

# Vícerozměrné metody - cvičení



RNDr. Eva Janoušová

Podzim 2014

# Cvičení 2

Vícerozměrné normální rozdělení a  
vícerozměrné statistické testy

# Osnova

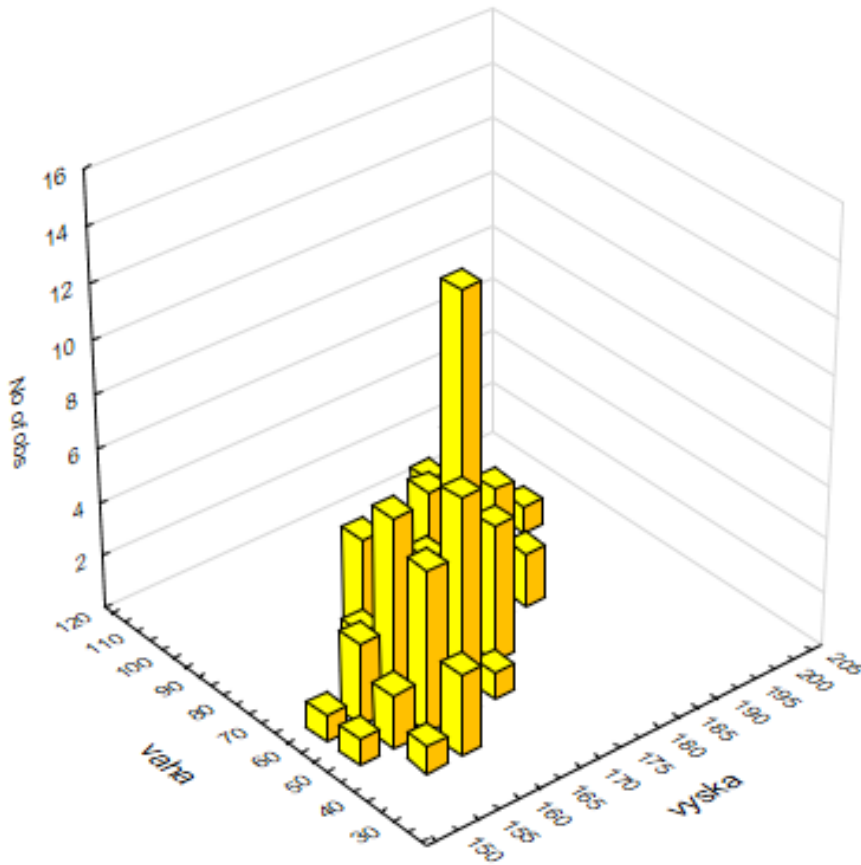
---

- vícerozměrné normální rozdělení
- vícerozměrný t-test
- vícerozměrná analýza rozptylu

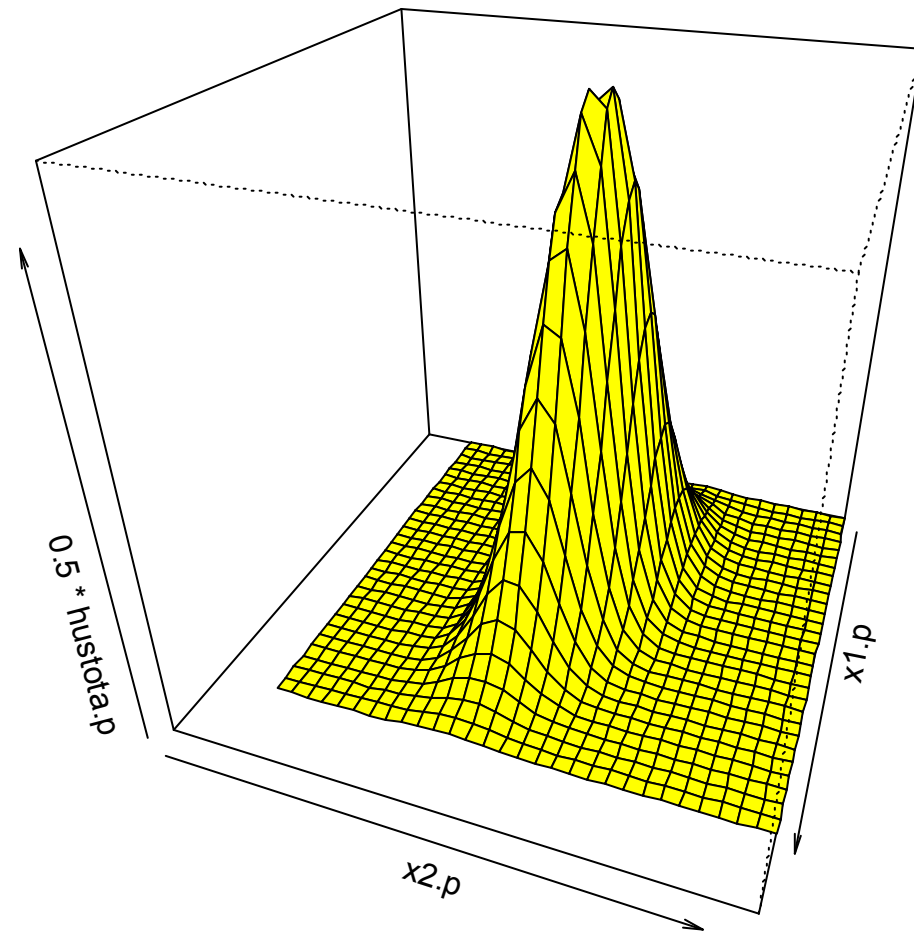
# Vícerozměrné normální rozdělení

# 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

# Vícerozměrný průměr a kovarianční matice

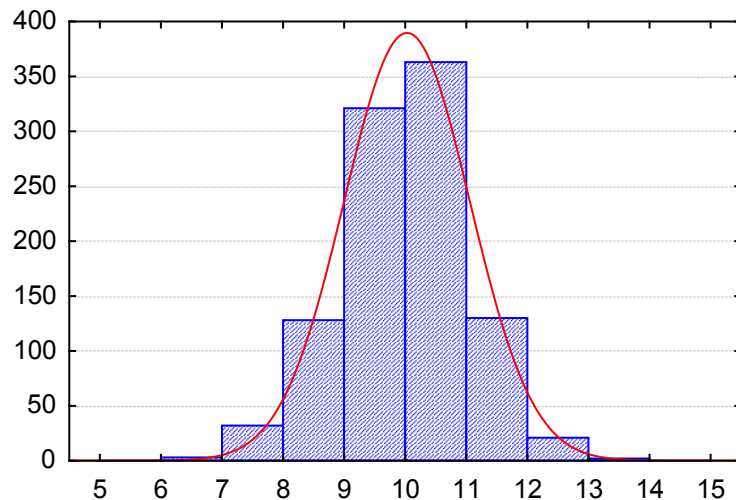
- vícerozměrný průměr (např. pro datový soubor se 2 proměnnými):

$$\bar{\mathbf{x}} = \begin{bmatrix} \frac{1}{n} \sum_{i=1}^n x_{i1} & \frac{1}{n} \sum_{i=1}^n x_{i2} \end{bmatrix}$$

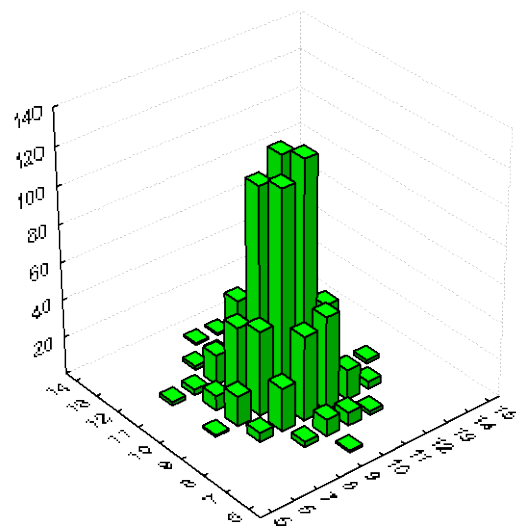
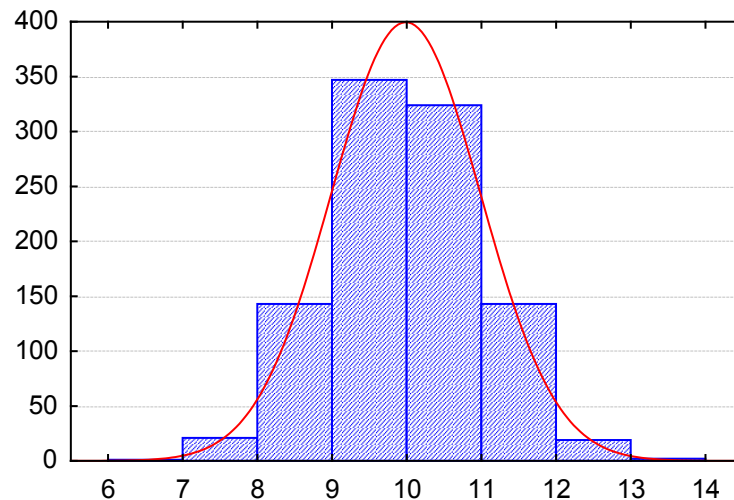
- výběrová kovarianční matice (např. pro datový soubor se 2 proměnnými):

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

# Je normalita v jednorozměrném prostoru jedinou podmínkou vícerozměrné normality?

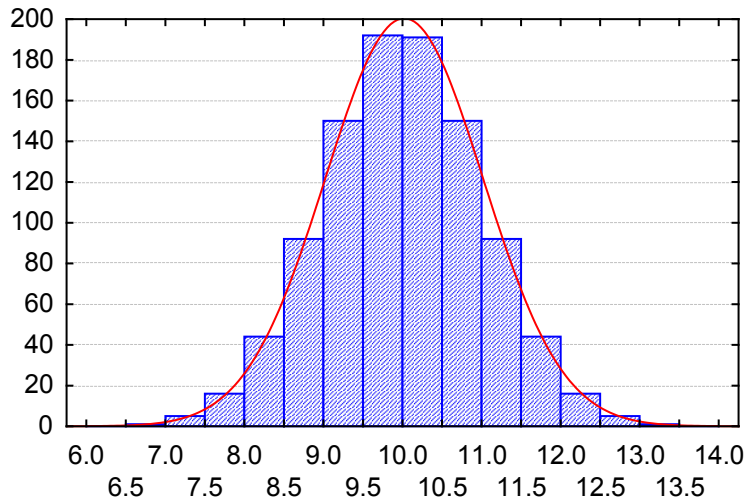


+

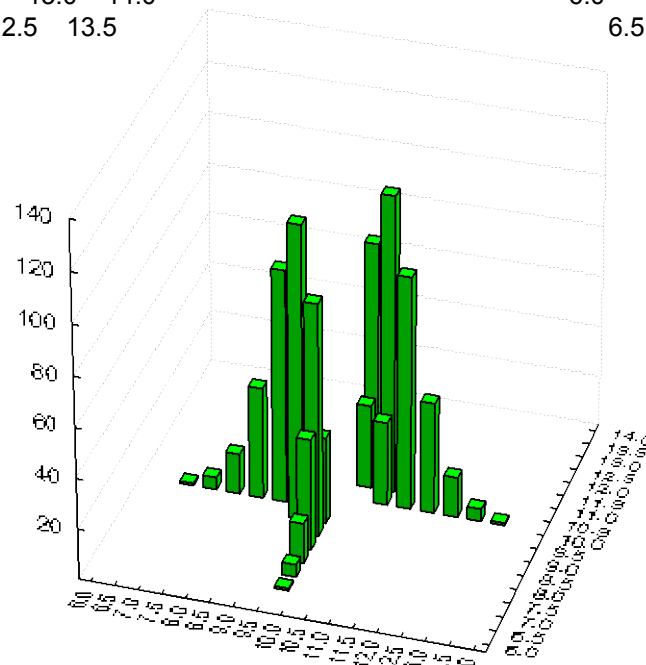
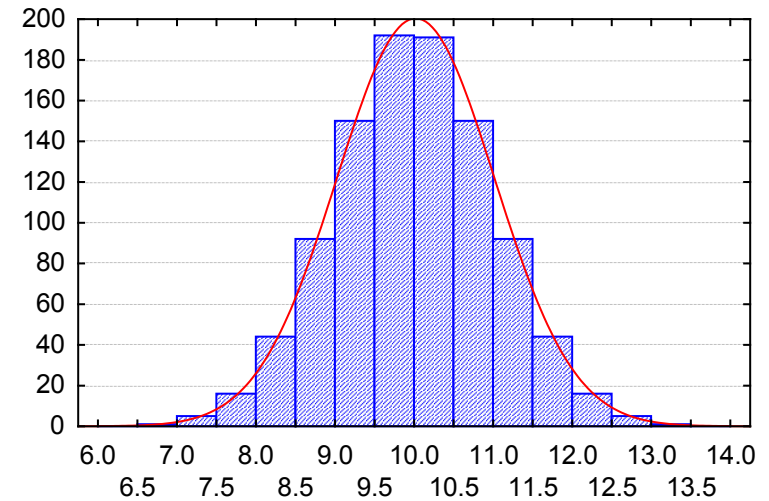




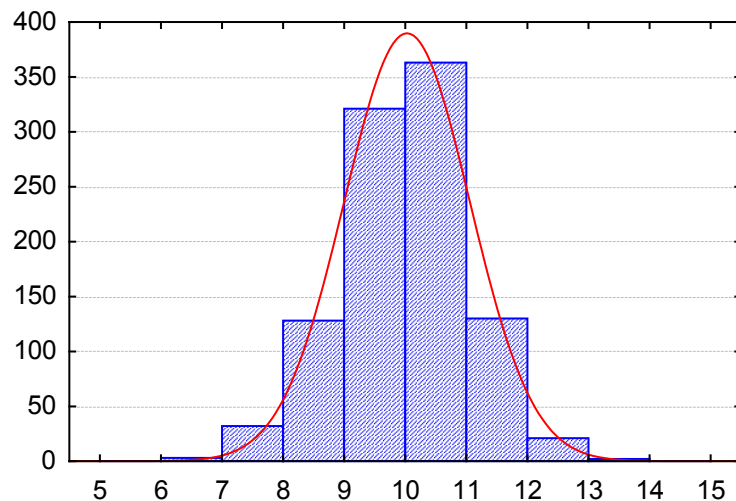
# Je normalita v jednorozměrném prostoru jedinou podmínkou vícerozměrné normality?



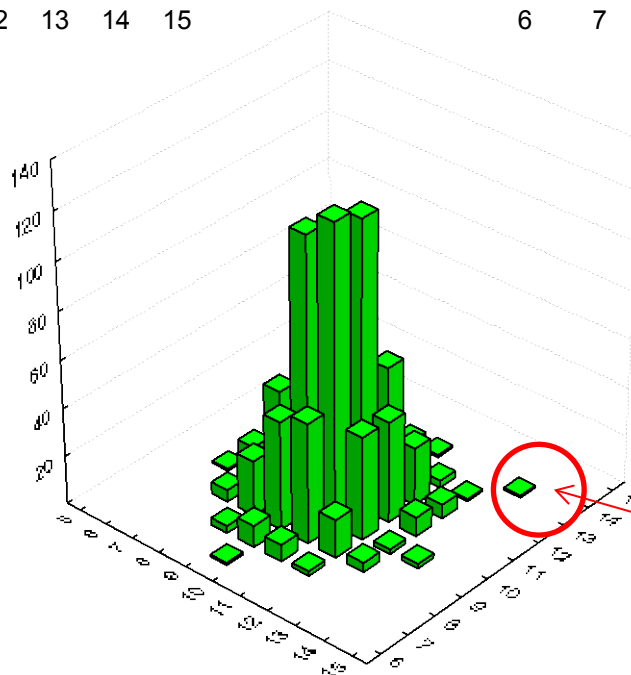
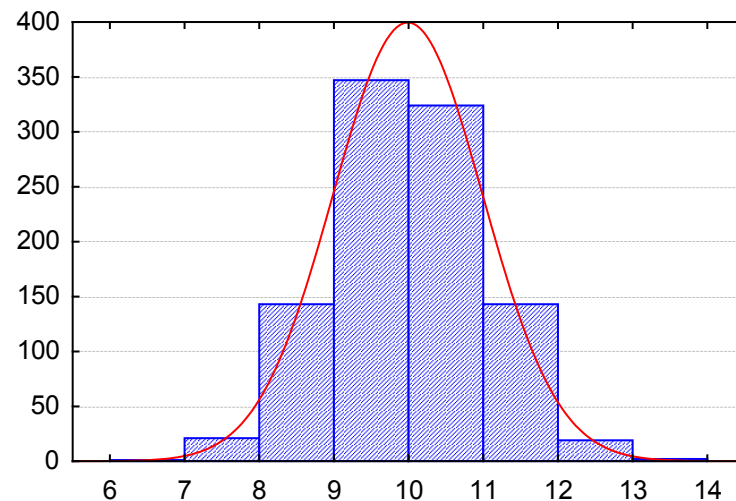
+



# Je normalita v jednorozměrném prostoru jedinou podmínkou vícerozměrné normality?



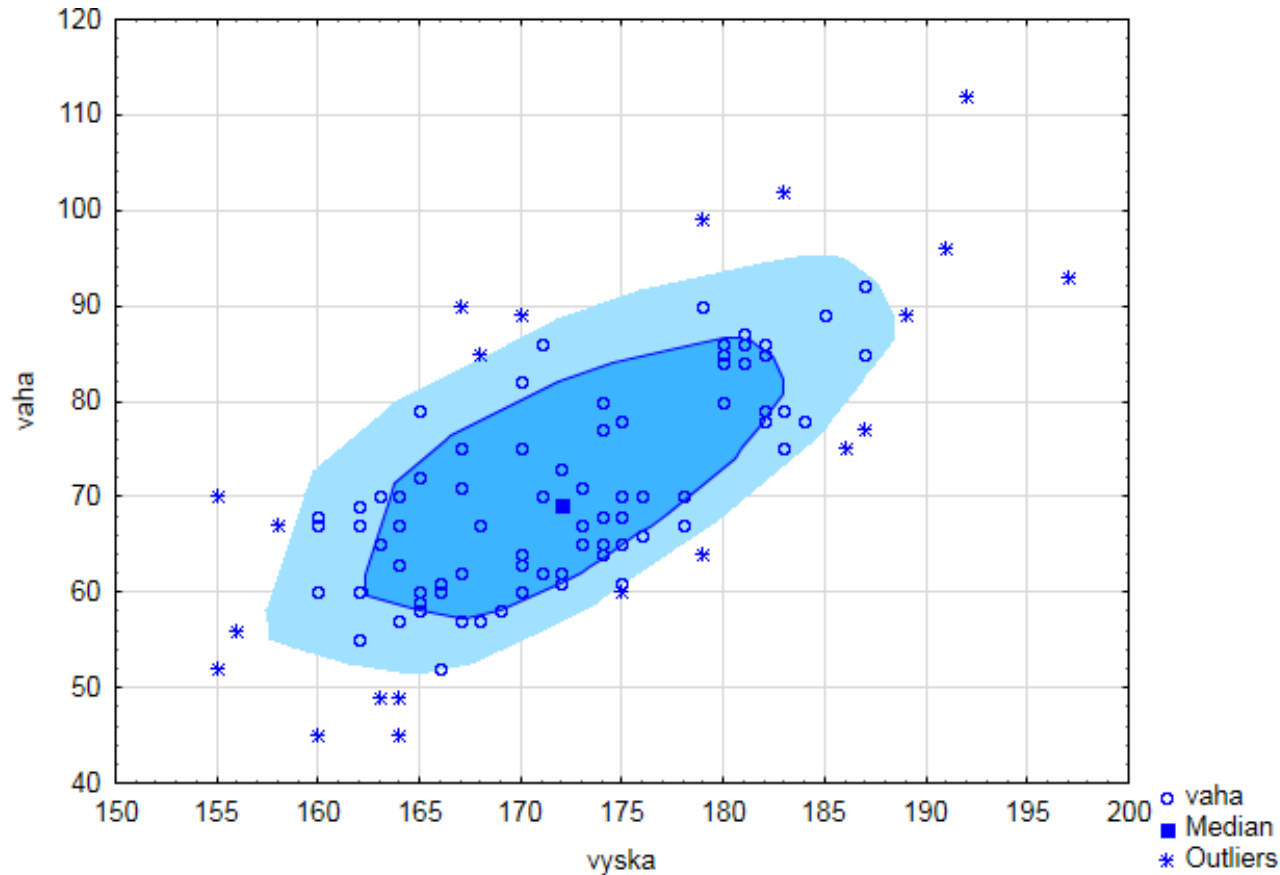
+



Vícerozměrná odlehlá hodnota (outlier)

# Ověření dvourozměrné normality

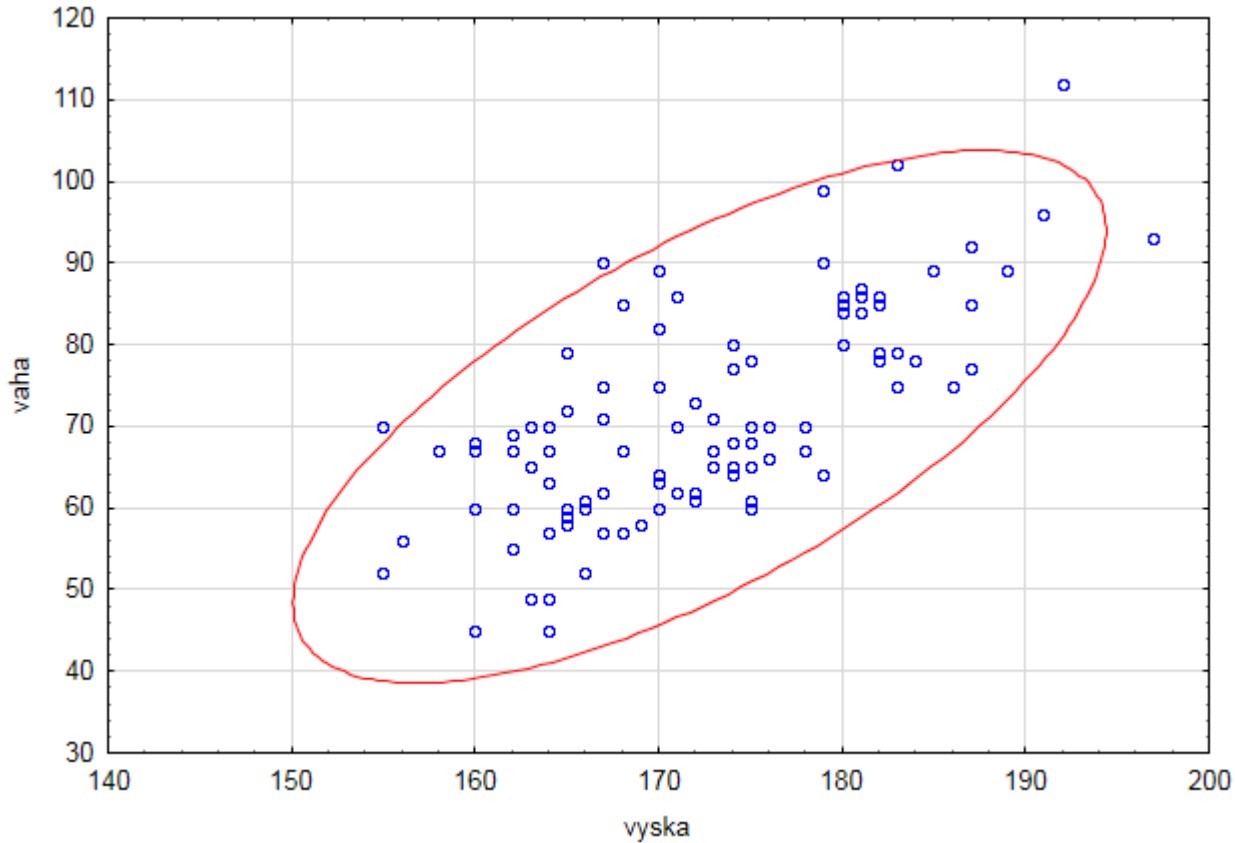
Bagplot = „bivariate boxplot“ (tzn. „dvourozměrný krabicový graf“)



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

# Ověření dvourozměrné normality

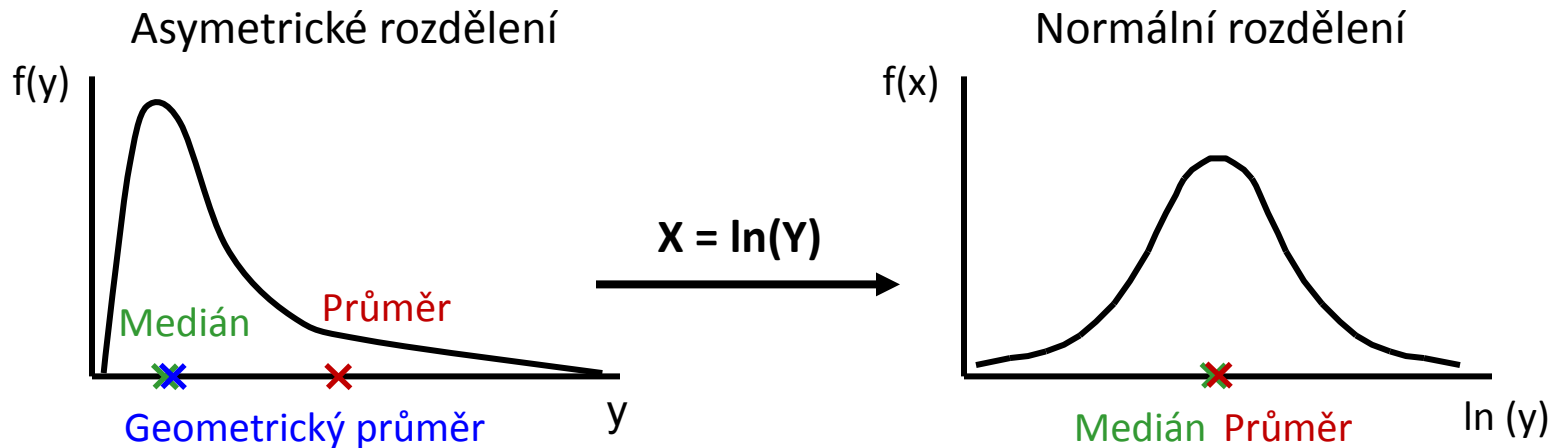
Vykreslení regulační elipsy („control“ ellipse):



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

# 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

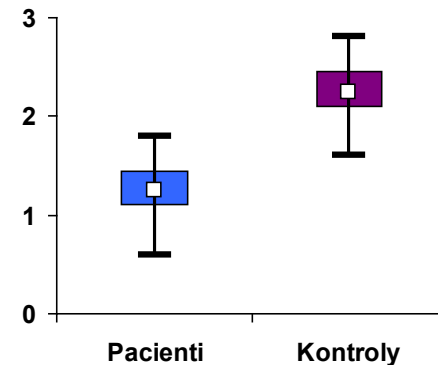
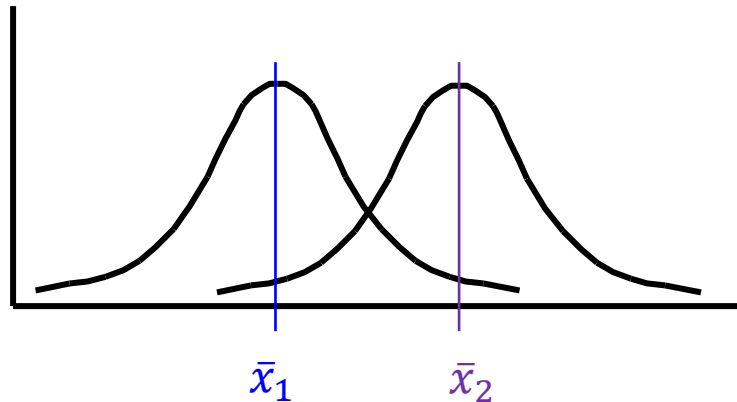


- Další příklady:
  - **odmocninová transf.** (pro proměnné s Poissonovým rozložením nebo obecně data typu počet jedinců, buněk apod.:  $X = \sqrt{Y}$  nebo  $X = \sqrt{Y + 1}$ )
  - **arcsin transformace** (pro proměnné s binomickým rozložením)
  - **Box-Coxova transformace**

# Vícerozměrný t-test

# Jednorozměrný dvouvýběrový t-test

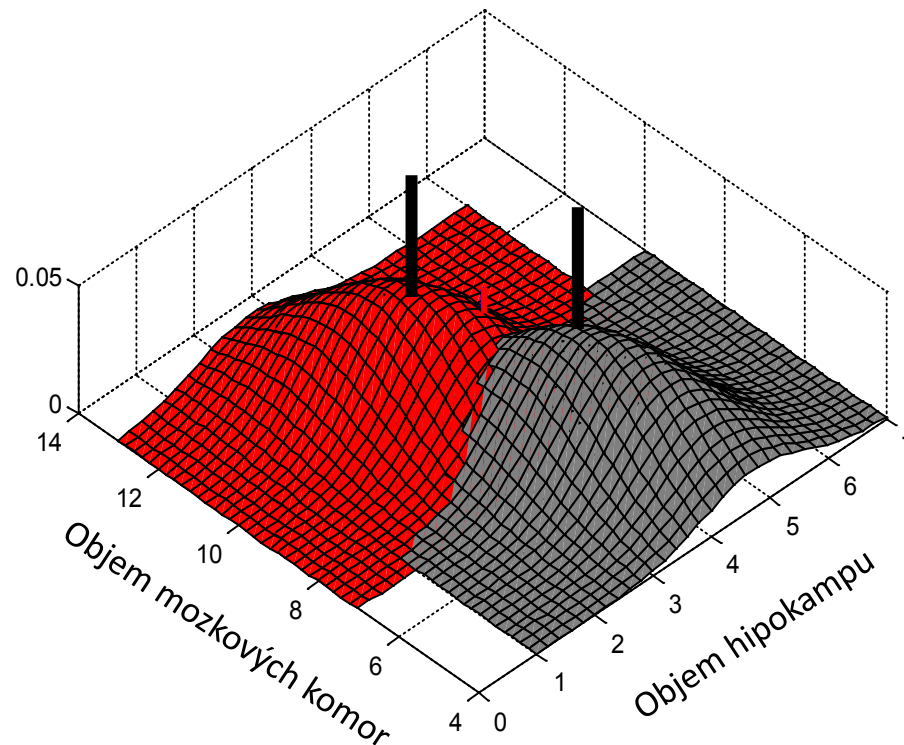
- Srovnáváme dvě skupiny dat, které jsou na sobě nezávislé – mezi objekty neexistuje vazba.
- Příklady: srovnání objemu hipokampu u mužů a u žen, srovnání kognitivního výkonu podle dvou kategorií věku,...



- 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:

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

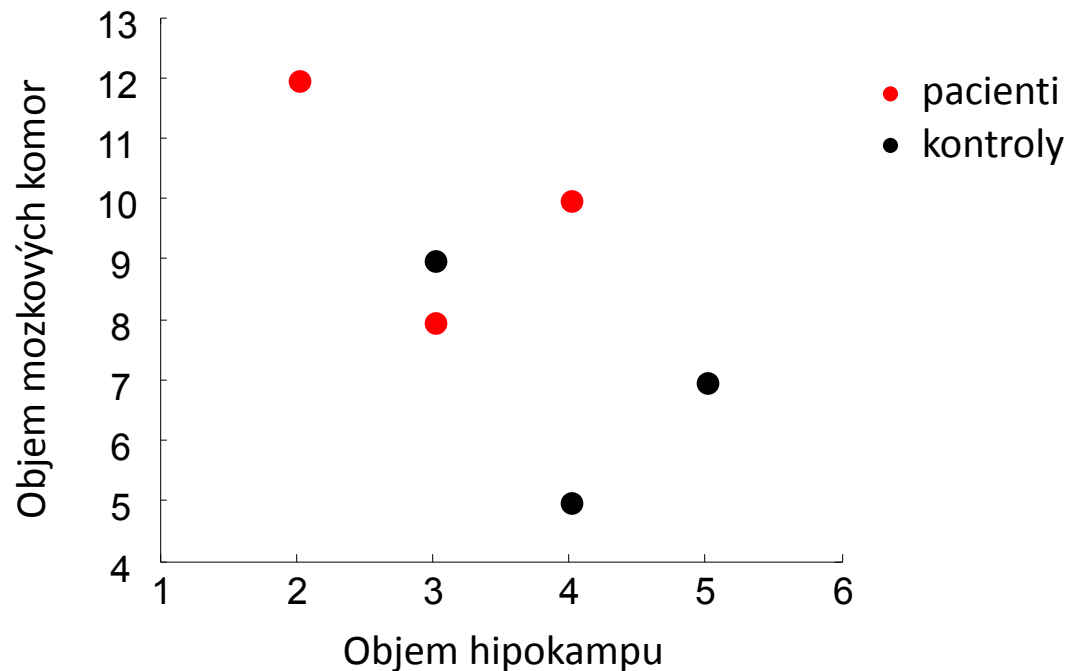
## Vícerozměrný t-test:

- 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

- 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:

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

# 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:

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

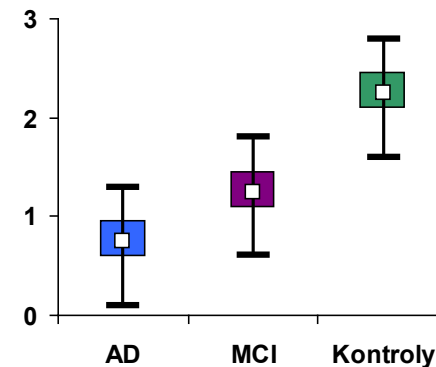
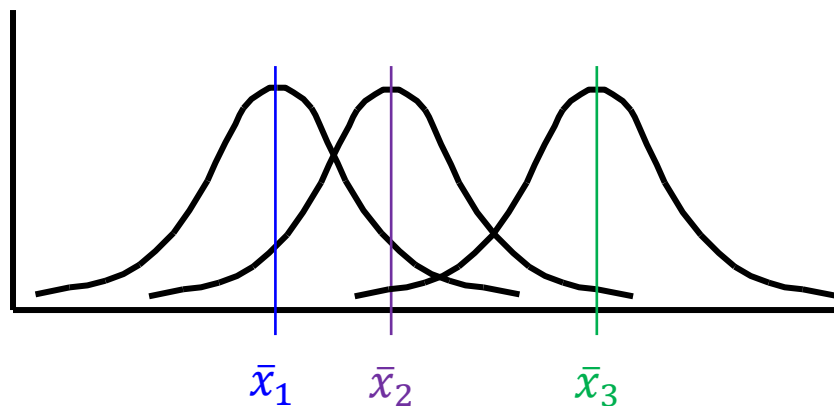
Vícerozměrný t-test:

n	5
k	2
T <sup>2</sup>	3,5
F	1,31
df1	2
df2	3
α	0,05
F-crit	9,55
p-hodnota	0,389

# Vícerozměrná analýza rozptylu

# Analýza rozptylu (ANOVA) jednoduchého třídění

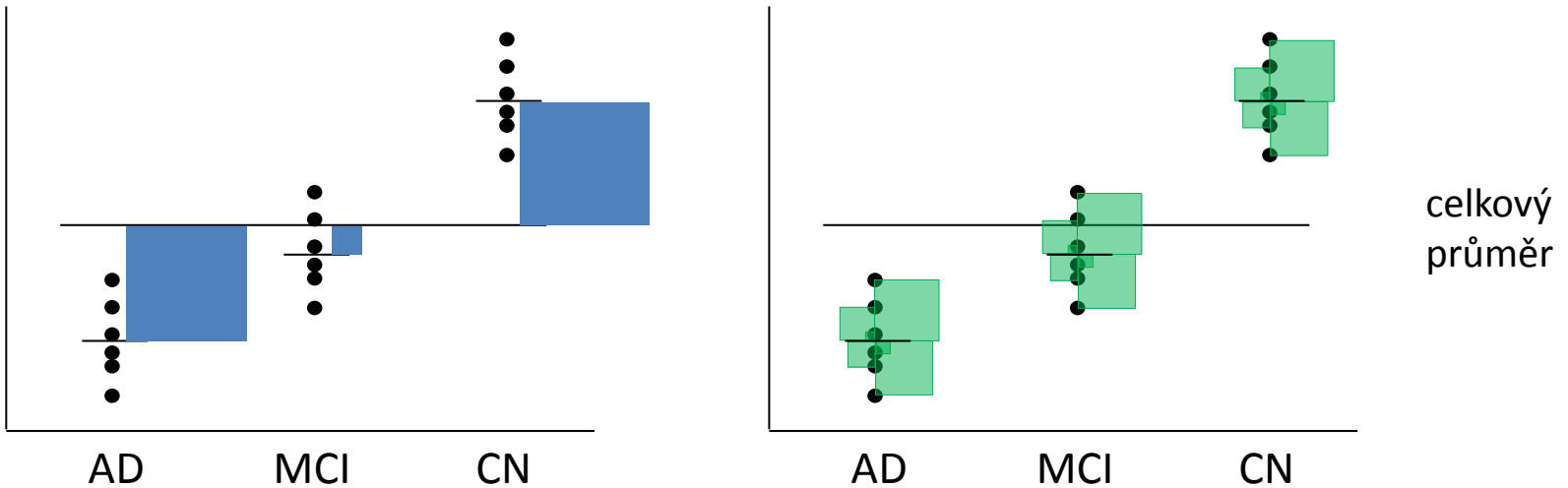
- Srovnáváme tři a více skupin dat, které jsou na sobě nezávislé (mezi objekty neexistuje vazba).
- Příklady: srovnání objemu hipokampu u pacientů s AD, pacientů s MCI a kontrol; srovnání kognitivního výkonu podle čtyř kategorií věku.



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

# 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 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 annotations for the model equation:
 

- Blue arrow from  $\mu$  to "Populační průměr"
- Blue arrow from  $\alpha_i$  to " $i$ -tý efekt faktoru A"
- Blue arrow from  $\beta_j$  to " $j$ -tý efekt faktoru B"
- Red arrow from  $\gamma_{ij}$  to "Interakce"
- Blue arrow from  $e_{ij}$  to "Reziduum"

- 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	$\rho$ -hodnota
Faktor A	$S_A$	$df_A = k - 1$	$MS_A = S_A / df_A$	$F_A$	$\rho$
Faktor B	$S_B$	$df_B = r - 1$	$MS_B = S_B / df_B$	$F_B$	$\rho$
Interakce AxB	$S_{AB}$	$df_{AB} = (k - 1)(r - 1)$	$MS_{AB} = S_{AB} / df_{AB}$	$F_{AB}$	$\rho$
Rezidua	$S_e$	$df_e = n - kr$	$MS_e = S_e / df_e$		
Celkem	$S_T$	$df_T = n - 1$			

# 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	1
1	2	1
1	3	6
2	1	3
2	2	4
2	3	9

# Úkol 2 - řešení

Pohlaví	Typ léku	Počet uzdrav. pacientů	
1	1	$X_{1..} = 8$ $M_{1..} = 8/3$	$\left\{ \begin{array}{l} 1 \\ 1 \\ 6 \end{array} \right.$
1	2		
1	3		
2	1	$X_{2..} = 16$ $M_{2..} = 16/3$	$\left\{ \begin{array}{l} 3 \\ 4 \\ 9 \end{array} \right.$
2	2		
2	3		

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

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

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

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

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

**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

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

**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)