Vícerozměrné metody - cvičení

RNDr. Eva Koriťáková

Podzim 2016

© Institut biostatistiky a analýz

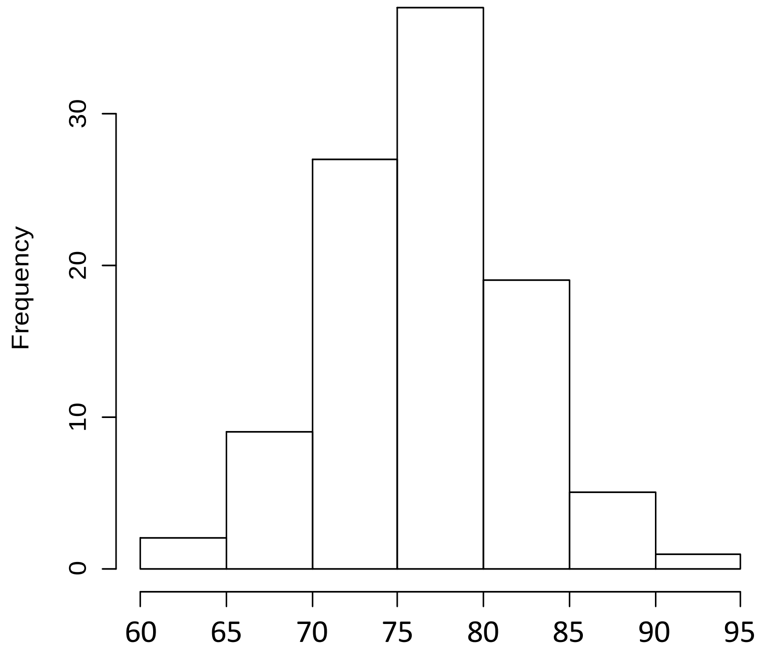
Cvičení 2

Vícerozměrné normální rozdělení
a vícerozměrný t-test

Vícerozměrné normální rozdělení

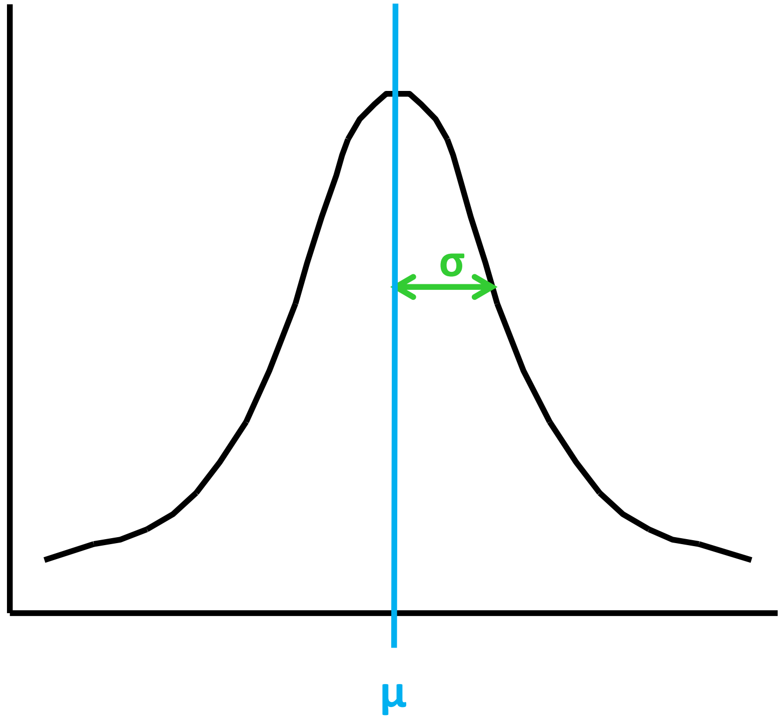
Motivace

Histogram



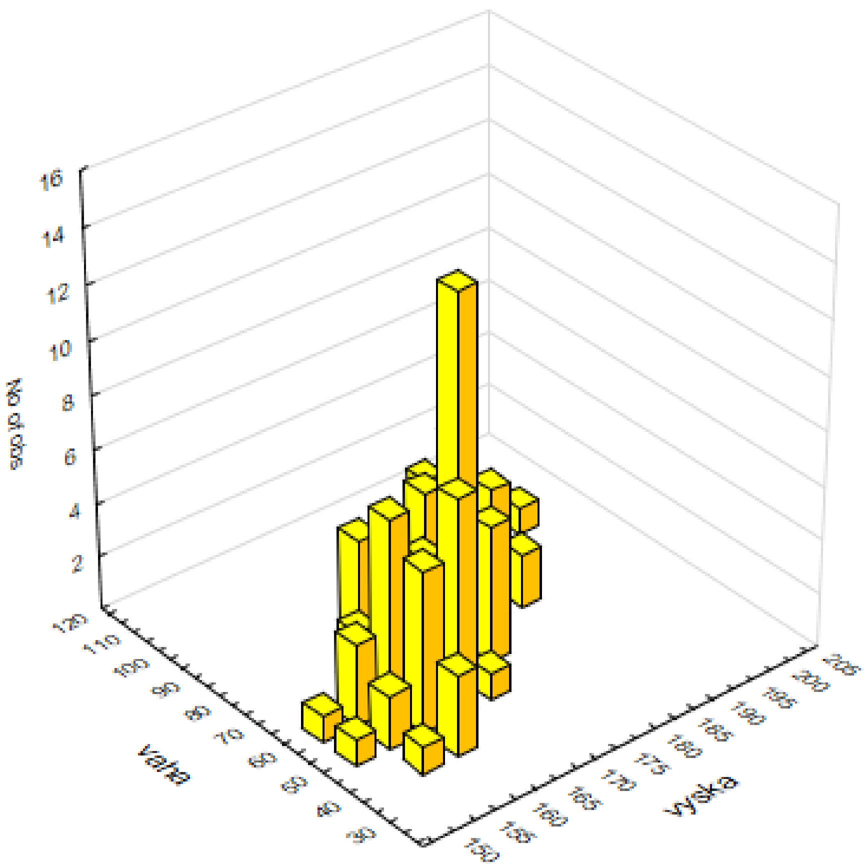
Diastolický tlak

Hustota jednorozměrného normálního rozdělení

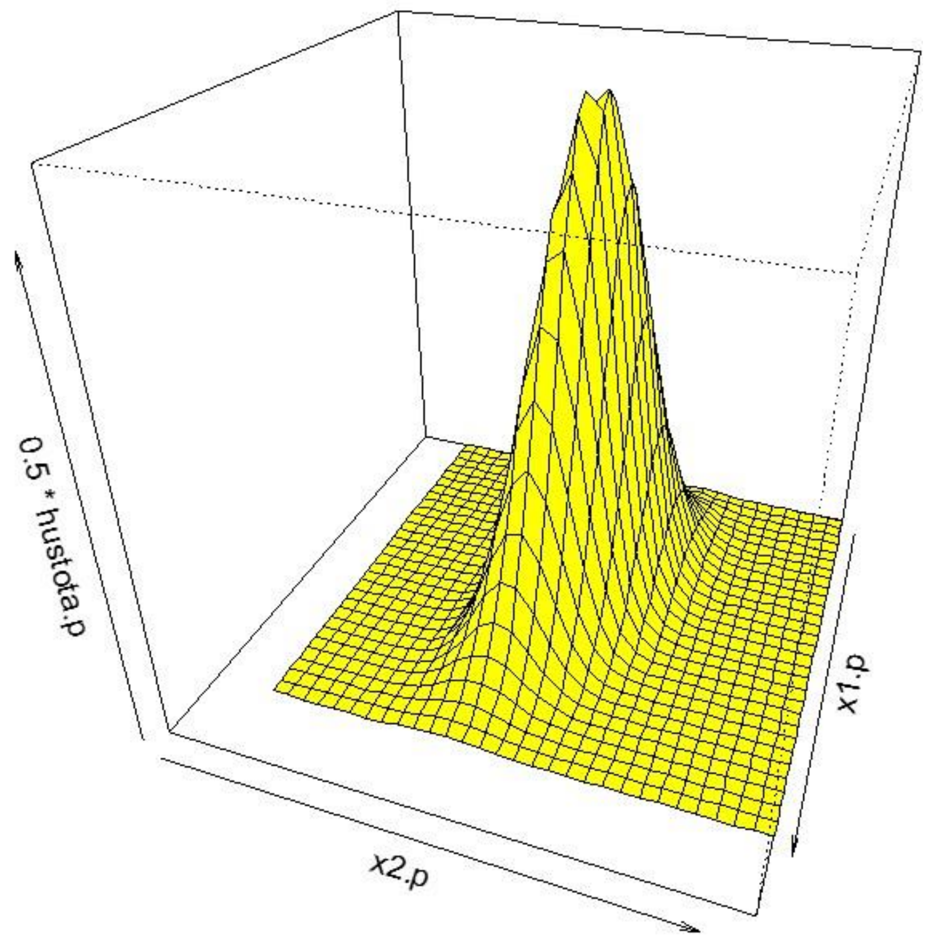


Motivace – pokračování

Dvourozměrný histogram



Hustota dvourozměrného normálního rozdělení



Vícerozměrné normální rozdělení

Hustota jednozměrného normálního rozdělení:

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

μ - střední hodnota σ^2 – rozptyl

Hustota vícerozměrného normálního rozdělení:

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

μ - vektor středních hodnot **Σ** - kovarianční matice

Hustota dvourozměrného normálního rozdělení:

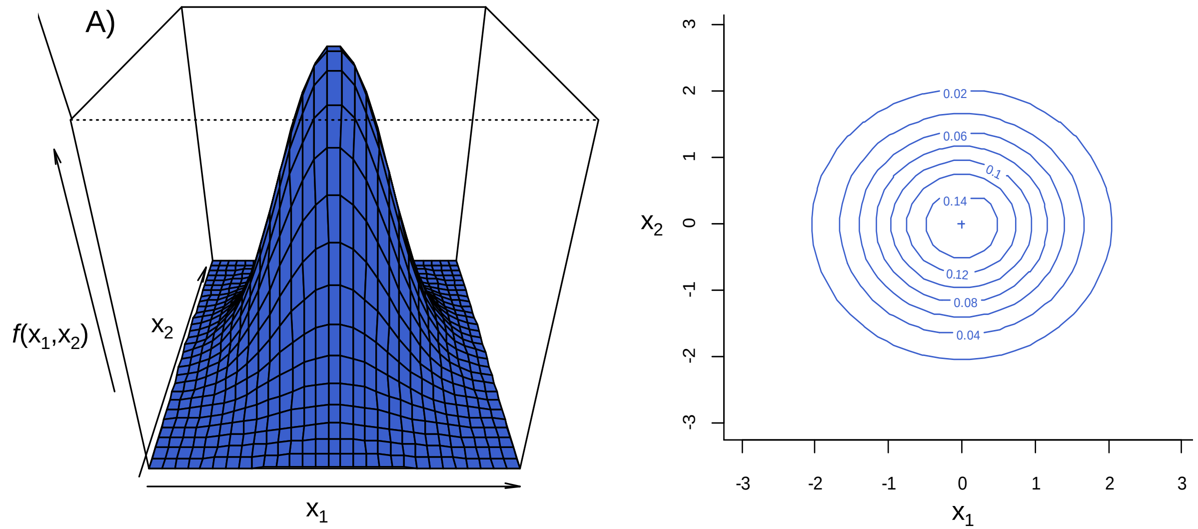
$$f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_x)^2}{\sigma_x^2} + \frac{(y-\mu_y)^2}{\sigma_y^2} - \frac{2\rho(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y}\right]\right),$$

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

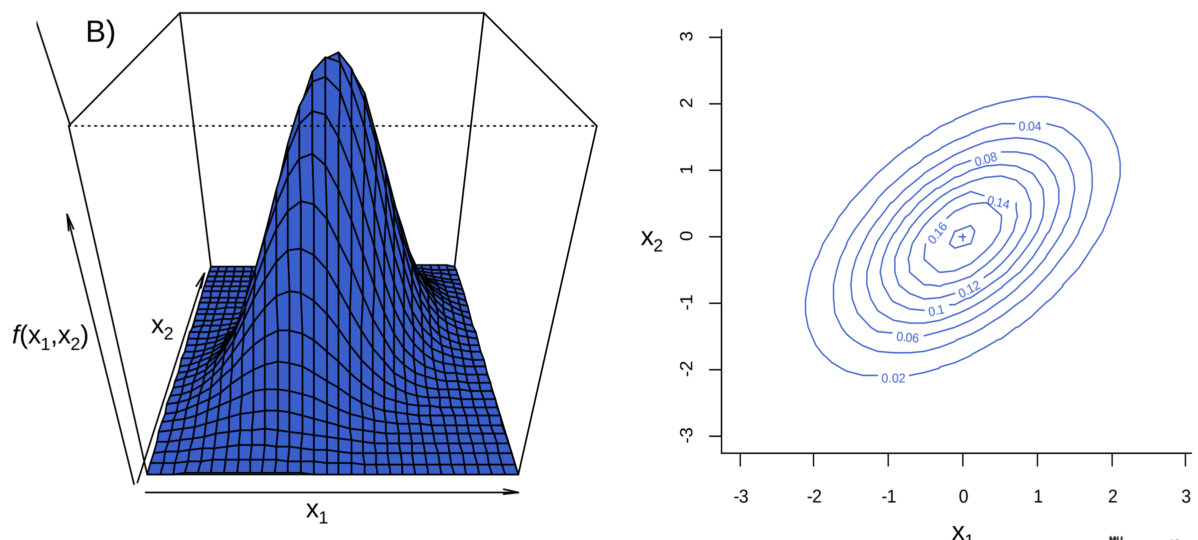
ρ - korelace mezi X a Y;
σ – směrodatná odchylka

Hustota u nekorelovaných a korelovaných proměnných

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



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



Vícerozměrný průměr a kovarianční matice

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

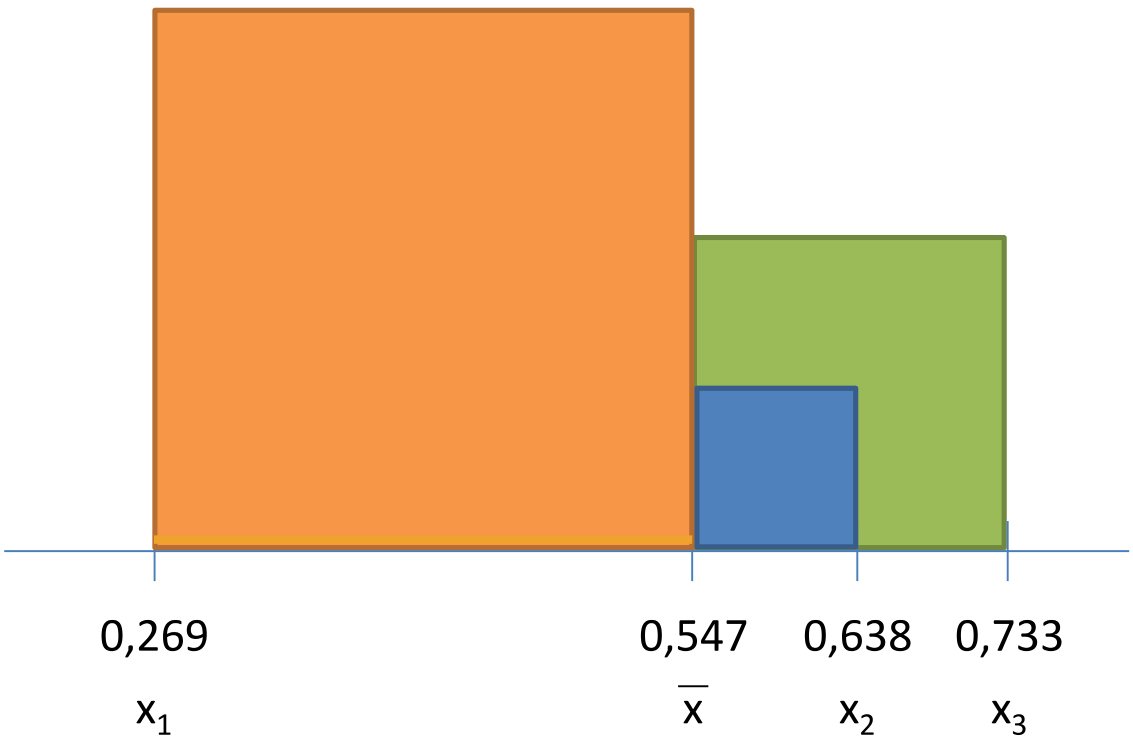
$$\bar{\mathbf{x}} = \left[\frac{1}{n}\sum_{i=1}^{n} \mathrm{x}_{i1} \quad \frac{1}{n}\sum_{i=1}^{n} \mathrm{x}_{i2}\right]$$

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

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

Výpočet rozptylu a směrodatné odchylky - opakování

- Příklad čtverců odchylek od průměru pro *n* = 3.
- Rozptyl je možno značně ovlivnit odlehlými pozorováními.



Rozptyl:

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

Směrodatná odchylka:

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

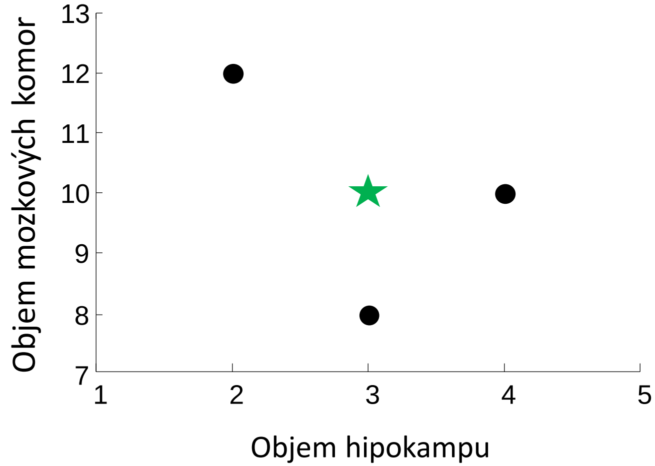
Úkol 1

- Spočtěte vícerozměrný průměr a výběrovou kovarianční matici pro soubor 3 subjektů, u nichž byly naměřeny hodnoty objemu hipokampu a mozkových komor, přičemž naměřené hodnoty byly zaznamenány do následující datové matice:

$$\mathbf{X} = \begin{bmatrix} 2 & 12 \\ 4 & 10 \\ 3 & 8 \end{bmatrix}$$

Úkol 1 - řešení

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



Vícerozměrný průměr:

$$\bar{\mathbf{x}} = \begin{bmatrix} \frac{1}{n}\sum_{i=1}^{n} \mathrm{x}_{i1} & \frac{1}{n}\sum_{i=1}^{n} \mathrm{x}_{i2} \end{bmatrix} = \begin{bmatrix} \frac{1}{3}(2+4+3) & \frac{1}{3}(12+10+8) \end{bmatrix} = \begin{bmatrix} 3 & 10 \end{bmatrix}$$

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

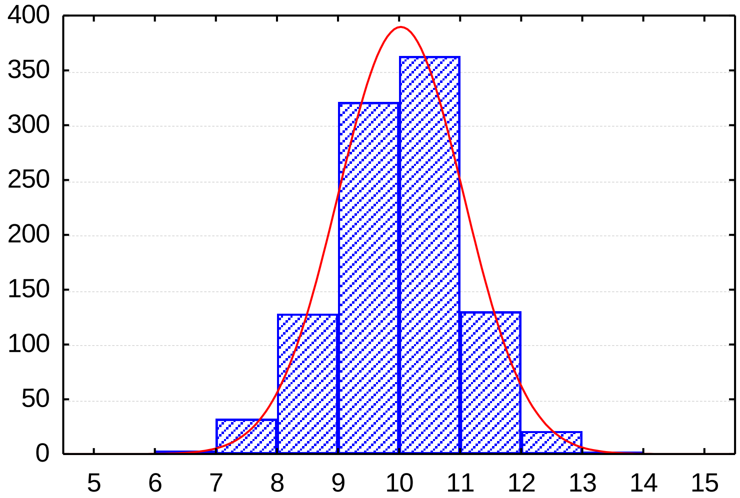
$$s_{11} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathrm{x}_{i1} - \bar{\mathrm{x}}_1)^2 = \frac{1}{3-1}\left((2-3)^2 + (4-3)^2 + (3-3)^2\right) = \frac{1}{2}(1+1+0) = 1$$

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

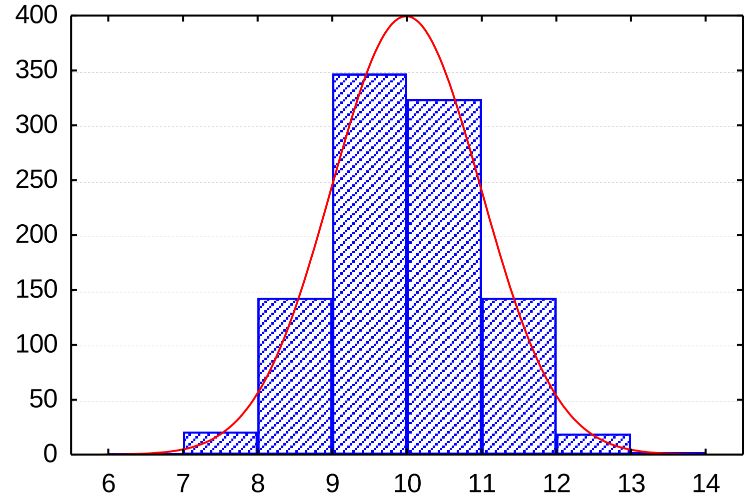
$$s_{21} = s_{12} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathrm{x}_{i1} - \bar{\mathrm{x}}_1)(\mathrm{x}_{i2} - \bar{\mathrm{x}}_2) = \frac{1}{3-1}\left((2-3)(12-\right.$$

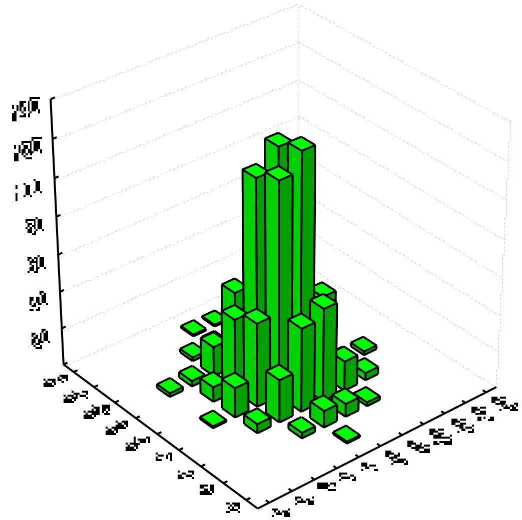
$$\rightarrow \mathbf{S} = \begin{bmatrix} 1 & -1 \\ -1 & 4 \end{bmatrix}$$

Je normalita v jednorozměrném prostoru jedinou podmínkou vícerozměrné normality?

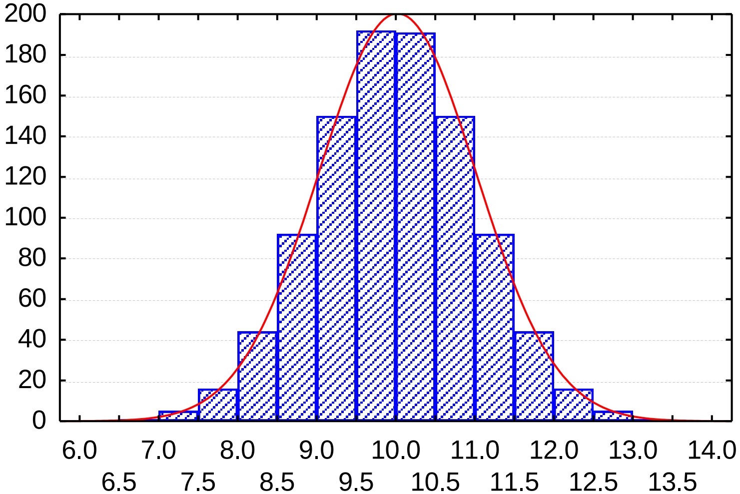


+

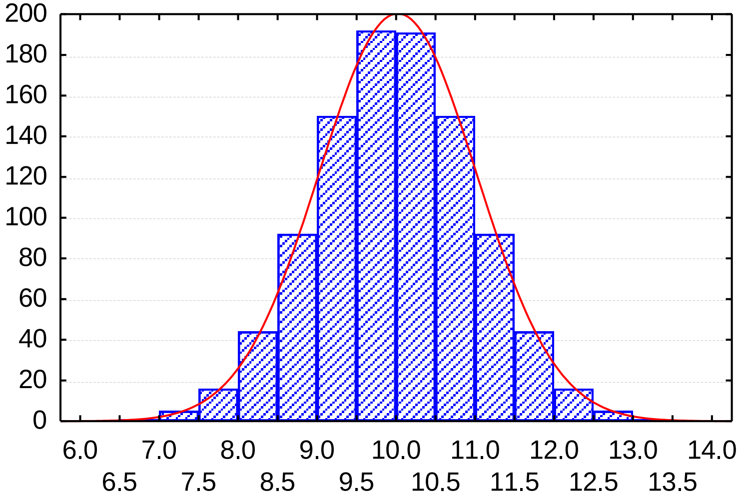


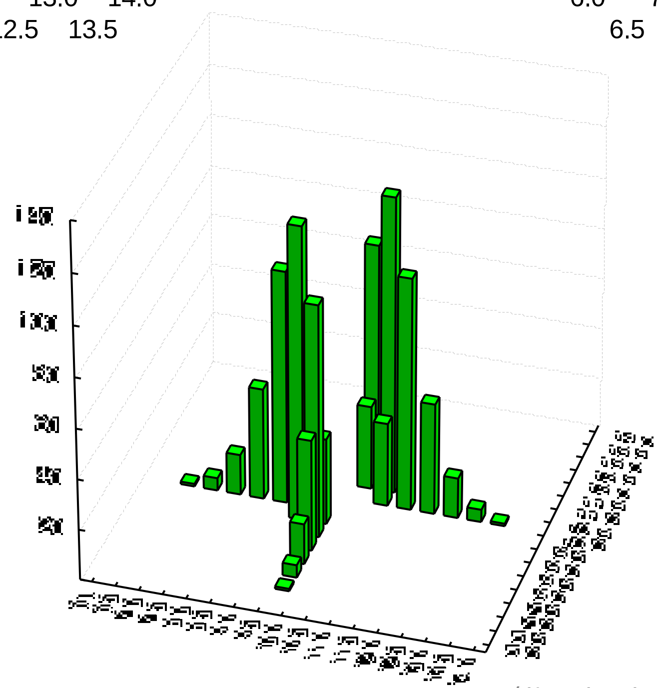


Je normalita v jednorozměrném prostoru jedinou podmínkou vícerozměrné normality?

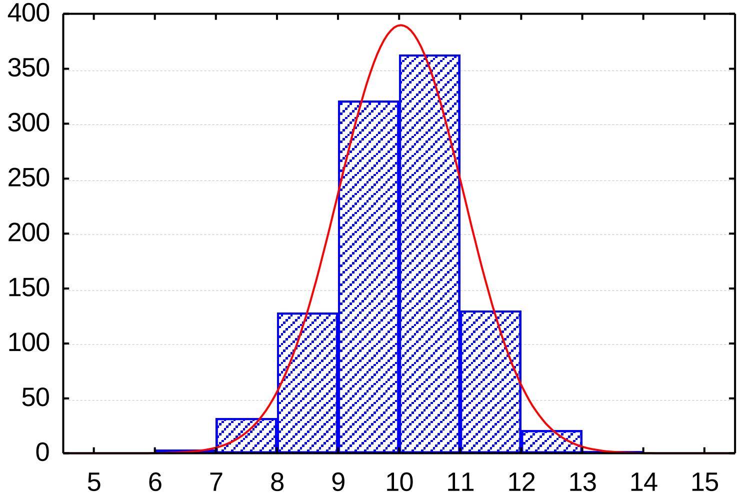


+

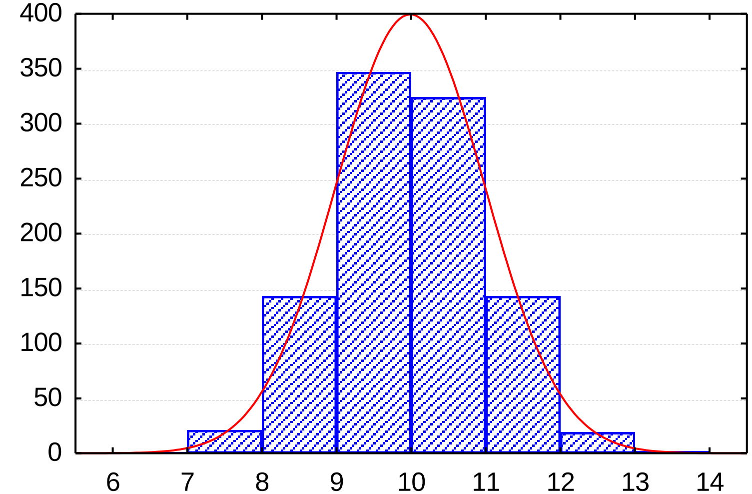


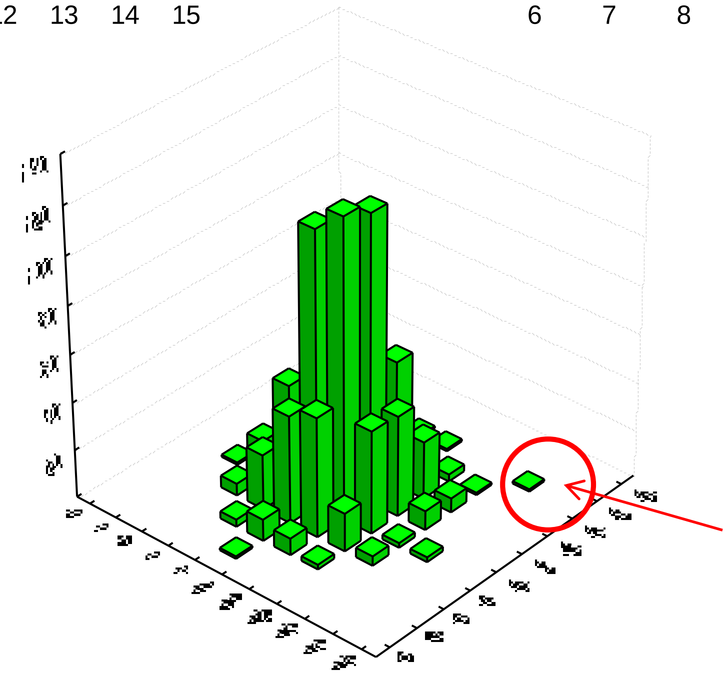


Je normalita v jednorozměrném prostoru jedinou podmínkou vícerozměrné normality?



+

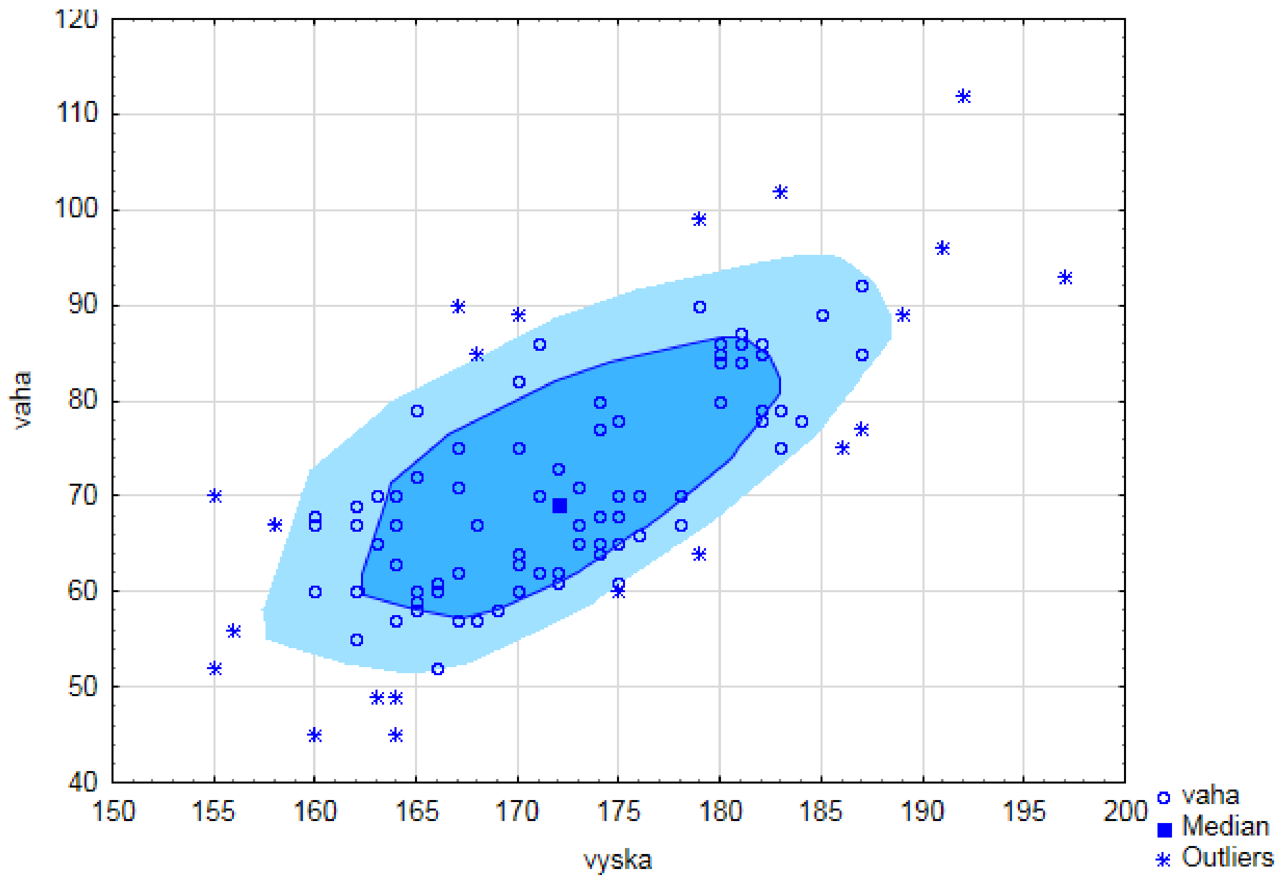




Vícerozměrná odlehlá hodnota (outlier)

Ověření dvourozměrné normality

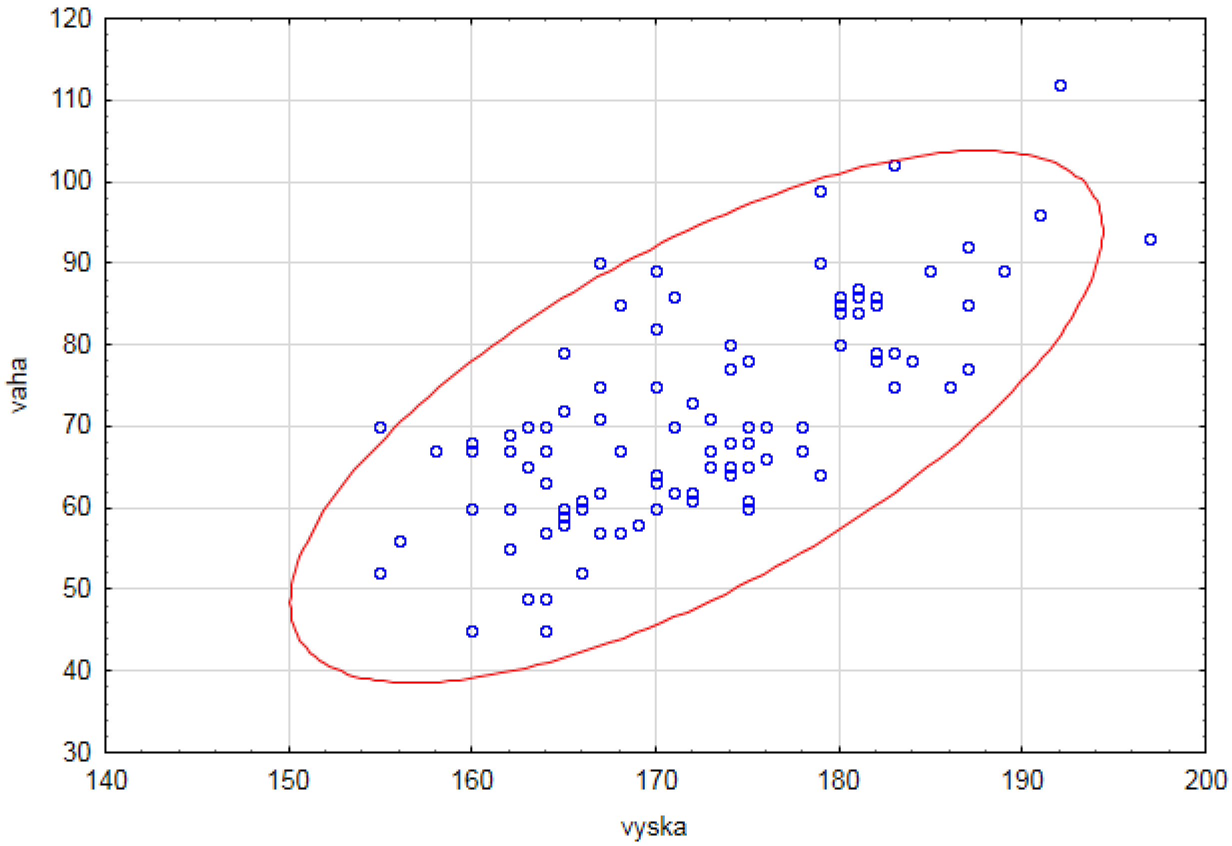
Bagplot = „bivariate boxplot“ (tzn. „dvourozměrný krabicový graf“)



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

Ověření dvourozměrné normality

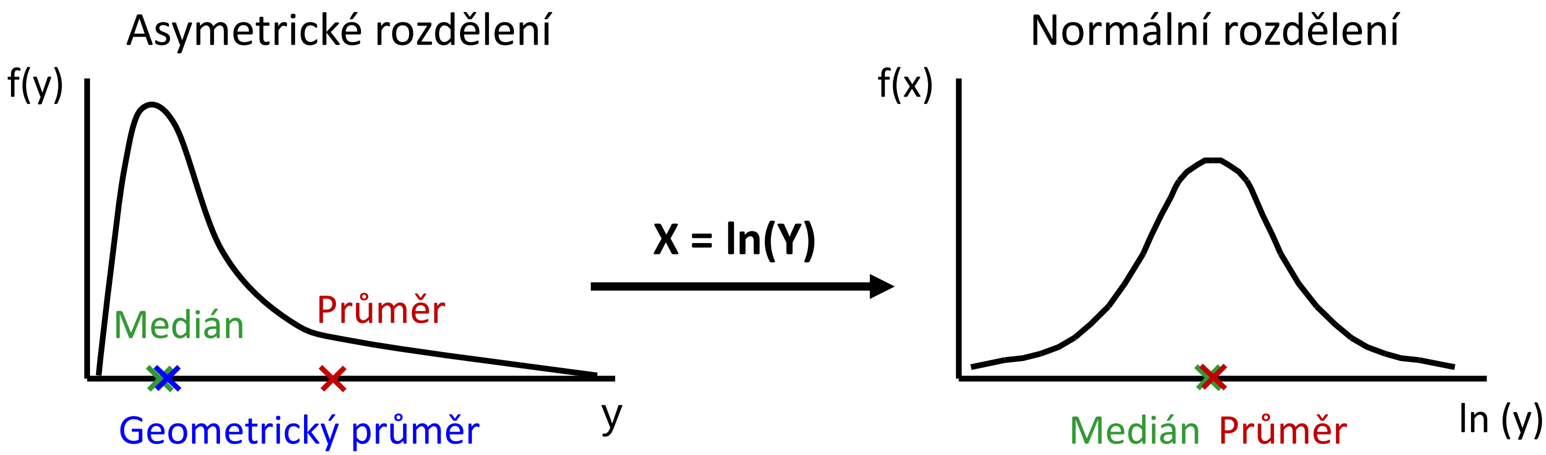
Vykreslení regulační elipsy („control“ elipse):



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

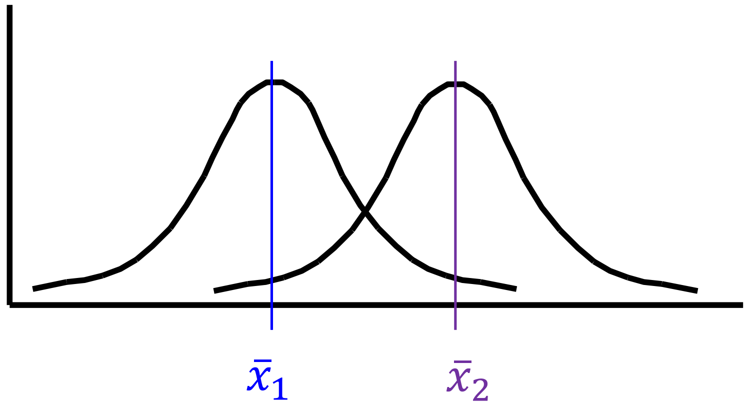


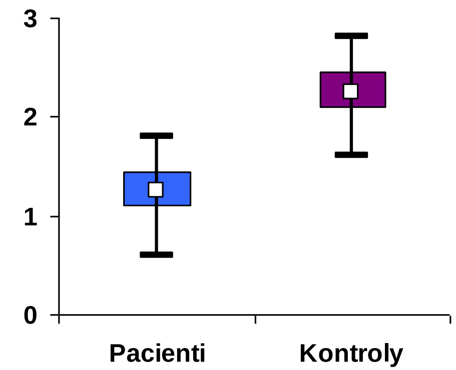
- Další příklady:
  - **odmocninová transf.** (pro proměnné s Poissonovým rozložením nebo obecně data typu počet jedinců, buněk apod.: $X = \sqrt{Y}$ nebo $X = \sqrt{Y+1}$
  - **arcsin transfomace** (pro proměnné s binomickým rozložením)
  - **Box-Coxova tranformace**

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,...



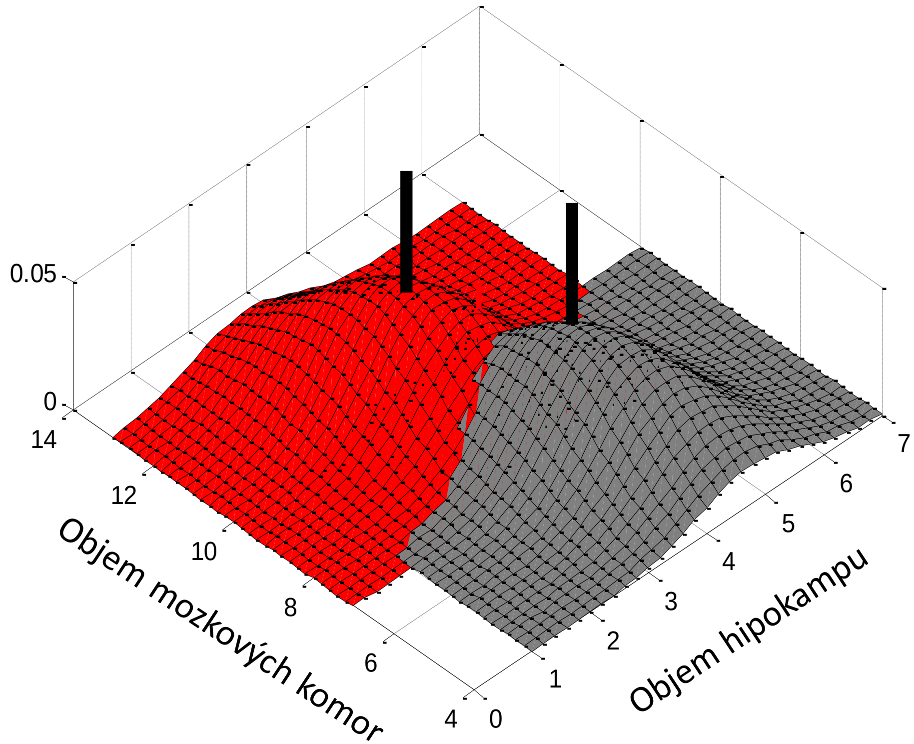


- Předpoklad: **normalita dat v OBOU skupinách, shodnost (homogenita) rozptylů** v obou skupinách
- Testová statistika: $t = \dfrac{\bar{x}_1 - \bar{x}_2 - c}{s_* \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$, kde s_* je vážená směrodatná odchylka,

  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:

Studentovo rozdělení

- testová statistika: $T = \frac{(\bar{x}_D - \bar{x}_H) - c}{s_* \sqrt{\frac{1}{n_D} + \frac{1}{n_H}}}$, kde $T \sim t(n_D + n_H - 2)$
- 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 $\mathrm{c} = 0$)
- nulová hypotéza zamítnuta, pokud $|T| > t_{1-\alpha/2}(n_D + n_H - 2)$

Je ekvivalentní testu:

F rozdělení

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

Vícerozměrný t-test:

- Hotellingova T^2 testová statistika: $T^2 = (\bar{\boldsymbol{x}}_D - \bar{\boldsymbol{x}}_H - \boldsymbol{c})^T \left[\mathbf{S}_* \left(\frac{1}{n_D} + \frac{1}{n_H} \right) \right]^{-1} (\bar{\boldsymbol{x}}_D - \bar{\boldsymbol{x}}_H - \boldsymbol{c})$
- kde $\mathbf{S}_*$ je vážená kovarianční matice: $\mathbf{S}_* = \frac{(n_D - 1)\mathbf{S}_D + (n_H - 1)\mathbf{S}_H}{(n_D - 1) + (n_H - 1)}$

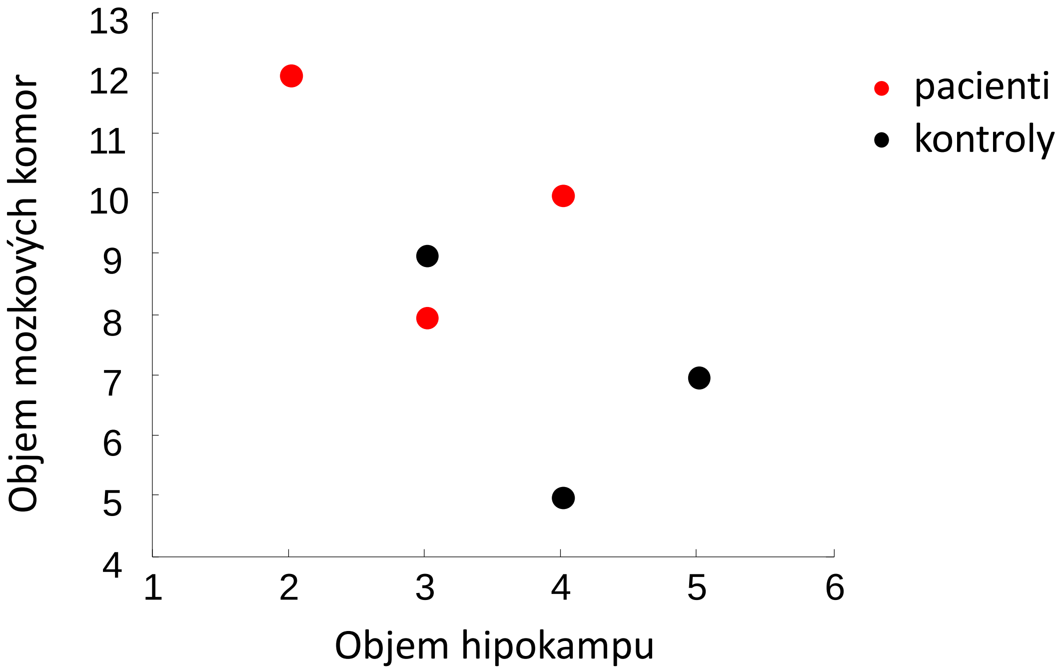
Hotellingovo rozdělení

- $T^2 \sim T^2(p, n-p-1)$; pro malé n_D a n_H je lepší použít: $F = \frac{n-p-1}{p} \frac{T^2}{n-2}$, kde $n = n_D + n_H$
- nulová hypotéza zamítnuta, když $F > F_{1-\alpha}(p, n-p-1)$ ← F rozdělení

Úkol 2

- 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 2

- 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 2 - řešení

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

Vícerozměrné průměry: $\bar{\mathbf{x}}_D =$

$$\left[\frac{1}{n_D}\sum_{i=1}^{n_D} \mathrm{x}_{i1} \quad \frac{1}{n_D}\sum_{i=1}^{n_D} \mathrm{x}_{i2}\right] = [3 \quad 10]$$

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

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 2 - řešení

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

Vícerozměrné průměry: $\overline{\mathbf{x}}_D = \left[\frac{1}{n_D} \sum_{i=1}^{n_D} \mathrm{x}_{i1} \quad \frac{1}{n_D} \sum_{i=1}^{n_D} \mathrm{x}_{i2} \right] = [3 \quad 10]$

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

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
α	0,05
F-crit	9,55
p-hodnota	0,389

Úkol 2 – řešení v softwaru R

```
install.packages("ICSNP")
library("ICSNP")
Xd=matrix(c(2,4,3,12,10,8),3,2)
Xh=matrix(c(5,3,4,7,9,5),3,2)
HotellingsT2(Xd, Xh)
```

```
        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)
```

Použití softwaru R jako kalkulačky:

```
S=solve(2/3*matrix(c(1,-1,-1,4),2,2))  # výpočet inverzní matice
b=matrix(c(-1,3),1,2)                  # vektor s hodnotami rozdílu souřadnic centroidů
t2=b%*%S%*%t(b)                        # výpočet testové statistiky T^2
F=(3/2)*(t2/4)                         # výpočet testové statistiky F
qf(0.95,2,3)                           # 95% kvantil F rozdělení pro stupně volnosti 2 a 3
1-pf(F,2,3)                            # p-hodnota
```