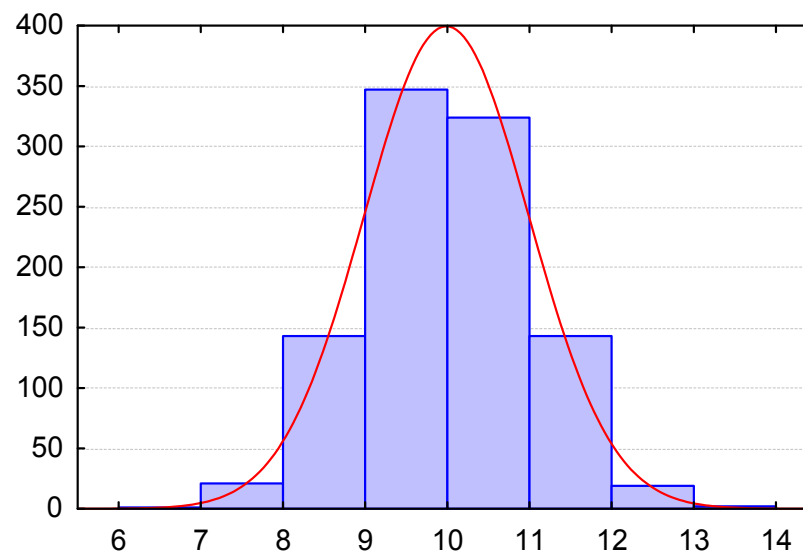
$$\mu = \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix}$$

ρ - korelace mezi X a Y; σ – směrodatná odchylka

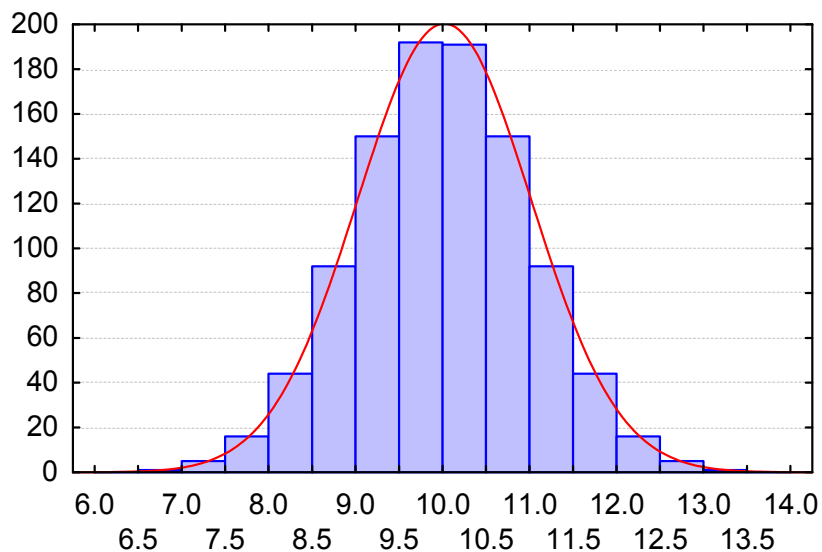
Je normalita v jednorozměrném prostoru jedinou podmínkou vícerozměrné normality?



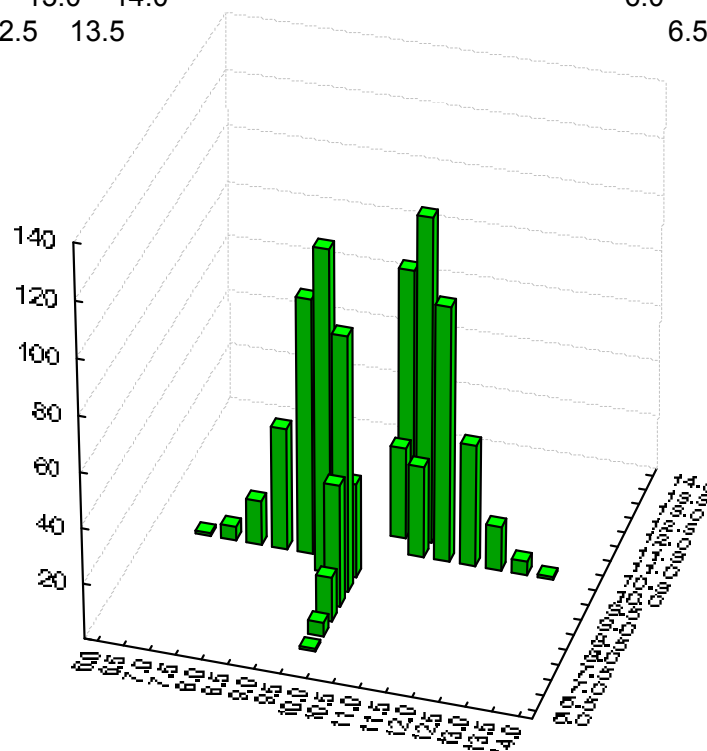
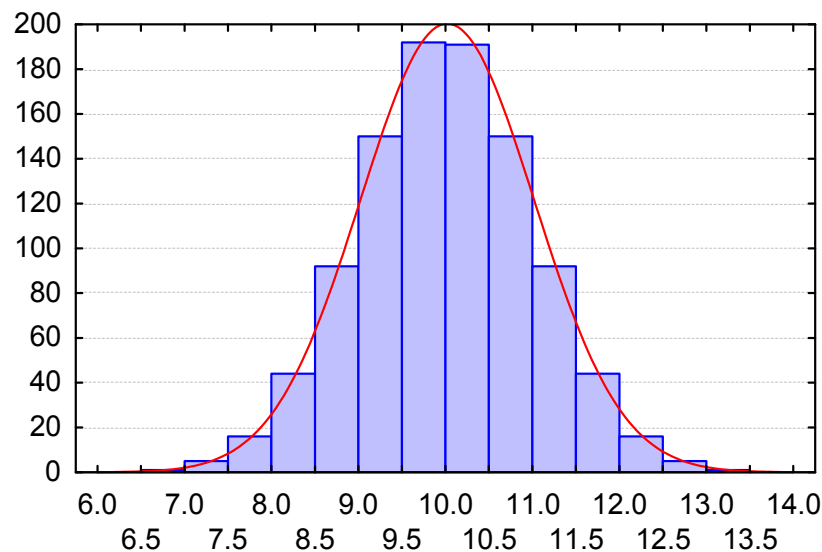
+



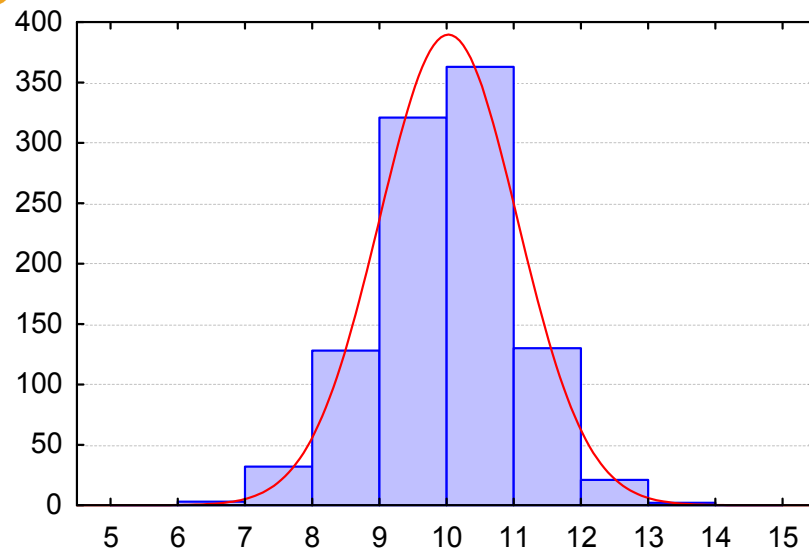
Je normalita v jednorozměrném prostoru jedinou podmínkou vícerozměrné normality?



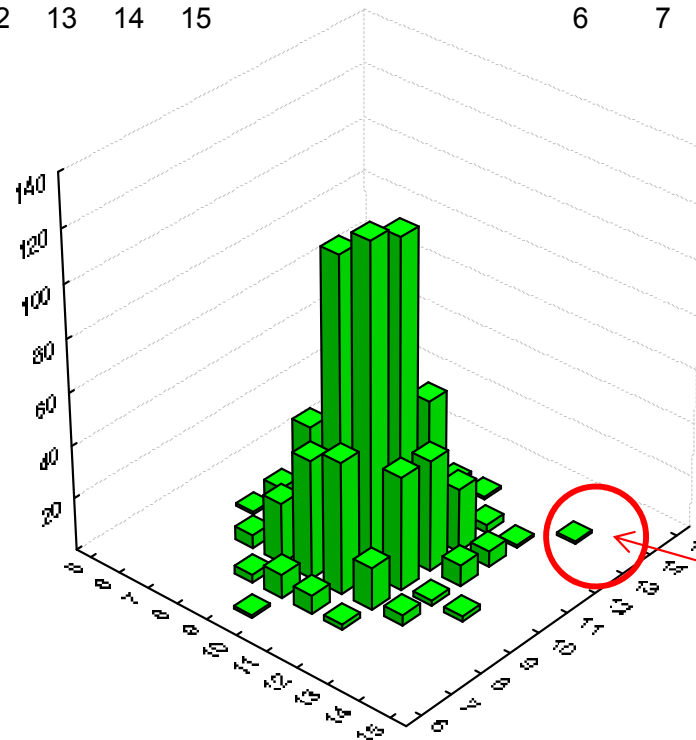
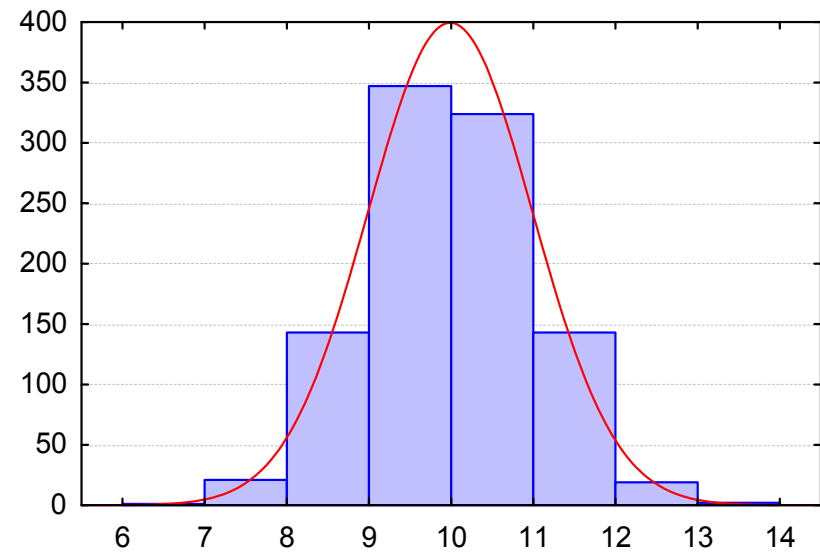
+



Je normalita v jednorozměrném prostoru jedinou podmínkou vícerozměrné normality?



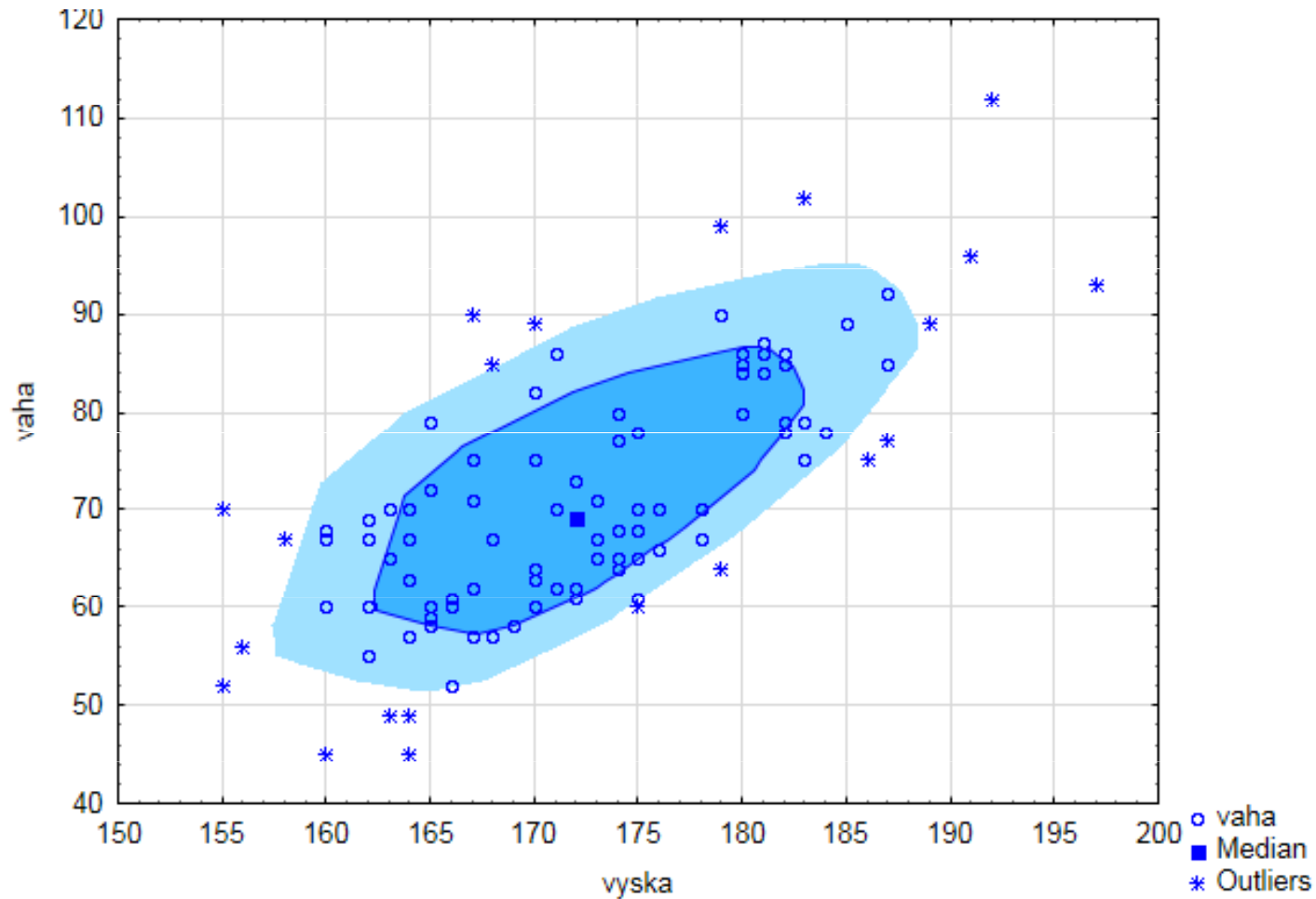
+



Vícerozměrný outlier

Ověření dvourozměrné normality

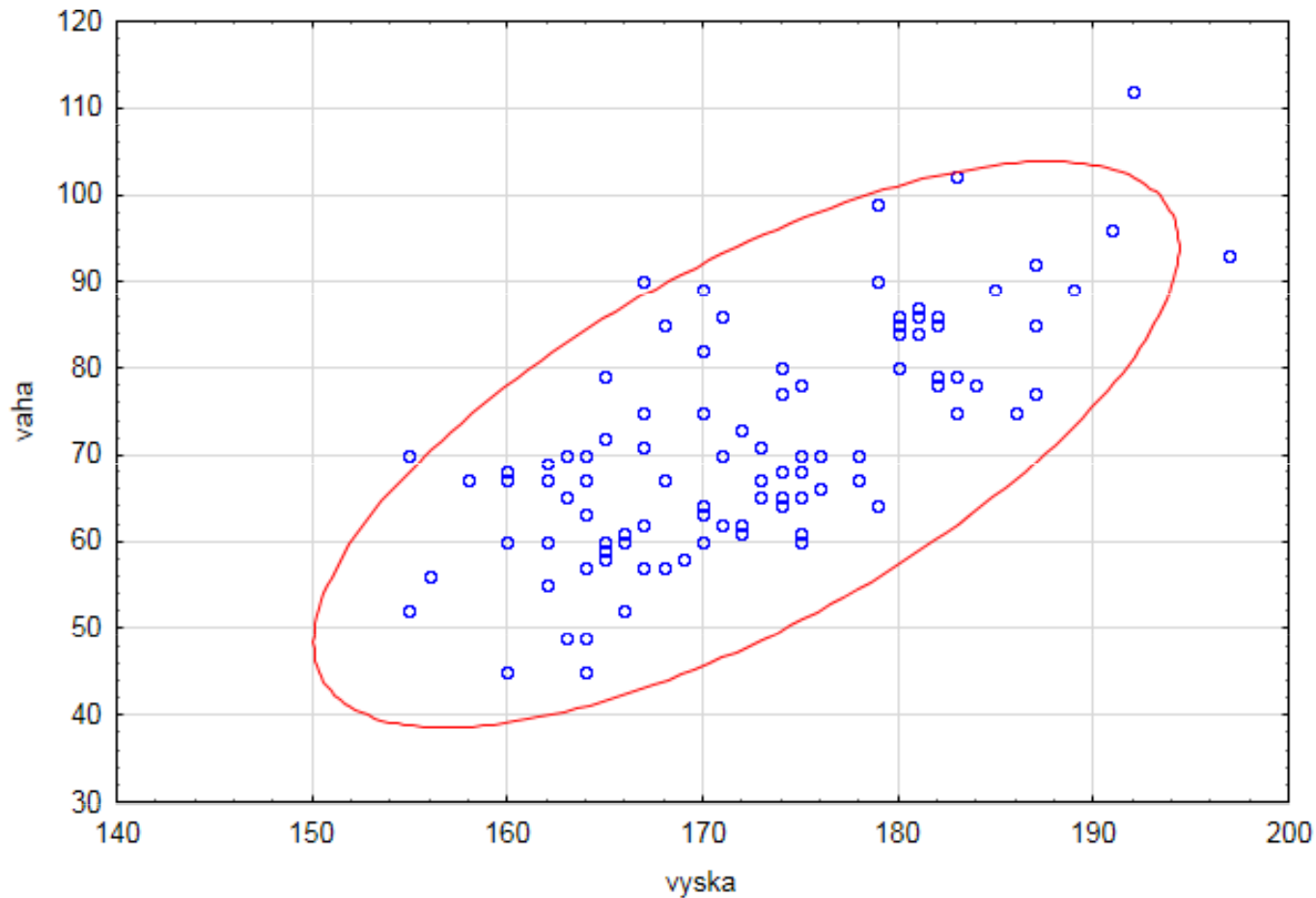
Bagplot = „bivariate boxplot“ (tzn. „dvourozměrný krabicový graf“)



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

Ověření dvourozměrné normality

Vykreslení regulační elipsy („control“ ellipse):

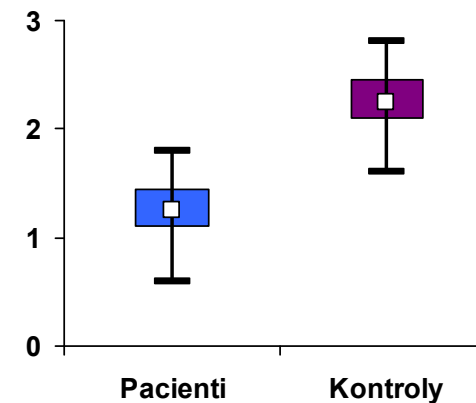
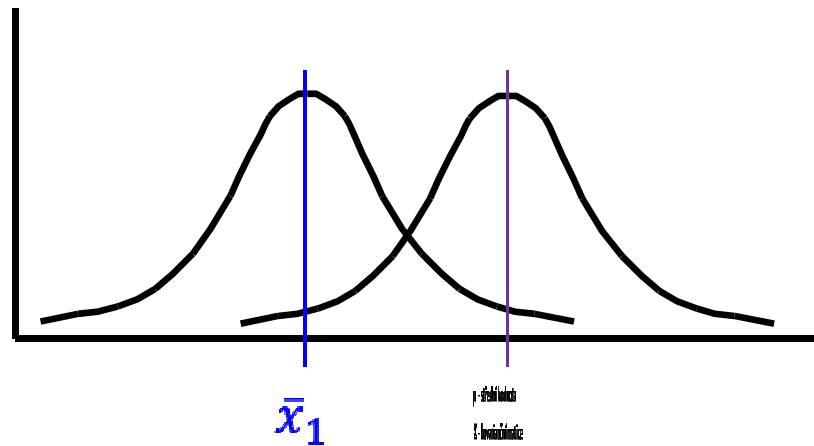


v softwaru Statistica: Graphs – Scatterplots – na záložce Advanced zvolit Ellipse Normal

Vícerozměrný t-test

Jednorozměrný dvouvýběrový t-test

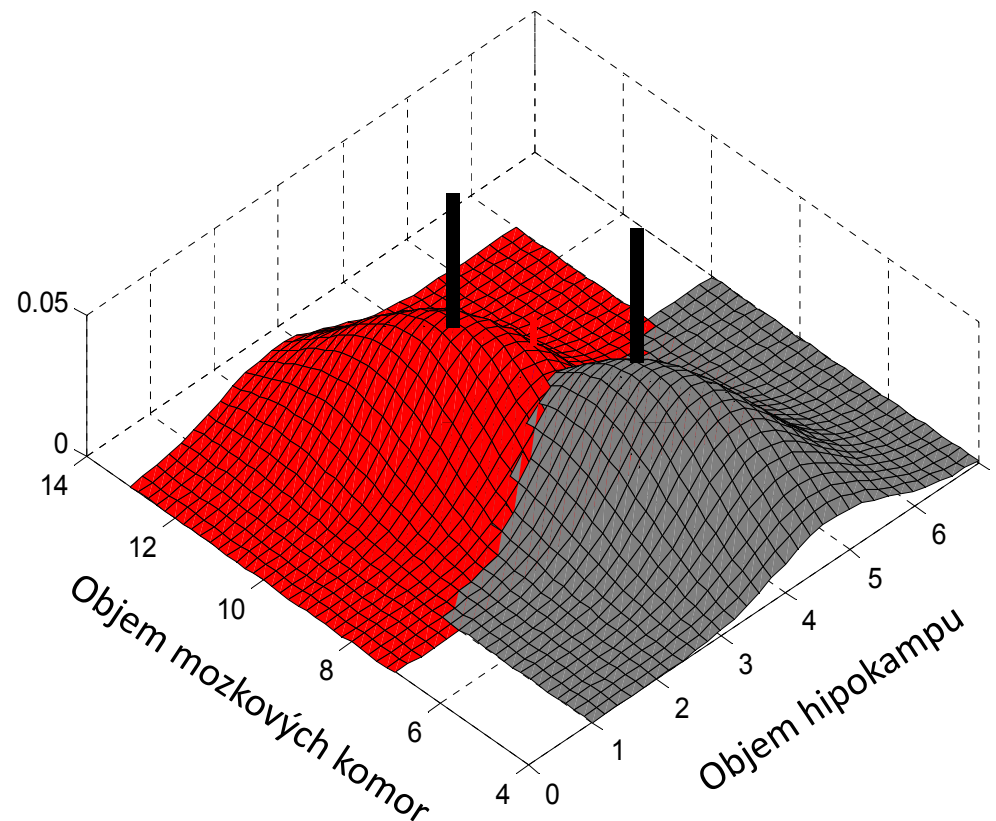
- Srovnáváme dvě skupiny dat, které jsou na sobě nezávislé – mezi objekty neexistuje vazba.
- Příklady: srovnání objemu hipokampu u mužů a u žen, srovnání kognitivního výkonu podle dvou kategorií věku,...



- 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: $t = \frac{(\bar{x} - \bar{y}) - (\mu_x - \mu_y)}{s \sqrt{\frac{1}{n_x} + \frac{1}{n_y}}}$, kde $t \sim T(n_x + n_y - 2)$ ← Studentovo rozdělení
- s je vážená směrodatná odchylka $s^2 = \frac{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}{(n_x - 1) + (n_y - 1)}$
- $(\mu_x - \mu_y) = c$ je konstanta, o kterou se rozdíl průměrů má lišit (většinou $c = 0$)
- nulová hypotéza zamítnuta, pokud $|t| > t_{crit}$

Je ekvivalentní testu:

- $t^2 = \left(\frac{(\bar{x} - \bar{y}) - (\mu_x - \mu_y)}{s \sqrt{\frac{1}{n_x} + \frac{1}{n_y}}} \right)^2 = (z - \mu_z) \left[s^2 \left(\frac{1}{n_x} + \frac{1}{n_y} \right) \right]^{-1} (z - \mu_z)$, kde $t^2 \sim F(1, n_x + n_y - 2)$ ← F rozdělení
 $\bar{z} = \bar{x} - \bar{y}$ a $\mu_z = \mu_x - \mu_y$

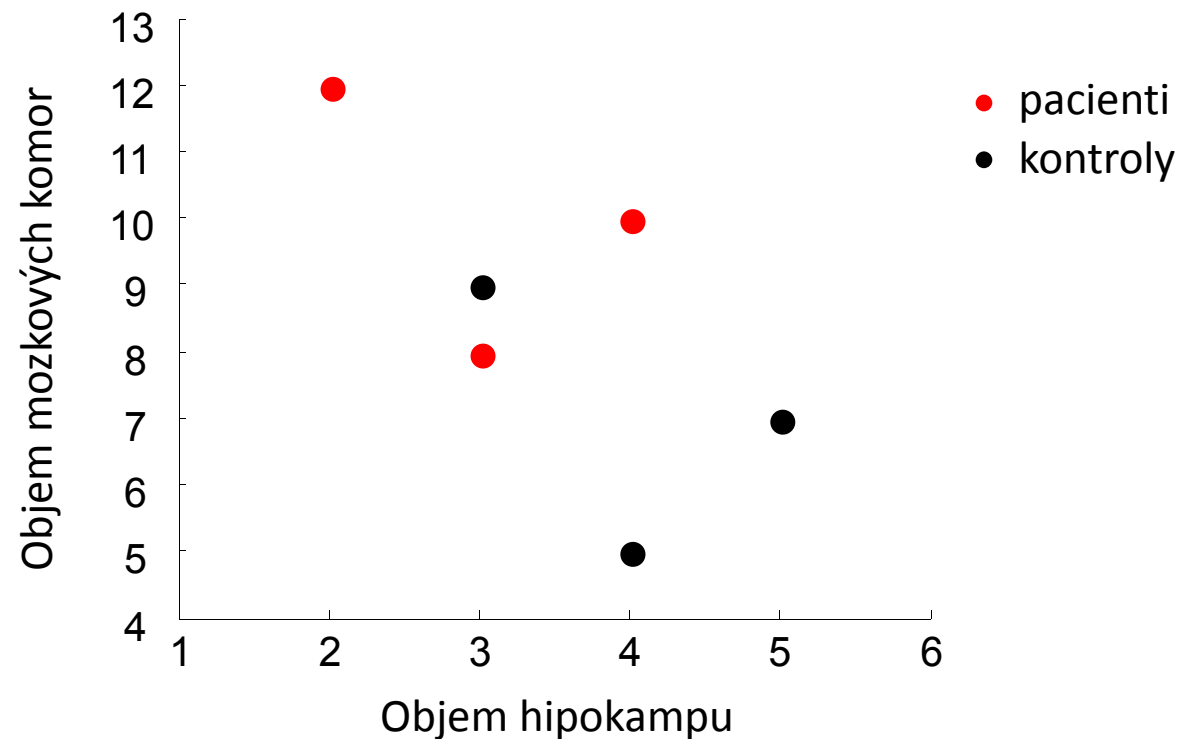
Vícerozměrný t-test:

- dvouvýběrová Hotellingova T^2 testová statistika: $T^2 = (\bar{X} - \bar{Y})^T \left[S \left(\frac{1}{n_x} + \frac{1}{n_y} \right) \right]^{-1} (\bar{X} - \bar{Y})$
- kde S je vážená kovarianční matice $S = \frac{(n_x - 1)S_X + (n_y - 1)S_Y}{(n_x - 1) + (n_y - 1)}$
- $T^2 \sim \chi^2(k)$; pro malé n_x a n_y je lepší použít: $F = \frac{n - k}{k(n - 1)} T^2 \sim F(k, n - k)$, kde $n = n_x + n_y - 1$ ← F rozdělení
- nulová hypotéza zamítnuta, pokud $F > F_{crit}$

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	5
k	2
T^2	3,5
F	1,31
df1	2
df2	3
α	0,05
F-crit	9,55
p-hodnota	0,389

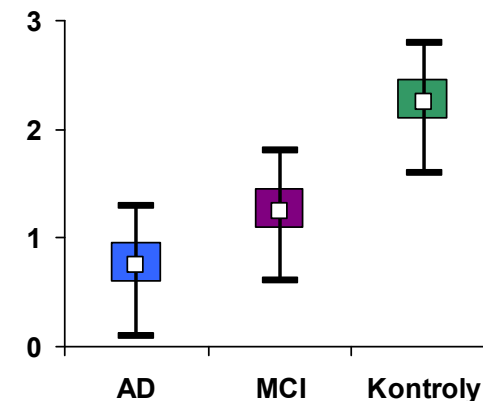
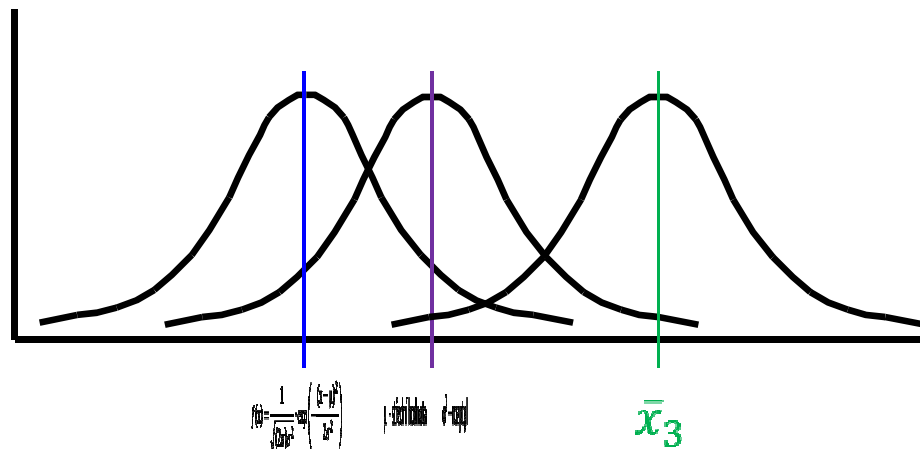
$$T^2 = (\bar{X} - \bar{Y})^T \left[S \left(\frac{1}{n_x} + \frac{1}{n_y} \right) \right]^{-1} (\bar{X} - \bar{Y})$$

$$F = \frac{n - k}{k(n - 1)} T^2 \sim F(k, n - k)$$

Vícerozměrná analýza rozptylu

Analýza rozptylu (ANOVA) jednoduchého třídění

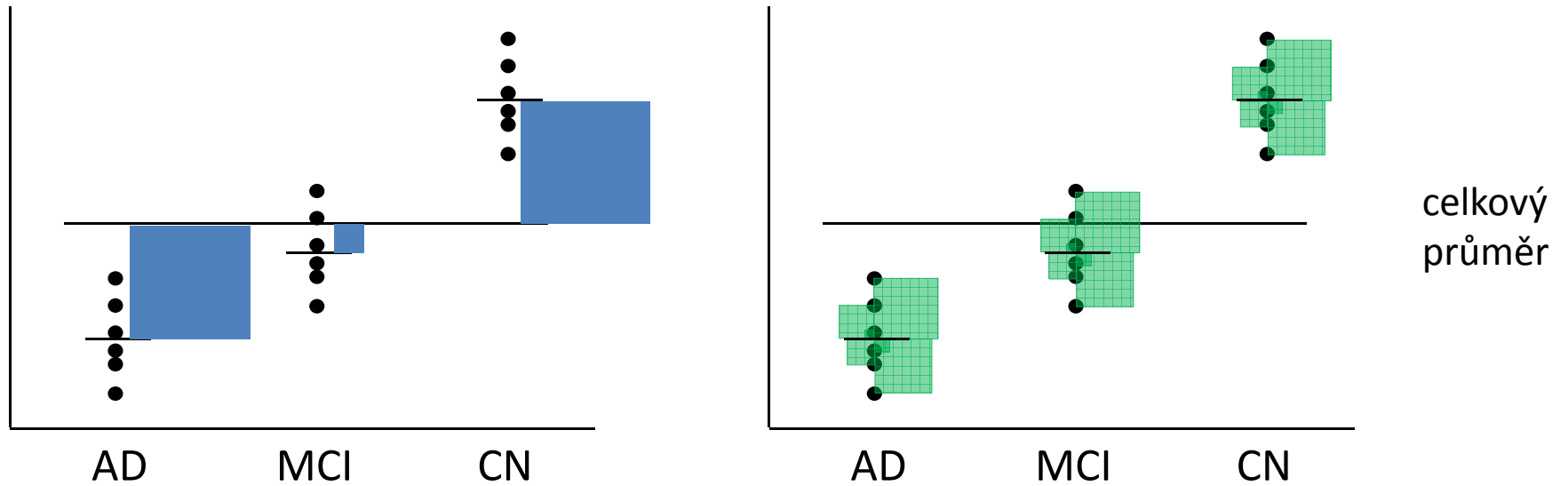
- Srovnáváme tři a více skupin dat, které jsou na sobě nezávislé (mezi objekty neexistuje vazba).
- Příklady: srovnání objemu hipokampu u pacientů s AD, pacientů s MCI a kontrol; srovnání kognitivního výkonu podle čtyř kategorií věku.



- 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 (arrow to μ)

i -tý efekt faktoru A (arrow to α_i)

Reziduum (arrow to e_{ij})

- 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 dvojného třídění

- Uvažujeme dvě vysvětlující proměnné zároveň.
- Zápis modelu:

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

↑ Populační průměr
 ↑ i -tý efekt faktoru A
 ↑ j -tý efekt faktoru B
 ← Reziduum

- 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	$df_A = k - 1$	$MS_A = S_A / df_A$	F_A	p
Faktor B	S_B	$df_B = r - 1$	$MS_B = S_B / df_B$	F_B	p
Rezidua	S_e	$df_e = (k - 1)(r - 1)$	$MS_e = S_e / df_e$		
Celkem	S_T	$df_T = n - 1 = kr - 1$			

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:

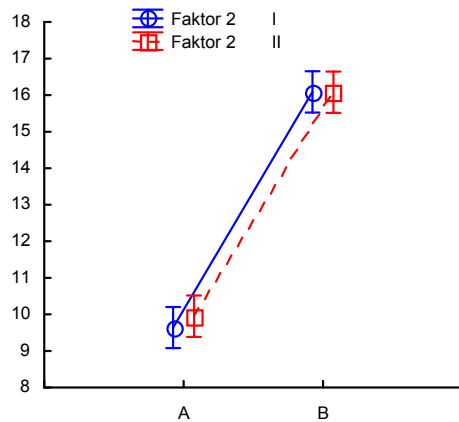
- μ : Populační průměr (Population mean)
- α_i : i -tý efekt faktoru A
- β_j : j -tý efekt faktoru B
- γ_{ij} : Interakce (Interaction)
- e_{ij} : Reziduum (Residual)

- Nulové hypotézy pak máme tři:

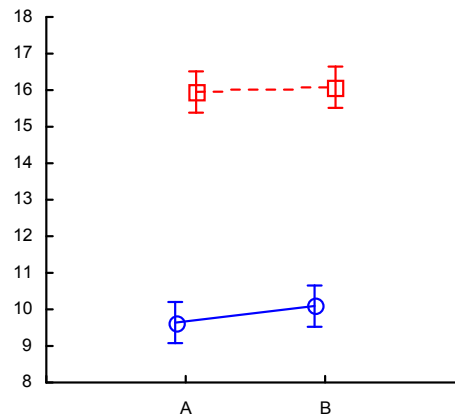
$$H_{01} : \gamma_{11} = \gamma_{12} = \dots = \gamma_{kr} \quad H_{02} : \alpha_1 = \alpha_2 = \dots = \alpha_k \quad H_{03} : \beta_1 = \beta_2 = \dots = \beta_r$$

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

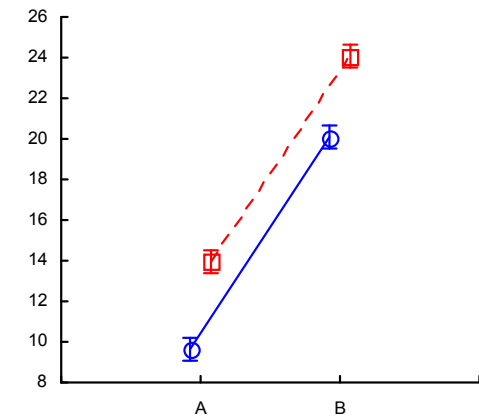
Hlavní efekty a interakce



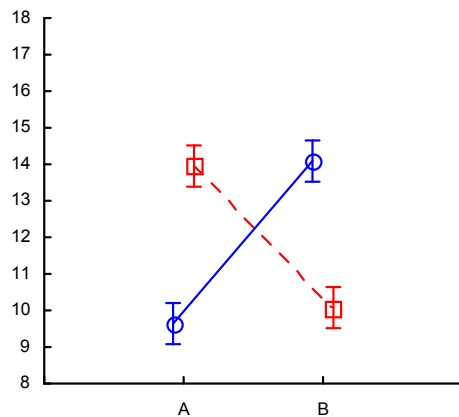
	SS	D.f.	MS	F	p
Faktor 1	1978	1	1978	482.2	0.000
Faktor 2	1	1	1	0.3	0.602
F1*F2	1	1	1	0.3	0.570
Error	804	196	4		



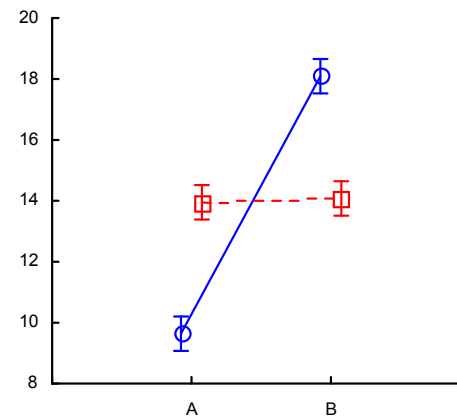
	SS	D.f.	MS	F	p
Faktor 1	4	1	4	1.0	0.314
Faktor 2	1891	1	1891	461.1	0.000
F1*F2	1	1	1	0.3	0.570
Error	804	196	4		



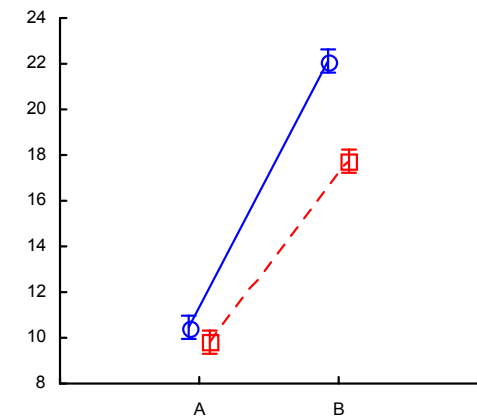
	SS	D.f.	MS	F	p
Faktor 1	5293	1	5293	1290.7	0.000
Faktor 2	861	1	861	209.9	0.000
F1*F2	1	1	1	0.3	0.570
Error	804	196	4		



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



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



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

Úkol 2

Zjistěte, zda má vliv pohlaví a typ léku na počet uzdravených pacientů s leukémií.

Pohlaví	Typ léku	Počet uzdravených pacientů
M	placebo	1
M	lék 1	1
M	lék 2	6
Z	placebo	3
Z	lék 1	4
Z	lék 2	9

Úkol 2 – řešení

Zjistěte, zda má vliv pohlaví a typ léku na počet uzdravených pacientů s leukémií.

Překódování:

Pohlaví	Typ léku	Počet uzdravených pacientů
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=placebo
2=lék 1
3=lék 2

Úkol 2 – řešení

Pohlaví	Typ léku	Počet uzdrav. pacientů	
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 :

počet stupňů volnosti: $f_E = n - a -$

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

$$b + 1 = 2$$

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 uzdravených pacientů

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

Úkol 2 – řešení v softwaru STATISTICA

Zjistěte, zda má vliv pohlaví a typ léku na počet uzdravených pacientů s leukémií.

Pohlaví	Typ léku	Počet uzdrav. pacientů
M	placebo	1
M	lék 1	1
M	lék 2	6
Z	placebo	3
Z	lék 1	4
Z	lék 2	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 – 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 uzdravených pacientů s leukémií.

Pohlaví	Typ léku	Počet uzdrav. pacientů
M	placebo	1
M	lék 1	1
M	lék 2	6
Z	placebo	3
Z	lék 1	4
Z	lék 2	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 uzdravených 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)
```

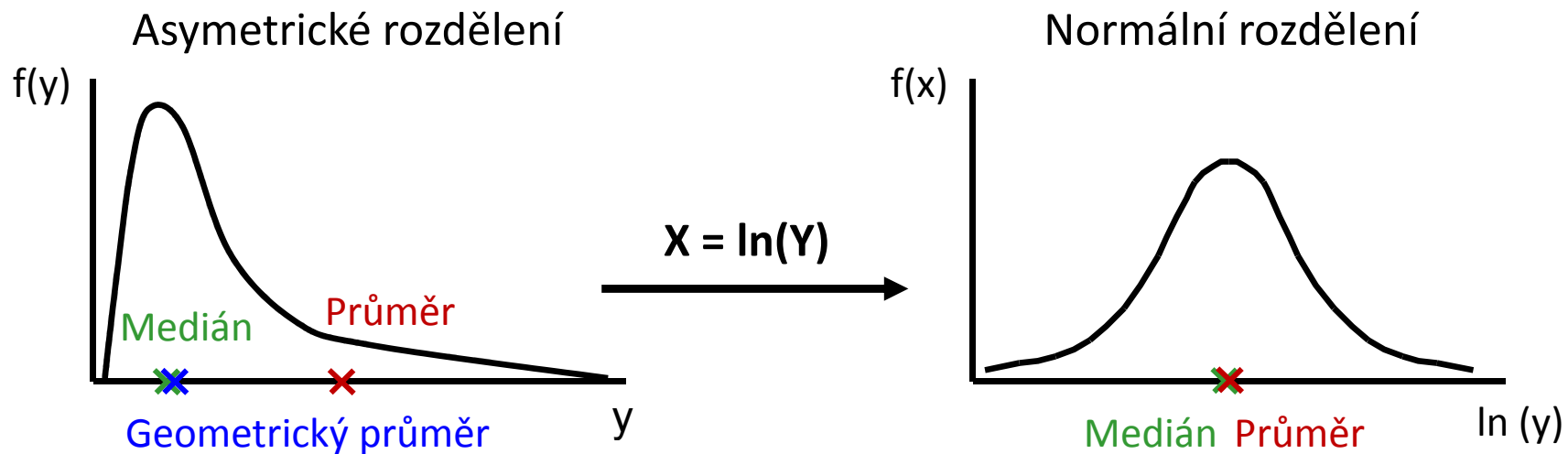
Transformace a jiné úpravy vícerozměrných dat

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

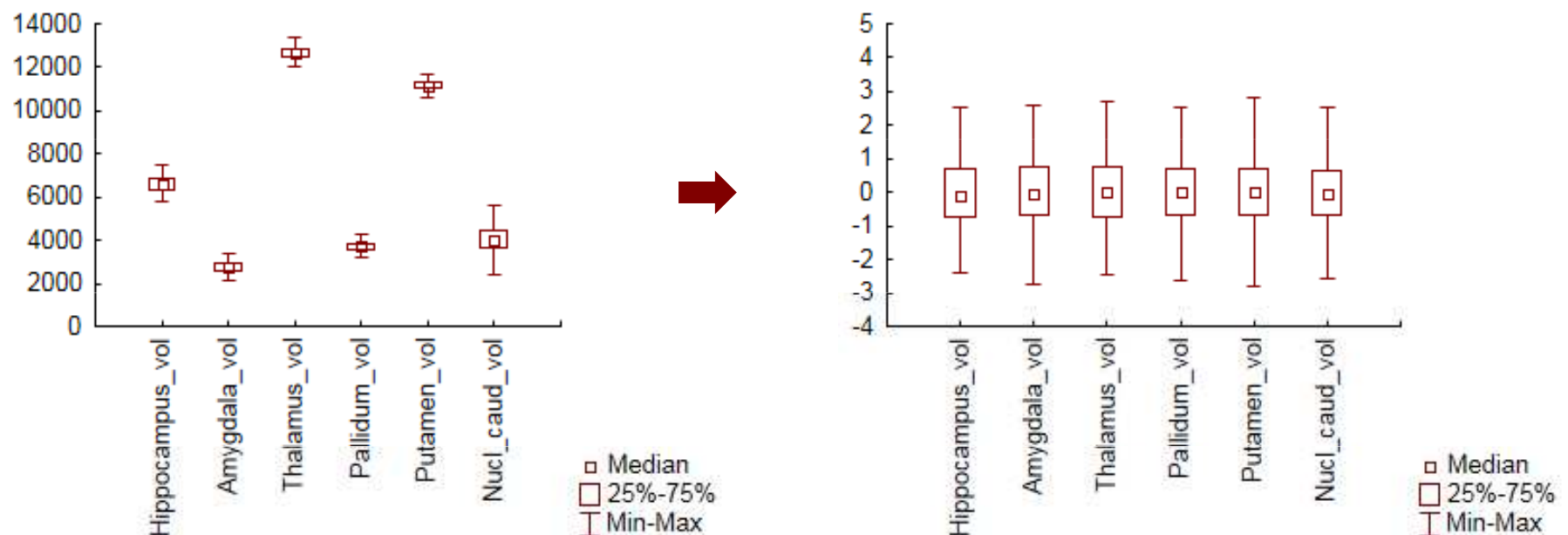
- převod na normální rozdělení (normalita je předpokladem řady statistických testů).
- např. **logaritmická transformace**: $X = \ln(Y)$ nebo $X = \ln(Y+1)$, pokud data obsahují hodnotu 0



- 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**

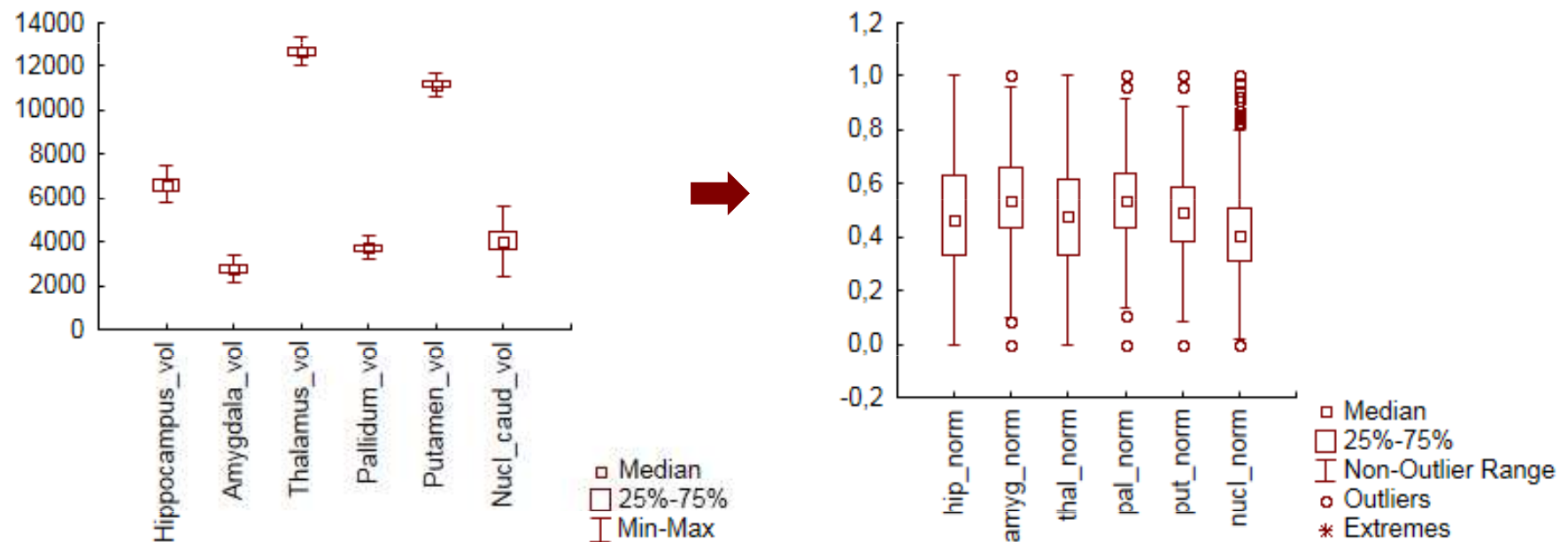
Standardizace dat

- 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
- **pozor: standardizace je nevhodná v případě, že proměnné nemají normální rozdělení a že se v datech vyskytují odlehlé hodnoty!!!**



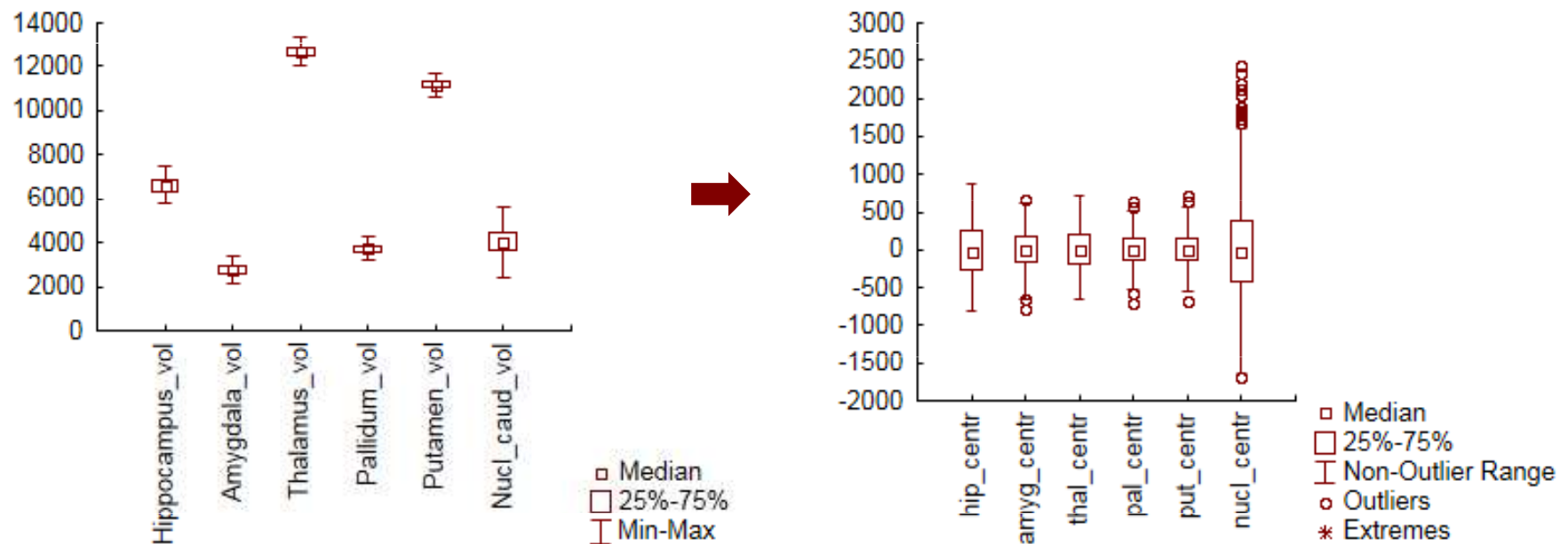
Min-max normalizace

- 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

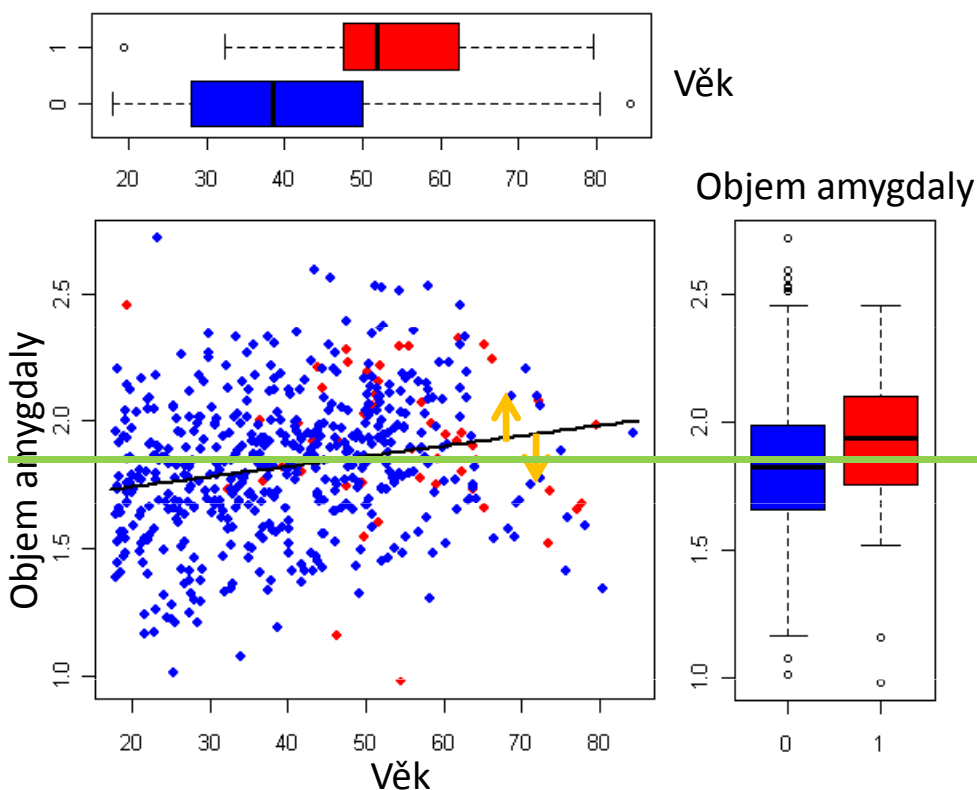
- odečtení průměru od dat – získáme novou proměnnou, která bude mít průměr roven nule
- důvod: centrování je důležitou podmínkou některých pokročilých statistických metod (např. klasifikačních)
- centrování: $z_i = x_i - \bar{x}$



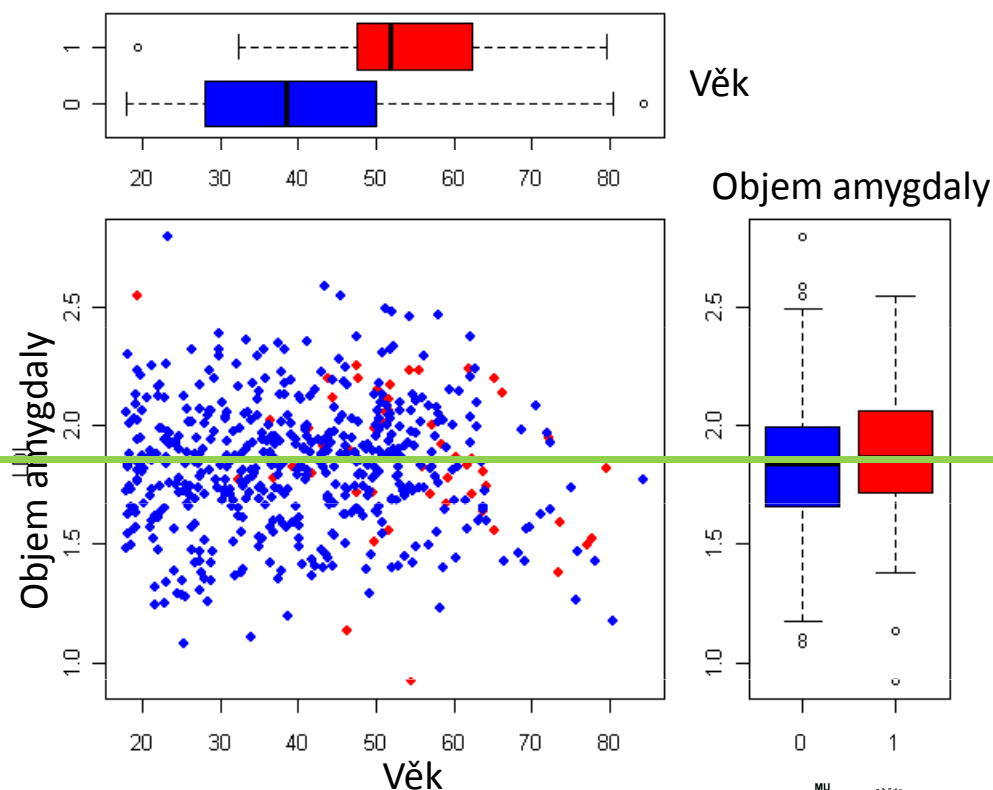
Odstranění vlivu kovariát na jiné proměnné (tzv. adjustace)

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 -----
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“

