

用SPSS作聚类分析

一、聚类分析（Cluster Analysis）简介

聚类分析是直接比较各事物之间的性质，将性质相近的归为一类，将性质差别较大的归入不同的类的分析技术。

常言道：“物以类聚”，对事物分门别类进行研究，有利于我们做出正确的判断。日常生活中，我们不自觉地用定性方法将人分为“好人”、“坏人”；按熟悉程度分为“朋友”、“熟人”、“陌生人”等等。

数理统计中的数值分类有两种问题：

- 判别分析：已知分类情况，将未知个体归入正确类别
- 聚类分析：分类情况未知，对数据结构进行分类

通过分类，有利于我们抓住重点，从总体上去把握事物，找出解决问题的方法。例如将股票进行分类，可以为我们投资提供参考。

二、聚类对象

要做聚类分析，首先得按照我们聚类的目的，从对象中**提取**出能表现这个目的**特征指标**；然后根据亲疏程度进行分类。

聚类分析根据分类对象的不同可分为**Q型**和**R型**两大类

Q型是对样本进行分类处理，其作用在于：

1. 能利用多个变量对样本进行分类
2. 分类结果直观，聚类谱系图能明确、清楚地表达其数值分类结果
3. 所得结果比传统的定性分类方法更细致、全面、合理

R型是对变量进行分类处理，其作用在于：

1. 可以了解变量间及变量组合间的亲疏关系
2. 可以根据变量的聚类结果及它们之间的关系，选择主要变量进行回归分析或**Q型**聚类分析

三、聚类过程与方法

聚类的主要过程一般可分为如下四个步骤：

1. 数据预处理（标准化）
2. 构造关系矩阵（亲疏关系的描述）
3. 聚类（根据不同方法进行分类）
4. 确定最佳分类（类别数）

下面我们结合实际例子分步进行讨论。

例、下表给出了1982年全国28个省、市、自治区农民家庭收支情况，有六个指标，是利用调查资料进行聚类分析，为经济发展决策提供依据。

（详见文件1982“农民生活消费聚类.sav”）

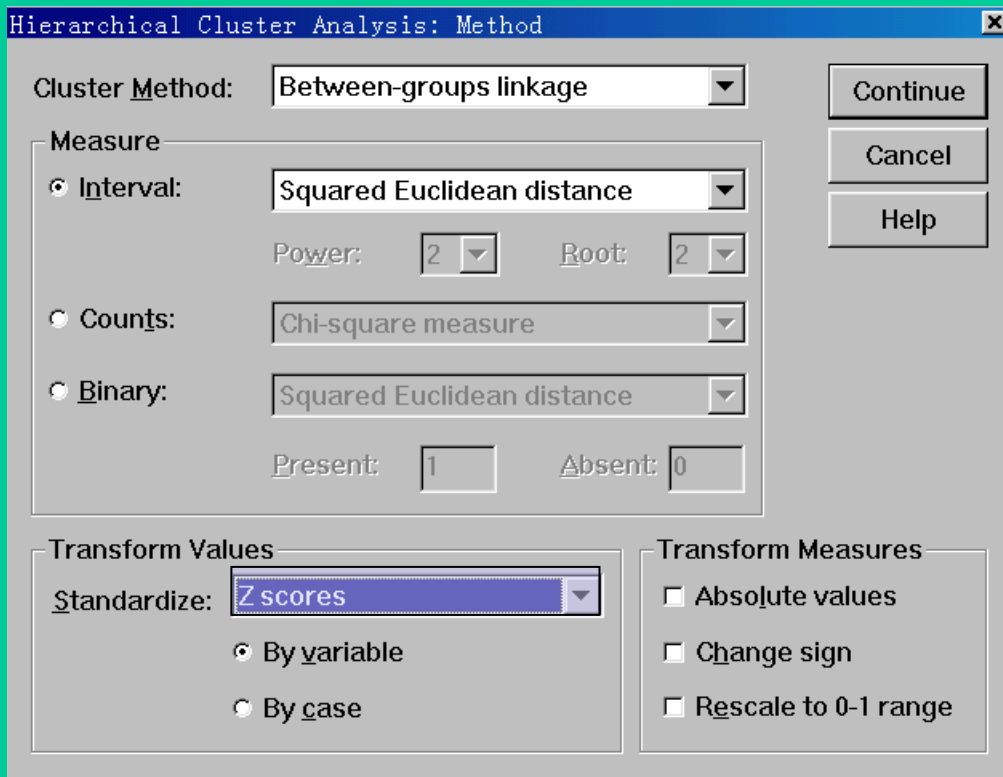
1. 数据预处理（标准化）

1) 为什么要做数据变换

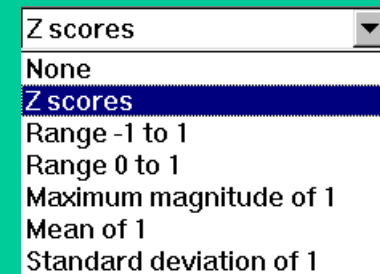
→ 指标变量的量纲不同或数量级相差很大，为了使这些数据能放到一起加以比较，常需做变换。

2) 在SPSS中如何选择标准化方法：

→ **Analyze** → **Classify** → **Hierachical Cluster Analysis**
→ **Method** 然后从对话框中进行如下选择



从**Transform Values**框中点击向下箭头，将出现如下可选项，从中选一即可：



3) 常用标准化方法（选项说明）：

a) None: 不进行标准化，这是系统默认值

为了便于后面的说明，作如下假设：

所有样本表示为

$$X = \begin{bmatrix} x_{11} & \cdots & x_{1m} \\ \vdots & \vdots & \vdots \\ x_{n1} & \cdots & x_{nm} \end{bmatrix}$$

均值表示为

$$\bar{x}_j = \frac{1}{n} \sum_{i=1}^n x_{ij}$$

标准差表示为

$$S_j = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_{ij} - \bar{x}_j)^2}$$

极差表示为

$$R_j = \max_{1 \leq i \leq n} x_{ij} - \min_{1 \leq i \leq n} x_{ij}$$

b) Z Scores: 标准化变换

$$x_{ij}^* = \begin{cases} \frac{x_{ij} - \bar{x}_j}{S_j} & \text{若 } S_j \neq 0 \\ 0 & \text{若 } S_j = 0 \end{cases} \quad \begin{pmatrix} i=1,2,\dots,n \\ j=1,2,\dots,m \end{pmatrix}$$

作用：变换后的数据均值为0，标准差为1，消去了量纲的影响；当抽样样本改变时，它仍能保持相对稳定性。这是最常用的方法。

c) Range -1 to 1: 极差标准化变换

$$x_{ij}^* = \begin{cases} \frac{x_{ij} - \bar{x}_j}{R_j} & \text{若 } R_j \neq 0 \\ x_{ij} & \text{若 } R_j = 0 \end{cases} \quad \begin{pmatrix} i=1,2,\dots,n \\ j=1,2,\dots,m \end{pmatrix}$$

作用：变换后的数据均值为0，极差为1，且 $|x_{ij}^*| < 1$ ，消去了量纲的影响；在以后的分析计算中可以减少误差的产生。该方法要求变量值中须含有负数。

d) Maximum magnitude of 1

$$x_{ij}^* = \begin{cases} \frac{x_{ij}}{\max_{1 \leq i \leq n} x_{ij}} & \text{若 } \max_{1 \leq i \leq n} x_{ij} \neq 0 \\ \frac{x_{ij}}{\left| \min_{1 \leq i \leq n} x_{ij} \right|} + 1 & \text{若 } \max_{1 \leq i \leq n} x_{ij} = 0 \end{cases} \quad \begin{pmatrix} i=1,2,\dots,n \\ j=1,2,\dots,m \end{pmatrix}$$

作用：变换后的数据最大值为1。

e) Range 0 to 1（极差正规化变换 / 规格化变换）

$$x_{ij}^* = \begin{cases} \frac{x_{ij} - \min_{1 \leq i \leq n} x_{ij}}{R_j} & \text{若 } R_j \neq 0 \\ 0.5 & \text{若 } R_j = 0 \end{cases} \quad \begin{pmatrix} i=1,2,\dots,n \\ j=1,2,\dots,m \end{pmatrix}$$

作用：变换后的数据最小为0，最大为1，其余在区间[0, 1]内，极差为1，无量纲。

f) Mean of 1

$$x_{ij}^* = \begin{cases} \frac{x_{ij}}{\bar{x}_j} & \text{若 } \bar{x}_j \neq 0 \\ x_{ij} + 1 & \text{若 } \bar{x}_j = 0 \end{cases} \quad \begin{pmatrix} i=1,2,\dots,n \\ j=1,2,\dots,m \end{pmatrix}$$

作用：变换后的数据均值为1。

g) Standard deviation of 1

$$x_{ij}^* = \begin{cases} \frac{x_{ij}}{S_j} & \text{若 } S_j \neq 0 \\ x_{ij} & \text{若 } S_j = 0 \end{cases} \quad \begin{pmatrix} i=1,2,\dots,n \\ j=1,2,\dots,m \end{pmatrix}$$

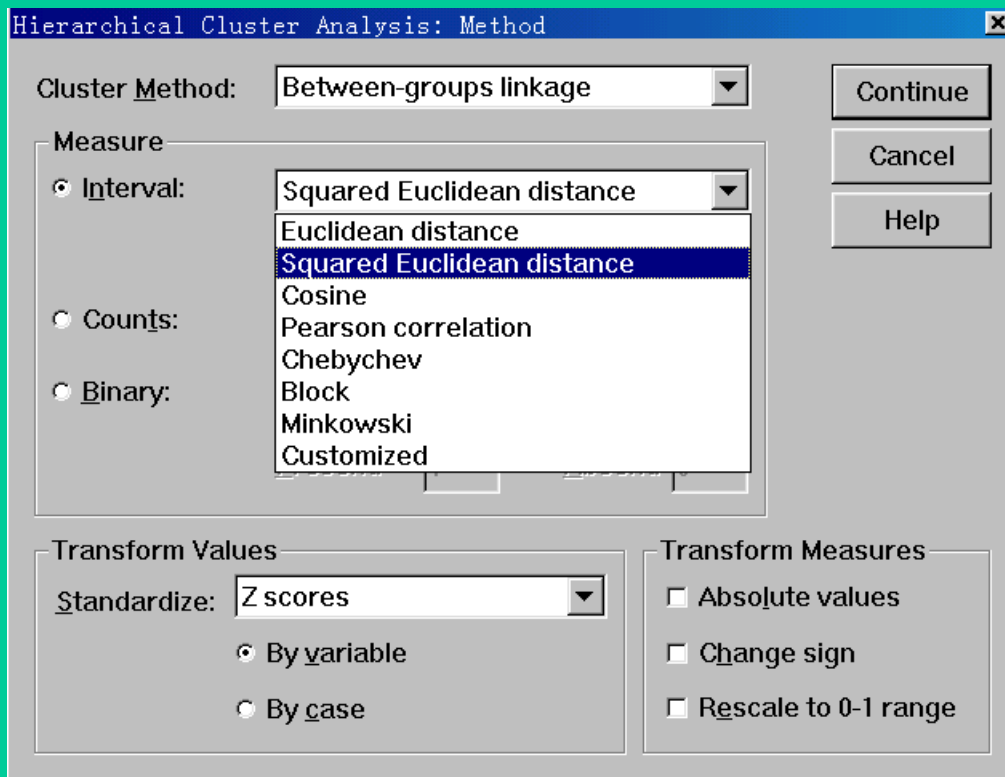
作用：变换后的数据标准差为1。

2. 构造关系矩阵

- 1) 描述变量或样本的亲疏程度的数量指标有两种：
 - 相似系数——性质越接近的样品，相似系数越接近于1或-1；彼此无关的样品相似系数则接近于0，聚类时相似的样品聚为一类
 - 距离——将每一个样品看作m维空间的一个点，在这m维空间中定义距离，距离较近的点归为一类。
 - ❖ 相似系数与距离有40多种，但常用的只是少数

2) 在SPSS中如何选择测度：

→ **Analyze** → **Classify** → **Hierarchical Cluster Analysis**
→ **Method** 然后从对话框中进行如下选择



从Measure框中点击Interval项的向下箭头，将出现如左可选项，从中选一即可。

3) 常用测度（选项说明）：

a) Euclidean distance: 欧氏距离
（二阶Minkowski距离）

$$d(x, y) = \sqrt{\sum_i (x_i - y_i)^2}$$

用途： 聚类分析中用得最广泛的距离

但与各变量的量纲有关，未考虑指标间的相关性，也未考虑各变量方差的不同

b) Squared Euclidean distance: 平方欧氏距离

$$d(x, y) = \sum_i (x_i - y_i)^2$$

用途： 聚类分析中用得最广泛的距离

c) Cosine: 夹角余弦(相似性测度)

$$\cos(x, y) = \frac{\sum_i x_i y_i}{\sqrt{\sum_i x_i^2 \cdot \sum_i y_i^2}}$$

用途: 计算两个向量在 origin 处的夹角余弦。当两夹角为 0° 时, 取值为1, 说明极相似; 当夹角为 90° 时, 取值为0, 说明两者不相关。取值范围: 0~1

d) Pearson correlation: 皮尔逊相关系数

$$r_{xy} = \frac{n \sum XY - \sum X \sum Y}{\sqrt{\left[N \sum X^2 - (\sum X)^2 \right] \left[N \sum Y^2 - (\sum Y)^2 \right]}}$$

用途: 计算两个向量的皮尔逊相关系数

e) Chebychev: 切比雪夫距离

$$d_{\infty}(x, y) = \max_i |x_i - y_i|$$

用途: 计算两个向量的切比雪夫距离

f) Block: 绝对值距离（一阶Minkowski度量）
（又称Manhattan度量或网格度量、马氏距离、广义欧氏距离）

$$d_1(x, y) = \sum_i |x_i - y_i|$$

用途: 计算两个向量的绝对值距离

g) Minkowski: 明科夫斯基距离

$$d_q(x, y) = \left[\sum_i |x_i - y_i|^q \right]^{1/q}$$

用途: 计算两个向量的明科夫斯基距离

h) Customized: 自定义距离

$$d_q(x, y) = \left[\sum_i |x_i - y_i|^q \right]^{1/r}$$

用途: 计算两个向量的自定义距离

3. 选择聚类方法

确定了样品或变量间的距离或相似系数后，就要对样品或变量进行分类。分类的一种方法是系统聚类法（又称谱系聚类）；另一种方法是调优法（如动态聚类法就属于这种类型）。此外还有模糊聚类、图论聚类、聚类预报等多种方法。

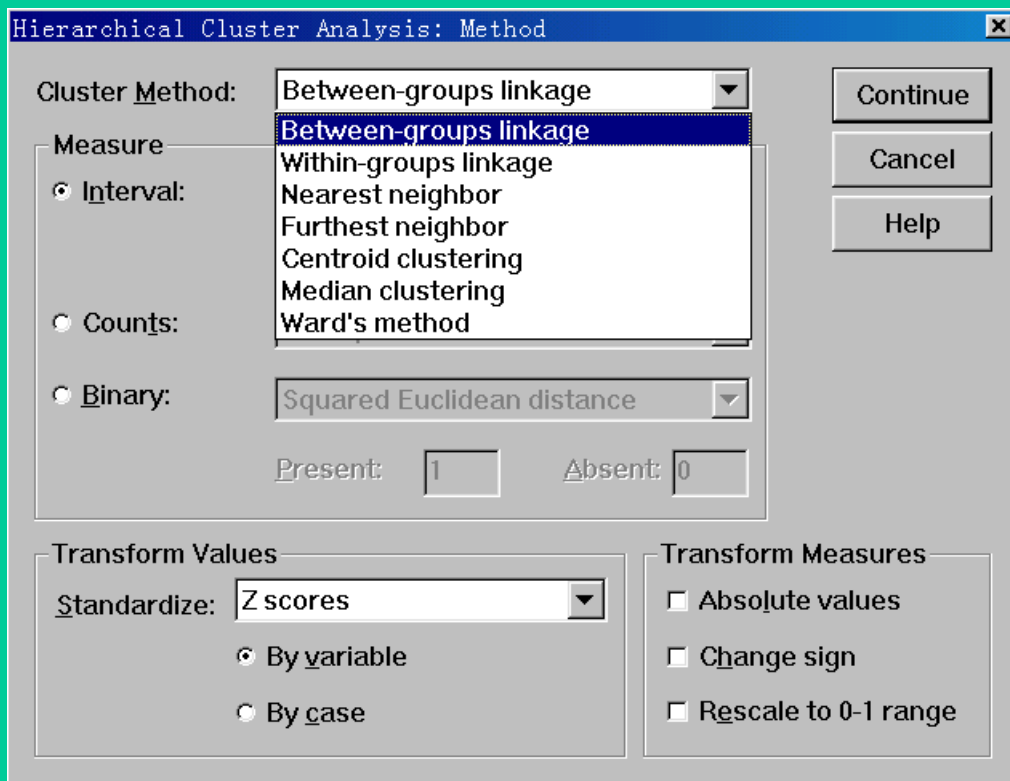
我们主要介绍系统聚类法(实际应用中使用最多)。

系统聚类法的基本思想：令 n 个样品自成一类，计算出相似性测度，此时类间距离与样品间距离是等价的，把测度最小的两个类合并；然后按照某种聚类方法计算类间的距离，再按最小距离准则并类；这样每次减少一类，持续下去直到所有样品都归为一类为止。聚类过程可做成聚类谱系图(Hierarchical diagram)。

1) 系统聚类法的产生

系统聚类法的聚类原则决定于样品间的距离（或相似系数）及类间距离的定义，类间距离的不同定义就产生了不同的系统聚类分析方法。

2) SPSS中如何选择系统聚类法



从Cluster Method框中点击向下箭头，将出现如左可选项，从中选一即可。

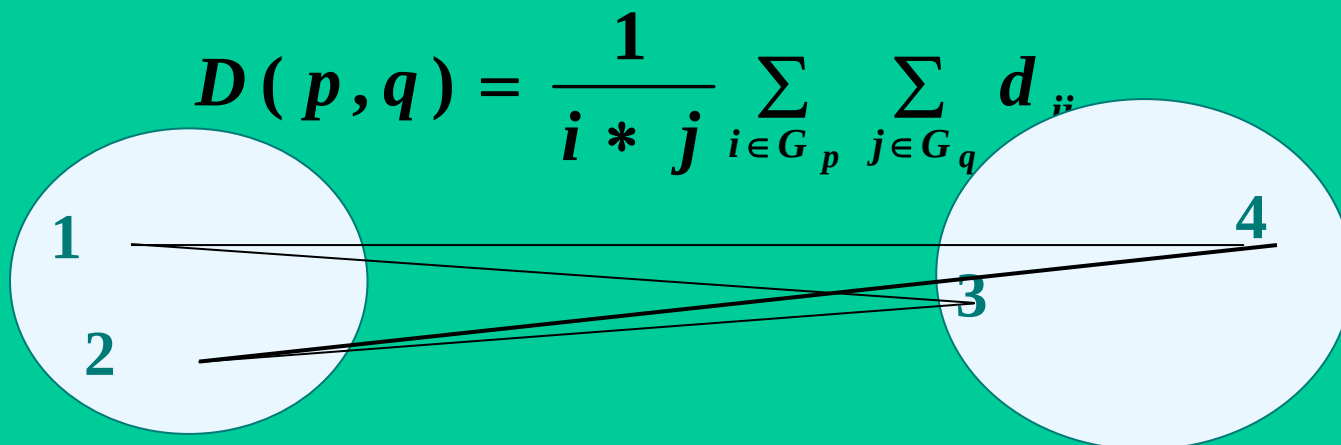
3) 常用系统聚类方法

用 $D(p,q)$ 表示类 p 和类 q 之间的距离

a) Between-groups linkage 组间平均距离连接法

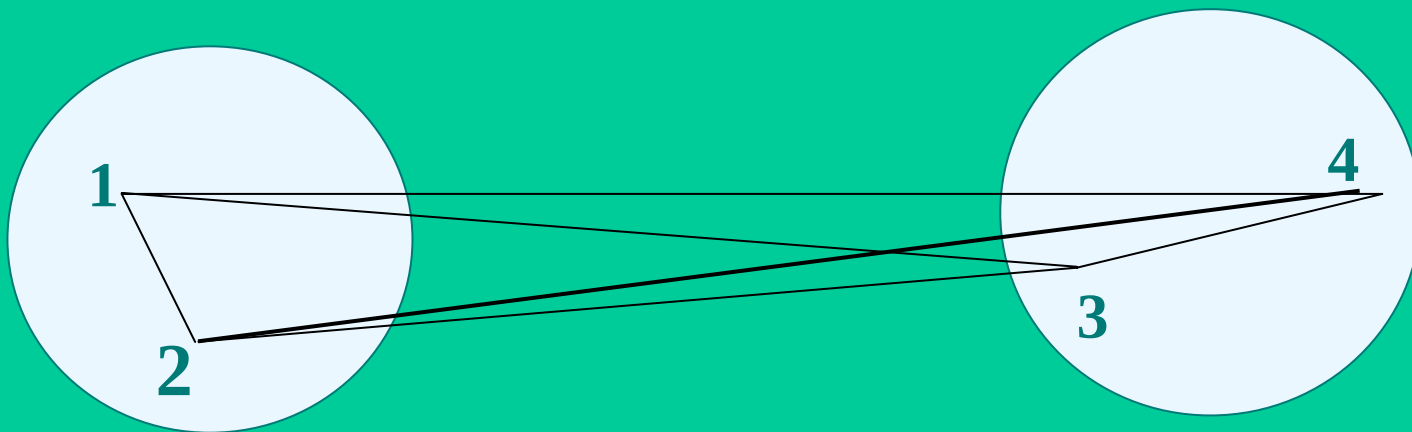
方法简述：将两个类所有的样本对（样本对的两个成员分属于不同的类）的平均距离作为两类的距离，合并距离最近或相关系数最大的两类。此方法利用了两个类中所有的样本信息。

特点：非最大距离，也非最小距离



b) Within-groups linkage 组内平均连接法

方法简述：两类合并为一类后，合并后的类中所有项之间的平均距离最小,包括两个类之间的样本对以及两个类内的样本对



c) Nearest neighbor 最近邻法（最短距离法）

方法简述：用两类中所有样本对的距离的最小值作为两类的距离，合并最近或最相似的两项

特点：样品有链接聚合的趋势，这是其缺点，不适合一般数据的分类处理，除去特殊数据外，不提倡用这种方法。

d) Furthest neighbor 最远邻法（最长距离法）

方法简述：用两类之间最远点的距离代表两类之间的距离，也称之为完全连接法

e) Centroid clustering 重心聚类法

方法简述：两类间的距离定义为两类重心之间的距离，对样品分类而言，每一类中心就是属于该类样品的均值

特点：该距离随聚类地进行不断缩小。该法的谱系树状图很难跟踪，且符号改变频繁，计算较烦。

f) Median clustering 中位数法

方法简述：两类间的距离既不采用两类间的最近距离，也不采用最远距离，而采用介于两者间的距离

特点：图形将出现递转，谱系树状图很难跟踪，因而这个方法几乎不被人们采用。

g) Ward's method 离差平方和法

方法简述：基于方差分析思想，如果分类合理，则同类样品间离差平方和应当较小，类与类间离差平方和应当较大

特点：实际应用中分类效果较好，应用较广；要求样品间的距离必须是欧氏距离。

四、谱系分类的确定

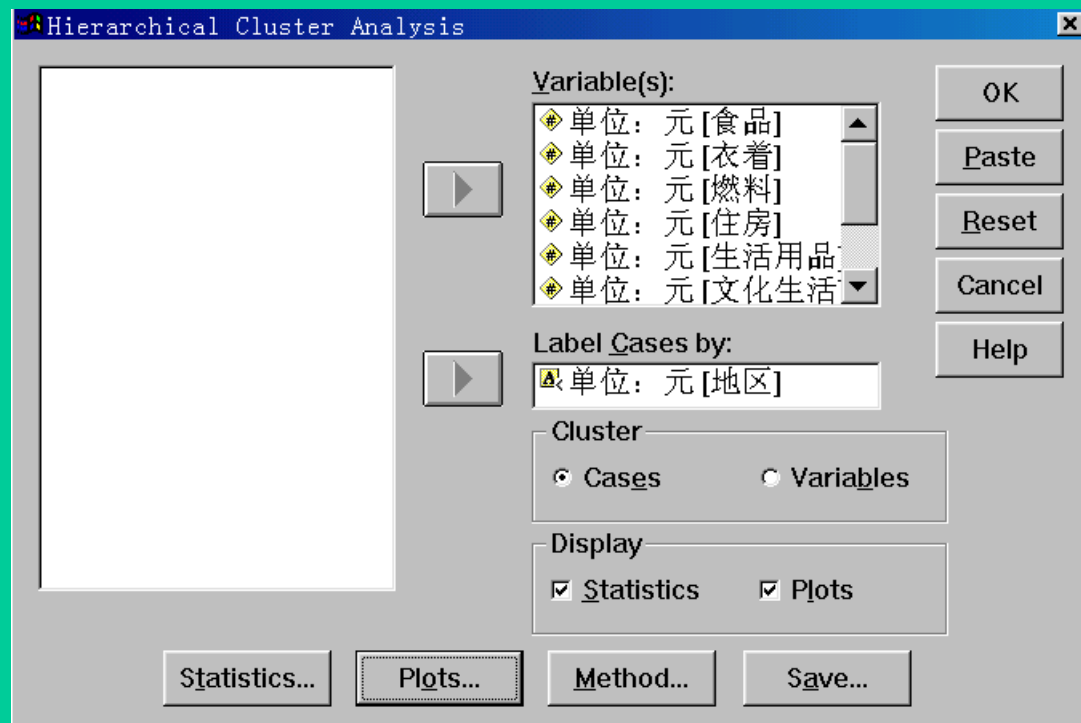
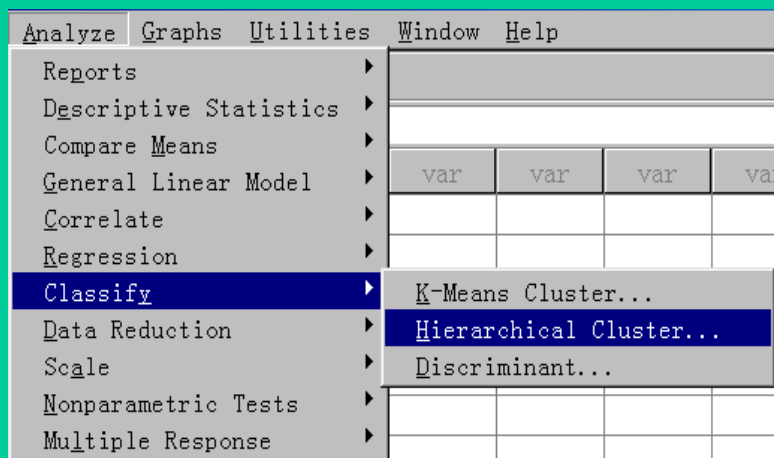
经过系统聚类法处理后，得到聚类树状谱系图，**Demirmen(1972)**提出了应根据研究的目的来确定适当的分类方法，并提出了一些根据谱系图来分类的准则：

- A.** 任何类都必须在临近各类中是突出的，即各类重心间距离必须极大
- B.** 确定的类中，各类所包含的元素都不要过分地多
- C.** 分类的数目必须符合实用目的
- D.** 若采用几种不同的聚类方法处理，则在各自的聚类图中应发现相同的类

SPSS中其他选项（通过实例演示）

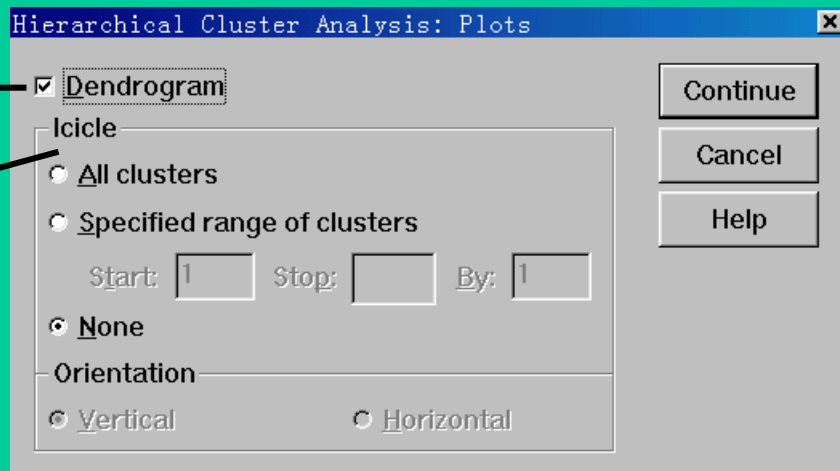
例、下表给出了1982年全国28个省、市、自治区农民家庭收支情况，有六个指标，是利用调查资料进行聚类分析，为经济发展决策提供依据。

（详见文件1982“农民生活消费聚类.sav”）



生成树形图

生成冰柱图

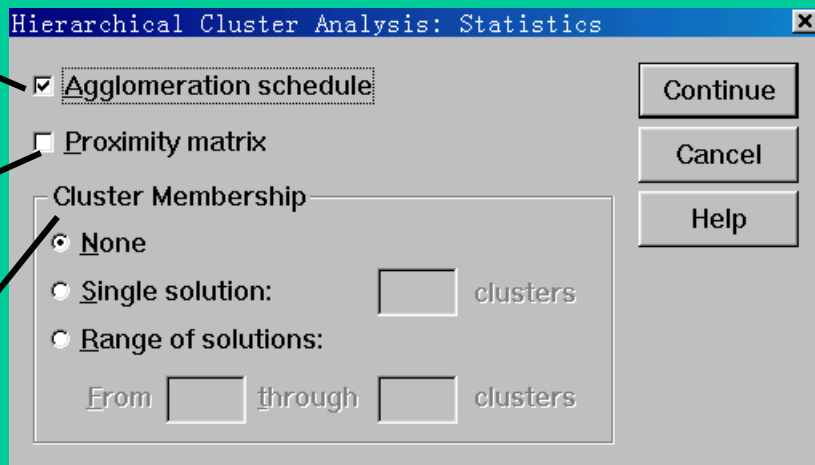


Plots...

凝聚状态表，显示聚类过程

各项间的距离矩阵

类成员栏



Statistics...

结果分析：（方法选择如下）

Hierarchical Cluster Analysis: Method

Cluster Method: Ward's method

Measure

☒ Interval: Euclidean distance

Power: 2 Root: 2

☐ Counts: Chi-square measure

☐ Binary: Squared Euclidean distance

Present: 1 Absent: 0

Transform Values

Standardize: Z scores

☒ By variable

☐ By case

Transform Measure

☐ Absolute values

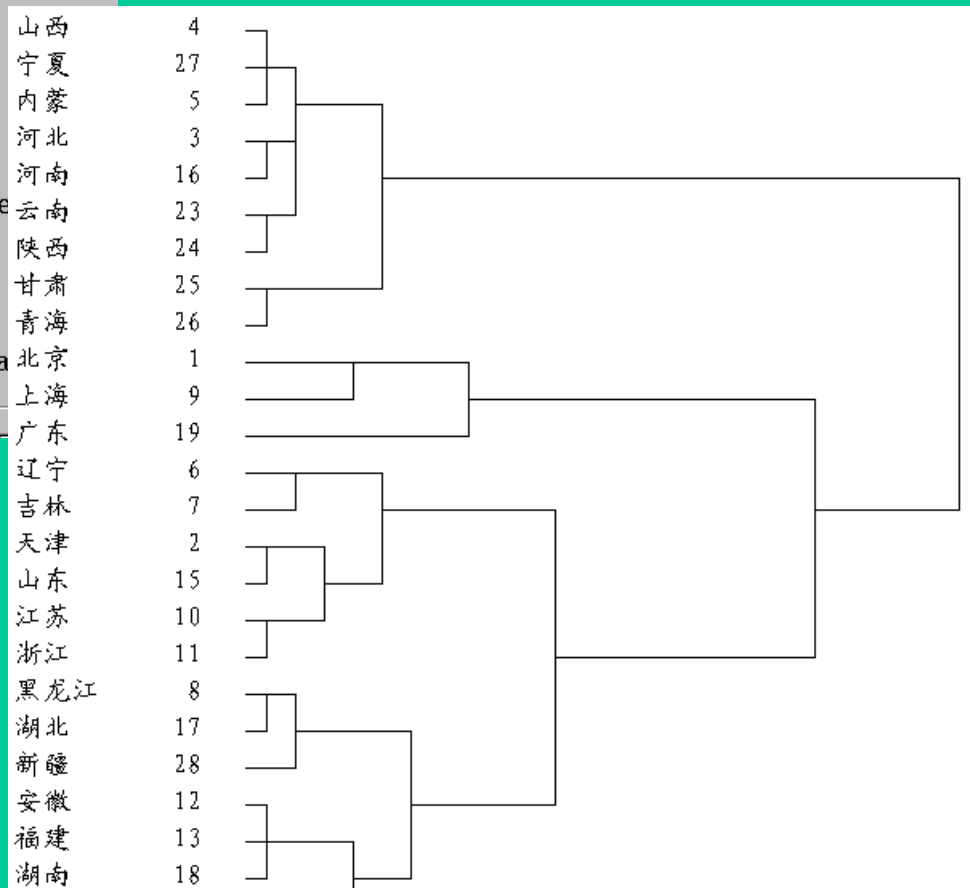
☐ Change sign

☐ Rescale to 0-1 range

Continue

Cancel

Help



- 通过比较，可知离差平方和法（Ward's method）分类结果较好，将28各样本分为三类：
 1. 第一类包含6个元素：2、15、10、11、6、7
 2. 第二类包含10个元素：8、17、28、12、13、18、14、20、21、22
 3. 第三类包含9个元素：3、16、23、24、4、27、5、25、26
- ❖ 另有三个元素1、9、19为孤立点。
- 从分类结果可以看出：1、9、19表示北京、上海、广东三地农民属高消费生活水平；天津等第一类的农民生活水平较高；安徽等第二类的农民生活水平为中等；陕西等地的农民生活水平较低。

练习：“各地区农村居民家庭平均每人生活消费支出”聚类分析

- 数据来源：中华人民共和国国家统计局
- 网址：<http://www.stats.gov.cn/>
- 数据选择：
 - “统计数据”—〉“年度数据”
 - 〉“能源生产和消费”—〉“九、人民生活”
 - 〉“各地区农村居民家庭平均每人生活消费支出”

数据预处理:

1. 文本数据处理:

- 将网页文本复制到“记事本”中（说明部分不复制）
（注：直接复制到Excel中会以网页形式出现）
 - 将地区名中的空格去掉
 - 在第一行加入数据项名称（变量名）
- ### 2. 将“记事本”中文档全选，复制粘贴到Excel中
- ### 3. 此时会出现如下标记，选择“使用文本导入向导”

26	广东	Guangdong	2255.01
27	广西	Guangxi	1202.91
28	海南	Hainan	1080.46
29			

就绪 粘贴选项

26	广东	Gu	2255.01	228.00
27	广西	Gu	1202.91	60.26
28	海南	Ha	1080.46	37.21
29				

就绪

4. “文本导入向导”中步骤1选择“分隔符号”，步骤2种分隔符号选择“空格”
5. 导入后对数据作简单修正，剔除无用的行列后存盘即可
6. 用SPSS打开该Excel文档（文件类型选“*.xls”即可）
7. 聚类分析

试用不同方法对变量进行聚类，并分析结果的含义

