第十三讲（第17章）　判别分析

Discriminant Analysis

本章概要

1) 概况
2) 基本概念
3) 回归分析和方差分析的关系
4) 判别分析模型
5) 与判别分析有关的统计量
6) 进行判别分析
   i. 定义问题
   ii. 估计判别函数系数
   iii. 确定判别函数的显著性
   iv. 解释结果
   v. 有效性

本章概要

7) 多组判别分析

i. 定义问题

ii. 估计判别函数系数

iii. 确定判别函数的显著性

iv. 解释结果

v. 有效性

8）逐步判别分析

9）因特网与计算机的应用

ANOVA、回归分析和判别分析的异同

	ANOVA	回归分析	判别分析
相同点			
因变量个数	一个	一个	一个
自变量个数	多个	多个	多个
不同点			
因变量性质	定量	定量	定类
自变量性质	定类	定量	定量

判别分析

判别分析适用于因变量（标准变量）为定类数据，自变量（预测变量）为定距数据的情形。

判别分析的目标:

- 建立能最大限度区分因变量的类别（组别）的判别函数，该函数是自变量的线性组合。
- 考察自变量的组间差异是否显著。
- 判断哪些自变量对组间差异的贡献较大。
- 根据自变量的取值将样本分类。
- 评估分类的准确程度。

判别分析

- 两组判别分析（**two-group discriminant analysis**）
- 多组判别分析（**multiple discriminant analysis**）

判别分析模型

判别分析模型涉及如下形式的线性组合：

$$D = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + \ldots + b_kX_k$$

where

D	=	判别分
b's	=	判别系数或权重
X's	=	自变量（预测变量）

- 估计系数的目的是使各组在判别函数值上的差异最大化。
- 同时保证组间平方和与组内平方和的判别分比率最大化。

判别分析中的有关统计量

- 典型相关系数（**Canonical correlation**），指判别分与组别的关联程度，是测量单个判别函数与代表组别的一组虚拟变量之间联系程度的指标。
- 重心（**Centroid**），指某一组判别分的均值。
- 分类矩阵（**Classification matrix**），该矩阵由正确分类和错误分类的样本数构成，对角线元素是正确分类数，非对角线元素是未正确分类数。

判别分析中的有关统计量

- 判别函数系数（**Discriminant function coefficients**），指判别分析模型中变量的乘数。
- 判别分（**Discriminant scores**），指判别分析模型的得分。
- 特征值（**Eigenvalue**），对于每个判别函数，特征值是组间平方和和组内平方和的比率。

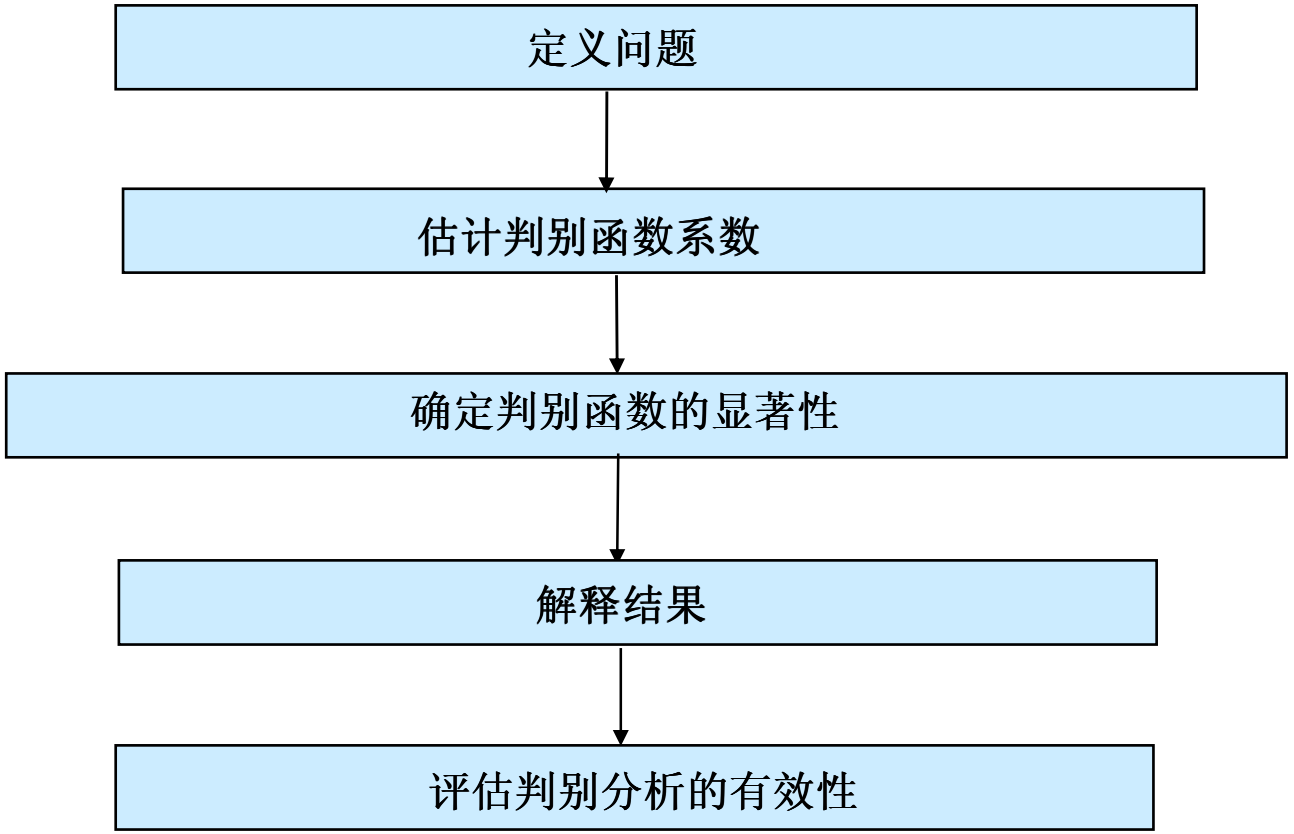
判别分析中的有关统计量

- ***F*值及其显著性（*F* values and their significance），分组变量作为ANOVA中的自变量，每个自变量轮流作为定量因变量。**
- 组均值和组标准差（**Group means and group standard deviations**），根据每个组中的每个自变量计算。
- 组内相关系数矩阵（**Pooled within-group correlation matrix**），根据所有组的各自的协方差矩阵平均计算出来。

判别分析中的有关统计量

- 标准化判别函数系数（**Standardized discriminant function coefficients**）
- 结构相关系数（**Structure correlations**），指自变量与判别函数之间的简单相关系数。
- 总相关系数矩阵（**Total correlation matrix**）， 如果样本都来自于单一样本并计算出相关系数，就得到总相关系数矩阵
- **Wilks'** λ，每个自变量的**Wilks'** λ就是组内平方和总平方和的比率， λ值越大，说明组均值无差异， λ值越小，说明组均值有差异。

判别分析过程



定义问题

- 确认分析目的、因变量和自变量。
- 因变量需包含两个或两个以上相互独立且覆盖所有情况的类别。
- 自变量应该根据理论模型或以前的研究加以选择，或者根据探索性研究以及经验来选择。
- 将样本分为两部分，一部分作为估计样本（**analysis sample**），用于估计判别函数；另一部分作为验证样本（**validation sample**），用于验证判别函数。
- 估计样本和验证样本的分布与总体样本分布相同。

关于度假的信息：估计样本

编号	是否旅游过	收入(千美元)	旅行	度假	家庭规模	年龄	家庭度假花费
1	1	50.2	5	8	3	43	M (2)
2	1	70.3	6	7	4	61	H (3)
3	1	62.9	7	5	6	52	H (3)
4	1	48.5	7	5	5	36	L (1)
5	1	52.7	6	6	4	55	H (3)
6	1	75.0	8	7	5	68	H (3)
7	1	46.2	5	3	3	62	M (2)
8	1	57.0	2	4	6	51	M (2)
9	1	64.1	7	5	4	57	H (3)
10	1	68.1	7	6	5	45	H (3)
11	1	73.4	6	7	5	44	H (3)
12	1	71.9	5	8	4	64	H (3)
13	1	56.2	1	8	6	54	M (2)
14	1	49.3	4	2	3	56	H (3)
15	1	62.0	5	6	2	58	H (3)

关于度假的信息：估计样本

编号	是否旅游过	收入(千美元)	旅行	度假	家庭规模	年龄	家庭度假花费
16	2	32.1	5	4	3	58	L (1)
17	2	36.2	4	3	2	55	L (1)
18	2	43.2	2	5	2	57	M (2)
19	2	50.4	5	2	4	37	M (2)
20	2	44.1	6	6	3	42	M (2)
21	2	38.3	6	6	2	45	L (1)
22	2	55.0	1	2	2	57	M (2)
23	2	46.1	3	5	3	51	L (1)
24	2	35.0	6	4	5	64	L (1)
25	2	37.3	2	7	4	54	L (1)
26	2	41.8	5	1	3	56	M (2)
27	2	57.0	8	3	2	36	M (2)
28	2	33.4	6	8	2	50	L (1)
29	2	37.5	3	2	3	48	L (1)
30	2	41.3	3	3	2	42	L (1)

关于度假信息：验证样本

编号	是否旅游过	收入(千美元)	旅行	度假	家庭规模	年龄	家庭度假花费
1	1	50.8	4	7	3	45	M(2)
2	1	63.6	7	4	7	55	H (3)
3	1	54.0	6	7	4	58	M(2)
4	1	45.0	5	4	3	60	M(2)
5	1	68.0	6	6	6	46	H (3)
6	1	62.1	5	6	3	56	H (3)
7	2	35.0	4	3	4	54	L (1)
8	2	49.6	5	3	5	39	L (1)
9	2	39.4	6	5	3	44	H (3)
10	2	37.0	2	6	5	51	L (1)
11	2	54.5	7	3	3	37	M(2)
12	2	38.2	2	2	3	49	L (1)

估计判别函数系数

- 直接法（ **direct method**），指将所有的自变量包括在方程中估计判别函数。
- 逐步判别分析（**stepwise discriminant analysis**），指根据自变量的的判别能力，逐步将其引入。

两组判别分析结果

组均值

是否旅游过	收入	旅行	度假	家庭规模	年龄
1	60.52000	5.40000	5.80000	4.33333	53.73333
2	41.91333	4.33333	4.06667	2.80000	50.13333
Total	51.21667	4.86667	4.9333	3.56667	51.93333

组标准差

1	9.83065	1.91982	1.82052	1.23443	8.77062
2	7.55115	1.95180	2.05171	.94112	8.27101
Total	12.79523	1.97804	2.09981	1.33089	8.57395

组内相关系数矩阵

	收入	旅行	度假	家庭规模	年龄
收入	1.00000				
旅行	0.19745	1.00000			
度假	0.09148	0.08434	1.00000		
家庭规模	0.08887	-0.01681	0.07046	1.00000	
年龄	- 0.01431	-0.19709	0.01742	-0.04301	1.00000

自由度为1和28的Wilks's λ（U统计量）和单变量F比率

变量	Wilks's λ	F	显著性
收入	0.45310	33.800	0.0000
旅行	0.92479	2.277	0.1425
度假	0.82377	5.990	0.0209
家庭规模	0.65672	14.640	0.0007
年龄	0.95441	1.338	0.2572

两组判别分析结果

典型判别函数

函数	特征值	方差百分比	累计百分比	典型相关系数
1*	1.7862	100.00	100.00	0.8007

函数后	Wilks's λ	卡方	自由度	显著性
0	0 .3589	26.130	5	0.0001

* 表示分析中第1个典型判别函数

标准典型判别函数系数

	函数 1
收入	0.74301
旅行	0.09611
度假	0.23329
家庭规模	0.46911
年龄	0.20922

结构矩阵

判别变量和典型判别函数之间的组内相关系数（变量次序依据函数相关系数排列）

	函数 1
收入	0.82202
家庭规模	0.54096
度假	0.34607
旅行	0.21337
年龄	0.16354

两组判别分析结果

非标准化典型判别函数系数

	函数 1
收入	0.8476710E-01
旅行	0.4964455E-01
度假	0.1202813
家庭规模	0.4273893
年龄	0.2454380E-01
(常数)	-7.975476

根据组均值估的典型判别系数（组重心）

组	函数 1
1	1.29118
2	-1.29118

分析样本的分类结果

	实际组	合计	预测组1	预测组2
组	1	15	12	3
			80.0%	20.0%
组	2	15	0	15
			0.0%	100.0%

原始分类的正确率是90.00%
交叉验证的正确分类比率是80.0%

两组判别分析结果

验证（保留）样本的分类结果

	实际组别	样本数	预测组1	预测组2
组	1	6	4 66.7%	2 33.3%
组	2	6	0 0.0%	6 100.0%

正确分类的样本百分比：83.33%.

确定判别函数的显著性

- 零假设：总体中各组所有判别函数的均值相等。
- 在**SPSS**中，可以通过将**Wilks′s** λ 转换成卡方进行检验
- 如果零假设被拒绝，说明判别函数是显著的。

解释结果

- 对判别系数的解释与多元回归分析类似。
- 如果自变量存在多重共线性，就无法明确测量自变量在判别组中的相对重要性。
- 因为存在这个问题，我们可以根据标准化判别函数系数的绝对值大小初步判断自变量的相对重要性。
- 通过结构相关系数，即自变量和判别函数之间的简单相关系数，也可以粗略判断自变量的相对重要性。
- 在解释判别分析结果还可以对每个组进行特征描述（**characteristic profile**），即计算每组自变量的均值。

评估判别分析的有效性

- **SPSS**提供了去掉一个样本的交叉验证。
- 命中率（**hit ratio**），是分类矩阵中对角线元素之和与总样本数的比例。
- 一些学者认为，判别分析的命中率应该比随机分类的命中率提高**25%**以上。

三组判别分析结果

组均值

数量	收入	旅行	度假	家庭规模	年龄
1	38.57000	4.50000	4.70000	3.10000	50.30000
2	50.11000	4.00000	4.20000	3.40000	49.50000
3	64.97000	6.10000	5.90000	4.20000	56.00000
Total	51.21667	4.86667	4.93333	3.56667	51.93333

组标准差

1	5.29718	1.71594	1.88856	1.19722	8.09732
2	6.00231	2.35702	2.48551	1.50555	9.25263
3	8.61434	1.19722	1.66333	1.13529	7.60117
Total	12.79523	1.97804	2.09981	1.33089	8.57395

组间样关系数矩阵

	收入	旅行	度假	家庭规模	年龄
收入	1.00000				
旅行	0.05120	1.00000			
度假	0.30681	0.03588	1.00000		
家庭规模	0.38050	0.00474	0.22080	1.00000	
收入	-0.20939	-0.34022	-0.01326	-0.02512	1.00000

三组判别分析结果

自由度为2和27的Wilks's λ（U统计量）和单变量F比率

变量	Wilks's λ	F	显著性
INCOME	0.26215	38.00	0.0000
TRAVEL	0.78790	3.634	0.0400
VACATION	0.88060	1.830	0.1797
HSIZE	0.87411	1.944	0.1626
AGE	0.88214	1.804	0.1840

典型判别函数

函数	特征值	方差%	累计%	典型相关系数
1*	3.8190	93.93	93.93	0.8902
2*	0.2469	6.07	100.00	0.4450
函数后	Wilks's λ	卡方	自由度	显著性
0	0.1664	44.831	10	0.00
1	0.8020	5.517	4	0.24

* 表示分析中剩余的两个典型判别函数

标准化典型判别函数系数

	函数 1	函数 2
收入	1.04740	-0.42076
旅行	0.33991	0.76851
度假	-0.14198	0.53354
家庭规模	-0.16317	0.12932
年龄	0.49474	0.52447

三组判别分析结果

结构矩阵

判别变量和典型判别函数之间的组内相关系数（变量跟据函数相关系数大小排列）

	函数 1	函数 2
收入	0.85556*	-0.27833
家庭规模	0.19319*	0.07749
度假	0.21935	0.58829*
旅行	0.14899	0.45362*
年龄	0.16576	0.34079*

非标准化典型判别函数系数

	函数 1	函数 2
收入	0.1542658	-0.6197148E-01
旅行	0.1867977	0.4223430
度假	-0.6952264E-01	0.2612652
家庭规模	-0.1265334	0.1002796
年龄	0.5928055E-01	0.6284206E-01
(常数)	-11.09442	-3.791600

根据判别函数均值估计（组重心）

组	函数 1	函数 2
1	-2.04100	0.41847
2	-0.40479	-0.65867
3	2.44578	0.24020

三组判别分析结果

分析样本的分类结果

实际组	合计	预测组1	预测组2	预测组3
组 1	10	9 90.0%	1 10.0%	0 0.0%
组 2	10	1 10.0%	9 90.0%	0 0.0%
组 3	10	0 0.0%	2 20.0%	8 80.0%

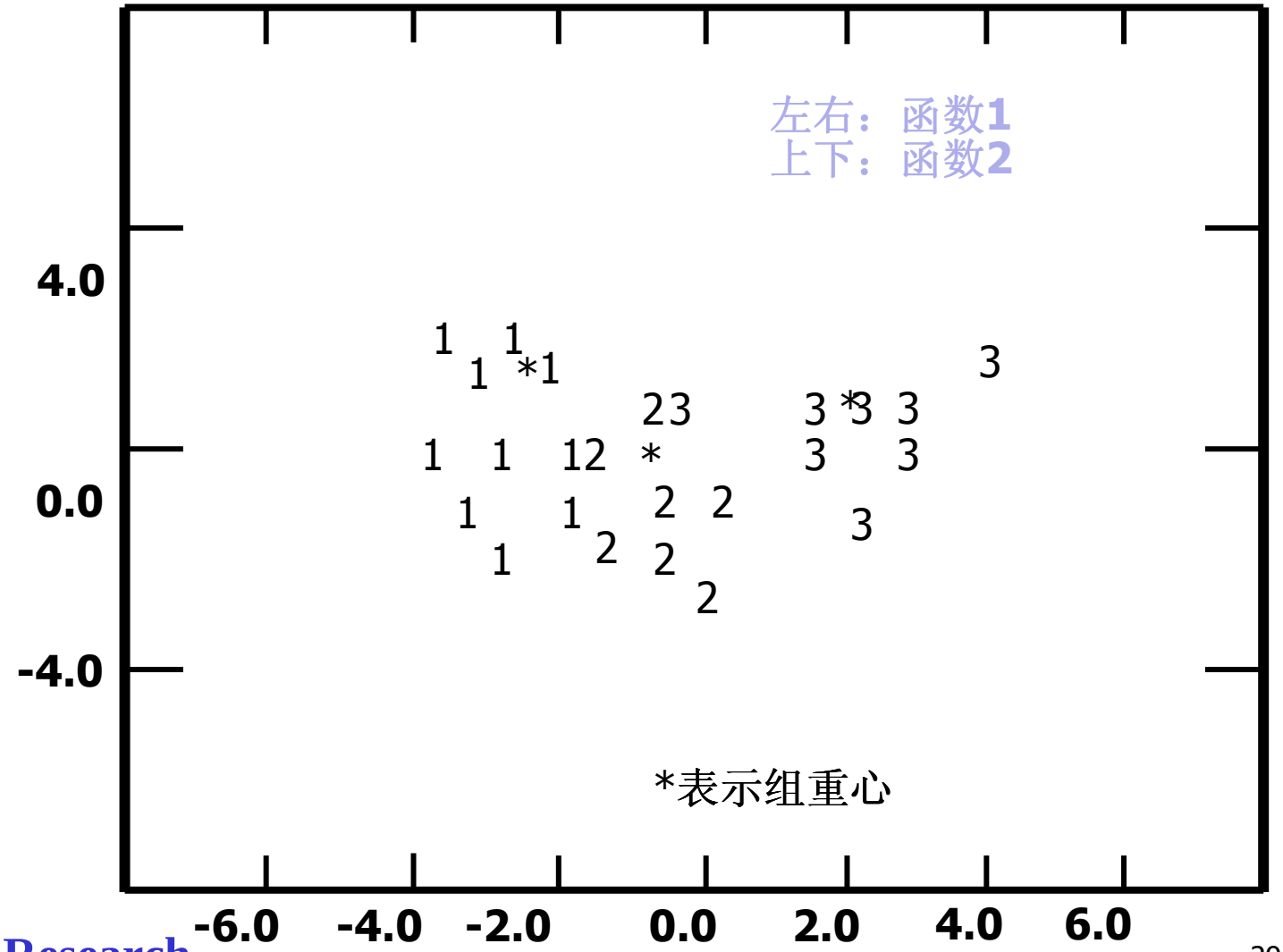
原始分类正确分组的比例是：86.67%

保留样本的分类结果

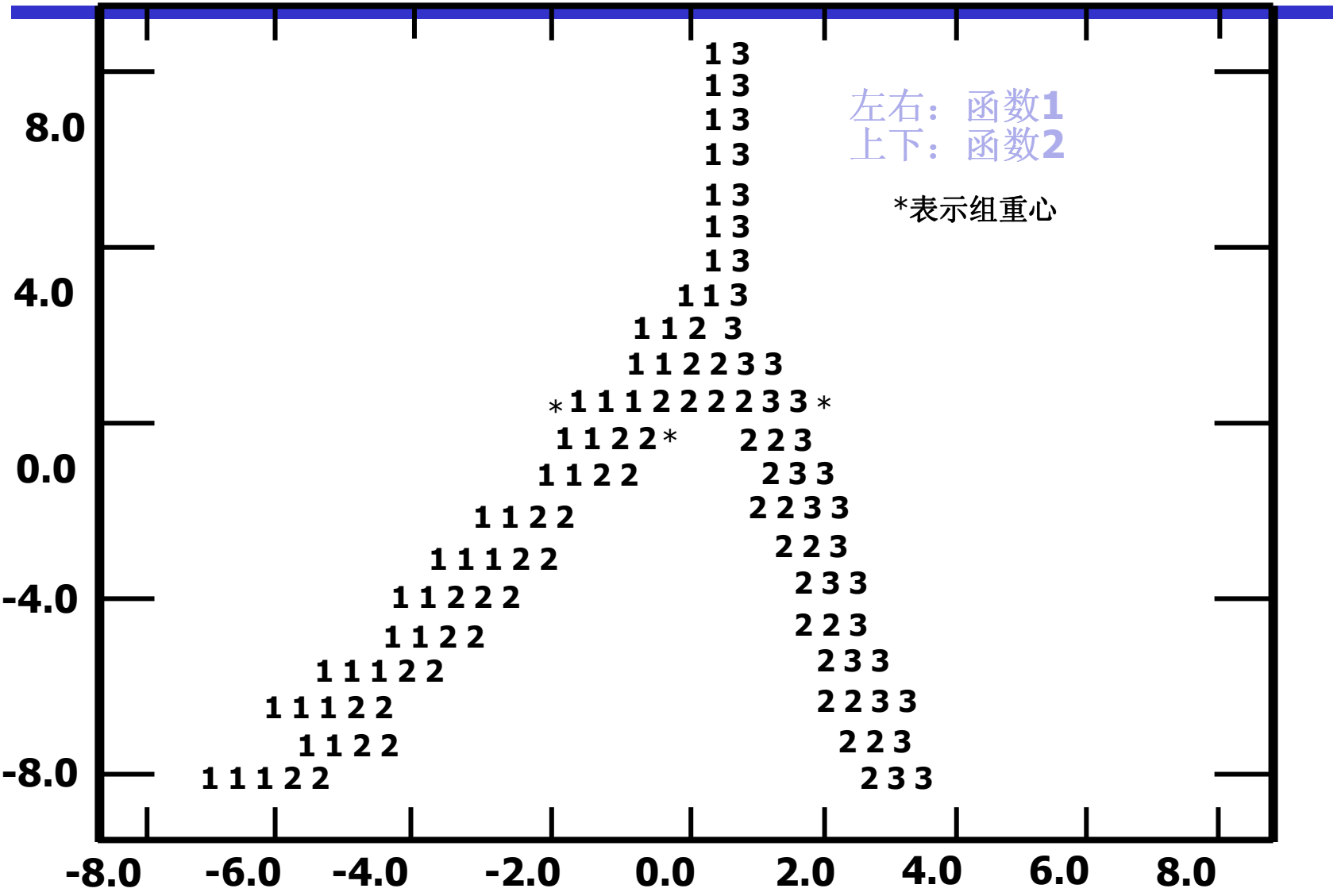
实际组	合计	预测组1	预测组2	预测组3
组 1	4	3 75.0%	1 25.0%	0 0.0%
组 2	4	0 0.0%	3 75.0%	1 25.0%
组 3	4	1 25.0%	0 0.0%	3 75.0%

正确分组的样本比例：75.00%

所有组的散点图



分组区域图



逐步判别分析

- 逐步判别分析与逐步多元回归很相似，都是根据自变量的判别能力逐步引入方程。
- 首先经组别作为定类变量，自变量作为因变量，进行单元方差分析，计算每个自变量的**F**比率。
- **F**比率最高的自变量最先进入判别方程。
- 第二个进入判别方程的自变量是根据调整的或偏**F**比率的高低来定。

逐步判别分析

- 每一个被选中的预测变量都需要根据其与其他入选预测变量的相关度计算保留价值
- 选择与保留的过程不断进行，直到将所有符合进入的显著性标准的变量全部引入判别方程
- 选择逐步判别过程的原因就在于使选择标准最优化，Mahalanobis方法的基础就在于使两组之间的广义距离最大化。
- 被选中的预测变量也表明了他们在组中辨别的重要性

SPSS 窗口

DISCRIMINANT 程序可以进行两组和多组的判别分析，要运行SPSS for Windows 的该程序，点击:

Analyze>Classify>Discriminant …