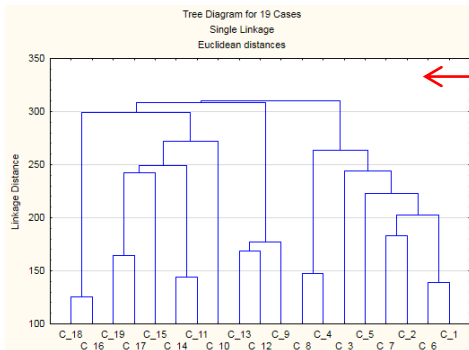
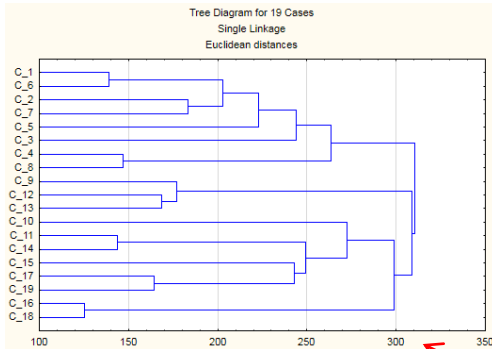


# Výpočet shlukové analýzy v softwarech

# STATISTICA – hierarchické aglomerativní shlukování

- Statistics – Multivariate Exploratory Techniques – Cluster Analysis – Joining (tree clustering) – OK
- Variables: výběr proměnných (např. objem hipokampu, amygdaly a pallida)
- Cluster: zvolit, zda chceme shlukovat proměnné (Variables (columns)) či subjekty (Cases (rows))
- Amalgamation (linkage) rule = volba shlukovacího algoritmu:
  - Single Linkage – metoda nejbližšího souseda
  - Complete Linkage – metoda nejvzdálenějšího souseda
  - Unweighted pair-group average – metoda průměrné vazby (nevážená)
  - Weighted pair-group average – metoda průměrné vazby (vážená)
  - Unweighted pair-group centroid – centroidová metoda (nevážená)
  - Weighted pair-group centroid (median) – centroidová metoda (vážená) = mediánová metoda
  - Ward's method – Wardova metoda
- Distance measure = volba metrik vzdáleností objektů (subjektů):
  - Squared Euclidean distances – čtverec Euklidovy vzdálenosti
  - Euclidean distances – Euklidova metrika
  - City-block (Manhattan) distances – Hammingova (manhattanská) metrika
  - Chebychev distance metric – Čebyševova metrika
  - Power:  $\text{SUM}(\text{ABS}(x-y)**p)**1/r$  – pokud  $r=p$ , jde o Minkovského metriku
  - Percent disagreement
  - 1-Pearson r – jedna mínus Pearsonův korelační koeficient

# STATISTICA – hierarch. aglom. shluk. – pokračování



Joining Results: Data\_neuro\_shlukovky

Number of variables: 3  
 Number of cases: 19  
 Joining of cases  
 Missing data were casewise deleted  
 Amalgamation (joining) rule: Single Linkage  
 Distance metric is: Euclidean distances (non-standardized)

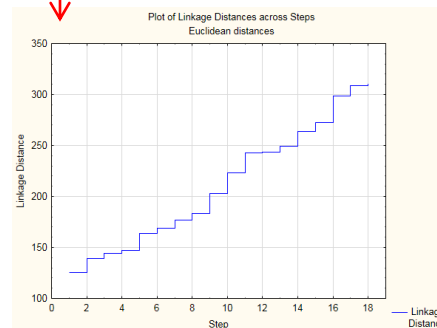
Quick | Advanced

- Horizontal hierarchical tree plot
- Vertical icicle plot
- Rectangular branches
- Scale tree to dlink/dmax\*100
- Amalgamation schedule
- Graph of amalgamation schedule
- Distance matrix
- Descriptive statistics
- Matrix
- Save classifications
- Sort by cluster membership

Summary | Cancel | Options | By Group

Amalgamation Schedule (Data\_neuro\_shlukovky)  
 Single Linkage  
 Euclidean distances

linkage distance	Obj. No. 1	Obj. No. 2	Obj. No. 3	Obj. No. 4	Obj. No. 5	O
125.1972	C_16	C_18				
139.2640	C_1	C_6				
143.9270	C_11	C_14				
147.0873	C_4	C_8				
164.1363	C_17	C_19				
168.6528	C_12	C_13				
176.9954	C_9	C_12	C_13			
183.2707	C_2	C_7				
202.7584	C_1	C_6	C_2	C_7		
223.0460	C_1	C_6	C_2	C_7	C_5	
249.7229	C_1	C_6	C_2	C_7	C_5	C_9



asociační matice Euklidových vzdáleností

Euclidean distances (Data\_neuro\_shlukovky)

Case No.	C_1	C_2	C_3	C_4	C_5	C_6	C
C_1	0	291	299	490	271	139	
C_2	291	0	244	264	454	251	
C_3	299	244	0	500	527	311	
C_4	490	264	500	0	535	410	
C_5	271	454	527	535	0	223	
C_6	139	251	311	410	223	0	
C_7	307	183	262	328	399	203	
C_9	574	297	619	447	554	479	

# STATISTICA – nehierarchické shlukování

- Statistics – Multivariate Exploratory Techniques – Cluster Analysis – K-means clustering – OK – přepnout se na záložku Advanced
- Variables: výběr proměnných (např. objem hipokampu, amygdaly a pallida)
- Cluster: zvolit, zda chceme shlukovat proměnné (Variables (columns)) či subjekty (Cases (rows))
- Number of clusters: zvolit počet shluků (např. 3)
- Number of iterations: volba počtu iterací (metoda  $k$ -průměrů je iterativní metoda)
- Initial cluster centers: volba počátečních středů shluků
  
- příslušnost jednotlivých subjektů do shluků nalezneme na záložce Advanced v „Members of each cluster & distances“

# SPSS – hierarchické aglomerativní shlukování

- Analyze – Classify – Hierarchical Cluster...
- Cluster: zvolit, zda chceme shlukovat proměnné (Variables) či subjekty (Cases)
- Statistics...: zatrhnout Proximity matrix (= asociační matice vzdáleností či podobností)
- Plots...: zatrhnout Dendrogram (možnost volby Vertical či Horizontal)
- Method...:
  - Cluster Method = volba shlukovacího algoritmu:
    - Between-groups linkage – metoda průměrné vazby mezi skupinami
    - Within-groups linkage – metoda průměrné vazby uvnitř skupin
    - Nearest neighbor – metoda nejbližšího souseda
    - Furthest neighbor – metoda nejvzdálenějšího souseda
    - Centroid clustering – centroidová metoda (nevážená)
    - Median clustering – centroidová metoda (vážená) = mediánová metoda
    - Ward's method – Wardova metoda
  - Distance measure: volba metrik vzdáleností objektů (subjektů):
    - Euclidean distance – Euklidova metrika
    - Squared Euclidean distance – čtverec Euklidovy vzdálenosti
    - Cosine – kosinová metrika
    - Pearson correlation – Pearsonův korelační koeficient
    - Chebychev – Čebyševova metrika
    - Block – Hammingova (manhattanská) metrika
    - Minkowski – Minkovského metrika
    - Customized – výpočet pomocí  $\text{SUM}(\text{ABS}(x-y)**p)**1/r$
  - Transform Values, Transform Measure – je možno transformovat původní data nebo vypočtené vzdálenosti

# SPSS – nehierarchické shlukování

- Analyze – Classify – K-Means Cluster...
- Variables: výběr proměnných (např. objem hipokampu, amygdaly a pallida)
- Number of clusters: zvolit počet shluků (např. 3)
- Method: přepnout na „Classify only“ v případě, že známe středy shluků, které můžeme načíst pomocí „Read initial“
- Iterate... – Maximum Iterations (volba počtu iterací – metoda  $k$ -průměrů je iterativní metoda)
- Options... – zatrhnout „Cluster information for each case“, abychom získali tabulku, do kterého shluku patří který subjekt

# Software R – hierarchické aglomerativní shlukování

- funkce *dist* na výpočet vzdáleností objektů (či subjektů) :
  - „euclidean“ – Euklidovská metrika
  - „maximum“ – Čebyševova metrika
  - „manhattan“ – Hammingova (manhattanská) metrika
  - „canberra“ – Canberrská metrika
  - „minkowski“ – Minkovského metrika
- funkce *hclust* na výpočet shlukové analýzy:
  - „ward.D“ a „ward.D2“ – dva algoritmy pro Wardovu metodu
  - „single“ – metoda nejbližšího souseda (single linkage)
  - „complete“ – metoda nejvzdálenějšího souseda (complete linkage)
  - „average“ – metoda průměrné vazby (nevážená) (average linkage)
  - „mcquitty“ – metoda průměrné vazby (vážená)
  - „median“ – centroidová metoda (vážená) = mediánová metoda
  - „centroid“ – centroidová metoda (nevážená)
- podrobná ukázka v souboru Shlukovky\_skript.R

# Software R – nehierarchické shlukování

- funkce *kmeans*
- ukázka:

```
cl <- kmeans(data.vyber, 3) # provedeni shlukove analyzy  
table(cl$cluster,groupCodes) # zjisteni, kolik subjektu bylo spatne zarazenych
```



# Matlab – hierarchické aglomerativní shlukování

- funkce *linkage*, která umožňuje volbu shlukovacího algoritmu i volbu metriky vzdálenosti mezi objekty (subjekty)
- volba shlukovacího algoritmu:
  - „average“ – metoda průměrné vazby (nevážená) (average linkage)
  - „centroid“ – centroidová metoda (nevážená)
  - „complete“ – metoda nejvzdálenějšího souseda (complete linkage)
  - „median“ – centroidová metoda (vážená) = mediánová metoda
  - „single“ – metoda nejbližšího souseda (single linkage)
  - „ward“ – Wardova metoda
  - „weighted“ – metoda průměrné vazby (vážená)
- volba metriky vzdáleností – stejná nabídka jako u funkce *pdist*
- ukázka:

```
[num, txt] = xlsread('Data_neuro_shlukovky.xlsx',1);  
data=num(:,[23,24,26]);
```

```
Z=linkage(data,'complete','euclidean'); % provedeni shlukove analyzy  
dendrogram(Z) % vykresleni dendrogramu
```

```
c=cluster(Z,'maxclust',3); % vytvoreni definovaneho poctu shluku  
crosstab(c,num(:,3)) % zjistení, kolik subjektu bylo spatne zarazenych
```

# Matlab – nehierarchické shlukování

- funkce *kmeans*
- ukázka:

```
[idx,C]=kmeans(data,3); % provedeni shlukove analyzy (matice C – centroidy skupin)  
crosstab(idx,num(:,3)) % zjisteni, kolik subjektu bylo spatne zarazenych
```

- funkce *kmedoids*
- bohužel není ve starých verzích Matlabu
- ukázka:

```
[idx,C]=kmedoids(data,3); % provedeni shlukove analyzy (matice C – medoidy skupin)  
crosstab(idx,num(:,3)) % zjisteni, kolik subjektu bylo spatne zarazenych
```