

# 2. cvičení

14.10.2014

# Grafy ve statistice – 2D

The image shows the 'Graphs' menu in Minitab software. The menu is open, showing various options for creating 2D graphs. Arrows point from text labels to specific menu items:

- Histogramy** points to 'Histograms...'
- XY grafy** points to 'Scatterplots...'
- Bag ploty** points to 'Bag Plots...'
- Box ploty** points to 'Box Plots...'
- Kombinované grafy** points to 'Scatterplots w/Histograms...' and 'Scatterplots w/Box Plots...'
- Grafy na ověřování normality** points to 'Normal Probability Plots...', 'Quantile-Quantile Plots...', and 'Probability-Probability Plots...'
- Sloupcový graf** points to 'Bar/Column Plots...'
- Matrix plot** points to 'Matrix Plots...'
- Kategorizované grafy** points to 'Categorized Graphs...'

The background shows a data table with columns 'ALLEN' and 'PETAL' and rows of numerical values.

# Opakování

- Datový soubor BMI (list: pacienti)
- Overit normalitu
- Korelace
- Kategorizace
- Kontingenční 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  $(x_{0.25} - 1,5q, x_{0.25} - 3q)$
  - extrémní 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  $n \times n$  vzdušnou čarou
  - Odvodit asociační matici  $n \times n$  po silnici
  - Ukázat opět xy graf a komentář, že jde o značně obtížnější problém
- Horní trojúhelníkové 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

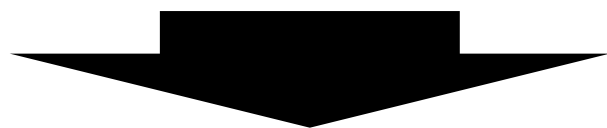
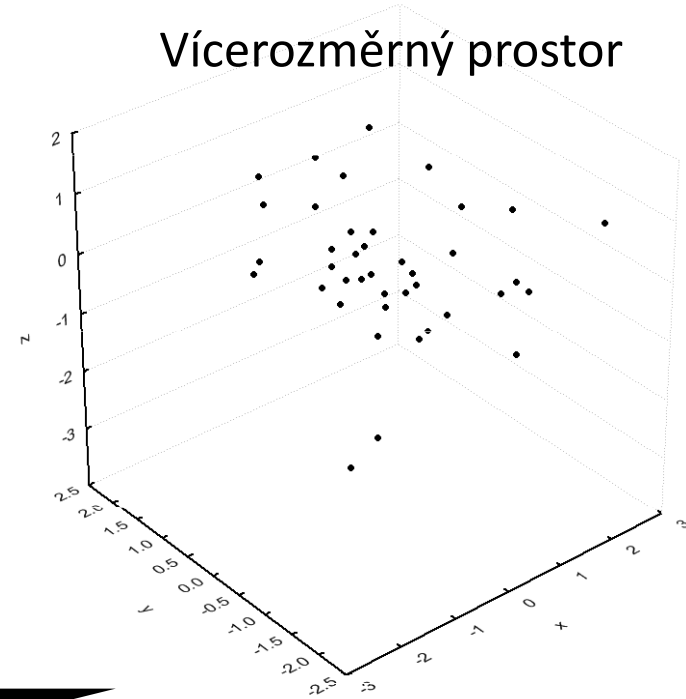
Add to Workbook Add to Report

Arial 10 **B I U**

Advertising Effectiveness Study.

	1	2	3	4	5
	ADVERT	MEASURE01	MEASURE02	MEASURE03	MEASURE04
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

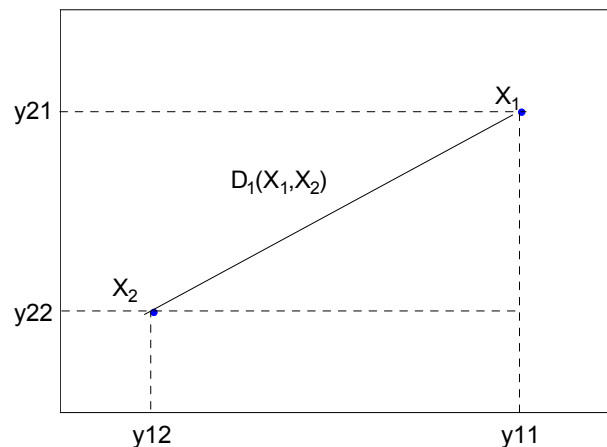
Case No.	Euclidean distances (multidimensional_normal_distribution)											
	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_10	C_11	C_12
C_1	0.00	2.58	3.73	0.95	1.46	1.85	3.78	1.85	1.59	1.80	0.46	2.82
C_2	2.58	0.00	2.17	2.58	1.90	0.82	2.32	1.99	1.14	1.57	2.26	1.73
C_3	3.73	2.17	0.00	3.15	3.54	2.19	1.66	2.49	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

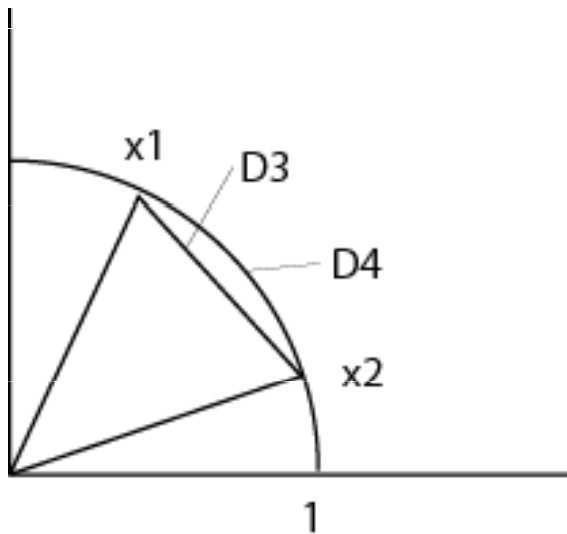
- Euklidovská vzdálenost je přepočítána na počet parametrů (druhů v případě vzdálenosti společenstev odběrů).

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

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

## 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.

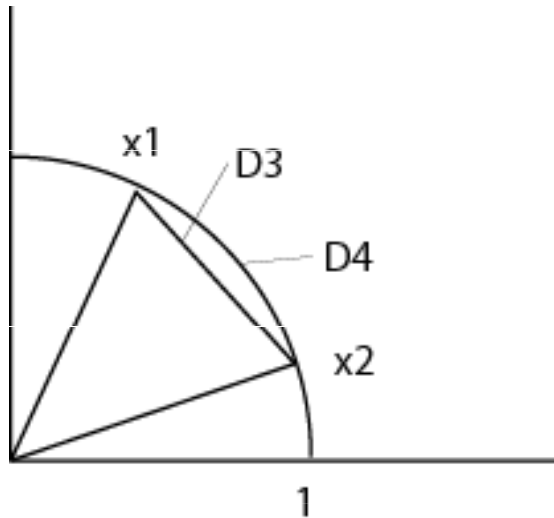


$$D_3(x_1, x_2) = \sqrt{2 \left( 1 - \frac{\sum_{j=1}^p y_{1j} y_{2j}}{\sqrt{\sum_{j=1}^p y_{1j}^2} \sqrt{\sum_{j=1}^p y_{2j}^2}} \right)}$$

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

# Geodetická metrika

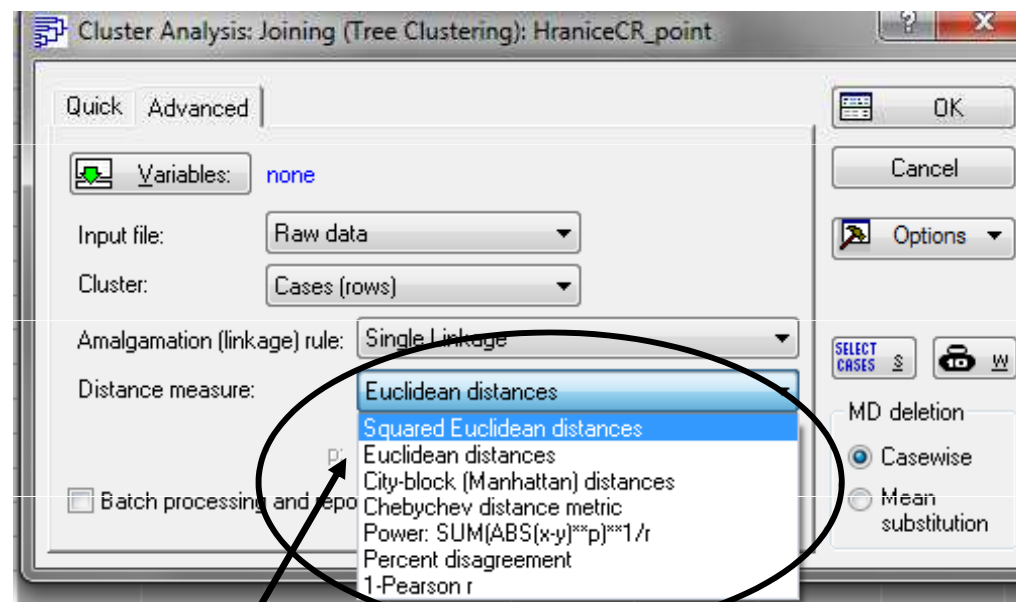
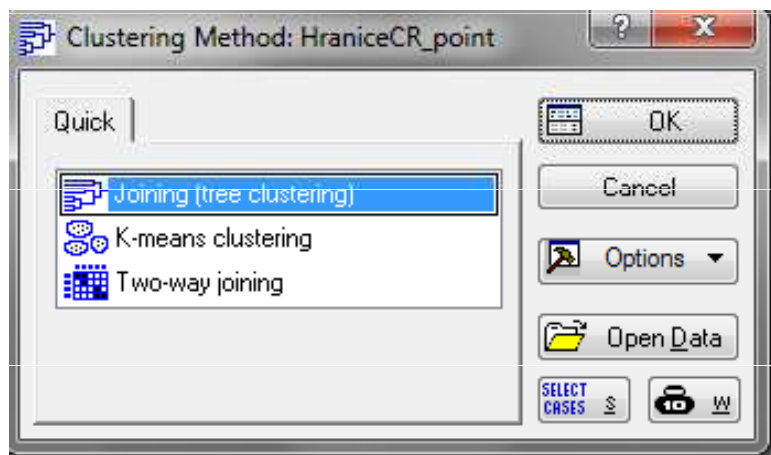
- Počítá délku výseče jednotkové kružnice mezi normalizovanými vektory (viz. Chord distance).



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

# Asociační matice 1

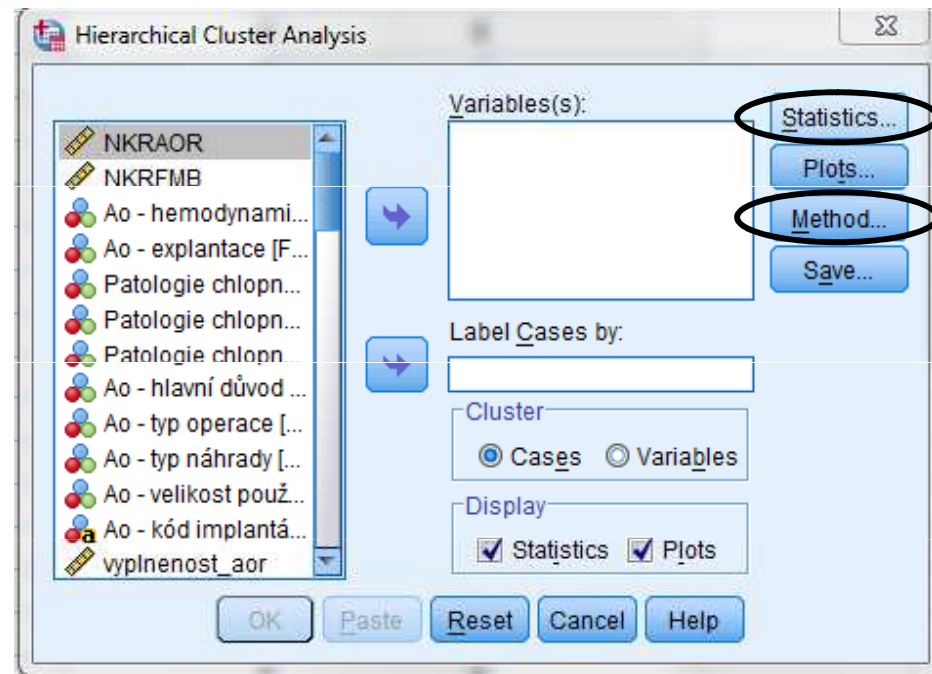
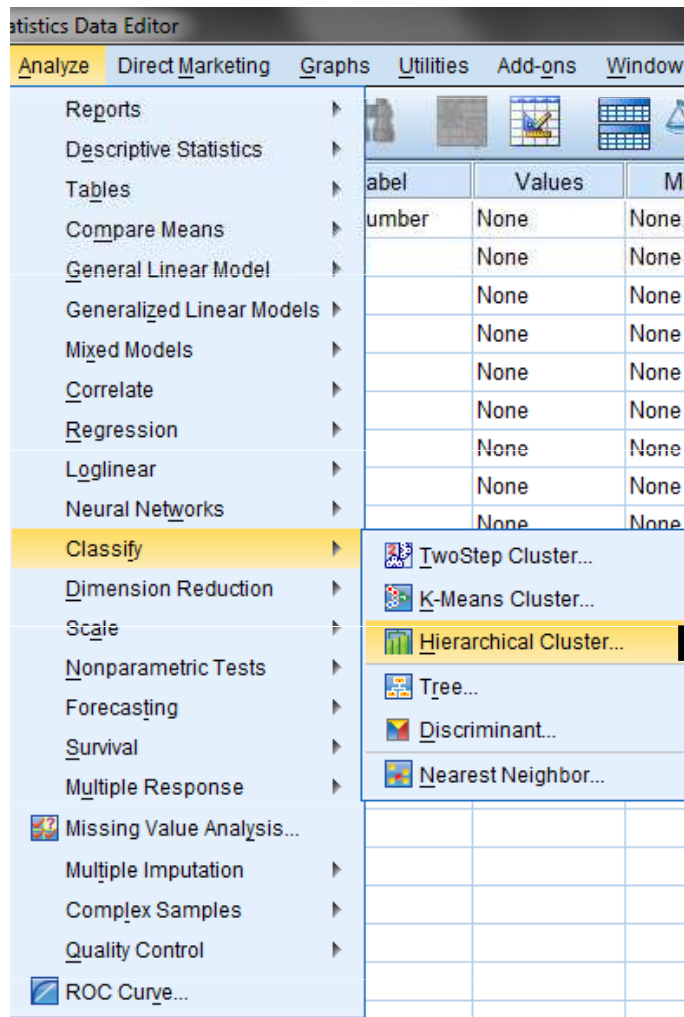
- SW Statistika



Metriky

# Asociační matice 2a

- SW SPSS



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

# Asociační matice 2b

- SW SPSS

Hierarchical Cluster Analysis

Variables(s):

Label Cases by:

Cluster:  Cases  Variables

Display:  Statistics  Plots

OK Paste Reset Cancel Help

Hierarchical Cluster Analysis: Statistics

Agglomeration schedule

Proximity matrix

Cluster solution:

None

Single solution

Number of clusters:

Range of solutions

Minimum number of clusters:

Maximum number of clusters:

Continue Cancel Help

Hierarchical Cluster Analysis: Method

Cluster Method: Between-groups linkage

Measure:

Interval: Squared Euclidean distance

Euclidean distance

Squared Euclidean distance

Counts: Cosine

Pearson correlation

Binary: Chebychev

Block

Minkowski

Customized

Transform Variables:

Standardize: None

By variable

By case:

Absolute values

Change sign

Rescale to

Continue Cancel Help

Metriky

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

# 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

CZ	Znojmo	Zlín	Ústí nad Labem	Uherské Hradiště	Teplice	Šumperk	Svitavy	Příbram	Přerov	Praha	Pízeň	Písek	Pardubice	Ostrava	Opava	Olomouc	Mladá Boleslav	Mariánské Lázně	Liberec	Klatovy	Kladno	Karlovy Vary	Jihlava	Chomutov	Čeb	Hradec Králové	Hodonín	Havlíčkův Brod	České Budějovice				
Brno	67	100	294	73	296	133	67	262	81	202	296	212	138	165	152	78	217	366	239	287	233	335	93	298	377	142	61	104	186				
České Budějovice	146	201	232	254	234	276	210	101	262	140	133	52	196	346	333	259	195	201	242	106	166	216	126	235	232	217	239	132					
Havlíčkův Brod	100	199	206	172	208	144	78	148	165	114	208	124	64	264	209	155	113	276	160	199	145	247	25	210	289	85	157						
Hodonín	108	71	347	44	349	177	128	315	91	255	349	265	191	175	169	112	270	419	300	340	286	388	146	351	430	203							
Hradec Králové	185	212	166	215	206	113	75	172	170	112	206	217	21	240	205	149	81	276	97	248	143	245	110	208	287								
Čeb	373	472	164	445	145	398	352	161	453	175	101	180	279	537	524	450	230	32	247	130	156	42	298	98									
Chomutov	294	393	66	368	47	319	273	156	374	96	102	183	200	458	445	371	132	112	149	144	74	56	219										
Jihlava	75	188	215	161	217	169	103	142	169	123	186	118	89	253	240	166	138	256	185	193	154	256											
Karlovy Vary	331	430	122	403	103	356	310	143	411	133	83	164	237	495	482	408	188	56	205	125	114												
Kladno	229	328	81	301	78	254	208	65	309	31	87	114	135	393	380	306	66	151	133	129													
Klatovy	252	382	188	355	169	359	277	102	363	136	42	75	240	447	434	360	191	99	238														
Liberec	260	309	92	332	102	208	172	162	267	107	196	207	118	335	300	246	47	266															
Mariánské Lázně	362	461	178	434	159	387	341	130	442	164	70	149	268	626	513	139	219													485	Český Tešín – PL		
Mladá Boleslav	213	312	85	285	104	102	156	115	251	55	149	160	96	319	284	230													398	287	Dolní Dvořiště – A		
Olomouc	140	63	367	68	369	65	79	335	21	275	369	285	147	93	74													308	377	159	Harrachov – PL		
Opava	214	111	369	125	388	92	131	409	78	349	443	359	210	35														298	194	253	317	Hatě – A	
Ostrava	227	104	454	131	456	127	172	422	84	382	456	372	240															79	316	248	229	354	Mikulov – A
Pardubice	164	210	196	206	198	118	73	164	168	104	198	188																					
Písek	193	307	197	280	199	268	202	49	288	105	81																						
Pízeň	292	391	146	364	127	317	271	60	372	94																							
Praha	198	297	92	270	94	223	177	60	278																								
Přerov	143	42	370	47	372	86	100	338																									
Příbram	217	357	152	330	154	283	237																										
Svitavy	134	142	269	140	271	66																											
Šumperk	195	128	315	133	317																												
Teplice	292	391	19	364																													
Uherské Hradiště	135	27	362																														
Ústí nad Labem	290	388																															
Zlín	162																																

TABULKA VZDÁLENOSTÍ MĚST ČR

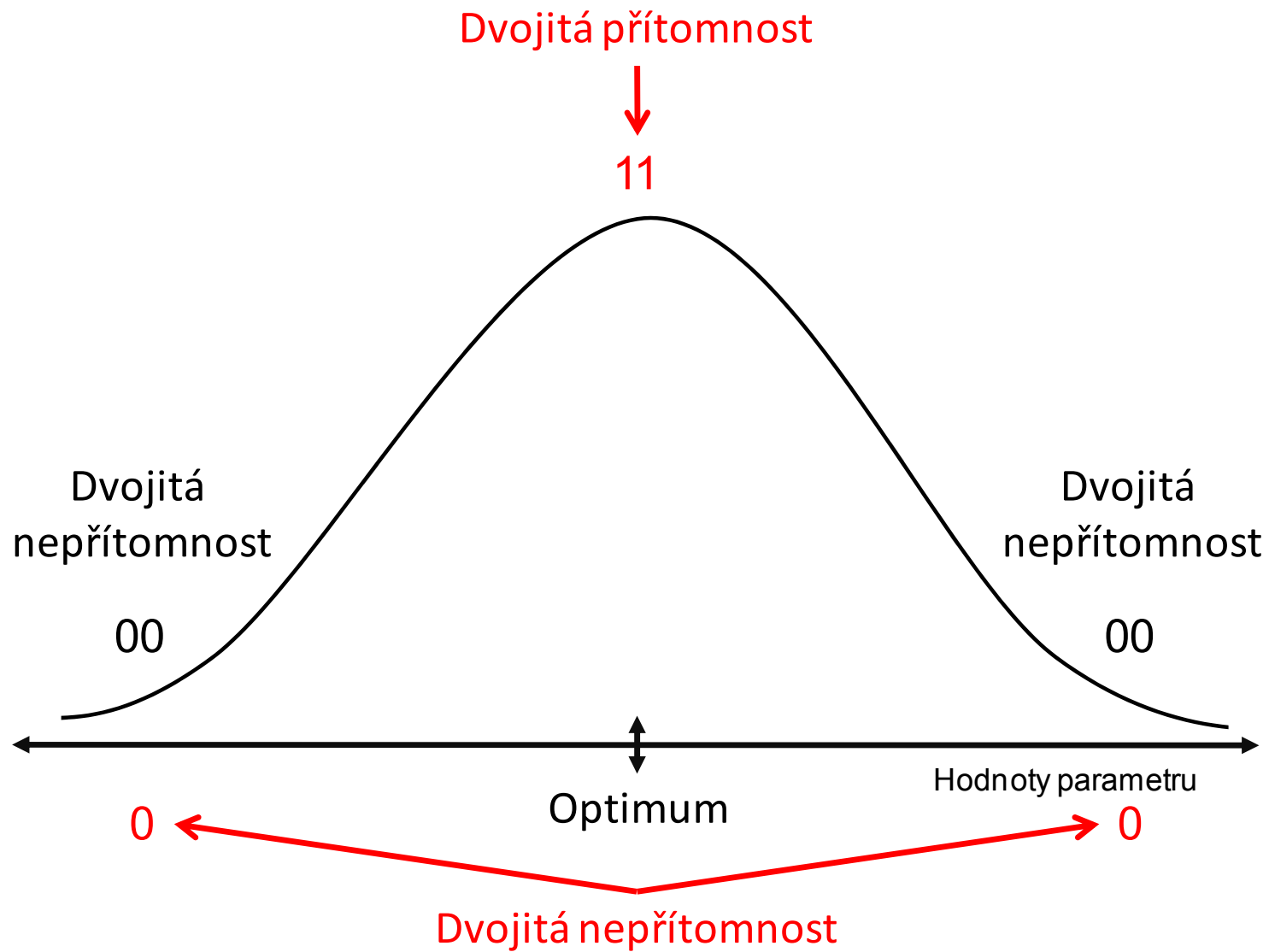
Vojtanov – D	197																																		
Rozvadov – D		122																																	
PRAMA			165																																
Pomezí nad Ohří – D				186																															
Náchod-Běloves – PL					146																														
Mikulov – A						332																													
Hatě – A							432																												
Harrachov – PL								395																											
Dolní Dvořiště – A									315																										
Český Tešín – PL										298																									
Činovec – D											186																								

Hraniční přechody  
Grenzübergang  
Border de frontière  
Passagierfrontiera



# Binární koeficienty

# Doble-zero problem



# Binární koeficienty

		Společenstvo 1		
		1	0	
Společenstvo 2	1	a	b	$a + b$
	0	c	d	$c + d$
		$a + c$	$b + d$	

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

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