对应分析

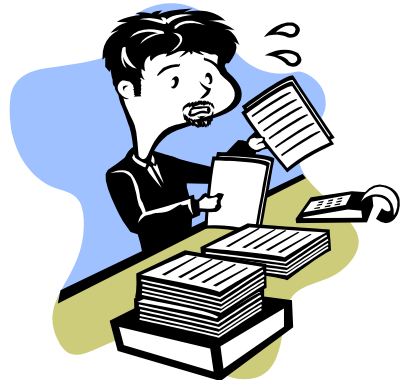
Correspondence Analysis

主要内容

- 对应分析的基本概念
- 对应分析方法的主要步骤
- 简单对应分析方法
- 多元对应分析方法
- 对应分析方法的应用实例
- 对应分析-R软件使用

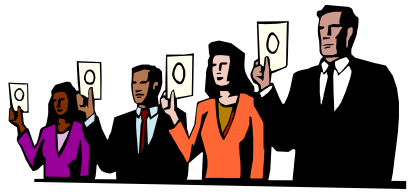
问题的提出

- 如何更加有效地探索数据？-多元统计分析
- 我们已经掌握的高级统计分析方法
  - 多元回归分析 （Multiple linear regression / logistic regression）
  - 主成分分析 (Principal Component Analysis)
  - 因子分析 (Factor Analysis)
  - 聚类分析 (Cluster Analysis)
  - 对数线性模型 (log-linear Model)
  - 方差分析 (ANOVA Analysis)
  - 判别分析 (Discriminant Analysis)
  - 典型相关分析(Canonical correlation analysis)
  - 结构方程式模型 (Structural Equation Modeling)
- 定量数据的处理方法——定性数据的处理？
- 多变量数据的图示化技术
  - 对维尺度偏好分析 (Multidimensional Preference Analysis)
  - 多维尺度分析 (Multidimensional Scaling MDS)
  - 对应分析 （Correspondence Analysis)



对应分析简介

- 一种新的多元统计分析技术。
- 一种主要分析**定性数据**（Category Data）方法。
- 一种强有力的数据图示化技术。
- 一种定性数据定量化分析的技术。
- 一种强有力的市场研究分析技术。



本课的目标

- 了解对应分析的基本概念
- 了解对应分析方法如何帮助探索数据的
- 分析列联表和卡方的独立性检验
- 解释对应图
- 对应分析对数据的格式要求
- 学会如何使用R软件作对应分析

什么是对应分析?

- 经常也称作 *Brand Mapping 或 CORAN Mapping*
  - Brand Mapping = Correspondence Analysis (usually)
- 相关性分析图
  - 一种非常有用的市场研究工具，可以表述一个市场的侧面（市场细分，品牌定位等）
    - 可以在2维空间内同时表达多维的属性
    - 可以更好的理解品牌和属性之间的关系

- 帮助客户/市场决策者
  - 为实施市场战略而去发现市场的空隙和优化产品的定位（对于新品牌或新产品的开发/延伸）
  - 发现市场上决定性的或显著的属性，例如对于选择不同品牌的重要和有显著区别的属性

什么是Brand Mapping?

对应分析的基本概念

- **对应分析**是一种数据分析技术，它能够帮助我们研究**由定性变量构成的交互汇总表**来揭示变量间的联系。
- 交互表的信息以图形的方式展示。
- 是强有力的探索数据技术，主要适用于有多个类别的定性变量。
- **可以揭示同一个变量的各个类别之间的差异，以及不同变量各个类别之间的对应关系。**
- 适用于两个或多个定性（分类）变量。

举例：不同性别-年龄段对不同户外运动的调研结果

运动种类	年龄段-性别	总计									
≤20M	≤30M	≤40M	≤50M	≤60M	≤20F	≤30F	≤40F	≤50F	≤60F		
乒乓球	1200	2050	2500	1600	950	1000	1800	2300	2200	1580	21108
自行车	350	900	650	170	50	45	90	100	50	30	3029
健身操（舞）	1500	2800	3900	3600	3000	200	600	1500	2000	1400	25290
徒步	100	200	300	200	250	30	150	400	700	500	3424
马拉松	500	900	910	500	300	30	70	50	40	10	4125
登高	200	400	300	500	300	150	300	350	400	390	4021
合计	4802	9046	10681	8197	6051	1814	3754	5863	6724	4065	60997

对应分析的商业应用

应用领域:

- **概念发展　(Concept Development)**
- **新产品开发　(New Product Development)**
- **市场细分　　(Market Segmentation)**
- **竞争分析　　(Competitive Analysis)**
- **广告研究　　(Advertisement Research)**

对应分析可以回答以下问题

- **谁是我的用户？**
- **还有谁是我的用户？**
- **谁是我竞争对手的用户？**
- **相对于我的竞争对手的产品，我的产品的定位如何？**
- **与竞争对手有何差异？**
- **我还应该开发哪些新产品？**
- **对于我的新产品，我应该将目标指向哪些消费者？**

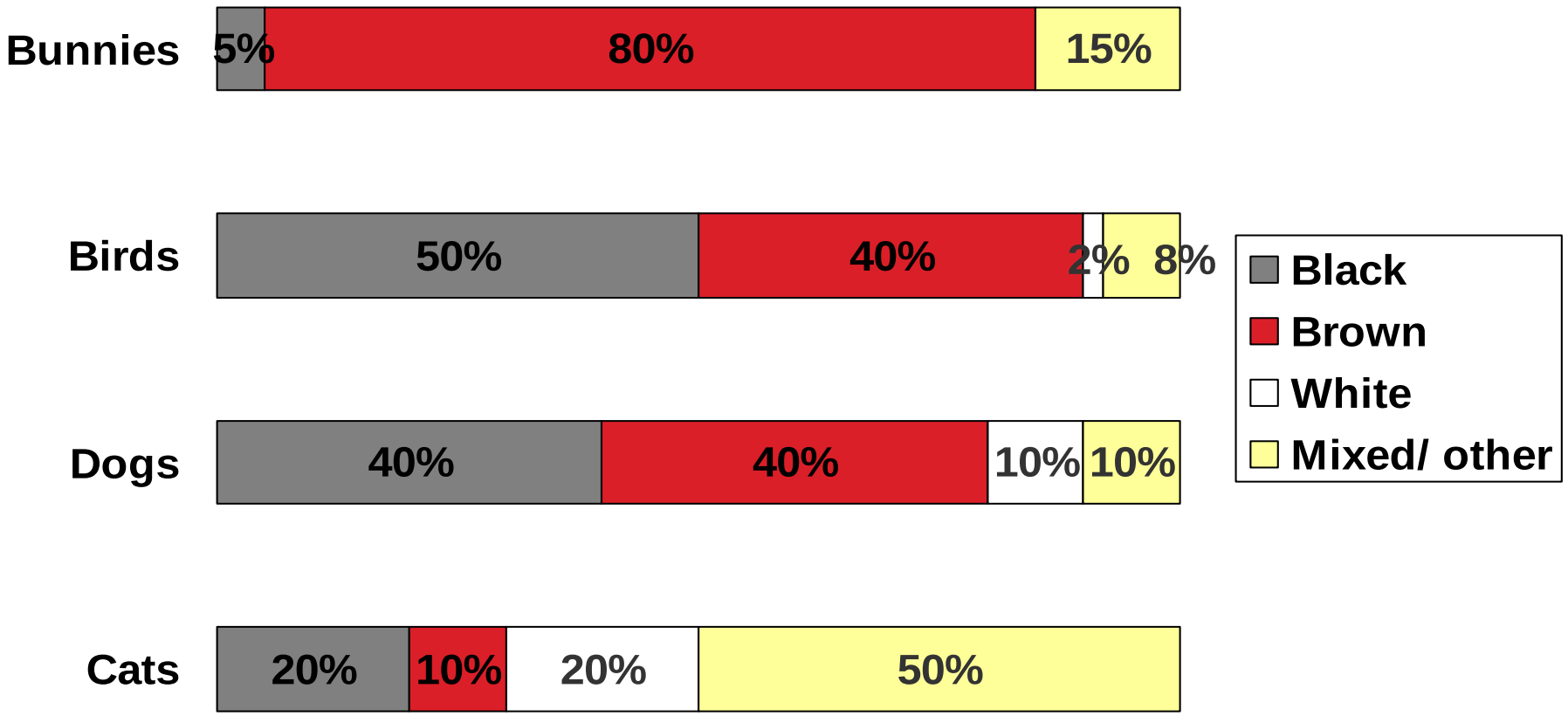
一个例子 -- 原始数据

- 以下这张表显示不同家庭宠物的颜色

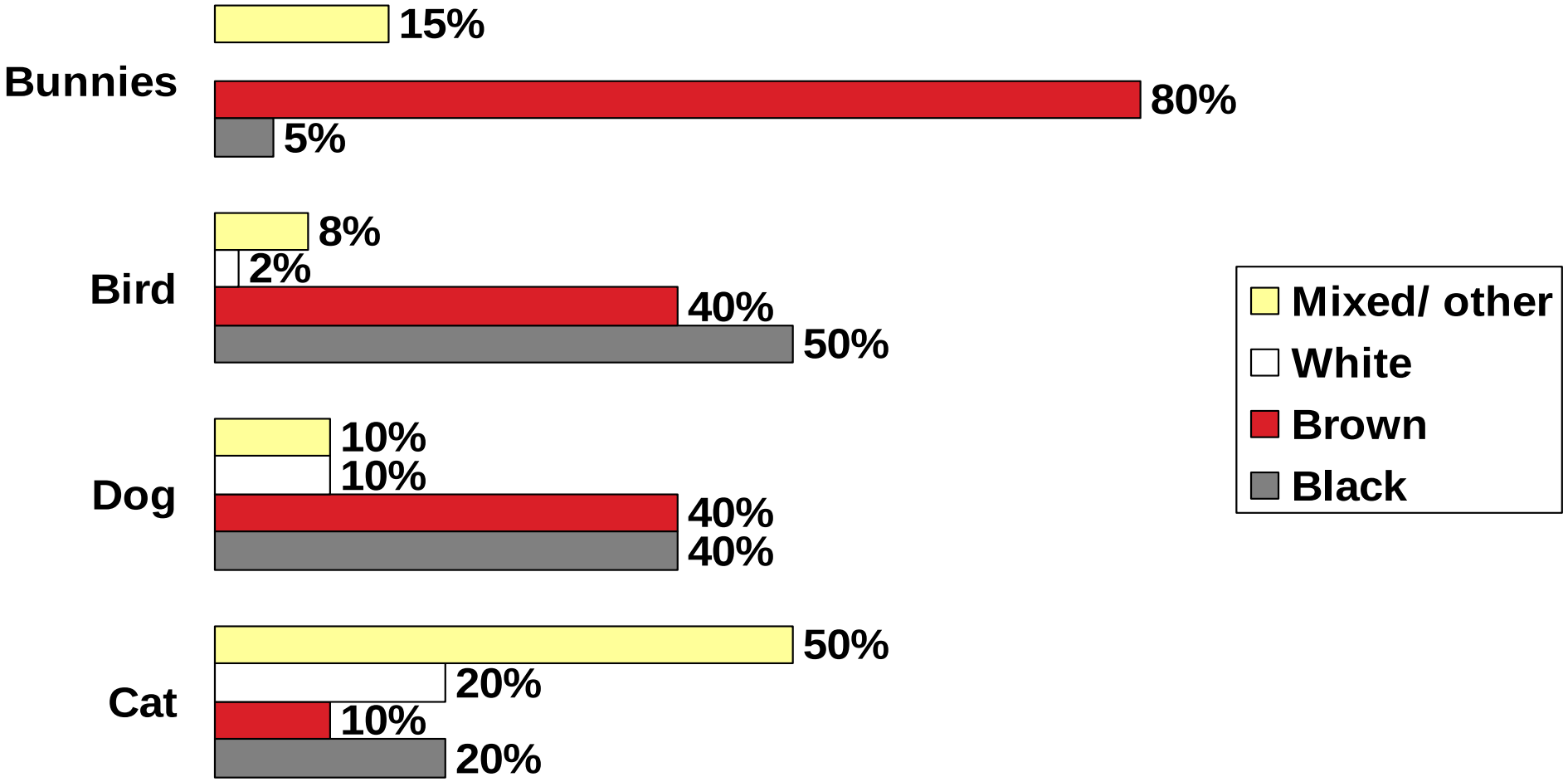
	Cats	Dogs	Birds	Bunnies
Black	20%	40%	50%	5%
Brown	10%	40%	40%	80%
White	30%	10%	2%	0%
Mixed/ other	50%	10%	8%	15%

- **思考：**
- **针对于上表，你会进行哪些分析？**
- **可以分析出什么？**
- **如何分析？**

你可能会制作的分析图...


Bunnies
5%
80%
15%
Birds
50%
40%
2%
8%
Dogs
40%
40%
10%
10%
Cats
20%
10%
20%
50%
Black
Brown
White
Mixed/ other

你可能制作的分析图...

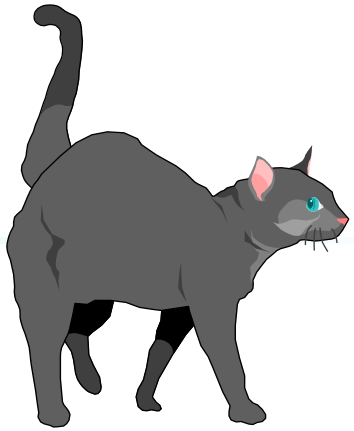

Bunnies
15%
80%
5%
Bird
8%
2%
40%
50%
Dog
10%
10%
40%
40%
Cat
50%
20%
10%
20%
Mixed/ other
White
Brown
Black

现在我们用颜色和动物名称两个变量来做2-维的图表

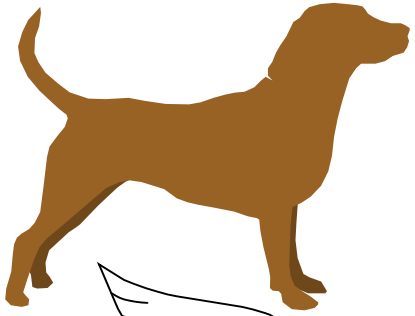
努力来显示..

- 哪些动物在颜色方面最相似，哪些区别最大？
- 哪些颜色更倾向哪类动物？
- 哪些动物和哪些颜色有更强的对应性，哪些对应性很弱？

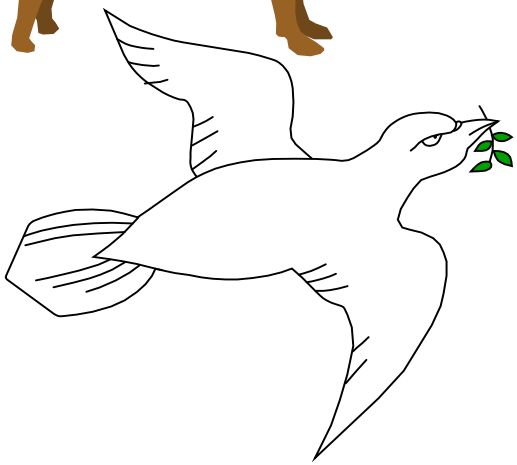
WHITE



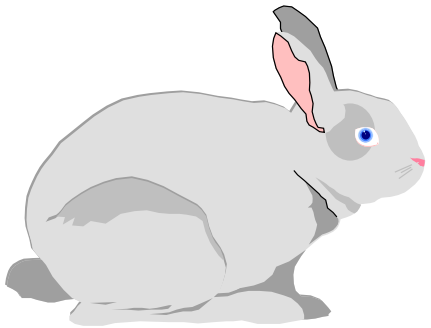
MIXED



BLACK

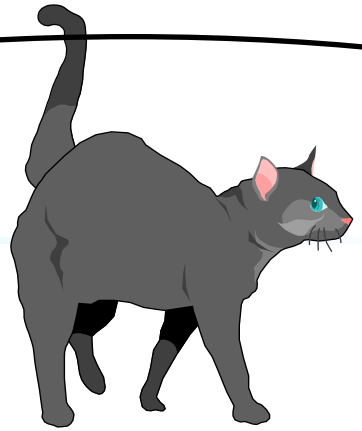


BROWN



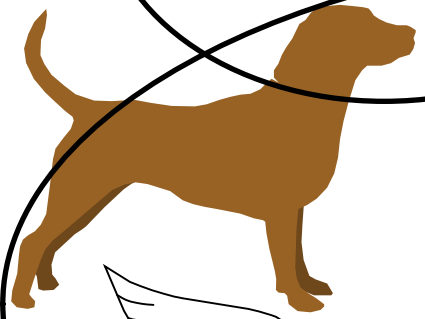
MIXED

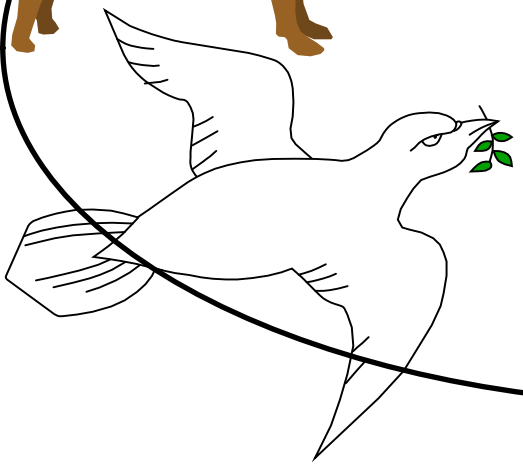
WHITE

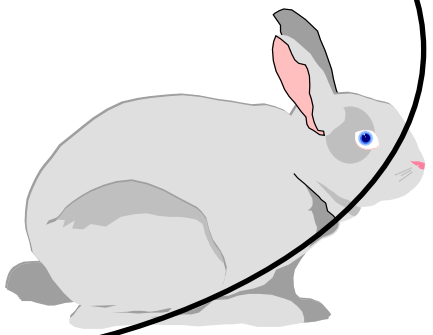


BLACK

BROWN





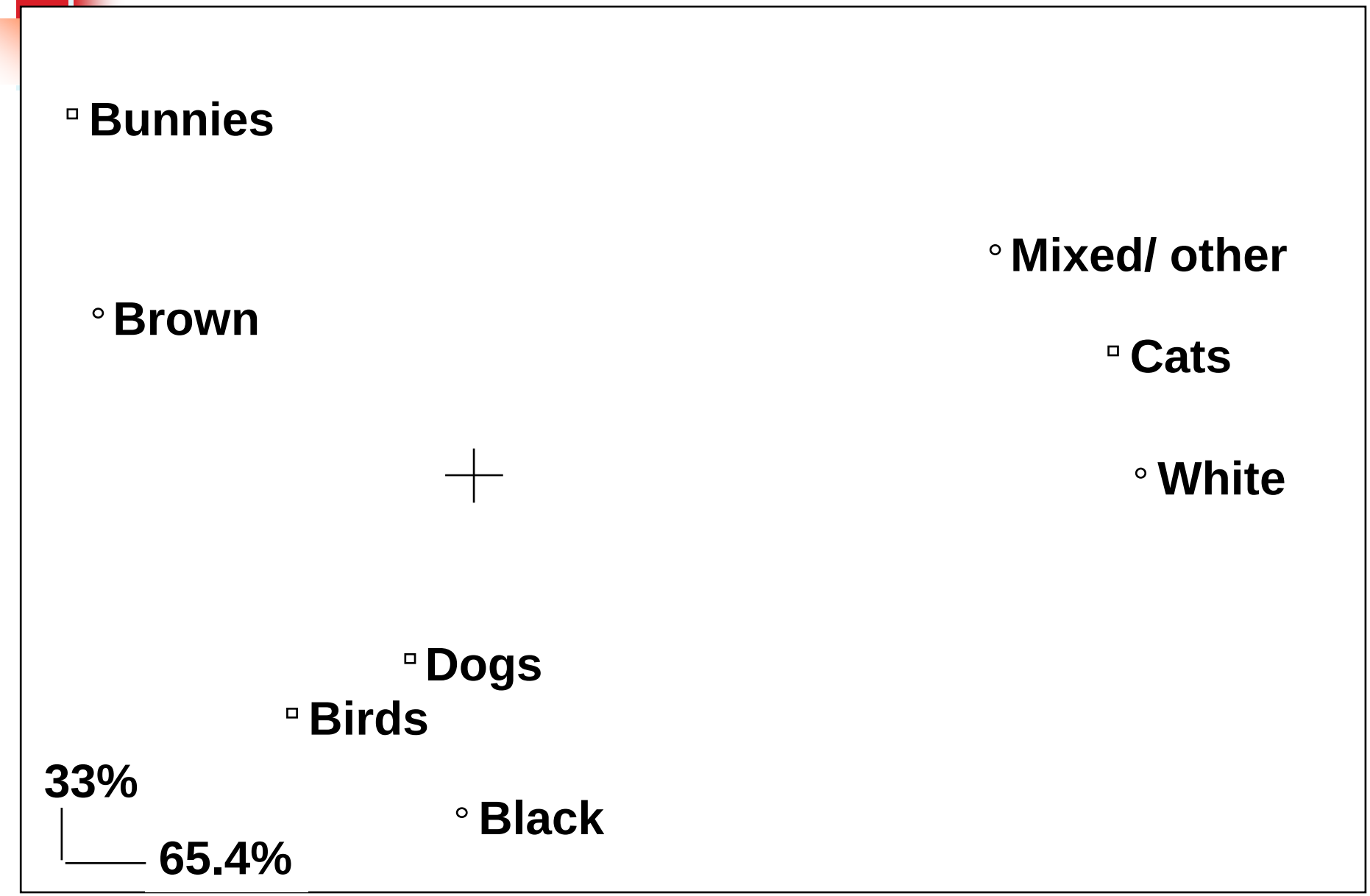


为了建立这种立体的图表你不得不...

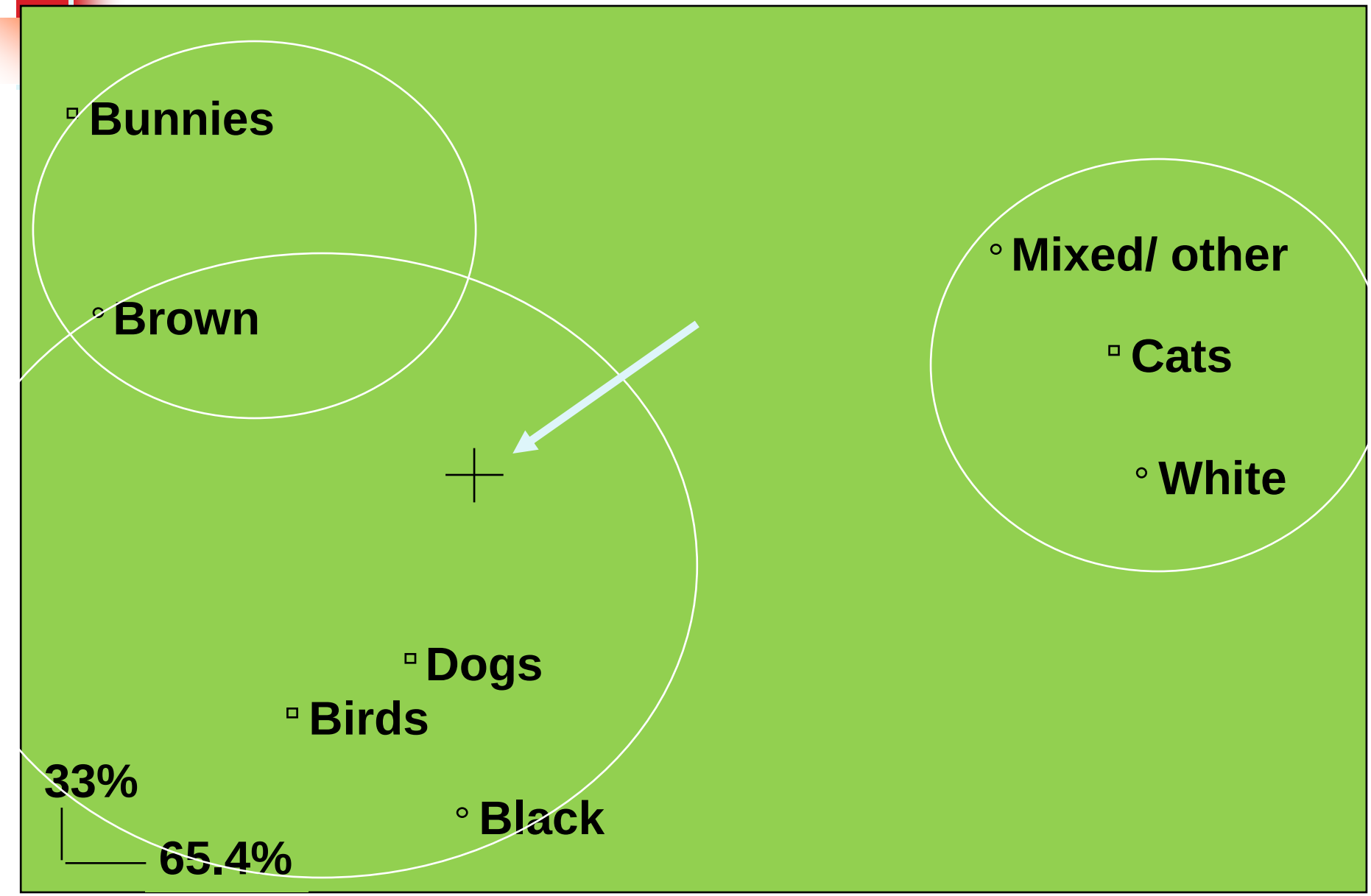
- 把那些与较多动物相关联的颜色放置在图的中央位置
- 把那些与较多动物相关联的颜色放置在图的边缘位置
- 如果一种颜色同时与超过二种以上的动物强相关，这些动物将会在图中更接近

非常简单——这就是相关性分析所做的事

以下这张表就是依据原始数据生成的…



以下这张表就是依据原始数据生成的…



实用软件

- **SPSS Categories 模块**
- **SAS Marketing 模块**
- **R 语言 MASS 包**

例2：起名为“波澜”恰当吗

中美纯水有限公司欲为其新推出的一种纯水产品起一个合适的名字，为此专门委托了当地的策划咨询公司，取了一个名字“波澜”。一个好的名字至少应该满足两个条件：

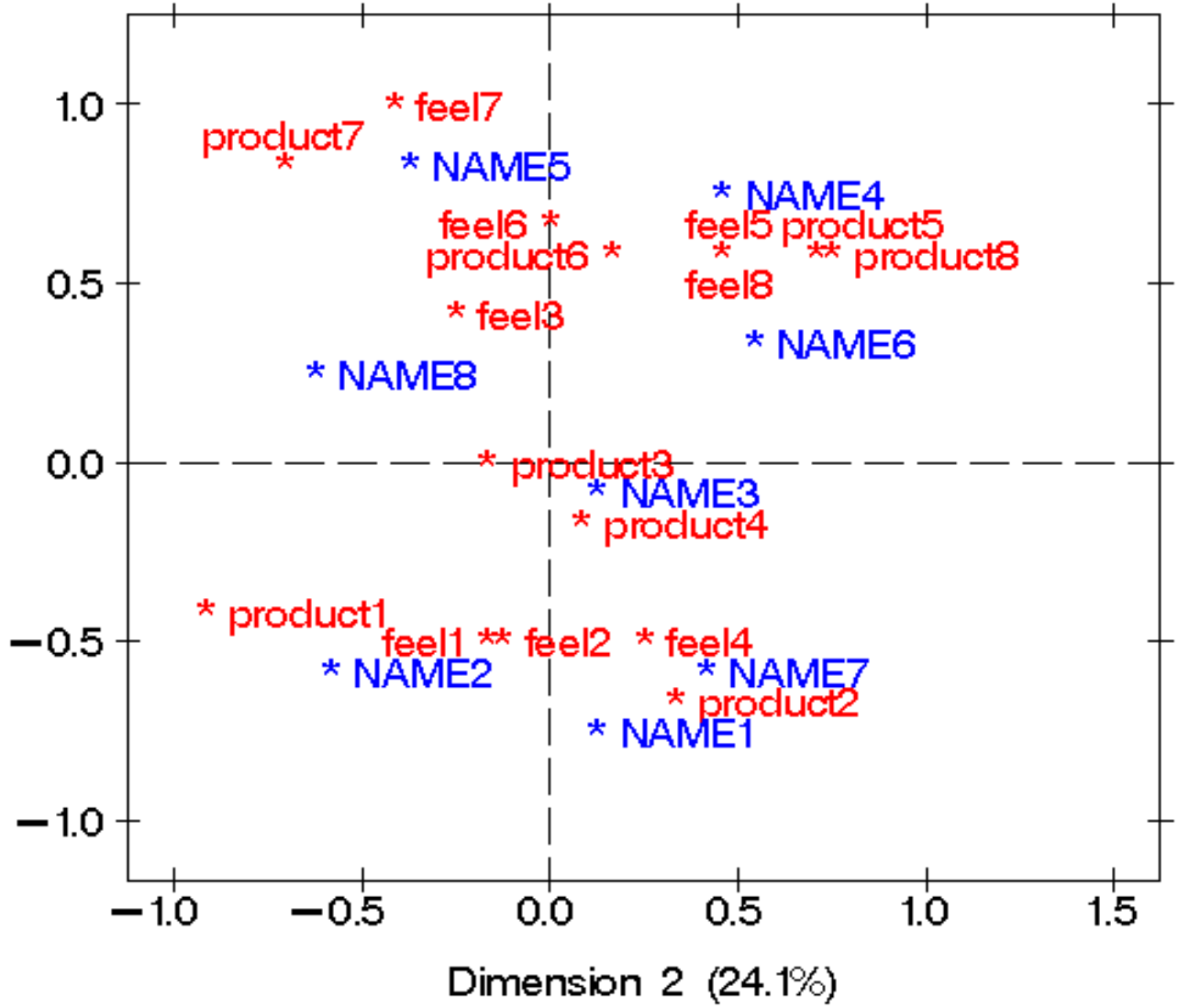
1）会使消费者联想到正确的产品“纯水”；

2）会使消费者产生与正确产品密切相关的联想，如“纯净”、“清爽”等。

后来中美纯水有限公司委托调查统计研究所，进行了一次全面的市场研究，在调查中还包括简单的名称测试。调查的代码和含义如下：

代码	含义	代码	含义	代码	含义
Name1	玉泉	Product1	雪糕	Feel1	清爽
Name2	雪源	Product2	纯水	Feel2	甘甜
Name3	春溪	Product3	碳酸饮料	Feel3	欢快
Name4	期望	Product4	果汁饮料	Feel4	纯净
Name5	波澜	Product5	保健食品	Feel5	安闲
Name6	天山绿	Product6	空调	Feel6	个性
Name7	中美纯	Product7	洗衣机	Feel7	兴奋
Name8	雪浪花	Product8	毛毯	Feel8	高档

	name1	name2	name3	name4	name5	name6	name7	name8
product1	50	442	27	21	14	50	30	258
product2	508	110	272	51	83	88	605	79
product3	55	68	93	36	71	47	37	77
product4	109	95	149	41	36	125	44	65
product5	34	29	45	302	37	135	42	18
product6	11	28	112	146	113	39	28	31
product7	30	12	54	64	365	42	8	316
product8	2	4	17	36	29	272	9	35
feel1	368	322	167	53	57	129	149	170
feel2	217	237	142	41	34	95	119	116
feel3	19	25	185	105	123	44	22	193
feel4	142	140	128	47	38	123	330	68
feel5	16	16	106	166	81	164	21	36
feel6	2	14	9	72	94	41	37	42
feel7	4	11	10	78	248	35	17	81
feel8	3	5	19	107	63	126	63	49


* feel7
product7
*
* NAME5
* NAME4
feel6 *
feel5 product5
product6 *
*
** product8
feel8
* feel3
* NAME6
* NAME8
* product3
* NAME3
* product4
* product1
feel1 ** feel2
* feel4
* NAME2
* NAME7
* product2
* NAME1
1.0
0.5
0.0
−0.5
−1.0
−1.0
−0.5
0.0
0.5
1.0
1.5
Dimension 2 (24.1%)
Dimension 1

对应分析图

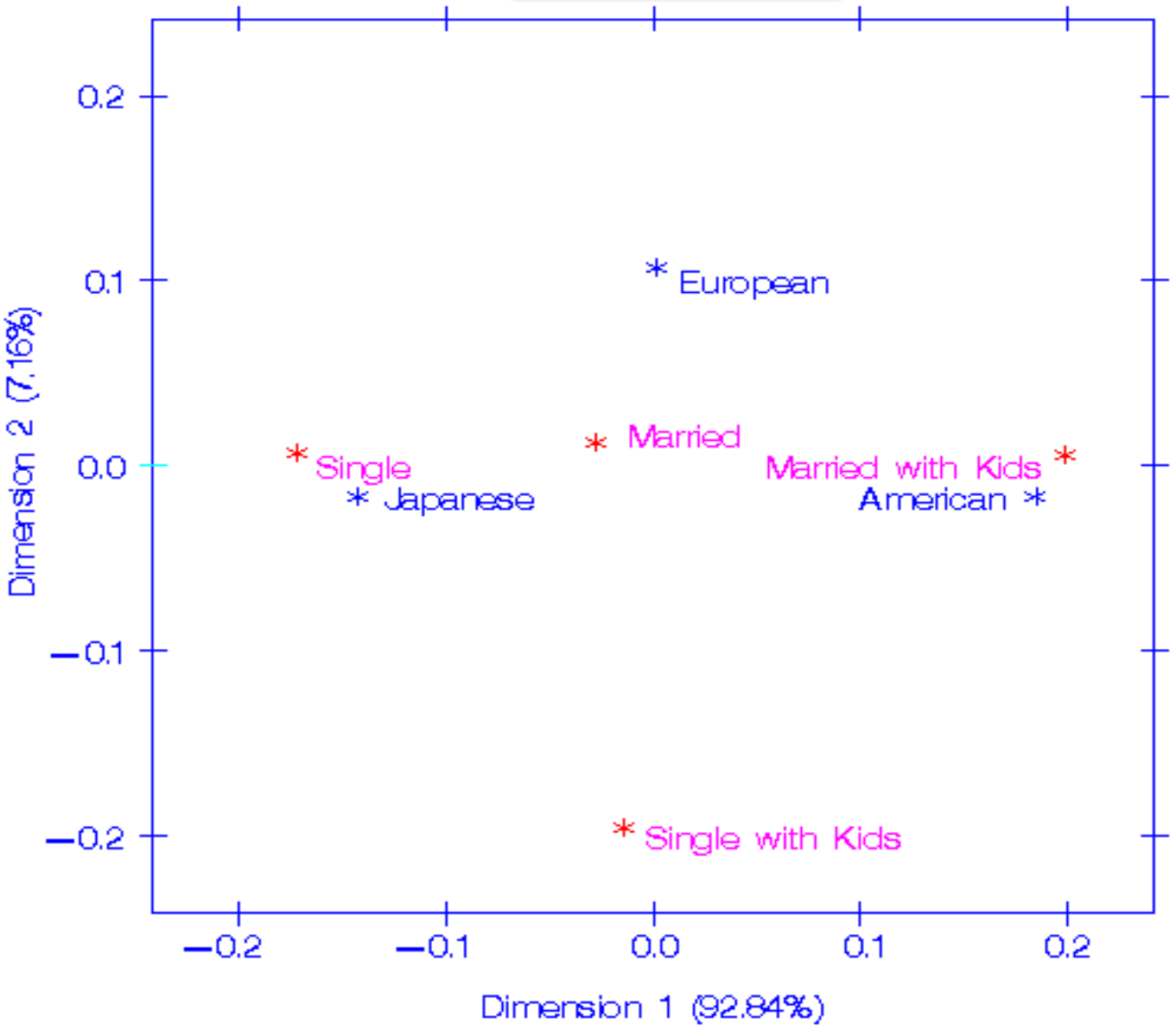
- 由直观图可以看出，“波澜”（Name5）与“洗衣机”（Product7）产品相联系，引起的感觉是“兴奋”，因此“波澜”不是合适的纯净水品牌名称。
- 中美纯水公司的产品是“纯水”（Product2），如果想要使该名称给人们一种“纯净”（Feel4）的感觉，那么“中美纯”（Name7）将是最好的商品名称。
- 如果想要使该名称给人们一种“清爽”（Feel1）的感觉，那么“玉泉”（Name1）将是最好的商品名称。

车主的车型及车主特征

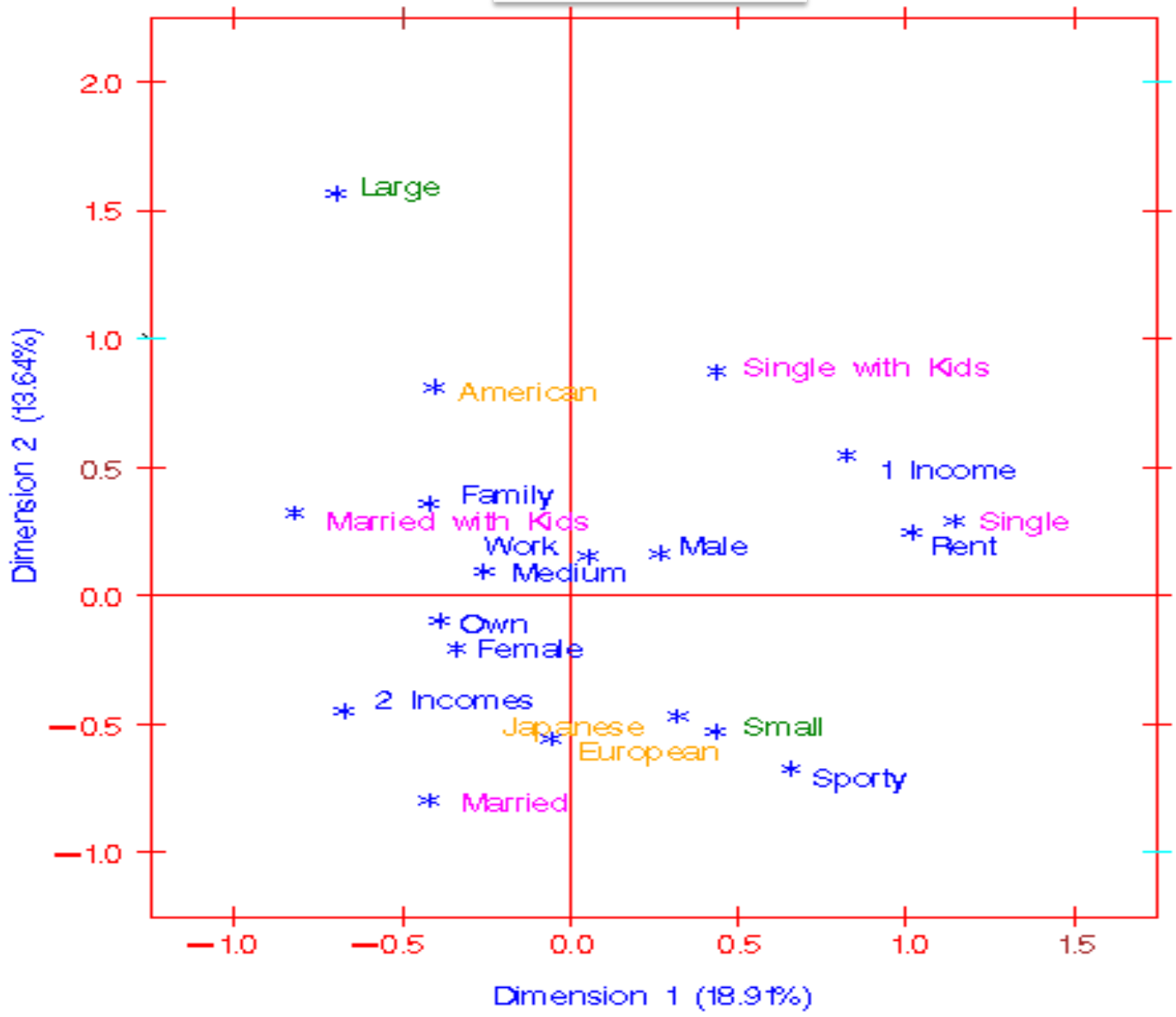
产地	1 = 'American' 2 = 'Japanese' 3 = 'European';
轿车的尺寸	1 = 'Small' 2 = 'Medium' 3 = 'Large';
车型	1 = 'Family' 2 = 'Sporty' 3 = 'Work';
拥有方式	1 = 'Own' 2 = 'Rent';
车主的性别	1 = 'Male' 2 = 'Female';
收入	1 = '1 Income' 2 = '2 Incomes';
婚姻状况	1 = 'Single with Kids' 2 = 'Married with Kids' 3 = 'Single' 4 = 'Married';

	American	European	Japanese	Large	Medium	Small	Family
American	125	0	0	36	60	29	81
European	0	44	0	4	20	20	17
Japanese	0	0	165	2	61	102	76
Large	36	4	2	42	0	0	30
Medium	60	20	61	0	141	0	89
Small	29	20	102	0	0	151	55
Family	81	17	76	30	89	55	174
Sporty	24	23	59	1	39	66	0
Work	20	4	30	11	13	30	0
1 Income	58	18	74	20	57	73	69
2 Incomes	67	26	91	22	84	78	105
Own	93	38	111	35	106	101	130
Rent	32	6	54	7	35	50	44
Married	37	13	51	9	42	50	50
Married with Kids	50	15	44	21	51	37	79
Single	32	15	62	11	40	58	35
Single with Kids	6	1	8	1	8	6	10
Female	58	21	70	17	70	62	83
Male	67	23	95	25	71	89	91

	Sporty	Work	1 Income	Incomes	Own	Rent	Married
American	24	20	58	67	93	32	37
European	23	4	18	26	38	6	13
Japanese	59	30	74	91	111	54	51
Large	1	11	20	22	35	7	9
Medium	39	13	57	84	106	35	42
Small	66	30	73	78	101	50	50
Family	0	0	69	105	130	44	50
Sporty	106	0	55	51	71	35	35
Work	0	54	26	28	41	13	16
1 Income	55	26	150	0	80	70	10
2 Incomes	51	28	0	184	162	22	91
Own	71	41	80	162	242	0	76
Rent	35	13	70	22	0	92	25
Married	35	16	10	91	76	25	101
Married with Kids	12	18	27	82	106	3	0
Single	57	17	99	10	52	57	0



从对应图可以看出，已婚带孩子和拥有美国车存在一定的关系；单身与日本车有关系。


Large
American
Single with Kids
1 Income
Family
Married with Kids
Single
Rent
Work
Male
Medium
Own
Female
2 Incomes
Japanese
European
Small
Sporty
Married
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.0
−0.5
0.0
0.5
1.0
1.5
Dimension 2 (13.64%)
Dimension 1 (18.91%)

图中的右上象限表明“单身”、“租用的”、“一项收入”和“单身带孩子”有关系；

在右下象限“跑车”、“小型”和“日本车”有关；

在左下象限表明“已婚”、“自己的”、“两项收入”和“女性”有关系；

左上象限表明“已婚带孩子”、和拥有一辆“大型”“美国”产“家用车”相对应。

这些信息对于市场调研部门确定广告的宣传对象很有用。

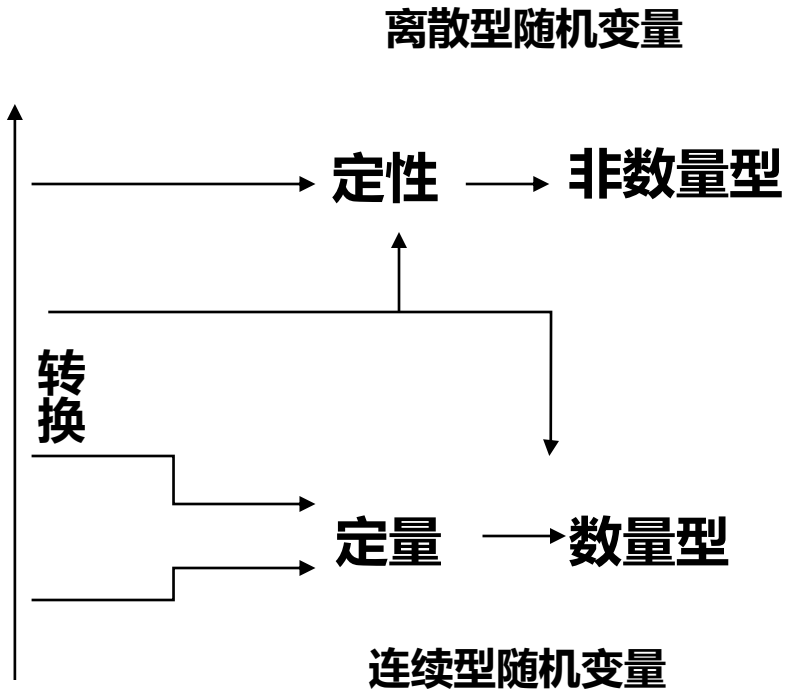
对应分析理论基础

变量的基本概念——测量等级

低

1. 名称级----定类变量
2. 顺序级----定序变量
3. 间隔级----定距变量
4. 比例级----定比变量

高



统计分析方法的应用有时候按变量的测量等级来划分。

在社会调查和市场调查中，面临着大量的定性数据（分类变量）。

- **识别消费者群体的变量：**
  - 年龄
  - 收入
  - 婚姻/家庭状况
  - 性别
  - 教育程度
  - 职业
- **区分你的产品和竞争对手的产品变量：**
  - 品牌
  - 大小
  - 颜色
  - 产地
  - 评价

传统的分析方法——交互（列联表）分析

- 在市场研究中，对于定类变量的分析，最常用、最简单的方法是**交互分析**。
- 下面的列联表显示了三个地区的120名随机样本对四种牙膏品牌的使用情况：

	地区1	地区2	地区3	合计
品牌A	5	5	30	40
品牌B	5	25	5	35
品牌C	15	5	5	25
品牌D	15	5	0	20
合计	40	40	40	120

从直观来看，品牌A在地区3占统治地位；品牌B在地区2占统治地位；地区1的消费者比较偏好品牌C和D；品牌D在地区3没有支持者。

BRAND * AREA Crosstabulation

			AREA			Total
			1 地区1	2 地区2	3 地区3	
BRAND	1 品牌A	Count	5	5	30	40
		Expected Count	13.3	13.0	13.7	40.0
		% within BRAND	12.5%	12.5%	75.0%	100.0%
		% within AREA	12.5%	12.8%	73.2%	33.3%
	2 品牌B	Count	5	25	5	35
		Expected Count	11.7	11.4	12.0	35.0
		% within BRAND	14.3%	71.4%	14.3%	100.0%
		% within AREA	12.5%	64.1%	12.2%	29.2%
	3 品牌C	Count	15	5	5	25
		Expected Count	8.3	8.1	8.5	25.0
		% within BRAND	60.0%	20.0%	20.0%	100.0%
		% within AREA	37.5%	12.8%	12.2%	20.8%
	4 品牌D	Count	15	4	1	20
		Expected Count	6.7	6.5	6.8	20.0
		% within BRAND	75.0%	20.0%	5.0%	100.0%
		% within AREA	37.5%	10.3%	2.4%	16.7%
Total		Count	40	39	41	120
		Expected Count	40.0	39.0	41.0	120.0
		% within BRAND	33.3%	32.5%	34.2%	100.0%
		% within AREA	100.0%	100.0%	100.0%	100.0%

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	78.192[a]	6	.000
Likelihood Ratio	74.014	6	.000
Linear-by-Linear Association	41.993	1	.000
N of Valid Cases	120		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 6.50.

Symmetric Measures

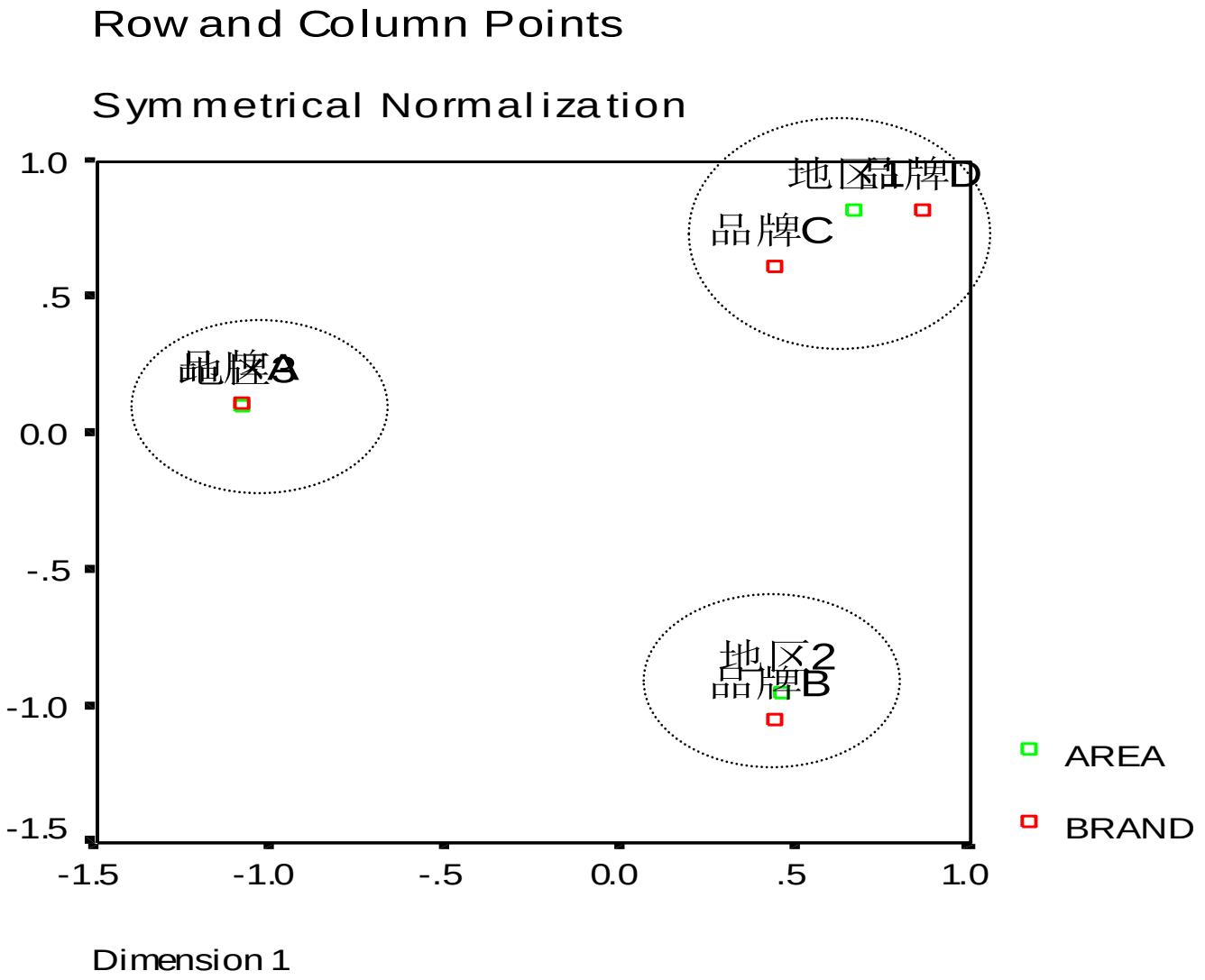
		Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Nominal by Nominal	Contingency Coefficient	.628			.000
Interval by Interval	Pearson's R	-.594	.070	-8.022	.000[c]
Ordinal by Ordinal	Spearman Correlation	-.602	.072	-8.181	.000[c]
N of Valid Cases		120			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

c. Based on normal approximation.

对应分析的结果


Row and Column Points
Symmetrical Normalization
1.0
.5
0.0
-.5
-1.0
-1.5
-1.5
-1.0
-.5
0.0
.5
1.0
地区1品牌D
品牌C
品牌A
地区2
品牌B
AREA
BRAND
Dimension 1

- 对应分析(correspondence analysis)是用于寻求列联表的行和列之间联系的一种低维图形表示法，它可以从直觉上揭示出同一分类变量的各个类别之间的差异，以及不同分类变量各个类别之间的对应关系。

- 在对应分析中，列联表的每一行对应（通常是二维）图中的一点，每一列也对应同一图中的一点。本质上，这些点都是列联表的各行各列向一个二维欧式空间的投影，这种投影最大限度地保持了各行（或各列）之间的关系。

- 对应分析是由法国人Benzecri于1970年提出的，起初在法国和日本最为流行，然后引入美国。

一、列联表及列联表分析

一、列联表

二、对应矩阵

三、行、列轮廓

列联表(contingency table)：综合了两个变量的联合分布表，同时汇总两个变量的数据的方法。又称交叉分组表（ crosstabulation)

家庭状况	青少年行为	
	犯罪	未犯罪
破裂	**146**	**45**
和好	**334**	**499**

一、列联表

表 9.1.1　　$p \times q$ 列联表

行 \ 列	1	2	…	q	合计
1	n_{11}	n_{12}	…	n_{1q}	$n_{1\cdot}$
2	n_{21}	n_{22}	…	n_{2q}	$n_{2\cdot}$
⋮	⋮	⋮		⋮	⋮
p	n_{p1}	n_{p2}	…	n_{pq}	$n_{p\cdot}$
合计	$n_{\cdot 1}$	$n_{\cdot 2}$	…	$n_{\cdot q}$	n

其中，n_{ij}是第i 行、第j 列类别组合的频数，

$i=1,2,\cdots,p,\ j=1,2,\cdots,q$

$n_{i\cdot}=\sum_{j=1}^{q}n_{ij}$ 为第i 行的频数之和，$i=1,2,\cdots,p$ ；

$n_{\cdot j}=\sum_{i=1}^{p}n_{ij}$ 为第j 列的频数之和，$j=1,2,\cdots,q$；

$$n=\sum_{i=1}^{p}n_{i\cdot}=\sum_{j=1}^{q}n_{\cdot j}=\sum_{i=1}^{p}\sum_{j=1}^{q}n_{ij}$$

为所有类别组合的频数总和。

二、对应矩阵

表 9.1.2　　对应矩阵

行 \ 列	1	2	…	q	合计
1	p_{11}	p_{12}	…	p_{1q}	$p_{1\cdot}$
2	p_{21}	p_{22}	…	p_{2q}	$p_{2\cdot}$
⋮	⋮	⋮		⋮	⋮
p	p_{p1}	p_{p2}	…	p_{pq}	$p_{p\cdot}$
合计	$p_{\cdot 1}$	$p_{\cdot 2}$	…	$p_{\cdot q}$	1

这里，$p_{ij}=\dfrac{n_{ij}}{n}$，$p_{i\cdot}=\sum_{j=1}^{q}p_{ij}=\sum_{j=1}^{q}\dfrac{n_{ij}}{n}$，$p_{\cdot j}=\sum_{i=1}^{p}p_{ij}=\sum_{i=1}^{p}\dfrac{n_{ij}}{n}$

显然有　$\sum_{i=1}^{p}p_{i\cdot}=\sum_{j=1}^{q}p_{\cdot j}=1$

称 $\boldsymbol{P}=\left(p_{ij}\right)=\left(n_{ij}/n\right)$ 为对应矩阵。将对应矩阵表中的最后一列用 $\mathbf{r}$ 表示，即

$$\mathbf{r}=\mathbf{P1}=\left(p_{1\cdot},p_{2\cdot},\cdots,p_{p\cdot}\right)'$$

其中 $\mathbf{1}=(1,1,\cdots,1)'$ 是元素均为1的 q 维向量，最后一行用表示 $\mathbf{c}'$ 即

$$\mathbf{c}'=\mathbf{1}'\mathbf{P}=\left(p_{\cdot 1},p_{\cdot 2},\cdots,p_{\cdot q}\right)$$

其中 $1=(1,1,\cdots,1)'$ 是元素均为1的 p 维向量，向量 $\mathbf{r}$ 和 $\mathbf{c}$ 的元素有时称为行和列密度（masses）。

三、行、列独立的检验

在列联表中，检验行变量和列变量相互独立假设的统计量为

$$\chi^2 = n\sum_{i=1}^{p}\sum_{j=1}^{q}\frac{\left(p_{ij}-p_{i\cdot}p_{\cdot j}\right)^2}{p_{i\cdot}p_{\cdot j}}$$

当独立性的原假设为真，且样本容量 n 充分大，期望频数 $np_{i\cdot}p_{\cdot j}\geq 5,\quad i=1,2,\cdots,p,\quad j=1,2,\cdots,q$ 时，χ^2 近似服从自由度为 $(p-1)(q-1)$ 的卡方分布。拒绝规则为 $\chi^2\geq\chi_\alpha^2(p-1,q-1)$

若 $\chi^2\geq\chi_\alpha^2(p-1,q-1)$ 则拒绝独立性的原假设，其中 $\chi_\alpha^2(p-1,q-1)$ 是 $\chi^2(p-1,q-1)$ 的上分位点。

例 某医师研究用兰芩口服液与银黄口服液治疗慢性咽炎疗效有无差别，将病情相似的80名患者随机分成两组，分别用两种药物治疗，

慢性咽炎两种药物疗效资料

$$T_{ij}=\frac{n_i m_j}{n}=\frac{45\times 65}{80}=36.56$$

药物	有效	无效	
兰芩口服液	41(36.56)	4(8.44)	45(固定值)
银黄口服液	24(28.44)	11(6.56)	35(固定值)
合计	65	15	80

完全随机设计下两组频数分布的四格表

处理	属性		合计
	阳性	阴性	
1 组	$A_{11}(T_{11})$	$A_{12}(T_{12})$	n_1(固定值)
2 组	$A_{21}(T_{21})$	$A_{22}(T_{22})$	n_2(固定值)
合计	m_1	m_2	n

二、对应分析的基本理论

1、有关概念

2、对应分析的基本理论

1、行、列轮廓（剖面）

- 第 i 行轮廓：

$$\mathbf{r}_i' = \left(\frac{p_{i1}}{p_{i\cdot}}, \frac{p_{i2}}{p_{i\cdot}}, \cdots, \frac{p_{iq}}{p_{i\cdot}} \right) = \left(\frac{n_{i1}}{n_{i\cdot}}, \frac{n_{i2}}{n_{i\cdot}}, \cdots, \frac{n_{iq}}{n_{i\cdot}} \right)$$

其各元素之和等于1，即 $\mathbf{r}_i'\mathbf{1} = 1, \quad i = 1, 2, \cdots, p$

- 第 j 列轮廓：

$$\mathbf{c}_j = \left(\frac{p_{1j}}{p_{\cdot j}}, \frac{p_{2j}}{p_{\cdot j}}, \cdots, \frac{p_{pj}}{p_{\cdot j}} \right)' = \left(\frac{n_{1j}}{n_{\cdot j}}, \frac{n_{2j}}{n_{\cdot j}}, \cdots, \frac{n_{pj}}{n_{\cdot j}} \right)'$$

其各元素之和等于1，即 $\mathbf{1}'\mathbf{c}_j = 1, \quad j = 1, 2, \cdots, q$

表 9.1.2　　对应矩阵

行 \ 列	1	2	…	q	合计
1	p_{11}	p_{12}	…	p_{1q}	$p_{1\cdot}$
2	p_{21}	p_{22}	…	p_{2q}	$p_{2\cdot}$
⋮	⋮	⋮		⋮	⋮
p	p_{p1}	p_{p2}	…	p_{pq}	$p_{p\cdot}$
合计	$p_{\cdot 1}$	$p_{\cdot 2}$	…	$p_{\cdot q}$	1

- 行列条件概率

行轮廓矩阵

$$\mathbf{R}=\mathbf{D}_r^{-1}\mathbf{P}=\begin{pmatrix}\mathbf{r}_1'\\ \mathbf{r}_2'\\ \vdots\\ \mathbf{r}_p'\end{pmatrix}=\begin{pmatrix}\frac{p_{11}}{p_{1\cdot}} & \frac{p_{12}}{p_{1\cdot}} & \cdots & \frac{p_{1q}}{p_{1\cdot}}\\ \frac{p_{21}}{p_{2\cdot}} & \frac{p_{22}}{p_{2\cdot}} & \cdots & \frac{p_{2q}}{p_{2\cdot}}\\ \vdots & \vdots & & \vdots\\ \frac{p_{p1}}{p_{p\cdot}} & \frac{p_{p2}}{p_{p\cdot}} & \cdots & \frac{p_{pq}}{p_{p\cdot}}\end{pmatrix}$$

其中 $\mathbf{D}_r=\mathrm{diag}\left(p_{1\cdot},p_{2\cdot},\cdots,p_{p\cdot}\right)$ 。

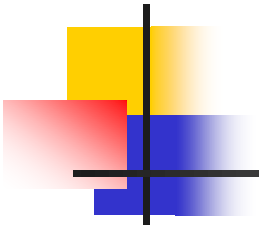
列轮廓矩阵

$$\mathbf{C}=\mathbf{P}\mathbf{D}_c^{-1}=\left(\mathbf{c}_1,\mathbf{c}_2,\cdots,\mathbf{c}_q\right)=\begin{pmatrix}\frac{p_{11}}{p_{\cdot 1}} & \frac{p_{12}}{p_{\cdot 2}} & \cdots & \frac{p_{1q}}{p_{\cdot q}}\\ \frac{p_{21}}{p_{\cdot 1}} & \frac{p_{22}}{p_{\cdot 2}} & \cdots & \frac{p_{2q}}{p_{\cdot q}}\\ \vdots & \vdots & & \vdots\\ \frac{p_{p1}}{p_{\cdot 1}} & \frac{p_{p2}}{p_{\cdot 2}} & \cdots & \frac{p_{pq}}{p_{\cdot q}}\end{pmatrix}$$

其中 $\mathbf{D}_c=\mathrm{diag}\left(p_{\cdot 1},p_{\cdot 2},\cdots,p_{\cdot q}\right)$。

说明：

- 行变量的各个取值的情况可以用**p**个点表示出来，同理，列变量的各个取值的情况可以用**q**个点表示出来



- 而对应分析就是利用降维的思想，既把行的各个状态表现在一张二维图上，又把列的各个状态表现在一张二维图上。
- 通过后面的分析可以看到，这两张二维图的坐标轴有相同的含义。
- 即可以把行与列同时在一张二维图上显示出来。

$$\mathbf{r} = \mathbf{P1} = \left(\mathbf{P}\mathbf{D}_c^{-1}\right)\left(\mathbf{D}_c\mathbf{1}\right) = \left(\mathbf{c}_1, \mathbf{c}_2, \cdots, \mathbf{c}_q\right)\begin{pmatrix} p_{\cdot 1} \\ p_{\cdot 2} \\ \vdots \\ p_{\cdot q} \end{pmatrix} = \sum_{j=1}^{q} p_{\cdot j}\mathbf{c}_j$$

可见，**r**可以表示成各列轮廓的加权平均。类似地，

$$\mathbf{c}' = \mathbf{1}'\mathbf{P} = \left(\mathbf{1}'\mathbf{D}_r\right)\left(\mathbf{D}_r^{-1}\mathbf{P}\right) = \sum_{i=1}^{p} p_{i\cdot}\mathbf{r}_{i\cdot}'$$

即 **c**′可以表示成各行轮廓的加权平均。

例

将由个人组成的样本按心理健康状况与社会经济状况进行交叉分类，分类结果见表9.1.3。

表 9.1.3　　心理健康状况—社会经济状况数据

心理健康状况 \ 社会经济状况	A(高)	B	C	D	E(低)
0(好)	121	57	72	36	21
1(轻微症状形成)	188	105	141	97	71
2(中等症状形成)	112	65	77	54	54
3(受损)	86	60	94	78	71

将表9.1.3中的数据除以，得到对应矩阵，列于表9.1.4中。表9.1.4给出的行密度和列密度向量为

$$\mathbf{r}=\begin{pmatrix}0.185\\0.363\\0.218\\0.234\end{pmatrix},\quad \mathbf{c}'=(0.305,0.173,0.231,0.160,0.131)$$

表 9.1.4　　从表 9.1.3 算得的对应矩阵

社会经济状况 / 心理健康状况	A(高)	B	C	D	E(低)	合计
0(好)	0.073	0.034	0.043	0.022	0.013	0.185
1(轻微症状形成)	0.113	0.063	0.085	0.058	0.043	0.363
2(中等症状形成)	0.067	0.039	0.046	0.033	0.033	0.218
3(受损)	0.052	0.036	0.057	0.047	0.043	0.234
合计	0.305	0.173	0.231	0.160	0.131	1.000

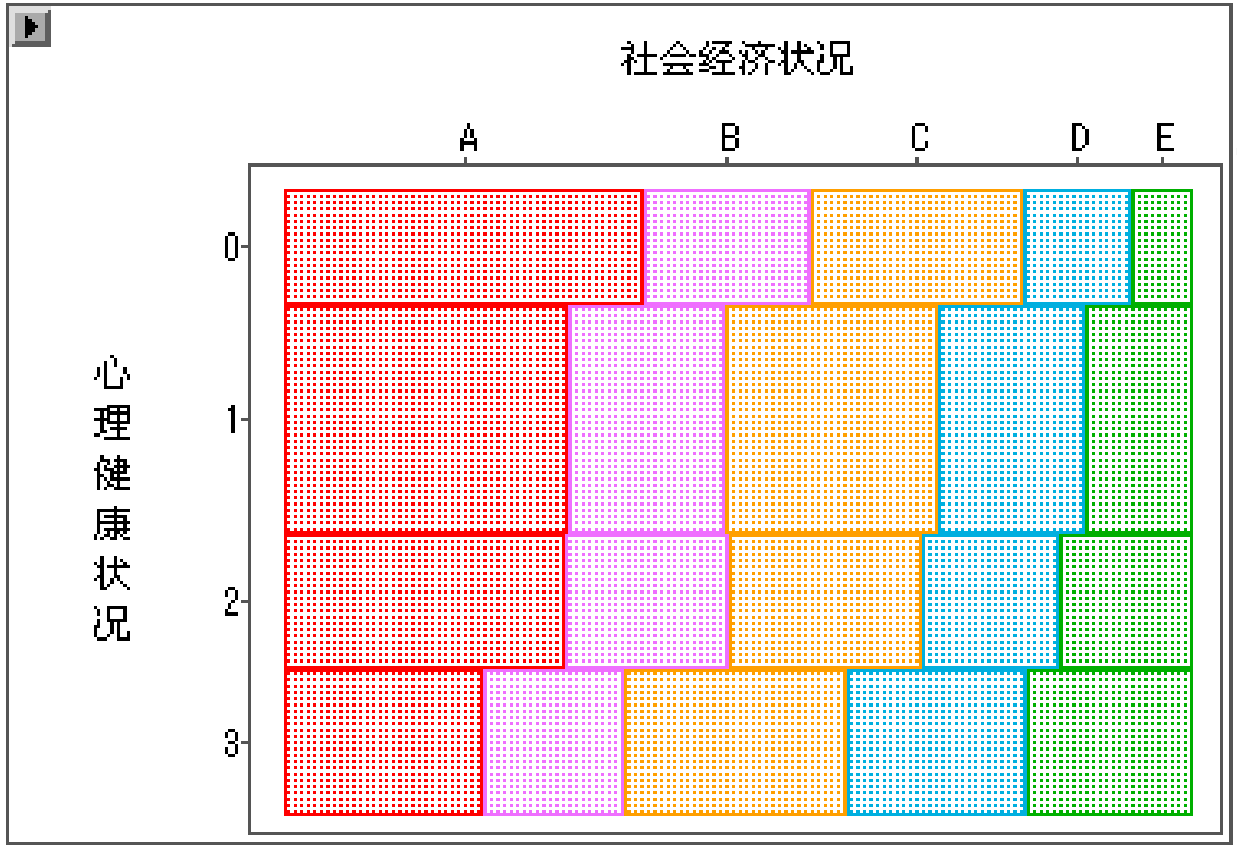
行轮廓矩阵为

$$\mathbf{R}=\mathbf{D}_r^{-1}\mathbf{P}=\begin{pmatrix}0.394 & 0.186 & 0.235 & 0.117 & 0.068\\ 0.312 & 0.174 & 0.234 & 0.161 & 0.118\\ 0.309 & 0.180 & 0.213 & 0.149 & 0.149\\ 0.221 & 0.154 & 0.242 & 0.201 & 0.183\end{pmatrix}$$

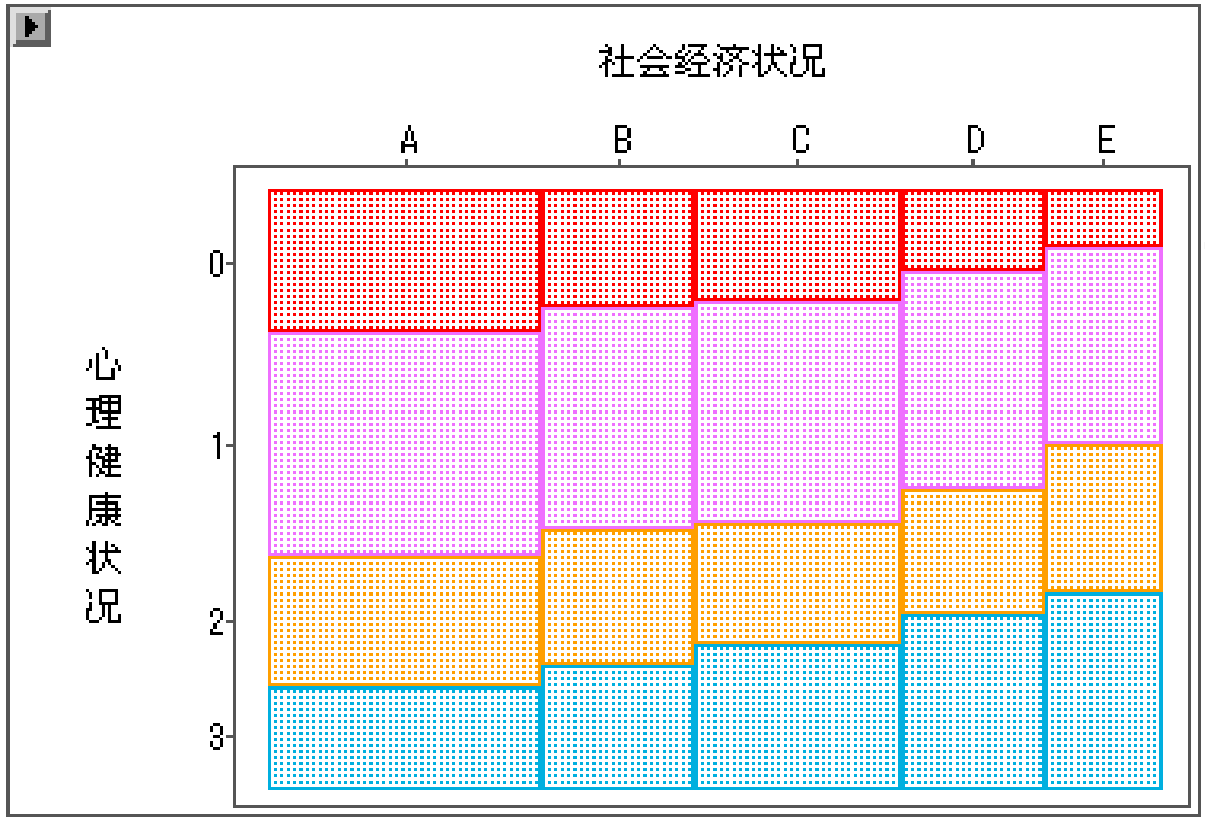
列轮廓矩阵为

$$\mathbf{C}=\mathbf{P}\mathbf{D}_c^{-1}=\begin{pmatrix}0.239 & 0.199 & 0.188 & 0.136 & 0.097\\ 0.371 & 0.366 & 0.367 & 0.366 & 0.327\\ 0.221 & 0.226 & 0.201 & 0.204 & 0.249\\ 0.170 & 0.209 & 0.245 & 0.294 & 0.327\end{pmatrix}$$

两个马赛克图



对心理健康的每一种状况，A、B、C、D、E五个小方块的宽度显示了行轮廓，0、1、2、3四种心理健康状况的小方块高度显示了行密度。



对社会经济的每一种状况，0、1、2、3四个小方块的高度显示了列轮廓，A、B、C、D、E五种社会经济状况的小方块宽度显示了列密度。

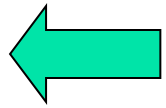
2、行（列）点之间的距离

如果两个行（列）点接近，则表明相应的两个行（列）轮廓是类似的；反之，如果两个行（列）点远离，则表明相应的两个行（列）轮廓是非常不同的。

因此，引入距离来分别描述各种状态之间的接近程度。需要指出的是，行点与列点之间并没有直接的距离关系。

这里只对行做详细说明

$$\begin{pmatrix} \frac{p_{11}}{p_{1.}} & \frac{p_{12}}{p_{1.}} & \cdots & \frac{p_{1p}}{p_{1.}} \\ & \ddots & & \\ & & \ddots & \\ \frac{p_{n1}}{p_{n.}} & \frac{p_{n2}}{p_{n.}} & \cdots & \frac{p_{np}}{p_{n.}} \end{pmatrix}$$



将每一行视为P维空间中的一个样本点，每个样本点的坐标是个变量在该样本点的相对比例。经过这个变换后对n个样本的研究就变成了对n个样本点的相对关系的研究。

任意两个样本 K 与 L 之间的距离：

$$D^2(K,L)=\sum_{j=1}^{p}\left(\frac{p_{Kj}}{p_{K.}}-\frac{p_{Lj}}{p_{L.}}\right)^2$$

加权距离

消除变量的数量级

$$D_*^2(K,L)=\sum_{j=1}^{p}\left(\frac{p_{Kj}}{p_{k.}}-\frac{p_{Lj}}{p_{L.}}\right)^2 / p_{.j}$$

$$=\sum_{j=1}^{p}\left(\frac{p_{Kj}}{\sqrt{p_{.j}}\,p_{k.}}-\frac{p_{Lj}}{\sqrt{p_{.j}}\,p_{L.}}\right)^2$$

等价于我们有了一个新的矩阵：

$$p^{*}=\begin{pmatrix} \frac{p_{11}}{\sqrt{p_{.1}p_{1.}}} & \frac{p_{12}}{\sqrt{p_{.2}p_{1.}}} & \cdots & \frac{p_{1p}}{\sqrt{p_{p.}p_{1.}}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{p_{n1}}{\sqrt{p_{.1}p_{n.}}} & \frac{p_{n2}}{\sqrt{p_{.2}p_{n.}}} & \cdots & \frac{p_{np}}{\sqrt{p_{.p}p_{n.}}} \end{pmatrix}$$

类似的，可将p个变量看成n为空间的点，按照同样的方法即可得到两个变量间的加权距离：

$$D_*^2(i,j)=\sum_{k=1}^{n}\left(\frac{p_{ki}}{\sqrt{p_{k.}p_{.i}}}-\frac{p_{kj}}{\sqrt{p_{k.}p_{.j}}}\right)^2$$

也即是等价于我们有了一个新的矩阵：

$$p^{**}=\begin{pmatrix} \dfrac{p_{11}}{\sqrt{p_{1.}p_{.1}}} & \dfrac{p_{12}}{\sqrt{p_{1.}p_{.2}}} & \cdots & \dfrac{p_{1p}}{\sqrt{p_{1.}p_{.p}}} \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{p_{n1}}{\sqrt{p_{n.}p_{.1}}} & \dfrac{p_{n2}}{\sqrt{p_{n.}p_{.2}}} & \cdots & \dfrac{p_{np}}{\sqrt{p_{n.}p_{.p}}} \end{pmatrix}$$

$$p^{*}=\begin{pmatrix} \dfrac{p_{11}}{\sqrt{p_{.1}p_{1.}}} & \dfrac{p_{12}}{\sqrt{p_{.2}p_{1.}}} & \cdots & \dfrac{p_{1p}}{\sqrt{p_{p.}p_{1.}}} \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{p_{n1}}{\sqrt{p_{.1}p_{n.}}} & \dfrac{p_{n2}}{\sqrt{p_{.2}p_{n.}}} & \cdots & \dfrac{p_{np}}{\sqrt{p_{.p}p_{n.}}} \end{pmatrix}$$

3、总惯量

$$\text{总惯量} = \frac{\chi^2}{n} = \sum_{i=1}^{p}\sum_{j=1}^{q}\frac{\left(p_{ij} - p_{i\cdot}p_{\cdot j}\right)^2}{p_{i\cdot}p_{\cdot j}}$$

总惯量还可以行轮廓和列轮廓的形式表达如下：

$$\text{总惯量} = \sum_{i=1}^{p} p_{i\cdot}\sum_{j=1}^{q}\frac{\left(p_{ij}/p_{i\cdot} - p_{\cdot j}\right)^2}{p_{\cdot j}} = \sum_{i=1}^{p} p_{i\cdot}\left(\mathbf{r}_i - \mathbf{c}\right)'\mathbf{D}_c^{-1}\left(\mathbf{r}_i - \mathbf{c}\right)$$

$$\text{总惯量} = \sum_{j=1}^{q} p_{\cdot j}\sum_{i=1}^{p}\frac{\left(p_{ij}/p_{\cdot j} - p_{i\cdot}\right)^2}{p_{i\cdot}} = \sum_{j=1}^{q} p_{\cdot j}\left(\mathbf{c}_j - \mathbf{r}\right)'\mathbf{D}_r^{-1}\left(\mathbf{c}_j - \mathbf{r}\right)$$

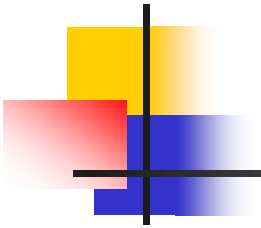
其中

$$\left(\mathbf{r}_i-\mathbf{c}\right)' \mathbf{D}_c^{-1}\left(\mathbf{r}_i-\mathbf{c}\right)=\sum_{j=1}^{q}\frac{\left(p_{ij}/p_{i\cdot}-p_{\cdot j}\right)^2}{p_{\cdot j}}$$

称为第 i行轮廓 $\mathbf{r}_i$到行轮廓中心 $\mathbf{c}$的卡方（ χ^2距离，它可看作是一个加权的平方欧氏距离。同样，

$$\left(\mathbf{c}_j-\mathbf{r}\right)' \mathbf{D}_r^{-1}\left(\mathbf{c}_j-\mathbf{r}\right)=\sum_{i=1}^{p}\frac{\left(p_{ij}/p_{\cdot j}-p_{i\cdot}\right)^2}{p_{i\cdot}}$$

是第 j 列轮廓 $\mathbf{c}_j$到列轮廓中心 $\mathbf{r}$的卡方距离。故总惯量可看成是**行轮廓到其中心的卡方距离的加权平均**，也可看成是**列轮廓到其中心的卡方距离的加权平均。它既度量了行轮廓之间的总变差，也度量了列轮廓之间的总变差。**



对应分析就是在总惯量信息损失最小的前提下，简化数据结构以反映量属性变量之间的相关系数。

实际上，总惯量的概念类似于主成分分析或因子分析中总方差总和的概念。

3、对应分析的基本理论

？ 样本与变量间的关系

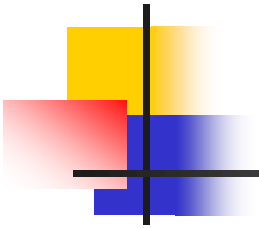
要通过样本来获得变量的观测值，反之又要通过变量来对样本进行刻画和解释。

对变量进行因子分析称R型因子分析，对样本进行因子分析称Q型因子分析

对应分析是将R型因子分析与Q型因子分析结合起来进行统计分析的统计方法。

对应分析从R型因子分析出发，而直接获得Q型因子的分析结果。

根据R型和Q型分析的内在联系，将变量和指标同时反映到相同坐标轴的一张图形上，旨在以简洁、明了的方式揭示属性变量之间及属性变量各种状态之间的相关关系。便 于对问题分析。



对应分析提供三个方面的信息：

指标之间的信息

样本之间的信息

指标与样本之间的信息

这些关系是通过作图来表示的。

样本数据

$$X=\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \cdot & \cdot & \cdot & \cdot \\ x_{n1} & x_{n2} & & x_{np} \end{pmatrix} \qquad \begin{pmatrix} x_{1.} \\ x_{2.} \\ \cdots \\ x_{n.} \end{pmatrix}$$

$$\begin{pmatrix} x_{.1} & x_{.2} & \cdots & x_{.p} \end{pmatrix}$$

$$x_{i.}=\sum_{j=1}^{p} x_{ij} \qquad x_{.j}=\sum_{i=1}^{n} x_{ij}$$

$$x_{..} = \sum_{i=1}^{n} \sum_{j=1}^{p} x_{ij}$$

$$p_{ij} = \frac{x_{ij}}{x_{..}}$$

$$1 = \sum_{i=1}^{n} \sum_{j=1}^{p} p_{ij}$$

$$p_{i.} = \sum_{j=1}^{p} p_{ij}$$

$$p_{.j} = \sum_{i=1}^{n} p_{ij}$$

概率矩阵

$$P=\begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1p} \\ p_{21} & p_{22} & \cdots & p_{2p} \\ . & . & . & . \\ p_{n1} & p_{n2} & & p_{np} \end{pmatrix} \begin{pmatrix} p_{1.} \\ p_{2.} \\ \cdots \\ p_{n.} \end{pmatrix}$$

$$\begin{pmatrix} p_{.1} & p_{.2} & \cdots & p_{.p} \end{pmatrix}$$

$$p_{i.}=\sum_{j=1}^{p} p_{ij} \qquad \sum_{j=1}^{p} p_{.j}=1$$

$$p_{.j}=\sum_{i=1}^{n} p_{.j} \qquad \sum_{i=1}^{n} p_{i.}=1$$

$$\left(\frac{p_{i1}}{p_{i.}},\frac{p_{i2}}{p_{i.}},\ldots,\frac{p_{ip}}{p_{i.}}\right)$$ i=1,2,…,n

称为n个p维空间中样本点，$\sum_{j=1}^{p}\frac{p_{ij}}{p_{i.}}=1$

研究两个样本点K, L之间的欧氏距离。

$$D^2(K,L)=\sum_{j=1}^{p}\left(\frac{p_{kj}}{p_{k.}}-\frac{p_{Lj}}{p_{L.}}\right)^2$$

加权距离，可以消除数量级的影响，

$$D^2.(K,L)=\sum_{j=1}^{p}\left(\frac{p_{kj}}{p_{k.}}-\frac{p_{Lj}}{p_{L.}}\right)^2 / p_{.j}$$

$$=\sum_{j=1}^{p}\left(\frac{p_{kj}}{\sqrt{p_{.j}\,p_{k.}}}-\frac{p_{Lj}}{\sqrt{p_{.j}\,p_{k.}}}\right)^2$$

可以理解成n个样本点第 i个样本的座标变为

$$\left(\frac{p_{i1}}{\sqrt{p_{.1}\,p_{i.}}},\frac{p_{i2}}{\sqrt{p_{.2}\,p_{i.}}},\cdots\frac{p_{ip}}{\sqrt{p_{.p}\,p_{i.}}}\right)\quad \text{i=1,2,…,n}$$

两个样本点K,L的距离为

$$D_{.}^{2}(K,L)=\sum_{j=1}^{p}\left(\frac{p_{kj}}{\sqrt{p_{.j}\,p_{k.}}}-\frac{p_{Lj}}{\sqrt{p_{.j}\,p_{k.}}}\right)^{2}$$

同理从列的方向看，可以将样本矩阵中的列看成是 n维空间中的点，变量Xi列为

$$P=\begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1p} \\ p_{21} & p_{22} & \cdots & p_{2p} \\ \cdot & \cdot & \cdot & \cdot \\ p_{n1} & p_{n2} & & p_{np} \end{pmatrix} \Rightarrow \begin{pmatrix} \frac{p_{11}}{p_{.1}} & \frac{p_{12}}{p_{.2}} & \cdots & \frac{p_{1p}}{p_{.p}} \\ \frac{p_{21}}{p_{.1}} & \frac{p_{22}}{p_{.2}} & \cdots & \frac{p_{2p}}{p_{.p}} \\ \cdot & \cdot & \cdot & \cdot \\ \frac{p_{n1}}{p_{.1}} & \frac{p_{n2}}{p_{.2}} & & \frac{p_{np}}{p_{.p}} \end{pmatrix}$$

设两个变量Xi与Xj的距离为

$$D_{\cdot}^{2}(i,j)=\sum_{k=1}^{n}\left(\frac{p_{ki}}{p_{\cdot i}}-\frac{p_{kj}}{p_{\cdot j}}\right)^{2}$$

加权距离，可以消除数量级的影响，

$$D_{\cdot}^{2}(i,j)=\sum_{k=1}^{n}\left(\frac{p_{ki}}{p_{\cdot i}}-\frac{p_{kj}}{p_{\cdot j}}\right)^{2}/\,p_{k\cdot}$$

$$=\sum_{k=1}^{n}\left(\frac{p_{ki}}{\sqrt{p_{k\cdot}\,p_{\cdot i}}}-\frac{p_{kj}}{\sqrt{p_{k\cdot}\,p_{\cdot j}}}\right)^{2}$$

矩阵变为

$$\Rightarrow \begin{pmatrix} \frac{p_{11}}{\sqrt{p_{1.}p_{.1}}} & \frac{p_{12}}{\sqrt{p_{1.}p_{.2}}} & \cdots & \frac{p_{1p}}{\sqrt{p_{1.}p_{.p}}} \\ \frac{p_{21}}{\sqrt{p_{2.}p_{.1}}} & \frac{p_{22}}{\sqrt{p_{2.}p_{.2}}} & \cdots & \frac{p_{2p}}{\sqrt{p_{2.}p_{.p}}} \\ \cdot & \cdot & \cdot & \cdot \\ \frac{p_{n1}}{\sqrt{p_{n.}p_{.1}}} & \frac{p_{n2}}{\sqrt{p_{n.}p_{.2}}} & & \frac{p_{np}}{\sqrt{p_{n.}p_{.p}}} \end{pmatrix}$$

求各列的加权平均值

$$\left(\sqrt{p_{.1}} \quad \sqrt{p_{.2}} \quad \cdots \quad \sqrt{p_{.p}}\right)$$

因为变量均值,由于 $\sum_{i=1}^{n} p_{i.} = 1$

$$\sum_{i=1}^{n} \frac{p_{ij}}{\sqrt{p_{.j} \cdot p_{i.}}} p_{i.} = \frac{1}{\sqrt{p_{.j}}} \sum_{i=1}^{n} p_{ij}$$

$$= \frac{p_{.j}}{\sqrt{p_{.j}}} = \sqrt{p_{.j}} \qquad E(x) = \sum_{i=1}^{n} p_i x_i$$

这是按概率加权平均

因为 协方差公式为

$$\operatorname{cov}(X,Y)=E(x-\bar{x})(y-\bar{y})$$

$$=\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})\,p_i$$

所以从矩阵

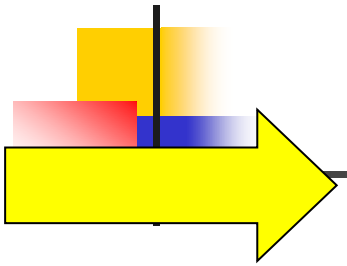
$$\begin{pmatrix} \frac{p_{11}}{\sqrt{p_{1.}\,p_{.1}}} & \frac{p_{12}}{\sqrt{p_{1.}\,p_{.2}}} & \cdots & \frac{p_{1p}}{\sqrt{p_{1.}\,p_{.p}}} \\ \frac{p_{21}}{\sqrt{p_{2.}\,p_{.1}}} & \frac{p_{22}}{\sqrt{p_{2.}\,p_{.2}}} & \cdots & \frac{p_{2p}}{\sqrt{p_{2.}\,p_{.p}}} \\ \cdot & \cdot & \cdot & \cdot \\ \frac{p_{n1}}{\sqrt{p_{n.}\,p_{.1}}} & \frac{p_{n2}}{\sqrt{p_{n.}\,p_{.2}}} & & \frac{p_{np}}{\sqrt{p_{n.}\,p_{.p}}} \end{pmatrix}$$

两个变量Xi, Xj的协方差

$$a_{ij}=\sum_{\alpha=1}^{n}\left(\frac{p_{\alpha i}}{\sqrt{p_{.j}\cdot p_{\alpha.}}}-\sqrt{p_{.i}}\right)\left(\frac{p_{\alpha j}}{\sqrt{p_{.j}\cdot p_{\alpha.}}}-\sqrt{p_{.j}}\right)p_{\alpha.}$$

$$=\sum_{\alpha=1}^{n}\left(\frac{p_{\alpha i}-p_{.i}p_{\alpha.}}{\sqrt{p_{.i}\cdot p_{\alpha.}}}\right)\left(\frac{p_{\alpha j}-p_{.j}p_{\alpha.}}{\sqrt{p_{.j}\cdot p_{\alpha.}}}\right)$$

P个变量的协方差 $A_{p\times p}=(a_{ij})$



$$\begin{pmatrix} \frac{p_{11}-p_{.1}p_{1.}}{\sqrt{p_{.1}\cdot p_{1.}}} & \frac{p_{12}-p_{.2}p_{1.}}{\sqrt{p_{.1}\cdot p_{1.}}} & \cdots & \frac{p_{1p}-p_{.p}p_{1.}}{\sqrt{p_{.p}\cdot p_{1.}}} \\ \frac{p_{21}-p_{.1}p_{2.}}{\sqrt{p_{.1}\cdot p_{2.}}} & \frac{p_{22}-p_{.2}p_{2.}}{\sqrt{p_{.2}\cdot p_{2.}}} & \cdots & \frac{p_{2p}-p_{.p}p_{2.}}{\sqrt{p_{.p}\cdot p_{2.}}} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{p_{n1}-p_{.1}p_{n.}}{\sqrt{p_{.1}\cdot p_{n.}}} & \frac{p_{n2}-p_{.2}p_{n.}}{\sqrt{p_{.2}\cdot p_{n.}}} & & \frac{p_{np}-p_{.p}p_{n.}}{\sqrt{p_{.p}\cdot p_{n.}}} \end{pmatrix}$$

$$= \begin{pmatrix} z_{11} & z_{12} & \cdots & z_{1p} \\ z_{21} & z_{22} & \cdots & z_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ z_{n1} & z_{n2} & \cdots & z_{np} \end{pmatrix}$$

$$a_{ij} = \sum_{\alpha=1}^{n} \left(\frac{p_{\alpha i} - p_{.i} p_{\alpha .}}{\sqrt{p_{.i} \cdot p_{\alpha .}}} \right) \left(\frac{p_{\alpha i} - p_{.i} p_{\alpha .}}{\sqrt{p_{.i} \cdot p_{\alpha .}}} \right)$$

$$= \sum_{\alpha=1}^{n} z_{\alpha i} z_{\alpha j}$$

$$z_{\alpha i} = \frac{p_{\alpha i} - p_{.i} p_{\alpha .}}{\sqrt{p_{.i} \cdot p_{\alpha .}}} = \frac{\frac{x_{\alpha i}}{x_{..}} - \frac{x_{.i}}{x_{..}} \cdot \frac{x_{\alpha .}}{x_{..}}}{\sqrt{\frac{x_{.i}}{x_{..}} \cdot \frac{x_{\alpha .}}{x_{..}}}}$$

$$= \frac{x_{\alpha i} - \frac{x_{.i} x_{\alpha .}}{x_{..}}}{\sqrt{x_{.i} x_{\alpha .}}}$$

$$Z = (z_{ij})_{n \times p}$$

$$a_{ij} = \sum_{\alpha=1}^{n} z_{\alpha i} z_{\alpha j} = \begin{pmatrix} z_{1i} & z_{2i} & \cdots & z_{ni} \end{pmatrix} \begin{pmatrix} z_{1i} \\ z_{2j} \\ \cdots \\ z_{nj} \end{pmatrix}$$

$$\underset{p \times p}{A} = (a_{ij}) = Z'Z$$

类似对变量的方法,对样本点好可得协方差矩阵

$$\underset{n\times n}{B} = \underset{n\times p}{Z}\ \underset{p\times n}{Z'}$$

A与B通过Z矩阵联系起来了,存在对应关系

$$\underset{p\times p}{A} = Z'Z$$

A与B的非零特征根相同

因为有

$$Z'ZU = \lambda U$$

$$ZZ'(ZU) = \lambda(ZU)$$ 两边右乘Z

$$AU = \lambda U$$

$$B(ZU) = \lambda(ZU)$$ 有相同的特征根

A=Z'Z特征值 $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$

对A进行因子分析，求得因子载荷矩阵

$$F = \begin{pmatrix} u_{11}\sqrt{\lambda_1} & u_{12}\sqrt{\lambda_2} & \ldots & u_{1m}\sqrt{\lambda_m} \\ u_{21}\sqrt{\lambda_1} & u_{22}\sqrt{\lambda_2} & \ldots & u_{2m}\sqrt{\lambda_m} \\ \ldots & \ldots & \ldots & \ldots \\ u_{p1}\sqrt{\lambda_1} & u_{p2}\sqrt{\lambda_2} & \ldots & u_{mm}\sqrt{\lambda_m} \end{pmatrix}$$

对前两个因子载荷作图。

特征值 $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$

对B进行因子分析，求得因子载荷矩阵

$$G = \begin{pmatrix} v_{11}\sqrt{\lambda_1} & v_{12}\sqrt{\lambda_2} & \cdots & v_{1m}\sqrt{\lambda_m} \\ v_{21}\sqrt{\lambda_1} & v_{22}\sqrt{\lambda_2} & \cdots & v_{2m}\sqrt{\lambda_m} \\ \cdots & \cdots & \cdots & \cdots \\ v_{p1}\sqrt{\lambda_1} & v_{p2}\sqrt{\lambda_2} & \cdots & v_{mm}\sqrt{\lambda_m} \end{pmatrix}$$

对前两个因子载荷 作图。

对应图

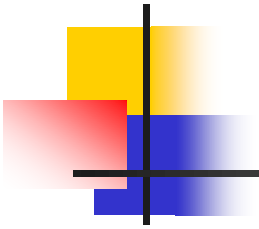
我们知道因子载荷矩阵的含义是原始数据与公共因子之间的相关系数，所以如果我们构造一个平面直角坐标系，将第一公共因子的载荷与第二个公共因子的载荷看成平面上的点，在坐标系中绘制散点图，则构成对应图。

某地环境检测部门对该地所属8个地区的大气污染状况进行了系统的的检测，每天4次同时在各个地区抽取大气样品，则定其中的氯、硫化氢、二氧化硫、碳4、环氧氯丙烷、环已烷6种气体的浓度。数据资料略。

	特征根	贡献率（%）	累 积 贡 献 率（%）
1	**0.50668**	**70.00**	**70.00**
2	**0.12213**	**16.87**	**86.87**
3	**0.05658**	**7.82**	**94.69**

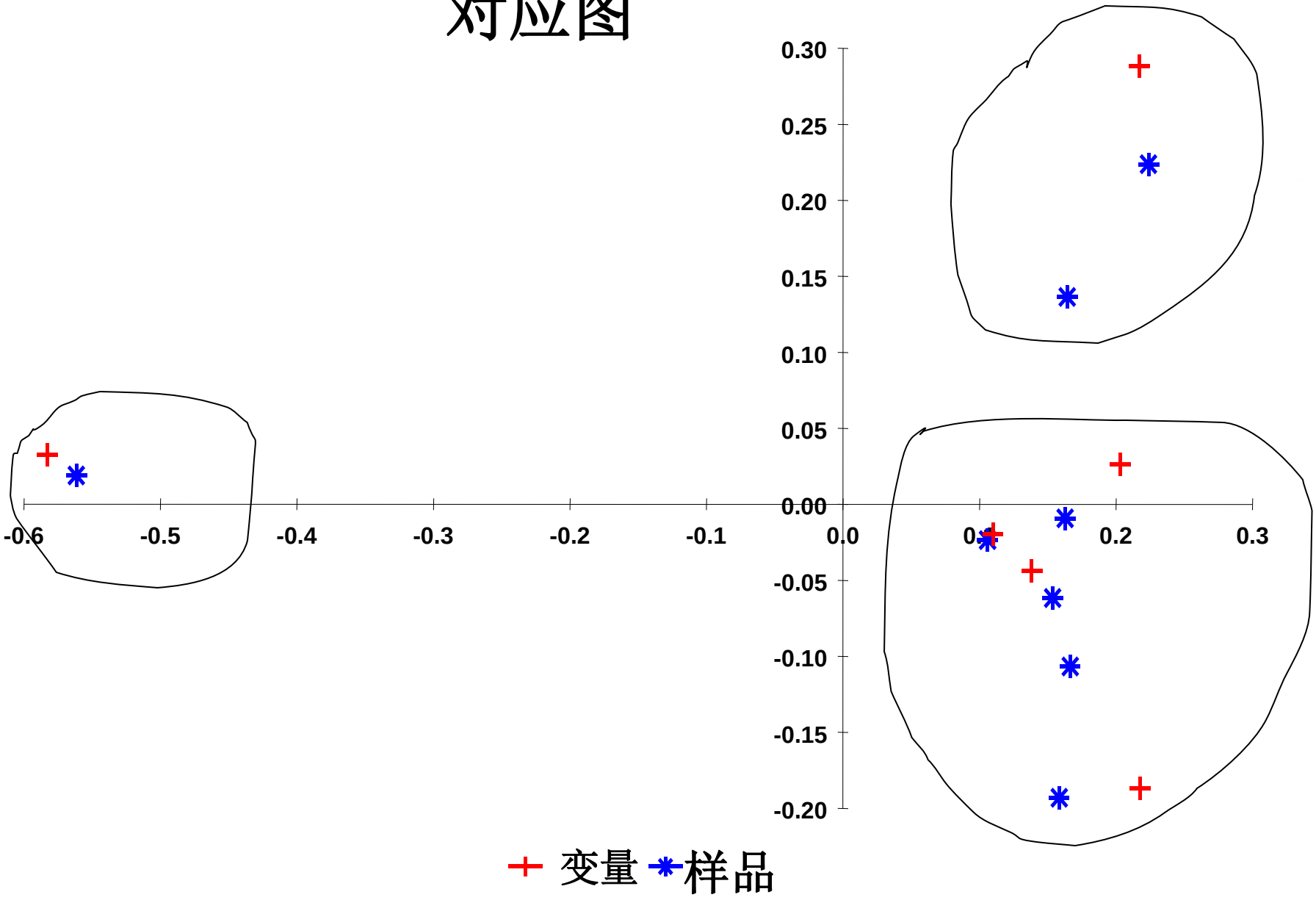
R型因子分析的载荷	
F1	F2
0.13831	-0.04385
0.20333	0.02650
0.11003	-0.01985
0.21754	-0.18687
0.21720	0.28831
-0.58275	0.03279

Q型因子分析的载荷	
G1	G2
0.10599	-0.02354
0.15369	-0.06164
0.16284	-0.00928
0.22388	0.22377
0.15853	-0.19307
-0.56153	0.01900
0.16656	-0.10664
0.16429	0.13644



在同一个直角坐标系中作出两种因子的载荷图，这种图称为对应图。

对应图


0.30
0.25
0.20
0.15
0.10
0.05
0.00
-0.05
-0.10
-0.15
-0.20
-0.6
-0.5
-0.4
-0.3
-0.2
-0.1
0.0
0.2
0.3
变量
样品

由图我们可以看出，全部变量与样品分为3类。每一类聚合一些变量和样品。

第一类：聚合了环氧氯丙烷X5和D和H两个地区，表明D和H两个地区主要大气污染物为环氧氯丙烷。

第二类：包含变量X1，X2，X3，X4和样品A，B，C，E和G地区，这5个地区的主要污染物是氯、硫化氢、二氧化硫、碳4。

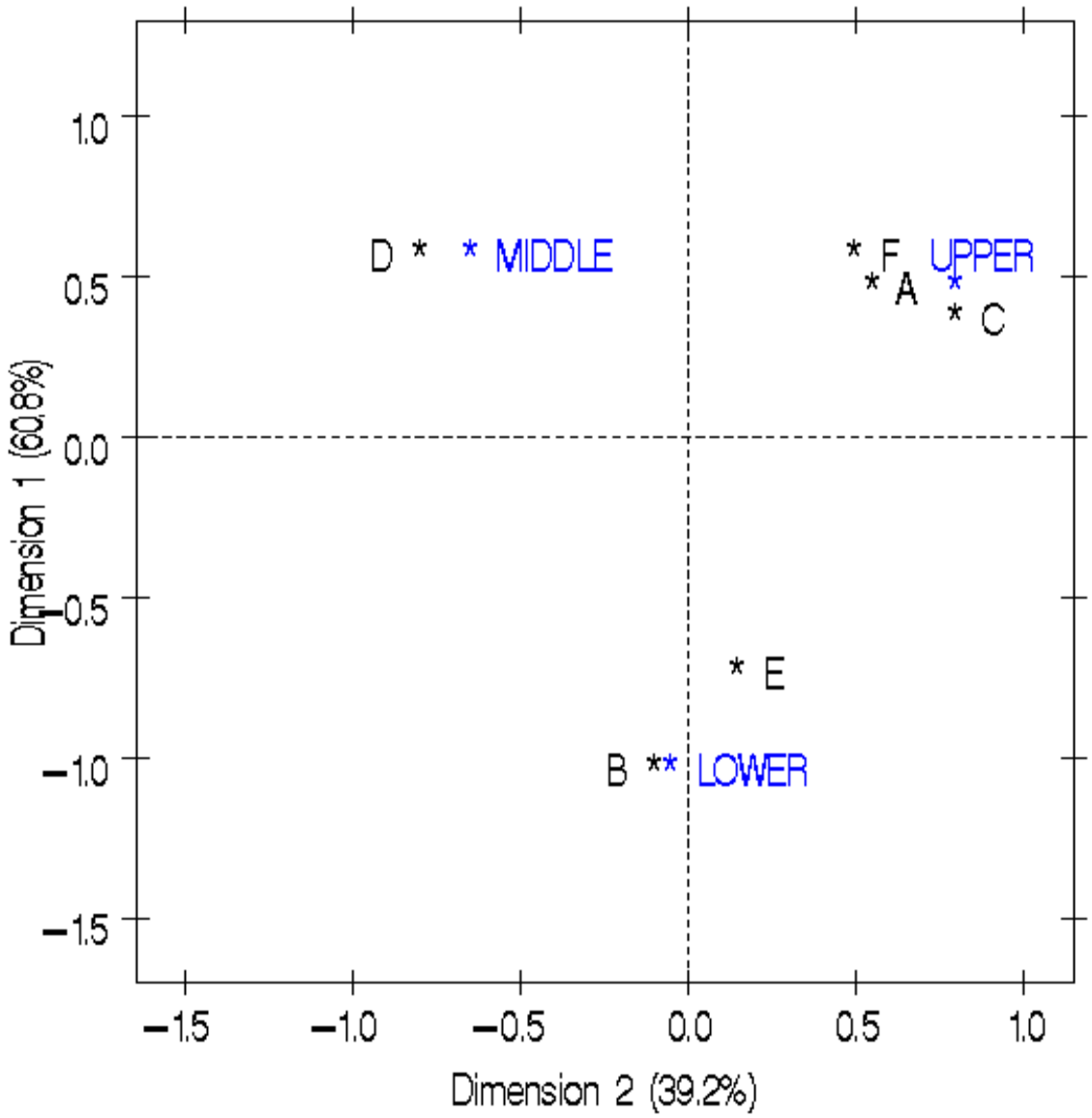
第三类：包含X6和地区F，该地区的主要污染物是环已烷。

案例：检验不同收入的消费者对品牌的选择

BRAND	LOWER	MIDDLE	UPPER	TOTAL
A	2	7	16	25
B	49	7	3	59
C	4	5	23	32
D	4	49	5	58
E	15	2	5	22
F	1	7	14	22
TOTAL	75	77	66	218

1. 低收入消费者与品牌B和E
2. 中等收入消费者与品牌D
3. 高收入消费者与品牌A、C和 F

SAS Market 模块



对应分析的基本原理

- 列联表(contingency table)

表中的每一行或每一列分别对应于一个行向量（点）或列向量（点）；分别将行和列的概率（百分比）看成是空间行点和列点的分量，称这些点为行轮廓和列轮廓。

列联表（交互表）是最常见的对应表的一种形式，通过交互分析得到行、列百分比，对于仅含有少量类的变量，通过这样的简单统计，就可以看出行、列变量之间的一些关系。

BRAND * AREA Crosstabulation

			AREA			Total
			1 地区1	2 地区2	3 地区3	
BRAND	1 品牌A	Count	5	5	30	40
		Expected Count	13.3	13.0	13.7	40.0
		% within BRAND	12.5%	12.5%	75.0%	100.0%
		% within AREA	12.5%	12.8%	73.2%	33.3%
	2 品牌B	Count	5	25	5	35
		Expected Count	11.7	11.4	12.0	35.0
		% within BRAND	14.3%	71.4%	14.3%	100.0%
		% within AREA	12.5%	64.1%	12.2%	29.2%
	3 品牌C	Count	15	5	5	25
		Expected Count	8.3	8.1	8.5	25.0
		% within BRAND	60.0%	20.0%	20.0%	100.0%
		% within AREA	37.5%	12.8%	12.2%	20.8%
	4 品牌D	Count	15	4	1	20
		Expected Count	6.7	6.5	6.8	20.0
		% within BRAND	75.0%	20.0%	5.0%	100.0%
		% within AREA	37.5%	10.3%	2.4%	16.7%
Total		Count	40	39	41	120
		Expected Count	40.0	39.0	41.0	120.0
		% within BRAND	33.3%	32.5%	34.2%	100.0%
		% within AREA	100.0%	100.0%	100.0%	100.0%

行轮廓

列轮廓

- 主成分（Principal components)

通过主成分分析，在以两个主成分为坐标轴的空间中，表出行轮廓或列轮廓，或同时标出行、列轮廓，从而探索它们之间的关系。

这种近似地表示行列轮廓的图形叫对应图（correspondence plot)。

- **惯量(inertias)和特征值(eigenvalues)**
  - 惯量是度量行轮廓和列轮廓的变差的统计量，总惯量表示轮廓点的全部变差；
  - 作图用的前两维度分别对应两个主惯量（principal inertias)，表示在坐标轴方向上的变差；
  - 主惯量就是对行轮廓和列轮廓作主成分分析时得到的特征值，特征值的平方根叫做奇异值（singular values)。

- 卡方(Chi-square)和列联系数(contingency coefficient)

  是检验对应分析显著性或近似效果的统计量

- 数据的格式要求

  对应分析数据的典型格式是列联表或交叉频数表。

  常表示不同背景的消费者对若干产品或产品的属性的选择频率。

  背景变量或属性变量可以并列使用或单独使用。
  两个变量间——简单对应分析。
  多个变量间——多元对应分析。

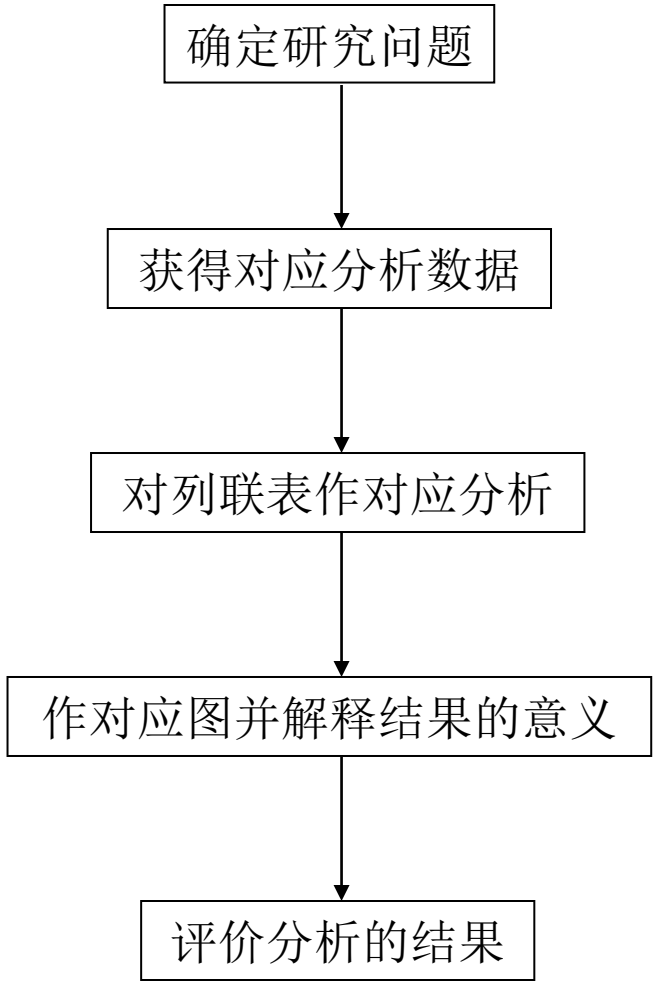
■对应图的维度（坐标轴）

分析的二维表中有r行、c列，即行变量有r类，列变量有c类，那么对应分析中所用的维度数目应为（r-1），（c-1）中的最小值，

我们将其记作：min（r，c）-1。也就是说，我们可以在min（r，c）-1维空间中非常好地描绘行变量的r类和列变量的c类。

例如，某一列联表中有5行，4列，则维度的最大值是min（5，4）-1=3。但是从实用的角度来讲，我们可以在较低维度下，例如用二维空间来描绘行变量和列变量的类别，很显然，二维空间非常易于理解，而多维空间则不然。在通常情况下，两个维度就可以比较好地解释行变量与列变量。

对应分析的主要步骤



简单的对应分析——二维的列联表

案例：一家音像连锁店，为了广告目的，需要了解不同类型的顾客在选择借租节目带类型的关系。

性别 gender	年龄 age	节目名称 movies	数量 count
F	50	MYSTERY 神话	37
F	60	DRAMA 戏曲	702
F	35	HORROR 恐怖	44
M	35	FAMILY 家庭	84
M	45	ACTION 动作	347
F	30	ROMANCE 爱情	30
F	20	SPORTS 体育	24
….	….	….	….
F	65	COMEDY 喜剧	86

三个变量——gender age movies

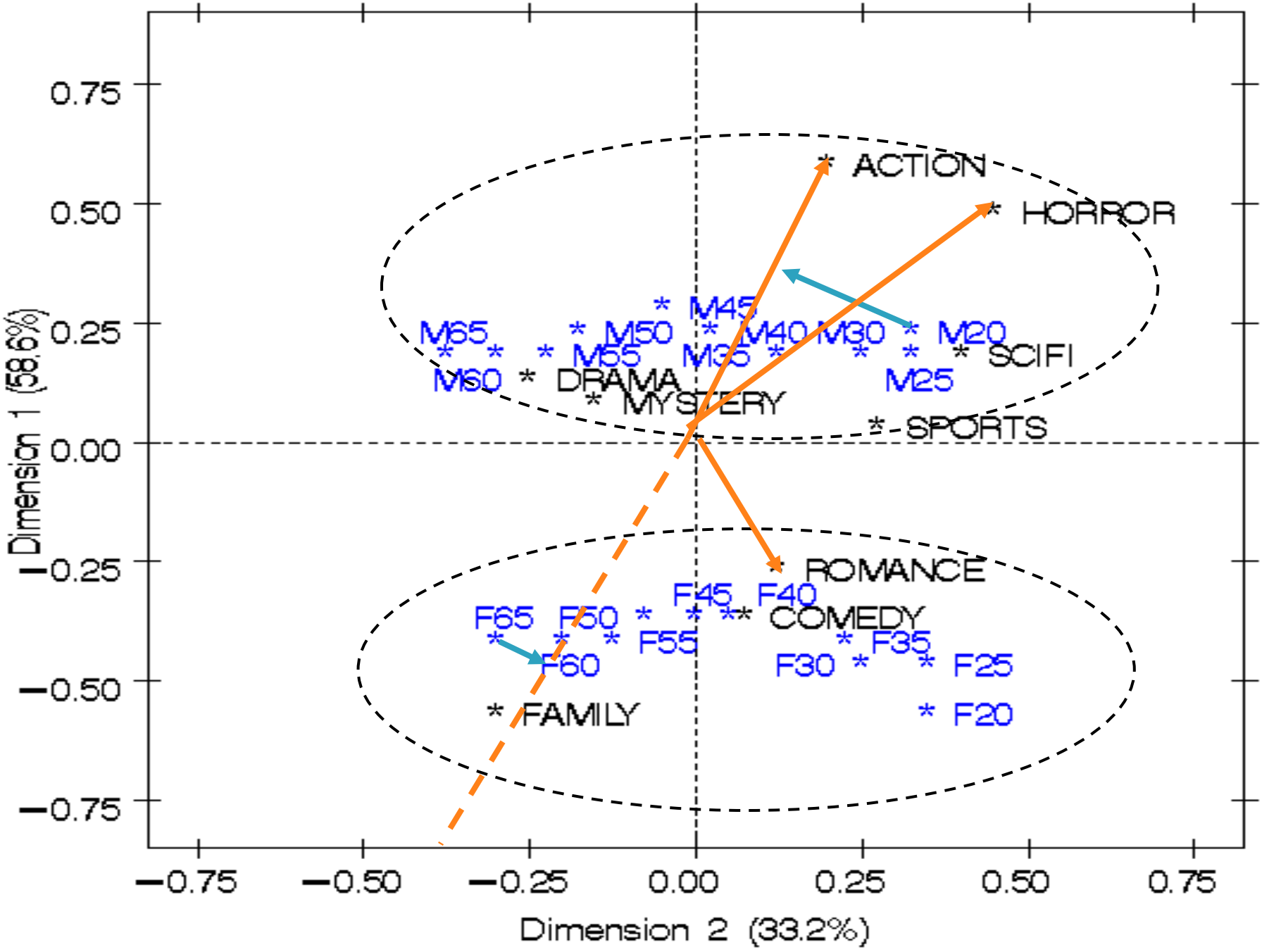
	F20	F25	M20	M25
DRAMA				
ROMANCE				

男Males

	20	25
DRAMA		
ROMANCE		

女Females

	20	25
DRAMA		
ROMANCE		

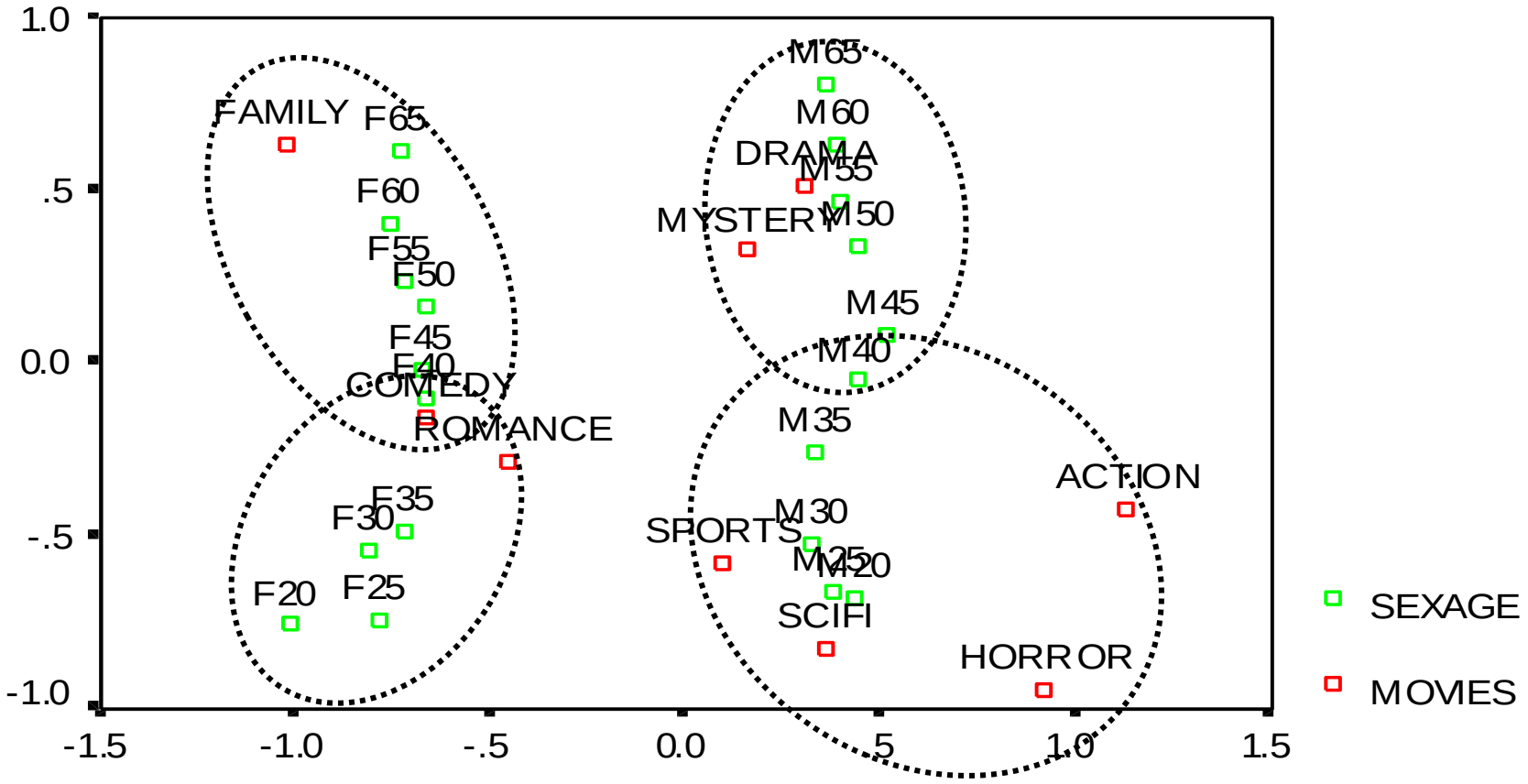

0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
Dimension 1 (58.6%)
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
Dimension 2 (33.2%)
ACTION
HORROR
SCIFI
DRAMA
MYSTERY
SPORTS
ROMANCE
COMEDY
FAMILY
M65
M60
M55
M50
M45
M40
M35
M30
M25
M20
F65
F60
F55
F50
F45
F40
F35
F30
F25
F20

同样的方法采用spss

- 我们一起来操作

Row and Column Points

Symmetrical Normalization



Summary

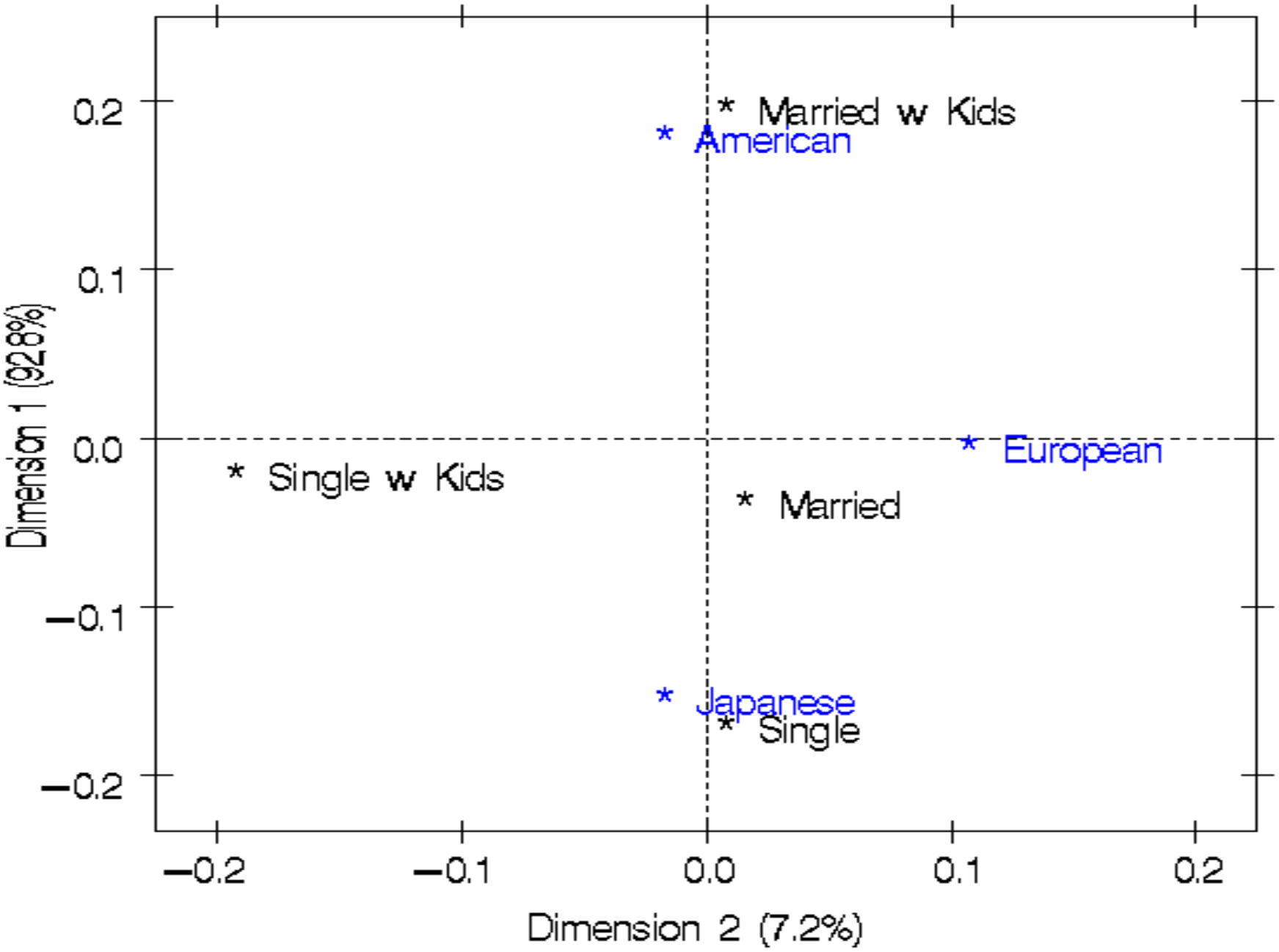
Dimension	Singular Value	Inertia	Chi Square	Sig.	Proportion of Inertia		Confidence Singular Value	
					Accounted for	Cumulative	Standard Deviation	Correlation 2
1	.299	.089			.586	.586	.004	.003
2	.225	.051			.332	.918	.005	
3	.089	.008			.052	.970		
4	.048	.002			.015	.986		
5	.029	.001			.006	.991		
6	.026	.001			.004	.996		
7	.023	.001			.003	.999		
8	.012	.000			.001	1.000		
Total		.152	6386.864	.000[a]	1.000	1.000		

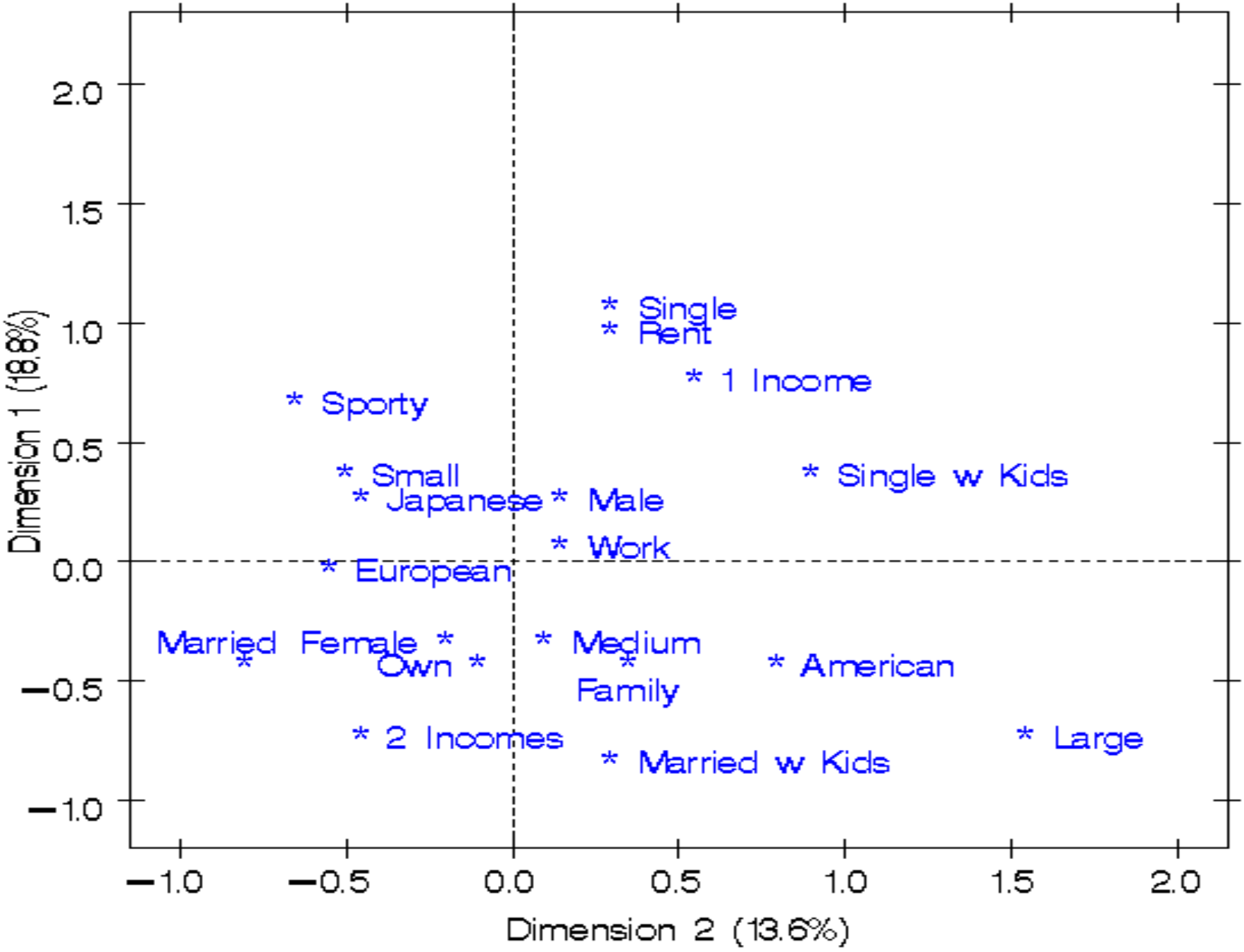
a. 152 degrees of freedom.

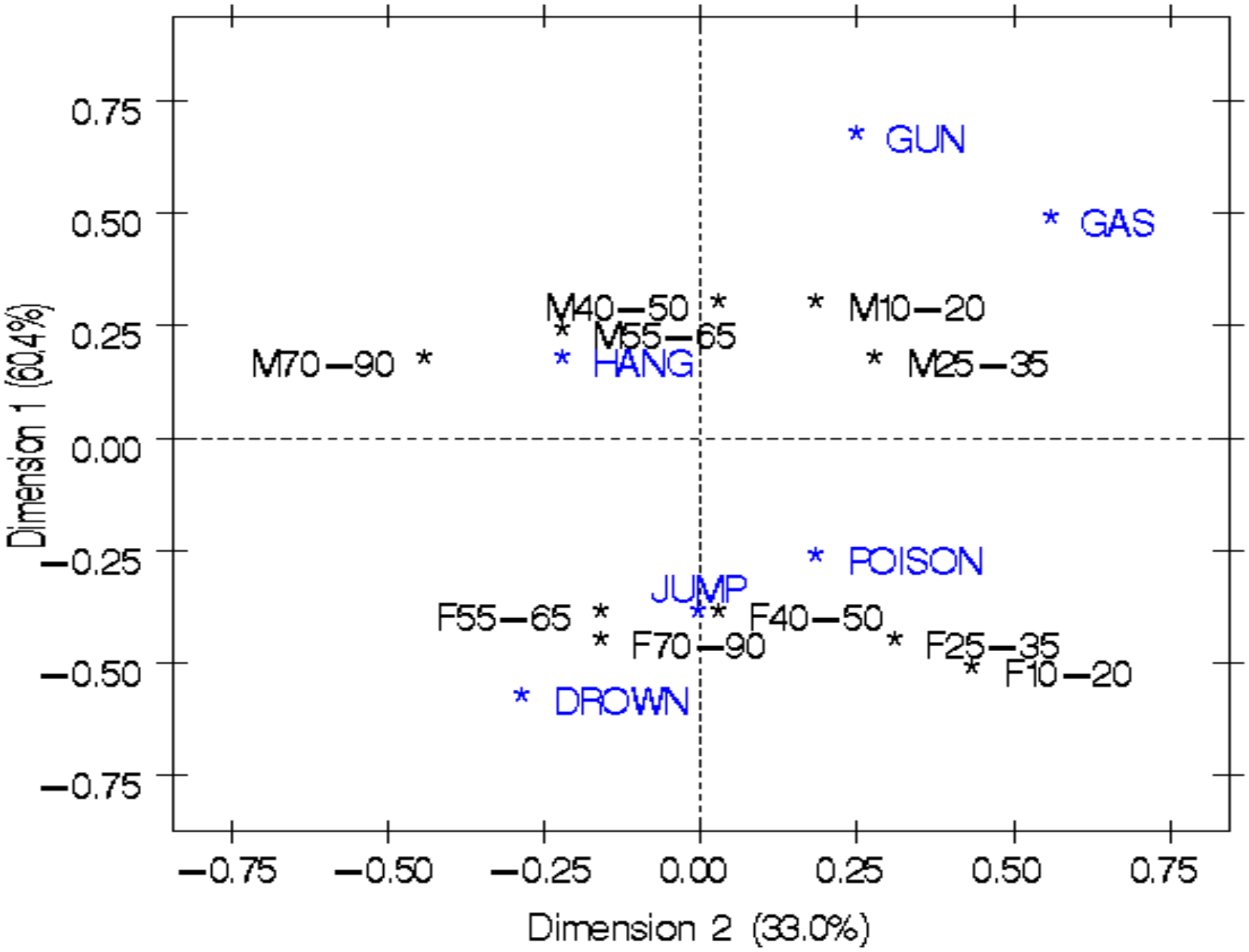
多元对应分析

案例：不同家庭背景的消费者对产地的汽车选择
SAS market 模块

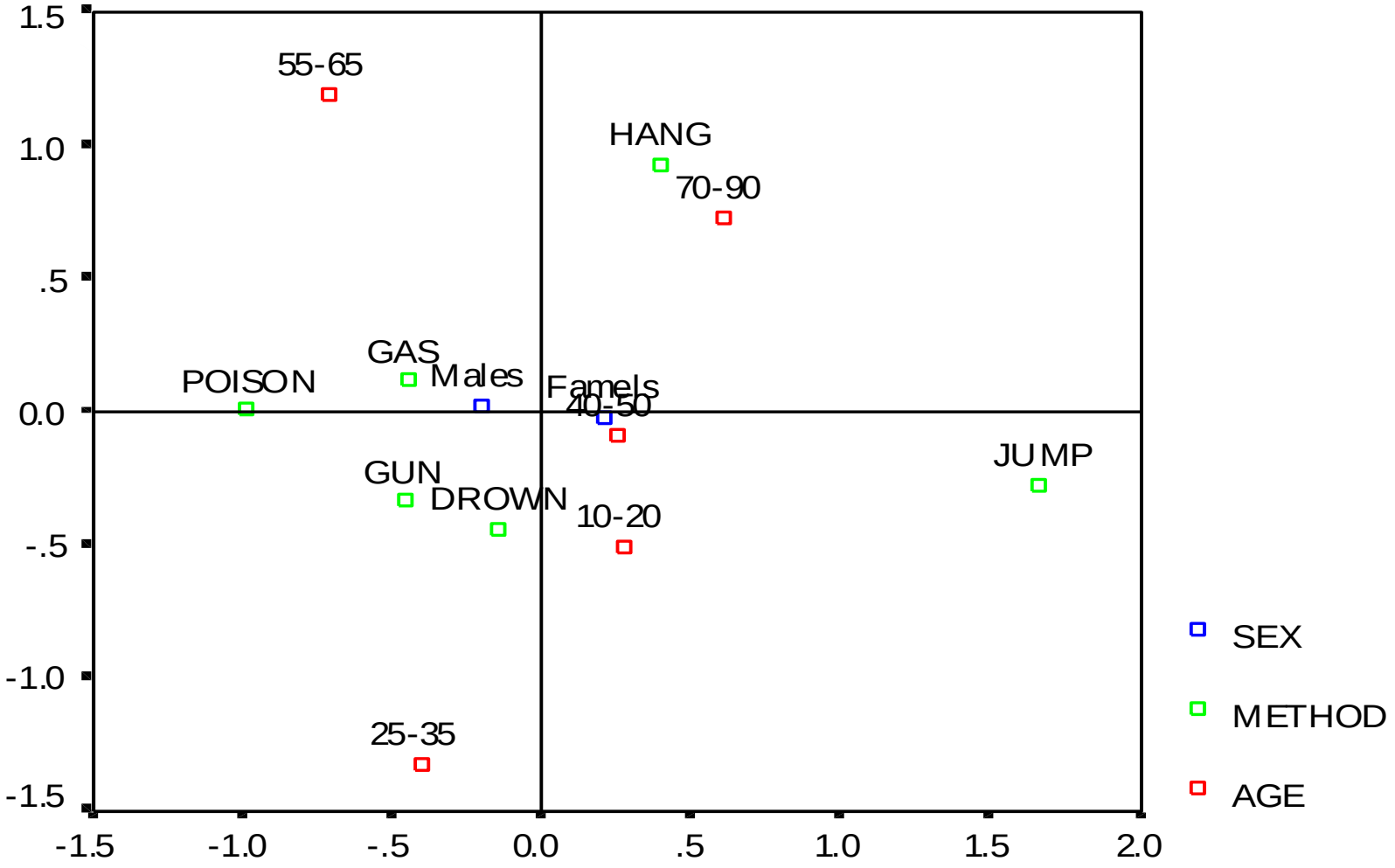
案例：不同年龄和性别人选择自杀的方法
SPSS correspondence analysis 模块


* Married w Kids
* American
* European
* Single w Kids
* Married
* Japanese
* Single
0.2
0.1
0.0
−0.1
−0.2
−0.2
−0.1
0.0
0.1
0.2
Dimension 1 (92.8%)
Dimension 2 (7.2%)


Single
Rent
1 Income
Sporty
Small
Japanese
Male
Single w Kids
Work
European
Married
Female
Own
Medium
Family
American
2 Incomes
Married w Kids
Large
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
Dimension 1 (18.8%)
Dimension 2 (13.6%)

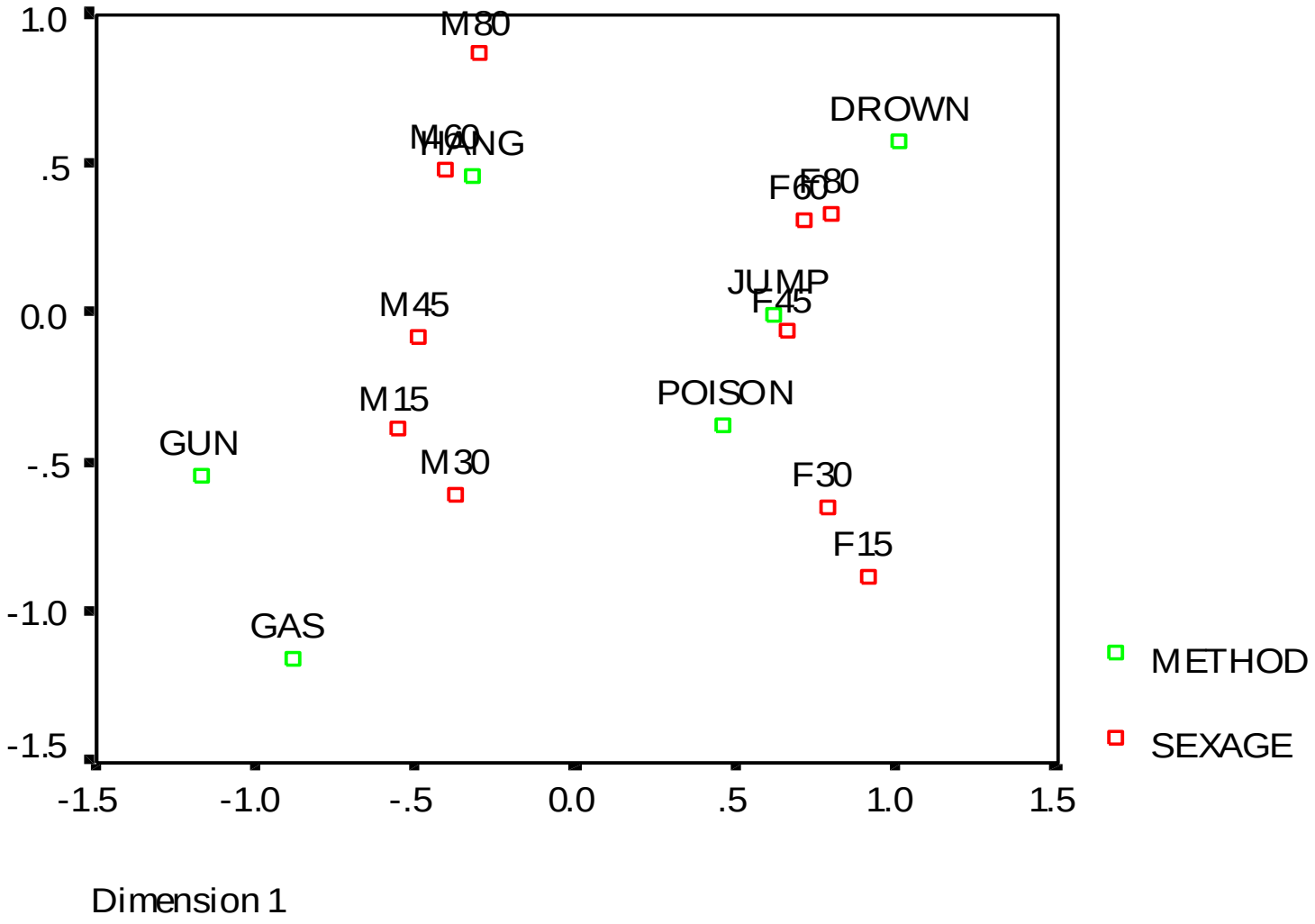

0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
Dimension 1 (60.4%)
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
Dimension 2 (33.0%)
GUN
GAS
M40−50
M10−20
M55−65
M70−90
HANG
M25−35
POISON
JUMP
F55−65
F40−50
F70−90
F25−35
F10−20
DROWN

Category Quantifications



Row and Column Points

Symmetrical Normalization



对应分析的优点

1. 定性变量划分的类别越多，这种方法的优势越明显。
2. 揭示行变量类别间与列变量类别间的联系。
3. 将类别联系直观地表现在二维图形中（对应图）。
4. 可以将名义变量或次序变量转变为间距变量。

对应分析的局限

1. 不能用于相关关系的假设检验。
2. 维度要由研究者决定。
3. 有时候对应图解释比较困难。
4. 对极端值比较敏感。

一个实际应用案例——新产品名称的测试

对新产品来说，产品名称是消费者认识和识别该产品的核心要素，是形成品牌概念的基础。为新产品起一个好的名字是非常重要的，好的名字至少应该满足下列两个条件：

1. 名字应该使消费者联想到正确的产品。

2. 名字应该使消费者有最接近正确产品的感觉。

拟定中的新产品名称“波澜”同其它7个模拟的名称一起测试。

问卷中的问题如下：

下面我将列出一些名词：

请您判断一下它们最象什么商品的名称？*（出示卡片，只选一项）*

1. 雪糕　2. 纯水　3. 碳酸饮料　4. 果汁饮料　5. 保健食品　6. 空调
7. 洗衣机　8. 毛毯　9. 其它

这些名称最能使您产生什么感觉？*（出示卡片，只选一项）*

1. 清爽　2. 甘甜　3. 欢快　4. 纯净　5. 安闲　6. 个性
7. 兴奋　8. 高档　9. 其它

名称 X、产品联想 Y 和感觉 Z 的对应表----频数

16×8 的列联表

	玉泉	雪源	春溪	期望	波澜	天山绿	中美纯	雪浪花
雪糕	50	442	27	21	14	50	20	258
纯水	508	110	272	51	83	88	605	79
碳酸饮料	55	68	93	36	71	47	37	77
果汁饮料	109	95	149	41	36	125	43	65
保健食品	34	29	45	302	37	135	42	18
空调	11	28	112	146	113	39	20	31
洗衣机	20	12	54	64	365	13	8	210
毛毯	2	4	17	36	29	272	9	35
清爽	368	322	167	53	57	129	149	170
甘甜	237	237	142	41	34	95	119	116
欢快	19	25	185	105	123	44	22	193
纯净	142	140	128	47	38	123	330	68
安闲	16	16	106	166	81	164	21	36
个性	2	14	9	72	94	41	37	42
兴奋	4	11	10	78	248	35	17	81
高档	3	5	19	107	63	126	63	49

对于一个16×8的列联表，用7-维空间才能完满地对其进行解释（100%），但是，可能有70%的对应表的惯量刚好保留在二维空间，从实用角度讲，用二维空间解释是可以接受的，也更易于理解。

在品牌测试中，有三个名义变量X，Y，Z，我们可以简单地将第三个变量Z附加在第二个变量Y上，构造一个新的二维表格，

	X
Y	
Z	

对应分析可以采用两种方法处理上面的表格：

（1）在分析过程中，将变量Y和Z作为一个新的行变量（处置1）；

（2）对应分析只基于变量X和Y，而将变量Z作为附加行变量（处置2），

这意味着变量Z的各类只是空间中的附加点，并不影响二维空间的属性，包括行、列得分，维度，坐标轴和方向。

这里X=名称，Y=产品，Z=感觉，其中Y+Z=形象，作为一个新的变量。

	X
Y	

（1）

	X
Z	

（2）

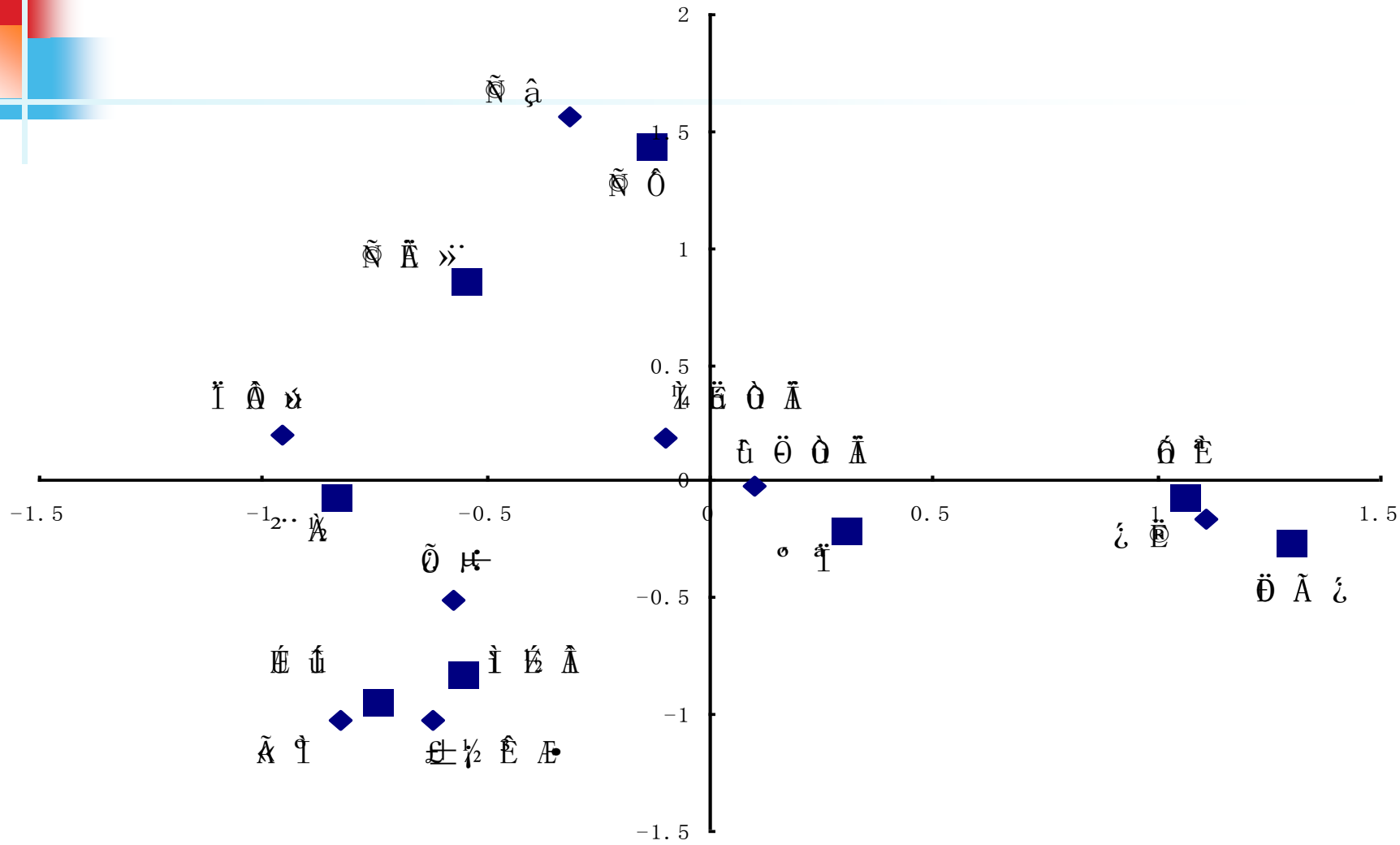
两种情况都有8行和8列，如果数据是完全随机的没有显著的依赖关系，则从每个轴抽取的平均惯量应该能解释总惯量的100/（8-1）=14.3%，因此，任何贡献大于14.3%的轴都被认为是重要的、不宜省略的，应该包括在解的空间中。

	x
Y	
Z	

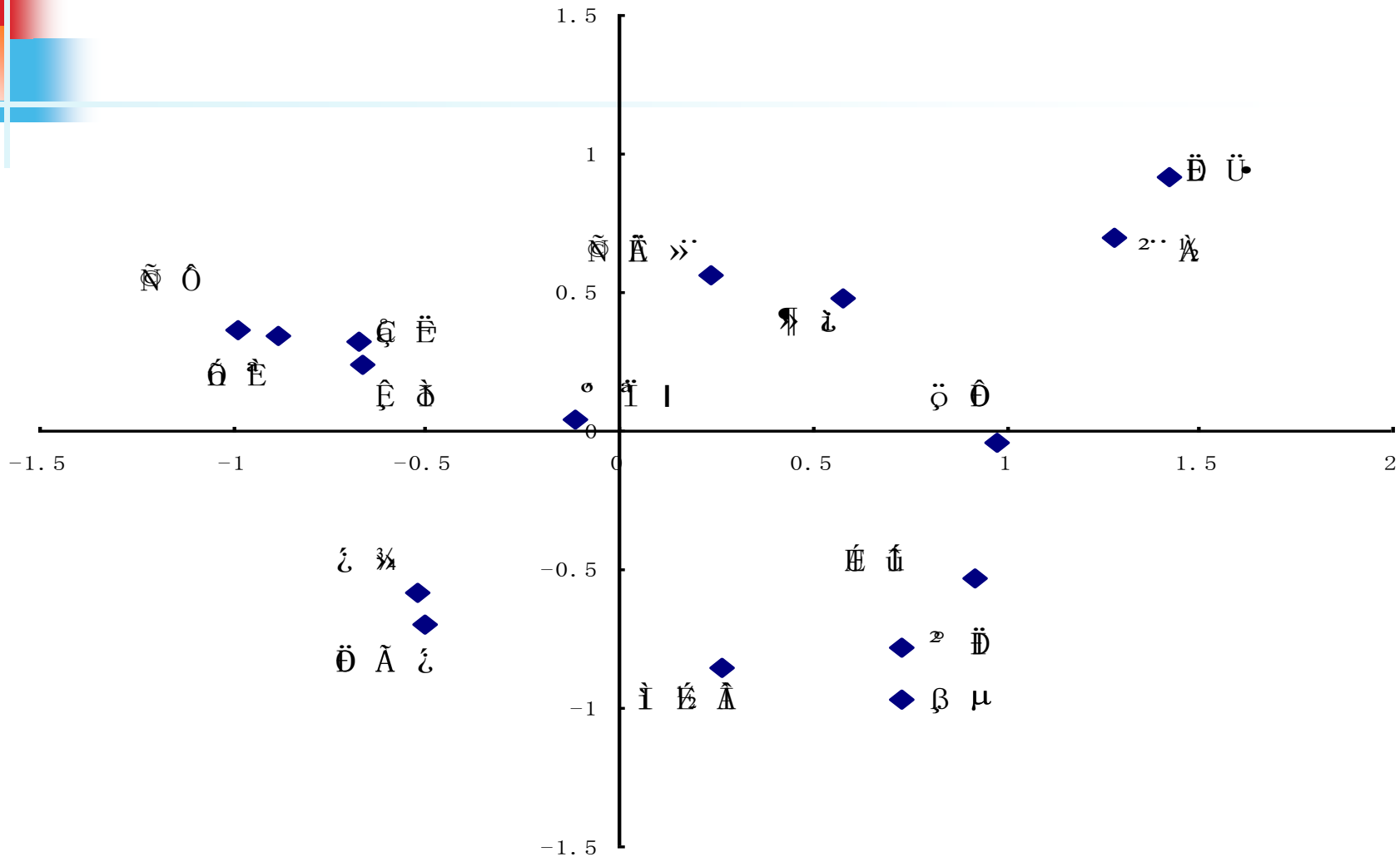
（3_1）

	x
Y	
Z	

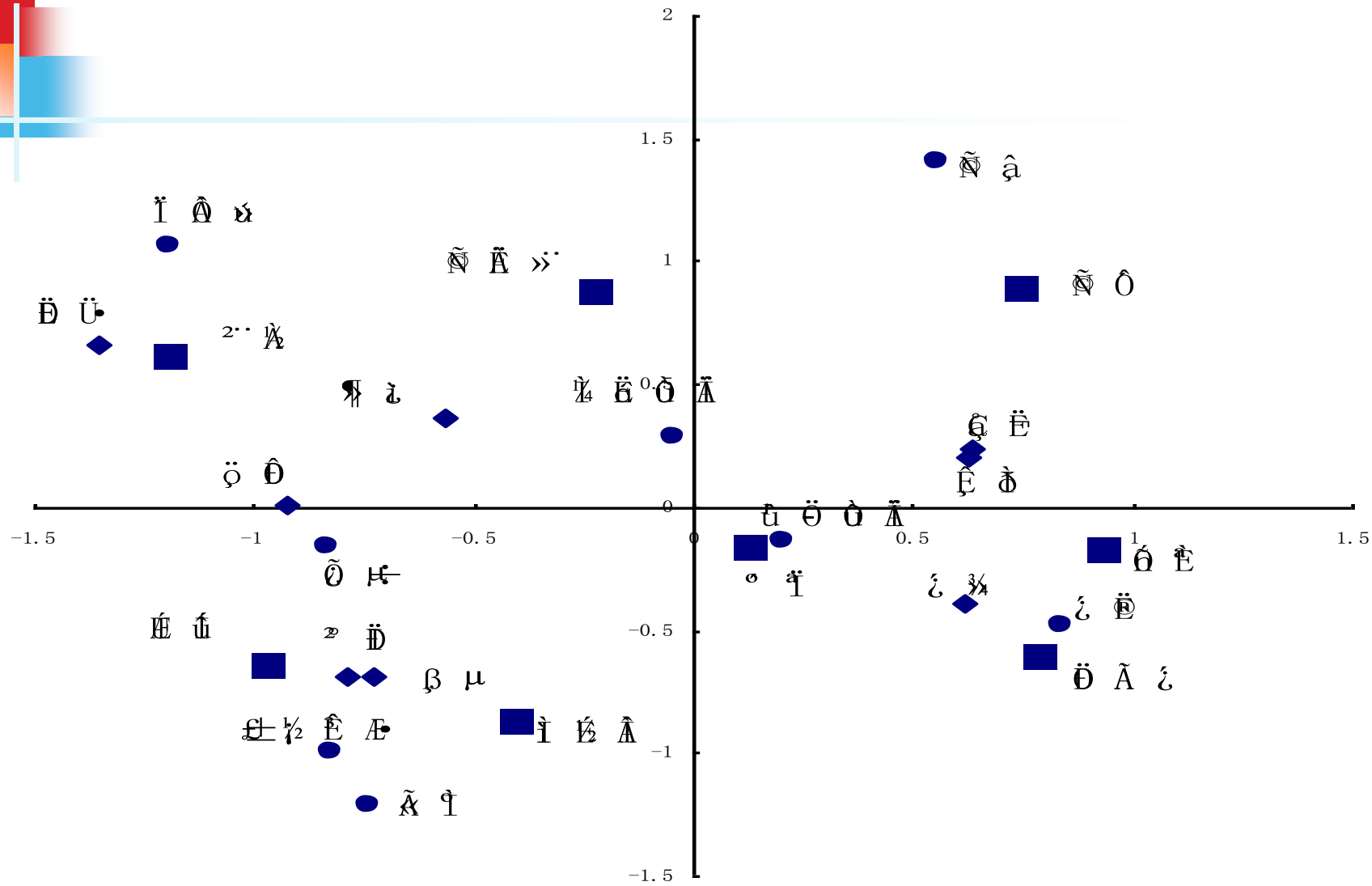
（3_2）



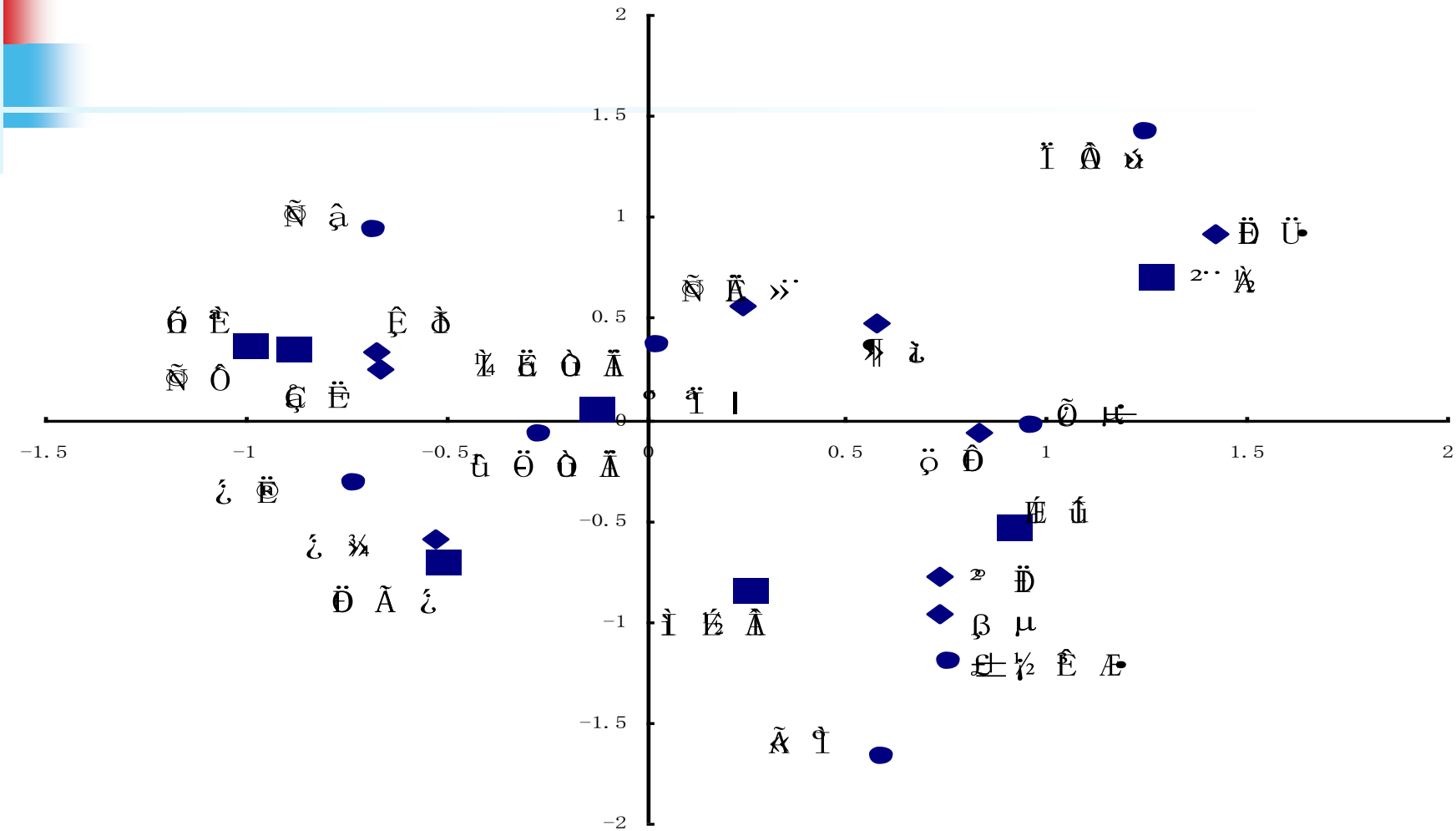
第一种情况中的行、列变量类得分（可解释61.7%）



第二种情况的行、列变量类得分（可解释77.7%）



3-1 行、列变量类得分（可解释63.5%，处置1）



3-2 的行、列变量类得分（可解释77.7%，处置2）

维度（Dimension）	奇异值（Singular Value）	惯量（Inertia）	惯量比例（Proportion of Inertia）	
			比例	累计比例
1	.572	.328	.594	.594
2	.318	.101	.183	.777
3	.267	.071	.129	.906
4	.210	.044	.080	.986
5	.077	.006	.011	.997
6	.038	.001	.003	.999
7	.018	.000	.001	.1.000
Total		.552	1.000	1.000
卡方（Chi Square） = 3312.225		概率值（Significance p）=.000		