

# 法律声明

---

□ 本课件包括：演示文稿，示例，代码，题库，视频和声音等，小象学院拥有完全知识产权的权利；只限于善意学习者在本课程使用，不得在课程范围外向任何第三方散播。任何其他人或机构不得盗版、复制、仿造其中的创意，我们将保留一切通过法律手段追究违反者的权利。

□ 课程详情请咨询

■ 微信公众号：小象

■ 新浪微博：ChinaHadoop



# 第一章 多元统计分析概述

---

主讲教师 周仕君

- 
- ❄ 一、为什么要学习多元统计分析
  - ❄ 二、多元统计分析是什么？
  - ❄ 三、多元统计分析有什么？
  - ❄ 四、多元统计分析能干什么？



# 一、为什么要学习多元统计学？

---

## 因为多元数据广泛存在！

- 比如：了解国际竞争力就需要从企业管理、国内经济实力、国际化程度、政府作用、金融环境、基础设施、科技和国民素质等八个方面进行，而每个方面又由若干个具体指标组成；
- 对一个企业的科学认识，需要了解企业的规模、产量、产值、利润、税收、所属行政所在地区、员工数、组织等多方面的信息；
- 考查一个中学生的学习情况，需要了解学生在几个主要科目的成绩。



- 面对复杂的多元数据，应该分析和认识高维复杂数据集合中的内在**规律**性，简明扼要地把握系统的**本质**特征；对高维数据集合进行最佳综合，迅速将隐没在其中的重要**信息**集中提取出来；充分发掘数据中的丰富内涵，清晰地展示系统结构。
- 利用统计学和数学方法，对多维复杂数据集合进行科学分析的理论和方法，就是多元统计数据分析研究的基本内容。



## 二、多元统计分析是什么？

---

- 多元统计分析是运用数理统计的方法来研究多变量问题的理论和方法，是一元统计学的推广。



---

□ 例如：考察学生的学习情况时，就需要了解学生在几个主要科目的考试成绩。（一元？多元？）



如果用一元统计方法，每次分析处理一门课程的成绩，会忽视课程之间可能存在的相关性，丢失的信息太多，分析的结果不能客观全面地反映学生的学习情况。

| 序号 | 政治  | 语文 | 外语 | 数学  | 物理  |
|----|-----|----|----|-----|-----|
| 1  | 99  | 94 | 93 | 100 | 100 |
| 2  | 99  | 88 | 96 | 99  | 97  |
| 3  | 100 | 98 | 81 | 96  | 100 |
| 4  | 93  | 88 | 88 | 99  | 96  |
| 5  | 100 | 91 | 72 | 96  | 78  |
| 6  | 90  | 78 | 82 | 75  | 97  |
| 7  | 75  | 73 | 88 | 97  | 89  |
| 8  | 93  | 84 | 83 | 68  | 88  |
| 9  | 87  | 73 | 60 | 76  | 84  |
| 10 | 95  | 82 | 90 | 62  | 39  |



- 
- 运用多元统计分析方法，同时对多门课程的成绩进行分析，对诸课程间的相关性、相依性和相当重要性等都能提供有用的信息。
  - ✓ 用各科成绩的总和作为综合指标，来比较学生学习成绩的好坏；
  - ✓ 根据各科成绩相近程度对学生分类（如成绩好的与成绩差的，又如文科成绩好的与理科成绩好的）；
  - ✓ 研究各科成绩之间的关系（如物理与数学成绩的关系，文科成绩与理科成绩的关系）等。



---

□ 因此，一元统计分析是研究一个随机变量统计规律性的学科，而多元统计分析是研究多个随机变量之间相互依赖关系以及内在统计规律性的一门统计学科。



# 三、多元统计分析有什么？

---

- ☐ 判别分析
- ☐ 聚类分析
- ☐ 主成分分析
- ☐ 因子分析
- ☐ 典型相关分析
- ☐ 相应分析
- ☐ 多重回归分析
- ☐ 可视化分析



□ 事实上，多元统计分析是以 $p$ 个变量的 $n$ 次观测数据所组成的数据矩阵

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & & \vdots \\ x_{p1} & x_{p2} & \cdots & x_{pn} \end{bmatrix}$$

为依据，根据实际问题的需要，所给出的种种方法。多元统计分析所研究的内容可以概括为以下几个方面：



## 1. 简化数据结构（降维问题）

- 简化数据结构即是将某些较复杂的数据结构通过变量变换等方法使相互依赖的变量变成互不相关的；或把高维空间的数据投影到低维空间，使问题得到简化而损失的信息又不太多的。
- 主成分分析、因子分析以及对应分析等。



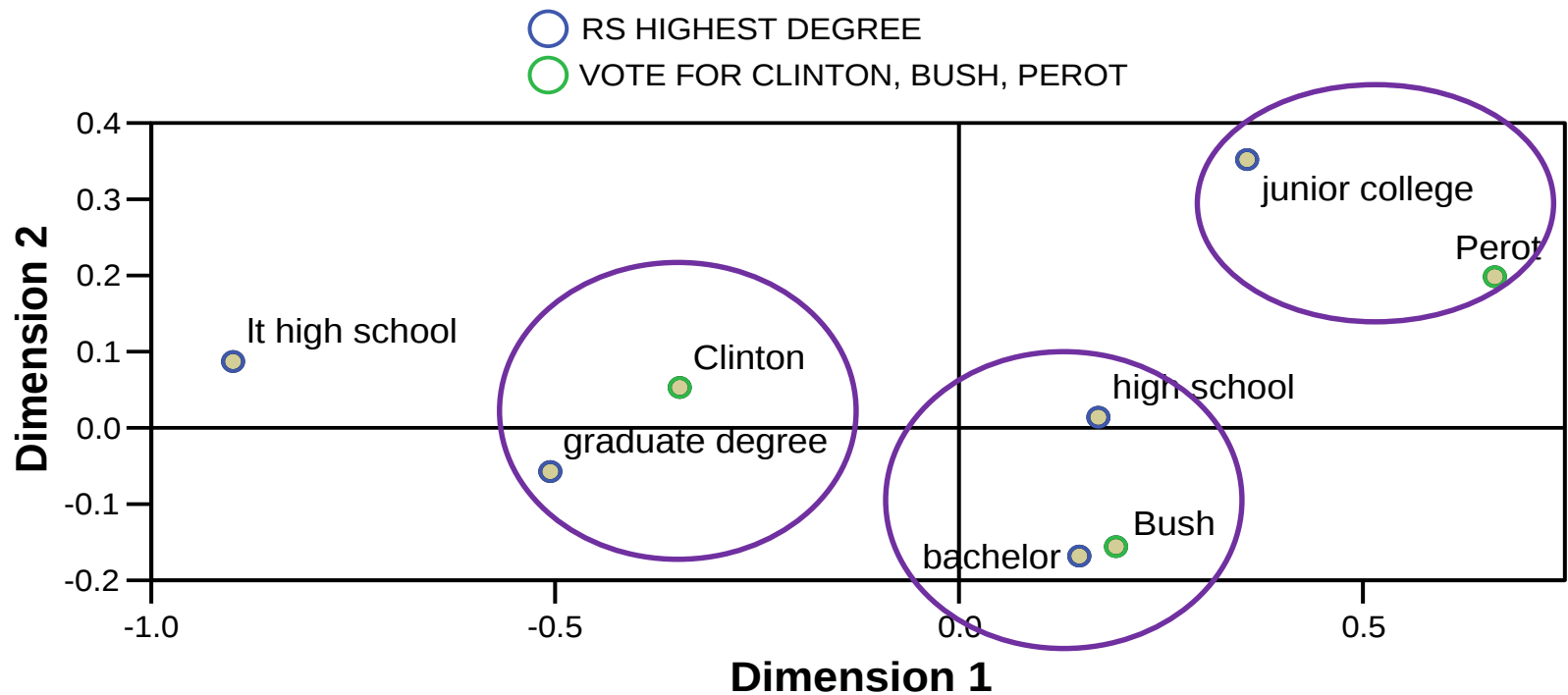
# 相应分析案例

---

- 数据为1992年美国大选的部分数据。要求对选民的最高学历水平（degree）和所支持的总统候选人（pres92）进行相应分析。



## Row and Column Points Symmetrical Normalization



# 因子分析案例

- 学生高考成绩，要考语文、数学、物理、化学、生物、历史、地理、外语等科目。从教育学上来看，学生各科成绩的好坏主要是各个学生的阅读能力、抽象能力、记忆能力和学习刻苦努力程度支配的。因此，学生各科成绩这些指标的变动就取决于上述四个共同因素和其它一些特殊因素的作用。
- 再例：对国民经济系统运行的观测指标有国内生产总值、国民收入、社会商品零售额、库存额、物价指数、货币发行量、失业率、固定资产投资率、城乡居民储蓄存款额、固定资产总额、财政收入、财政支出、进出口总额、国际收支余额等。这些指标的数值大小取决于该国的社会经济制度、银行制度、市场机制、人口数量、科学技术水平等因素。



# 主成份分析的经典案例

---

- ❑ 美国的统计学家斯通在1947年所做的关于国民经济的研究中，曾利用美国1929年-1938年各年的数据，得到17个反映国民收入与支出的变量要素，如雇主补贴、消费资料和生产资料、纯公共支出、净增库存、股息、利息、外贸平衡等。
- ❑ 在他进行主成份分析后，发现能够以97.4%的精度，用3个新变量取代原来的17个变量。这3个新变量可以分别命名为总收入、总收入变化率 and 经济发展或衰退的趋势。
- ❑ 更有趣的是，这3个变量其实都是可以直接测量的。



---

## 2. 分类与判别（归类问题）

- 归类问题是对所考察的观测点（或变量）按相似程度进行分类（或归类）。
- 方法：聚类分析和判别分析。



- 根据城市发展的情况，对城市进行分类。
- 对不同客户，根据其消费信贷情况，对其进行分类。
- 对作品著作权的归属进行分析。
- 市场细分。市场细分是指营销者通过市场调研，依据消费者的需要、购买行为和购买习惯等方面的特征差异，把某一产品的市场整体划分为若干消费者群的市场分类过程。每一个消费者群就是一个细分市场，每一个细分市场都是具有类似需求倾向的消费者构成的群体。



### 3. 变量间的相互联系

- 相互依赖关系：分析一个或几个变量的变化是否依赖于另一些变量的变化？如果有依赖关系，则建立变量间的定量关系式，并用于预测或控制——回归分析。
- 变量间的相互关系：分析两组变量间的相互关系——典型相关分析。



- 在对社会经济现象问题的研究中，通常需要考察多个变量与多个变量之间即两组变量之间的相关性。
- 例如，为了研究证券市场走势与宏观国民经济之间的关系，需要分析反映证券市场状况的指标，如股票价格指数、股票市值、股票融资量等与反映宏观经济的指标如经济增长率、物价指数、固定资产投资、失业率、进出口额等两组变量之间的相关关系。



---

## 4.预测与决策

- 通过统计模型或最优准则，对未来进行预见或判断。
- 方法：多元回归、判别分析、聚类分析、可视化分析。



- 
- 利用公司会计数据信息，构造识别具有潜在财务危机上市公司的方法。
  - 根据产品销售状况与企业投放广告情况、价格水平、促销力度、产品质量、竞争产品等因素的关系，对某产品的销售情况进行预测。



# 四、多元统计分析能干什么？

---

 统计学的生命力在于应用

 多元统计分析方法的应用



- 
- 统计方法是科学研究的一种重要工具，其应用颇为广泛。特别地，多元统计分析方法常常被应用于自然科学、社会科学等领域的问题中。



# (一) 教育学

---

- 假如 $n$ 个考生参加 $p$ 门课程的考试。那么可以对这些成绩作多元统计分析。
- 利用 $n$ 个学生在高中学习期间 $m$ 门主要课程的考试成绩，可对学生进行分类，并分析每一类学生的特点。



## （二）医学

---

- 医生对病人的诊断是靠对病人观测若干症状后来综合评定的。如一个人发高烧，医生根据他的体温高低、白血球数目及其他症状来判断他是患感冒、肺炎还是其他疾病。
- 为了判断准确可靠，事先应有一批经专家确诊或手术后经病理化验确诊的病例资料，根据这批资料利用多元统计方法可建立诊断的准则（专家系统）。对来就诊的病人，按专家系统的要求观测若干项指标后，根据诊断准则作出诊断。



### (三) 气象学

---

- 全国各地建立了很多气象站，在不同时间各气象站都记录了降雨量、气温、气压、湿度、风速、风向等气象指标资料。对这些资料作统计分析：
- 例如降雨量与前一天的气温、气压、湿度等的都有关系，利用该关系可对降雨量作预报。



## （四）环境科学

---

- 例如：为了了解某大型化工厂对环境的污染程度，在厂区建立很多监测点，每天定时测定各种污染气体的浓度。
- 用统计分析方法分析处理这些资料，可对厂区按污染情况分为几类，如分为严重污染、一般污染和轻污染三类；并为今后监测点的布局提供既合理又经济的方案。



## （五）经济学

---

- 为了研究不同地区农民收支的分布规律，抽样调查了全国31个省市自治区的农民生活消费支出情况。用聚类分析方法31个地区分类，根据分类结果还可进一步研究各类地区农民的生活水平、富裕程度，以便进一步研究经济发展对策。
- 根据人均国民收入、人均工农业产值、人均消费水平等多种指标来判定一个国家的经济发展程度所属类型。



## （六）服装业—服装的定型分类

---

- 多元统计分析方法在服装业上的应用。首先在各地做抽样调查，对被调查人测量其身体的几十个部位的尺寸；然后对庞大的调查资料用多元统计方法分析处理，确定一种服装究竟需要有多少种型号，每种型号服装的比例是多少，由身体的哪几个主要部位的尺寸决定。



## (七) 信用评级、农业及其他

例如：多元统计分析方法在其他很多领域中也  
有它的应用。比如体育科研、军事科学、生  
物学、心理学、生态学、保险科学、火警预  
报、地震预报、中医阴阳学说研究等。



# 五、多元统计分析的工具

---

## □ R语言



# 疑问

---

□ 问题答疑：<http://www.xxwenda.com/>

■ 可邀请老师或者其他人回复问题

# 联系我们

---

## 小象学院：互联网新技术在线教育领航者

- 微信公众号：小象
- 新浪微博：ChinaHadoop

