7.4　聚类分析

2014年6月25日

聚类分析的基本思想

上一节介绍的判别分析,其特点是：事先知道研究对象分为几个类别，而且有一些类别已知的样品，从这些类别已知的样品数据出发，建立一种判别方法，以此,对类别未知的样品可以判别其分类

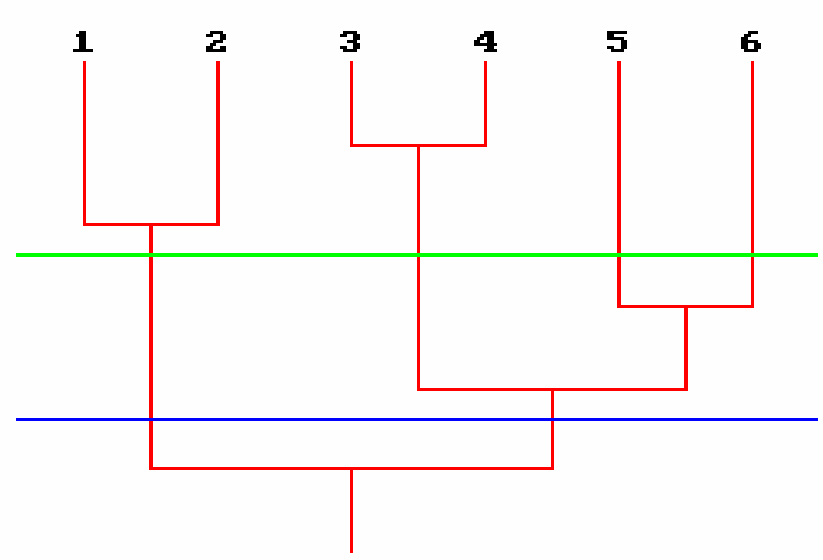
聚类分析（**Cluster Analysis**）是指:　有一些样品需要分类，但是它们可以分成哪几类，是什么样的类型，事先都是不知道的，也没有什么已知类别的样品可以作为分类的参考。我们只能根据“物以类聚”的原则，把特性比较接近的样品聚集在一起，以此对样品进行分类的方法

例如，　动植物的分类。采集了一大批某种动物或植物的标本，事先不知道它们可以分为几类，只是根据从标本测得的各种数据(例如 动物的各种体形特征，植物的各种外形尺寸)考虑把特征相近的标本聚集在一起，分成几种类型。这是一个聚类分析问题

又如, 上市股票的分类。在一个股市中，有成百上千种股票对每一种股票，都有一大批数据（例如股票价格、成交量、市盈率、公司资本、负债、产值、利润等等），要求把特征相近的股票聚集在一起，分成几种类型。这也是一个聚类分析问题

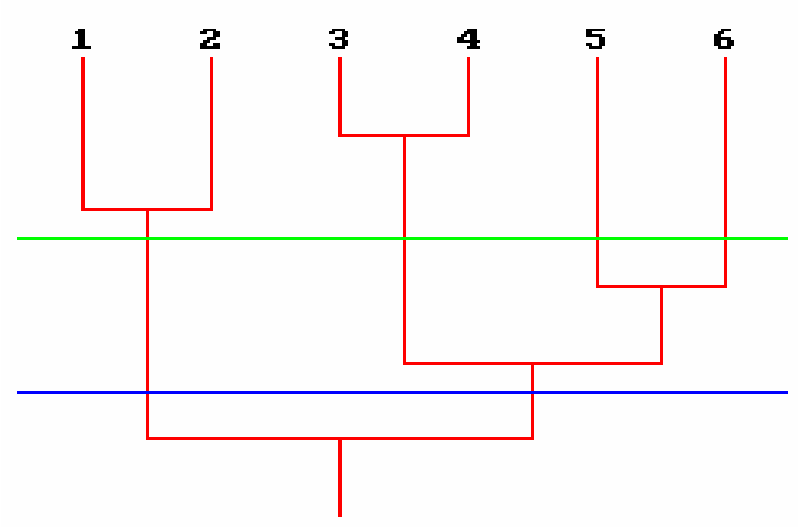
再如，不同气象情况年份的分类。对某地积累了许多年的气象资料，对每一年，都有一大批数据（例如各个月份的平均气温、降水量、年最高气温、年最低气温等等），要求把气象情况相近的年份聚集在一起，分成几种类型。这也是一个聚类分析问题

聚类的方法很多，比如: 系统聚类, 动态聚类, 模糊聚类 等等
这里我们只介绍最常用的、也是比较成熟的一种方法是**系统聚类法（Hierarchical Clustering Method，**又称**谱系聚类法**）。系统聚类法的基本思想是：开始时将每个样品单独作为一类，类与类之间的距离也就是样品与样品之间的距离。然后，找出距离最近的两类，将它们合并成一类，再找出距离最近的两类，将它们合并为一类，……，这样一直下去，每次类别的个数减少1，直到所有的样品合并成为一类为止



聚类分析的结果，可以用一个图的形式表示出来，这种图像一棵树的样子(如图)，称为聚类图

要知道分成 m 类的聚类结果，可以在聚类图中与 m 条竖线相交的高度处画一条水平线，这 m 条竖线对应的就是用系统聚类法分成的 m个 类 (如图)



在上图中，如果我们希望知道分成2类的聚类结果，可以在与2条竖线相交的高度处画一条水平线，可以看出，分成的2类是：{1，2}，{3，4，5，6}

如果我们希望知道分成4类的聚类结果，可以在与4条竖线相交的高度处画一条水平线，可以看出，分成的4类是：

{1，2}，{3，4}，{5}，{6}

系统聚类法中类与类之间的距离

在系统聚类法的每一步中，都要寻找距离最近的两类，所以，必须对类与类之间的距离作出定义

我们用小写的 d_{ij} ，$i=1,2,\cdots,n$，$j=1,2,\cdots,n$

表示第 i 个样品与第 j 个样品之间的距离:

d_{ij} 可以是闵氏距离 $d_{ij}=\left(\sum_{k=1}^{m}\left|x_{ik}-x_{jk}\right|^{q}\right)^{\frac{1}{q}}$

也可以是马氏距离 $d_{ij}=\sqrt{(x_i-x_j)^T S^{-1}(x_i-x_j)}$

设 $G_1,G_2,\cdots,G_g$ 表示分成的各个类

我们用大写: D_{pq} 表示 G_p 类与 G_q 类之间的距离

系统聚类一开始很简单，每一个类只有一个样品，类与类之间的距离也就是样品与样品之间的距离

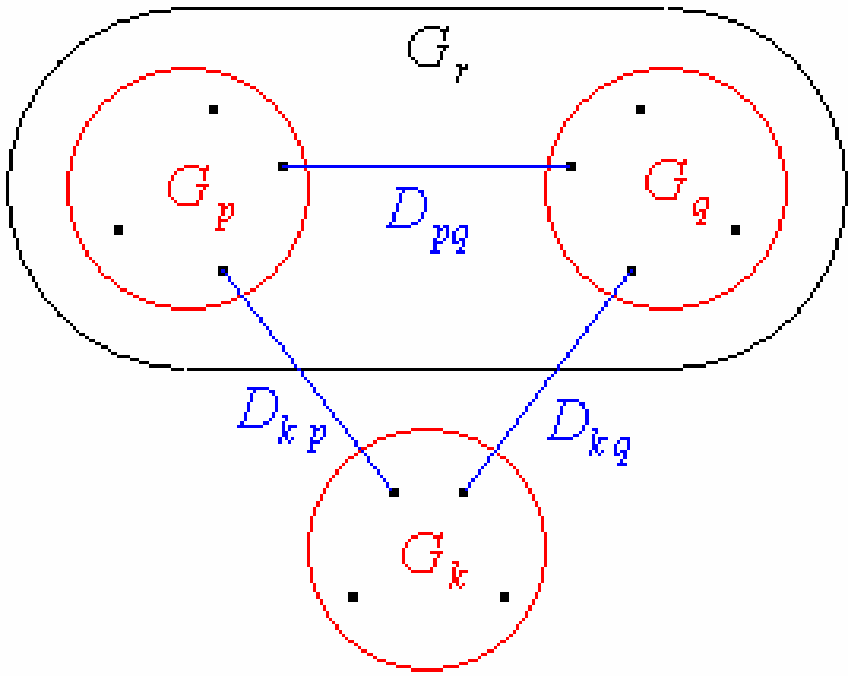
从第二步开始就不一样了,要计算类与类的距离了，下面介绍几种常用的类与类之间距离的定义:

（1）最短距离法

定义 G_p 类与 G_q 类之间的距离 D_{pq} 为 G_p, G_q 这两类中最近的两个样品之间的距离，即 $D_{pq} = \min_{x_i \in G_p,\, x_j \in G_q} d_{ij}$

类与类距离定义的最短距离法：

$$D_{pq} = \min_{x_i \in G_p,\, x_j \in G_q} d_{ij}$$



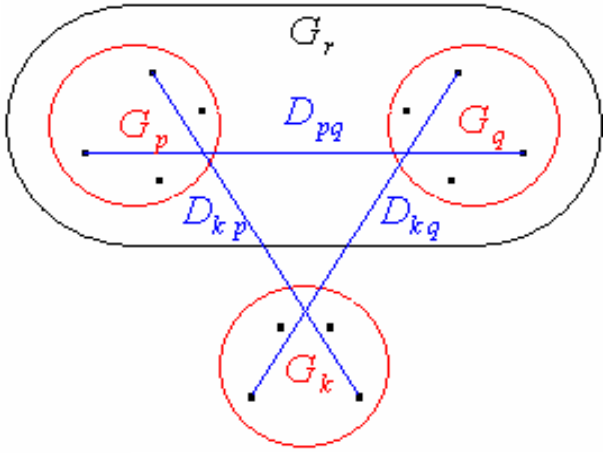
如果将 G_p, G_q 合并成一个新的类 G_r，这时，

其他的类 G_k 到 G_r 的距离显然可由下式求出：

$$D_{kr} = \min_{x_i \in G_k,\, x_j \in G_r} d_{ij} = \min\left\{ \min_{x_i \in G_k,\, x_j \in G_p} d_{ij},\ \min_{x_i \in G_k,\, x_j \in G_q} d_{ij} \right\} = \min\{ D_{kp}, D_{kq} \}$$

（2）最长距离法

定义 G_p 类与 G_q 类之间的距离 D_{pq} 为 G_p, G_q 这两类中最远的两个样品之间的距离，即 $D_{pq} = \max_{x_i \in G_p, x_j \in G_q} d_{ij}$



如果将 G_p, G_q 合并成一个新的类 G_r，这时，其他的类 G_k 到 G_r 的距离显然可由下式求出：

$$D_{kr} = \max_{x_i \in G_k, x_j \in G_r} d_{ij} = \max\left\{ \max_{x_i \in G_k, x_j \in G_p} d_{ij}, \max_{x_i \in G_k, x_j \in G_q} d_{ij} \right\}$$

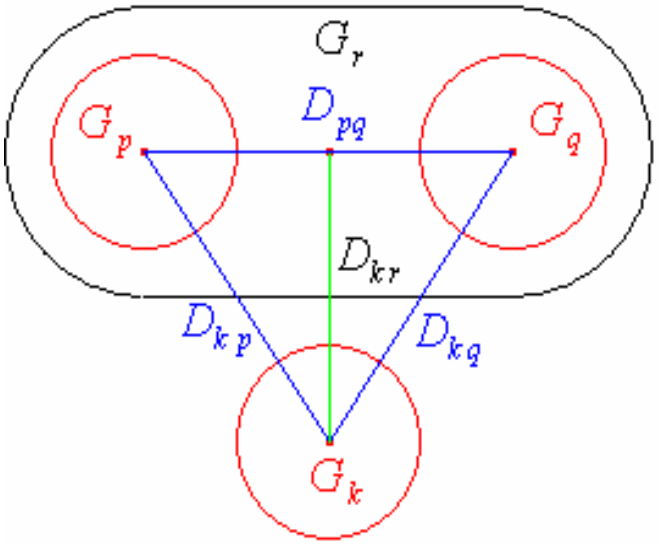
$$= \max\{D_{kp}, D_{kq}\} \quad 。$$

（3）中间距离法

首先，对每一类都可以确定一个中心：如果一个类中只有一个样品，则中心就是这个样品，如果将两类合并，则合并后的类的中心，就是原来两类中心的连线的中点

定义 G_p 类与 G_q 类之间的距离 D_{pq} 为 G_p 的中心与 G_q 的中心之间的欧氏距离。

如果将 G_p, G_q 合并成一个新的类 G_r，这时

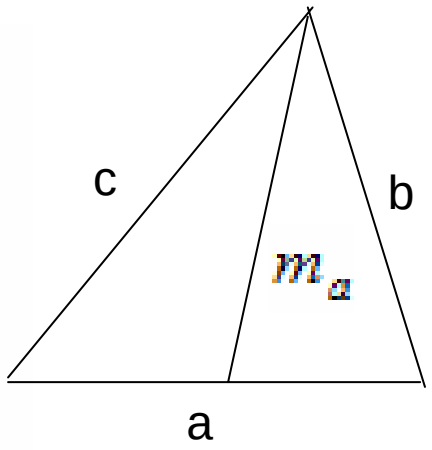


其他的类 G_k 到 G_r 的距离可由下式求出：

$$D_{kr}^2 = \frac{1}{2}D_{kp}^2 + \frac{1}{2}D_{kq}^2 - \frac{1}{4}D_{pq}^2$$

类 G_k 到 G_r 的距离的平方公式是什么含意呢?

$$D_{kr}^2 = \frac{1}{2}D_{kp}^2 + \frac{1}{2}D_{kq}^2 - \frac{1}{4}D_{pq}^2$$

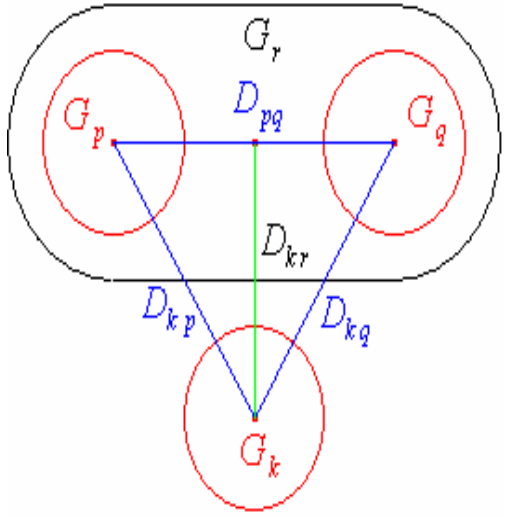


由平面几何可知，在边长为 a,b,c 的三角形中，

a 边上的中线长度 m_a 等于 $m_a = \sqrt{\frac{1}{2}b^2 + \frac{1}{2}c^2 - \frac{1}{4}a^2}$

而 D_{pq} 相当于 a，D_{kp} 相当于 b，D_{kq} 相当于 c，

D_{kr} 相当于 a 边上的中线长度 m_a



（4）重心法

首先，对每一类 G_p 都可以确定一个重心：重心就是属于这一类的样品的观测值的样本均值 $\bar{x}_p = \frac{1}{n_p}\sum_{x_i \in G_p} x_i$ ，

其中 n_p 是 G_p 类中的样品数，$x_i = \begin{bmatrix} x_{i1} \\ \vdots \\ x_{im} \end{bmatrix}$ 是第 i 个样品的观测值。

如果一个类中只有一个样品，则重心就是这个样品，如果将 G_p, G_q 两类合并，则合并后的类 G_r 的重心为：

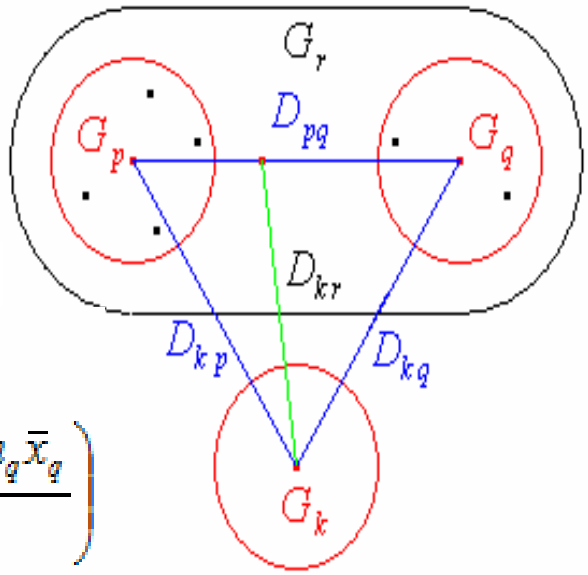
$\bar{x}_r = \frac{n_p \bar{x}_p + n_q \bar{x}_q}{n_r}$ ，其中 n_p, n_q 是 G_p, G_q 中的样品数

$n_r = n_p + n_q$ 是 G_r 中的样品数。

定义 G_p 类与 G_q 类之间的距离 D_{pq} 为 G_p 的重心与 G_q 的重心之间的欧氏距离。

如果将 G_p, G_q 合并成一个新的类 G_r，这时，其他的类 G_k 到 G_r 的距离可由下式求出：

$$D_{kr}^2 = \frac{n_p}{n_r} D_{kp}^2 + \frac{n_q}{n_r} D_{kq}^2 - \frac{n_p n_q}{n_r^2} D_{pq}^2$$



这是因为

$$D_{kr}^2 = (\bar{x}_k - \bar{x}_r)^T(\bar{x}_k - \bar{x}_r) = \left(\bar{x}_k - \frac{n_p\bar{x}_p + n_q\bar{x}_q}{n_r}\right)^T\left(\bar{x}_k - \frac{n_p\bar{x}_p + n_q\bar{x}_q}{n_r}\right)$$

$$= \bar{x}_k^T\bar{x}_k - \frac{2n_p\bar{x}_k^T\bar{x}_p}{n_r} - \frac{2n_q\bar{x}_k^T\bar{x}_q}{n_r} + \frac{n_p^2\bar{x}_p^T\bar{x}_p + 2n_pn_q\bar{x}_p^T\bar{x}_q + n_q^2\bar{x}_q^T\bar{x}_q}{n_r^2}$$

$$D_{kr}^2 = (\bar{x}_k - \bar{x}_r)^T(\bar{x}_k - \bar{x}_r) = \left(\bar{x}_k - \frac{n_p\bar{x}_p + n_q\bar{x}_q}{n_r}\right)^T\left(\bar{x}_k - \frac{n_p\bar{x}_p + n_q\bar{x}_q}{n_r}\right)$$

$$= \bar{x}_k^T\bar{x}_k - \frac{2n_p\bar{x}_k^T\bar{x}_p}{n_r} - \frac{2n_q\bar{x}_k^T\bar{x}_q}{n_r} + \frac{n_p^2\bar{x}_p^T\bar{x}_p + 2n_pn_q\bar{x}_p^T\bar{x}_q + n_q^2\bar{x}_q^T\bar{x}_q}{n_r^2}$$

$$= \frac{n_p(\bar{x}_k^T\bar{x}_k - 2\bar{x}_k^T\bar{x}_p + \bar{x}_p^T\bar{x}_p)}{n_r} + \frac{n_q(\bar{x}_k^T\bar{x}_k - 2\bar{x}_k^T\bar{x}_q + \bar{x}_q^T\bar{x}_q)}{n_r} - \frac{n_pn_q(\bar{x}_p^T\bar{x}_p - 2\bar{x}_p^T\bar{x}_q + +\bar{x}_q^T\bar{x}_q)}{n_r^2}$$

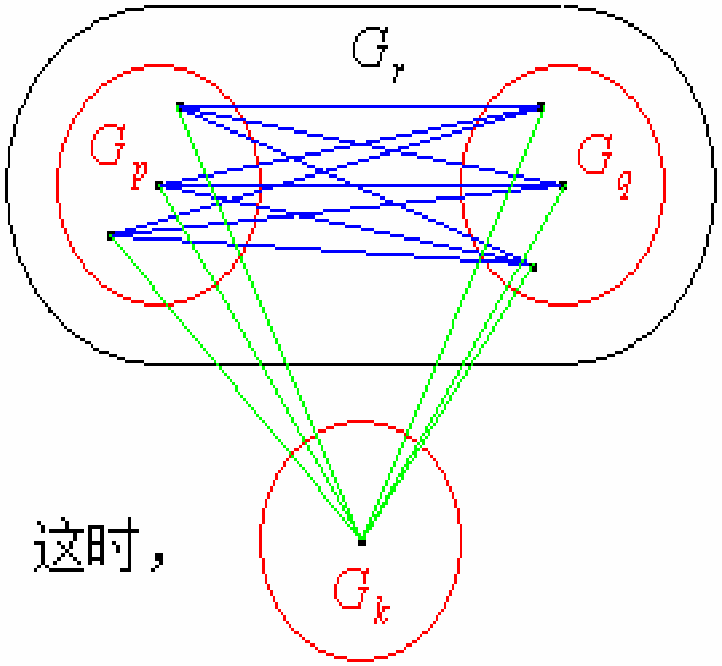
$$= \frac{n_p}{n_r}(\bar{x}_k - \bar{x}_p)^T(\bar{x}_k - \bar{x}_p) + \frac{n_q}{n_r}(\bar{x}_k - \bar{x}_q)^T(\bar{x}_k - \bar{x}_q) - \frac{n_pn_q}{n_r^2}(\bar{x}_p - \bar{x}_q)^T(\bar{x}_p - \bar{x}_q)$$

$$= \frac{n_p}{n_r}D_{kp}^2 + \frac{n_q}{n_r}D_{kq}^2 - \frac{n_pn_q}{n_r^2}D_{pq}^2$$

(5) 类平均法

定义 G_p 类与 G_q 类之间的距离 D_{pq} 的平方为 G_p 中样品与 G_q 中样品的距离的平方的平均值

$$D_{pq}^2 = \frac{1}{n_p n_q} \sum_{x_i \in G_i} \sum_{x_j \in G_q} d_{ij}^2$$



如果将 G_p, G_q 合并成一个新的类 G_r，这时，

其他的类 G_k 到 G_r 的距离显然可由下式求出：

$$D_{kr}^2 = \frac{n_p}{n_r} D_{kp}^2 + \frac{n_q}{n_r} D_{kq}^2$$

（6）离差平方和法（Ward 法）

对每一类 G_p 都可以定义一个离差平方和

$$SS_p = \sum_{x_i \in G_p} (x_i - \bar{x}_p)^T (x_i - \bar{x}_p) = \sum_{x_i \in G_p} x_i^T x_i - n_p \bar{x}_p^T \bar{x}_p$$

如果将 G_p, G_q 合并成一个新的类 G_r，G_r 的离差平方和 $SS_r > SS_p + SS_q$，

G_p 离 G_q 越远，SS_r 越大，G_p 离 G_q 越近，SS_r 越小，所以，可以定义：

G_p 类与 G_q 类之间的距离平方为 $D_{pq}^2 = SS_r - SS_p - SS_q$

可以证明：

如果将 G_p, G_q 合并成一个新的类 G_r，这时，其他的类 G_k 到 G_r 的距离可由下式求出：

$$D_{kr}^2 = \frac{n_k + n_p}{n_k + n_r} D_{kp}^2 + \frac{n_k + n_q}{n_k + n_r} D_{kq}^2 - \frac{n_k}{n_k + n_r} D_{pq}^2$$

系统聚类法的统一公式和计算步骤

前面介绍了6 种常用的系统聚类法，这些方法的区别在于：它们对类与类之间的距离有不同的定义.

如果将 G_p, G_q 两类合并成一个新的类 G_r，这时，其他的类 G_k 到 G_r 的距离就有不同的计算公式。

1969 年，Wishart 发现这些公式可以统一起来，写成下列统一形式：

$$D_{kr}^2 = \alpha_p D_{kp}^2 + \alpha_q D_{kq}^2 + \beta D_{pq}^2 + \gamma \left| D_{kp}^2 - D_{kq}^2 \right|$$

其中系数 $\alpha_p, \alpha_q, \beta, \gamma$ 对于不同的方法有不同的取值

下表列出了**Wishart**的统一计算公式中参数对应的取值：

方法	α_p	α_q	β	γ
最短距离法	$1/2$	$1/2$	0	$-1/2$
最长距离法	$1/2$	$1/2$	0	$1/2$
中间距离法	$1/2$	$1/2$	$-1/4$	0
重心法	$\frac{n_p}{n_r}$	$\frac{n_q}{n_r}$	$-\frac{n_p n_q}{n_r^2}$	0
类平均法	$\frac{n_p}{n_r}$	$\frac{n_q}{n_r}$	0	0
离差平方和法	$\frac{n_k+n_p}{n_k+n_r}$	$\frac{n_k+n_q}{n_k+n_r}$	$-\frac{n_k}{n_k+n_r}$	0

系统聚类的步骤

建立一个 D^2 矩阵，其中元素是类与类之间距离的平方：

$$\begin{bmatrix} 0 & D_{12}^2 & \cdots & D_{1n}^2 \\ D_{21}^2 & 0 & \cdots & D_{2n}^2 \\ \vdots & \vdots & \ddots & \vdots \\ D_{n1}^2 & D_{n2}^2 & \cdots & 0 \end{bmatrix}$$

一开始，每一个样品单独作为一类，类与类之间的距离就是样品与样品之间的距离。 即有 $D_{pq}^2 = d_{pq}^2$。

然后，在 D^2 矩阵的非对角元素中找出一个最小值 D_{pq}^2 。

D_{pq}^2 在所有非对角元素中最小，说明在现有的各类中 G_p 类与 G_q 类距离最近，将 G_p, G_q 两类合并成一个新的类 G_r 按照前面给出的统一计算公式，可以求出其他的类 G_k 到 G_r 的距离，从而建立一个新的 D^2 矩阵。

然后，再在新的 D^2 矩阵的非对角元素中找出一个最小值 D_{pq}^2，

将 G_p, G_q 两类合并成一个新的类 G_r，按照前面给出的统一计算公式，

可以求出其他的类 G_k 到 G_r 的距离，再建立一个新的 D^2 矩阵，……

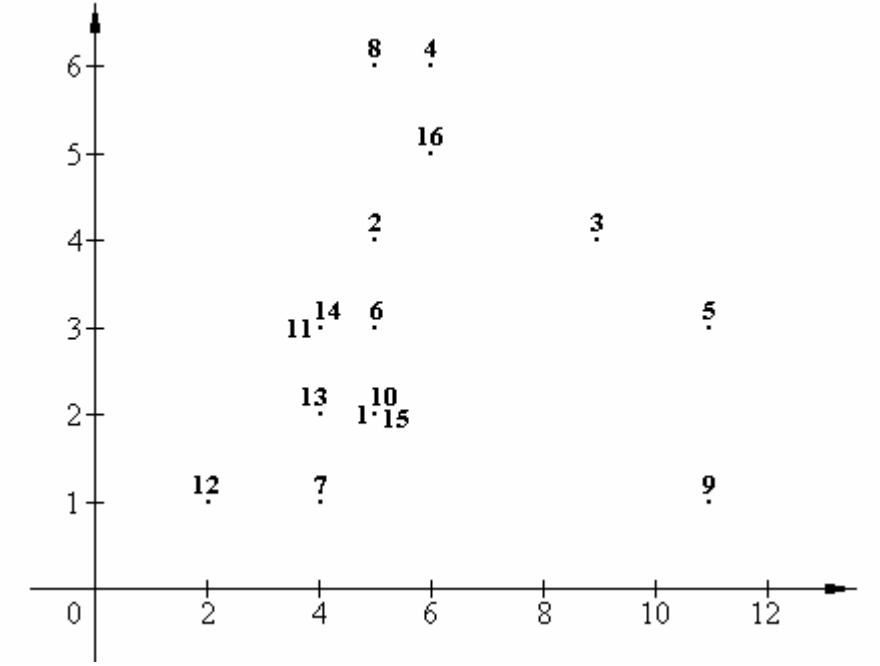
就这样，一直下去，每次类别的个数减少 1，直到所有的样品合并成为一类为止

记录下全部合并过程，就能画出聚类图。从聚类图就可以得到聚类分析的结果

聚类分析应用实例

例 **2002**年足球世界杯赛**16**强

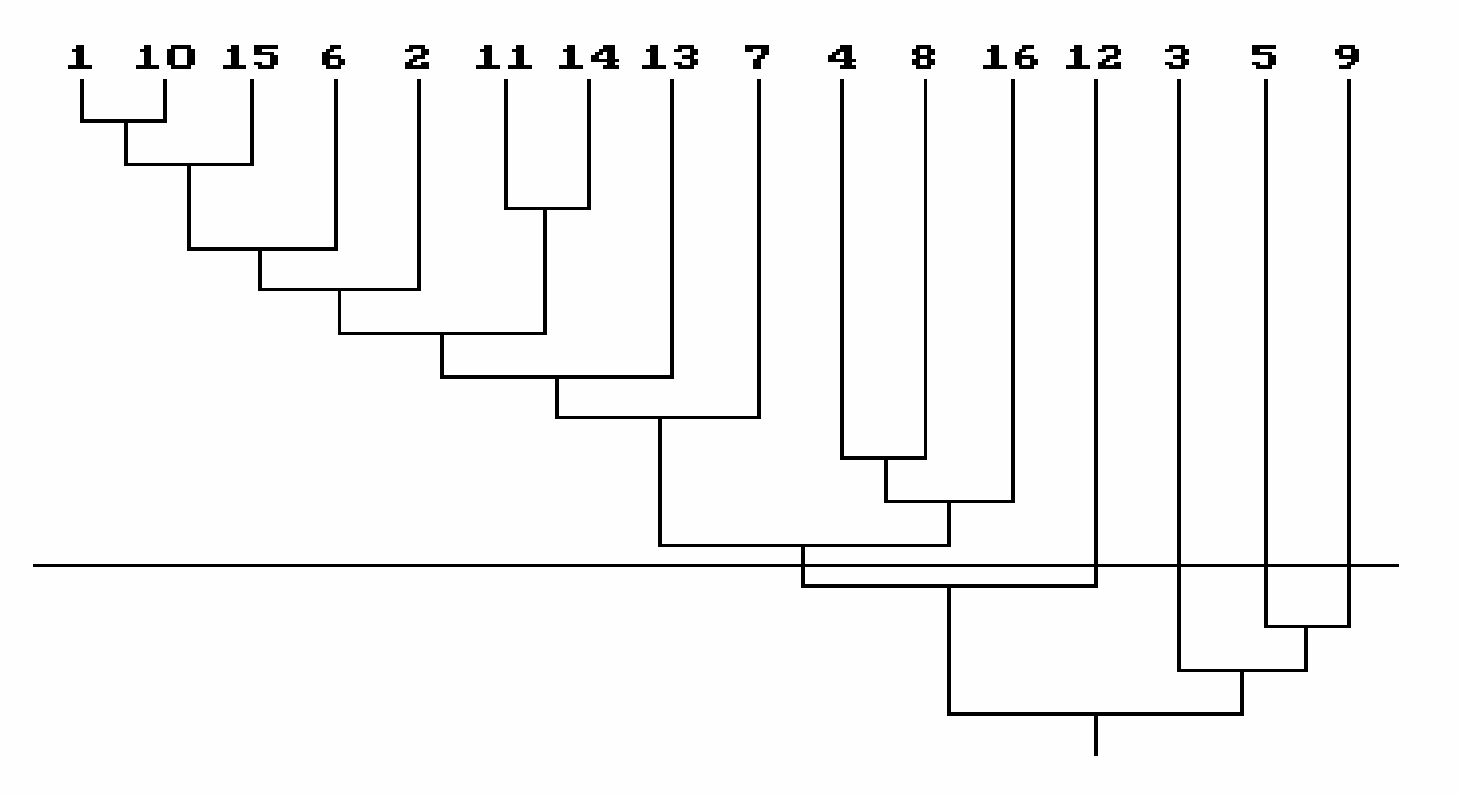
2002年足球世界杯赛，最后有16支球队进入前16名，这些球队在进入前16名以前的分组赛中的进球数和失球数统计如右表,作图如下：



编号	球队名称	X_1进球数	X_2失球数
1	丹麦	5	2
2	塞内加尔	5	4
3	西班牙	9	4
4	巴拉圭	6	6
5	巴西	11	3
6	土耳其	5	3
7	韩国	4	1
8	美国	5	6
9	德国	11	1
10	爱尔兰	5	2
11	瑞典	4	3
12	英格兰	2	1
13	墨西哥	4	2
14	意大利	4	3
15	日本	5	2
16	比利时	6	5

下面对这16支球队进行系统聚类分析。
因为进球数和失球数是同一类型的变量，所以不必对它们进行标准化处理。选用欧氏距离作为样品与样品之间的距离

（1）最短距离法 得到的聚类图:



从聚类图可以看出，如果分成5类，最短距离法有下列聚类结果：

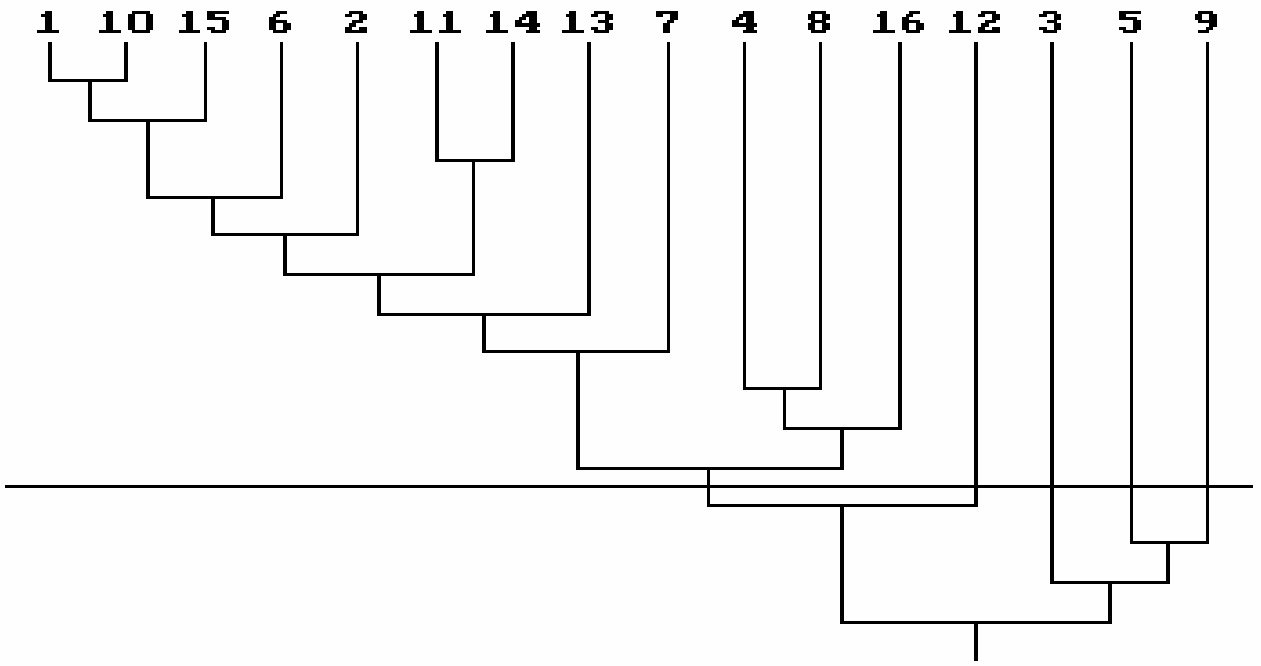
第1类：{ 9.德国 }

第2类：{ 5.巴西 }

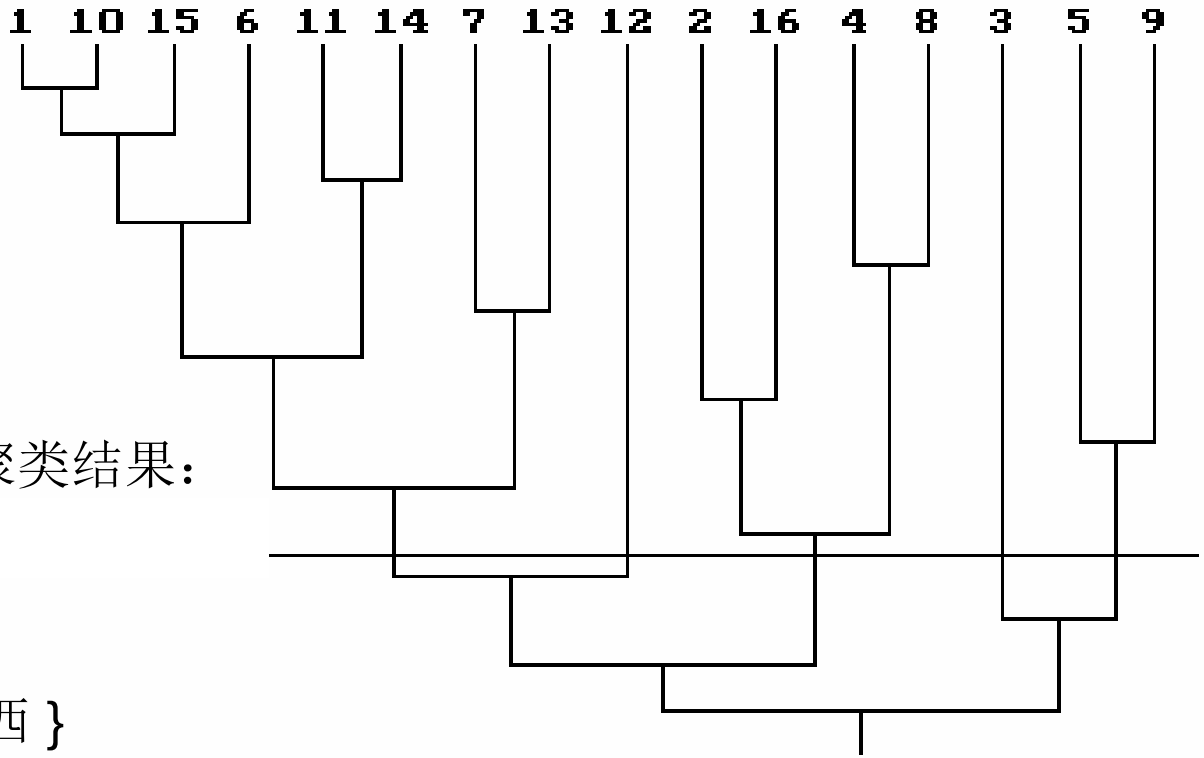
第3类：{ 3.西班牙 }

第4类：{ 12.英格兰 }

第5类：{ 16.比利时，8.美国，4.巴拉圭，7.韩国，13.墨西哥，14.意大利，
11.瑞典，2.塞内加尔，6.土耳其，15.日本，10.爱尔兰，1.丹麦 }



（2）最长距离法 的聚类结果:



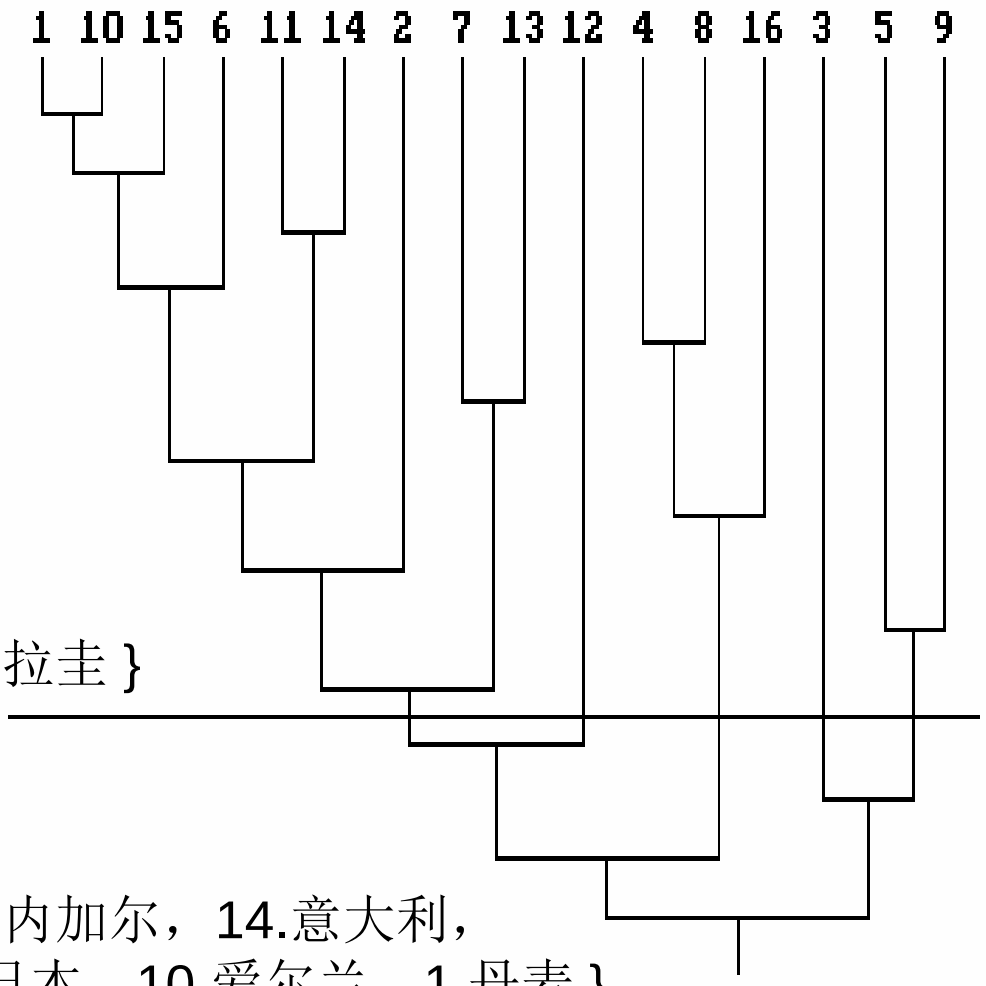
从聚类图可以看出

最长距离法分成5类的聚类结果：

第1类：{ 9.德国，5.巴西 }
第2类：{ 3.西班牙 }
第3类：{ 8.美国，4.巴拉圭，16.比利时，2.塞内加尔 }
第4类：{ 12.英格兰 }
第5类：{ 13.墨西哥，7.韩国， 14.意大利，11.瑞典，6.土耳其，15.日本，
10.爱尔兰，1.丹麦 }

（3）中间距离法 得到的聚类结果

从聚类图可以看出



中间距离法分成5类的聚类结果：

第1类：{ 9.德国，5.巴西 }

第2类：{ 3.西班牙 }

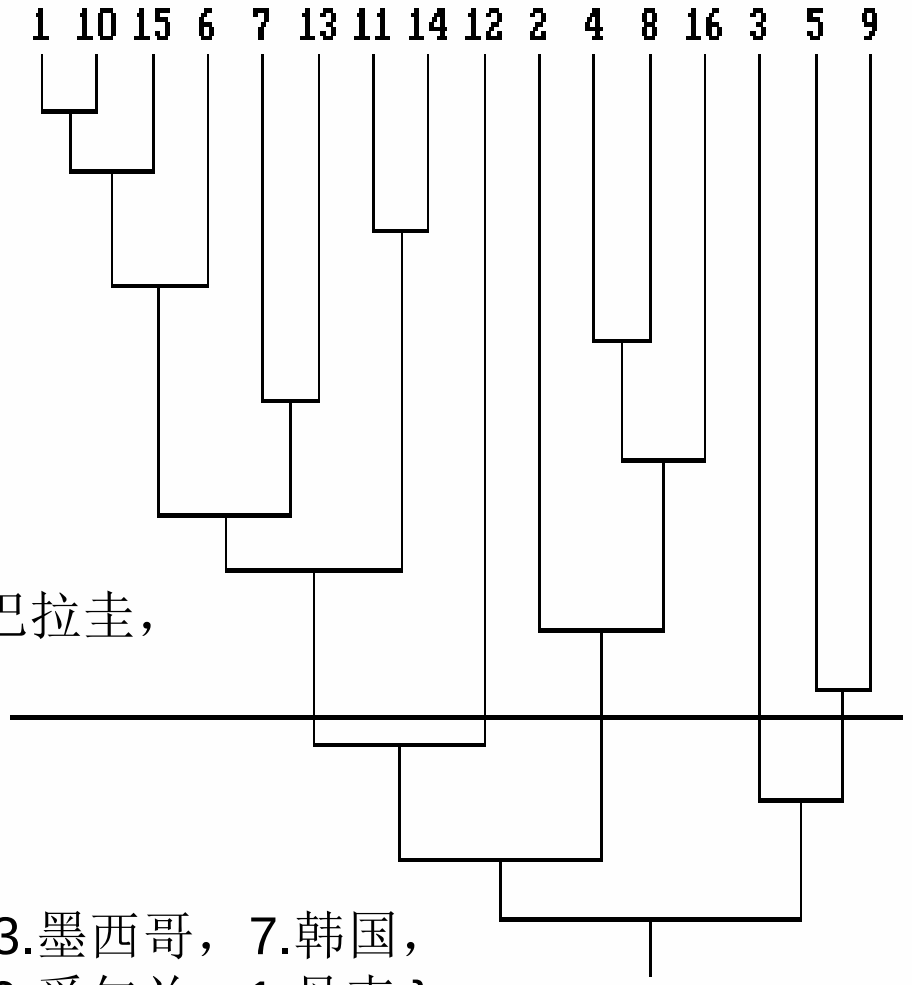
第3类：{ 16.比利时，8.美国，4.巴拉圭 }

第4类：{ 12.英格兰 }

第5类：{ 13.墨西哥，7.韩国，2.塞内加尔，14.意大利，
11.瑞典，6.土耳其，15.日本，10.爱尔兰，1.丹麦 }

（4）重心法 得到的聚类结果:

从聚类图可以看出



重心法分成5类的结果：

第1类：{ 9.德国，5.巴西 }

第2类：{ 3.西班牙 }

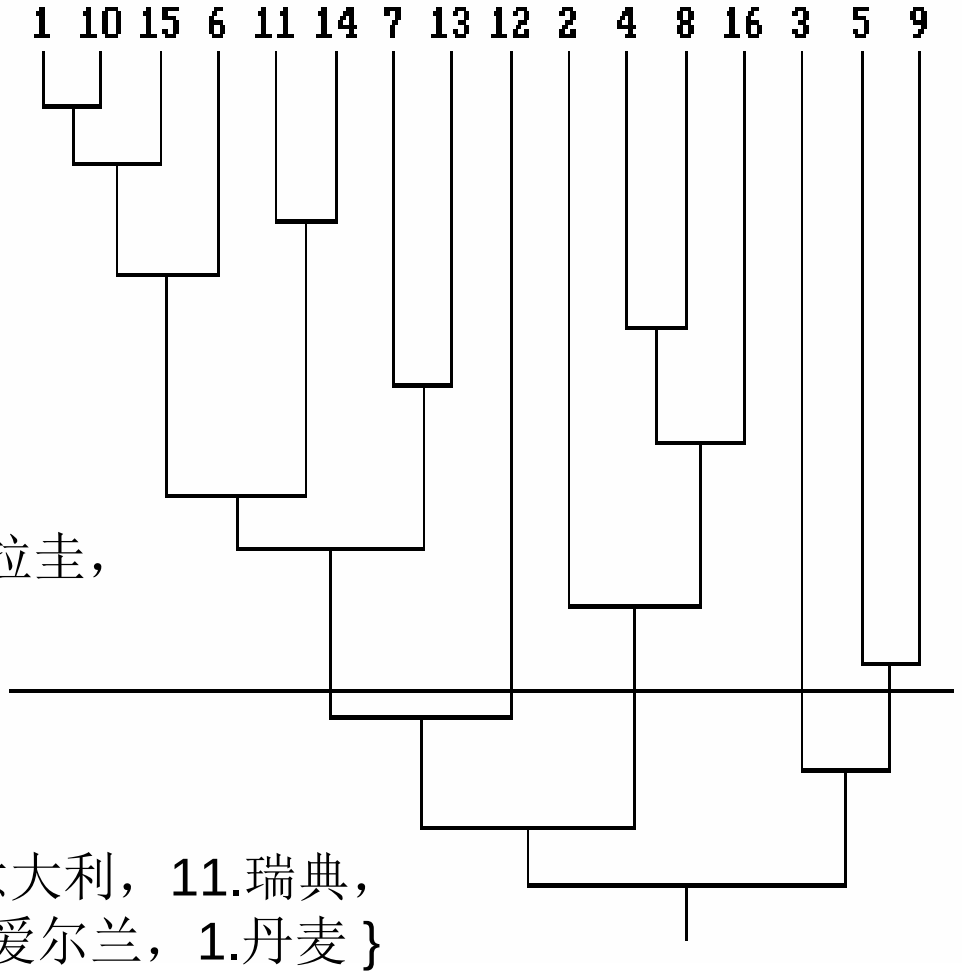
第3类：{ 16.比利时，8.美国，4.巴拉圭，
2.塞内加尔 }

第4类：{ 12.英格兰 }

第5类：{ 14.意大利，11.瑞典，13.墨西哥，7.韩国，
6.土耳其，15.日本，10.爱尔兰，1.丹麦 }

（5）类平均法 得到的聚类结果:

类平均法分成5类有下列聚类结果：



第1类：{ 9.德国，5.巴西 }

第2类：{ 3.西班牙 }

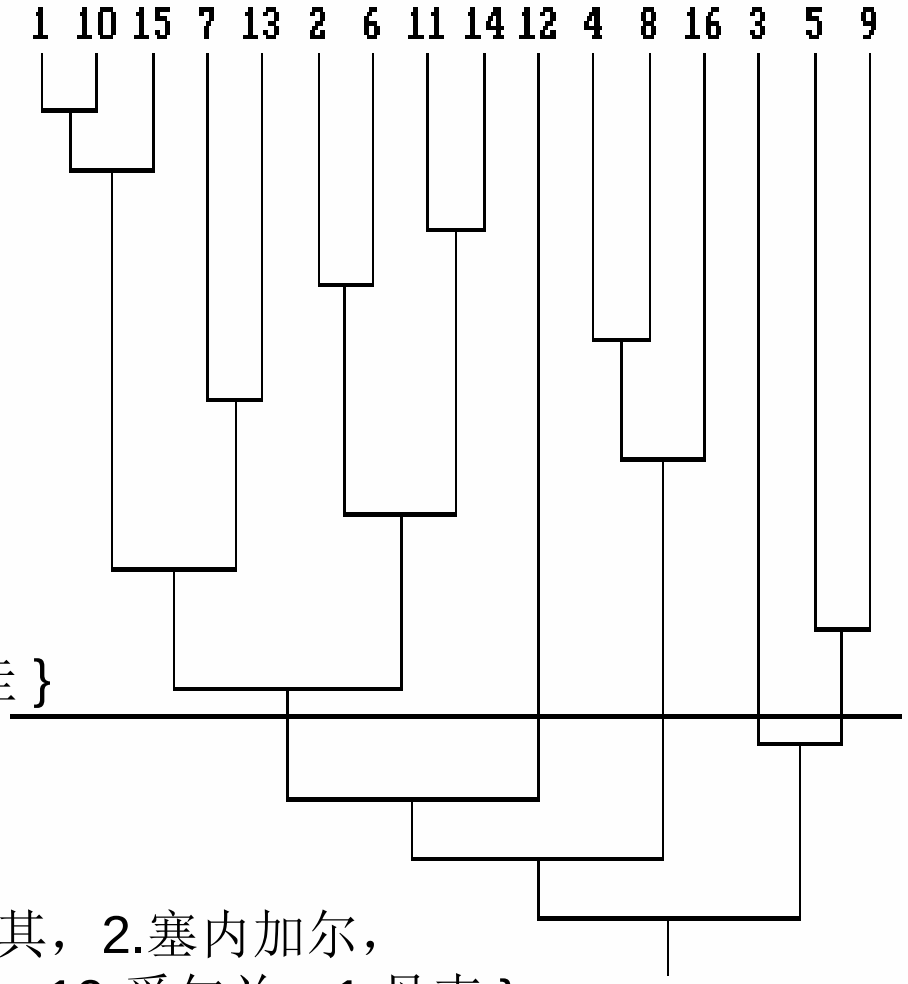
第3类：{ 16.比利时，8.美国，4.巴拉圭，
2.塞内加尔 }

第4类：{ 12.英格兰 }

第5类：{ 13.墨西哥，7.韩国，14.意大利，11.瑞典，
6.土耳其，15.日本，10.爱尔兰，1.丹麦 }

（6）离差平方和法 得到的聚类结果

从聚类图可以看出，如果分成5类



离差平方和法有下列聚类结果：

第1类：{ 9.德国，5.巴西 }

第2类：{ 3.西班牙 }

第3类：{ 16.比利时，8.美国，4.巴拉圭 }

第4类：{ 12.英格兰 }

第5类：{ 14.意大利，11.瑞典，6.土耳其，2.塞内加尔，
13.墨西哥，7.韩国，15.日本，10.爱尔兰，1.丹麦 }

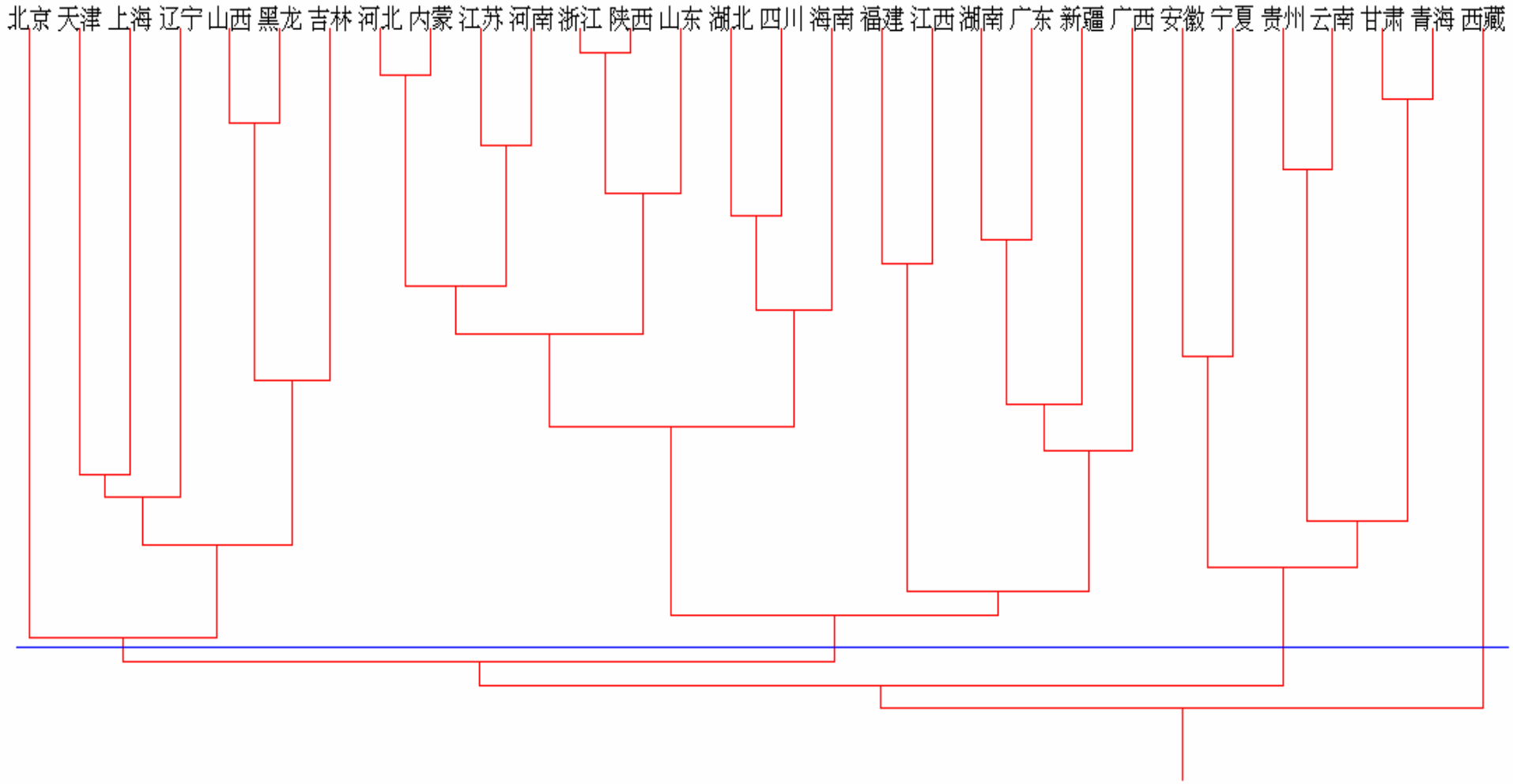
例 全国各地文化程度状况分类

1990年全国人口普查对全国30个省、直辖市、自治区各种文化程度的人口比例作了统计得到数据如下

编号	地区	大学以上	初中	文盲半文盲	编号	地区	大学以上	初中	文盲半文盲
1	北京	9.30	30.55	8.70	16	河南	0.85	26.55	16.15
2	天津	4.67	29.38	8.92	17	湖北	1.57	23.16	15.79
3	河北	0.96	24.69	15.21	18	湖南	1.14	22.57	12.10
4	山西	1.38	29.24	11.30	19	广东	1.34	23.04	10.45
5	内蒙古	1.48	25.47	15.39	20	广西	0.79	19.14	10.61
6	辽宁	2.60	32.32	8.81	21	海南	1.24	22.53	13.97
7	吉林	2.15	26.31	10.49	22	四川	0.96	21.65	16.24
8	黑龙江	2.14	28.46	10.87	23	贵州	0.78	14.65	24.27
9	上海	6.53	31.59	11.04	24	云南	0.81	13.85	25.44
10	江苏	1.47	26.43	17.23	25	西藏	0.57	3.85	44.43
11	浙江	1.17	23.74	17.46	26	陕西	1.67	24.36	17.62
12	安徽	0.88	19.97	24.43	27	甘肃	1.10	16.85	27.93
13	福建	1.23	16.87	15.63	28	青海	1.49	17.76	27.70
14	江西	0.99	18.84	16.22	29	宁夏	1.61	20.27	22.06
15	山东	0.98	25.18	16.87	30	新疆	1.85	20.66	12.75

要求根据文化程度状况统计数据对上述地区作聚类分析

用“类平均法”作系统聚类得到的聚类图:



从用“类平均法”作系统聚类的聚类图可以看出，如果分成4类则聚类结果为:

第1类：{ 1.北京，2.天津，9.上海，6.辽宁，4.山西，8.黑龙江，7.吉林 }

第2类：{ 3.河北，5.内蒙古，10.江苏，16.河南，11.浙江，26.陕西，
15.山东，17.湖北，22.四川，21.海南，13.福建，14.江西，
18.湖南，19.广东，30.新疆， 20.广西 }

第3类：{ 12.安徽，29.宁夏，23.贵州，24.云南，27.甘肃，28.青海 }

第4类：{ 25.西藏 }

教 材 内 容 讲 解 完 毕

作为复习,我们最后再讲解一套综合测试题

数理统计综合测试题-讲解

一． 选择题（每小题 4 分，共 36 分）

1．设总体的期望 μ 和方差 σ^2 均未知，从总体中抽取了一个容量为 n 的样本 $(X_1, X_2, \cdots, X_n)$，则下述选项中可以作为总体的期望 μ 和方差 σ^2 的无偏估计量的选项是（ A ）

(A) X_1 和 $\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2$　　(B) $\bar{X}$ 和 $\frac{1}{n-1}\sum_{i=1}^{n}X_i^2-\bar{X}^2$

(C) $\bar{X}$ 和 $\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2$　　(D) $\bar{X}$ 和 $\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2$

2．5 名评委对某歌手的打分分别是：63，65，70，71，95，根据打分，代表该歌手水平最合理的指标应是这些分值的（ B ）.

(A) 均值；　　(B) 中值；　　(C) 方差；　　(D) 众数

3．设总体期望为 μ，方差为 σ_0^2，$(X_1, X_2, \cdots, X_n)$ 为总体的一个容量为 n 的样本，$\overline{X}$ 为样本均值，则（ D ）.

（A） 当 n 充分大时，$\overline{X}$ 近似服从正态分布 $N(\mu, \sigma_0^2)$；

(B) 当 n 充分大时，$\overline{X}$ 的取值收敛于总体期望 μ；

（C） 因总体分布未知，无论 n 多大，$\overline{X}$ 都未必可视为服从正态分布；

(D) 当 n 充分大时，$\overline{X}$ 近似服从正态分布 $N(\mu, \sigma_0^2/n)$

4. 设总体 $\xi \sim N(1,2^2)$，$(X_1, X_2, \cdots, X_{10})$ 是 ξ 的样本

$Y = (X_1 - 1)^2 + (X_2 - 1)^2 + \ldots + (X_{10} - 1)^2$，则下述选项正确的是（ C ）.

（A） $Y \sim \chi^2(10)$；　　（B） $Y \sim N(10,\ 40)$；

（C） $\frac{Y}{4} \sim \chi^2(10)$；　　（D） $\frac{Y}{2} \sim \chi^2(10)$

5. 不考虑交互作用的正交试验，若问题中有 4 个因子，每个因子都是 2 个水平，应选取的正交表是（ B ）.

（A） $L_4(2^3)$；　（B） $L_8(2^7)$；　（C） $L_9(3^4)$；　（D） $L_{16}(2^{15})$

6. 设总体$\xi \sim N(\mu, \sigma_0^2)$，其中 σ_0^2 已知，$(X_1, X_2, \cdots, X_n)$ 是 ξ 的样本，总体期望 μ 的置信水平为 $1-\alpha$ 的置信区间的长度记为 L，则**错误**的选项是 （ C ）。

（A） L 与样本容量 n 有关；　　（B） L 与置信水平 $1-\alpha$ 有关；

（C） L 与样本$(X_1, X_2, \cdots, X_n)$的取值有关；　　（D） L 与总体方差σ_0^2有关.

7．显著性水平 α 下的某假设检验，原假设 H_0，则（ A ）.

（A） 犯第一类错误的概率一定不超过 α；

（B） 犯第二类错误的概率一定为 $1-\alpha$；

（C） 犯第一类错误的概率一定为 α；

（D） 要么犯第一类错误，要么犯第二类错误，二者必居其一

8．多元线性回归模型 $\mathrm{Y}=\mathrm{X}\beta+e$，其中 $\mathrm{e}\sim N(0,\sigma^2\mathrm{I})$，关于 β 的最小二乘估计 $\hat{\beta}$ 下述**错误**的选项是（ C ）。

（A） $\hat{\beta}=(X'X)^{-1}X'Y$

（B） $\mathrm{E}(\hat{\beta})=\beta$

（C） $\hat{\beta}\sim N(\beta,\sigma^2\mathrm{I})$

（D） $\hat{\beta}$ 与残差平方和 SS_e 相互独立

9. 根据变元的 n 组观测值来求 m 元线性回归的复相关系数，下述选项正确的是（ A ）

（A） $R=\sqrt{\dfrac{SS_R}{SS_T}}$

（B） $R=\sqrt{\dfrac{SS_e}{n-2}}$

（C） $R=\sqrt{\dfrac{SS_e}{n-m-1}}$

（D） $R=\sqrt{1-\dfrac{SS_R}{SS_T}}$

二、(本题 10 分) 立邦牌油漆的干燥时间 $\xi \sim N(\mu, \sigma^2)$，随机抽取 9 个样品，测得干燥时间（单位：小时）的样本均值为 6.2，**修正样本标准差**为 0.6928，分别求 μ, σ^2 的置信水平为 95%的置信区间。

解: (1) μ 的置信水平为 $1-\alpha$ 的置信区间为：

$$[\bar{X} - t_{1-\alpha/2}(n-1)\frac{S^*}{\sqrt{n}},\quad \bar{X} + t_{1-\alpha/2}(n-1)\frac{S^*}{\sqrt{n}}]$$

=[6.2-2.306*0.6928/3, 6.2+2.306*0.6928/3]=[5.6675, 6.7325]

(2) 由样本数据得到 $n=9$， $s_{n-1}^2 = 0.48$，对于 $\alpha = 0.05$，自由度为 8 有 $\chi^2_{0.025}(8) = 2.180$，$\chi^2_{0.975}(8) = 17.535$，所以

$$\frac{(n-1)S_{n-1}^2}{\chi^2_{0.975}(n-1)} = \frac{8\times 0.48}{17.535} = 0.2190; \qquad \frac{(n-1)S_{n-1}^2}{\chi^2_{0.025}(n-1)} = \frac{8\times 0.48}{2.180} = 1.7615$$

故 σ^2 的 95%的置信区间为 [0.2190, 1.7615]

三. (本题 10 分) 设$(X_1,\ldots,X_n)$是取自总体ξ的一个简单随机样本ξ的密度函数为

$$p(x)=\begin{cases} e^{-(x-\theta)}, & x\geq\theta \\ 0, & x<\theta \end{cases}$$ 其中$\theta>0$为未知参数,

（1） 求θ的矩法估计量$\hat{\theta}$,并说明$\hat{\theta}$是否为θ的无偏估计?

（2） 求θ的极大似然估计.

解：(1)先计算 $E\xi=\int_{\theta}^{+\infty}xe^{-(x-\theta)}dx=(-xe^{-(x-\theta)})\Big|_{\theta}^{+\infty}+\int_{\theta}^{+\infty}e^{-(x-\theta)}dx=\theta+1$

由于$E\xi=\overline{X}$，得到$\hat{\theta}=\overline{X}-1$

因 $\mathrm{E}\hat{\theta}=E(\overline{X}-1)=E\overline{X}-1=E\xi-1=(\theta+1)-1=\theta$,

故$\hat{\theta}=\overline{X}-1$是$\theta$无偏估计。

(2) 对于一组观测值$(x_1, x_2, \cdots x_n)$，设$x_1, \cdots, x_n \geq \theta$，此时似然函数

$$L(\theta) = \prod_{i=1}^{n} p(x_i) = \prod_{i=1}^{n} (e^{-(x_i - \theta)})$$

两边取对数，得对数似然函数 $\ln L(\theta) = -\sum_{i=1}^{n} x_i + n\theta$

分别关于θ求导，可得 $\frac{d \ln L(\theta)}{d\theta} = n > 0$ $\ln L(\theta)$关于θ严格单调递增

所以$\ln L(\theta)$的极大值应在θ取值的右面的边界点上取到，

故极大似然估计为 $\hat{\theta} = \min_{1 \leq i \leq n} x_i$。

四．(本题 10 分) 对某种合金材料的熔点作了四次测试，根据 4 次的测试数据算得样本均值为 $\bar{X}=1267$ (度)，修正样本标准差 $S^*=3.65$ (度)．设合金材料的熔点服从正态分布，在显著性水平 $\alpha=5\%$ 下：

(1) 能否认为该种合金的熔点符合厂家所公布的 1260 度？
(2) 能否认为该种合金熔点的标准差不超过 2 度？

解：1) 要检验的假设为 $H_0:\mu=1260$，$H_1:\mu\neq 1260$

检验用的统计量 $T=\dfrac{\bar{X}-\mu_0}{S^*/\sqrt{n}}\sim t(n-1)$，

拒绝域为 $|T|\geq t_{1-\frac{\alpha}{2}}(n-1)=t_{0.975}(3)=3.1824$.

$|T|=\dfrac{1267-1260}{3.65/\sqrt{4}}=3.836>3.1824$，落在拒绝域内，

故拒绝原假设 H_0，即不能认为结果符合公布的数字 1260°C.

(2) 要检验的假设为 $H_0:\sigma \le 2$，$H_1:\sigma > 2$

检验用的统计量 $\chi^2 = \dfrac{(n-1)S^{*2}}{\sigma_0^2} \sim \chi^2(n-1)$，

拒绝域 $\chi^2 > \chi^2_{1-\alpha}(n-1) = \chi^2_{0.95}(3) = 7.815$

$\chi^2 = 40/4 = 10 > 7.815$，落在拒绝域内，故拒绝原假设 H_0

即不能认为测定值的标准差不超过 2^0C.

五. (本题 10 分) 把一枚硬币连抛 100 次, 结果出现了 40 次正面向上,60 次反面向上在显著性水平显著性水平 $\alpha = 5\%$ 下, 能否认为这枚硬币是均匀的?

解: 假设硬币是均匀的, 令 X=0 表示反面向上,否则,X=1, 即:

$$H_0 : X \sim \begin{bmatrix} 0 & 1 \\ 0.5 & 0.5 \end{bmatrix}$$

$$\chi^2 = \frac{1}{n}\sum_{i=1}^{r}\frac{n_k^2}{p_k} - n \sim \chi^2(r-1); \quad \chi^2 = \frac{1}{100}\left(\frac{60^2}{0.5} + \frac{40^2}{0.5}\right) - 100 = 4$$

$\chi^2 = 4 > \chi^2_{1-\alpha}(r-1) = 3.841$, 故拒绝原假设,认为该硬币不均匀.

六.（本题 14 分）抽查 6 家企业，根据产量 x_i（台）与单位成本 y_i（万元）的统计数据得：

$$\sum x_i = 360,\quad \sum x_i^2 = 25000,\quad \sum y_i = 55,\quad \sum y_i^2 = 565,\quad \sum x_i y_i = 2860$$

（1）求单位成本与产量的相关系数；

（2）求单位成本关于产量的回归方程；

（3）求线性回归的残差平方和 SS_e 及估计的标准差 $\hat{\sigma}$；

(4) 在显著性水平 $\alpha = 0.05$ 下检验单位成本与产量是否有线性相关关系.

解：1） $\bar{x} = 60,$ $\bar{y} = 9.1667,$ $L_{xx} = \sum x_i^2 - n\bar{x}^2 = 3400.$

$$L_{yy} = 60.8333 \qquad L_{xy} = \sum x_i y_i - n\bar{x}\ \bar{y} = -440,$$

$$r = \frac{L_{xy}}{\sqrt{L_{xx} L_{yy}}} = -0.9675$$

2) $y = \beta_0 + \beta_1 x + \varepsilon, \qquad \varepsilon \sim N(0, \sigma)$

$$\hat{\beta}_1 = \frac{L_{xy}}{L_{xx}} = -0.1294,$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 16.931$$

回归方程为 y = 16.931-0.1294x

3) $SS_e = L_{yy} - \hat{\beta}_1 L_{xy} = 3.8973$

$$\hat{\sigma} = \sqrt{\frac{SS_e}{n-2}} = 0.9871$$

4) $H_0 : \beta_1 = 0$

$$T = \frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}} \sqrt{L_{xx}} \sim t(n-2)$$

$$T = \frac{-0.1294 - 0}{0.9871} \sqrt{3400} = -7.6439$$

$$| T | > t_{0.975}(4) = 2.7764$$

故拒绝原假设,即认为单位成本与产量有统计的线性相关关系.

七. （本题 10 分）为了研究一天中的不同工作时间对工作效率的影响，随机抽取 12 人，等分成三组，A 组做早班，B 组做晚班，C 组做夜班，分别记录他们完成同一种工作的完工时间，数据如下：

组别	完	工	时	间
A 早班	5.2	5.6	5.8	5.4
B 晚班	5.4	4.9	6.1	6.6
C 夜班	6.1	5.8	5.9	7.2

试利用方差分析的方法，在显著性水平 $\alpha = 0.05$ 下分析不同的班次对工作效率是否有显著性影响？

解：方差分析的前提是：假设不同班次的完工时间服从正态分布，且方差相等，即 $\xi_i \sim N(\mu_i, \sigma^2)$， i=1, 2, 3.

检验班次对工作效率是否有影响，相当于检验： $H_0 : \mu_1 = \mu_2 = \mu_3$

方差分析：单因素方差分析

组	计数	求和	平均	方差	
行 1	4	22	5.5	0.066667	
行 2	4	23	5.75	0.563333	
行 3	4	25	6.25	0.416667	

方差分析

差异源	SS	df	MS	F	F crit
组间	1.166667	2	0.583333	1.671975	4.256492
组内	3.14	9	0.348889		
总计	4.306667	11			

F < F crit = 4.26，故 接受原假设，即在显著性水平 0.05 下认为不同的班次对工作效率无显著性影响.

谢 谢 你 的 配 合

祝 你 考 出 好 成 绩

By K. Zhu