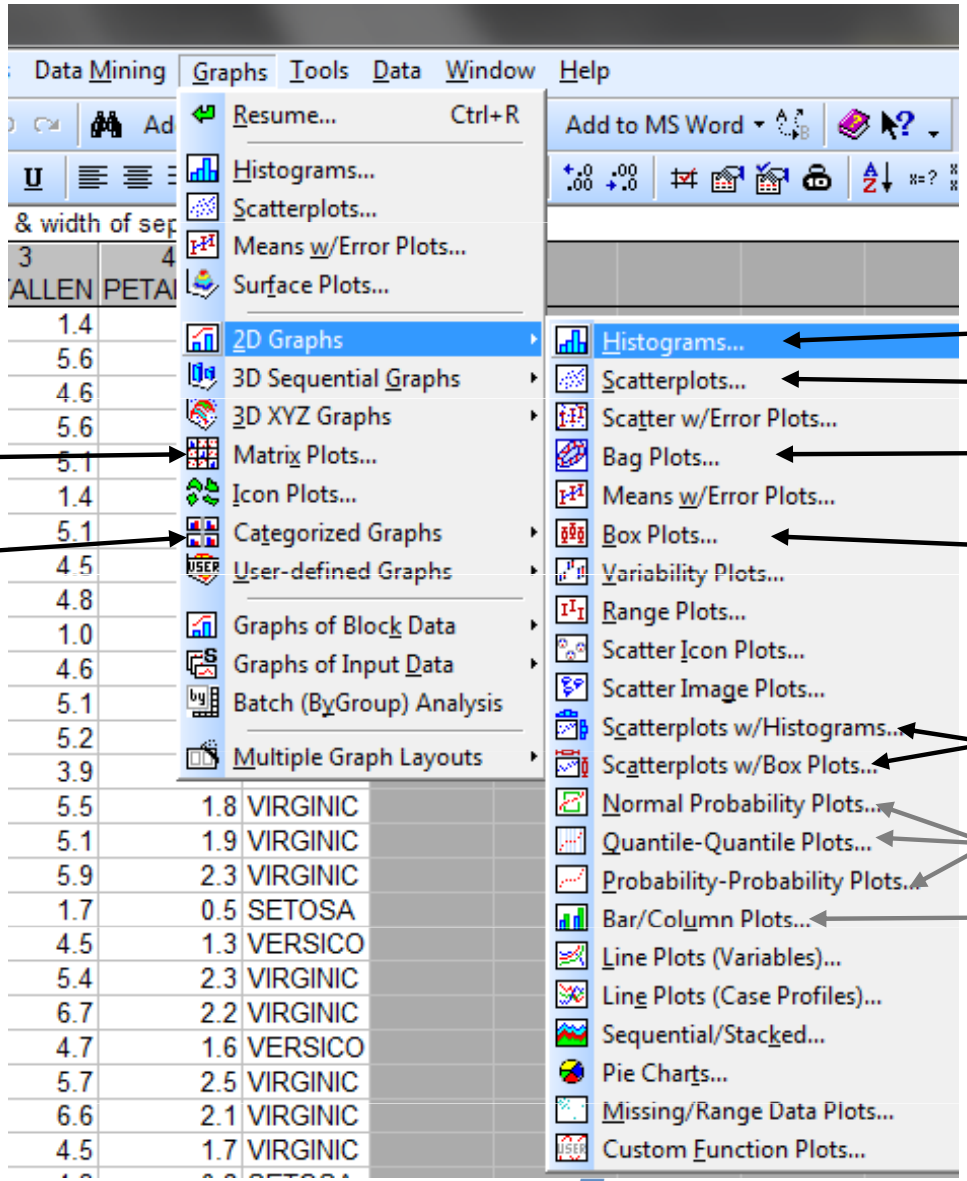
2. cvičení

14.10.2014

Grafy ve statistice – 2D



Opakování

- Datový soubor BMI (list: pacienti)
- Oveřit normalitu
- Korelace
- Kategorizace
- Kontigenční tabulky

2D grafy

- Histogram
  - „správný“ histogram
    - obsah jednoho sloupečku je relativní četnost daného intervalu, a výška sloupečku je hustota četnosti
  - „používaný“ histogram
    - výška sloupečku je absolutní nebo relativní četnost daného intervalu
  - většina SW kreslí „používaný“ histogram
- Matrix plot
  - Kombinovaný graf
- Box plot (Krabicový graf)
  - umožňuje posoudit symetrii a variabilitu datového souboru a odlehlé a extrémní hodnoty
  - odlehlá hodnota: $(x_{0.75} + 1{,}5q, x_{0.75} + 3q)$ nebo $(x_{0.25} - 1{,}5q, x_{0.25} - 3q)$
  - extremní hodnota: $(x_{0.75} + 3q, \infty)$ nebo $(-\infty, x_{0.25} - 3q)$
  - SW Statistka umožňuje vlastní nastavení

Asociace ve vícerozměrném prostoru

Obsah

- Principy asociace ve vícerozměrném prostoru
- Euklidovská vzdálenost, Manhattan distance
  - Odvodit asociační matici 5x5
  - Pythagorova věta (excel, statistka, SPSS)
  - Pomocí makra v excelu horní trojúhelníkovou matici zlinearizovat a vykreslit do histogramu
- Soubor s množstvím bodů (opět např. města)
  - Odvodit asociační matici nxn vzdušnou čarou
  - Odvodit asociační matici nxn po silnici
  - Ukázat opět xy graf a komentář, že jde o značně obtížnější problém
- Horní trohúhelníkové matice zlinerizovat a dát do xy grafu proti sobe

Asociace ve vícerozměrném prostoru

Data

STATISTICA - [Data: Adstudy* (24v by 50c)]

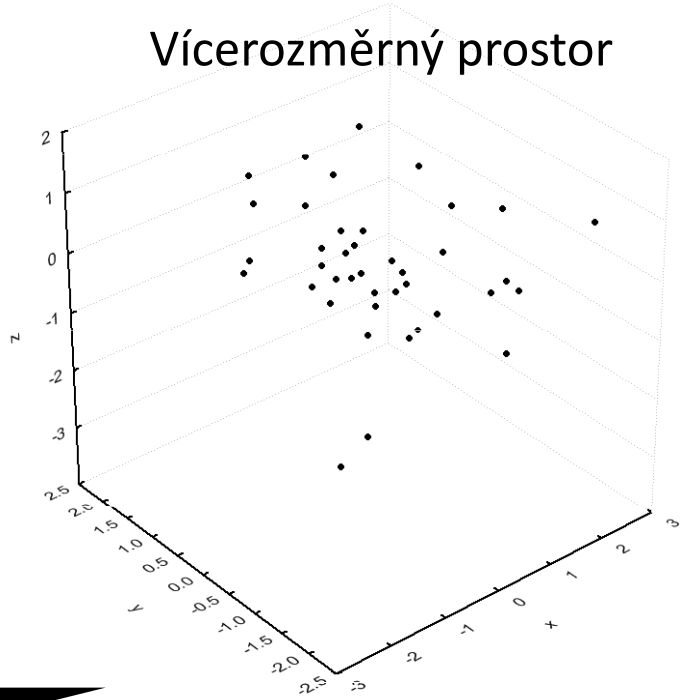
File Edit View Insert Format Statistics Data Mining Graphs Tools Data Window

Add to Workbook Add to Report

Arial 10 B I U

	Advertising Effectiveness Study.				
	1 ADVERT	2 MEASURE01	3 MEASURE02	4 MEASURE03	5 MEASU
R. Rafuse	id_1	9	1	6	
T. Leiker	id_2	6	7	1	
E. Bizot	id_3	9	8	2	
K. French	id_4	7	9	0	
E. Van Landuyt	id_5	7	1	6	
K. Harrell	id_6	6	0	0	
W. Noren	id_7	7	4	3	
W. Willden	id_8	9	9	2	
S. Kohut	id_9	7	8	2	
B. Madden	id_10	6	6	2	

Vícerozměrný prostor



Asociační matice

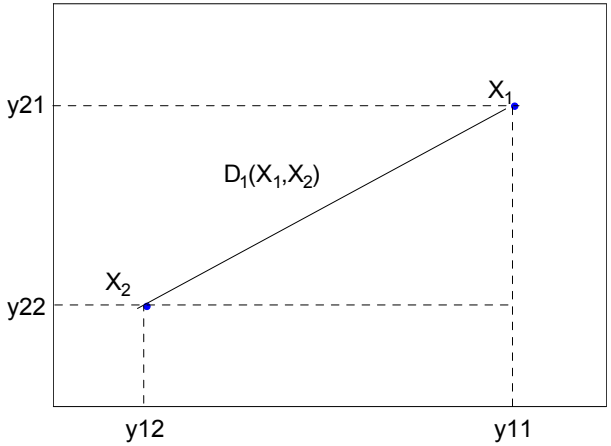
Case No.	Euclidean distances (multidinensional_normal_distribution)											
	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_10	C_11	C_12
C_1	0.00	2.58	3.73	0.95	1.46	1.85	3.78	1.85	1.59	1.80	0.46	2.82
C_2	2.58	0.00	2.17	2.58	1.90	0.82	2.32	1.99	1.14	1.57	2.26	1.73
C_3	3.73	2.17	0.00	3.15	3.54	2.19	1.66	2.49	2.34	2.58	3.53	2.88
C_4	0.95	2.58	3.15	0.00	1.85	1.77	3.45	1.31	1.53	1.56	1.01	3.00
C_5	1.46	1.90	3.54	1.85	0.00	1.50	3.88	1.58	1.58	1.12	1.02	2.90
C_6	1.85	0.82	2.19	1.77	1.50	0.00	2.40	1.38	0.41	1.07	1.56	1.81
C_7	3.78	2.32	1.66	3.45	3.88	2.40	0.00	3.31	2.36	3.28	3.70	1.86
C_8	1.85	1.99	2.49	1.31	1.58	1.38	3.31	0.00	1.48	0.59	1.53	3.14
C_9	1.59	1.14	2.34	1.53	1.58	0.41	2.36	1.48	0.00	1.27	1.39	1.68

Euklidovská vzdálenost

- Jde o základní metrické měřítko vzdálenosti a počítá vzdálenost objektů obdobně jako Pythagorova věta počítá přeponu pravoúhlého trojúhelníku. Metoda je citlivá na rozdílný rozsah hodnot vstupujících proměnných (vhodným řešením může být standardizace) a double zero problém. Nemá horní hranici hodnot.

$$D_1(x_1, x_2) = \sqrt{\sum_{j=1}^{p} (y_{1j} - y_{2j})^2}$$

- Jako další měřítko se používá také čtverec této vzdálenosti. . Jeho nevýhodou jsou semimetrické vlastnosti.



$$D_1^2(x_1, x_2) = \sum_{j=1}^{p} (y_{1j} - y_{2j})^2$$

Průměrná vzdálenost

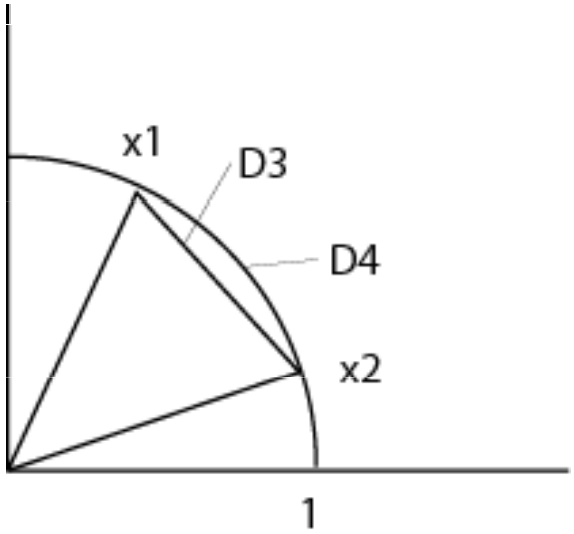
- Euklidovská vzdálenost je přepočítána na počet parametrů (druhů v případě vzdálenosti společenstev odběrů).

$$D_2^2(x_1, x_2) = \frac{1}{p} \sum_{j=1}^{p} (y_{1j} - y_{2j})^2$$

$$D_2(x_1, x_2) = \sqrt{D_2^2}$$

Chord distance (Orlóci, 1967)

- Odstraňuje double zero problém a vliv rozdílného počtu jedinců druhů ve vzorcích při výpočtu Euklidovské vzdálenosti. Její maximální hodnota je druhá odmocnina ze dvou a minimum 0. Při výpočtu počítá pouze s poměry druhů v rámci jednotlivých vzorků. Jde vlastně o Euklidovskou vzdálenost počítanou pro vektory vzorků standardizované na délku 1, nebo je možný přímý výpočet už zahrnující standardizaci. Vnitřní část výpočtu je vlastně cosinus úhlu svíraného vektory, zápis vzorce je možný i v této formě.

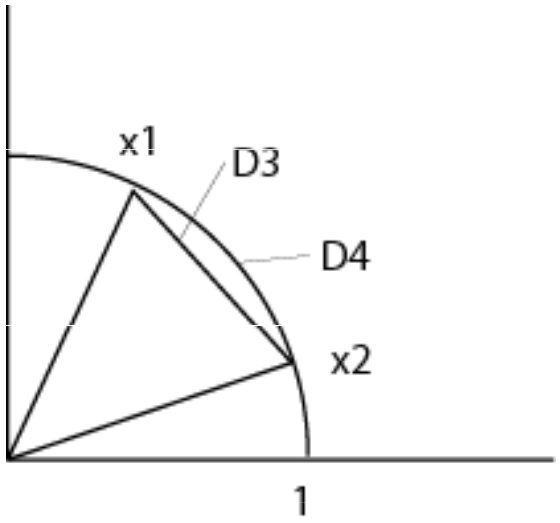


$$D_3(x_1, x_2) = \sqrt{2\left(1 - \frac{\sum_{j=1}^{p} y_{1j} y_{2j}}{\sqrt{\sum_{j=1}^{p} y_{1j}^2 \sum_{j=1}^{p} y_{2j}^2}}\right)}$$

$$D_3 = \sqrt{2(1 - \cos\theta)}$$

Geodetická metrika

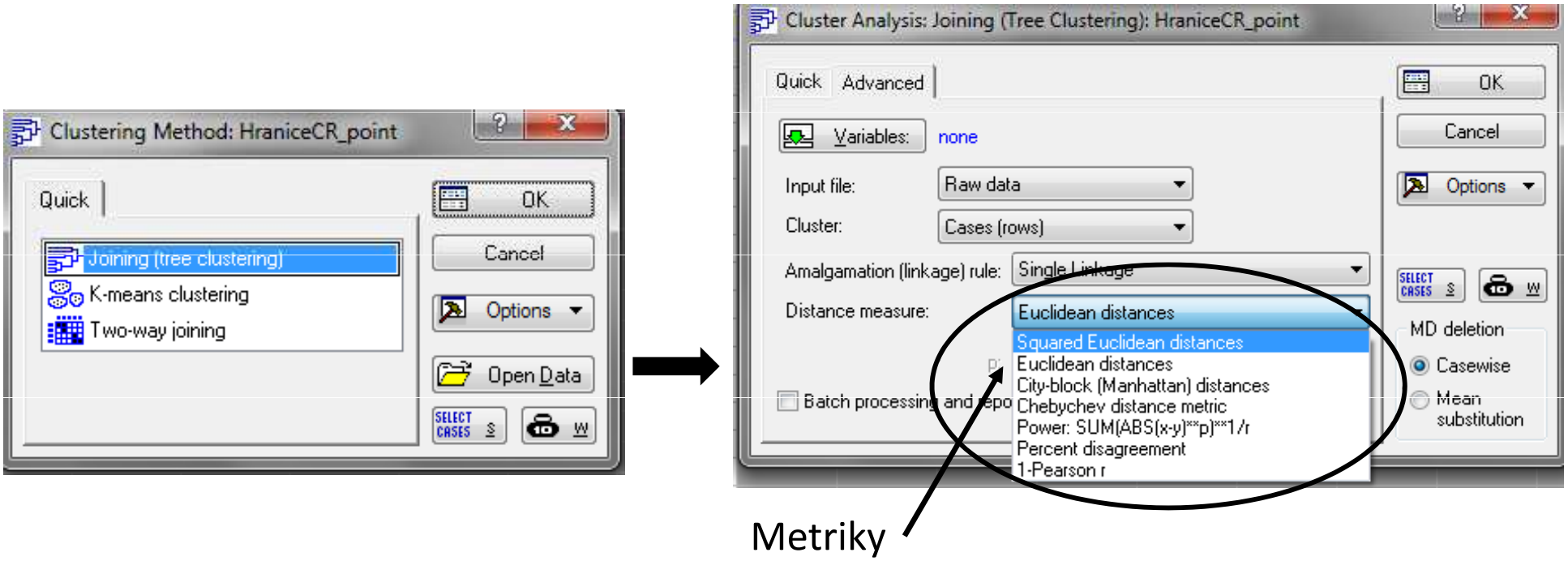
- Počítá délku výseče jednotkové kružnice mezi normalizovanými vektory (viz. Chord distance).



$$D_4(x_{1,}x_2) = \arccos\left[1 - \frac{D_3^2(x_1, x_2)}{2}\right]$$

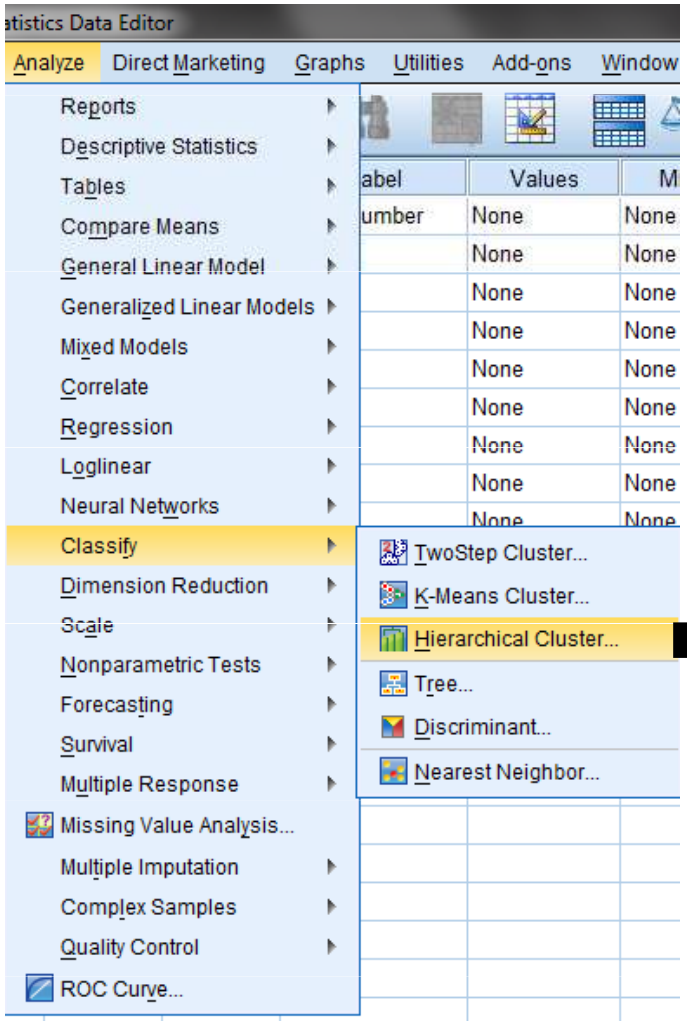
Asociační matice 1

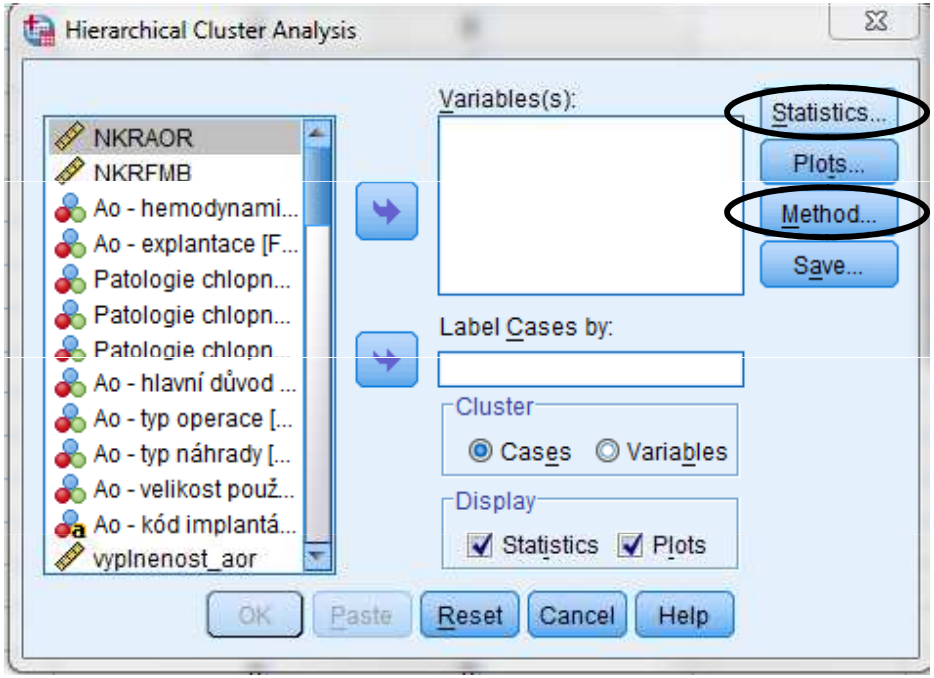
- **SW Statistika**



Asociační matice 2a

- SW SPSS

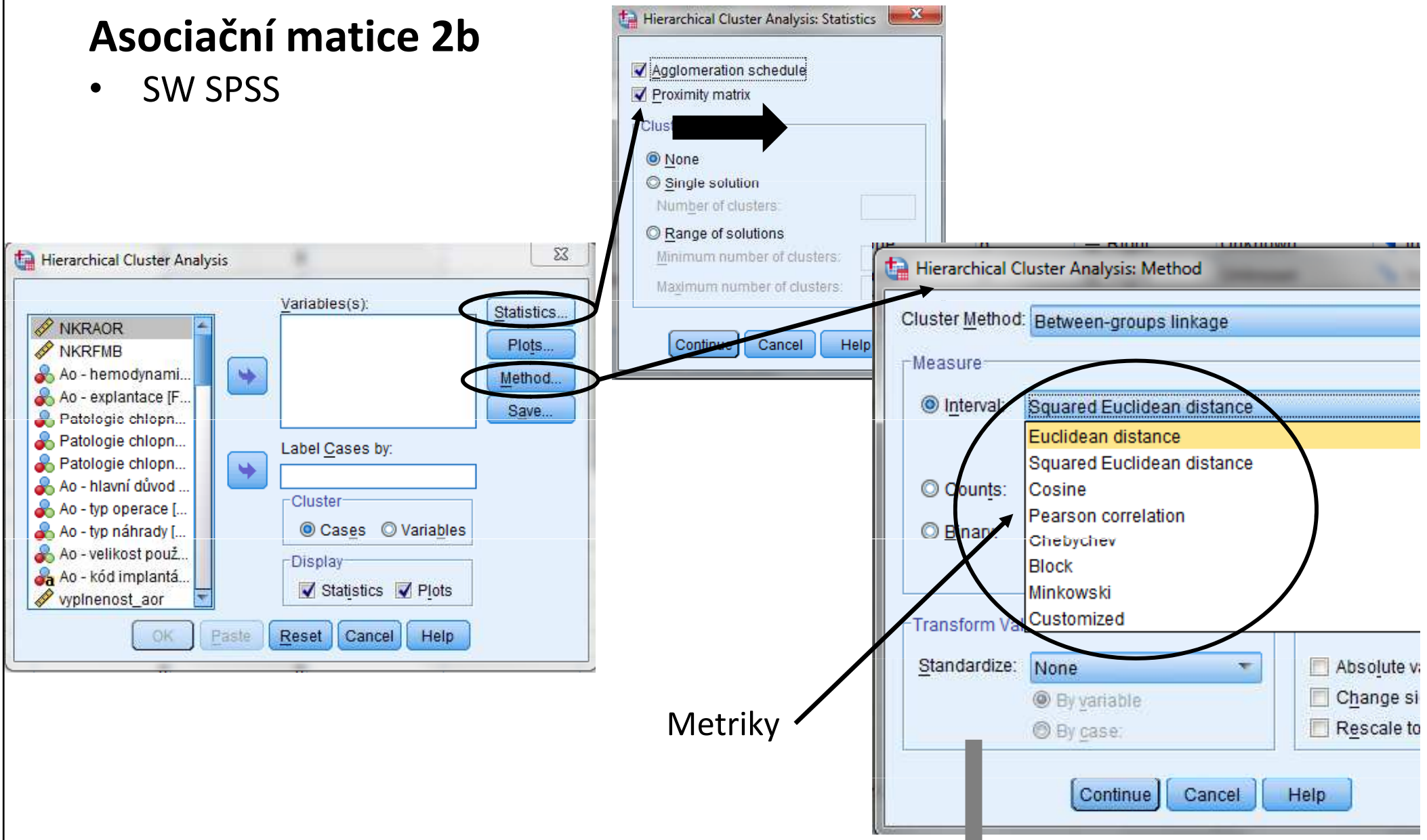





```
CLUSTER   seznam promennych
  /MEASURE=SEUCLID
  /PRINT DISTANCE
  /MATRIX OUT('C:\cesta a nazev.sav').
```

Asociační matice 2b

- SW SPSS



R-ko

- Volně přístupný na http://www.r-project.org/

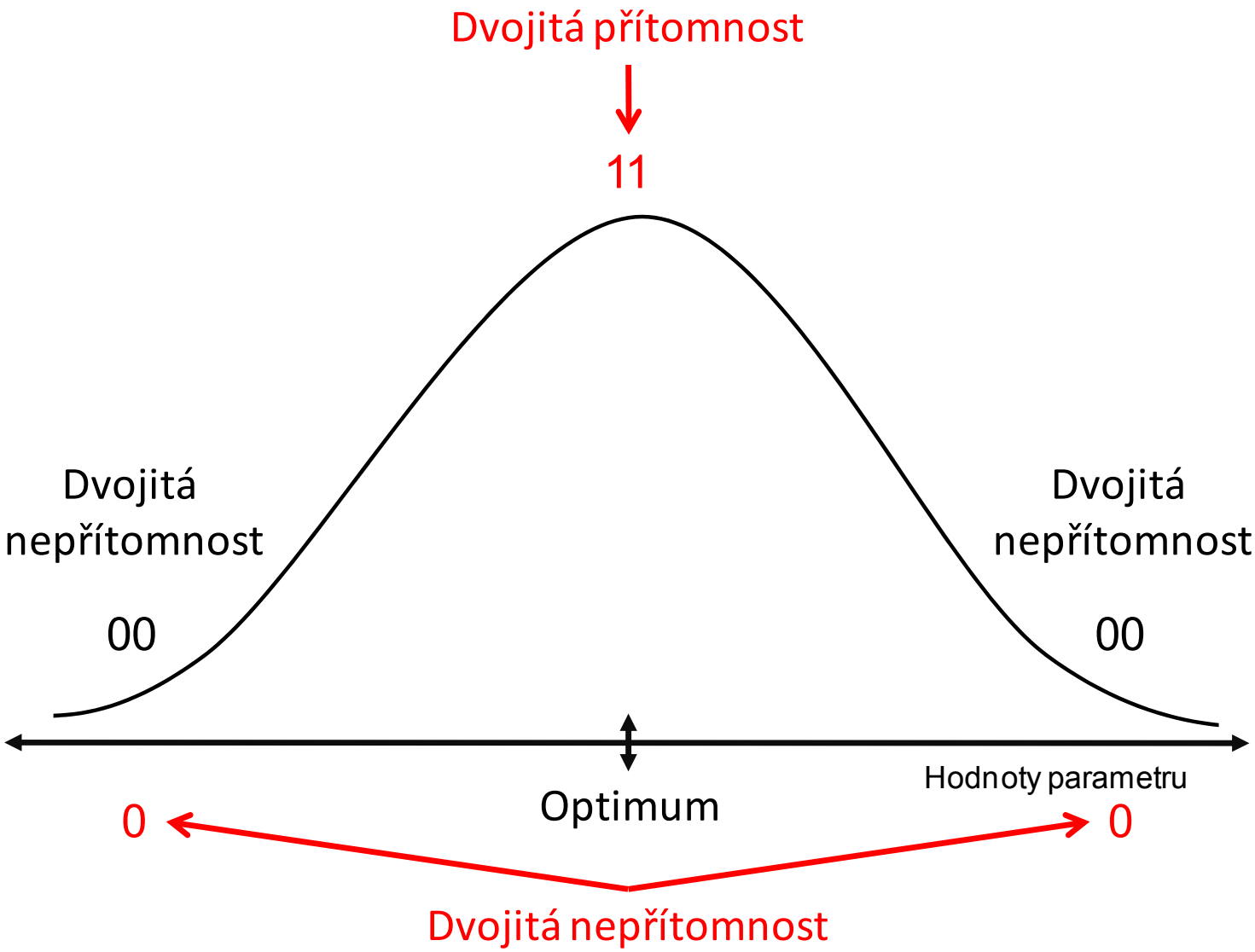
- Klady
  - Velké množství základních i pokročilejších funkcí pro statistickou analýzu
  - Velké možnosti v úpravě grafů

- Zápory
  - Příkazový řádek
  - Nevidíme data

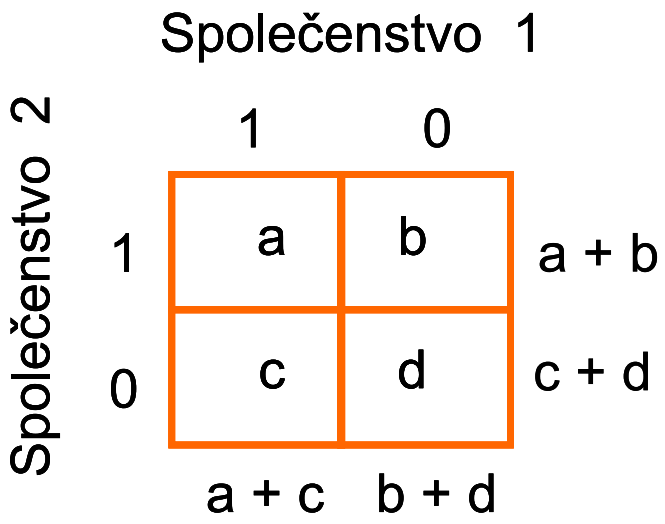
Vzdálenosti měst ČR

Binární koeficienty

Doble-zero problem



Binární koeficienty



a, b, c, d – počet případů kdy souhlasí binární charakteristika společenstev

a + b + c + d = n

- **Symetrické binární koeficienty**
  - nerozlišují mezi případy 0-0, 1–1, jsou citlivé na double-zero problém
- **Asymetrické binární koeficienty**
  - rozlišují mezi případy 0–0 a 1–1, tímto vylučují problém double -zero. Tyto koeficienty mohou být použity ve shlukové analýze

Symetrické binární koeficienty

- Simple matching koeficient

$$S_1(x_1, x_2) = \frac{a+d}{a+b+c+d}$$

- Rogers & Tanimoto koeficient

$$S_2(x_1, x_2) = \frac{a+d}{a+2b+2c+d}$$

Asymetrické binární koeficienty

- Jaccardův koeficient

$$S_7 = \frac{a}{a+b+c}$$

- Sørensenův koeficient

$$S_8 = \frac{2a}{2a+b+c}$$