# 信息浓缩技术——主成分分析、因子分析

@文彤老师

# 概 述

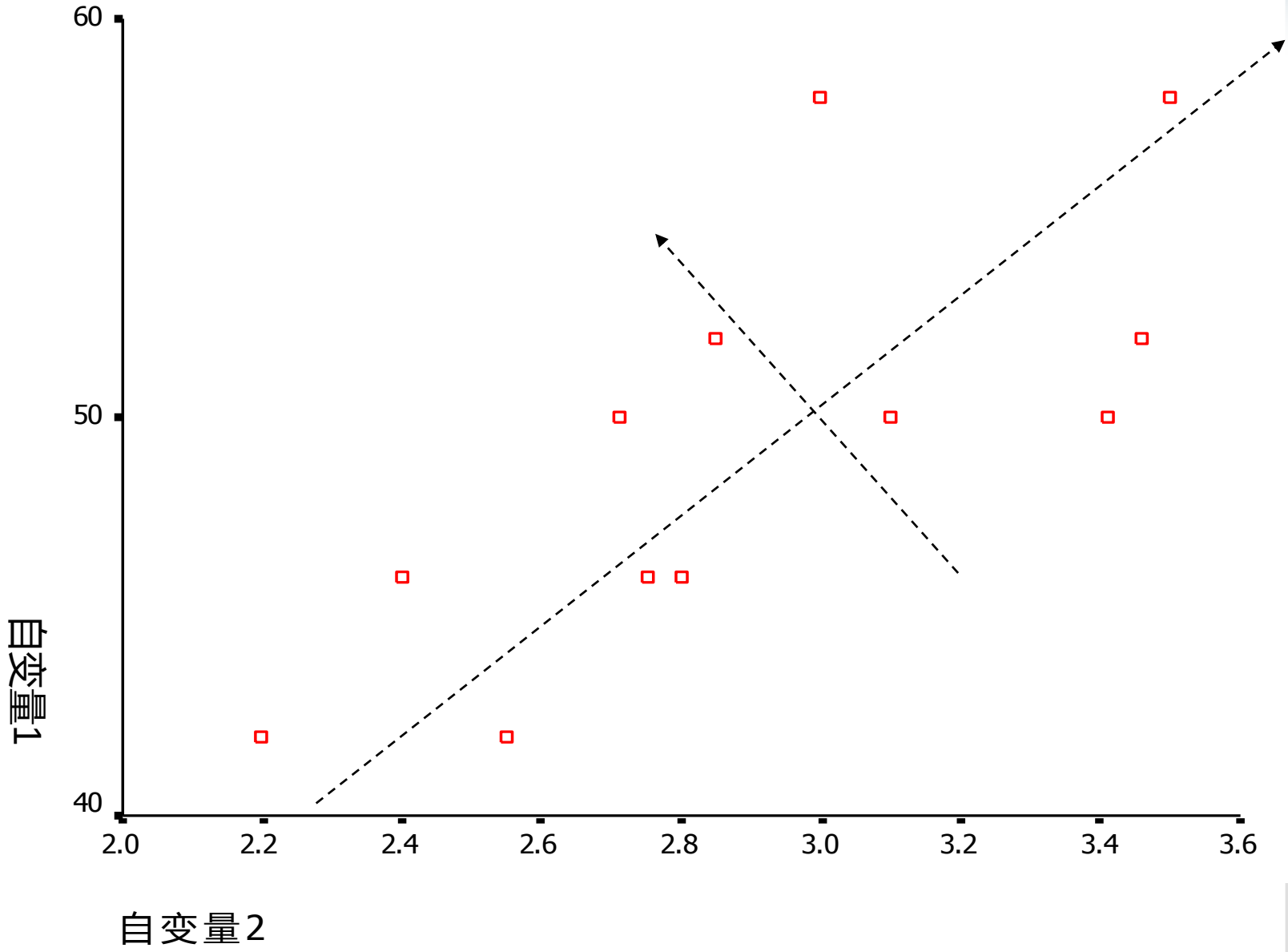
- 解决变量间多重共线性(data reduction)
  - 有太多的变量，希望能够消减变量，用一个新的、更小的由原始变量集组合成的新变量集作进一步分析。
  - 新的变量集能够更好的说明问题，利于简化和解释问题。
- 探讨变量内在联系和结构(structure detection)
  - 观测变量之间的存在相互依赖关系
  - 由量表所反映出的受访者对卫生服务的真实满意度

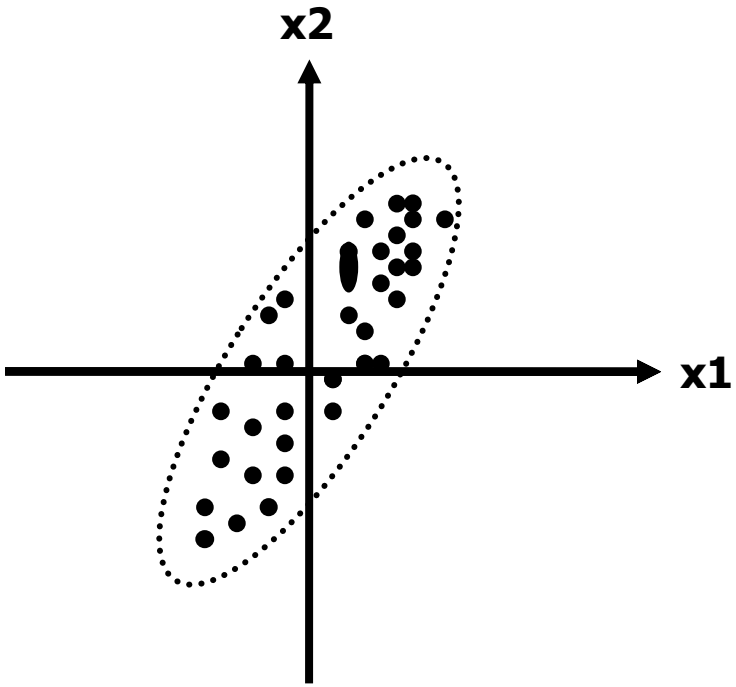
# 概 述

- 本讲课程中涉及的方法其实质均为数据化简、信息浓缩，即将分散在多个变量中的同类信息集中、提纯，从而便于分析、解释和利用。
  - 目的为浓缩信息
  - 目的为探讨内在结构
- 正因如此，这些信息浓缩方法、特别是其中的因子分析方法，往往成为更复杂的多元分析方法的基石。

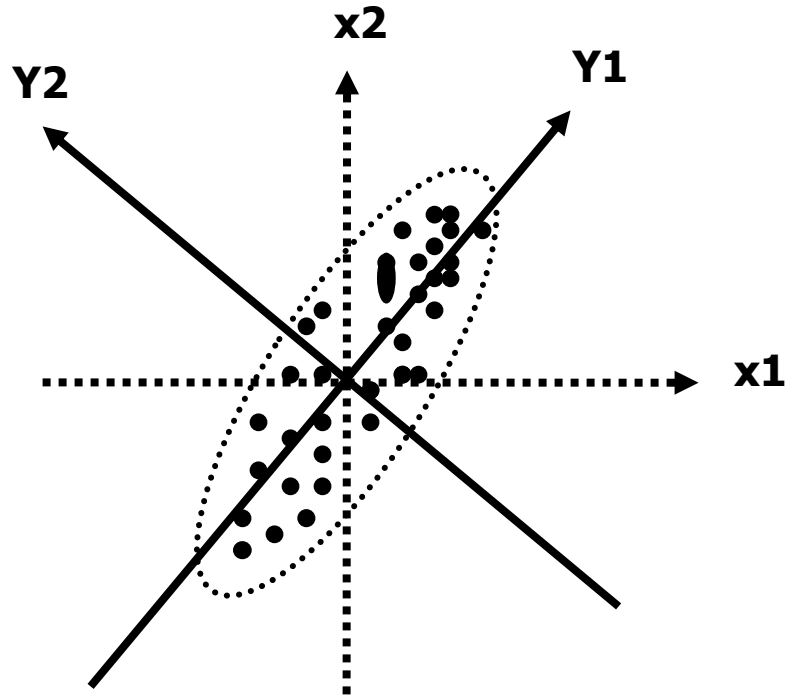
# 主成分分析

- 只是一种中间手段，其背景是研究中经常会遇到多指标的问题，这些指标间往往存在一定的相关，直接纳入分析不仅复杂，变量间难以取舍，而且可能因多元共线性而无法得出正确结论。
- 主成分分析的目的就是通过线性变换，将原来的多个指标组合成相互独立的少数几个能充分反映总体信息的指标，便于进一步分析。
  - 尽可能保留原始变量的信息，且彼此不相关。


60
50
40
2.0
2.2
2.4
2.6
2.8
3.0
3.2
3.4
3.6
自变量1
自变量2




X1与x2相关




Y1与Y2不相关

# 主成分分析

- 在主成分分析中，提取出的每个主成分都是原来多个指标的线性组合
- 如有两个原始变量x1和x2，则一共可提取出两个主成分如下：
  - $z1=b_{11}x1+b_{21}x2$
  - $z2=b_{12}x1+b_{22}x2$

# 主成分分析

- 原则上如果有n个变量，则最多可以提取出n个主成分，但如果将它们全部提取出来就失去了该方法简化数据的实际意义。多数情况下提取出前2~3个主成分已包含了90%以上的信息，其他的可以忽略不计。
- 在进行主成分回归时，提取出的主成分能包含主要信息即可，不一定非要有准确的实际含义。

# 主成分分析

- 方法用途
  - 主成分评价：当进行多指标的综合评价时，应用主成分方法将多指标中的信息集中为若干个主成分，然后加权求和，得到综合评价指数。
  - 主成分回归：通过对存在共线性的自变量进行主成分分析，从而在提取多数信息的同时解决共线性问题。

# 案例：各省经济发展情况综合评价

- 现希望根据全国30个省市自治区经济发展基本情况的八项指标对其进行分析和排序。具体指标有：GDP、居民消费水平、固定资产投资、职工平均工资、货物周转量、居民消费价格指数、商品零售价格指数、工业总产值，数据文件见factor1.sav。

# 因子分析

- 是一种多变量化简技术。目的是分解原始变量，从中归纳出潜在的“类别”，相关性较强的指标归为一类，不同类间变量的相关性较低。每一类变量代表了一个“共同因子”，即一种内在结构，因子分析就是要寻找该结构。
  - 探索性因子分析
  - 证实性因子分析

# 因子分析

- 比如在市场调查中收集了食品的五项指标：味道、价格、风味、是否快餐食品、能量。经过因子分析后发现结果如下：
  - x1=0.02z1+0.99z2+ε1
  - x2=0.94z1-0.01z2+ε2
  - x3=0.13z1+0.98z2+ε3
  - x4=0.84z1+0.42z2+ε4
  - x5=0.97z1-0.02z2+ε5

# 因子分析

- 第一公因子主要影响价格、是否快餐食品和能量，代表“价廉”
- 第二公因子则主要影响味道和风味，代表“味美”
- ε代表特殊因子，只对当前变量有影响，表示该变量中独特的，不能被公因子所解释的特征

# 用途

- 研究设计阶段/问卷效果评估阶段
  - 评价问卷的结构效度
- 统计分析阶段
  - 解决变量间多重共线性
  - 寻找变量间潜在结构
  - 内在结构证实

# 适用条件

- 样本量
  - 样本量与变量数的比例应在5：1以上
  - 总样本量不得少于100，而且原则上越大越好
- 各变量间必须有相关性
  - KMO统计量：0.9最佳，0.7尚可，0.6很差，0.5以下放弃
  - Bartlett's球形检验
- 这些条件均是为了保证能够寻找出内在结构

# 分析步骤

- 判断是否需要进行因子分析，数据是否符合要求
- 进行分析，按一定标准确定提取的因子数目
- 如果进行的是主成分分析，则将主成分存为新变量用于继续分析，步骤到此结束
- 如果进行的是因子分析，则考察因子的可解释性，并在必要时进行因子旋转，以寻求最佳解释方式
- 如有必要，可计算出因子得分等中间指标供进一步分析使用

# 因子分析

- 公因子数量的确定
  - 主成分的累积贡献率：80~85％以上
  - 特征根：大于1
  - 综合判断
  - 因子分析时更重要的是因子的可解释性
    - 必要时可保留小于1的因子
    - 碎石图可以帮助确定因子数量

# 有关概念

- 因子负荷
  - 即表达式中各因子的系数值，用于反映因子和各个变量间的密切程度，其实质是两者间的相关系数
- 公因子方差比（Communalities）
  - 指的是提取公因子后，各变量中信息分别被提取出的比例，或者说原变量的信息量（方差）中由公因子决定的比例

# 有关概念

- 特征根（Eigenvalue）
  - 可以被看成是主成分影响力度的指标，代表引入该因子／主成分后可以解释平均多少原始变量的信息。
  - 正因如此，一般对特征根大于1的因子才加以注意

# 分析实例

- 案例：对各省经济数据的进一步分析
- 奥林匹克资料：olymp88.sav
  - 因子旋转
  - 因子负荷的排序和禁止输出
  - 因子计算公式的生成



# 对应分析

# 对应分析

- 问题的提出
  - 分析分类变量间关系时，卡方检验只能给出总体有无关联的结论，但不能进行精细分析，在变量类别极多时于事无补
  - Logistic模型在多分类时过于笨拙
- 解决方案
  - 精细建模：对数线性模型
  - 直观展示：对应分析

# 对应分析

- 特点
  - 是多维图示分析技术之一，结果直观、简单
  - 与因子分析有关
    - 分类资料的典型相关分析
  - 用于展示两个/多个分类变量各类间的关系
    - 高收入、黑人、男性倾向于反对开战
  - 研究较多分类变量间关系时较佳
  - 各个变量的类别较多时较佳
    - 均为四类以上

# 对应分析

- 实质
  - 就是对列联表中的数据信息进行浓缩，然后以易于阅读的图形方式呈现出来
  - 以默认的卡方测量方式为例，首先以列联表为分析基础，计算基于H0假设的标化单元格残差

$$\text{标准化残差}=\frac{\text{观察频数}-\text{理论频数}}{\sqrt{\text{理论频数}}}$$

# 对应分析

- 实质
  - 将每行看成是一条记录，基于列变量相关系数阵进行因子分析，计算出列变量各类的负荷值
  - 将每列看成是一条记录，基于行变量相关系数阵进行因子分析，计算出行变量各类的负荷值
- 局限性
  - 不能进行变量间相关关系的检验
    - 仍然只是一种统计描述方法
  - 解决方案的所需维度需要研究者决定
  - 对极端值敏感
    - 对于小样本不推荐使用

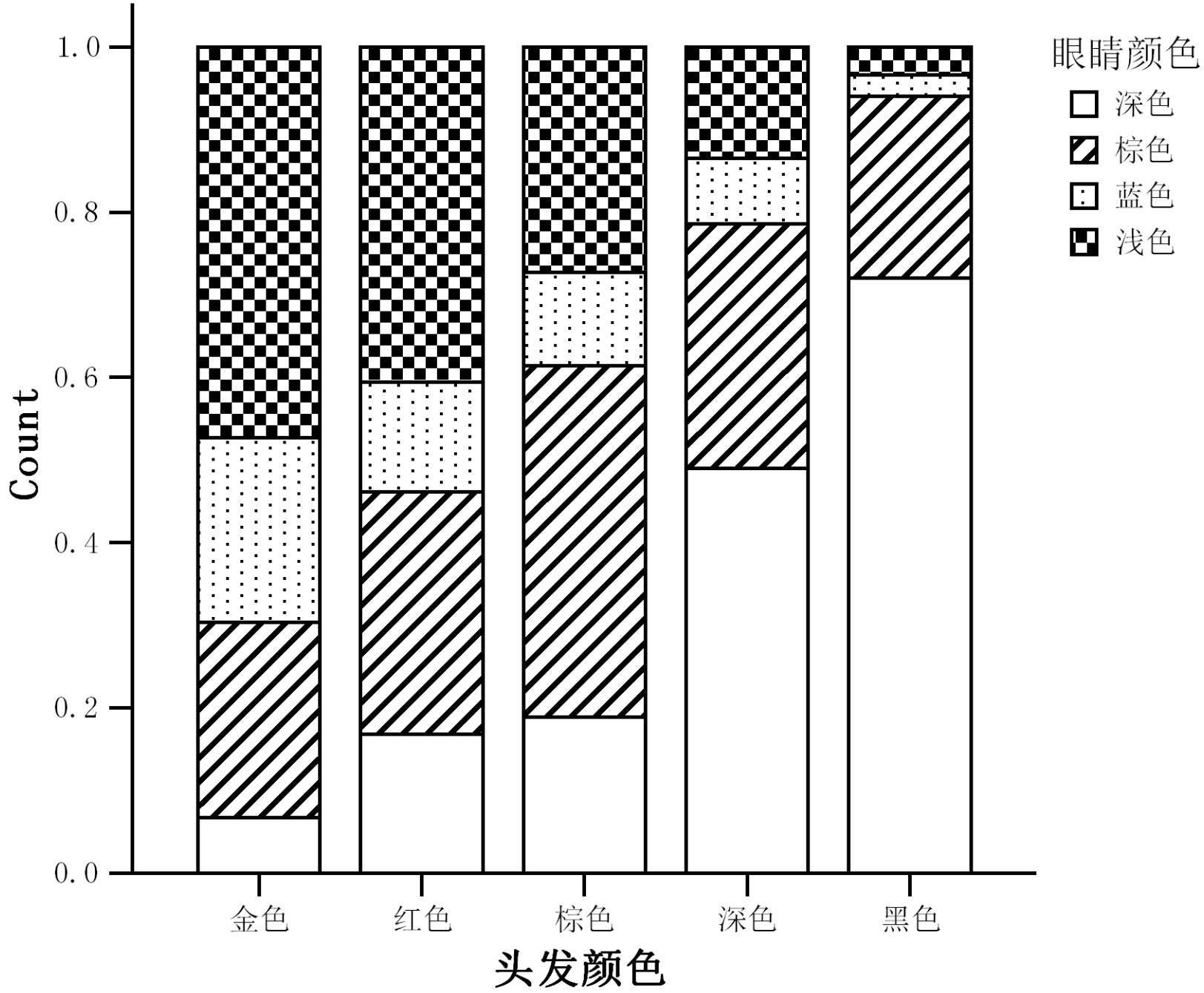
# 对应分析

- 实际应用（保险业）
  - 我们某个险种的用户是谁？
  - 另外还有什么特征的人群也倾向于成为我们的用户？
  - 什么特征的人群还没有找到满意的险种，或者说现有险种还有哪种人群未能覆盖？
  - 我们提供的险种系列是否全面，有无哪些险种过于重叠，是否还有种类空白需要填补？

研究者收集了苏格兰北部Caithness郡5387名小学生眼睛与头发颜色的数据，见下表，其中眼睛有深、棕、蓝、浅四种颜色，头发有金、红、棕、深、黑五种颜色。研究者希望知道头发和眼睛的颜色间存在何种关联，即某种头发颜色的人起眼睛更倾向于何种颜色？

**表 15.1　头发颜色与眼睛颜色的交叉表**

| | | 头发颜色 | | | | | 合 计 |
|---|---|---|---|---|---|---|---|
| | | 金色 | 红色 | 棕色 | 深色 | 黑色 | |
| 眼 | 深色 | 98 | 48 | 403 | 681 | 85 | 1315 |
| 睛 | 棕色 | 343 | 84 | 909 | 412 | 26 | 1774 |
| 颜 | 蓝色 | 326 | 38 | 241 | 110 | 3 | 718 |
| 色 | 浅色 | 688 | 116 | 584 | 188 | 4 | 1580 |
| 合计 | | 1455 | 286 | 2137 | 1391 | 118 | 5387 |

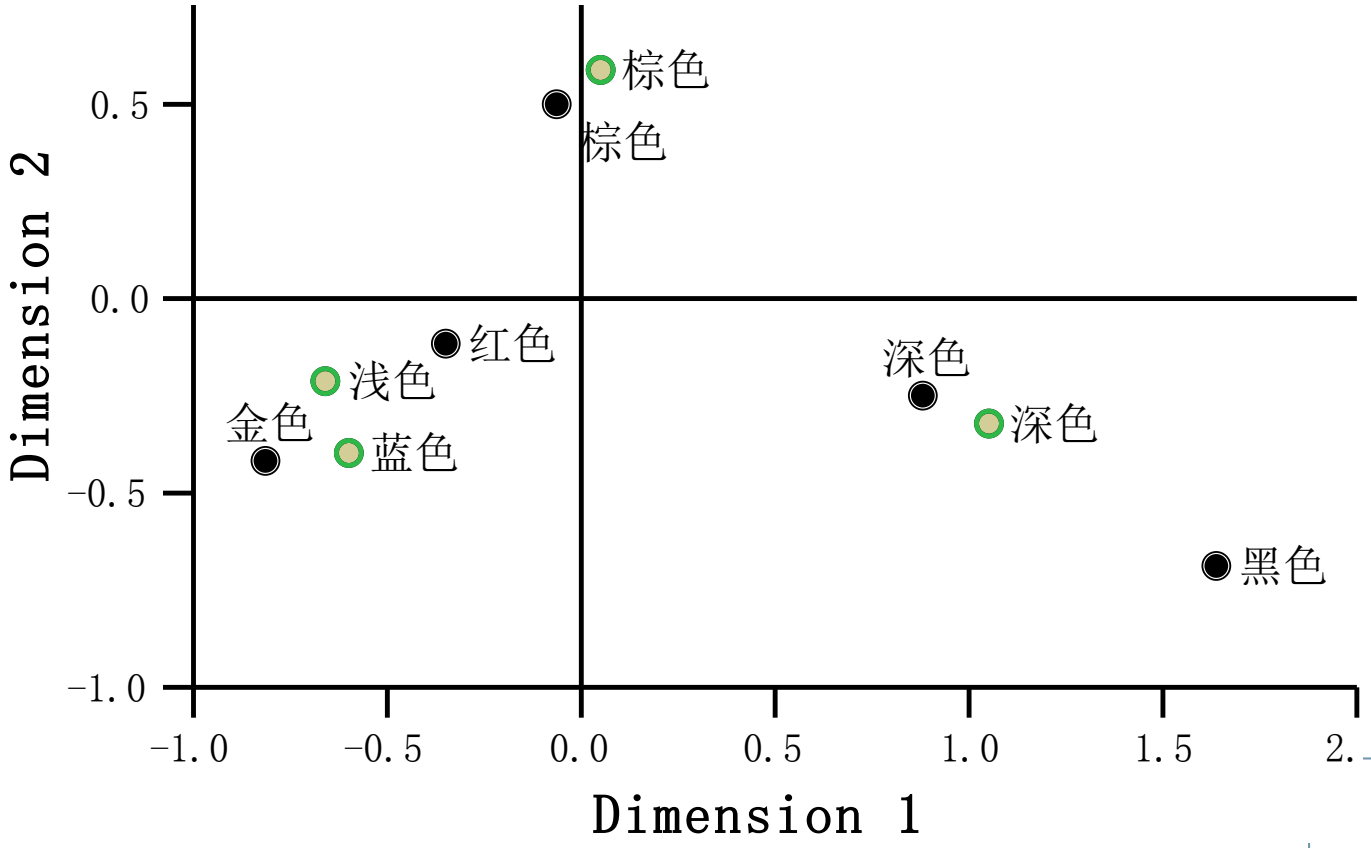

1.0
0.8
0.6
0.4
0.2
0.0
Count
金色
红色
棕色
深色
黑色
头发颜色
眼睛颜色
深色
棕色
蓝色
浅色

# 对应分析图的阅读

- 每个维度可能代表了一种特征
  - 实际上就是一个提取出的主成分，但由于分类变量的信息较少，可能找不到合理的解释
- 1. 考察同一变量的区分度：
  - 如果同一变量不同类别在某个方向上靠得较近，则说明这些类别在该维度上区别不大。
- 2. 考察不同变量的类别联系：
  - 一般而言，落在从图形原点(0,0)处出发相同方位上大致相同区域内的不同变量的分类点彼此有联系。散点间距离越近，说明关联倾向越明显；散点离原点越远，也说明关联倾向越明显。

# 结果的正确解释

- 错误的解释：金色头发的儿童中蓝色、浅色眼睛者居多
- 正确的解释：相对于平均水平而言，金色头发的儿童中蓝色、浅色眼睛的比例要高一些，也就是高于其他颜色头发的儿童



# 对应分析中应注意的问题

- 分析目的：重在观察行、列变量间的联系
- 数据类型：无序分类较佳，如果均为有序分类，且变量较多时，采用多维偏好分析更合适
- 样本量：对极端值敏感，分析时有必要去除频数过少的单元格
  - 对于小样本不推荐使用
- 变量间关联：不能将对应分析作为筛选相关变量的方法，变量纳入前最好先做卡方检验

# 最优尺度变换

- 许多时候我们所分析的变量并非连续性变量，如评分、等级等如果要按照连续变量来分析，则存在一个适当量化的问题
- 有时候虽然变量均为连续性，但变量间的关系并非简单的线性，而现有的多元分析方法几乎无一例外的是以线性关联为基础进行分析的
- 多个变量间的复杂联系如何能够简单明了的表现出来？

# 最优尺度变换

- 所谓最优尺度变换的本质，就是根据数据本身的关联，寻找出最佳的原始变量评分方法，将原始变量一律转化为相应的分值，并在转化时将变量间的关联一律变换为线性，这样就解决了以上问题
- 可以同时分析多个分类变量间的关系，并同样用图形方式表示出来
- 在变量种类上更加丰富，已可以处理各种类型的变量，如对无序多分类分析、有序多分类变量和连续性变量同时进行分析的问题

# 最优尺度变换

- 不能自动筛选变量，需要用户根据经验和分析结果进行耐心筛选
- 对样本量要求较大，特别是对少数极端值和罕见类别频数的变化非常敏感
- 由于结果往往以图形方式呈现，不加注意可能会得到完全错误地分析结果
- 所作的最优尺度变换是基于数据本身而来，当增减变量、或者对变量进行变换后重新拟合时，相应的结果可能完全不同

# 最优尺度变换

- Homogeneity (HOMALS)
  - 同质性分析，即多重对应分析
  - 以图形化方式展示多个分类变量间的关系
- Categorical Principal Components (CatPCA)
  - 其实质为分类变量的主成分分析
  - 当一些变量为名义测量外的其它测量（有序分类或连续性变量）时使用
  - 也就是MR中常用的多维偏好分析
- Nonlinear Canonical Correlation (OVERALS)
  - 非线性典型相关方法
  - 用于分析两个或多个变量集之间的关系
  - 允许变量为任何类型

# 案例

- 数据集mcorres.sav提供了某次调查得来的轿车特征与一些用户特征的数据，请分析汽车原产地（norigin）、汽车大小（nsize）、轿车类型（ntype）、是否租房（nhome）、有无双份收入（nincome）、性别（nsex）、婚姻状况（nmarit）之间的联系如何

# 注意事项

- 由于算法不同，当分析两个变量时，结果不会等同于简单对应分析，但是基本相同
- 不推荐同时分析过多变量
- 必要时应当对频数较少的类别加以合并或者剔除
- 得到结果后应当和原始表格加以仔细对照，以确保分析结果的正确性

# 多维尺度分析

@文彤老师

# 多维尺度分析简介

- 用于反映多个研究事物间的相似（不相似）程度
- 通过询问消费者对各种品牌（或者概念）的相似程度评分，我们可以评价那些品牌在消费者的心目中比较类似，而这些类似的品牌往往意味着市场定位重叠，即存在竞争关系
- 但是，直接采用原始数据加以考察非常麻烦，数据太多，不容易得出一个综合、全面的结论

# 多维尺度分析简介

- 同时所需的解释空间维度太高，不容易阅读
- 通过适当的降维方法，将这种相似（不相似）程度在低维度空间中用点与点之间的距离表示出来，并有可能帮助识别那些影响事物间相似性的潜在因素
- 最常见情况是在二维空间中将结果表现出来

# 分析目的

- 通过空间定位图，研究者可以得知
  - 消费者认为哪些品牌的产品类似于我们的产品？
  - 在这些品牌中消费者用于评价相似性的是哪些特征指标（属性）？

# 案例：城市航空距离

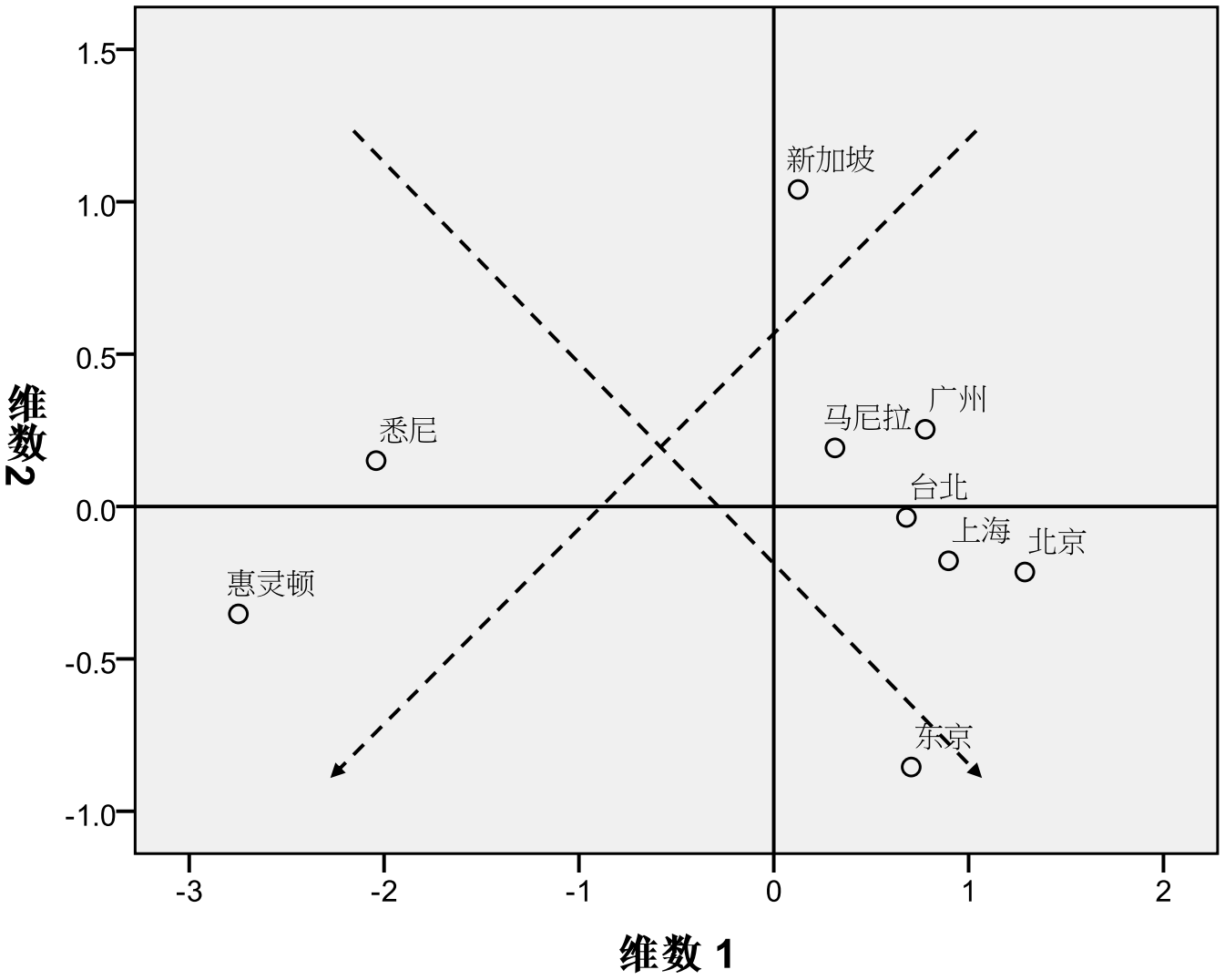
- 数据distance.sav是笔者基于地图测量的亚太地区9个城市的地面距离，请就此进行多维尺度分析

| | cityname | 北京 | 上海 | 台北 | 广州 | 东京 | 新加坡 | 马尼拉 | 悉尼 | 惠灵顿 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 北京 | 0 | . | . | . | . | . | . | . | . |
| 2 | 上海 | 1067 | 0 | . | . | . | . | . | . | . |
| 3 | 台北 | 1720 | 683 | 0 | . | . | . | . | . | . |
| 4 | 广州 | 1892 | 1212 | 869 | 0 | . | . | . | . | . |
| 5 | 东京 | 2096 | 1763 | . | 2909 | 0 | . | . | . | . |
| 6 | 新加坡 | 4507 | 3830 | 3272 | 2656 | 5339 | 0 | . | . | . |
| 7 | 马尼拉 | 2855 | 1855 | 1173 | 1251 | 3000 | 2402 | 0 | . | . |
| 8 | 悉尼 | 8952 | 7885 | 7269 | 7514 | 7836 | 6271 | 6273 | 0 | . |
| 9 | 惠灵顿 | 10792 | 9753 | 9203 | 9568 | 9285 | 8504 | 8318 | 2229 | 0 |

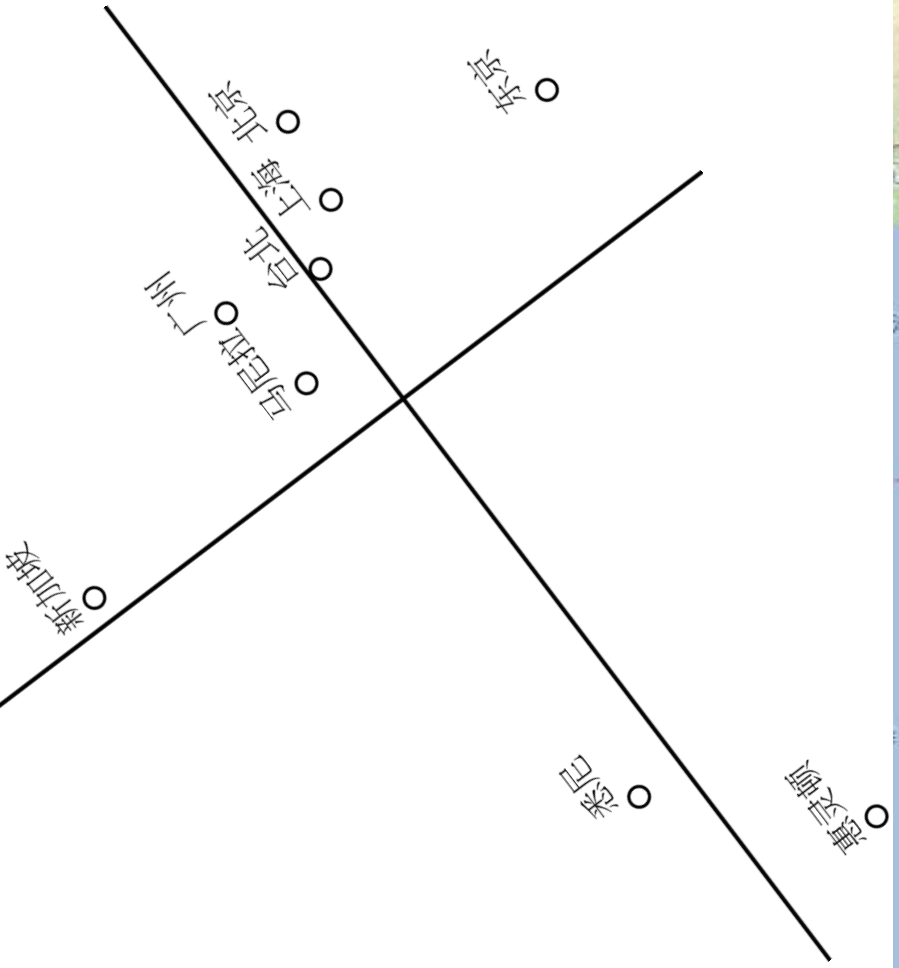
# 空间图（刺激匹配图）

- 哪些散点比较接近（相似），所有的散点大致被分为了几类
- 如果有可能，为每个维度找到一个合理的解释
- 寻找图形散点间相关性的合理解释

# 建模得到的地图结果



# 建模得到的地图结果


东京
北京
上海
台北
马尼拉
广州
新加坡
悉尼
奥克兰


日本海
日本
中国
韩国
缅甸
泰国
南海
安达曼海
泰国湾
越南
菲律宾
马来西亚
印度尼西亚
巴布亚新几内亚
所罗门海
珊瑚海
NT
QLD
澳大利亚
WA
SA
NSW
大澳洲湾
VIC
ACT
塔斯曼海
TAS
新西兰

# 案例：饮料的口味差异评价

- 数据文件为mds.sav，来自Schiffman，Reynolds和Young(1981)，里面包括了10位受访者对常见的10种饮料的口感差异性评分，分值在0～100之间，数值越高表明差异越大。每位个体的数据形成了一个距离阵，十个距离阵被纵向叠加在一起
- 请从中分析各种饮料的差异性，并进一步寻求导致这种差异性的合理解释

# 结果解释

- 散点分类：这十种饮料大致被分为了无糖饮料和普通饮料两组
- 维度解释：
  - 第一维度代表的是饮料的健康程度
  - 第二维度代表的是饮料的风味或者口感
- 散点定位的合理解释与建议
  - 在受访者心目中，pepper的口感非常独特，和其余饮料有着明显的区别
  - 可口可乐与百事可乐的差异更多的是营销活动所塑造出来的品牌形象，而在产品口感上这两者实际是没有本质区别的

# PROXSCAL过程

- ALSCAL提供的是比较经典的5个分析模型，而PROXSCAL是使用了Data Theory Scaling System Group (DTSS)的最优化数据转换的方法，提供了4个更高级的模型。两个过程的分析结果不全相等，但是有对应关系。
- 原理和最优尺度模型完全相同
- ALSCAL只能分析不相似性数据，而PROXSCAL对相似性数据或者不相似性数据都能进行分析。
- 提供了更加丰富的模型诊断、设置和结果输出。

# 多维展开模型

- 在MDS模型中，用于计算距离的各对象不存在分组，分析时直接考虑的是各对象两两间的距离远近
- 在有的研究问题中，研究对象可能会被分为两组，分析的核心目的是考察两组对象之间的距离远近，而同一组内各对象间的距离则并不在研究关心的范围之内，这时多维展开模型就是比较合适的选择

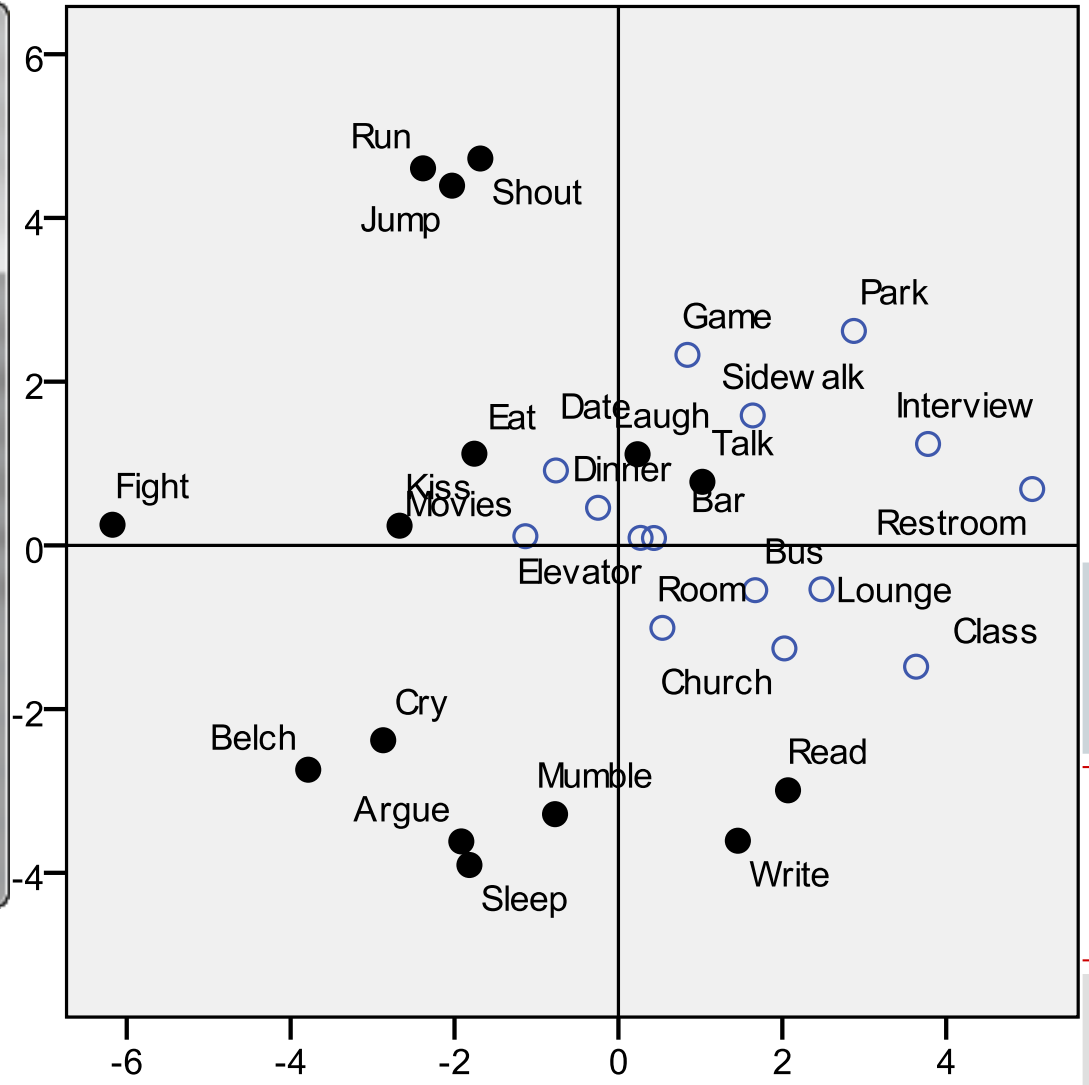
# 多维展开模型

- 数据排列格式
  - 仍然属于距离阵格式，但是此时并非对称距离阵，行变量和列变量并不相同，分别构成了行变量组合列变量组

| ROWID | Run | Talk | Kiss | Write | Eat | Sleep | Mumble | Read | Fight |
|---|---|---|---|---|---|---|---|---|---|
| Class | 6.48 | 2.79 | 6.90 | .83 | 4.77 | 5.40 | 5.38 | 1.73 | 7.79 |
| Date | 4.00 | .44 | .27 | 5.38 | 1.21 | 5.23 | 5.88 | 6.12 | 5.42 |
| Bus | 7.56 | .92 | 4.73 | 4.13 | 3.52 | 1.96 | 3.83 | 1.83 | 7.48 |
| Dinner | 6.44 | .48 | 4.08 | 6.42 | .56 | 6.71 | 6.46 | 5.04 | 7.33 |
| Park | 1.06 | .58 | 1.29 | 2.00 | .87 | 3.37 | 3.60 | 1.23 | 5.94 |
| Church | 7.62 | 5.71 | 6.62 | 6.15 | 7.62 | 7.23 | 5.48 | 5.42 | 8.38 |
| Interview | 7.06 | .54 | 7.92 | 4.15 | 7.27 | 8.25 | 7.69 | 6.52 | 7.96 |
| Sidewalk | 3.42 | .81 | 4.25 | 5.62 | 4.17 | 7.54 | 4.04 | 4.19 | 7.54 |
| Movies | 6.54 | 4.02 | 2.79 | 6.27 | 1.52 | 4.92 | 4.87 | 7.27 | 7.63 |
| Bar | 7.04 | .75 | 3.83 | 3.62 | 1.33 | 6.10 | 2.79 | 4.29 | 7.10 |
| Elevator | 7.37 | 1.60 | 4.21 | 5.96 | 3.90 | 7.69 | 3.88 | 4.52 | 7.42 |
| Restroom | 6.17 | 1.75 | 6.19 | 5.54 | 6.65 | 6.17 | 3.96 | 4.25 | 7.23 |

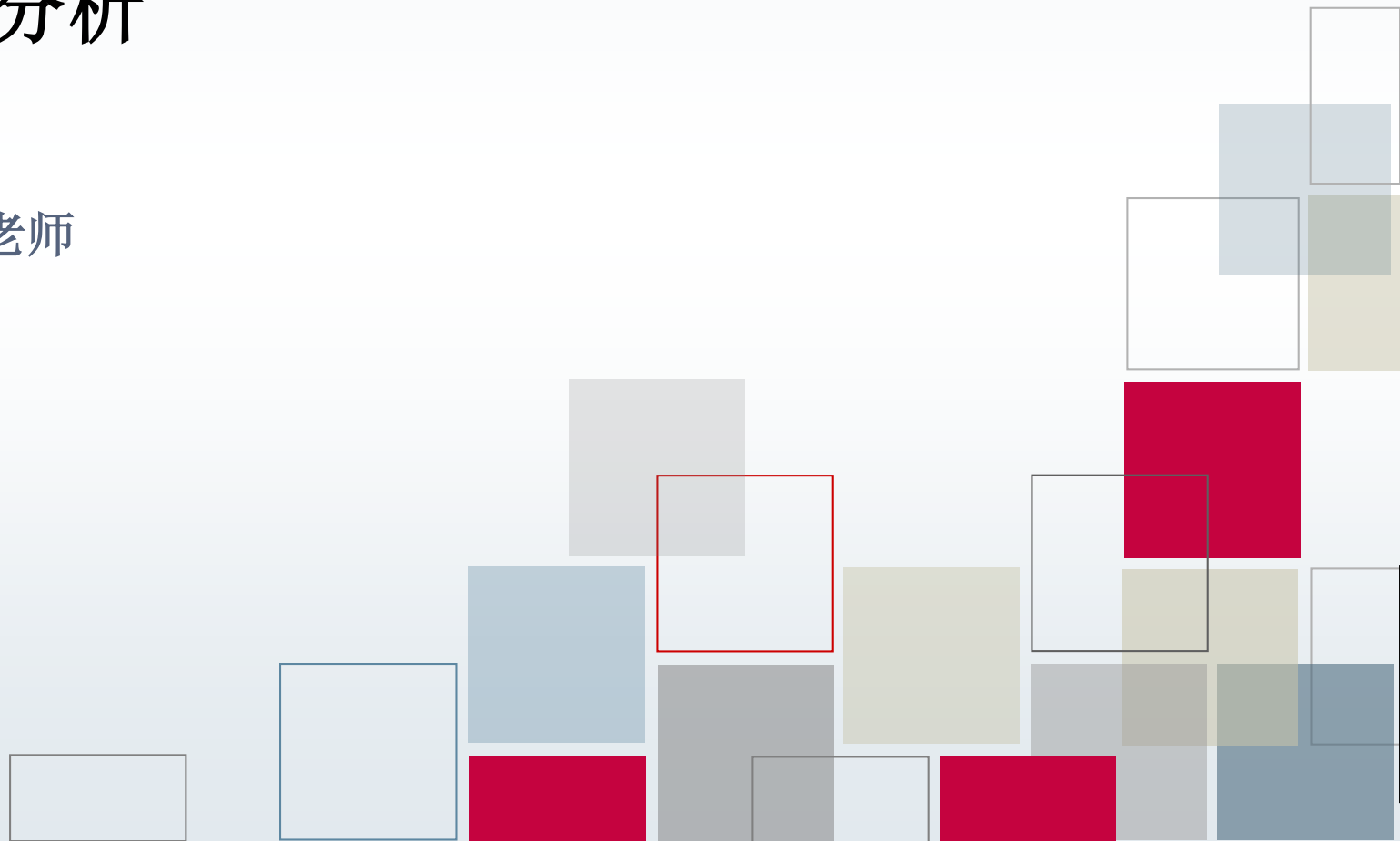
# 多维展开模型


近似值(X):
Run
Talk
Kiss
Write
权重(G):
行(W):
ROWID
源(S):
模型(M)...
限制(E)...
选项(O)...
绘制(T)...
输出(U)...
粘贴(P)
重置(R)
取消
帮助


Run
Shout
Jump
Park
Game
Sidew alk
Interview
Eat
Date
Laugh
Talk
Dinner
Fight
Kiss
Movies
Bar
Restroom
Bus
Elevator
Room
Lounge
Class
Church
Cry
Belch
Read
Mumble
Argue
Write
Sleep
6
4
2
0
-2
-4
-6
-4
-2
0
2
4

# 聚类分析

@文彤老师



# 方法原理

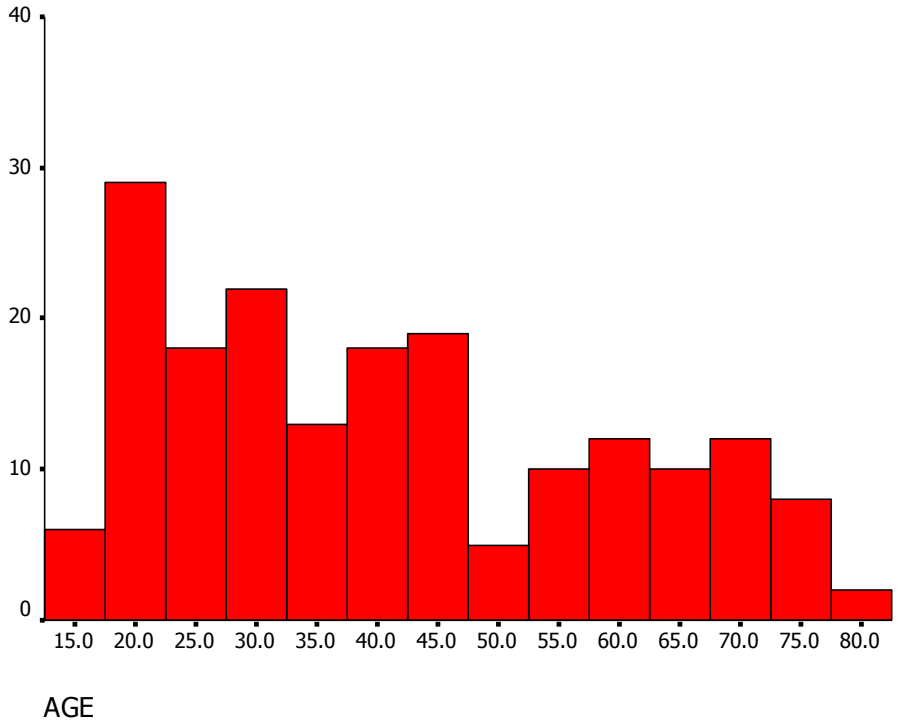
- 按照个体（记录）的特征将它们分类，使同一类别内的个体具有尽可能高的同质性，而类别之间则具有尽可能高的异质性。
- 为了得到比较合理的分类，首先要采用适当的指标来定量地描述研究对象之间的联系的紧密程度。
  - 直观的理解为按空间距离的远近来划分类别

# 方法原理

- 假定研究对象均用所谓的“点”来表示。
- 在聚类分析中，一般的规则是将“距离”较小的点归为同一类，将“距离”较大的点归为不同的类。
- 常见的是对个体分类，也可以对变量分类
  - 此时一般使用相似系数作为“距离”测量指标

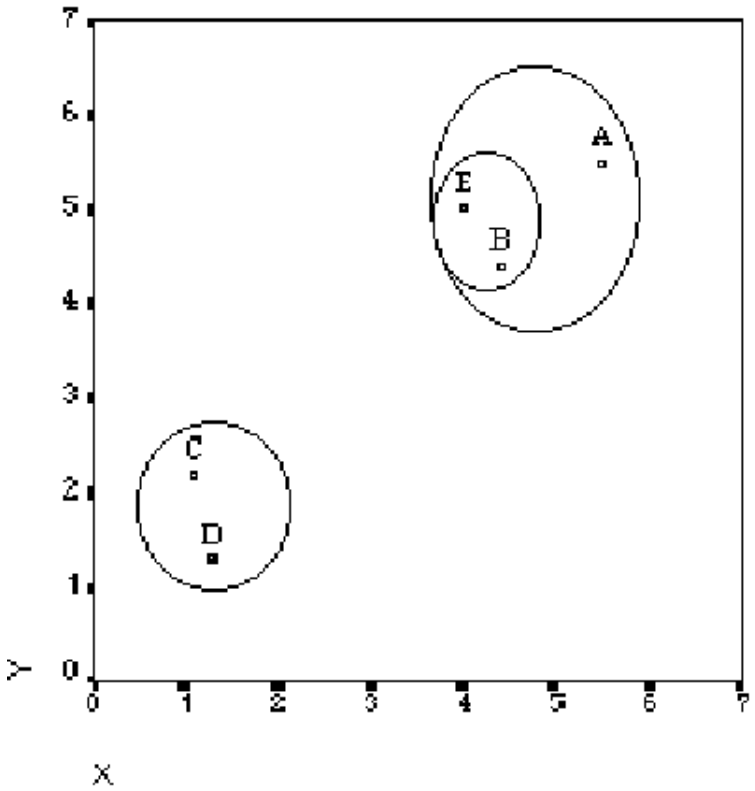
# 方法原理

- 例：根据年龄将人群分成适当的类，从图中可见，人群被分为两类是比较合适的。



# 方法原理

- 在右图中可以看到五个样品应当可能被分为两组或者三组，C/D组x和y的取值均偏低，而另三个所在组x和y的取值均偏高
- 分为两类或三类都是可接受的
- 从图中可以直观的理解“距离”的含义



# 方法原理

- 当用于聚类的变量逐渐增多时，分析思路完全相同，只是这样简单、清晰的图示展现类别情况变得逐渐不大可能
  - 多维空间中的观察
- 可能的解决方法
  - 放弃图示化观察，改用复杂的统计指标
  - 缩减维度，使得可以在低维度空间进行呈现

# 特 点

- 聚类分析前所有个体所属的类别是未知的，类别个数一般也是未知的，分析的依据就是原始数据，可能事先没有任何有关类别的信息可参考。
- 严格说来聚类分析并不是纯粹的统计技术，它不像其它多元分析法那样，需要从样本去推断总体。
  - 一般都涉及不到有关统计量的分布，也不需要进行显著性检验。
- 聚类分析更像是一种建立假设的方法，而对相关假设的检验还需要借助其它统计方法。

# 用 途

- 设计抽样方案
  - 分层抽样
- 预分析过程
  - 先通过聚类分析达到简化数据的目的，将众多的个体先聚集成比较好处理的几个类别或子集，然后再进行后续的多元分析。
- 细分市场、个体消费行为划分
  - 先聚类，然后再利用判别分析进一步研究各个群体之间的差异。

# 聚类分析的基本步骤总结

**步骤1：确定研究问题**

选择研究目的：

分类描述；数据简化；揭示相互联系选择分类变量

**步骤2：研究设计**

- 对个体聚类还是对变量聚类？
- 分类变量是什么类型？
- 选择用“距离”还是“相似系数”？是哪一种？
- 数据是否需要标准化？
- 是否需要删除奇异点？

**步骤3：考虑是否满足基本的假定**

- 样本对总体是否有代表性？
- 聚类变量是否存在共线性？
- 共线性是否足以影响结果？

# 聚类分析的基本步骤总结

**步骤4：选择聚类方法**

- 采用谱系聚类法、非谱系聚类法、还是两者的结合？
- 采用哪种具体的聚类方法？
- 应该聚成几个类别？

**步骤5：解释聚类分析的结果**

- 考察类别的中心是否存在显著的差异？
- 观察树形图和冰柱图
- 是否可以根据分类变量给各个类别命名？

**步骤6：评价聚类分析结果的有效性**

- 利用适当的结果变量进行评价；
- 利用其它描述性的变量描述各个类别的轮廓。

# K-means Cluster过程

- 属于非系统聚类法的一种
- 方法原理
  - 选择（或人为指定）某些记录作为凝聚点
  - 按就近原则将其余记录向凝聚点凝集
  - 计算出各个初始分类的中心位置（均值）
  - 用计算出的中心位置重新进行聚类
  - 如此反复循环，直到凝聚点位置收敛为止

# K-means Cluster过程

- 方法特点
  - 要求已知类别数
  - 可人为指定初始位置
  - 节省运算时间
  - 样本量过大时有必要考虑
  - 只能使用连续性变量



# 案例：移动通讯客户细分

- telco.sav数据包含6个变量
  - 是客户编号（Customer_ID）
  - 工作日上班时期电话时长（Peak_mins）
  - 工作日下班时期电话时长（OffPeak_mins）
  - 周末电话时长（Weekend_mins）
  - 国际电话时长（International_mins）
  - 总通话时长（Total_mins）
  - 平均每次通话时长（average_mins）
- 根据前期的调研，研究者认为移动用户应当被分为5个主要群体，现希望得到相应的定量聚类结果。

# 聚类结果

- 第一类：高端商用客户，总通话时间长，工作日上班时间通话比例高
- 第二类：少使用低端客户，总通话时间短，各时段通话时间都短
- 第三类：中端商用客户，总通话时间居中，工作日上班时间通话比例高
- 第四类：中端日常用客户，总通话时间居中，工作日下班时间通话比例高
- 第五类：长聊客户，每次通话时间长

# Hierarchical Cluster过程

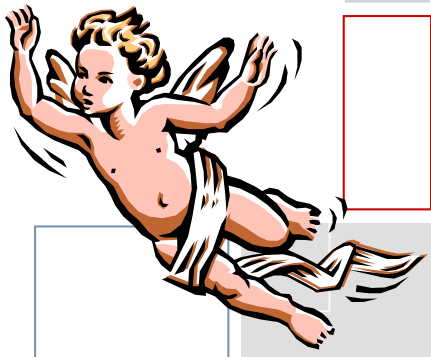
- 属于系统聚类法的一种
  - 其聚类过程可以用树形结构(treelike structure)来描绘的方法
- 方法原理
  - 先将所有n个变量／观测看成不同的n类
  - 然后将性质最接近（距离最近）的两类合并为一类
  - 再从这n-1类中找到最接近的两类加以合并
  - 依此类推，直到所有的变量／观测被合为一类
  - 使用者再根据具体的问题和聚类结果来决定应当分为几类

# Hierarchical Cluster过程

- 特点
  - 一旦记录/变量被划定类别，其分类结果就不会再进行更改
  - 可以对变量或记录进行聚类
  - 变量可以为连续或分类变量
  - 提供的距离测量方法非常丰富
  - 运算速度较慢

# 案例：体操裁判打分倾向聚类

- SPSS自带数据集judges.sav是中、美、法等七个国家的裁判和未经严格训练体育爱好者在评判体育比赛中对选手的评分情况。请根据在评分上的差异将它们分为适当的类。
  - 树状图的使用
  - 不同聚类结果的比较



# TwoStep Cluster过程

- 特点：
  - 处理对象：分类变量和连续变量
  - 自动决定最佳分类数
  - 快速处理大数据集
- 前提假设：
  - 变量间彼此独立
  - 分类变量服从多项分布，连续变量服从正态分布
  - 模型稳健

# 需要注意的问题

- 距离测量方法
  - 使用默认值即可
- 变量选择
  - 无关变量有时会引起严重的错分
  - 应当只引入在不同类间有显著差别的变量
  - 尽量只使用相同类型的变量进行分析
    - 使用连续变量，将分类变量用于结果解释
    - 新的聚类方法可以同时使用这些变量

# 需要注意的问题

- 共线性问题
  - 对记录聚类结果有较大的影响
    - 相当于某个变量在聚类中的权重大于其它变量
  - 最好先进行预处理
- 变量的标准化
  - 变量量纲／变异程度相差非常大时需要进行
  - 数理统计算法上要求一律标准化
  - 标准化后会削弱有用变量的作用

# 需要注意的问题

- 异常值
  - 影响较大
  - 还没有比较好的解决办法
  - 尽力避免
- 分类数
  - 从实用角度讲，2～8类比较合适
- 专业意义
  - 一定要结合专业知识进行分析

# 需要注意的问题

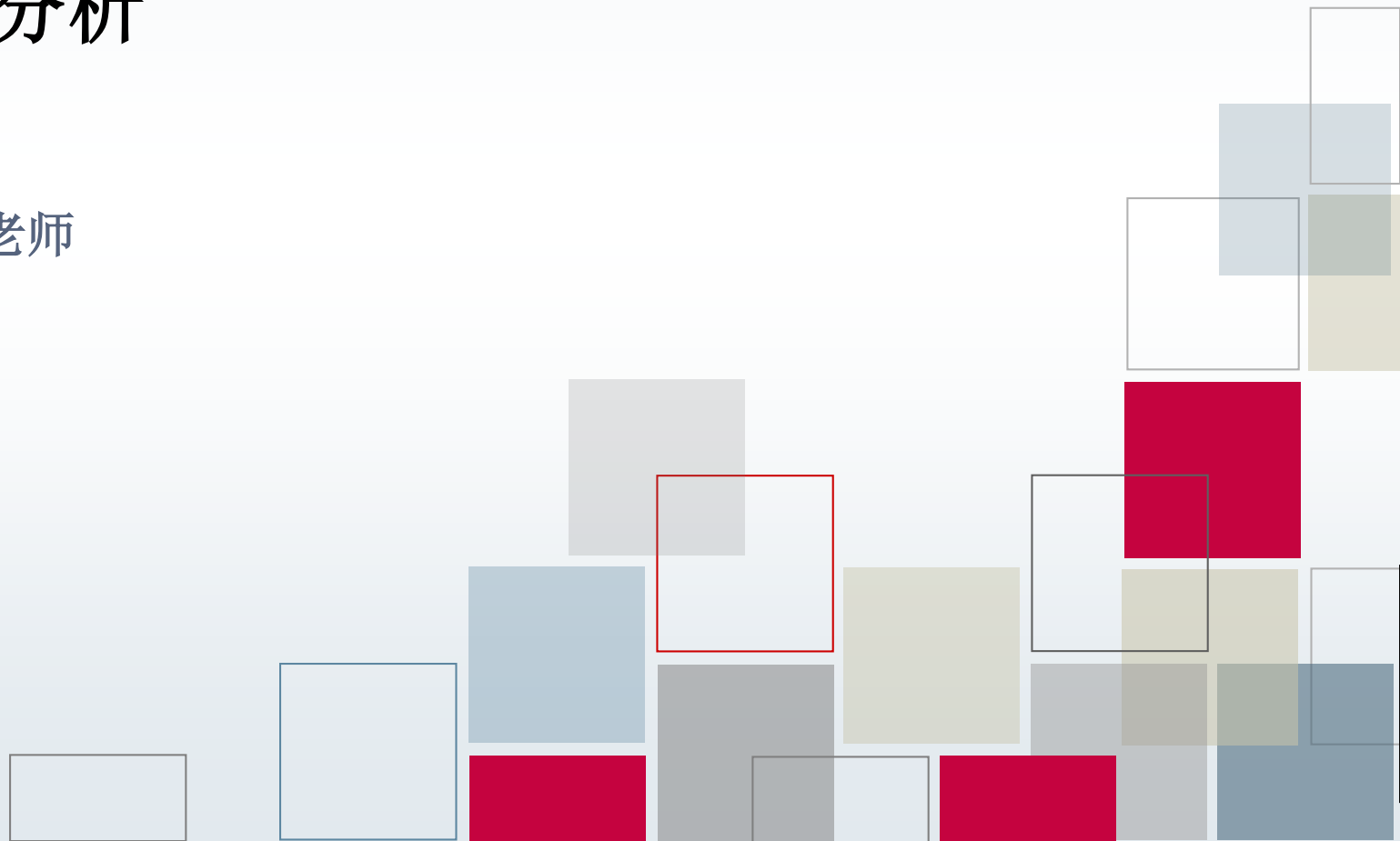
- 其他方面
  - 聚类分析主要应用于探索性的研究，其分析的结果可以提供多个可能的解，选择最终的解需要研究者的主观判断和后续的分析
  - 聚类分析的解完全依赖于研究者所选择的聚类变量，增加或删除一些变量对最终的解都可能产生实质性的影响
  - 不管实际数据中是否真正存在不同的类别，利用聚类分析都能得到分成若干类别的解

# 总结

| 内容\方法 | TwoStep | K-Means | Hierarchical |
|---|---|---|---|
| 聚类对象 | 记录 | 记录 | 记录、变量 |
| 变量类型 | 连续变量、分类变量 | 连续变量 | 连续变量、分类变量 |
| 样本量 | 大样本 (>1000) | 大样本 (>5000) | 小样本 (<1000) |
| 特点 | 自动确定最佳分类数 | 保存每个样本到类中心的距离 | 提供丰富的聚类方法和图形 |

# 判别分析

@文彤老师



# 判别分析

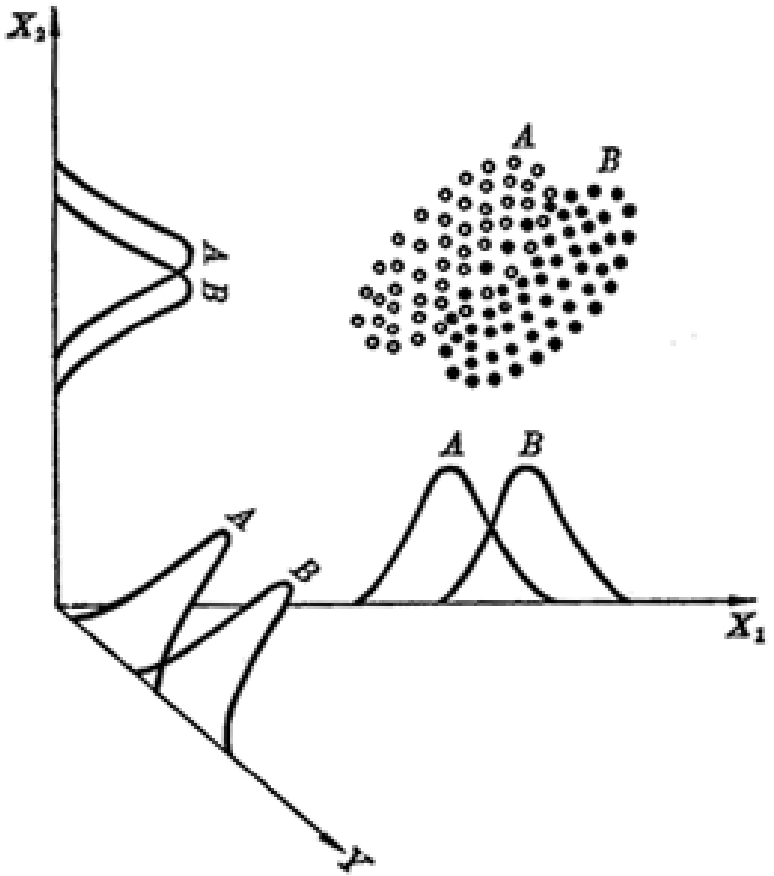
- 概述
  - 判别分析的一般形式：$y=a_1x_1+a_2x_2+\ldots\ldots+a_nx_n$
  - 非常明确共有几个类别
  - 目的是从已知样本中训练出判别函数
  - 训练样本必须由金标准确立分类
- 用途
  - 对客户进行信用预测
  - 寻找潜在客户

# 常用判别方法

- 最大似然法
  - 用于自变量均为分类变量的情况
- 距离判别
  - 对新样品求出他们离各个类别重心的距离远近
  - 适用于自变量均为连续变量的情况，对变量分布类型无严格要求

# 常用判别方法

- Fisher判别法
  - 与主成份分析有关
  - 对分布、方差等都没有什么限制
- Bayes判别
  - 计算该样品落入各个子域的概率
  - 强项是进行多类判别
  - 要求总体呈多元正态分布



# 判别分析的适用条件

- 各自变量为连续性或有序分类变量
- 样本来自一个多元正态总体
  - 该前提几乎做不到
- 各组的协方差矩阵相等
  - 类似与方差分析中的方差齐
- 变量间独立，无共线性
- 违反条件影响也不大

# 判别函数效果的验证方法

- 自身验证
- 外部数据验证
- 样本二分法
- 交互验证（Cross-Validation）
  - 刀切法

# 案例：鸢尾花数据

- Fisher在研究有关判别分析方法的时候所使用的资料，包含了刚毛、变色、弗吉尼亚这三种鸢尾花的花萼长、宽和花瓣长、宽，分析目的是希望能够使用这4个变量来对花的种类进行区分。
  - 标化函数
  - 领域图
  - 未标化典型判别函数
  - Bayes判别函数
  - 交互验证