Vícerozměrná data, jejich popis a vizualizace

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

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	
…						

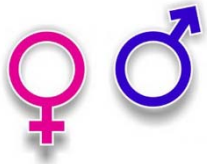
$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

maticový zápis datového souboru *n* objektů (subjektů), které jsou popsané *p* proměnnými

jeden prvek matice x_{ij} je hodnota *j*-té proměnné u *i*-tého objektu (subjektu), přičemž $j = 1, \ldots, p$ a $i = 1, \ldots, n$

Typy dat - opakování

- **Kvalitativní (kategoriální) data:**
  - Binární data



  - Nominální data

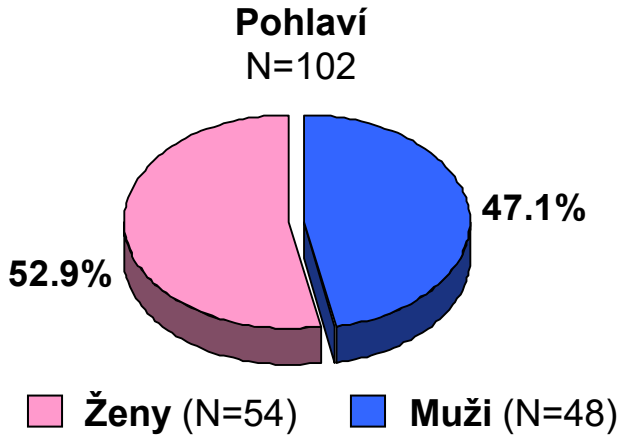
  - Ordinální data

- **Kvantitativní data:**
  - Intervalová data

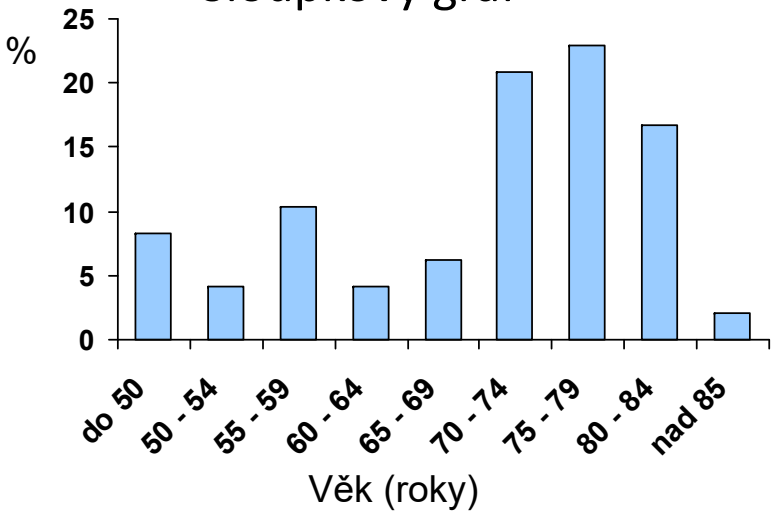
  - Poměrová data

Vizualizace jednorozměrných dat - opakování

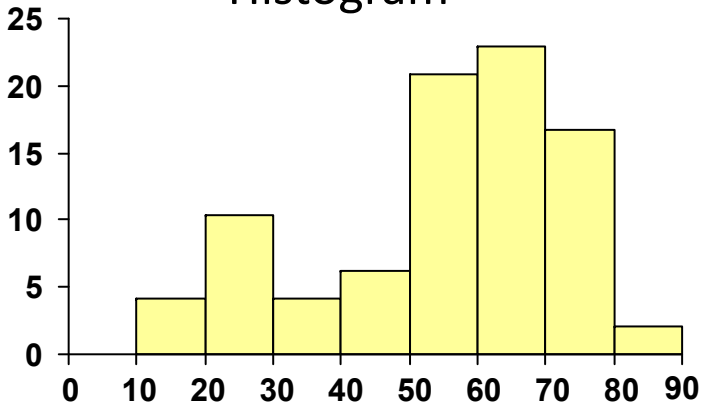
Koláčový graf



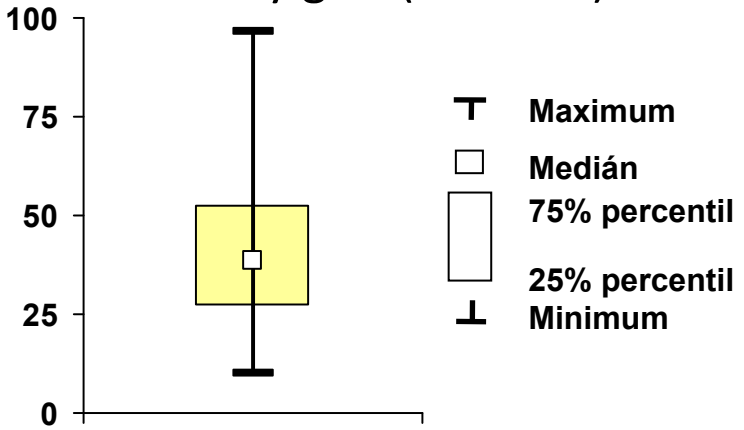
Sloupkový graf



Histogram



Krabicový graf (Box Plot)



K čemu nám může pomoci vizualizace dat?

→ odhalení problémů v datech

id	vek	pohlavi	cholesterol	vyska	vaha	obvod_pasu	obvod_boku	BMI	sys_tlak	dia_tlak
1	38	Z	4.6	164	45	60	87	16.7	120	80
2	36	Z	4.35	167	90	97	112	32.3	130	80
3	26	Z		178	70	72	94	22.1	127	80
4	25	Z	4.2	165	59	65	92	21.7	130	80
5	47	M	5.65	158		92	96	26.8	155	90
6	21	Z	6.35	172	61	69	98	20.6	135	80
7	23	Z	3.45	170	82	92	113	28.4	130	80
8	35	M	7.99	179	90	101	110	28.1	140	88
9	33	Z	4.88	167	57	70	92	20.4	140	85
10	48	Z	9.56	164	70	93	107	26.0	250	97
11	25	M	3.1	186	75	81	102	21.7	120	70
12	41	Z	10	167	62	71	101	22.2	140	90
13	29	ZZ	4.2	165	58	66	98	21.3	120	80
14	24	M	5.62	174	80	92	107	26.4	156	90
15	58	Z	7.9	164	63	73	100	23.4	135	90

Chybné hodnoty

Chybějící hodnoty

Odlehlé hodnoty

Problémy v datech – chybějící hodnoty

- snaha, aby v datech vůbec nenastaly
- pokud však nastanou, je silně nedoporučováno dělat každou analýzu na jinak velkém souboru (tzv. „casewise“ odstraňování objektů) → 3 možná řešení:

1. vyloučit z analýzy všechny objekty, u nichž se vyskytla nějaká chybějící hodnota (tzv. „listwise“ odstranění objektů):
   - pokud chybějících hodnot mnoho, zbyde pouze málo objektů
   - pozor na systematicky chybějící hodnoty – může dojít ke zkreslení výsledků analýz
   - občas vhodné odstranit proměnné s mnoha chybějícími hodnotami místo objektů, pokud proměnné nejsou důležité pro analýzu
2. definování souboru s vyplněnými „klíčovými“ proměnnými:
   - na tomto souboru provedena většina analýz
   - další analýzy dělány na podsouboru s menším počtem subjektů
3. doplnění chybějících hodnot (tzv. imputace):
   - doplnění průměrem z hodnot, které jsou pro danou proměnnou k dispozici
   - doplnění hodnot na základě regresních modelů
   - pozor! doplnění hodnot však může zkreslit výsledky analýz

Problémy v datech – odlehlé hodnoty

- k identifikaci odlehlých hodnot mohou pomoci např. tečkové, maticové či krabicové grafy
- je třeba rozlišovat:

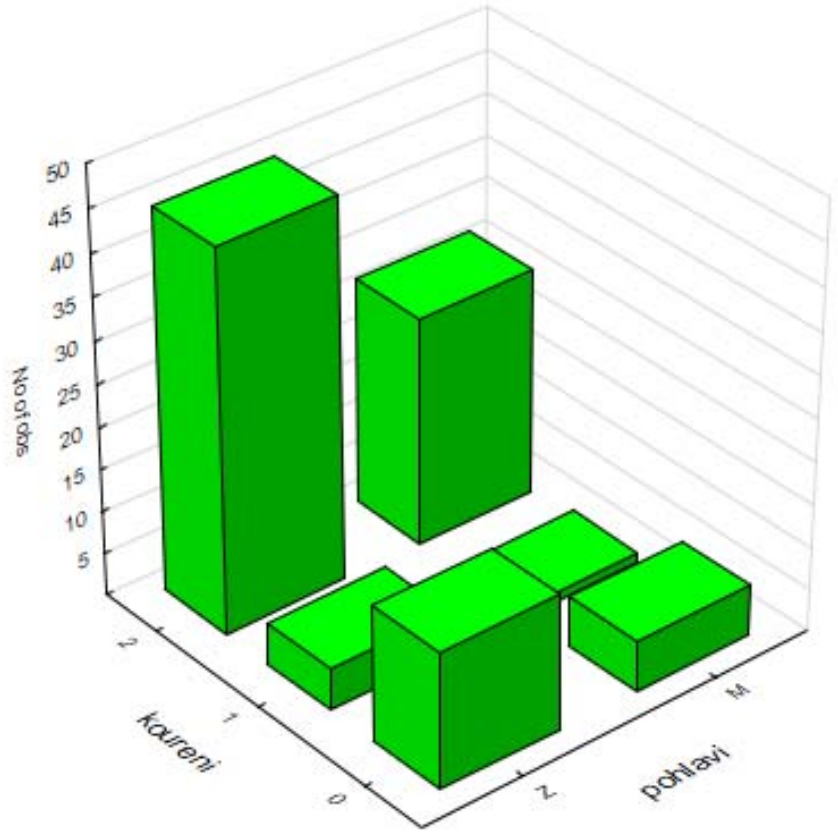
1. **odlehlé hodnoty, které jsou způsobeny chybou** (měřících přístrojů apod.) - jsou to většinou nereálné hodnoty → je vhodné je smazat a dále s nimi zacházet jako s chybějícími hodnotami
2. **odlehlé hodnoty, které jsou fyziologické** (tzn. jsou to reálné hodnoty) → je vhodné tyto hodnoty v datech ponechat, pokud je to možné a nezkreslí to analýzu a použít neparametrické metody analýzy dat
   - příklad, kdy je vhodné odlehlou hodnotu v souboru ponechat: pacienti Alzheimerovou chorobou v našem souboru mají hodnotu MMSE skóre větší než 15, jeden pacient má však hodnotu skóre 7 (je to reálná hodnota, smazáním bychom uměle snížili variabilitu)
   - příklad, kdy je nevhodné odlehlou hodnotu v souboru ponechat: chceme měřit výšku 15-letých dětí – dítě trpící nanismem měřící 80 cm by průměrnou výšku velice zkreslilo, proto ho ze souboru vyřadíme

Vizualizace vícerozměrných dat

- 3D sloupkové grafy
- dvourozměrný histogram
- maticové grafy
- krabicové grafy pro více proměnných
- ikonové (symbolové) grafy:
  - profilové sloupce
  - profily
  - paprskové (hvězdicové) grafy
  - polygony
  - pavučinové grafy
  - Chernoffovy tváře

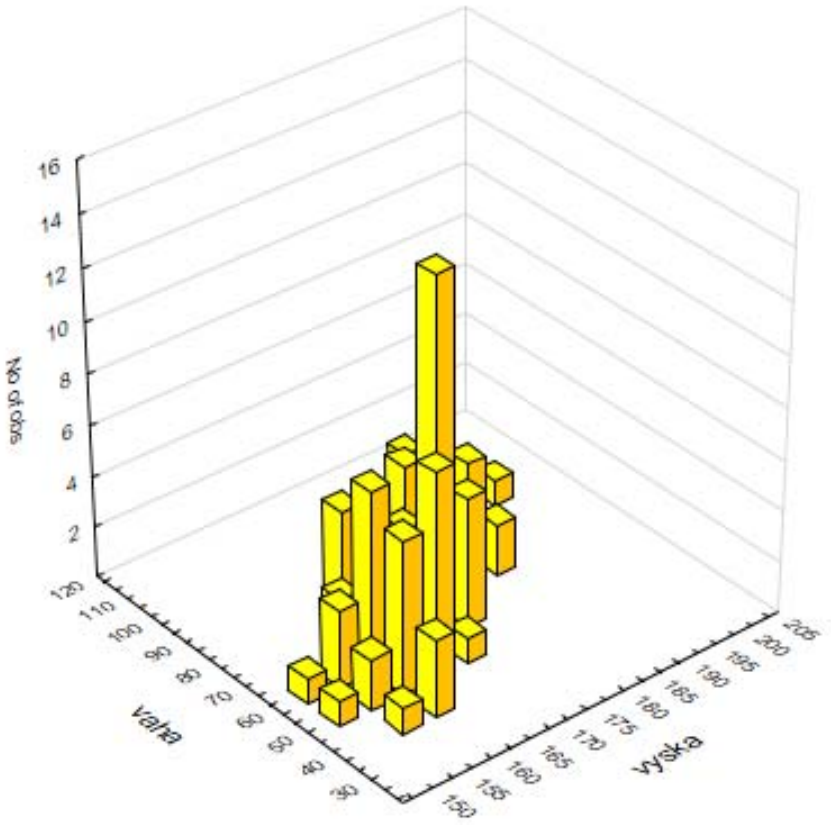
3D sloupkové grafy

- vzájemný výskyt kategorií dvou kategoriálních proměnných
- v softwaru Statistica: Graphs – 3D Sequential Graphs – Bivariate Histograms...



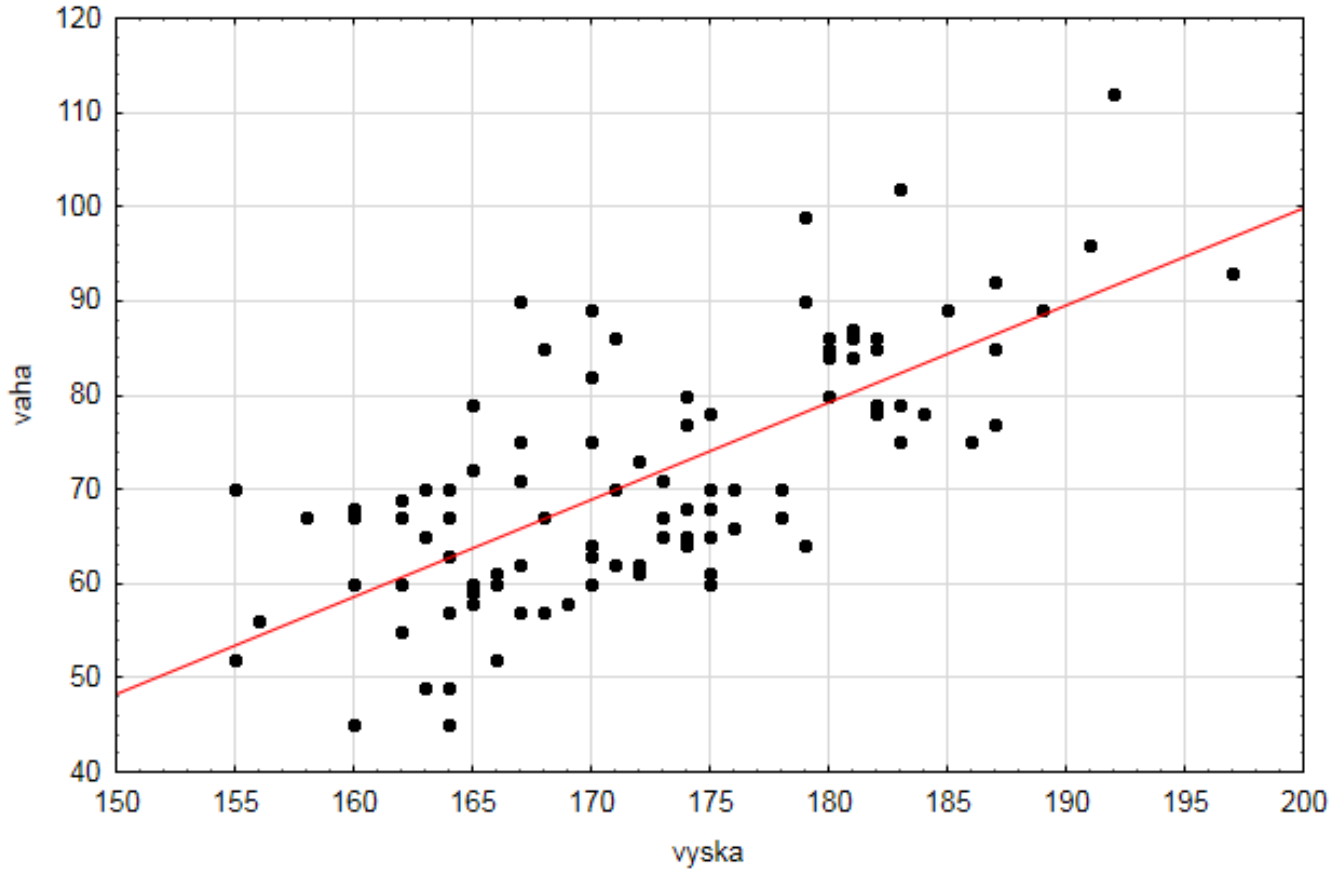
Dvourozměrný histogram

- pro vykreslení vztahu dvou spojitých proměnných
- v softwaru Statistica: Graphs – 3D Sequential Graphs – Bivariate Histograms...



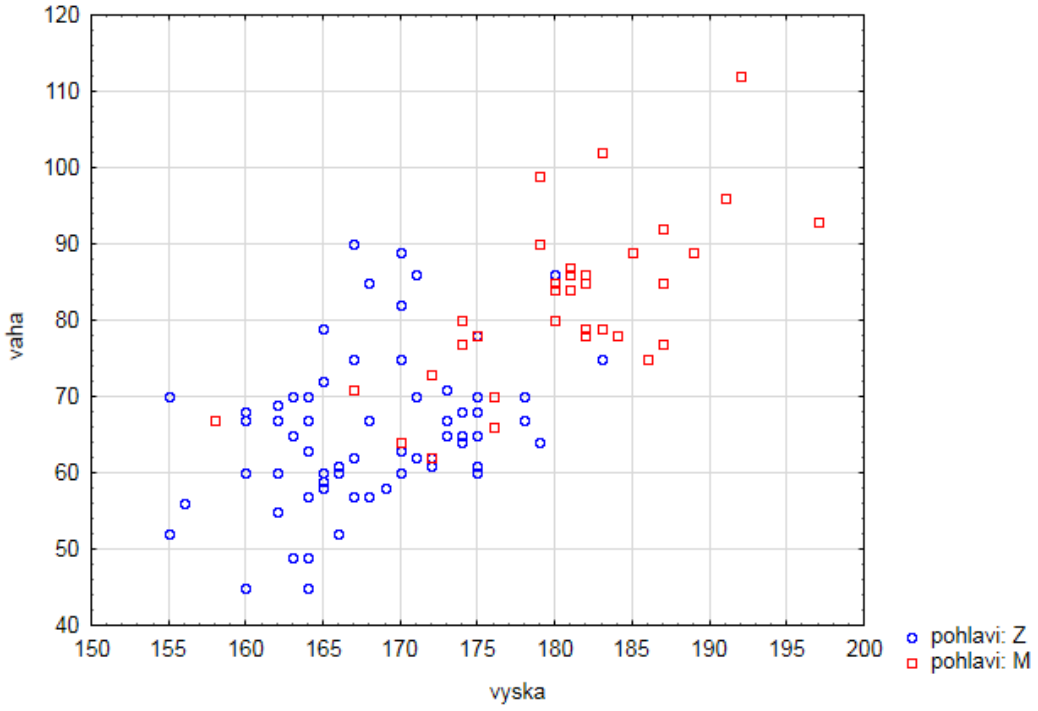
Tečkový graf

- rovněž pro vykreslení vztahu dvou spojitých proměnných
- v softwaru Statistica: Graphs – Scatterplots...



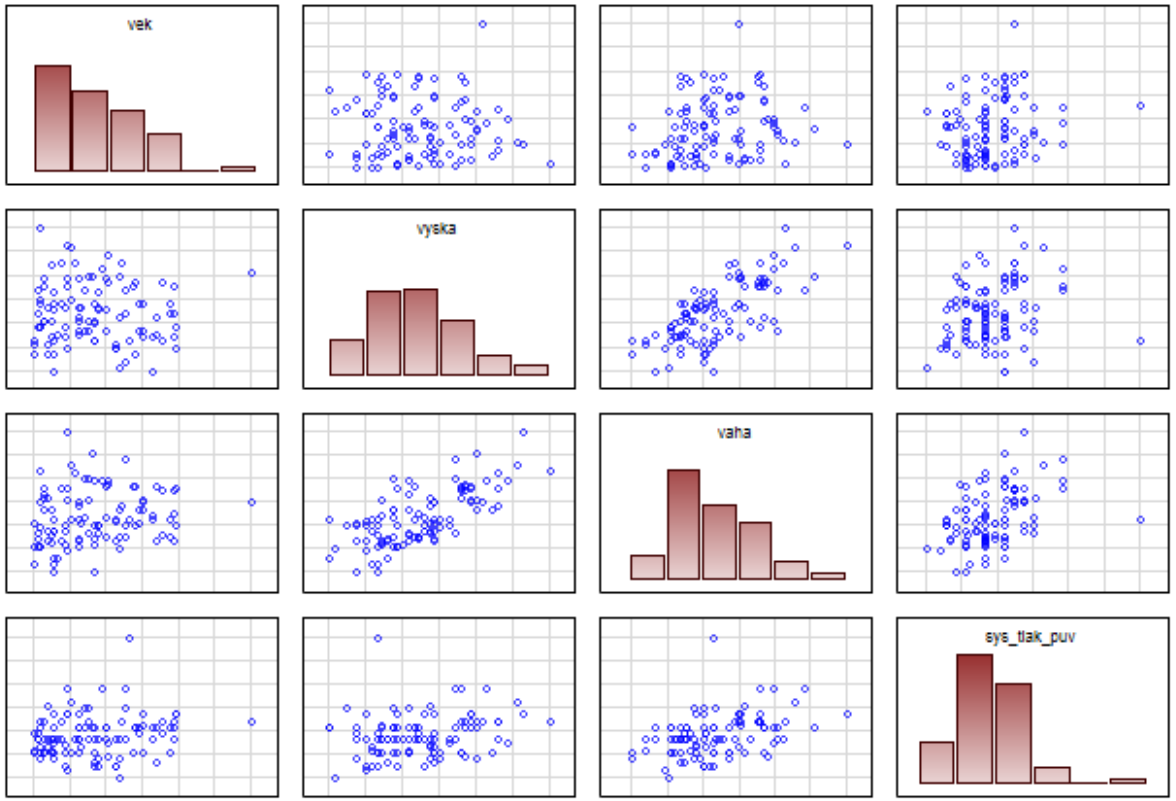
Tečkový graf – přidání kategoriální proměnné

- zahrnutí kategoriální proměnné do grafu použitím různých symbolů či barev pro jednotlivé skupiny určené danou kategoriální proměnnou
- v softwaru Statistica: Graphs – Scatterplots – na záložce Categorized zahrnout On u X-Categorized, vybrat kategoriální proměnnou pomocí Change Variable a změnit Layout na Overlaid



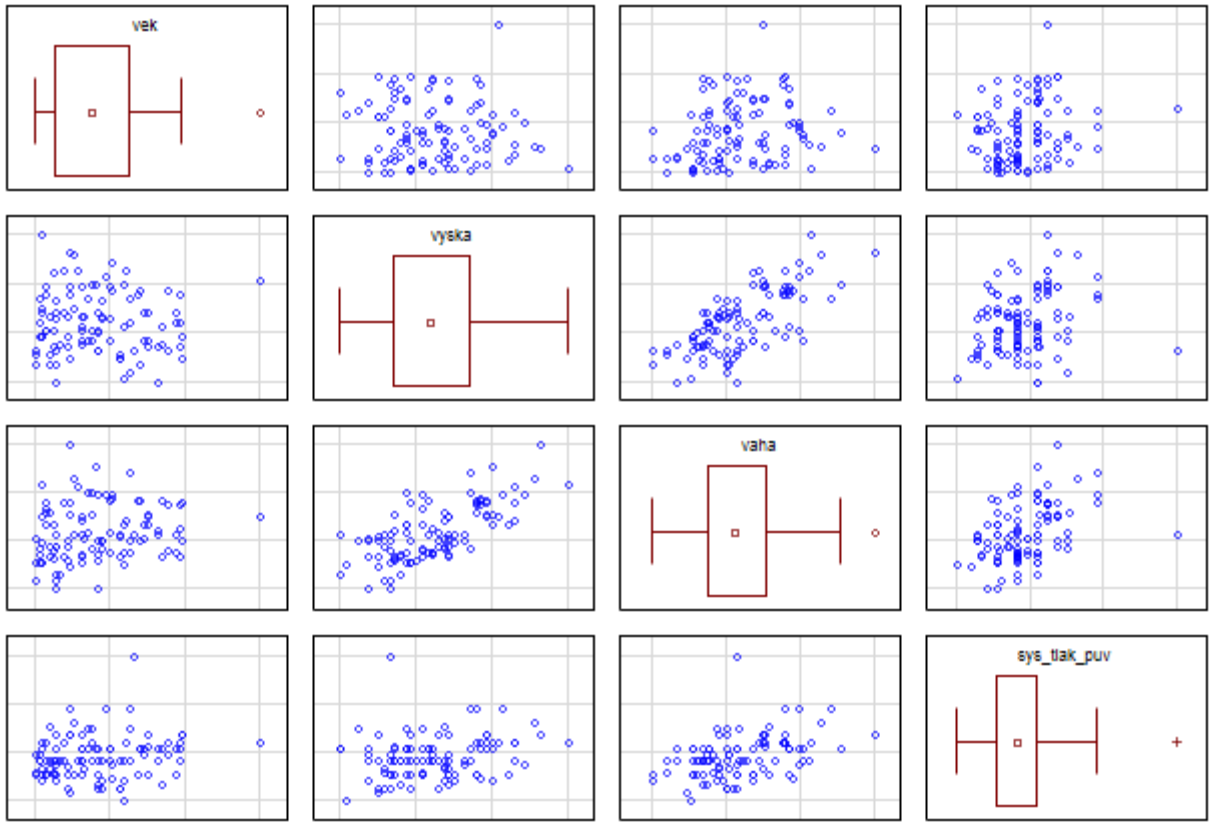
Maticový graf

- vykreslení vztahu více spojitých proměnných
- v softwaru Statistica: Graphs – Matrix Plots...
- upozornění: nastavení, jak se vypořádat s chybějícími hodnotami



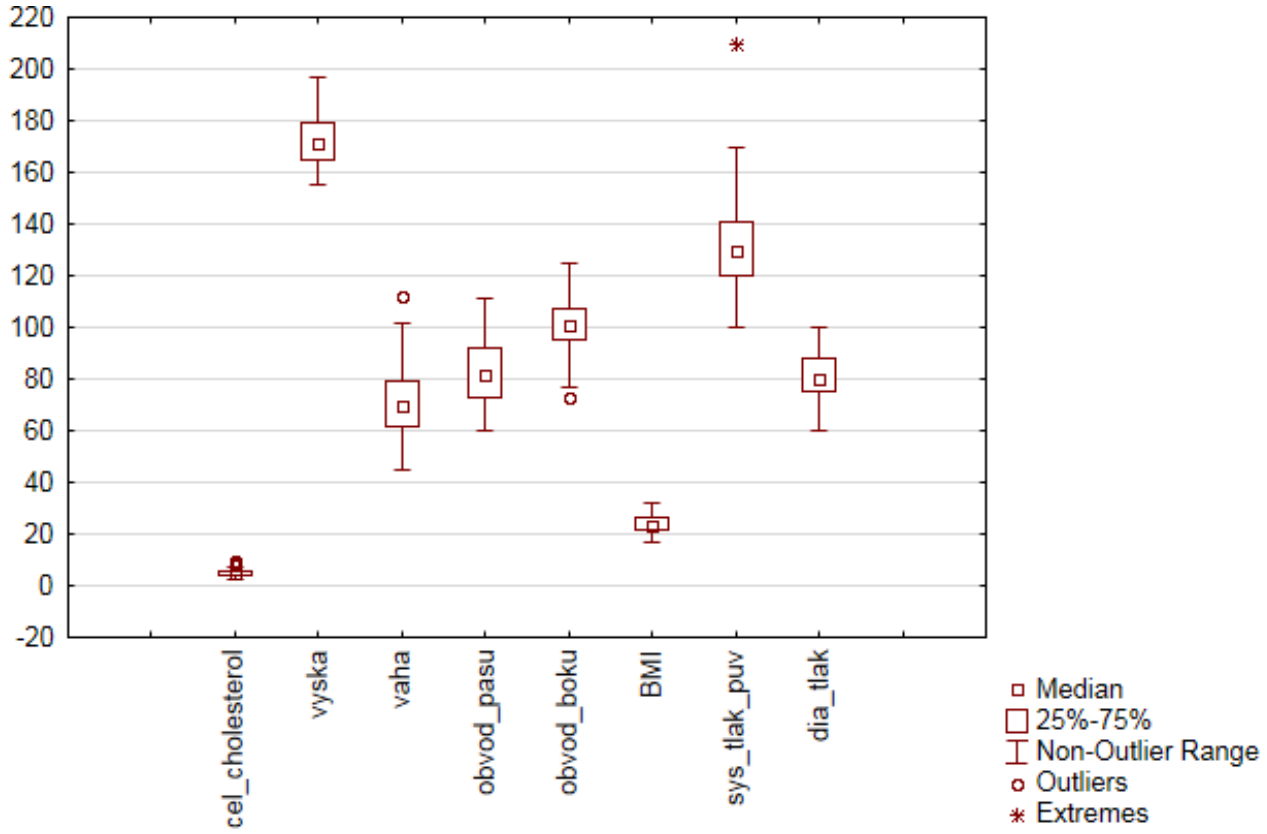
Maticový graf – na diagonále krabicové grafy

- v softwaru Statistica: Graphs – Matrix Plots...; na záložce Advanced zatrhnout Display: Box plot



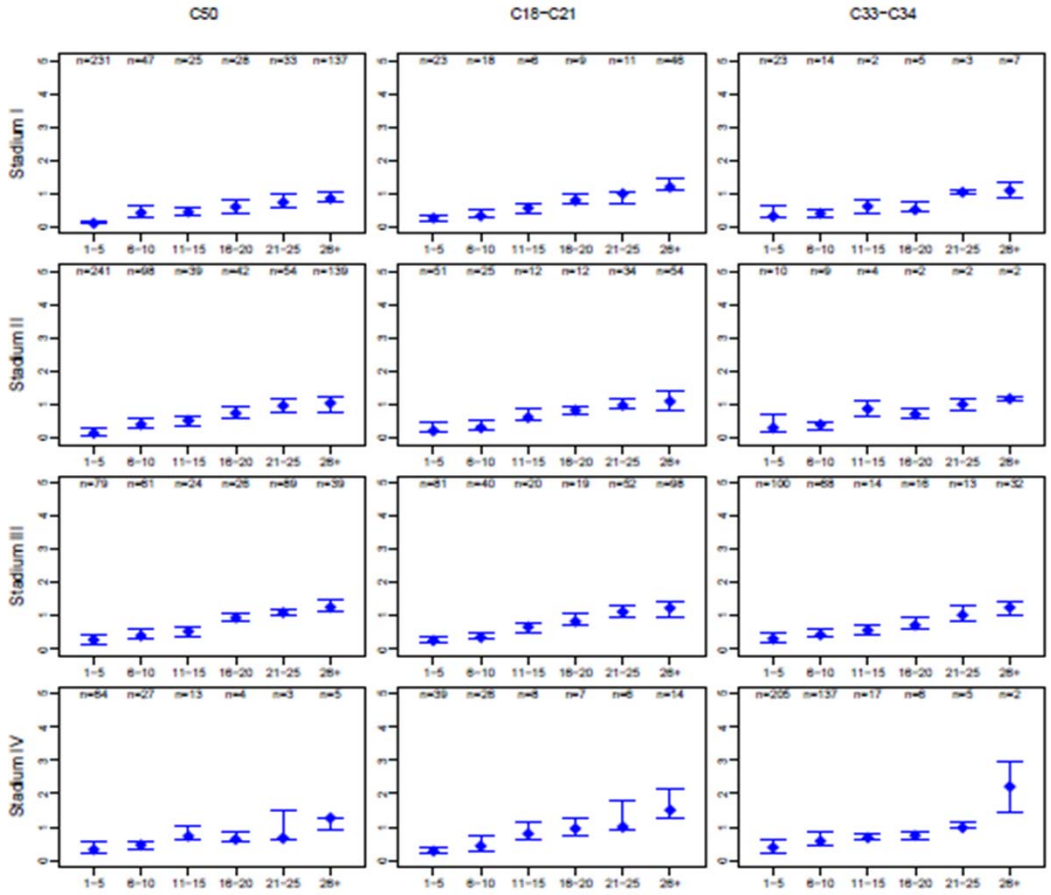
Krabicové grafy pro více proměnných

- ukáží nám, zda mají proměnné podobný rozsah hodnot
- v softwaru Statistica: označit příslušné sloupečky v datech – Graphs – Graphs of Block Data – Box Plot: Block columns



Vícenásobné krabicové grafy

- umožňují znázornění vztahu několika kvalitativních proměnných a jedné kvantitativní proměnné

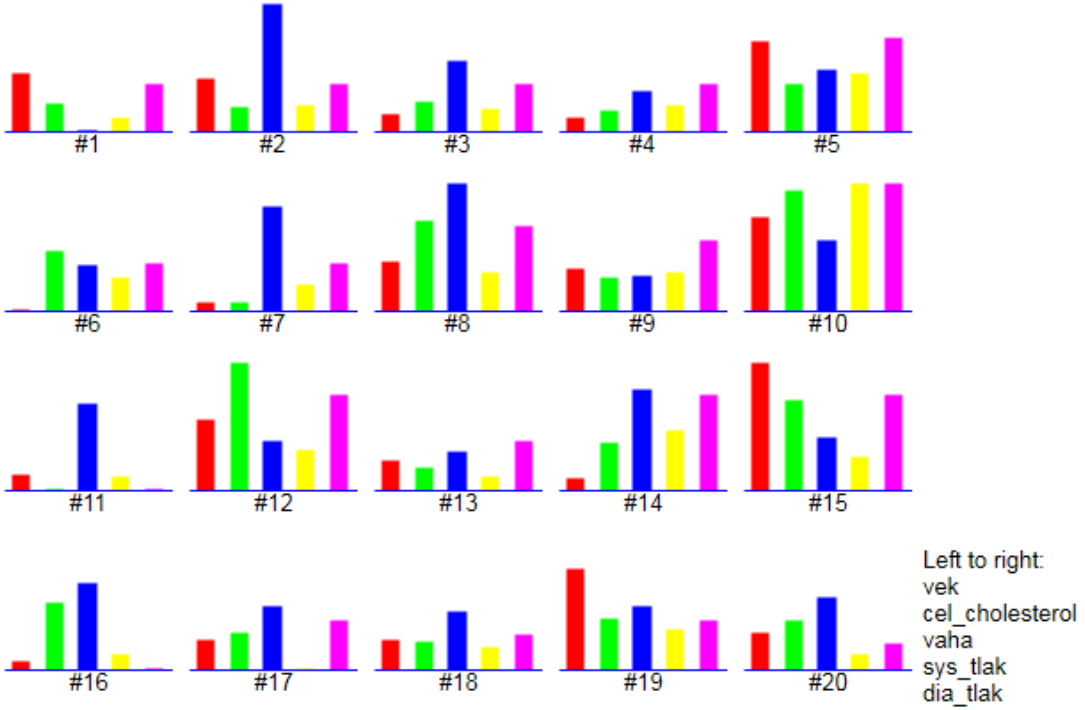


Ikonové (symbolové) grafy

- hodnoty znaků znázorněny jako geometrické útvary či symboly
- každému objektu (subjektu) odpovídá jeden obrazec složený z těchto geometrických útvarů či symbolů
- umožní vizuálně porovnat, které objekty (subjekty) jsou si podobné
- mnoho druhů, v softwaru Statistica např.:
  1. Profilové sloupce
  2. Profily
  3. Paprskové (hvězdicové) grafy
  4. Polygony
  5. Pavučinové grafy
  6. Chernoffovy tváře

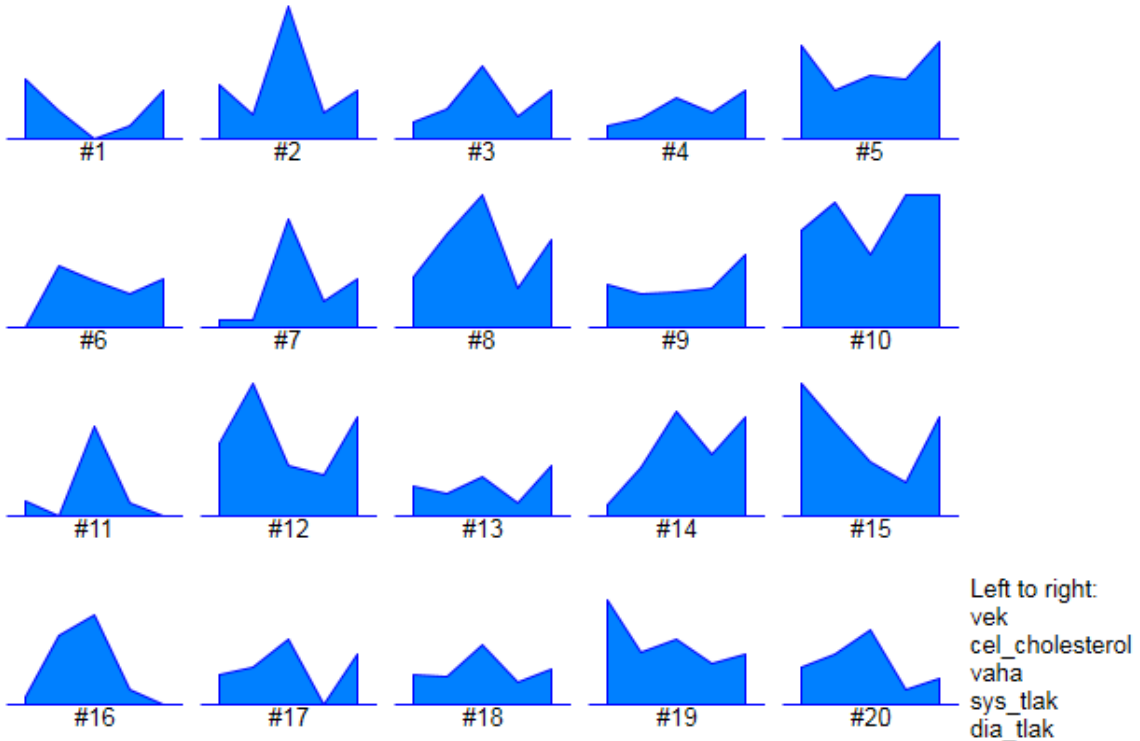
Ikonové grafy – profilové sloupce

- výšky sloupců odpovídají relativním hodnotám proměnných (relativní hodnota je podíl původní hodnoty a maxima z absolutních hodnot dané proměnné)
- v softwaru Statistica: Graphs – Icon Plots... – Graph type: **Columns**
  – zvolit proměnné – na záložce Options 1 zatrhnout „Display case labels“



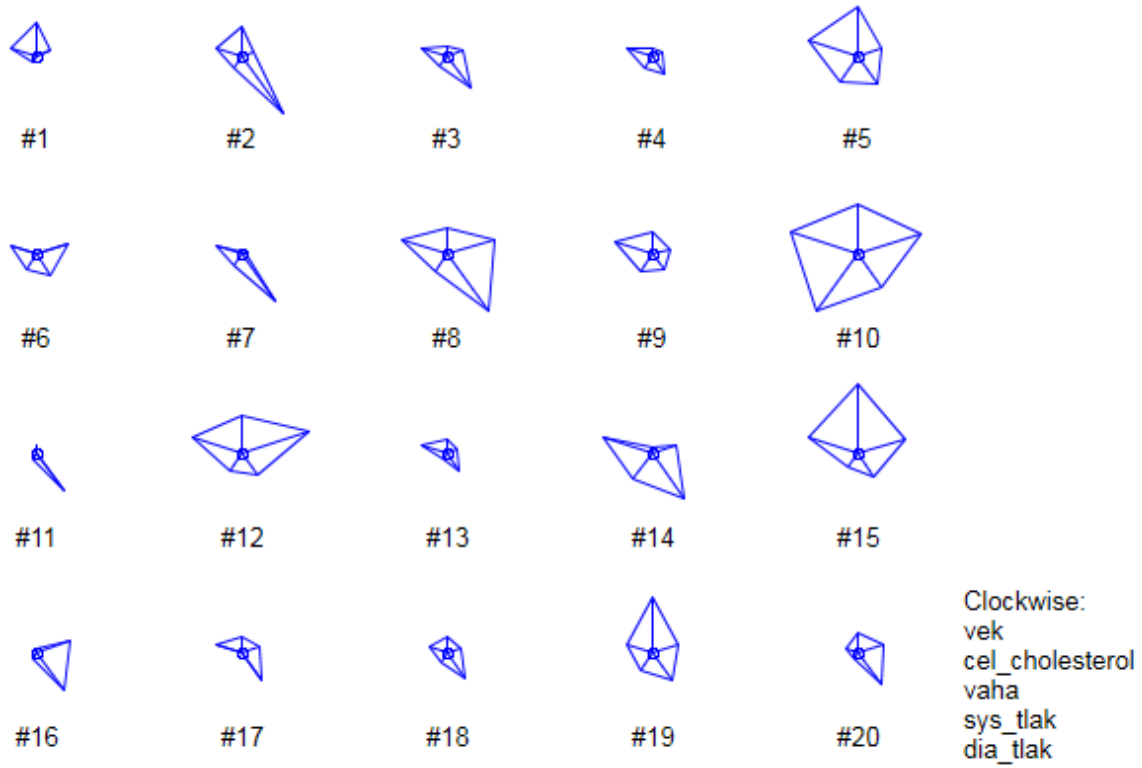
Ikonové grafy – profily

- obdoba profilových sloupců, jen se středy horních hran profilových sloupců spojí úsečkami
- v softwaru Statistica: Graphs – Icon Plots... – Graph type: **Profiles**
  – zvolit proměnné – na záložce Options 1 zatrhnout „Display case labels“



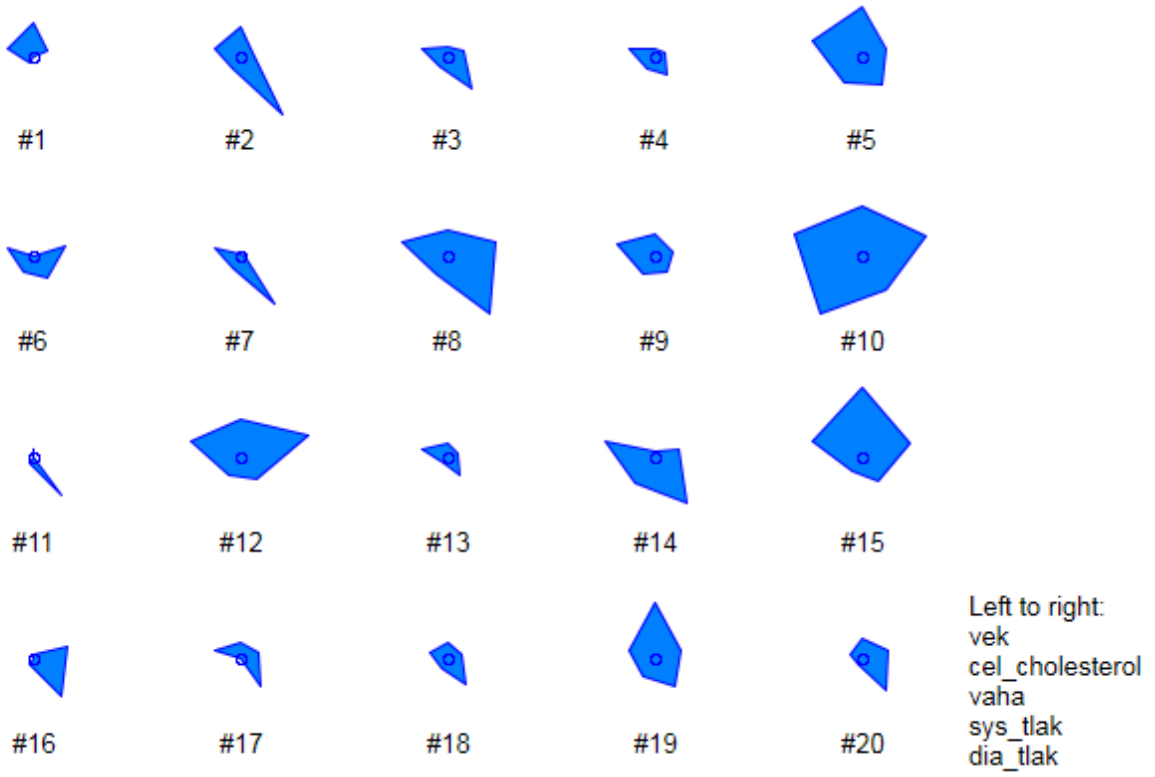
Ikonové grafy – paprskové (hvězdicové) grafy

- vzdálenosti od středu odpovídají relativním hodnotám proměnných
- v softwaru Statistica: Graphs – Icon Plots... – Graph type: **Stars**
  – zvolit proměnné – na záložce Options 1 zatrhnout „Display case labels“



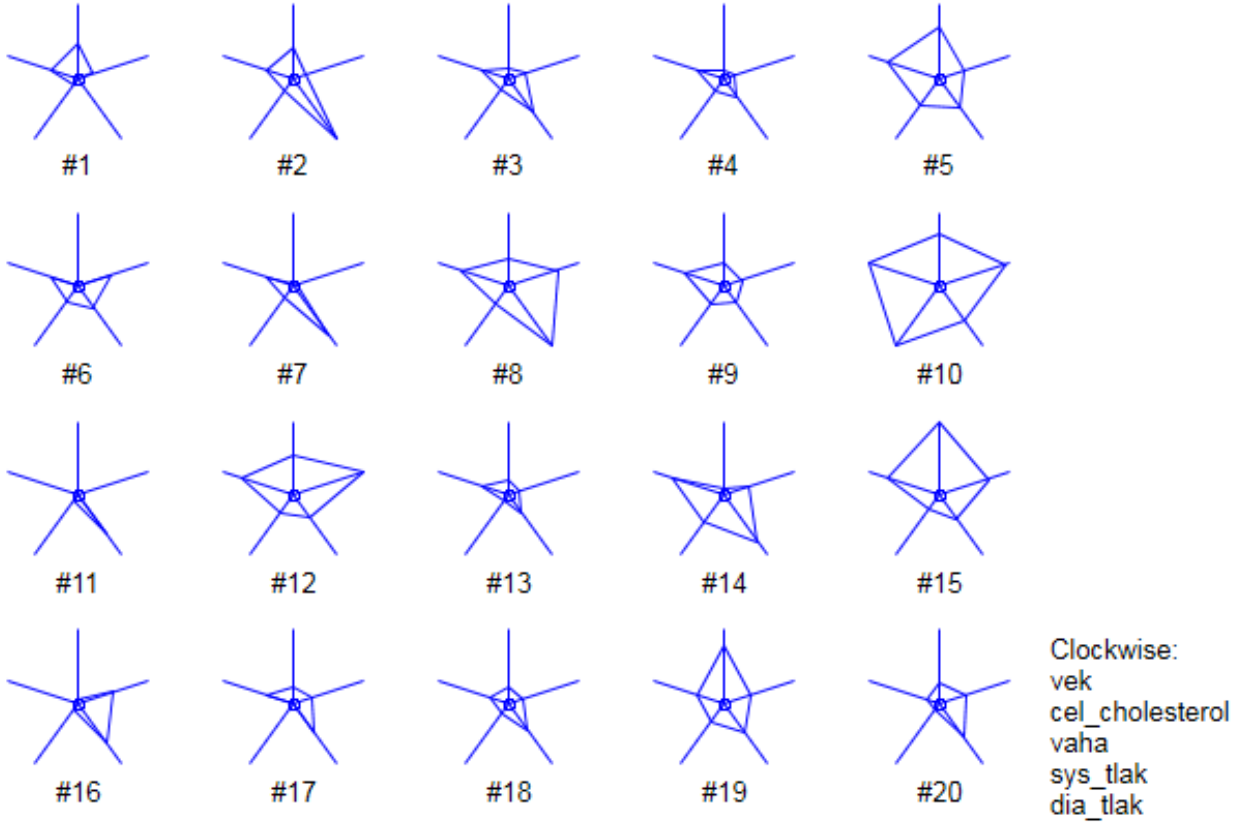
Ikonové grafy – polygony

- obdoba paprskových grafů, jen jsou vyplněné
- v softwaru Statistica: Graphs – Icon Plots... – Graph type: **Polygons**
  – zvolit proměnné – na záložce Options 1 zatrhnout „Display case labels“



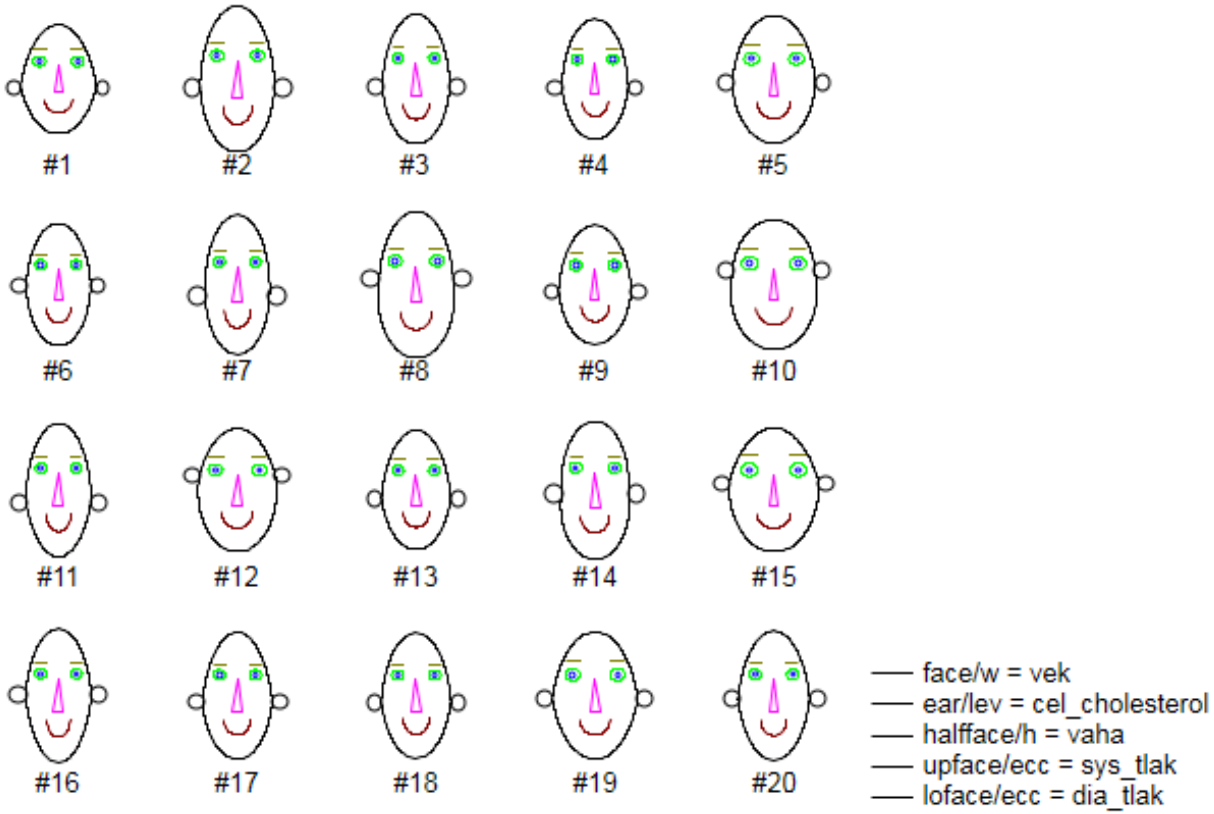
Ikonové grafy – pavučinové grafy

- obdoba paprskových grafů, přidáno znázornění maxima absolutních hodnot
- v softwaru Statistica: Graphs – Icon Plots... – Graph type: **Sun Rays**
  – zvolit proměnné – na záložce Options 1 zatrhnout „Display case labels“



Ikonové grafy – Chernoffovy tváře

- proměnné znázorněny jako části obličeje
- v softwaru Statistica: Graphs – Icon Plots... – Graph type: **Chernoff Faces** – zvolit proměnné – na záložce Options 1 zatrhnout „Display case labels“



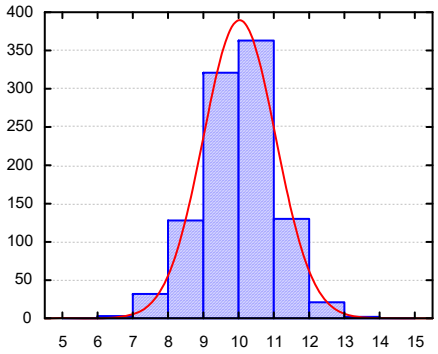
Vícerozměrné statistické rozdělení a testy

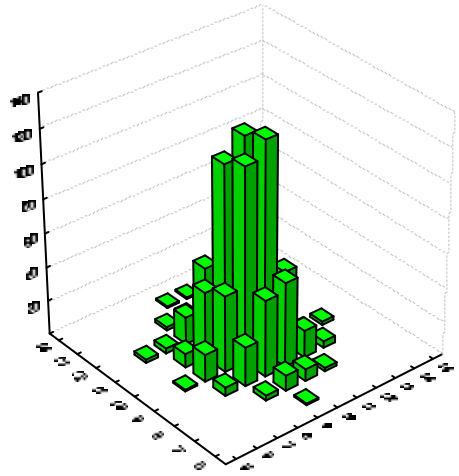
Význam rozdělení ve vícerozměrném prostoru

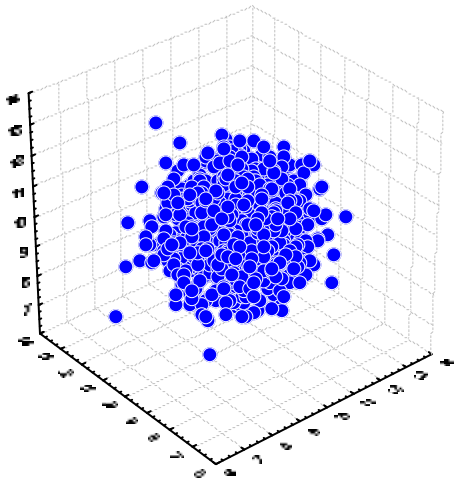
- Použitelnost mnohých klasických statistických metod a postupů vyžaduje předpoklad o normálním rozdělení sledovaných proměnných.
- Podmínka normality vyplývá z toho, že metody založené na tomto předpokladu mohou využít kompletní matematický aparát schovaný za danou statistickou metodou. Tyto metody 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

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

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

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

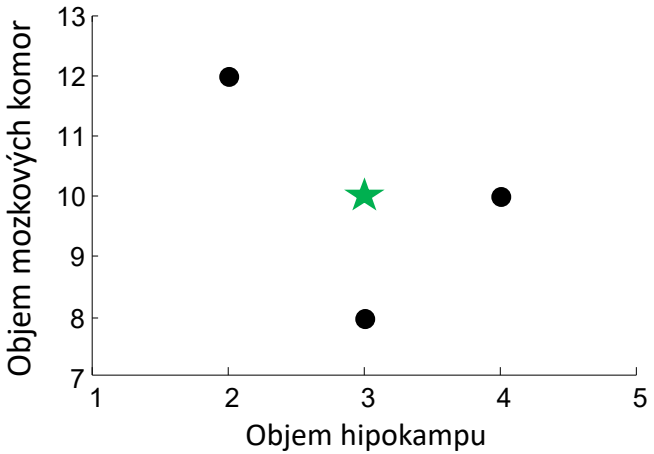
Příklad

- Spočtěte vektor středních hodnot 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:

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

Příklad - řešení

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



Vektor středních hodnot:

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

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

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

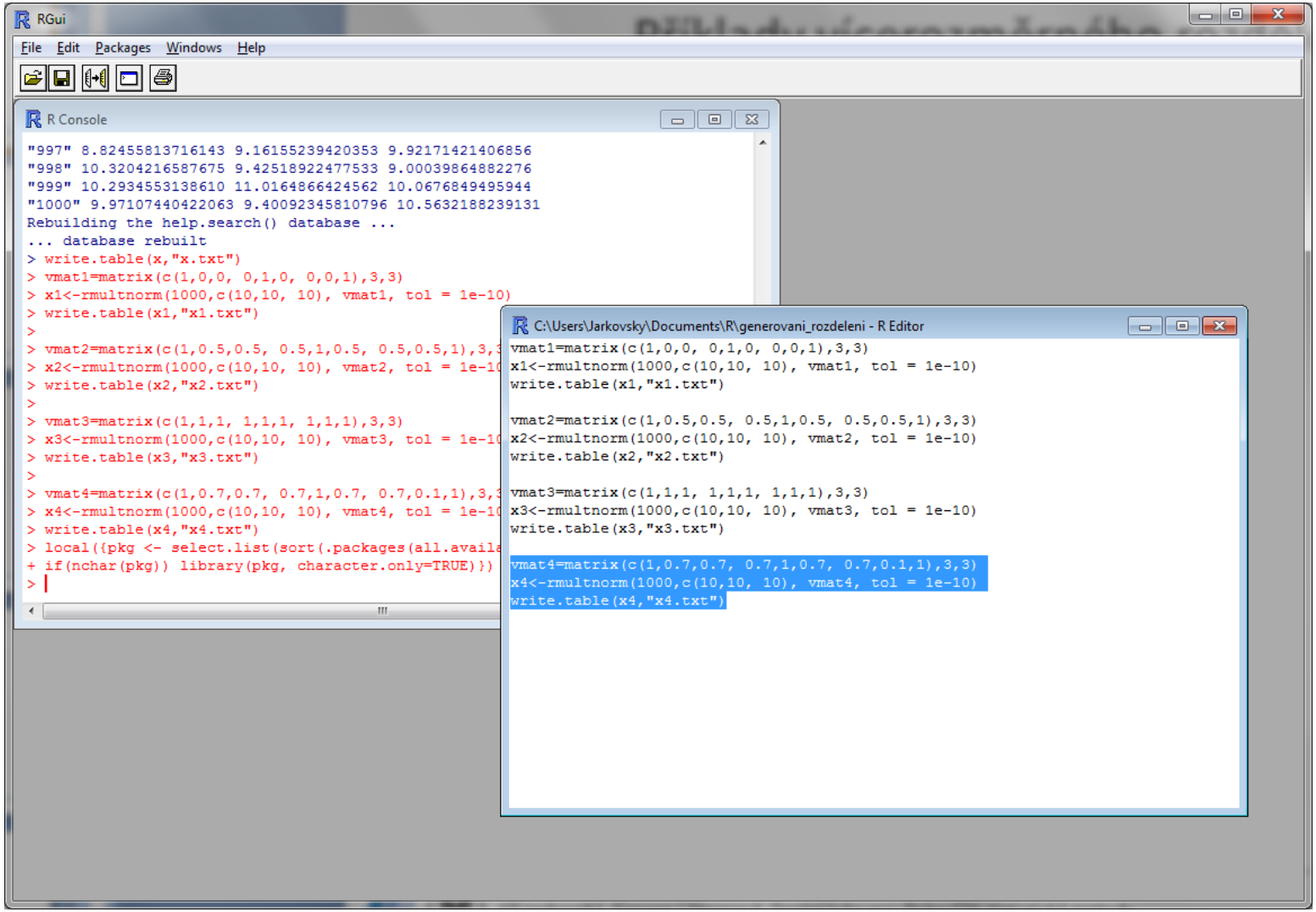
$$s_{22} = \frac{1}{n-1}\sum_{i=1}^{n}(x_{i2} - \bar{x}_2)^2 = \frac{1}{3-1}\left((12-10)^2 + (10-10)^2 + (8-10)^2\right) = 4$$

$$s_{21} = s_{12} = \frac{1}{n-1}\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2) = \frac{1}{3-1}\left((2-3)(12-10) + (4-3)(10-10) + (3-3)(8-10)\right) = -1$$

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

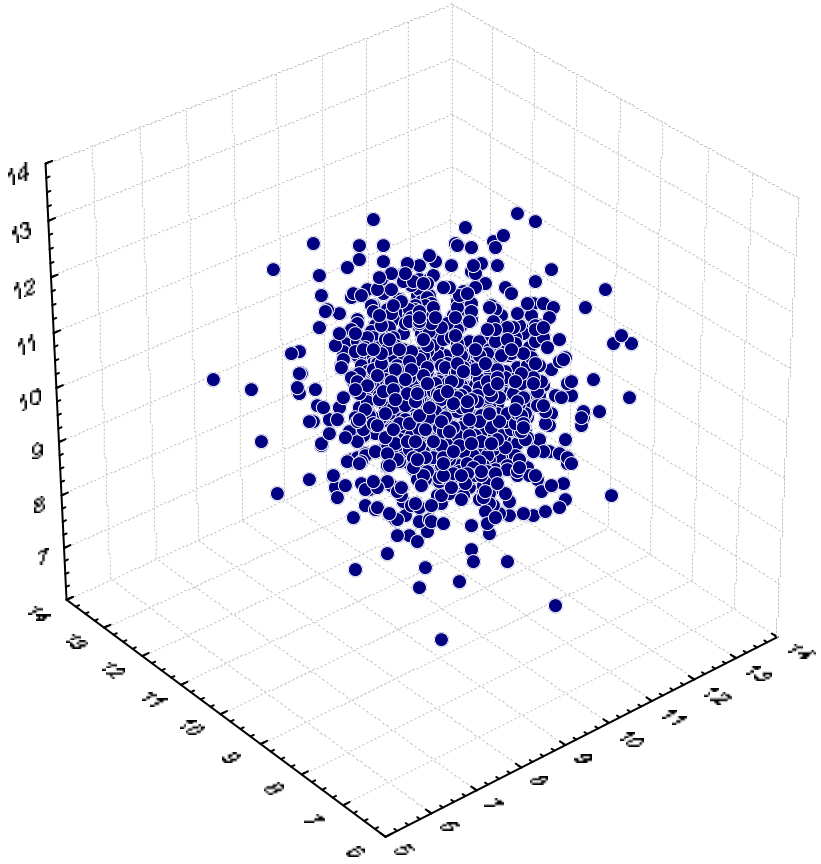
Příklady vícerozměrného rozdělení

- R – knihovna MSBVAR



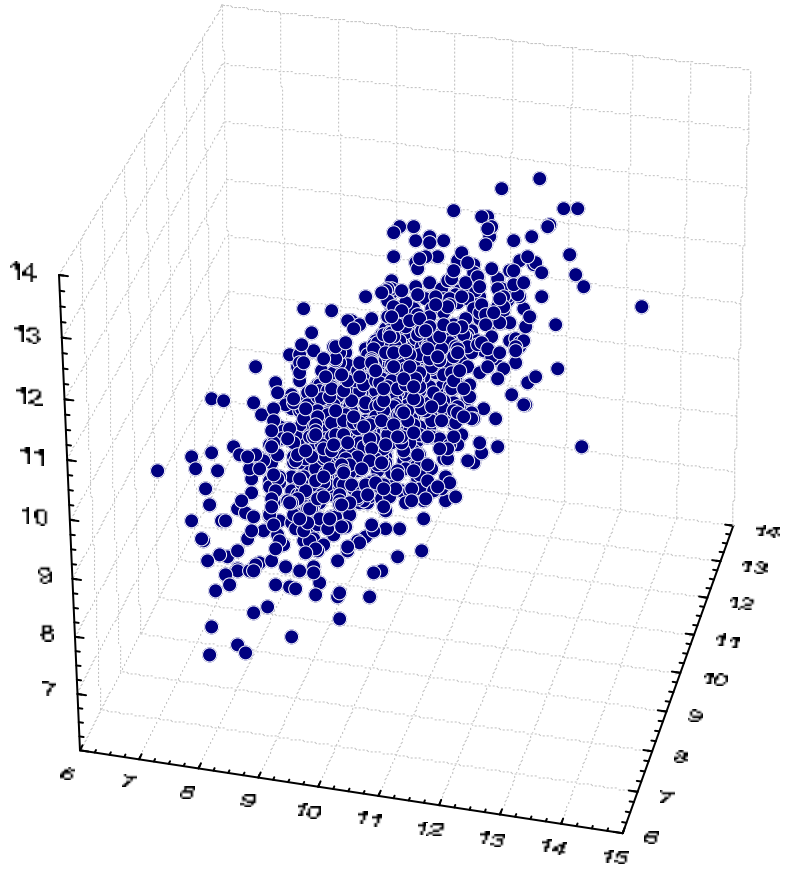
Příklad vícerozměrného rozdělení I

```
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")
```



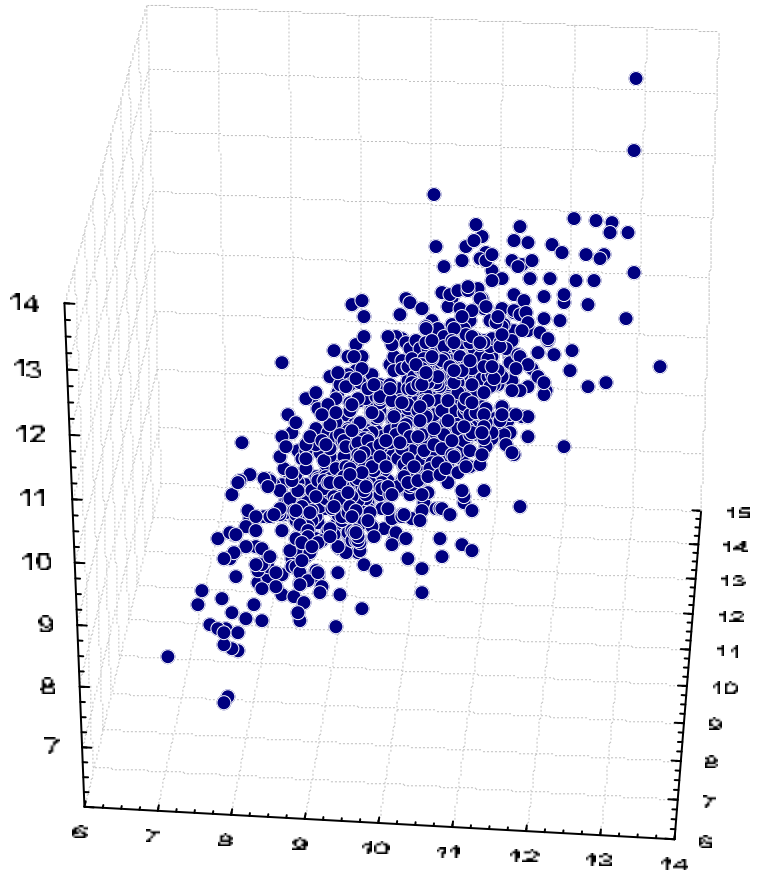
Příklad vícerozměrného rozdělení II

```
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")
```



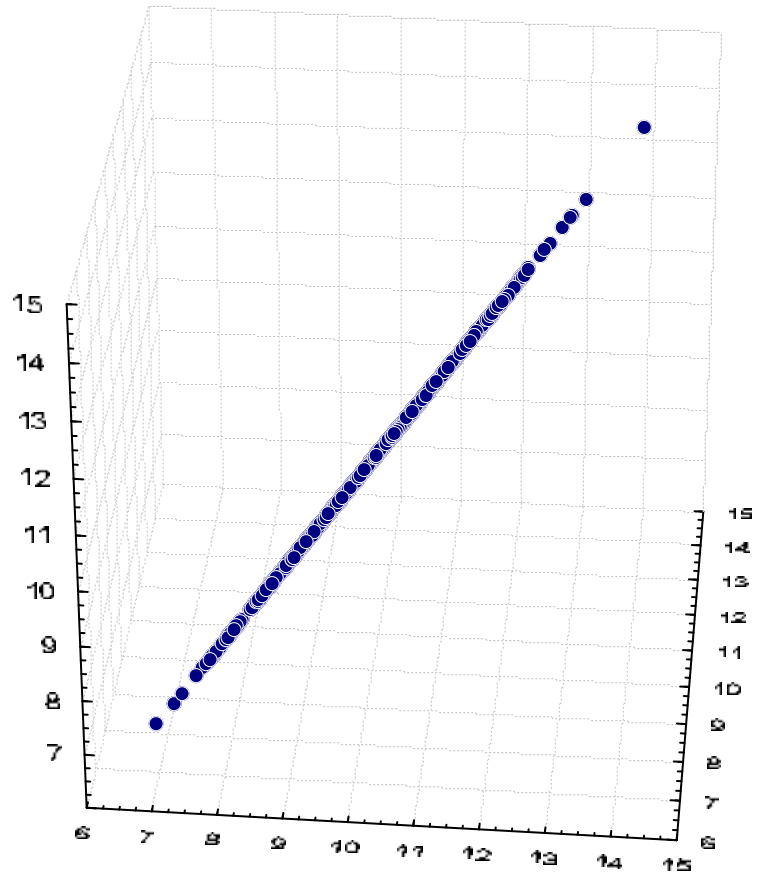
Příklad vícerozměrného rozdělení III

```
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")
```



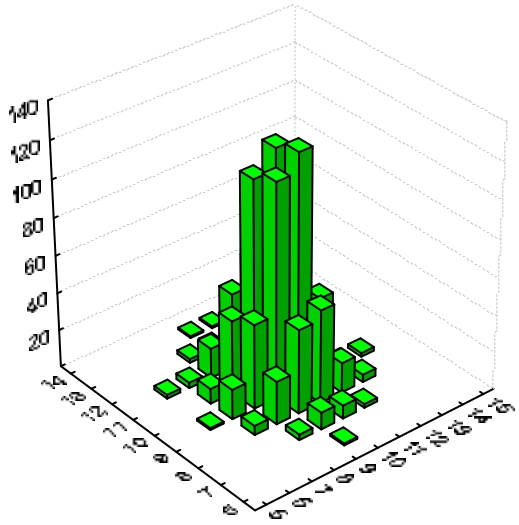
Wishartovo rozdělení

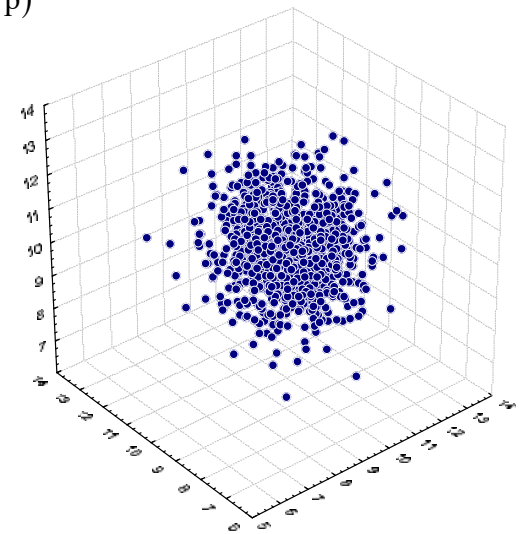
- Wishartovo rozdělení je vícerozměrným zobecněním chi-square rozdělení
- Při odvození některých důležitých algoritmů ve vícerozměrné statistické analýze se uplatňuje dále uvedená vlastnost Wishartova rozdělení.
- Součet nezávislých náhodných matic s Wishartovým rozdělením se shodnou střední hodnotou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

$$\left.\begin{array}{l} \mathbf{A} = \mathbf{A}_1 + \mathbf{A}_2 + \ldots + \mathbf{A}_H \\ \mathbf{A}_h \sim W_p(\nu_h, \boldsymbol{\Sigma}), h = 1,2,\ldots,H \end{array}\right\} \longrightarrow \mathbf{A}_h \sim W_p\left(\sum_{h=1}^{H} \nu_h, \boldsymbol{\Sigma}\right)$$

Hotellingovo rozdělení

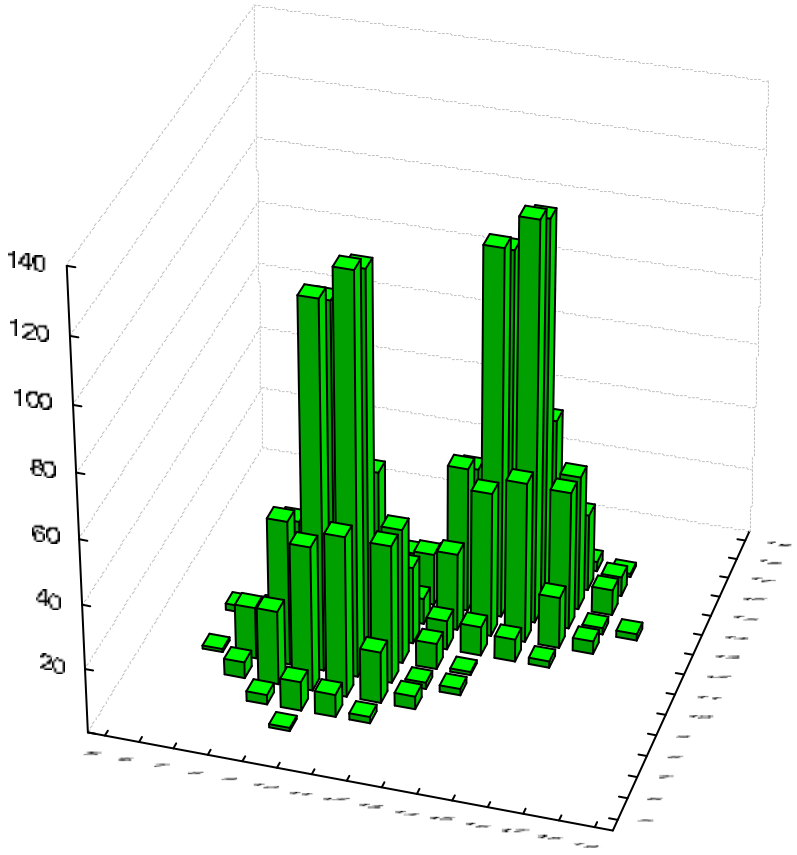
- Jedná se o zobecnění t- rozdělení pro *p*-rozměrný prostor
- Uvažujme regulární čtvercovou matici A *p*-tého řádu a rozdělením $W_p(\nu,\Sigma)$ a na A nezávislý *p*-položkový vektor a s rozdělením $N_p\left(\mathbf{o}_p,\Sigma/c\right)$ Potom kvadratická forma $Q_1 = c\nu\mathbf{a}^T\mathbf{A}^{-1}\mathbf{a}$ 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) $X \sim N(\mu,\sigma^2) \longrightarrow \dfrac{\bar{x}-\mu}{\sqrt{\dfrac{s^2(x)}{n}}} \sim t(n-1)$
- Druhou mocninu této statistiky můžeme upravit a zapsat ve tvaru $t^2 = n(\bar{x}-\mu)\left[s^2(x)\right]^{-1}(\bar{x}-\mu)$ 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 $x \sim N_p(\mu,\Sigma) \longrightarrow n(x-\mu)^T S^{-1} \sim T^2(p,n-p)$

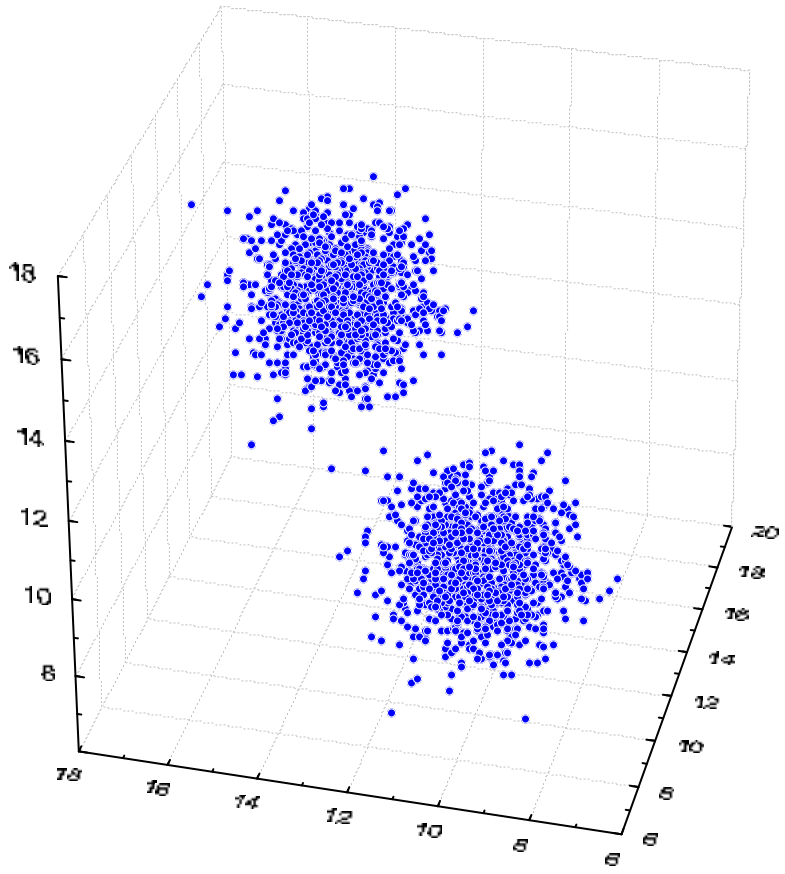




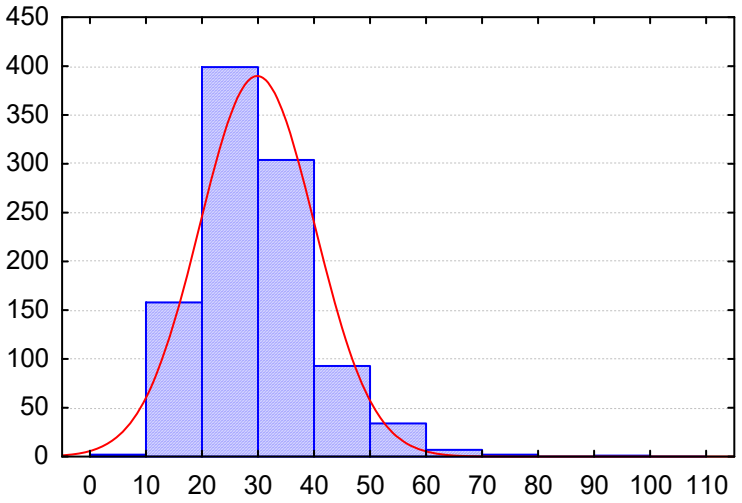
Normalita ve vícerozměrném prostoru

- Normalita ve vícerozměrném prostoru

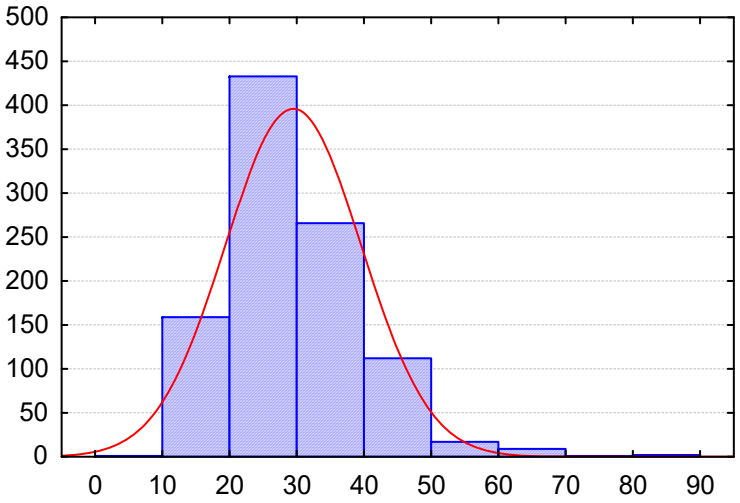


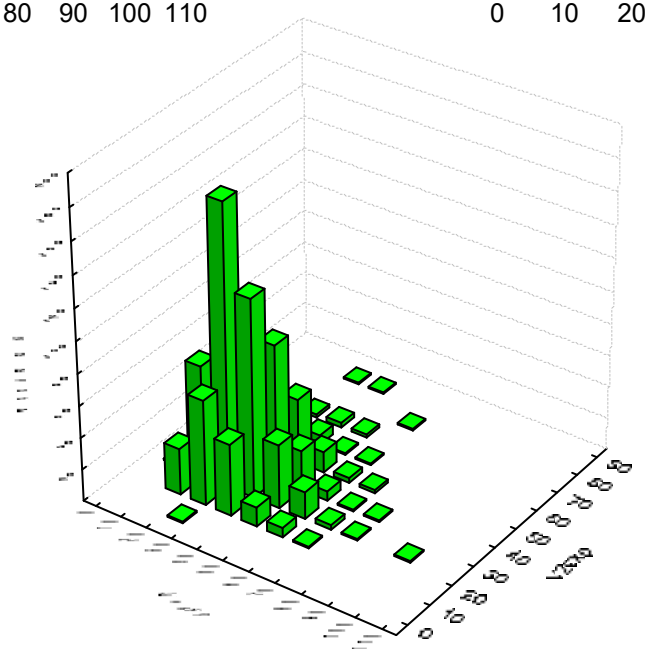


Nenormální rozložení ve vícerozměrném prostoru

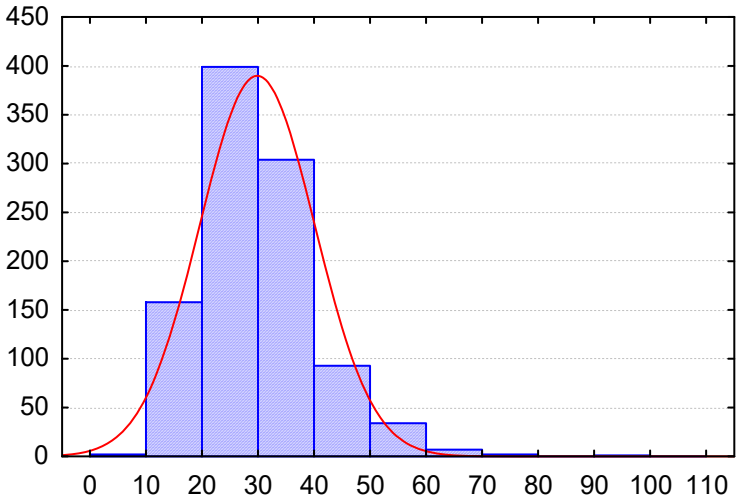


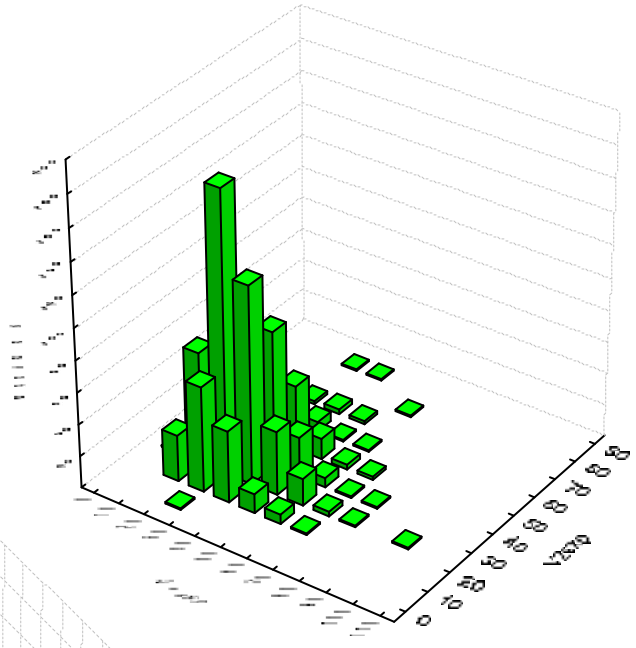
+

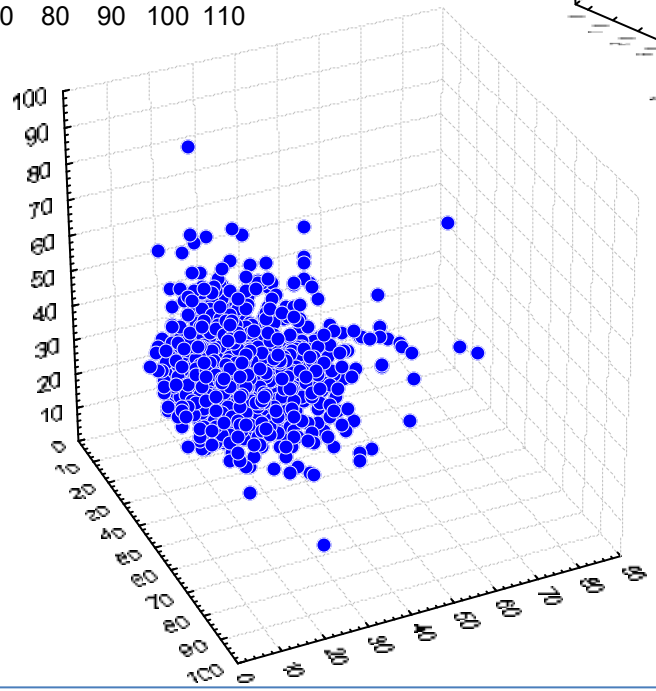




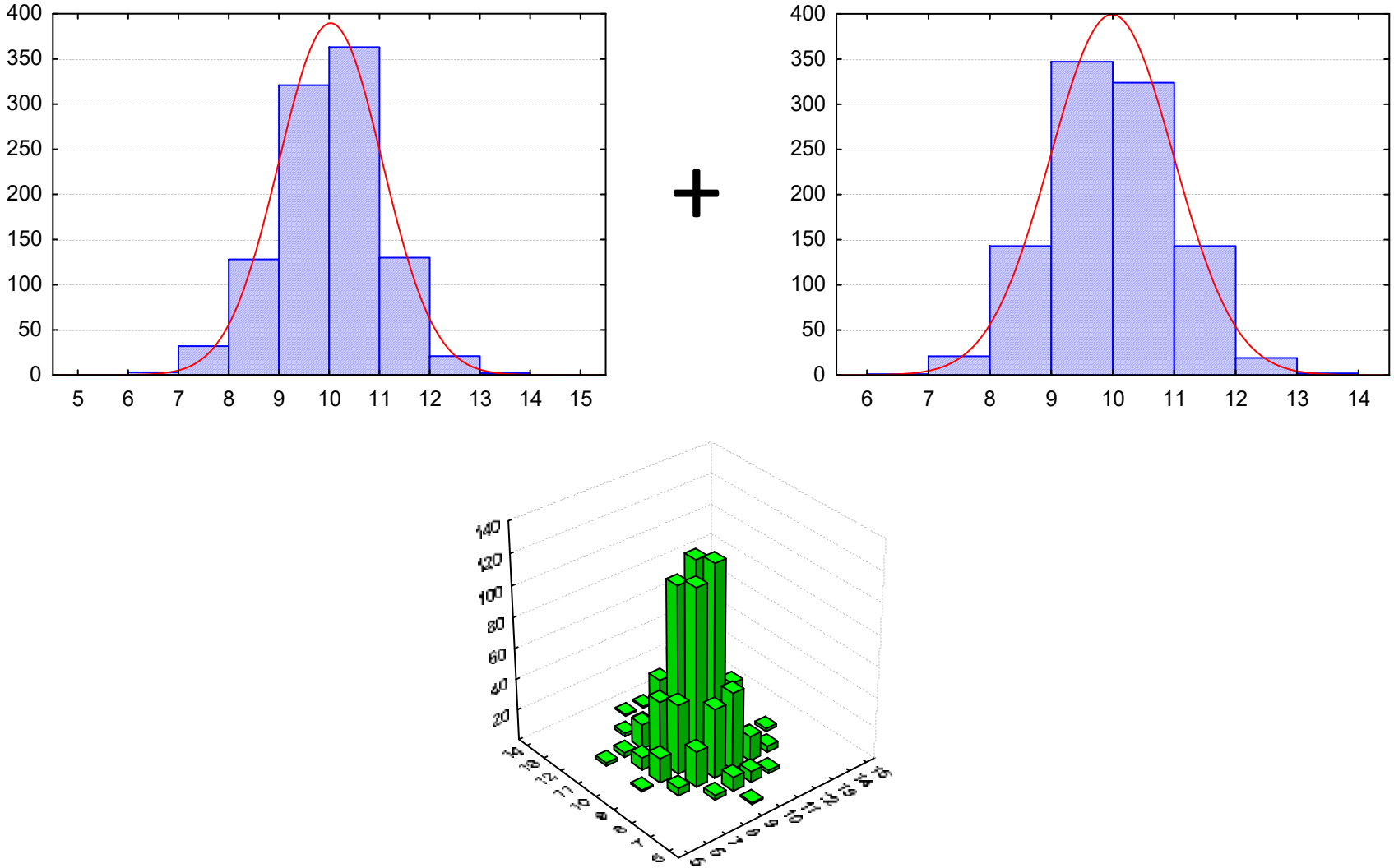
Nenormální rozložení ve vícerozměrném prostoru



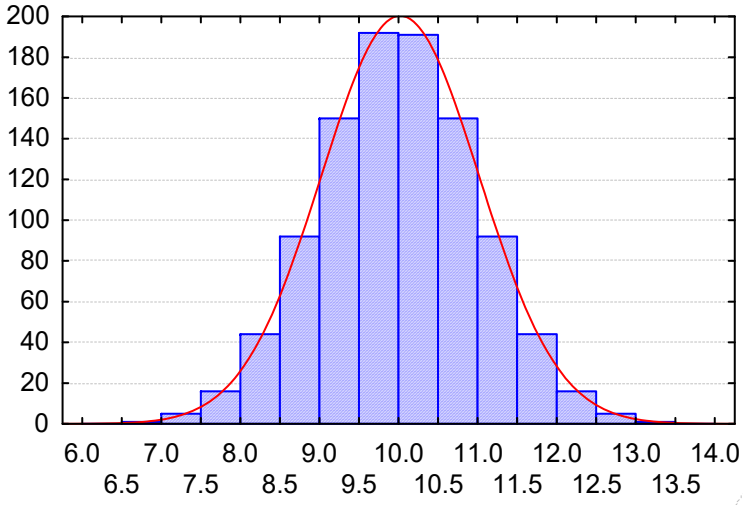




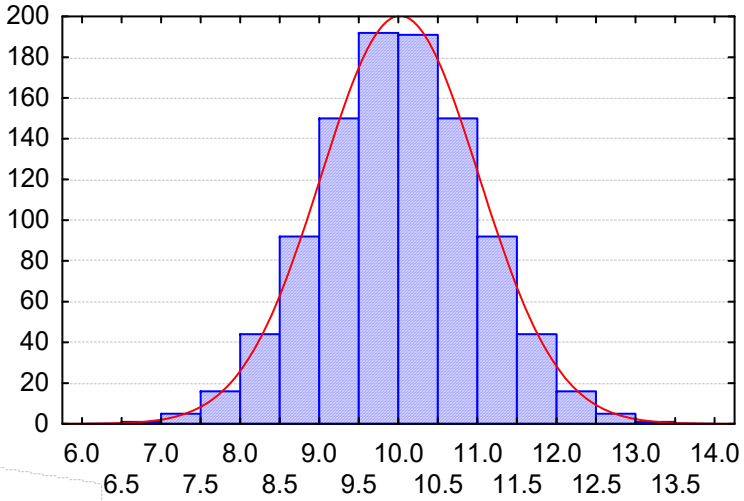
Je normalita v jednorozměrném prostoru jedinou podmínkou vícerozměrné normality?

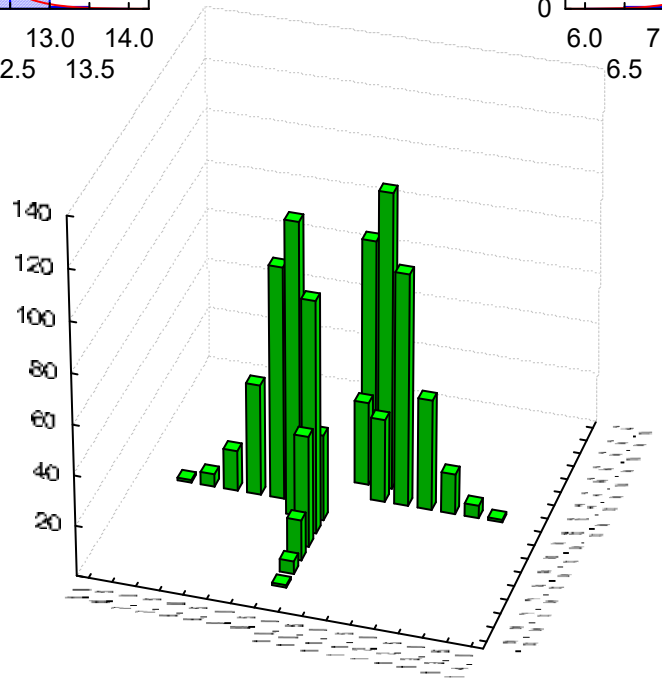


Je normalita v jednorozměrném prostoru jedinou podmínkou vícerozměrné normality?

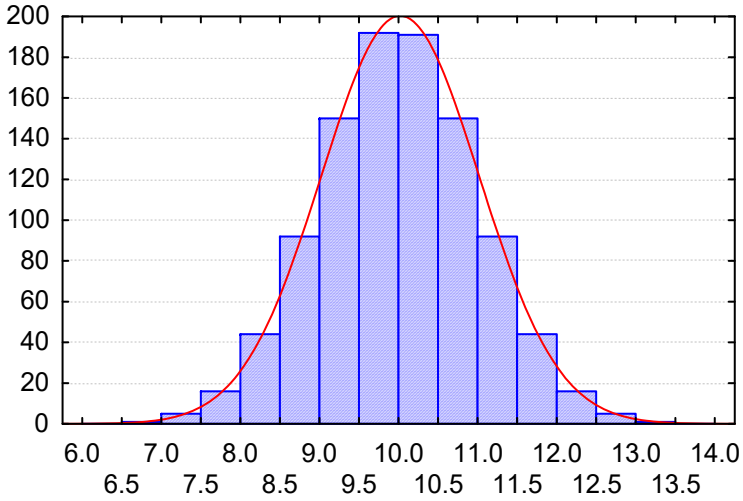


+

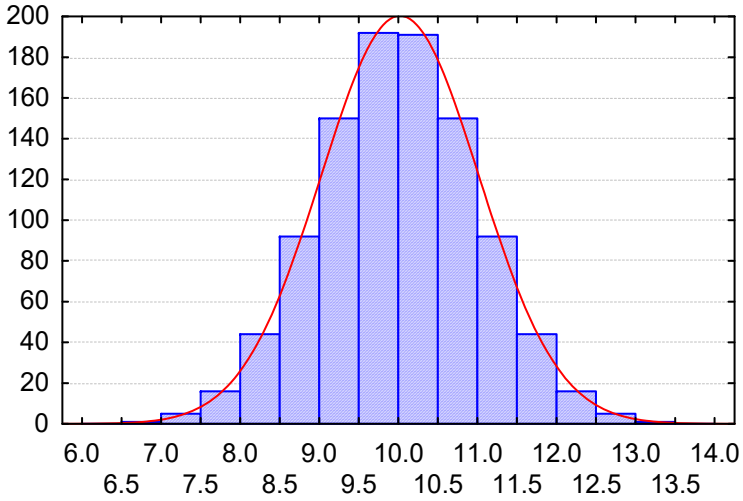


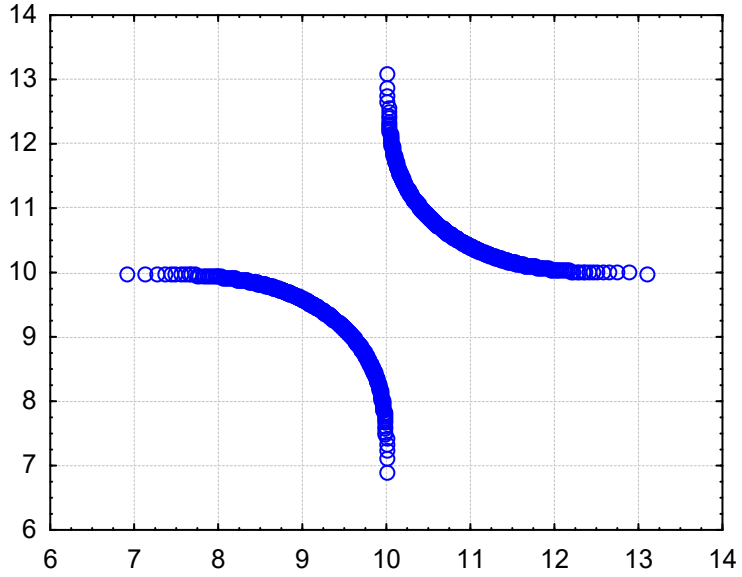


Je normalita v jednorozměrném prostoru jedinou podmínkou vícerozměrné normality?

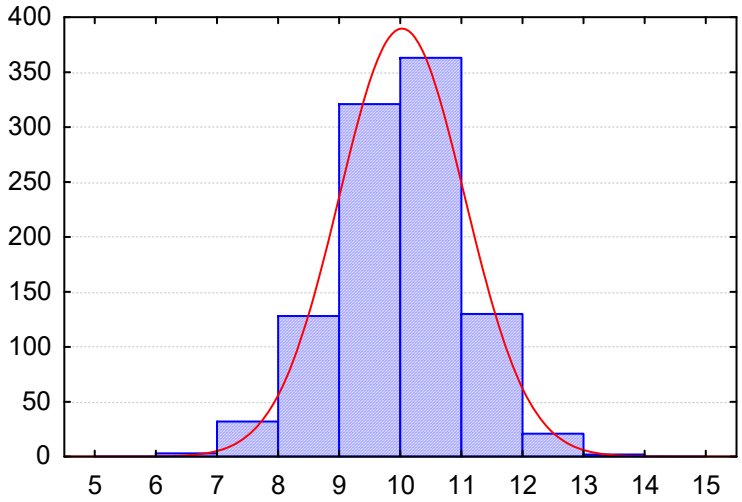


+

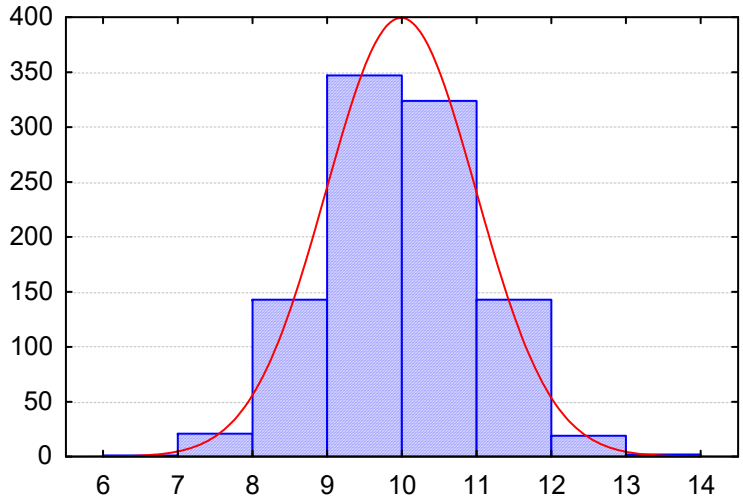


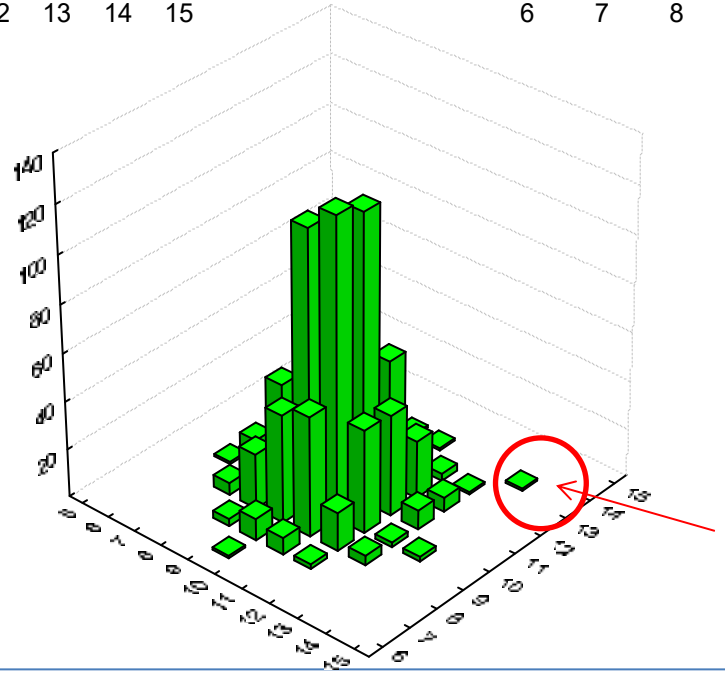


Vícerozměrný outlier



+

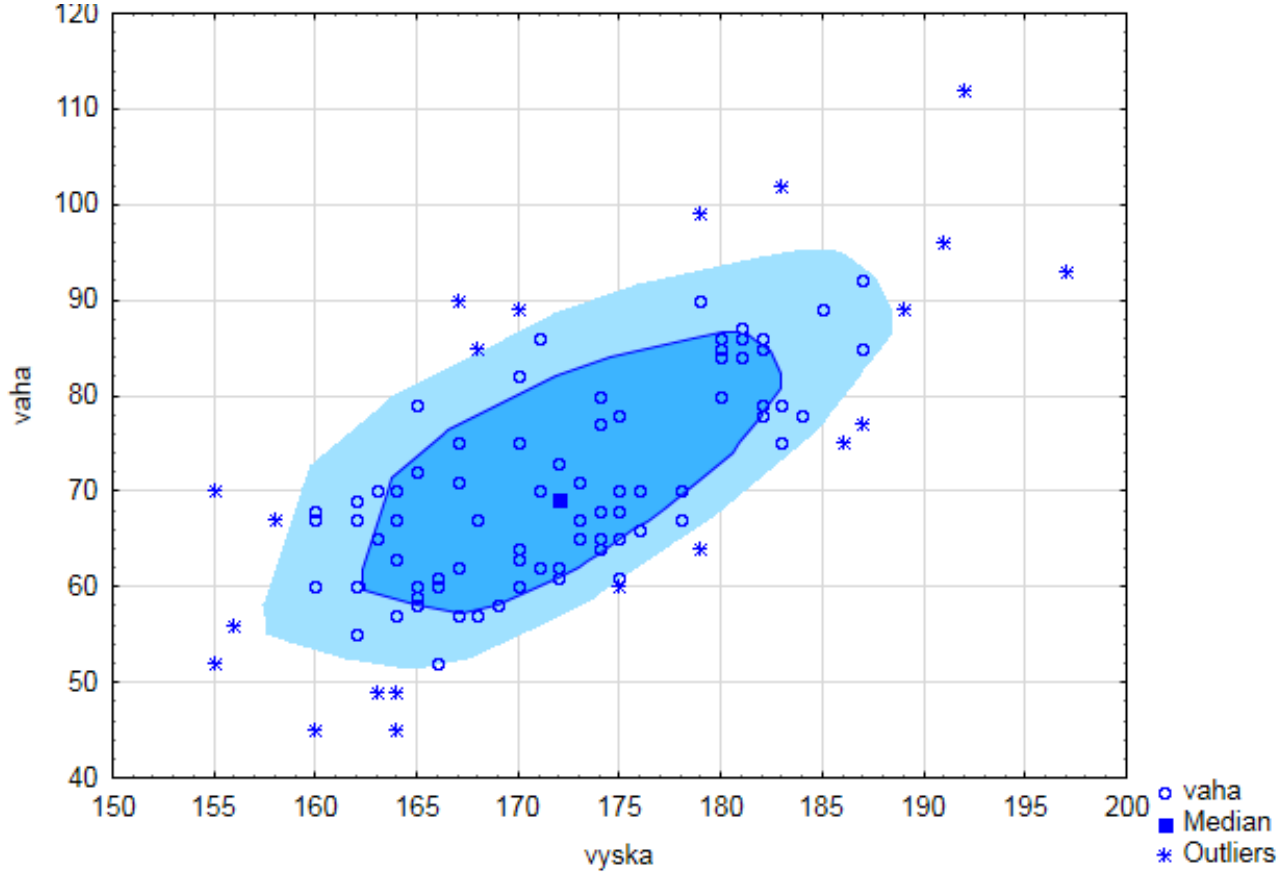




Vícerozměrná odlehlá hodnota (outlier)

Ověření dvourozměrné normality

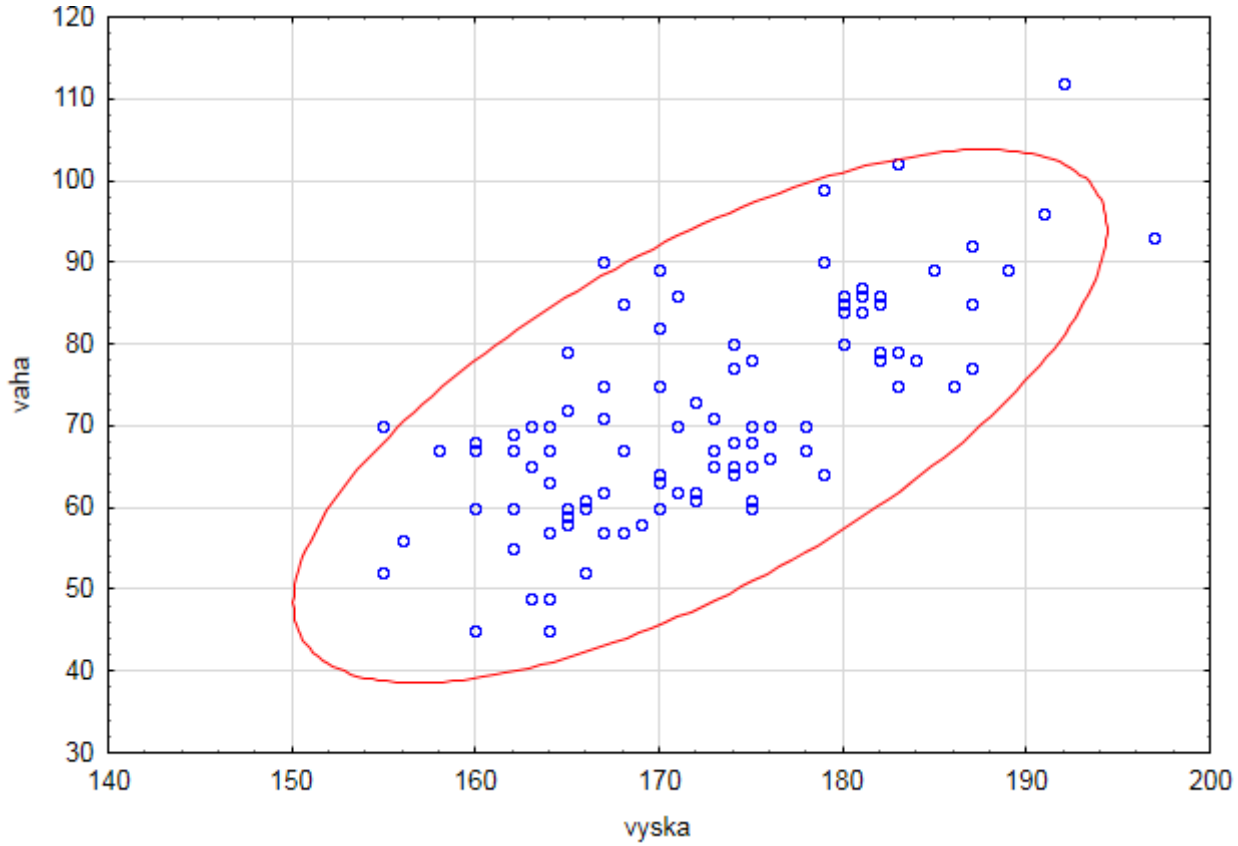
Bagplot = „bivariate boxplot“ (tzn. „dvourozměrný krabicový graf“)



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

Ověření dvourozměrné normality

Vykreslení regulační elipsy („control“ elipse):



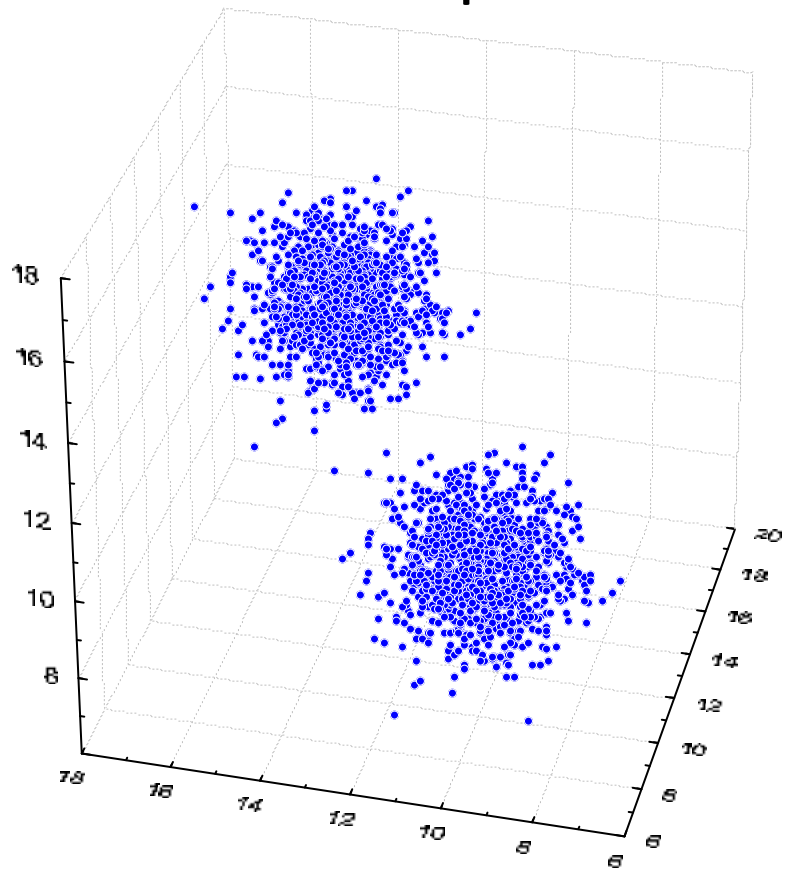
v softwaru Statistica: Graphs – Scatterplots – na záložce Advanced zvolit Elipse Normal

Srovnání průměrů ve vícerozměrném prostoru

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

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

- 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

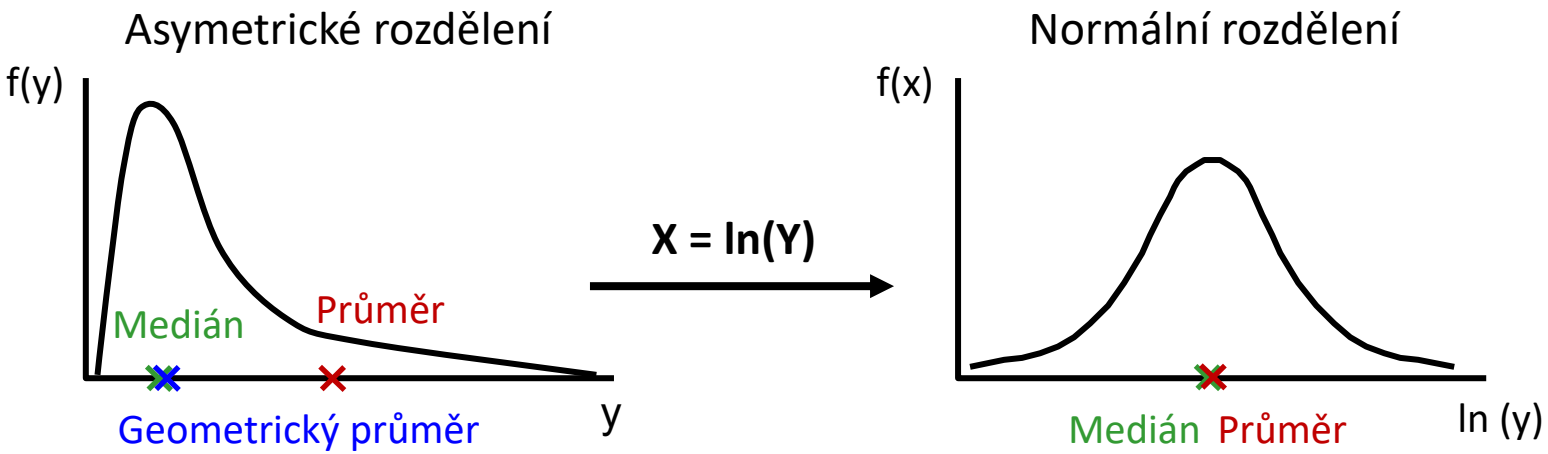
Variable	Mean 1	Mean 2	t-value	df	p	Valid N 1	Valid N 2	**Std.Dev. 1**	Std.Dev. 2	F-ratio Variances	p Variances
V1	10.00068	14.00068	-87.3755	1998	0.00	1000	1000	1.023659	1.023659	1.000000	1.000000
V2	9.96685	13.96685	-89.5768	1998	0.00	1000	1000	0.998503	0.998503	1.000000	1.000000
V3	10.00140	14.00140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

Typy transformací a jiných úprav vícerozměrných dat

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

Normalizace dat

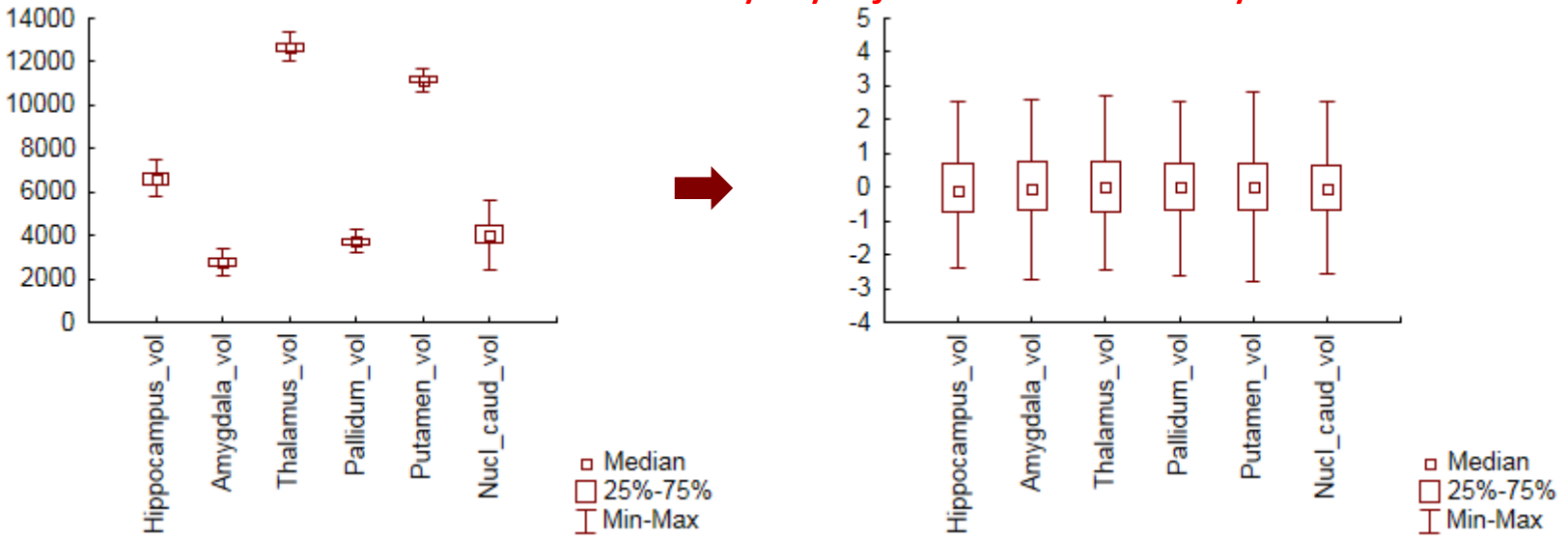
- převod na normální rozdělení (normalita je předpokladem řady statistických testů).
- např. **logaritmická transformace**: X = ln(Y) nebo X = ln(Y+1), pokud data obsahují hodnotu 0



- 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 transfomace** (pro proměnné s binomickým rozložením)
  - **Box-Coxova tranformace**

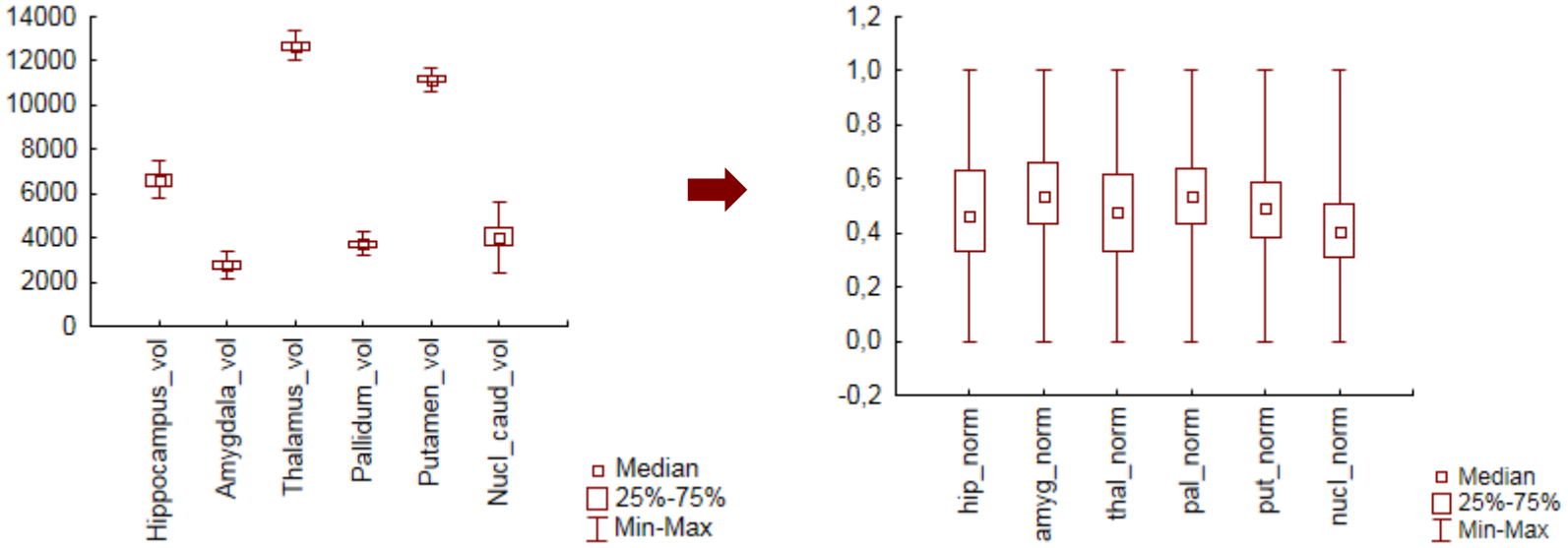
Standardizace dat

- 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, že proměnné nemají normální rozdělení a že se v datech vyskytují odlehlé hodnoty!!!



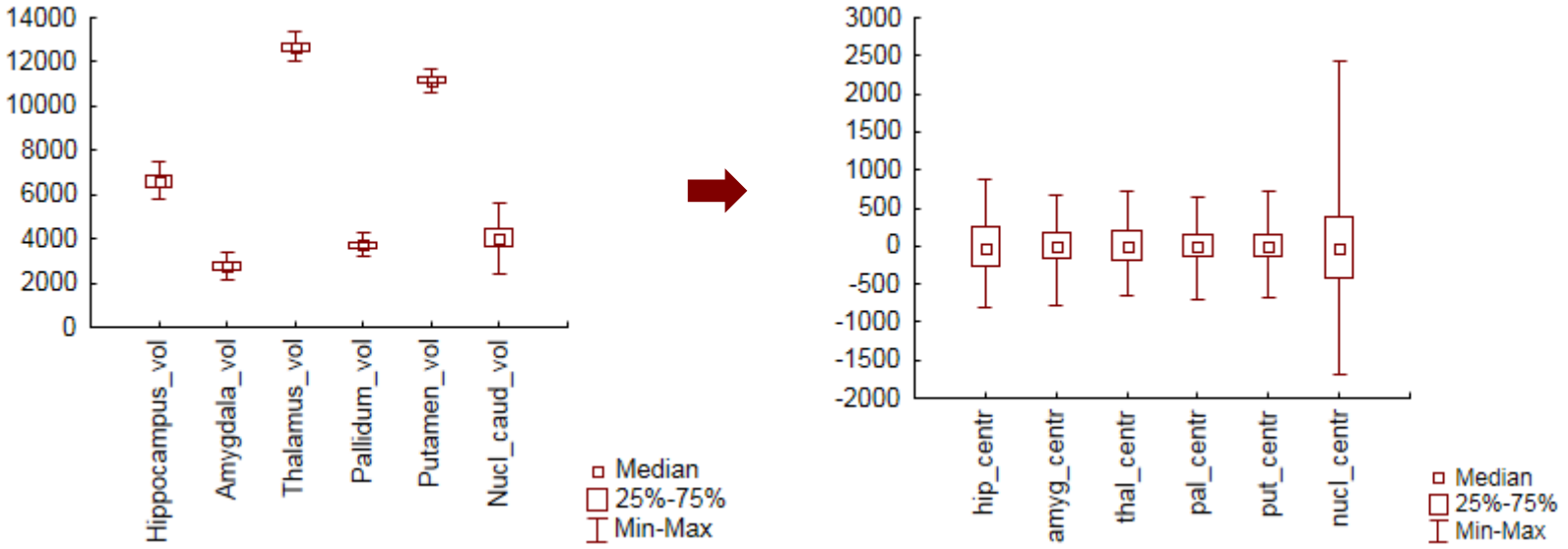
Min-max normalizace

- 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 (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 ↑↓
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