Vícerozměrné statistické metody

Vícerozměrné statistické rozdělení a testy, operace s vektory a maticemi

Jiří Jarkovský, Simona Littnerová

Vícerozměrné statistické metody

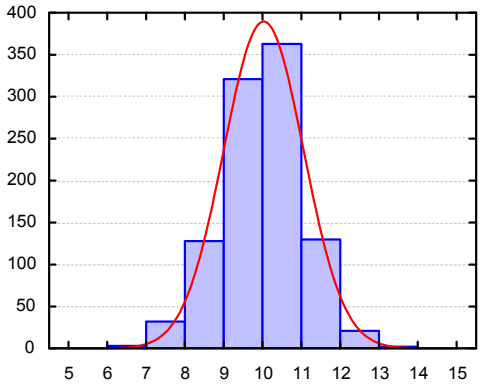
Vícerozměrné statistické rozdělení a testy

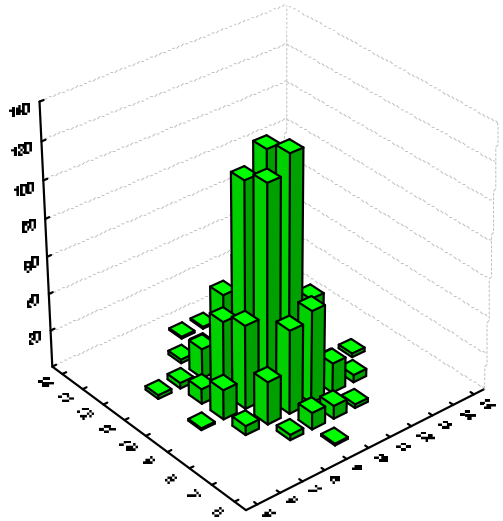
Význam rozdělení ve vícerozměrném prostoru

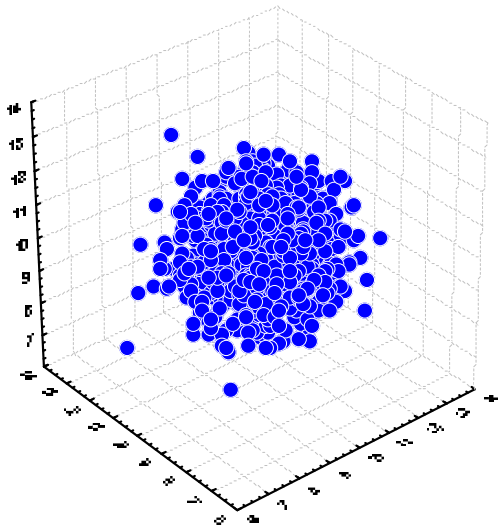
- Použitelnost mnohých klasických statistických metod a postupů vyžaduje předpoklad o normálním rozdělení sledovaných proměnných.
- Podmínka normality vyplývá z toho, že metody založené na tomto předpokladu mohou využít kompletní matematický aparát schovaný za danou statistickou metodou. Tyto metody jsou také relativně snadno pochopitelné a se získanými řešeními se dobře pracuje.
- Ovšem v reálném světě bývá obtížné předpoklad o normálním rozložení dodržet, v mnohých oblastech přírodních a mnohdy i technických oborů není tento předpoklad samozřejmostí.
- Předpokládejme však normalitu a předpoklad o jedné normálně rozložené náhodné proměnné můžeme rozšířit na předpoklad simultánního normálního rozložení dvou a více náhodných proměnných. Některé vícerozměrné postupy a metody vycházejí z předpokladu vícerozměrného normálního rozdělení. Vícerozměrné normální rozdělení může být také velmi užitečnou aproximací různých jiných simultánních rozdělení.

Rozdělení dat ve vícerozměrném prostoru

- Klasická jednorozměrná rozdělení a testy mají svůj protějšek ve vícerozměrném prostoru; analogii lze nalézt v podstatě ke každému z nich
- Obrázky zobrazují 1D, 2D a 3D normální rozdělení
- Při popisu vícerozměrných dat se uplatňují stejné charakteristiky jako při popisu dat jednorozměrných, nicméně nyní již ne jako jedno číslo, ale jako vektor







Pojmy popisu vícerozměrných rozdělení

- Centroid
  - průměr nebo medián nebo jiná charakteristika středu spočtená pro všechny dimenze
  - Je popsán vektorem charakteristik středu
  - Používán jako popisná statistika nebo i jako součást výpočtu shlukovacích metod
  - „virtuální střed vícerozměrného shluku“

- Medoid
  - Medoid je reprezentativní objekt datového souboru nebo shluku v datech, jehož průměr podobnosti od všech ostatních objektů v datech nebo ve shluku je minimální.
  - Medoid má podobný význam jako průměr nebo centroid, jen je vždy reprezentován reálným objektem z datového souboru.
  - Medoid bývá nejčastěji používán tam, kde není definován průměr nebo centroid (např. tří a vícerozměrný prostor). Tento termín se používá při shlukové analýze.

Vícerozměrné charakteristiky rozdělení

- Základní charakteristikou vícerozměrného rozdělení je vektor středních hodnot (vektor průměrů)

$$\mathrm{E}(\mathbf{X}) = \begin{bmatrix} \mathrm{E}(X_1) \\ \mathrm{E}(X_2) \\ \vdots \\ \mathrm{E}(X_p) \end{bmatrix}$$

- a kovariační matice

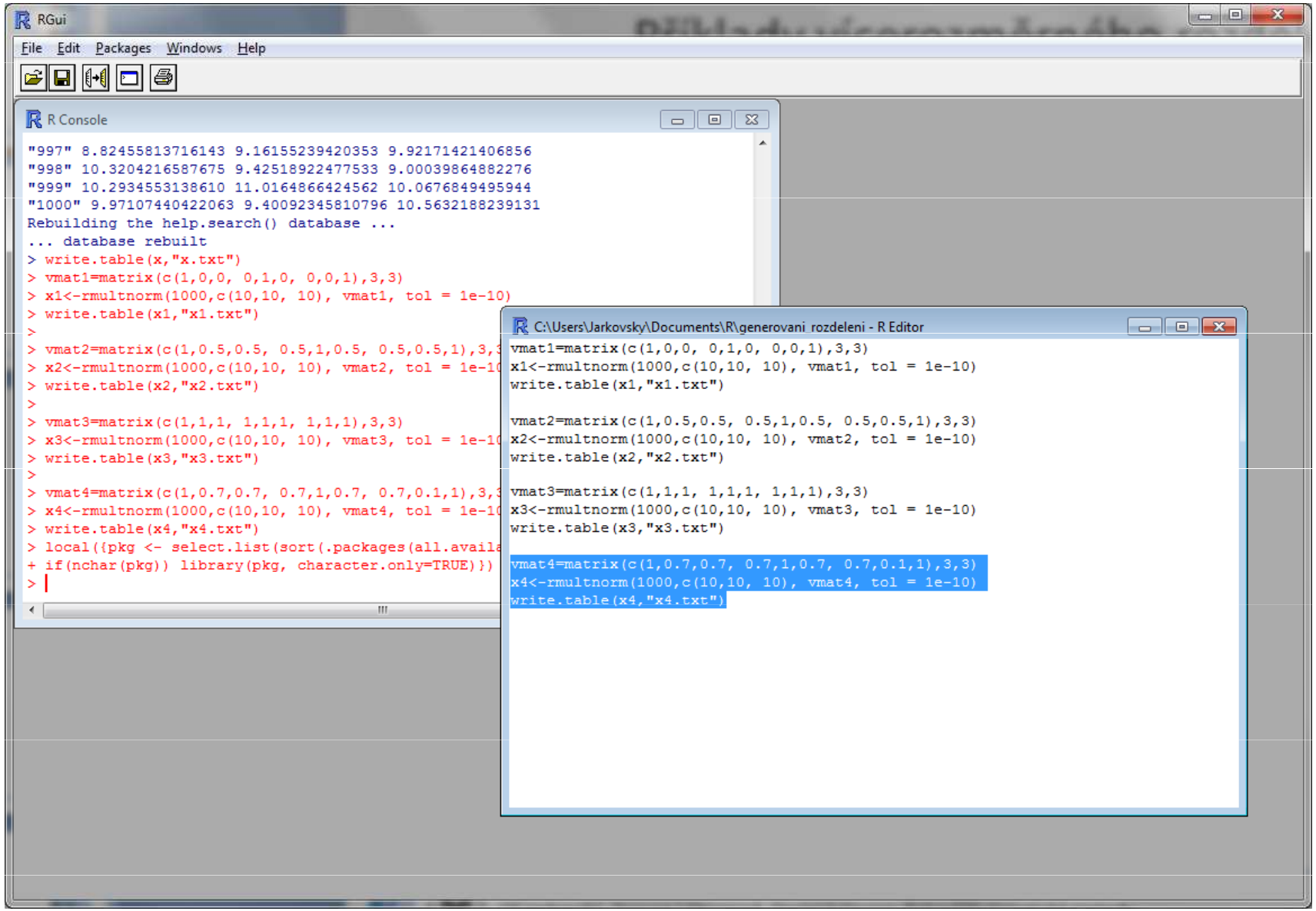
$$\mathbf{\Sigma} = \mathrm{var}(\mathbf{X}) = \mathrm{cov}(\mathbf{X}) = \begin{pmatrix} \sigma_1^2 & \sigma_1\sigma_2 & \cdots & \sigma_1\sigma_p \\ \sigma_2\sigma_1 & \sigma_2^2 & \cdots & \sigma_2\sigma_p \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_p\sigma_1 & \sigma_p\sigma_2 & \cdots & \sigma_p^2 \end{pmatrix}$$

- kde je σ_{ij} kovariance dvou náhodných veličin, tj.

$$\sigma_{ij} = \mathrm{cov}(X_i, X_j) = \mathrm{E}(X_i - \mathrm{E}(X_i))(X_j - \mathrm{E}(X_j))$$

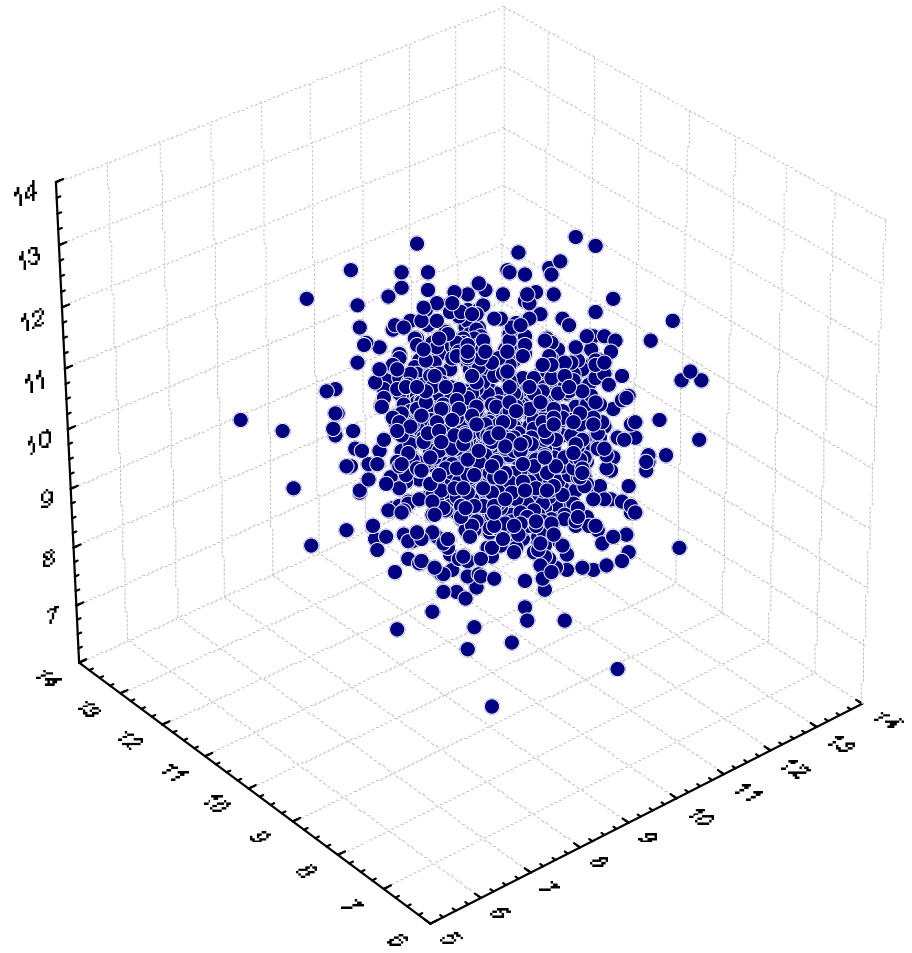
Příklady vícerozměrného rozdělení

- R – knihovna MSBVAR



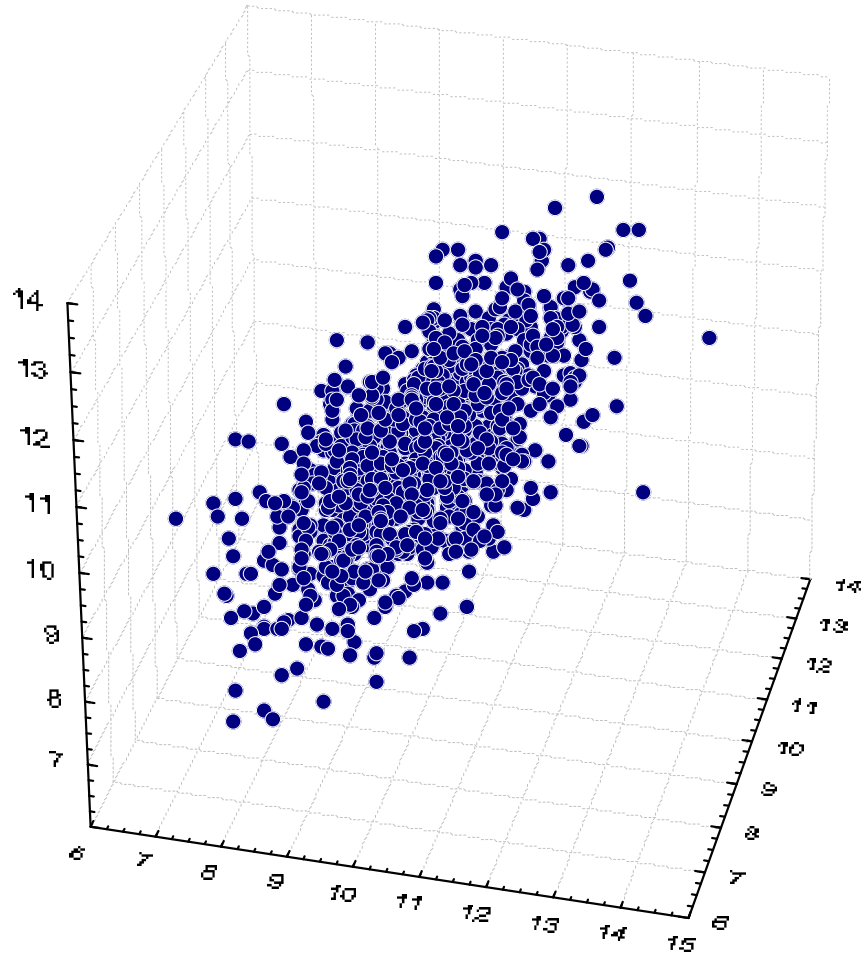
Příklad vícerozměrného rozdělení I

```
vmat1=matrix(c(1,0,0, 0,1,0, 0,0,1),3,3)
x1<-rmultnorm(1000,c(10,10, 10), vmat1, tol = 1e-10)
write.table(x1,"x1.txt")
```



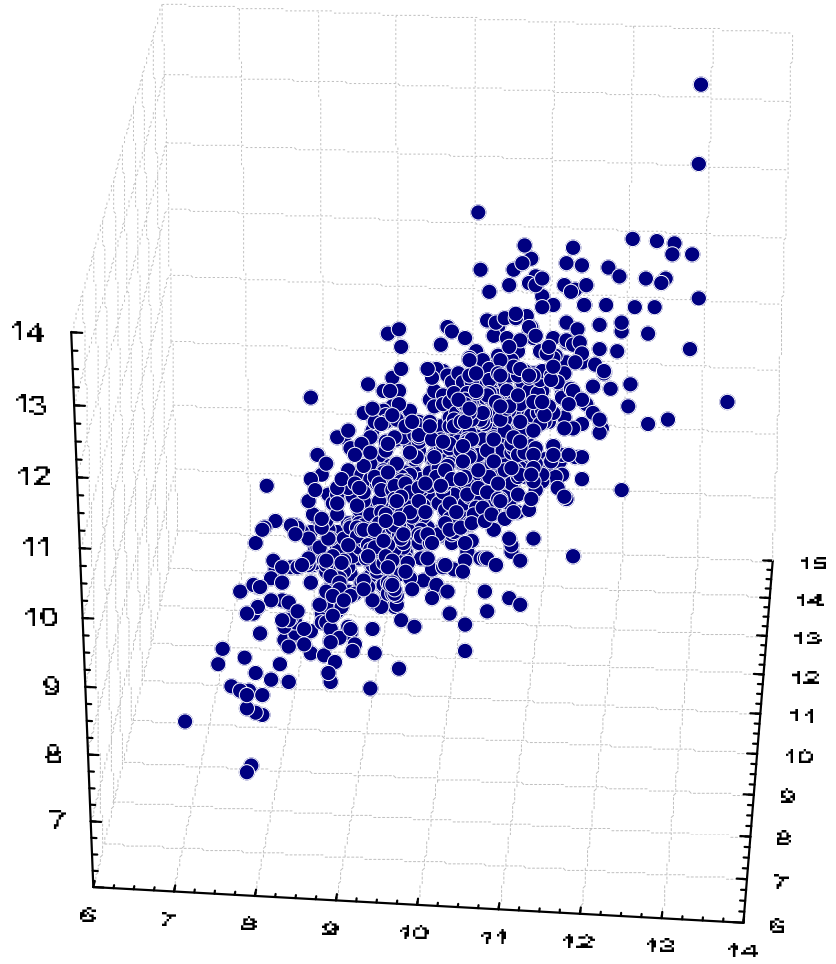
Příklad vícerozměrného rozdělení II

```
vmat2=matrix(c(1,0.5,0.5, 0.5,1,0.5, 0.5,0.5,1),3,3)
x2<-rmultnorm(1000,c(10,10, 10), vmat2, tol = 1e-10)
write.table(x2,"x2.txt")
```



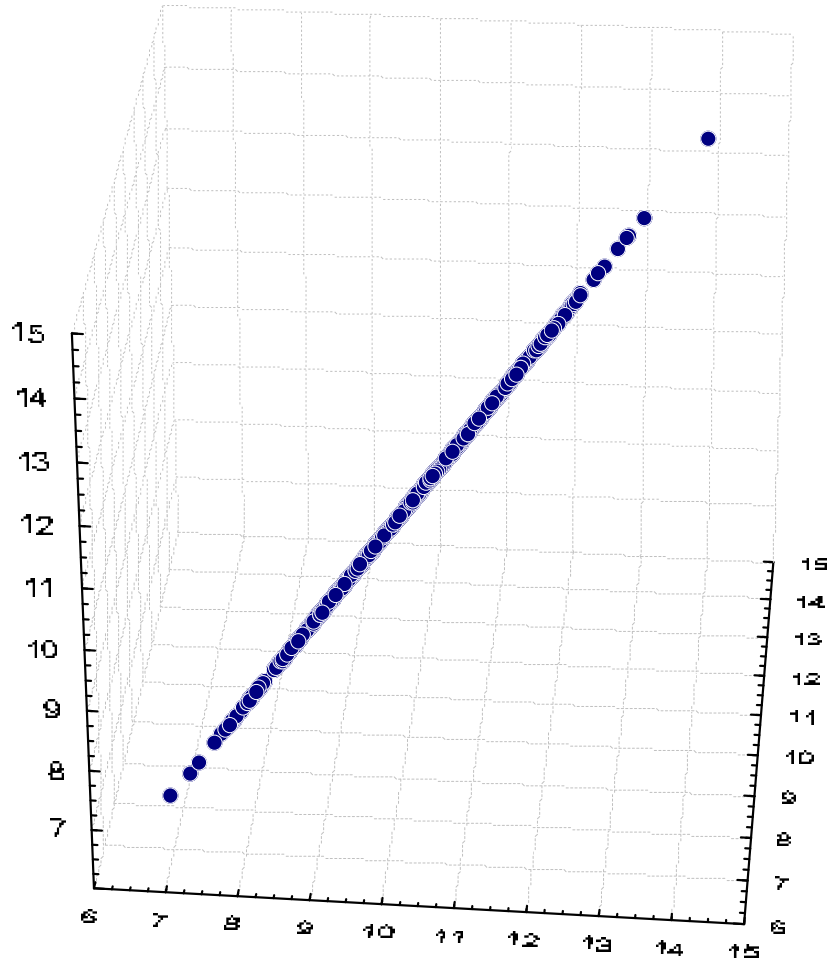
Příklad vícerozměrného rozdělení III

```
vmat4=matrix(c(1,0.7,0.7, 0.7,1,0.7, 0.7,0.1,1),3,3)
x4<-rmultnorm(1000,c(10,10, 10), vmat4, tol = 1e-10)
write.table(x4,"x4.txt")
```



Příklad vícerozměrného rozdělení IV

```
vmat3=matrix(c(1,1,1, 1,1,1, 1,1,1),3,3)
x3<-rmultnorm(1000,c(10,10, 10), vmat3, tol = 1e-10)
write.table(x3,"x3.txt")
```



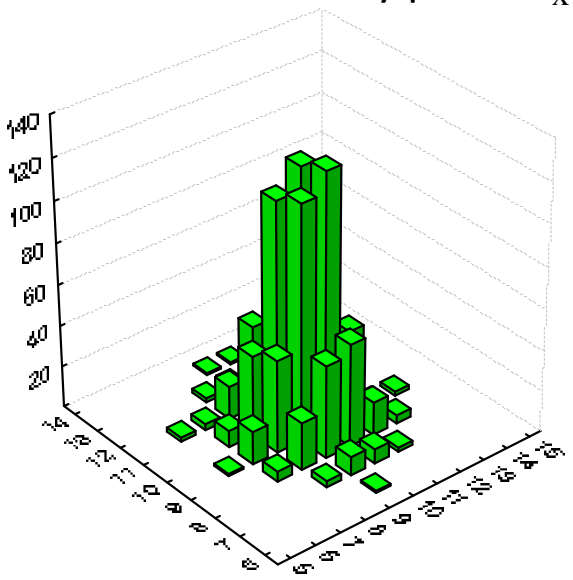
Wishartovo rozdělení

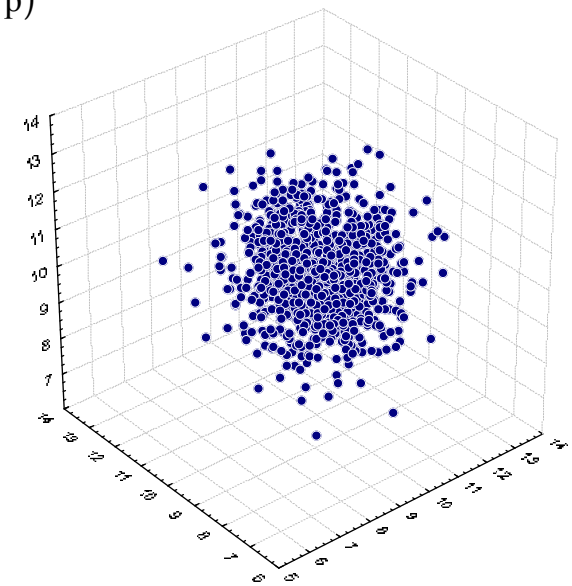
- Wishartovo rozdělení je vícerozměrným zobecněním chi-square rozdělení
- Při odvození některých důležitých algoritmů ve vícerozměrné statistické analýze se uplatňuje dále uvedená vlastnost Wishartova rozdělení.
- Součet nezávislých náhodných matic s Wishartovým rozdělením se shodnou střední hodnotou je rovněž Wishartovo rozdělení se stejnou střední hodnotou, přičemž stupně volnosti se sčítají.

$$\left.\begin{array}{c}\mathbf{A} = \mathbf{A_1} + \mathbf{A_2} + \ldots + \mathbf{A}_H \\ \mathbf{A}_h \sim W_p(\nu_h, \boldsymbol{\Sigma}), h = 1,2,\ldots,H\end{array}\right\} \longrightarrow \mathbf{A}_h \sim W_p\left(\sum_{h=1}^{H} \nu_h, \boldsymbol{\Sigma}\right)$$

Hotellingovo rozdělení

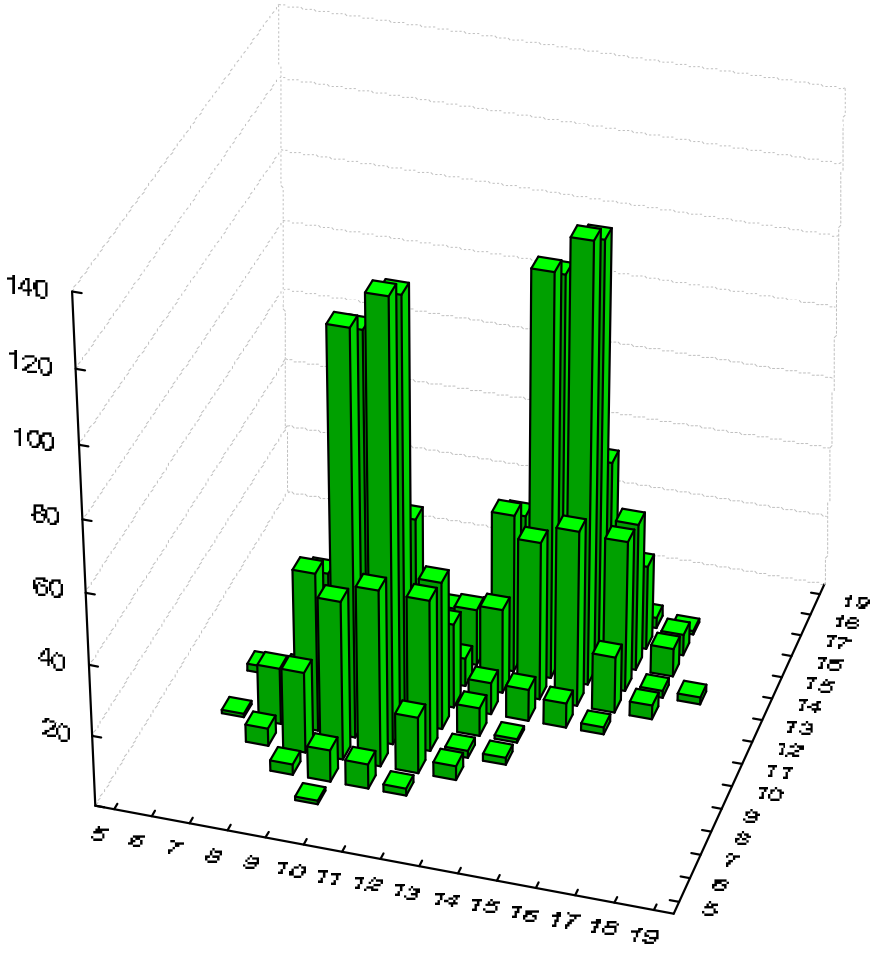
- Jedná se o zobecnění t- rozdělení pro *p*-rozměrný prostor
- Uvažujme regulární čtvercovou matici A *p*-tého řádu a rozdělením $W_p(\nu, \Sigma)$ a na A nezávislý *p*-položkový vektor a s rozdělením $N_p\left(\mathbf{o}_p, \frac{\Sigma}{c}\right)$ Potom kvadratická forma $Q_1 = c\nu\mathbf{a}^T\mathbf{A}^{-1}\mathbf{a}$ má Hotellingovo rozdělení T^2 (p, ν – p+1).
- V jednorozměrném normálním rozdělení se při testování hypotéz o střední hodnotě používá statistika (jednovýběrový t-test) $X \sim N(\mu, \sigma^2) \longrightarrow \frac{\bar{x} - \mu}{\sqrt{\frac{s^2(x)}{n}}} \sim t(n-1)$
- Druhou mocninu této statistiky můžeme upravit a zapsat ve tvaru $t^2 = n(\bar{x} - \mu)[s^2(x)]^{-1}(\bar{x} - \mu)$ Tento výraz odpovídá *p*-rozměrné statistice, vhodné k úsudku o μ, která má Hotellingovo rozdělení T^2 s p a n–p stupni volnosti, jedná se tedy o zobecnění t- rozdělení pro *p*-rozměrný prostor. Můžeme tedy psát $x \sim N_p(\mu, \Sigma) \longrightarrow n(x - \mu)^T S^{-1} \sim T^2(p, n - p)$

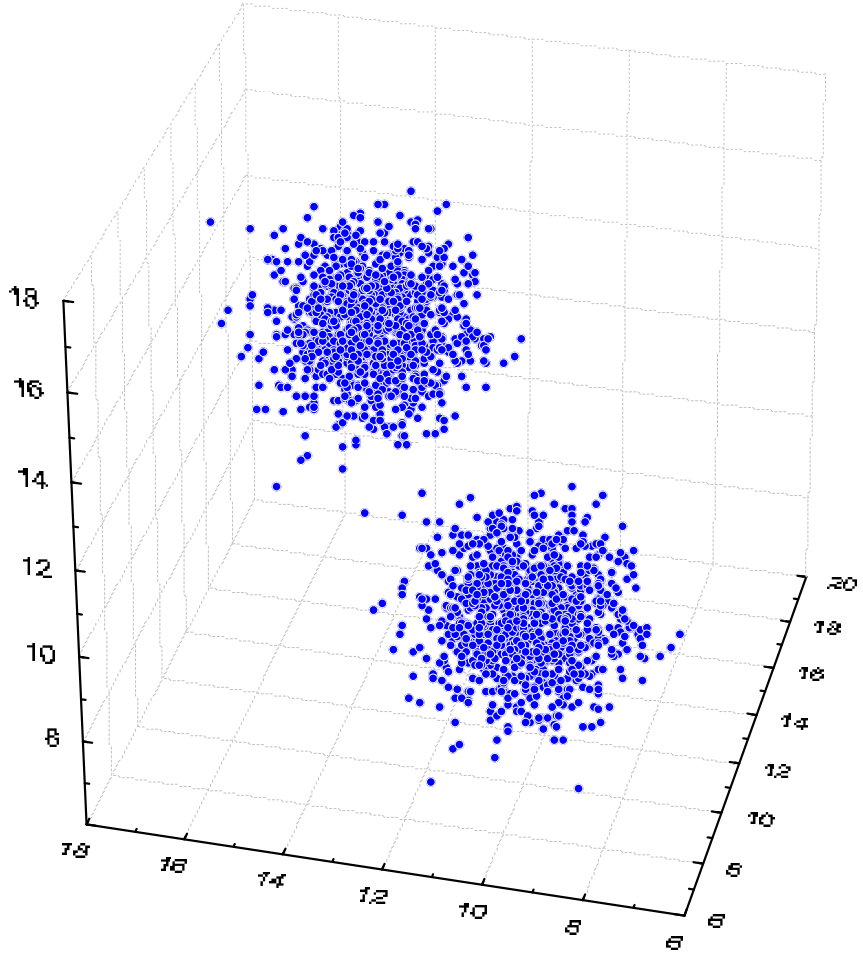




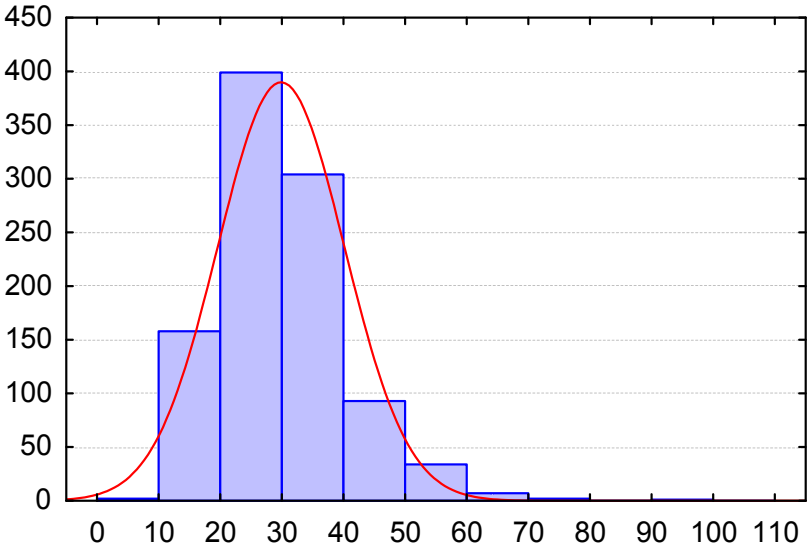
Normalita ve vícerozměrném prostoru

- Normalita ve vícerozměrném prostoru

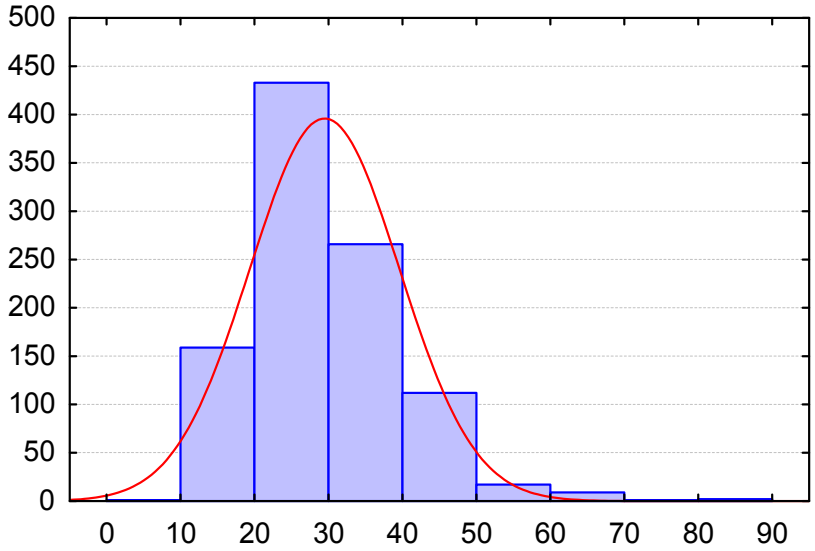


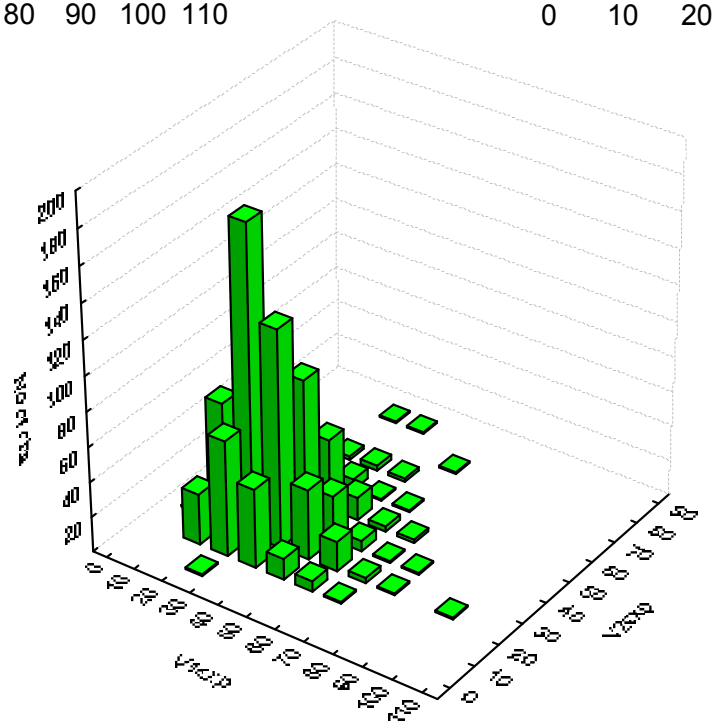


Nenormální rozložení ve vícerozměrném prostoru

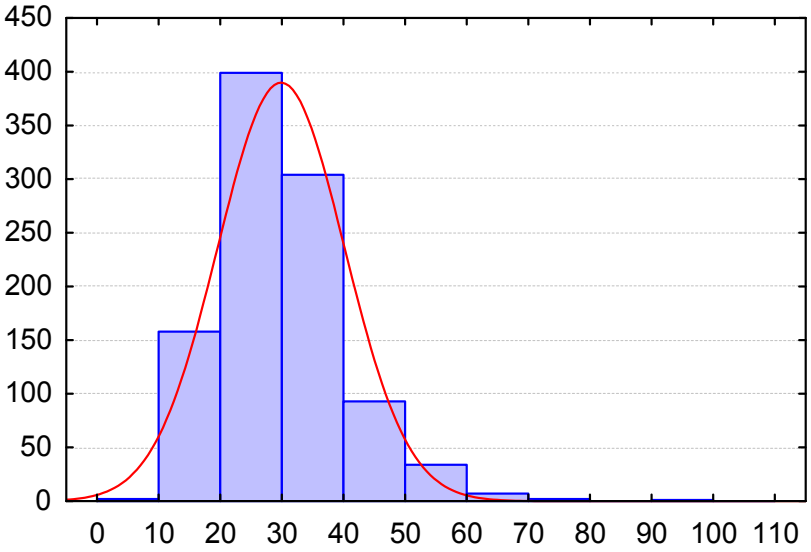


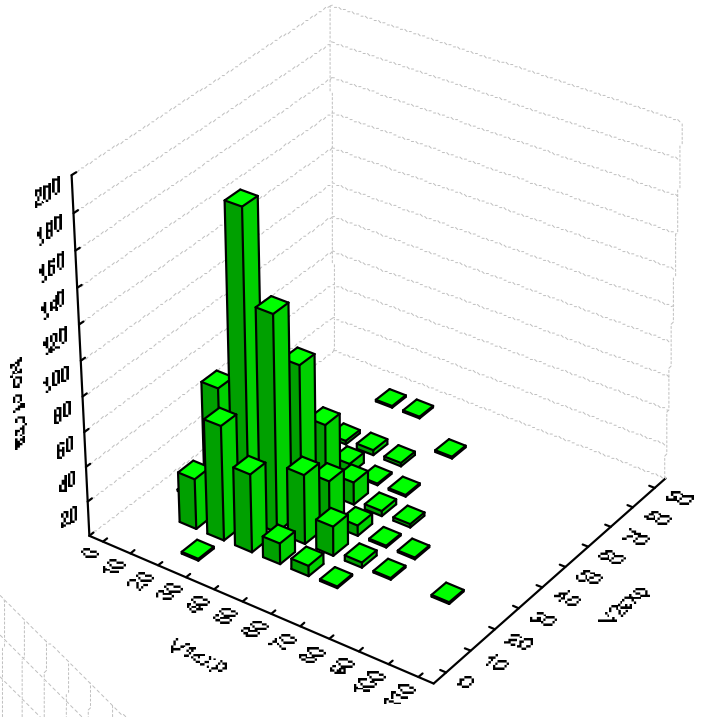
+

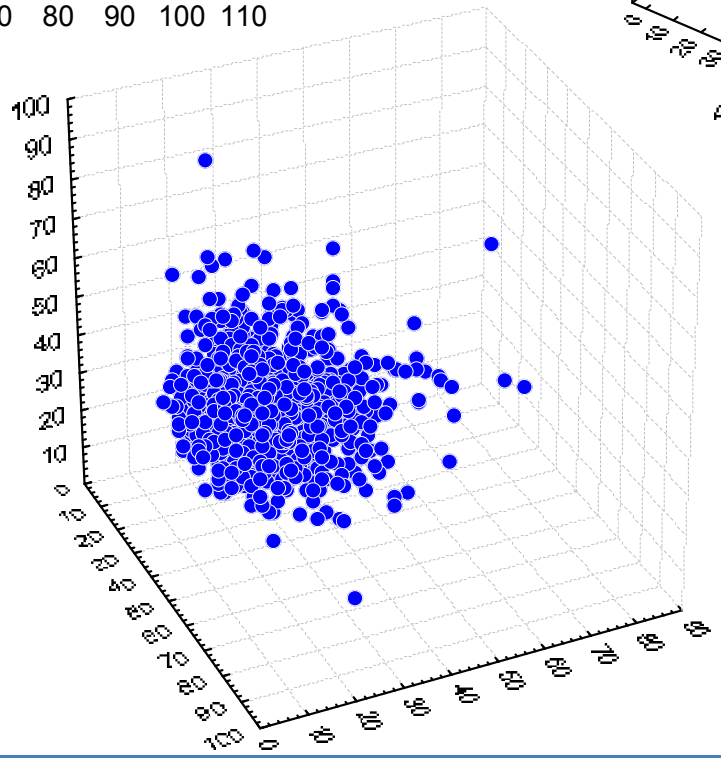




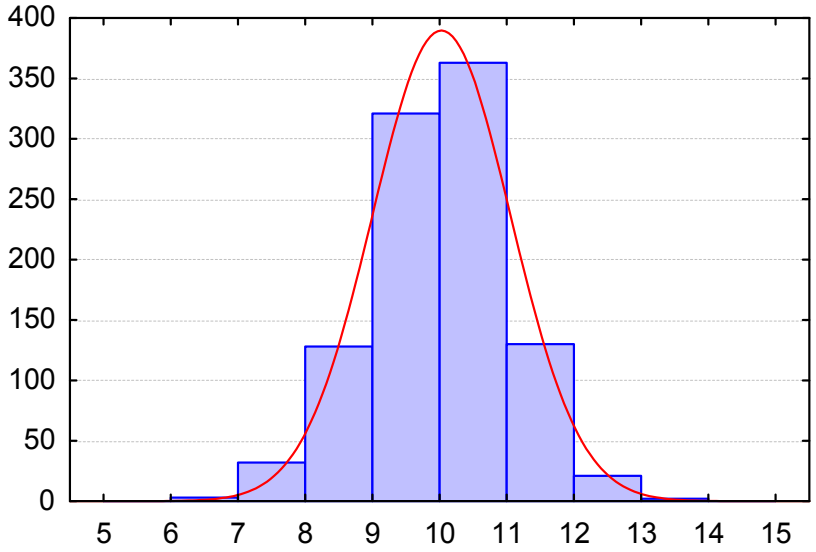
Nenormální rozložení ve vícerozměrném prostoru



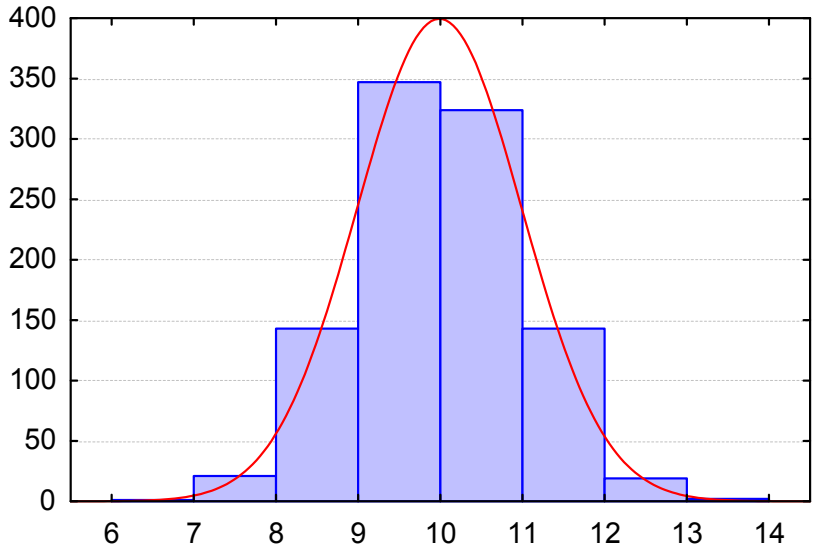


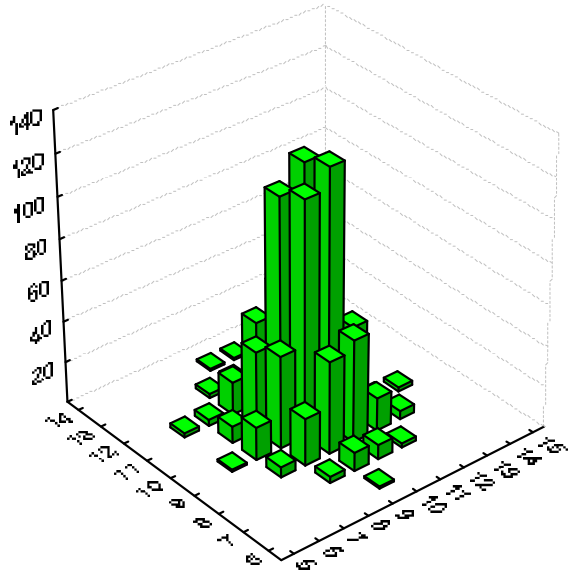


Je normalita v jednorozměrném prostoru jedinou podmínkou vícerozměrné normality?

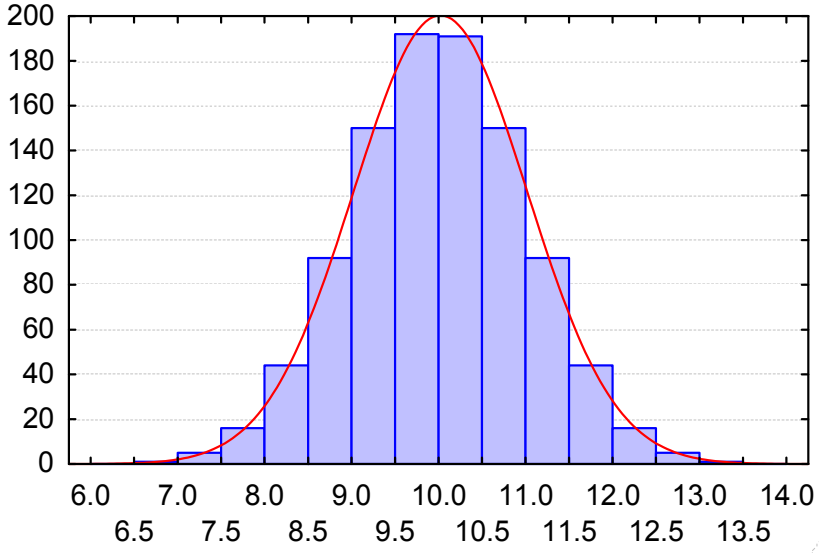


+

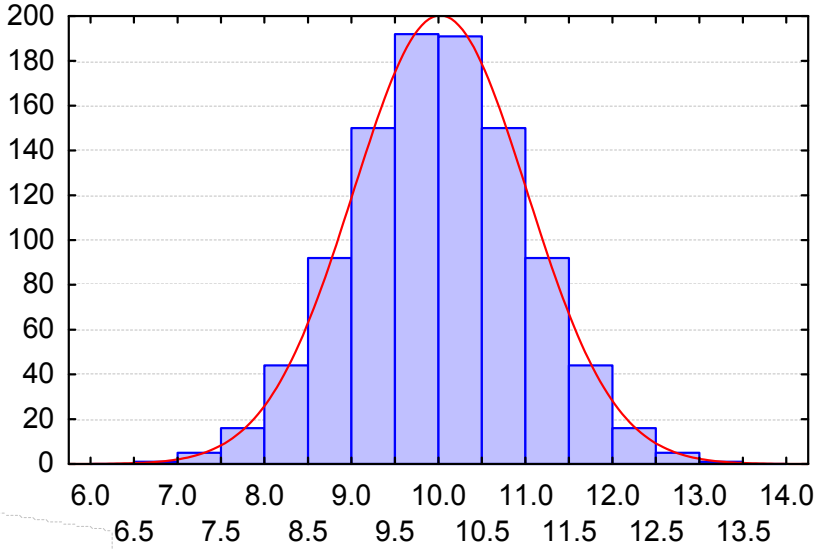


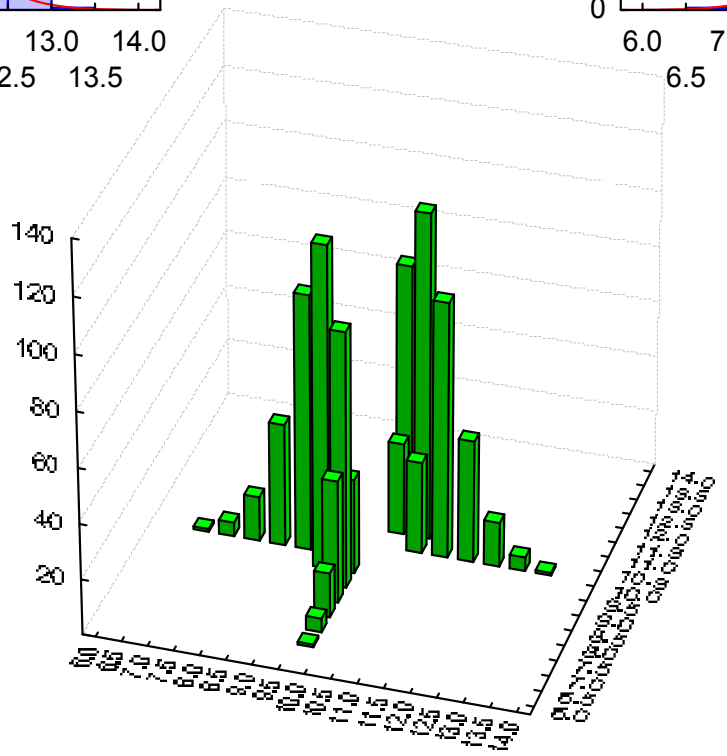


Je normalita v jednorozměrném prostoru jedinou podmínkou vícerozměrné normality?

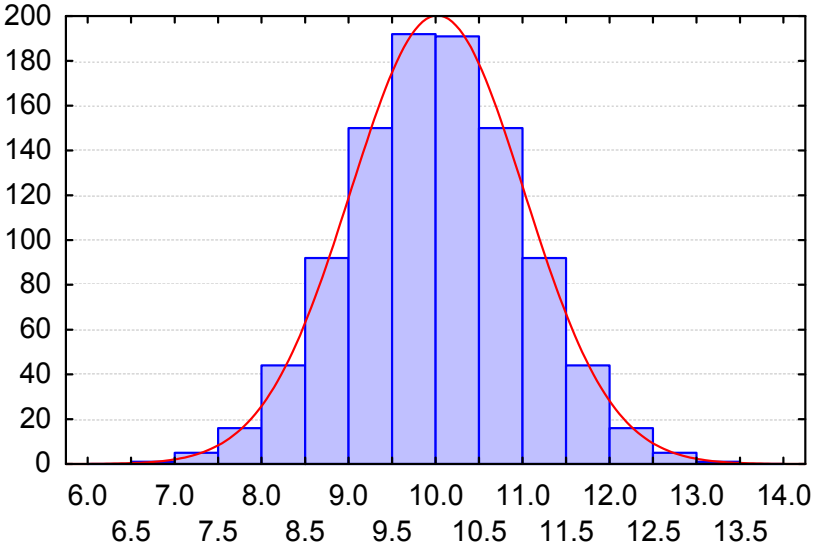


+

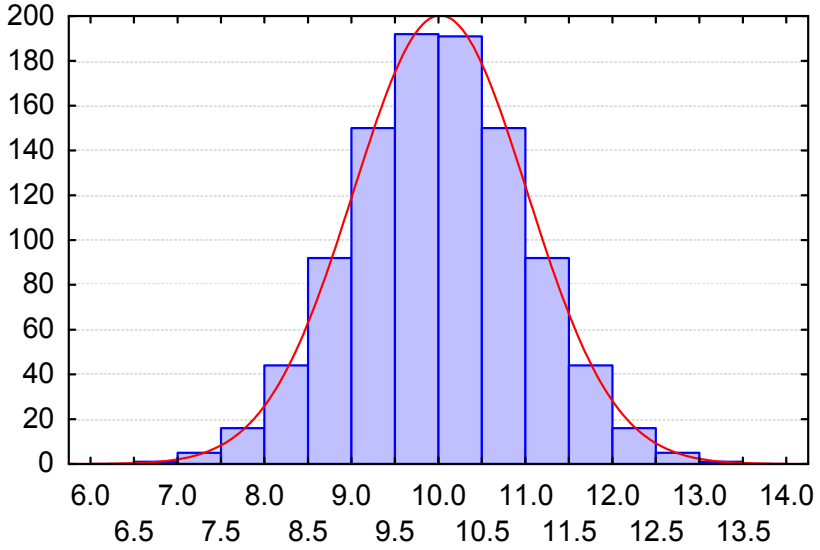


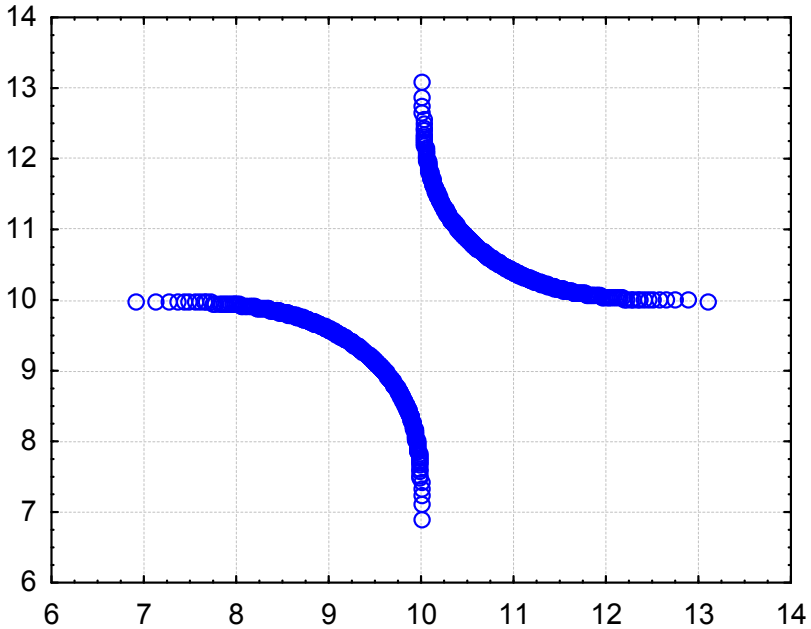


Je normalita v jednorozměrném prostoru jedinou podmínkou vícerozměrné normality?

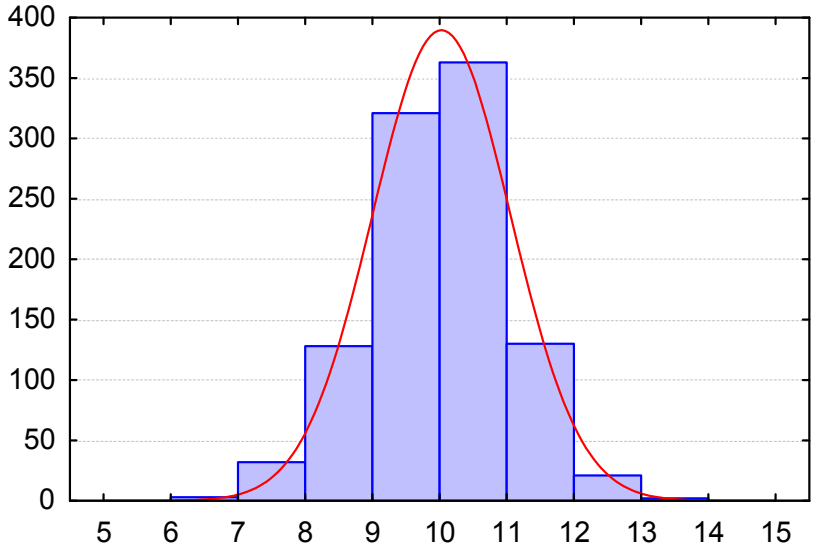


+

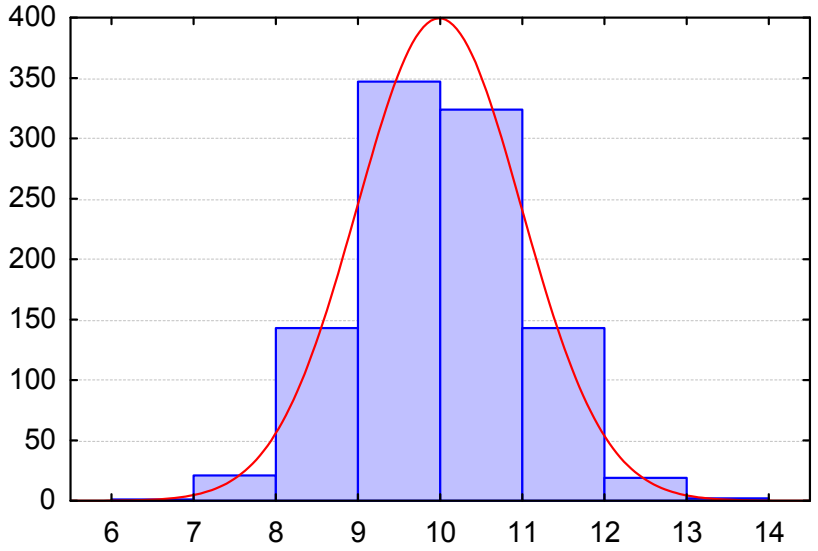


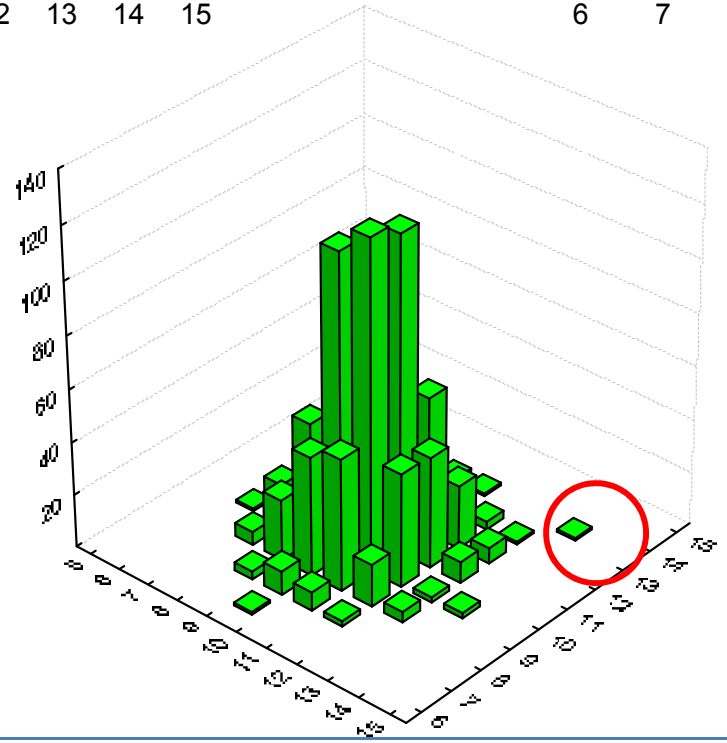


Vícerozměrný outlier



+



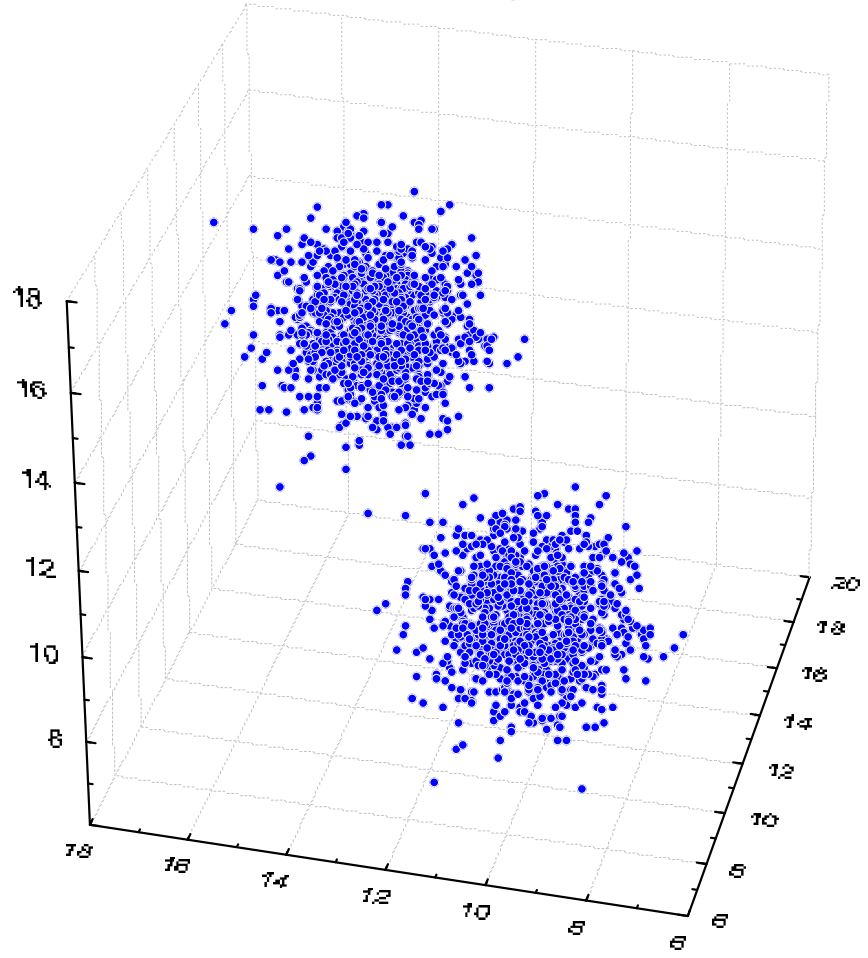


Srovnání průměrů ve vícerozměrném prostoru

- Pro zobecnění t-testu pro p rozměrů se využívá Hottelingovo rozdělení

$$T^2 = \frac{n_1 n_2}{n}\left(\overline{x_1} - \overline{x_2} - \delta\right)^T S^{-1}\left(\overline{x_1} - \overline{x_2} - \delta\right)$$

- kde $\delta = \mu_1 - \mu_2$ (nejčastěji δ = 0), má opět Hotellingovo rozdělení s parametry p, n – p –1



T-tests; Grouping: group (vicerozmerne_modelove)
Group 1: 1; Group 2: 2
Hotelling T2=23280.9 F(3,1996)=7752.5 p<0.0000

Variable	Mean 1	Mean 2	t-value	df	p	Valid N 1	Valid N 2	Std.Dev. 1	Std.Dev. 2	F-ratio Variances	p Variances
V1	10.00068	14.00068	-87.3755	1998	0.00	1000	1000	1.023659	1.023659	1.000000	1.000000
V2	9.96685	13.96685	-89.5768	1998	0.00	1000	1000	0.998503	0.998503	1.000000	1.000000
V3	10.00140	14.00140	-88.5272	1998	0.00	1000	1000	1.010342	1.010342	1.000000	1.000000

Vícerozměrné statistické metody

Operace s vektory a maticemi

Pojmy vícerozměrných analýz

- Vícerozměrné metody: Název vícerozměrné vychází z typu vstupních dat, tato data jsou tvořena jednotlivými objekty (i.e. klienti) a každý z nich je charakterizován svými parametry (věk, příjem atd.) a každý z těchto parametrů můžeme považovat za jeden rozměr objektu.
- Maticová algebra: Základem práce s daty a výpočtů vícerozměrných metod je maticová algebra, matice tvoří jak vstupní, tak výstupní data a probíhají na nich výpočty.
- NxP matice: N objektů s p parametry pak vytváří tzv. NxP matici, která je prvním typem vstupu dat do vícerozměrných analýz.
- Asociační matice: Na základě těchto matic jsou počítány matice asociační na nichž pak probíhají další výpočty, jde o čtvercové matice obsahující informace o podobnosti nebo rozdílnosti (tzv. metriky) buď objektů (Q mode analýza) nebo parametrů (R mode analýza).Měřítko podobnosti se liší podle použité metody a typu dat, některé metody umožňují použití uživatelských metrik.

Vstupní matice vícerozměrných analýz