判 别 分 析

Discriminant Analysis

银行贷款

- 银行需要决定是否同意申请者贷款

倾向违约组

按时还款组

§1 判别分析的基本思想

- **基本思想**

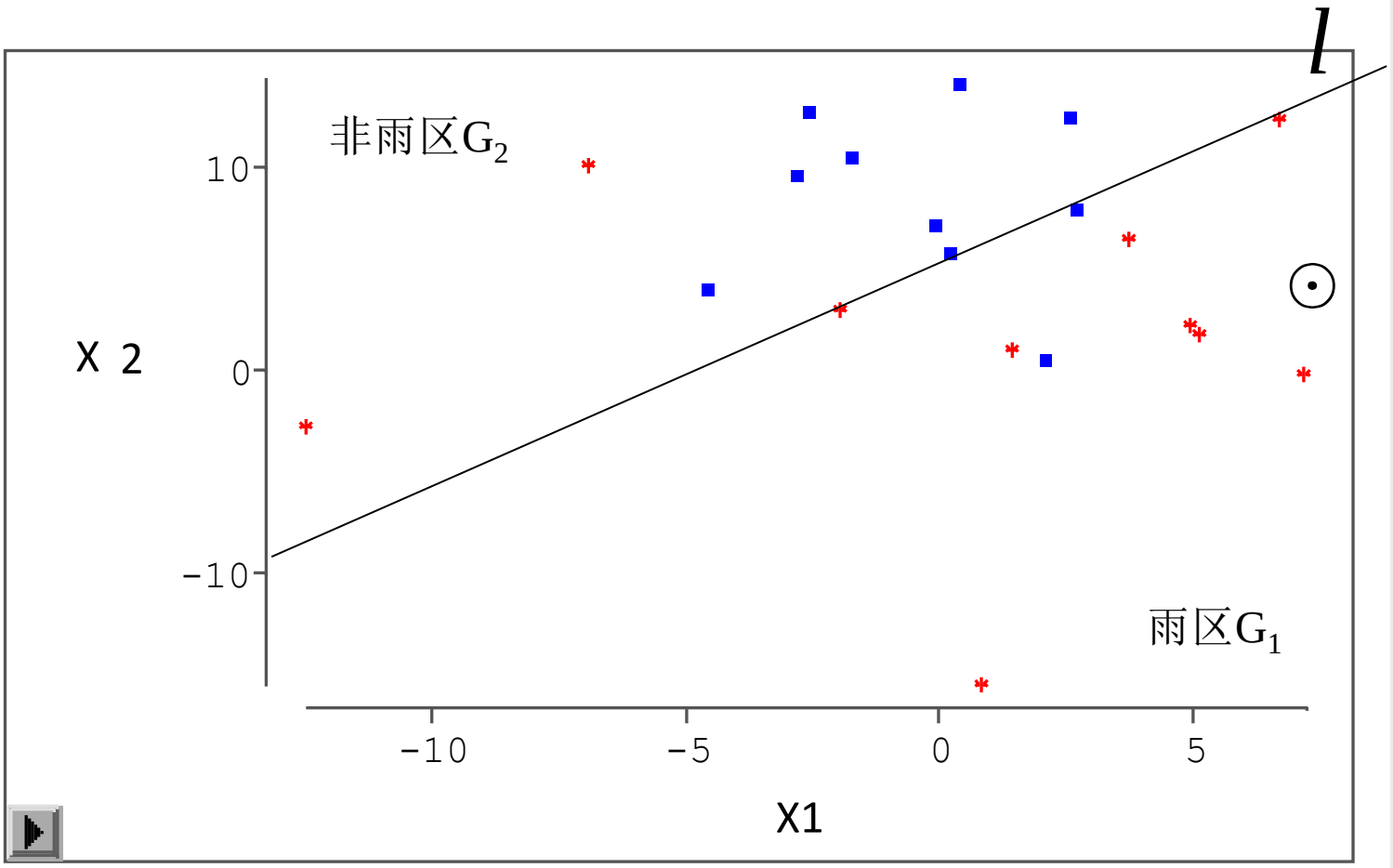
  根据**已知类别**的样本所提供的信息，总结出分类的规律性，建立**判别公式**和**判别准则**，判别新的样本点所属类型，是判别个体所属群体的一种统计方法。

根据经验，今天与昨天的湿度差及今天的压差（气压与温度之差）是预报明天下雨或不下雨的两个重要因素。今测得 X_1=8.1， X_2 =2.0，试问应预报明天下雨还是不下雨?

这个问题是两类判别问题，总体分为两类，用G_1表示下雨，G_2表示不下雨。为进行预报，应先收集一批资料，从已有的资料中找出规律，再作预报。

我们收集过去10个雨天和非雨天X_1和X_2的数值

雨天		非雨天	
X_1	X_2	X_1	X_2
-1.9	3.2	0.2	6.2
-6.9	10.4	-0.1	7.5
5.2	2.0	0.4	14.6
5.0	2.5	2.7	8.3
7.3	0.0	2.1	0.8
6.8	12.7	-4.6	4.3
0.9	-15.4	-1.7	10.9
-12.5	-2.5	-2.6	13.1
1.5	1.3	2.6	12.8
3.8	6.8	-2.8	10.0


l
非雨区G2
雨区G1
X 2
X1
10
0
-10
-10
-5
0
5

判别分析与聚类分析的区别

- **判别分析**　已知研究对象分为若干个**类别**，并且已 经取得每一类别的一批观测数据，在此基础上**寻求出分类的规律性**，建立判别准则，然后对未知类别的样品进行判别分类。
- **聚类分析**　一批样品**划分为几类事先并不知道**，正需要通过聚类分析来给以确定类型。

判别分析与聚类分析的联系

- **聚类分析**
- 用不同的聚类方法可能得到不同的结果，保留共性的聚类结果；对于用不同方法归类不同的少数样品，再结合判别分析加以判断归类。
- **判别分析**　将共性的聚类结果，作为**已知类别**的样本的信息（训练样本），对未知类别的样品（测试样本）进行判别分类。

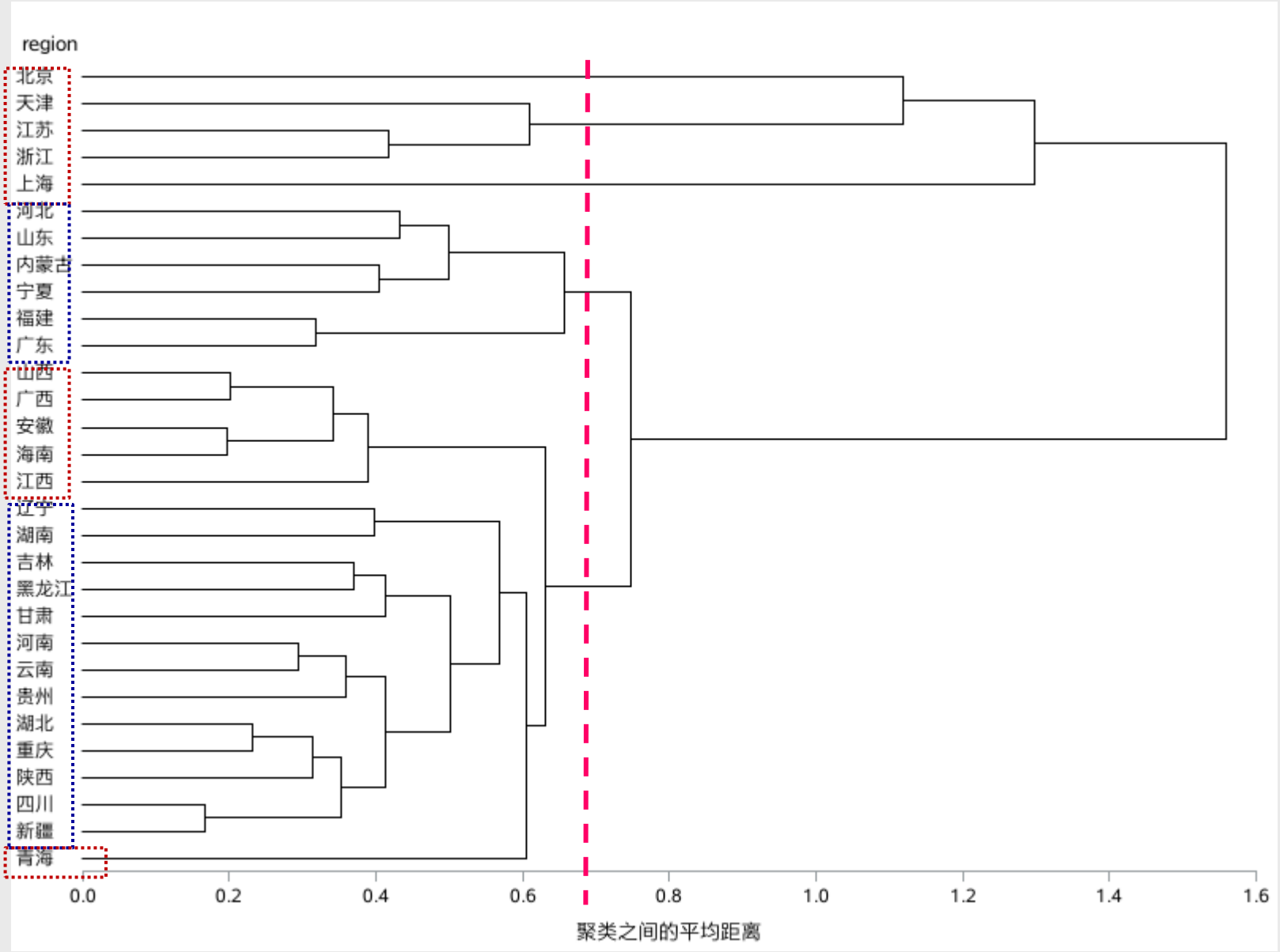

region
北京
天津
江苏
浙江
上海
河北
山东
内蒙古
宁夏
福建
广东
山西
广西
安徽
海南
江西
辽宁
湖南
吉林
黑龙江
甘肃
河南
云南
贵州
湖北
重庆
陕西
四川
新疆
青海
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
聚类之间的平均距离

图2 类平均法树形图

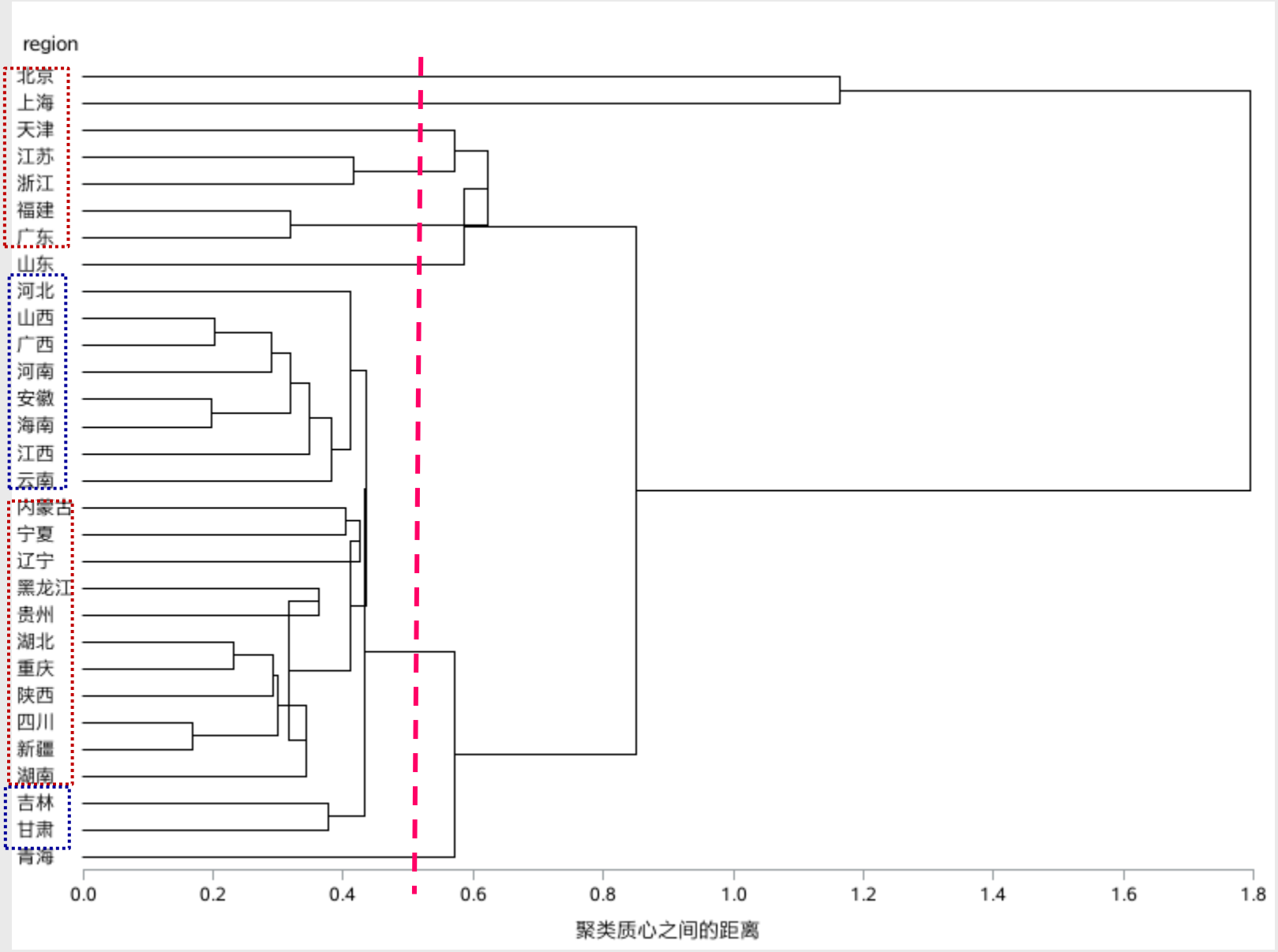

region
北京
上海
天津
江苏
浙江
福建
广东
山东
河北
山西
广西
河南
安徽
海南
江西
云南
内蒙古
宁夏
辽宁
黑龙江
贵州
湖北
重庆
陕西
四川
新疆
湖南
吉林
甘肃
青海
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
1.8
聚类质心之间的距离

图3 重心法树形图

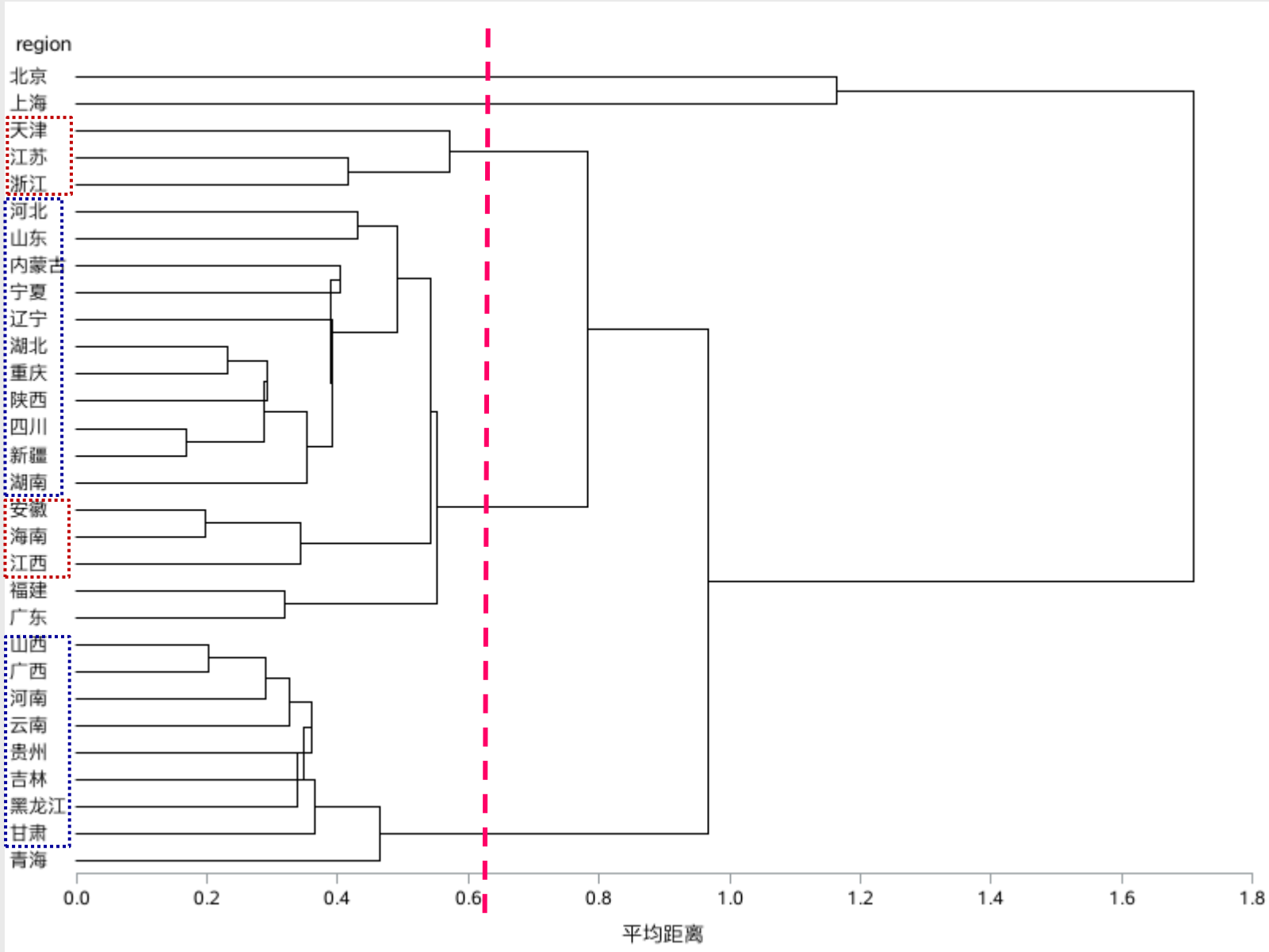

region
北京
上海
天津
江苏
浙江
河北
山东
内蒙古
宁夏
辽宁
湖北
重庆
陕西
四川
新疆
湖南
安徽
海南
江西
福建
广东
山西
广西
河南
云南
贵州
吉林
黑龙江
甘肃
青海
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
1.8
平均距离

图4 中间距离法树形图

输出2：树形图 (Ward法)

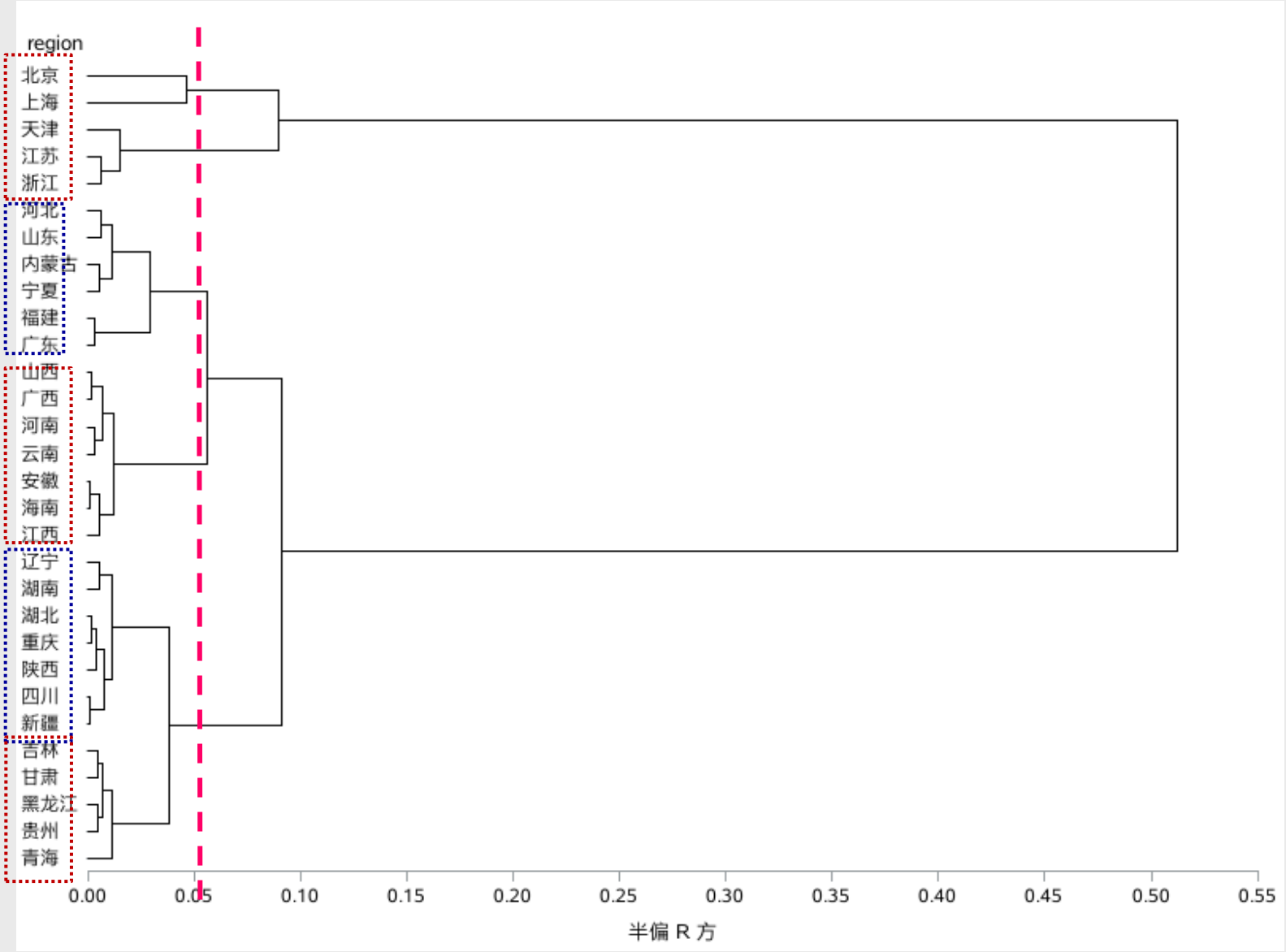


图5 Ward离差平方和法树形图

§2 距离判别

（一）距离判别法的基本思想

距离判别的最直观的想法：计算样品 x 到第 i 个类的距离 $d^2(x, G_i)$，哪个距离最小，就将它判归哪个总体。

所以，首先考虑，是否能够构造一个恰当的距离函数，通过计算样本点与某类别之间距离的大小，判别其所属类别。

判别分析中常用马氏距离

样品 x 和 G_i 类之间的马氏距离定义为 x 与 G_i 类重心间的距离：

$$d^2(x, G_i) = (x - \mu_i)' \Sigma^{-1} (x - \mu_i) \qquad i = 1, 2, \cdots, k$$

马氏距离不受变量间的相关性和量纲的影响

（二）总体协差阵相等时

1. 多总体判别

设有个K总体 $G_i\ (i = 1, 2, \cdots, k)$，分别有均值向量

μ_i(i=1,2,…,k) 和协方差矩阵 $\Sigma_i = \Sigma$。

又设 x 是一个待判样品。则 x 与G_i的距离为（即判别函数）

$$d^2(x, G_i) = (x - \mu_i)'\Sigma^{-1}(x - \mu_i)$$

$$= x'\Sigma^{-1}x - 2x'\Sigma^{-1}\mu_i + \mu_i'\Sigma^{-1}\mu_i$$

上式中的第一项 $x'\Sigma^{-1}x$ 与 i 无关，则舍去，得一个等价的函数

$$f_i(x) = -2x'\Sigma^{-1}\mu_i + \mu_i'\Sigma^{-1}\mu_i$$

将上式中提-2，得

$$f_i(x) = -2(x'\Sigma^{-1}\mu_i - \frac{1}{2}\mu_i'\Sigma^{-1}\mu_i)$$

则距离判别法的**判别函数**为：

$$f_i(x) = (x'\Sigma^{-1}\mu_i - \frac{1}{2}\mu_i'\Sigma^{-1}\mu_i)$$

判别规则为

$$f_l(x) = \max_{1\le i\le k} f_i(x), \quad 则 x \in G_l$$

2. **两总体**判别

设有两个协差阵Σ相同的p维正态总体，对给定的样品 X ，判别一个样品X 到底是来自哪一个总体。故我们用马氏距离来给定判别规则，有：

$$\begin{cases} x \in G_1, & \text{如} d^2(x, G_1) < d^2(x, G_2), \\ x \in G_2, & \text{如} d^2(x, G_2) < d^2(x, G_1) \\ \text{待判}, & \text{如} d^2(x, G_1) = d^2(x, G_2) \end{cases}$$

$$f_1(x) - f_2(x)$$

$$= x'\Sigma^{-1}\mu_1 - \frac{1}{2}\mu_1'\Sigma^{-1}\mu_1 - x'\Sigma^{-1}\mu_2 + \frac{1}{2}\mu_2'\Sigma^{-1}\mu_2$$

$$= x'\Sigma^{-1}(\mu_1 - \mu_2) - \frac{1}{2}(\mu_1 + \mu_2)'\Sigma^{-1}(\mu_1 - \mu_2)$$

$$= (x - \frac{\mu_1 + \mu_2}{2})'\Sigma^{-1}(\mu_1 - \mu_2)$$

$$\frac{1}{2}\mu_1'\Sigma^{-1}\mu_2$$

$$-\frac{1}{2}\mu_2'\Sigma^{-1}\mu_1$$

令 $\bar{\mu} = \frac{\mu_1 + \mu_2}{2}$ $\qquad \alpha = \Sigma^{-1}(\mu_1 - \mu_2) = (a_1, a_2, \cdots, a_p)'$

$$= (x - \bar{\mu})'\alpha$$

$$W(x) = (x - \frac{\mu_1 + \mu_2}{2})'\Sigma^{-1}(\mu_1 - \mu_2) = (x - \bar{\mu})'\alpha$$

$$= \alpha'(x - \bar{\mu})$$

$$= a_1(x_1 - \bar{\mu}_1) + \cdots + a_p(x_p - \bar{\mu}_p)$$

则前面的判别法则表示为 $f_1(x) - f_2(x)$

$$\begin{cases} x \in G_1, & \text{如} W(x) > 0, \\ x \in G_2, & \text{如} W(x) < 0。 \\ \text{待判}, & \text{如} W(x) = 0 \end{cases}$$

特别地，**当p=1时**，若两个总体分别为 $N(\mu_1,\sigma^2)$ 和 $N(\mu_2,\sigma^2)$

则判别函数为

$$W(x)=(x-\overline{\mu})\frac{1}{\sigma^2}(\mu_1-\mu_2)\ ,\quad \text{其中}\ \ \overline{\mu}=\frac{1}{2}(\mu_1+\mu_2)$$

不妨设 $\mu_1<\mu_2$

则 $W(x)$ 的符号取决于 $x>\overline{\mu}$ 还是 $x<\overline{\mu}$

因此判别规则可写成：

$$\begin{cases}\text{若}x<\overline{\mu}\ \ ，\text{则}x\in G_1\\ \text{若}x>\overline{\mu}\ \ ，\text{则}x\in G_2\end{cases}$$

我们看到用距离判别所得到的准则是颇为合理的，但用这个判别法有时会错判。如 x 来自 G_1 ，但却落入 D_2 ，

被判为属 G_2 ，错判的概率为图中阴影部分的面积，记为 $P(2/1)$ ，类似地有 $P(1/2)$

显然，

$$P(2/1)=1-\Phi\left(\frac{\bar{\mu}-\mu_1}{\sigma}\right)=1-\Phi\left(\frac{\mu_2-\mu_1}{2\sigma}\right)$$

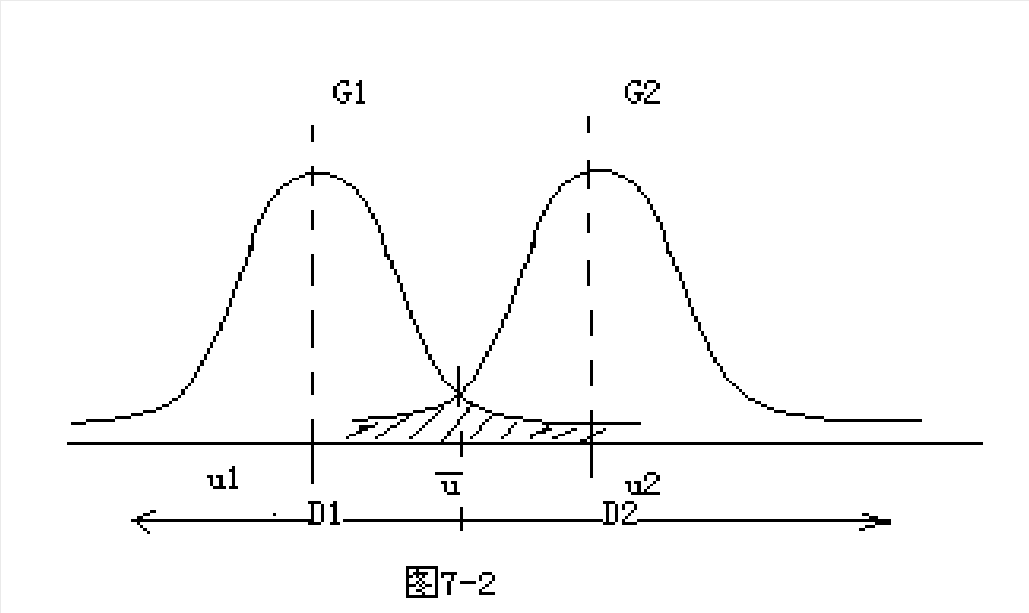


图7-2

。

例1 在企业的考核中，可以根据企业的生产经营情况把企业分为优秀企业和一般企业。考核企业经营状况的指标有：

资金利润率=利润总额/资金占用总额

劳动生产率=总产值/职工平均人数

产品净值率=净产值/总产值

三个指标的均值向量和协方差矩阵如下。

变量	均值向量		协方差矩阵		
	优秀	一般			
资金利润率	13.5	5.4	68.39	40.24	21.41
劳动生产率	40.7	29.8	40.24	54.58	11.67
产品净值率	10.7	6.2	21.41	11.67	7.90

现有二家企业，观测值分别为（7.8，39.1，9.6）和（8.1，34.2，6.9），问这两家企业应该属于哪一类？

线性判别函数：

$$W(x)=[x-\frac{(\mu_1+\mu_2)}{2}]'\Sigma^{-1}(\mu_1-\mu_2)$$

$$=(x_1-9.45 \quad x_2-35.25 \quad x_3-8.45)'\begin{pmatrix}68.39 & 40.24 & 21.41\\ 40.24 & 54.58 & 11.67\\ 21.41 & 11.67 & 7.90\end{pmatrix}^{-1}\begin{pmatrix}8.1\\ 10.9\\ 4.5\end{pmatrix}$$

$$=-0.60581x_1+0.25362x_2+1.83679x_3-18.7359$$

判别准则：

$$\begin{cases}x\in G_1, & \text{如}W(x)>0,\\ x\in G_2, & \text{如}W(x)<0。\\ \text{待判}, & \text{如}W(x)=0\end{cases}$$

$$y_1 = -0.60581 \times 7.8 + 0.25362 \times 39.1 + 1.83679 \times 9.6 - 18.73596$$

$$= 4.0892 > 0$$ 故属于优秀企业

$$y_2 = -0.60581 \times 8.1 + 0.25362 \times 34.2 + 1.83679 \times 6.9 - 18.73596$$

$$= -2.2956 < 0$$ 故属于一般企业

例2 两类判别在市场分析中的应用

某企业生产新式大衣，将新产品的样品分寄给九个城市百货公司的进货员，并附寄调查意见表征求对新产品的评价，评价分质量、款式、颜色三个方面，以十分制评分。结果五位喜欢，四位不喜欢。评价表如下：

		产品特性		
		x_1 质量	x_2款式	x_3颜色
喜欢组	1	8	9.5	7
	2	9	8.5	6
	3	7	8.0	9
	4	10	7.5	8.5
	5	8	6.5	7
不喜欢组	1	6	3	5.5
	2	3	4	3.5
	3	4	2	5
	4	3	5	4

（1）先求两类样本的均值

$$\bar{x}^{(1)}=\begin{pmatrix}8.4\\8.0\\7.5\end{pmatrix}\qquad \bar{x}^{(2)}=\begin{pmatrix}4.0\\3.5\\4.5\end{pmatrix}$$

$$\bar{x}^{(1)}-\bar{x}^{(2)}=\begin{pmatrix}4.4\\4.5\\3\end{pmatrix}\qquad \frac{\bar{x}^{(1)}+\bar{x}^{(2)}}{2}=\begin{pmatrix}6.2\\5.75\\6\end{pmatrix}$$

（2）计算样本协方差矩阵，从而求出 $\hat{\Sigma}$ 及 $\hat{\Sigma}^{-1}$

$$S_1=\frac{1}{4}\begin{pmatrix}5.2 & -0.5 & -1\\ -0.5 & 5 & -1.25\\ -1 & -1.25 & 6\end{pmatrix}=\begin{pmatrix}1.3 & & \\ -0.125 & 1.25 & \\ -0.25 & -0.3125 & 1.5\end{pmatrix}$$

$$S_2=\frac{1}{3}\begin{pmatrix}6 & -3 & 3.5\\ -3 & 5 & -2.5\\ 3.5 & -2.5 & 2.5\end{pmatrix}=\begin{pmatrix}2 & & \\ -1 & 1.667 & \\ 1.167 & -0.833 & 0.833\end{pmatrix}$$

$$\hat{\Sigma}=\frac{4S_1+3S_2}{5+4-2}=\frac{1}{7}\begin{pmatrix}11.2 & -3.5 & 2.5\\ -3.5 & 10 & -3.75\\ & & 8.5\end{pmatrix}$$

$$\hat{\Sigma}^{-1}=\begin{pmatrix}0.716 & 0.2056 & -0.1198\\ 0.2056 & 0.8978 & 0.3356\\ -0.1198 & 0.3356 & 1.0069\end{pmatrix}$$

（3） 求线性判别函数

$$W(x)=\left(\bar{x}^{(1)}-\bar{x}^{(2)}\right)'\hat{\Sigma}^{-1}\left(x-\frac{\bar{x}^{(1)}+\bar{x}^{(2)}}{2}\right)$$

$$=3.7166x_1+5.9520x_2+4.0048x_3-81.2956$$

（4）对已知类别的样品判别归类

对已知类别的样品（通常称为训练样本）用线性判别函数进行判别归类

样品	判别函数$W(x)$ 的值	原类号	判归类别
1	33.01	1	1
2	26.77	1	1
3	28.38	1	1
4	34.55	1	1
5	15.16	1	1
6	-19.11	2	2
7	-21.24	2	2
8	-32.32	2	2
9	-24.37	2	2

回代率为百之百，全部判对。

（5）对待判样品判别归类

如果有一潜在顾客，他对新产品的质量、款式、颜色的评价值为分别为6、8、8，则该顾客喜欢这款大衣吗？

$$W(x)=3.7166\times6+5.9520\times8+4.0048\times8-81.2956=20.66>0$$

故他属喜欢组

（三）总体协差阵不相等时

1. **多总体**判别

设有个K总体 $G_i\ (i = 1, 2, \cdots, k)$，分别有均值向量

μ_i(i=1,2,…,k) 和协方差阵Σ_i (i=1,2,…,k) 。

又设 x 是一个待判样品。则 x 与G_i的距离为（即判别函数）

$$d^2(x, G_i) = (x - \mu_i)'\Sigma_i^{-1}(x - \mu_i)$$

$$= x'\Sigma_i^{-1}x - 2x'\Sigma_i^{-1}\mu_i + \mu_i'\Sigma_i^{-1}\mu_i$$

2. **两总体**判别

$$\begin{cases} x \in G_1, & \text{如} d^2(x, G_1) < d^2(x, G_2), \\ x \in G_2, & \text{如} d^2(x, G_2) < d^2(x, G_1) \\ \text{待判}, & \text{如} d^2(x, G_1) = d^2(x, G_2) \end{cases}$$

$$W(x) = d^2(x, G_2) - d^2(x, G_1)$$

$$= (x - \mu_2)' \Sigma_2^{-1} (x - \mu_2) - (x - \mu_1)' \Sigma_1^{-1} (x - \mu_1)$$

判别准则：

$$\begin{cases} x \in G_1, & \text{如} W(x) > 0, \\ x \in G_2, & \text{如} W(x) < 0。 \\ \text{待判}, & \text{如} W(x) = 0 \end{cases}$$

特别地，当p=1时，若两个总体分别为$N\left(\mu_1,\sigma_1^2\right)$ 和 $N\left(\mu_2,\sigma_2^2\right)$ 则判别函数为

$$W(x)=\frac{(x-\mu_2)^2}{\sigma_2^2}-\frac{(x-\mu_1)^2}{\sigma_1^2}$$

当 $\mu_1<x<\mu_2$

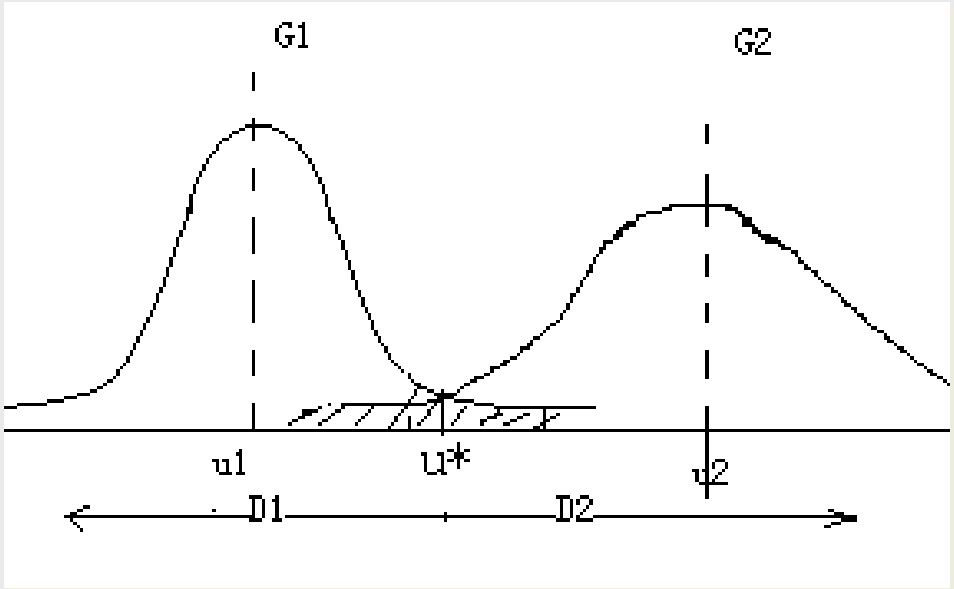
$$W(x)=\frac{\mu_2-x}{\sigma_2}-\frac{x-\mu_1}{\sigma_1}=\frac{\sigma_1\mu_2+\sigma_2\mu_1-x(\sigma_1+\sigma_2)}{\sigma_1\sigma_2}$$

$$=-\frac{\sigma_1+\sigma_2}{\sigma_1\sigma_2}\left(x-\frac{\sigma_1\mu_2+\sigma_2\mu_1}{\sigma_1+\sigma_2}\right)$$

令$\mu^*=\frac{\sigma_1\mu_2+\sigma_2\mu_1}{\sigma_1+\sigma_2}$

判别规则：

$$\begin{cases}\text{当}x<\mu^*\text{时}\quad，\quad\text{则}x\in G_1\\ \text{当}x>\mu^*\text{时}\quad，\quad\text{则}x\in G_2\end{cases}$$



$$\mu^* - \mu_1 = \frac{\frac{\sigma_2\mu_1 + \sigma_1\mu_2}{\sigma_1 + \sigma_2} - \mu_1}{\sigma_1} = \frac{\mu_2 - \mu_1}{\sigma_1 + \sigma_2}$$

$$\mu_2 - \mu^* = \frac{\mu_2 - \frac{\sigma_2\mu_1 + \sigma_1\mu_2}{\sigma_1 + \sigma_2}}{\sigma_2} = \frac{\mu_2 - \mu_1}{\sigma_1 + \sigma_2}$$

μ^* 到两个总体的马氏距离相等

- 距离判别只要求知道总体的数字特征，不涉及总体的分布函数，当总体均值和协差阵未知时，用样本的均值和协方差矩阵来估计。
- 距离判别方法简单实用，但没有考虑到每个总体出现的机会大小，即先验概率，没有考虑错判的损失。
- **贝叶斯判别法**正是为了解决这两个问题提出的判别分析方法。

贝叶斯公式是一个我们熟知的公式

$$P(B_i \mid A) = \frac{P(A \mid B_i)P(B_i)}{\sum_i P(A \mid B_i)P(B_i)}$$

设有总体 $G_i\ (i = 1, 2, \cdots, k)$ ，G_i 具有概率密度函数 $f_i(x)$。根据以往经验，知道 G_i 出现的概率为 q_i 。

即当样本 x_0 发生时，求他属于某类的概率。由贝叶斯公式计算后验概率，有：

$$P(G_i \mid x_0) = \frac{q_i f_i(x_0)}{\sum_i q_i f_i(x_0)}$$

若 $P(G_l \mid x_0) = \dfrac{q_l f_l(x_0)}{\sum_i q_i f_i(x_0)} = \max\limits_{1 \le i \le k} \dfrac{q_i f_i(x_0)}{\sum_i q_i f_i(x_0)}$

则 x_0 判给 G_l

$$q_l f_l(x_0) = \max_{1\le i\le k} q_i f_i(x_0), \quad \text{则 } x_0 \text{ 判给 } G_l$$

特别地，当总体服从正态分布时

$$若 f_i(x) = \frac{1}{(2\pi|\Sigma_i|)^{1/2}} \exp[-\frac{1}{2}(x-\mu^{(i)})'\Sigma_i^{-1}(x-\mu^{(i)})]$$

$$则, q_i f_i(x) = q_i \frac{1}{(2\pi|\Sigma_i|)^{1/2}} \exp[-\frac{1}{2}(x-\mu^{(i)})'\Sigma_i^{-1}(x-\mu^{(i)})]$$

上式两边取对数并去掉与i无关的项，则等价的判别函数为：

$$z_i(x) = \ln(q_i f_i(\mathbf{x}))$$

$$= \ln q_i - \frac{1}{2}\ln|\Sigma_i| - \frac{1}{2}(x-\mu^{(i)})'\Sigma_i^{-1}(x-\mu^{(i)})$$

问题转化为若 $Z_l(x) = \max\limits_{1\le i\le k}[Z_i(x)]$，则判 $x \in G_l$ 。

当**协方差阵相等时** $\Sigma_1 = \cdots \Sigma_k = \Sigma$

则判别函数退化为

$$z_i(x) = \ln q_i - \frac{1}{2}(x-\mu^{(i)})'\Sigma^{-1}(x-\mu^{(i)})$$

当先验概率相等时，完全成为距离判别法

§3　费歇判别法

（一）费歇判别的基本思想

费歇判别的基本思想是**投影**，将k组p维数据投影到某一个方向，使其投影的组与组之间尽可能地分开。

Fisher判别法由**Fisher**在1936年提出，是根据方差分析的思想建立起来的一种能较好区分各个总体的线性判别法，该判别方法对总体的分布不做任何要求。

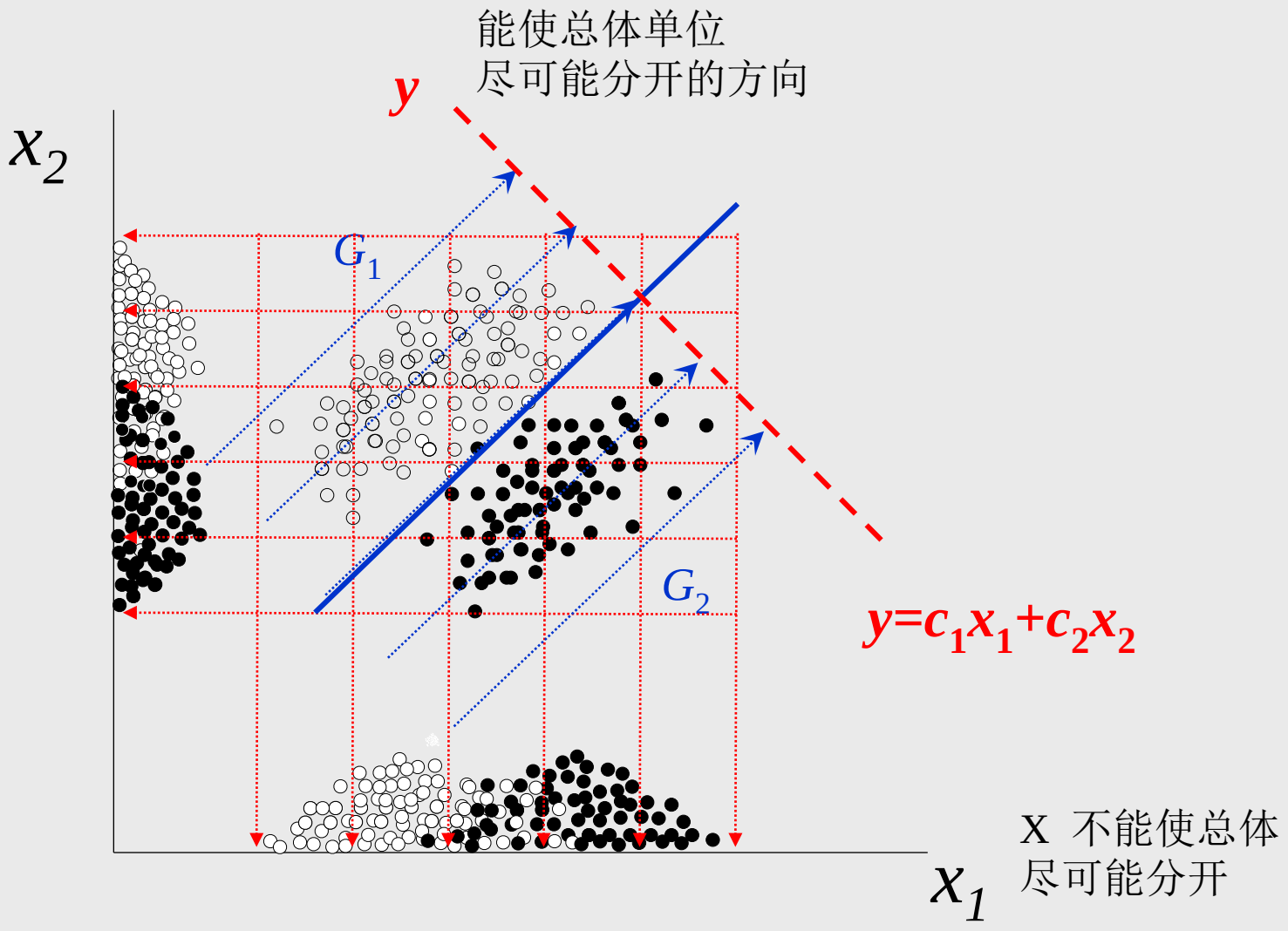
（二）两总体的费歇（Fisher）判别法

从两个总体中抽取具有p个指标的样品观测数据，借助于方差分析的思想构造一个线性判别函数：

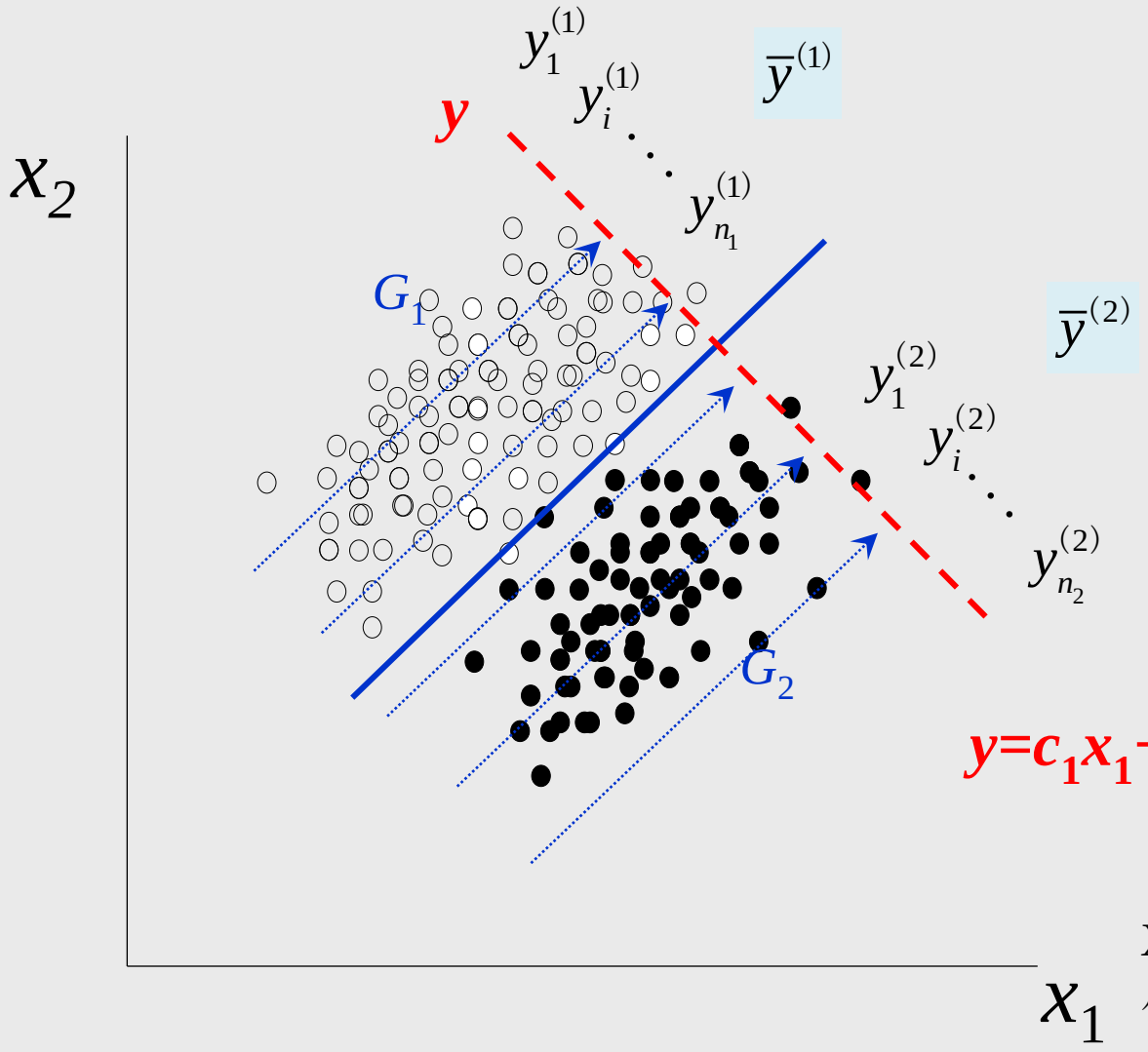
$$y = c_1 x_1 + c_2 x_2 + \cdots + c_p x_p$$

旋转坐标轴至 y 方向，使两总体尽可能分开，此时y 即为分类变量

两类*Fisher*判别示意图



两类*Fisher*判别示意图



x 不能使总体尽可能分开

系数 $c_1, c_2, \cdots, c_p$ 确定的**原则**

使组间离差平方和最大，而组内离差平方和最小。

假设我们可以得到一个线性判别函数：

$$y = c_1 x_1 + c_2 x_2 + \cdots + c_p x_p$$

我们把两个总体的样品数据代入上面的判别式

$$y_i^{(1)} = c_1 x_{i1}^{(1)} + c_2 x_{i2}^{(1)} + \cdots + c_p x_{ip}^{(1)} \qquad i = 1,2,\cdots,n_1$$

$$y_i^{(2)} = c_1 x_{i1}^{(2)} + c_2 x_{i2}^{(2)} + \cdots + c_p x_{ip}^{(2)} \qquad i = 1,2,\cdots,n_2$$

$$\bar{y}^{(1)} = \frac{1}{n_1}\sum_{i=1}^{n_1} y_i^{(1)} \qquad \bar{y}^{(1)} = c_1 \bar{x}_1^{(1)} + c_2 \bar{x}_2^{(1)} + \cdots + c_p \bar{x}_p^{(1)} = \sum_{k=1}^{p} c_k \bar{x}_k^{(1)}$$

$$\bar{y}^{(2)} = \frac{1}{n_2}\sum_{i=1}^{n_2} y_i^{(2)} \qquad \bar{y}^{(2)} = c_1 \bar{x}_1^{(2)} + c_2 \bar{x}_2^{(2)} + \cdots + c_p \bar{x}_p^{(2)} = \sum_{k=1}^{p} c_k \bar{x}_k^{(2)}$$

为了使判别函数能够很好地区分来自不同总体G_1和G_2的样品，自然希望：

(1）$\bar{y}^{(1)}$和 $\bar{y}^{(2)}$ 的差异越大越好

（2）来自同一总体的各个样品之间的差异越小越好。

即 $y_i^{(1)}(i=1,2,\cdots,n_1)$的离差平方和 $\sum_{i=1}^{n_1}\left(y_i^{(1)}-\bar{y}^{(1)}\right)^2$ 越小越好

即 $y_i^{(2)}(i=1,2,\cdots,n_2)$的离差平方和 $\sum_{i=1}^{n_2}\left(y_i^{(2)}-\bar{y}^{(2)}\right)^2$ 越小越好

$$Q=\left(\bar{y}^{(1)}-\bar{y}^{(2)}\right)^2 \quad \to \quad \max$$

$$R=\sum_{i=1}^{n_1}\left(y_i^{(1)}-\bar{y}^{(1)}\right)^2+\sum_{i=1}^{n_2}\left(y_i^{(2)}-\bar{y}^{(2)}\right)^2 \quad \to \quad \min$$

$$I=\frac{Q}{R} \quad \to \quad \max$$

$$I = \frac{Q}{R} \to \max$$

$$\ell nI = \ell nQ - \ell nR \to \max$$

令 $\dfrac{\partial \ell nI}{\partial c_k} = 0 \quad (k = 1,2,\cdots, p)$

由于 $\dfrac{\partial \ell nI}{\partial c_k} = \dfrac{1}{Q}\dfrac{\partial Q}{\partial c_k} - \dfrac{1}{R}\dfrac{\partial R}{\partial c_k} = 0$

故 $\dfrac{1}{I}\dfrac{\partial Q}{\partial c_k} = \dfrac{\partial R}{\partial c_k}$

$$Q=\left(\bar{y}^{(1)}-\bar{y}^{(2)}\right)^2=\left[\sum_{k=1}^{p}c_k\left(\bar{x}_k^{(1)}-\bar{x}_k^{(2)}\right)\right]^2\triangleq\left(\sum_{k=1}^{p}c_k d_k\right)^2$$

其中 $d_k=\bar{x}_k^{(1)}-\bar{x}_k^{(2)}\qquad k=1,2,\cdots,p$

即 $\begin{pmatrix}d_1\\d_2\\\vdots\\d_p\end{pmatrix}=\begin{pmatrix}\bar{x}_1^{(1)}-\bar{x}_1^{(2)}\\\bar{x}_2^{(1)}-\bar{x}_2^{(2)}\\\vdots\\\bar{x}_p^{(1)}-\bar{x}_p^{(2)}\end{pmatrix}=\left(\bar{x}^{(1)}-\bar{x}^{(2)}\right)$

为两类总体的样本均值差

$$\frac{\partial Q}{\partial c_k}=2\left(\sum_{l=1}^{p}c_l d_l\right)d_k$$

$$R = \sum_{i=1}^{n_1}\left(y_i^{(1)} - \bar{y}^{(1)}\right)^2 + \sum_{i=1}^{n_2}\left(y_i^{(2)} - \bar{y}^{(2)}\right)^2$$

$$= \sum_{i=1}^{n_1}\left[\sum_{k=1}^{p} c_k\left(x_{ik}^{(1)} - \bar{x}_k^{(1)}\right)\right]^2 + \sum_{i=1}^{n_2}\left[\sum_{k=1}^{p} c_k\left(x_{ik}^{(2)} - \bar{x}_k^{(2)}\right)\right]^2$$

$$= \sum_{i=1}^{n_1}\left[\sum_{k=1}^{p} c_k\left(x_{ik}^{(1)} - \bar{x}_k^{(1)}\right)\cdot\sum_{l=1}^{p} c_l\left(x_{il}^{(1)} - \bar{x}_l^{(1)}\right)\right]$$

$$+ \sum_{i=1}^{n_2}\left[\sum_{k=1}^{p} c_k\left(x_{ik}^{(2)} - \bar{x}_k^{(2)}\right)\cdot\sum_{l=1}^{p} c_l\left(x_{il}^{(2)} - \bar{x}_l^{(2)}\right)\right]$$

$$= \sum_{k=1}^{p}\sum_{l=1}^{p} c_k c_l \left[\sum_{i=1}^{n_1}\left(x_{ik}^{(1)} - \bar{x}_k^{(1)}\right)\left(x_{il}^{(1)} - \bar{x}_l^{(1)}\right) + \sum_{i=1}^{n_2}\left(x_{ik}^{(2)} - \bar{x}_k^{(2)}\right)\left(x_{il}^{(2)} - \bar{x}_l^{(2)}\right)\right]$$

$$\hat{=} \sum_{k=1}^{p}\sum_{l=1}^{p} c_k c_l s_{kl}$$

其中

$$s_{kl}=\sum_{i=1}^{n_1}\left(x_{ik}^{(1)}-\overline{x}_k^{(1)}\right)\left(x_{il}^{(1)}-\overline{x}_l^{(1)}\right)+\sum_{i=1}^{n_2}\left(x_{ik}^{(2)}-\overline{x}_k^{(2)}\right)\left(x_{il}^{(2)}-\overline{x}_l^{(2)}\right)$$

$$\frac{\partial R}{\partial c_k}=2\left(\sum_{l=1}^{p}c_l s_{kl}\right)=2c_1 s_{k1}+2c_2 s_{k2}+\cdots+2c_p s_{kp}$$

$$(k=1,2,\cdots,p)$$

$$\frac{2}{I}\left(\sum_{l=1}^{p}c_l d_l\right)d_k=2\sum_{l=1}^{p}c_l s_{kl}\qquad k=1,2,\cdots,p$$

令 $\beta=\frac{1}{I}\sum_{l=1}^{p}c_l d_l$

β 是常数因子，不依赖于 k

它对方程组只起共同扩大倍数的作用，不影响判别结果，不妨取 $\beta = 1$

于是得到 $\sum_{l=1}^{p} c_l s_{kl} = d_k \qquad k = 1,2,\cdots,p$

$$s_{k1}c_1 + s_{k2}c_2 + \cdots + s_{kp}c_p = d_k \qquad k = 1,2,\cdots,p$$

$$\begin{cases} s_{11}c_1 + s_{12}c_2 + \cdots + s_{1p}c_p = d_1 \\ s_{21}c_1 + s_{22}c_2 + \cdots + s_{2p}c_p = d_2 \\ \cdots \\ s_{p1}c_1 + s_{p2}c_2 + \cdots + s_{pp}c_p = d_p \end{cases}$$

用矩阵表示：

$$\begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ s_{p1} & s_{p2} & \cdots & s_{pp} \end{pmatrix} \cdot \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_p \end{pmatrix} = \begin{pmatrix} d_1 \\ d_2 \\ \vdots \\ d_p \end{pmatrix}$$

因此得到

$$\begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_p \end{pmatrix} = \begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ s_{p1} & s_{p2} & \cdots & s_{pp} \end{pmatrix}^{-1} \cdot \begin{pmatrix} d_1 \\ d_2 \\ \vdots \\ d_p \end{pmatrix}$$

$$\begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ \vdots \\ c_p \end{bmatrix} = \begin{bmatrix} s_{11} & s_{12} & \cdots & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & \cdots & s_{2p} \\ \vdots & \vdots & \cdots & \cdots & \vdots \\ \vdots & \vdots & & & \vdots \\ s_{p1} & s_{p2} & \cdots & \cdots & s_{pp} \end{bmatrix}^{-1} \begin{bmatrix} d_1 \\ d_2 \\ \vdots \\ \vdots \\ d_p \end{bmatrix} = E^{-1} \cdot d$$

$$\begin{bmatrix} d_1 \\ d_2 \\ \vdots \\ d_p \end{bmatrix} = \begin{bmatrix} \bar{x}_1^{(1)} - \bar{x}_1^{(2)} \\ \bar{x}_2^{(1)} - \bar{x}_2^{(2)} \\ \vdots \\ \bar{x}_p^{(1)} - \bar{x}_p^{(2)} \end{bmatrix}$$

两总体的
积差阵之和

称 $y(x) = c_1 x_1 + \cdots + c_p x_p$ 为**判别函数**.

判别临界值

定义临界点为 $y_c = \begin{cases} \dfrac{\bar{y}^{(1)} + \bar{y}^{(2)}}{2} & \text{两总体方差相等时} \\ \dfrac{\hat{\sigma}_2 \bar{y}^{(1)} + \hat{\sigma}_1 \bar{y}^{(2)}}{\hat{\sigma}_2 + \hat{\sigma}_1} & \text{两总体方差不相等时} \end{cases}$

其中

$$\hat{\sigma}_1 = \sqrt{\frac{1}{n_1 - 1} \sum_{i=1}^{n_1} \left(y_i^{(1)} - \bar{y}^{(1)} \right)^2}$$

$$\hat{\sigma}_2 = \sqrt{\frac{1}{n_2 - 1} \sum_{i=1}^{n_2} \left(y_i^{(2)} - \bar{y}^{(2)} \right)^2}$$

判别准则

不妨假定$\bar{y}^{(1)} > \bar{y}^{(2)}$，则判别准则为：

$$\begin{cases} 若y(x) > y_c, 则x \in G_1 \\ 若y(x) < y_c, 则x \in G_2 \\ 若y(x) = y_c, 待\quad 判 \end{cases}$$

例 某外贸公司为推销某一新产品，为保险起见，在新产品大量上市前将该产品的样品寄往12个国家的进口代理商，并附意见调查表，要求对该产品给予评估，评估的因素有式样、包装及耐久性三项。评分表用10分制，最后要求说明是否愿意购买，调查结果如下：

		x_1	x_2	x_3
购买组	1	9	8	7
	2	7	6	6
	3	10	7	8
	4	8	4	5
	5	9	9	3
	6	8	6	7
	7	7	5	6
非购买组	1	4	4	4
	2	3	6	6
	3	6	3	3
	4	2	4	5
	5	1	2	2

第13个国家的进口代理商评分（9，5，8），问该代理商是否愿意购买此产品。

1.求两总体的样本均值

$$\bar{x}^{(1)}=\begin{pmatrix}\bar{x}_1^{(1)}\\ \bar{x}_3^{(1)}\\ \bar{x}_3^{(1)}\end{pmatrix}=\begin{pmatrix}8.29\\ 6.43\\ 6.00\end{pmatrix}\qquad \bar{x}^{(2)}=\begin{pmatrix}\bar{x}_1^{(2)}\\ \bar{x}_2^{(2)}\\ \bar{x}_3^{(2)}\end{pmatrix}=\begin{pmatrix}3.20\\ 3.80\\ 4.00\end{pmatrix}$$

2. 求两总体样本均值之差

$$d=\bar{x}^{(1)}-\bar{x}^{(2)}=\begin{pmatrix}5.09\\ 2.63\\ 2.00\end{pmatrix}$$

3. 求两总体的样本离差平方和矩阵E

先求各 s_{kl}

$$s_{11}=\sum_{i=1}^{7}\left(x_{i1}^{(1)}-\overline{x}_1^{(1)}\right)^2+\sum_{i=1}^{5}\left(x_{i1}^{(2)}-\overline{x}_1^{(2)}\right)^2=22.22857$$

$$s_{12}=\sum_{i=1}^{7}\left(x_{1i}^{(1)}-\overline{x_1^{(1)}}\right)\left(x_{2i}^{(1)}-\overline{x_2^{(1)}}\right)+\sum_{i=1}^{5}\left(x_{1i}^{(2)}-\overline{x}_1^{(2)}\right)\left(x_{2i}^{(2)}-\overline{x}_2^{(2)}\right)=8.34288$$

$$E=\begin{pmatrix}22.22857 & 8.34288 & 2\\ & 26.51427 & 6\\ & & 26\end{pmatrix}$$

$$E^{-1}=\begin{pmatrix}0.05101 & & \\ -0.016 & 0.04481 & \\ -0.00023 & -0.00911 & 0.04058\end{pmatrix}$$

4. 求判别系数

$$a=(c_1,c_2,c_3)'=E^{-1}d$$

$$a=\begin{pmatrix}c_1\\c_2\\c_3\end{pmatrix}=\begin{pmatrix}0.05101 & -0.016 & -0.00023\\-0.016 & 0.04481 & -0.00911\\-0.00023 & -0.00911 & 0.04058\end{pmatrix}\cdot\begin{pmatrix}d_1\\d_2\\d_3\end{pmatrix}=\begin{pmatrix}0.21692\\0.0182\\0.05604\end{pmatrix}$$

5. 得判别函数

$$y(x)=a'x$$

$$=0.21692x_1+0.0182x_2+0.05604x_3$$

6. $\bar{y}^{(1)} = a'\bar{x}^{(1)}$

$=0.21692\times 8.29+0.0182\times 6.43+0.05604\times 6=2.251533$

$\bar{y}^{(2)} = a'\bar{x}^{(2)}$ $=0.987464$

判别的临界值

$$Y_c = \frac{\bar{y}^{(1)} + \bar{y}^{(2)}}{2} = 1.62$$

则判别准则为：

$$\begin{cases} 若y(x) > Y_c, 则x \in G_1 \\ 若y(x) < Y_c, 则x \in G_2 \\ 若y(x) = Y_c, 待 \quad 判 \end{cases}$$

7. 对已知类别的样品判别分类

对已知类别的样品（通常成为训练样品）用线性判别函数进行判别归类，结果如下表：

样品	$W(x)$	原类号	判归类别
1	2.49	1	1
2	1.96	1	1
3	2.74	1	1
4	2.09	1	1
5	2.28	1	1
6	2.24	1	1
7	1.95	1	1
1	1.16	2	2
2	1.10	2	2
3	1.52	2	2
4	0.79	2	2
5	0.37	2	2

回代率为百之百，全部判对。

- 对判别类别的样品判别归类

x =（9,5,8），

$$y(x)= 0.21692\times 9+0.0182\times 5+0.05604\times 8=2.4916>Y_c$$

故 X 属购买组 G_1

（三）多总体的Fisher判别法

Fisher判别法实际上是致力于寻找一个最能反映组和组之间差异的**投影**方向，即寻找线性判别函数 $Y(x)=c_1x_1+\cdots+c_px_p$，设有$k$个总体$G_1,G_2,\cdots,G_k$，分别有均值向量$\bar{\mu}_1$，$\bar{\mu}_2$，…，$\bar{\mu}_k$和协方差阵$\Sigma_1,\cdots,\Sigma_k$，分别从各总体中得到样品：

$$X_1^{(1)},\cdots,X_{n_1}^{(1)}$$
$$X_1^{(2)},\cdots,X_{n_2}^{(2)}$$
$$\cdots$$
$$X_1^{(k)},\cdots,X_{n_k}^{(k)}$$

$$n_1 + n_2 + \cdots + n_k = n$$

第i个总体的样本均值向量 $\bar{X}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} X_j^{(i)}$

综合的样本均值向量 $\bar{X}_i = \frac{1}{n}\sum_{i=1}^{k} n_i \bar{X}_i$

第i个总体样本组内离差平方和

$$V_i = \sum_{j=1}^{n_i} (X_j^{(i)} - \bar{X}_i)(X_j^{(i)} - \bar{X}_i)'$$

综合的组内离差平方和

$$E = V_1 + V_2 + \cdots + V_k = \sum_{i=1}^{k}\sum_{j=1}^{n_i} (X_j^{(i)} - \bar{X}_i)(X_j^{(i)} - \bar{X}_i)'$$

组间离差平方和 $$B=\sum_{i=1}^{k}n_i(\bar{X}_i-\bar{X})(\bar{X}_i-\bar{X})'$$

因为$Y(x)=c_1x_1+\cdots+c_px_p$

$$V_{iy}=\sum_{i=1}^{n_i}(Y_t^{(i)}-\bar{Y}_i)^2=\sum_{t=1}^{n_i}(Y_t^{(i)}-\bar{Y}_i)(Y_t^{(i)}-\bar{Y})'=C'V_iC$$

$$E_0=\sum_{i=1}^{k}V_{iy}=\sum_{i=1}^{k}C'V_iC=C'EC$$

$$B_0=\sum_{i=1}^{k}n_i(\bar{Y}_i-\bar{Y})^2=\sum_{i=1}^{n_i}n_i(\bar{Y}_i-\bar{Y})(\bar{Y}_i-\bar{Y})'=C'BC$$

如果判别分析是有效的，则所有的样品的线性组合$Y(x)=c_1x_1+\cdots+c_px_p$ 满足**组内离差平方和小，**而**组间离差平方和大**。则

$$\Delta^2(C)=\frac{B_0}{E_0}=\frac{C'BC}{C'EC}=\max$$

$\Delta^2(C)=\dfrac{B_0}{E_0}=\dfrac{C'BC}{C'EC}$的最大值是$B$相对于$E$最大的特征根$\lambda_1$。

而λ_1 所对应的特征向量即 $C_1=(c_{11},\cdots,c_{p1})'$。

Fisher 判别函数是

$$\hat{Y}_1(x)=\hat{c}_{11}x_1+\cdots+\hat{c}_{p1}x_p$$

然而，如果组数k太大，讨论的指标太多，则一个判别函数是不够的，这时需要寻找第二个，甚至第三个线性判别函数

$\Delta^2(C)=\dfrac{B_0}{E_0}=\dfrac{C'BC}{C'EC}$的最大值是$B$相对于$E$第二大的特征根

其特征向量构成第二个判别函数的系数。

$$C_2=(c_{12},\cdots,c_{p2})'$$

$$\hat{Y}_2(x)=\hat{c}_{12}x_1+\cdots+\hat{c}_{p2}x_p$$

类推得到m(m<k)个线性函数。

关于需要几个判别函数的问题，需要累计判别效率达到85％以上，即有

$$\Delta^2(C)=\frac{B_0}{E_0}=\frac{C'BC}{C'EC}$$

设 $\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_p$ 为B相对于E的特征根，则

$$\frac{\lambda_1+\lambda_2+\cdots+\lambda_k}{\lambda_1+\lambda_2+\cdots+\lambda_p}\leq 85\%$$

投影（变换）

$$\begin{cases} y_1 = c_{11}x_1 + c_{21}x_2 + \cdots c_{p1}x_p \\ \cdots \\ y_m = c_{1m}x_1 + c_{2m}x_2 + \cdots c_{pm}x_p \end{cases}$$

- 将原来p个变量综合成m个新变量

以m个线性判别函数得到的函数值为新的变量，再进行距离判别。

判别规则：

设$Y_i(x)$为第i个线性判别函数，$(i=1,2,\cdots,m)$

$$d(x,G_k)=\sum_{i=1}^{m}(y_i(x)-y_i(\overline{x}_k))^2$$

$$d(x,G_t)=\min_{1\le j\le k} d(x,G_k) \qquad \text{则} \quad x\in G_t$$

例　ST和非ST企业的距离判别

- 自2013年12月31日起，新三板正式接受全国企业挂牌，使中小高新技术企业的融资问题得到有效缓解。但新三板企业尤其成长型企业仍处在快速成长的发展阶段，容易遇到各种各样的经营问题
- 2020年上半年，有70%的新三板企业盈利，有2600多家公司同比增长，有9%左右的公司实现扭亏为盈，盈利能力和经营质量提升；有4家公司亏损超过10亿元，整体企业发展形势并不乐观。
- 2018年，新三板市场新增ST (Special treatment)公司114家； 2019年新增ST 公司58家 。
- 根据现有的ST和非ST企业经营状况，总结判别规律，以判断其他公司所属类别，准确预测财务危机，从而有针对性地采取措施防范风险。

例　ST和非ST企业的距离判别

- 根据公司价值的一般理论，公司价值主要由盈利报酬能力、资产管理能力、股东回报能力、偿付债务能力、成长能力等诸多因素决定，这些能力与要素最终都将影响企业的财务绩效。

例　ST和非ST企业的距离判别

- 指标体系

财务能力	财务指标	解释	编号
盈利能力	总资产收益率	净利润/平均资产总额	x_1
	净资产收益率	净利润/平均净资产	x_2
	营业净利率	净利润/营业收入	x_3
营运能力	应收账款周转率	当期销售净收入/平均应收账款余额	x_4
	总资产周转率	总营业额/总资产值	x_5
	流动资产周转率	主营业务收入净额/平均流动资产总额	x_6
发展能力	主营业务收入增长率	（主营业务收入-上期主营业务收入）/上期主营业务收入	x_7
	R&D 强度	研发支出/主营业务收入	x_8
偿债能力	流动比率	流动资产合计/流动负债合计	x_9
	速动比率	（流动资产-存货）/流动负债	x_{10}
	现金比率	（货币资金+有价证券）/流动负债	x_{11}

表1 训练样本数据

企业编号	总资产收益率	净资产收益率	营业净利率	应收账款周转率	总资产周转率	流动资产周转率	主营业务收入增长率	R&D强度	流动比率	速动比率	现金比率	类别
企业1	-0.32	-1.58	-0.38	5.64	0.96	1.19	0.26	0.14	0.91	0.50	0.25	非ST
企业2	-0.09	-0.22	-0.10	1.90	0.90	1.00	-0.57	0.08	3.53	2.78	0.46	非ST
企业3	-0.17	-0.20	-0.18	4.79	1.13	1.84	-0.27	0.00	9.46	9.46	8.24	非ST
企业4	-0.02	-0.14	1.00	-0.47	-0.02	-0.06	-1.02	0.00	0.33	0.32	0.08	非ST
企业5	-0.15	-0.31	-0.15	2.19	1.04	1.46	-0.22	0.00	1.37	1.29	0.41	非ST
企业6	-0.15	-0.17	-0.23	11.33	0.71	0.86	-0.35	0.16	6.27	5.77	2.16	非ST
企业7	0.00	0.00	0.00	6.56	0.51	4.14	0.17	0.00	0.83	0.82	0.09	非ST
企业8	0.12	0.44	0.13	3.59	0.89	1.89	0.32	0.00	0.69	0.62	0.22	非ST
企业9	-0.04	-0.05	-0.04	17.55	0.99	1.19	0.39	0.19	11.36	10.76	8.78	非ST
企业10	0.04	0.08	0.01	15.77	3.31	4.64	0.40	0.00	1.47	1.22	0.16	非ST
企业11	-0.07	-0.08	-0.11	2.46	0.62	0.65	-0.36	0.09	4.72	4.40	0.45	非ST
企业12	-0.22	-0.30	-6.43	1.49	0.03	0.11	-0.91	0.86	1.02	0.13	0.02	非ST
企业13	-0.27	-0.28	-0.56	6.48	0.56	1.27	0.07	0.06	8.08	8.05	5.81	非ST
企业14	-0.15	-0.70	-0.53	10.31	0.25	1.16	0.23	0.07	0.55	0.54	0.30	非ST
企业15	0.02	0.02	0.07	3.69	0.25	0.26	0.16	0.00	15.03	1.63	0.12	非ST
企业16	-0.05	-0.10	-0.04	3.83	1.46	1.91	-0.02	0.05	1.83	1.52	0.33	非ST
企业17	0.08	0.10	0.10	7.08	0.78	1.03	0.25	0.03	5.37	3.46	2.11	非ST
企业18	0.02	0.02	0.03	2.17	0.69	1.01	-0.03	0.02	5.09	3.56	0.10	非ST
企业19	0.04	0.15	0.08	4.28	0.45	1.15	0.94	0.02	0.53	0.28	0.02	非ST
企业20	0.18	0.23	0.30	1.59	0.55	0.69	0.08	0.00	11.03	10.92	3.90	非ST
企业21	0.03	0.03	0.10	0.43	0.30	0.39	-0.40	0.14	28.21	26.95	0.92	非ST
企业48	-0.03	-0.03	-0.12	1.21	0.86	9.02	-0.91	0.00	0.28	0.23	0.01	ST
企业49	-0.51	-3.85	-0.91	2.54	0.85	1.20	-0.47	0.00	0.59	0.40	0.04	ST
企业50	-0.07	-0.23	-0.22	2.72	0.35	0.94	-0.38	0.00	0.70	0.57	0.01	ST

表2　判别样本数据

企业编号	总资产收益率	净资产收益率	营业净利率	应收账款周转率	总资产周转率	流动资产周转率	主营业务收入增长率	R&D强度	流动比率	速动比率	现金比率
企业51	-0.04	-0.08	-0.12	0.64	0.30	0.37	-0.21	0.03	1.94	1.70	0.58
企业52	0.42	0.62	0.74	1.76	0.46	0.46	0.73	0.00	2.61	1.50	0.08
企业53	0.07	0.29	0.07	1.91	0.84	0.92	0.53	0.02	1.23	0.84	0.06
企业54	0.01	0.02	0.01	2.61	1.03	1.22	0.27	0.02	2.08	1.69	0.52
企业55	0.16	0.20	0.24	11.40	0.68	1.20	0.19	0.02	3.50	3.10	0.33
企业56	0.20	0.24	0.20	52.60	0.88	2.11	0.32	0.00	2.08	1.28	1.01
企业57	0.10	0.11	2.38	0.19	0.04	0.05	0.00	0.30	7.38	6.58	0.03
企业58	0.14	0.24	0.27	2.02	0.42	0.84	0.01	0.00	2.87	2.53	0.40
企业59	-0.02	-0.02	-0.06	4.05	0.36	0.62	1.06	0.00	6.45	5.92	2.89
企业60	0.02	0.06	0.02	4.04	0.71	1.52	0.41	0.01	1.03	0.77	0.15
企业61	0.12	0.13	0.33	1.03	0.34	0.44	0.33	0.00	10.39	9.64	4.12
企业62	0.11	0.16	0.19	1.80	0.42	0.58	1.22	0.06	2.49	2.16	0.68
企业63	-0.11	-0.41	-0.16	9.39	0.57	0.58	0.48	0.00	1.20	0.34	0.06
企业64	0.04	0.04	0.09	8.25	0.30	0.43	1.06	0.00	5.50	5.48	5.13
企业65	0.23	0.30	0.14	10.57	1.61	2.01	0.51	0.02	4.99	3.46	0.24
企业66	0.01	0.01	0.01	4.35	0.87	1.37	0.26	0.04	4.87	3.83	2.23
企业95	-0.13	-0.16	-0.68	3.81	0.20	0.35	-0.56	0.00	2.72	2.40	0.05
企业96	-0.29	-0.62	-5.81	4.50	0.05	0.20	-0.49	0.00	0.40	0.12	0.01
企业97	-0.20	-1.83	-1.01	4.10	0.20	0.63	-0.02	0.00	0.62	0.54	0.02

例　ST和非ST企业的距离判别

训练样本 enterprise_classified

判别样本数据 enterprise_unclassified

- **SAS程序**

```
proc discrim data=tmp1. enterprise_classified listerr
testdata=tmp1.enterprise_unclassified out=classified_out
testout=unclassified_out  outstat=os  pool=yes;
class type;
var X1-X11;
run;
```

到 TYPE 的广义平方距离		
从 TYPE	ST	非ST
ST	0	6.59359
非ST	6.59359	0

以下对象的线性判别函数: TYPE			
变量	标签	ST	非ST
常数		-2.92421	-2.07285
X1	总资产收益率	-6.80562	1.51080
X2	净资产收益率	0.03062	-0.04370
X3	营业净利率	-0.56583	-0.08834
X4	应收账款周转率	0.01389	0.00481
X5	总资产周转率	0.56413	3.05703
X6	流动资产周转率	0.59758	-0.10940
X7	主营业务收入增长率	-1.25878	-0.60597
X8	RandD强度	-4.33727	7.92585
X9	流动比率	0.19914	0.42368
X10	速动比率	-0.08679	-0.23302
X11	现金比率	-0.14221	0.17651

线性判别函数为：

$$f_1(x) = -2.9242 - 6.8056x_1 + 0.0306x_2 - 0.5658x_3 + 0.0139x_4 + 0.5641x_5 + 0.5976x_6 - 1.2588x_7 - 4.3373x_8 + 0.1991x_9 - 0.0868x_{10} - 0.1422x_{11}$$

$$f_2(x) = -2.0729 + 1.5108x_1 - 0.0437x_2 - 0.0883x_3 + 0.0048x_4 + 3.057x_5 - 0.1094x_6 - 0.606x_7 + 7.9259x_8 + 0.4237x_9 - 0.233x_{10} + 0.1765x_{11}$$

判别规则为：

若$f_1(x) > f_2(x)$，企业属于ST类型

企业	原始类型	判别类型		$f_1(x)$	$f_2(x)$	ST	非ST
企业1	非ST	非ST		-0.0870	1.6352	0.1516	0.8484
企业2	非ST	非ST		-0.3431	2.3490	0.0635	0.9365
企业3	非ST	非ST		0.3648	4.3854	0.0176	0.9824
企业4	非ST	非ST		-2.1119	-1.5349	0.3596	0.6404
企业5	非ST	非ST		0.0698	1.2497	0.2351	0.7649
企业6	非ST	非ST		-0.4845	3.0008	0.0297	0.9703
企业7	非ST	ST	*	-0.2111	-0.8469	0.6538	0.3462
企业8	非ST	非ST		-2.4690	0.6115	0.0439	0.9561
企业9	非ST	非ST		-2.3610	5.9784	0.0002	0.9998
企业10	非ST	非ST		1.3038	7.7963	0.0015	0.9985
企业11	非ST	非ST		-1.0887	1.6586	0.0603	0.9397
企业12	非ST	非ST		-0.0746	6.0205	0.0022	0.9978
企业25	非ST	非ST		-2.9920	2.4705	0.0042	0.9958
企业26	非ST	非ST		2.0663	4.3643	0.0911	0.9089
企业35	ST	非ST	*	-0.9546	-0.7750	0.4552	0.5448
企业36	ST	非ST	*	-3.3488	-1.5570	0.1428	0.8572
企业46	ST	非ST	*	-1.2498	-0.5336	0.3282	0.6718
企业47	ST	ST		5.9749	0.3425	0.9964	0.0036
企业48	ST	ST		4.4228	0.1558	0.9862	0.0138

SAS 系统

DISCRIM 过程
以下校准数据的分类汇总: TMP2.ENTERPRISE_CLASSIFIED
使用以下项的重新替换汇总: 线性判别函数

分入“TYPE”的观测数和百分比			
从 TYPE	ST	非ST	合计
ST	16 80.00	4 20.00	20 100.00
非ST	1 3.33	29 96.67	30 100.00
合计	17 34.00	33 66.00	50 100.00
先验	0.5	0.5	

“TYPE”的出错数估计			
	ST	非ST	合计
比率	0.2000	0.0333	0.1167
先验	0.5000	0.5000	

- 两总体协方差矩阵相等吗？

例　ST和非ST企业的距离判别

- **SAS程序**

```
proc discrim data=tmp1. enterprise_classified listerr
testdata=tmp1.enterprise_unclassified out=classified_out
testout=unclassified_out  outstat=os  pool=no;
class type;
var X1-X11;
run;
```

SAS 系统

DISCRIM 过程
以下校准数据的分类汇总: TMP2.ENTERPRISE_CLASSIFIED
使用以下项的重新替换汇总: 二次判别函数

分入“TYPE”的观测数和百分比

从 TYPE	ST	非ST	合计
ST	20 100.00	0 0.00	20 100.00
非ST	5 16.67	25 83.33	30 100.00
合计	25 50.00	25 50.00	50 100.00
先验	0.5	0.5	

“TYPE”的出错数估计

	ST	非ST	合计
比率	0.0000	0.1667	0.0833
先验	0.5000	0.5000	

- “ST”与“非ST”企业比例相当吗？
- 先验概率取相等，合理吗？

例　ST和非ST企业的距离判别

- **SAS程序**

```
proc discrim data=tmp1. enterprise_classified listerr
testdata=tmp1.enterprise_unclassified out=classified_out
testout=unclassified_out  outstat=os  pool=no;
class type;
var X1-X11;
priors 'ST'=0.1 '非ST'=0.9;
run;
```

SAS 系统

DISCRIM 过程
以下校准数据的分类汇总: TMP2.ENTERPRISE_CLASSIFIED
使用以下项的重新替换汇总: 二次判别函数

分入“TYPE”的观测数和百分比			
从 TYPE	ST	非ST	合计
ST	19 95.00	1 5.00	20 100.00
非ST	2 6.67	28 93.33	30 100.00
合计	21 42.00	29 58.00	50 100.00
先验	0.1	0.9	

“TYPE”的出错数估计			
	ST	非ST	合计
比率	0.0500	0.0667	0.0650
先验	0.1000	0.9000	

- 在统计分析过程中，充分运用所学方法
- 尊重客观事实，不歪曲数据特征，实事求是，严谨求真
- 注重培养耐心细致的工作作风和严肃认真的科学精神。

案例分析：鸢尾花分类

鸢尾花是法国的国花，setosa，versico-lor，virginica是三个最有名的品种，三种花的外形非常像，但是可以通过花萼长、花萼宽、花瓣长、花瓣宽的不同来判别花属于哪一种类型。

表1　数据集前15条数据示例（文件名：flower）

序号	类别（A）	花萼长（X1）	花萼宽（X2）	花瓣长（X3）	花瓣宽（X4）
1	1	50	33	14	2
2	3	64	28	56	22
3	2	65	28	46	15
4	3	67	31	56	24
5	3	63	28	51	15
6	1	46	34	14	3
7	3	69	31	51	23
8	2	62	22	45	15
9	2	59	32	48	18
10	1	46	36	10	2
11	2	61	30	46	14
12	2	60	27	51	16
13	3	65	30	52	20
14	2	56	25	39	11
15	3	65	30	55	18

SAS 程序：费歇判别

```
proc candisc data=flower  out=flowerout;
    class a;
    var X1-X4;
    run;
```

SAS 输出：费歇判别

输出1：特征值及方差贡献率

特征值: Inv(E)*H = CanRsq/(1-CanRsq)			
特征值	差分	比例	累积
32.1919	31.9065	0.9912	0.9912
0.2854		0.0088	1.0000

SAS 输出：费歇判别

输出2：原始典型系数

原始典型系数			
变量	标签	Can1	Can2
X1	花萼长	-.0829377642	0.0024102149
X2	花萼宽	-.1534473068	0.2164521235
X3	花瓣长	0.2201211656	-.0931921210
X4	花瓣宽	0.2810460309	0.2839187853

由此可得中心化的费歇判别函数：

$$y_1 = -0.0829(x_1 - 58.433) - 0.1534(x_2 - 30.573) + 0.2201(x_3 - 37.580) + 0.2810(x_4 - 11.993)$$

$$y_2 = 0.0024(x_1 - 58.433) + 0.2165(x_2 - 30.573) - 0.0932(x_3 - 37.580) + 0.2839(x_4 - 11.993)$$

SAS 输出：费歇判别

输出3：典型变量上的类均值

典型变量分类均值		
A	Can1	Can2
1	-7.607599927	0.215133017
2	1.825049490	-0.727899622
3	5.782550437	0.512766605

各组的重心：

类别1：（-7.6076，0.2151）

类别2：（1.8250，-0.7279）

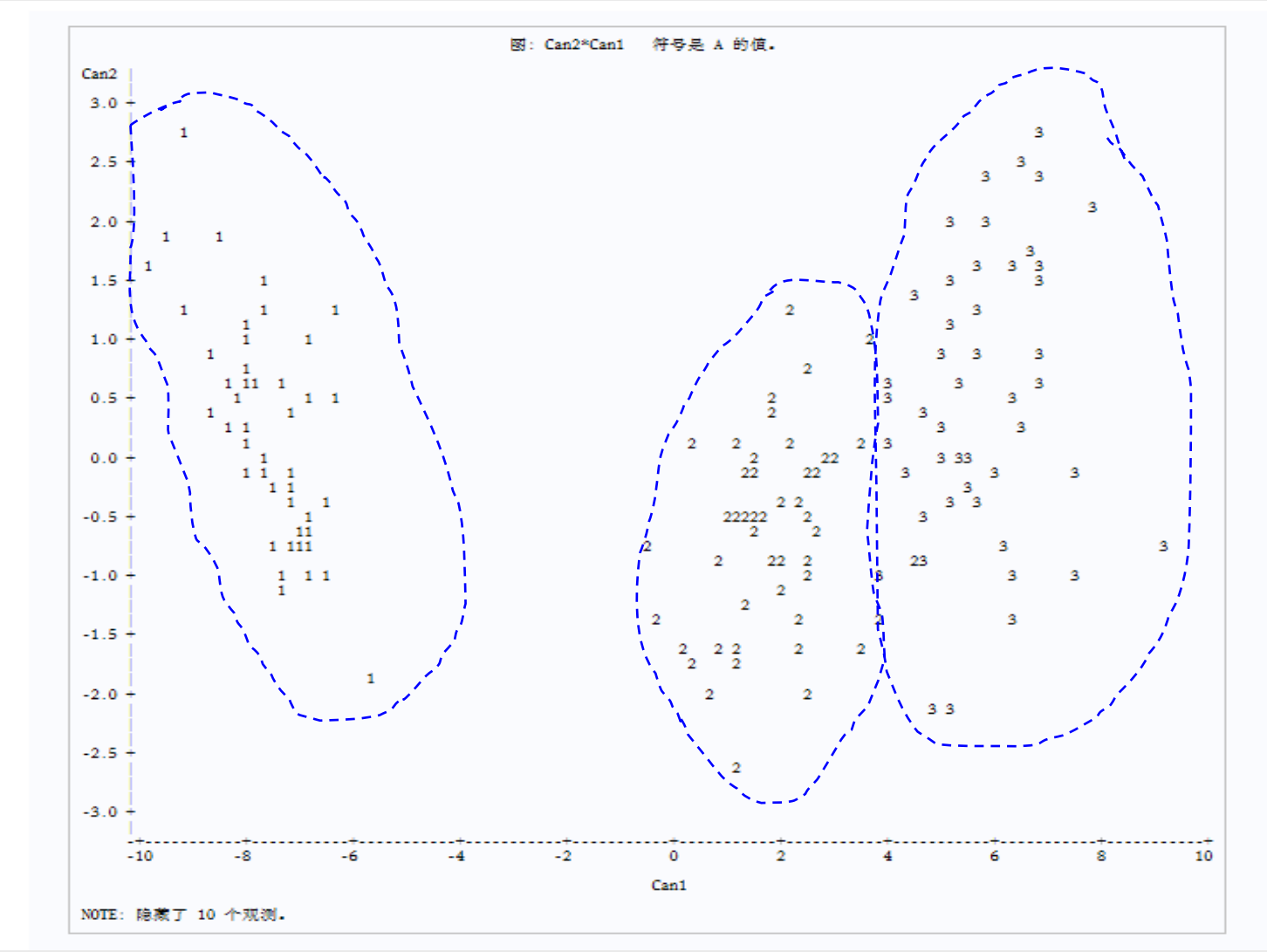
类别3：（5.7826，0.5128）

SAS 程序：费歇判别

```
proc candisc data=flower  out=flowerout;
    class a;
    var X1-X4;
    run;

proc plot data=flowerout;
    plot can2*can1=A;
    run;
```

SAS 输出：费歇判别


图：Can2*Can1　符号是 A 的值。
Can2
3.0
2.5
2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0
-2.5
-3.0
-10
-8
-6
-4
-2
0
2
4
6
8
10
Can1
NOTE：隐藏了 10 个观测。

从图中可以看出，三组分离的效果非常好，且分离的很大程度显示在can1上，这与一个判别函数解释的方差贡献率相符合。

因此，对于一个新的待判样本，通过计算判别得分，在图中标出坐标点，就可以判断出新样本点的所属类型。

SAS 程序 （距离判别）

Proc discrim data=SAS数据集 listerr testdata= SAS数据集
out= SAS数据集 testout= SAS数据集 outstat= SAS数据集
pool=yes/no;

线性判别函数，可缺省

二次判别函数

Class 变量名;　　DATA 中说明类别的变量

Var 变量名;　　分类根据的变量

Run;

DATA=	已分类的数据集
TESTDATA=	要分类的数据集
OUT=	已分类数据的回判结果
TESTOUT=	要分类数据的判别结果
OUTSTAT=	已分类数据的一些统计量

Listerr　列出被分到错误类别的观察值，可缺省

SAS 程序（Bayes判别）

Proc discrim data=SAS数据集 testdata= SAS数据集 out= SAS数据集 testout= SAS数据集 outstat= SAS数据集;

Class 变量名; DATA 中说明类别的变量

Priors 选项; 指定先验概率

Var 变量名; 分类根据的变量

Run;

Priors EQUAL; （可缺省，指先验概率相等）

Priors proportional | PROP; （要求用各类出现的比例计算各类的先验概率）

Priors '1'=**0.8** **'2'=0.2;** （指定具体的先验概率值）

SAS 程序 （Fisher 判别）

Proc candisc data=SAS数据集 out= SAS数据;

Class 变量名; DATA 中说明类别的变量

Var 变量名; 分类根据的变量

Run;

Proc plot;
plot can2*can1=变量名;
run;

画图：给出由典型判别分析得到的前两个典型变量的散点图，以直观地显示各类是否得以较好地区分开。

浙江工商大學
ZHEJIANG GONGSHANG UNIVERSITY

Thank You