多元统计分析
绪论

李高荣

北京师范大学统计学院

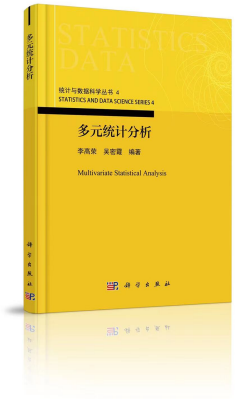
E-mail: *liqaorong@bnu.edu.cn*

微信公众号：BNUlgr

- 扫二维码获取在线课件和相关教学电子资源
- 请遵守电子资源使用协议

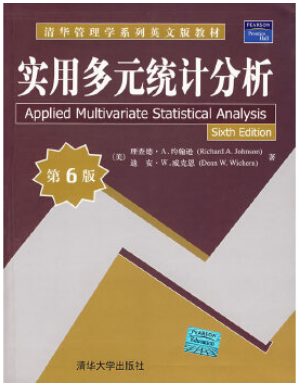
本课程使用教材

- 李高荣，吴密霞 (2021). 多元统计分析. 北京：科学出版社.



主要参考教材

- Johnson, R. A. and Wichern, D. W. (2008). *Applied Multivariate Statistical Analysis*(Sixth Edition). 北京：清华大学出版社.
- 高惠璇 (2004). 应用多元统计分析. 北京：北京大学出版社.



主要参考教材

- 张尧庭, 方开泰(1982). 多元统计分析引论. 北京: 科学出版社.
- 张润楚(2006). 多元统计分析. 北京: 科学出版社.
- 王静龙(2008). 多元统计分析. 北京: 科学出版社.
- 姜丹丹, 白志东(2014). 大维矩阵谱理论在多元统计分析中的应用. 北京: 知识产权出版社.
- 吴密霞, 刘春玲(2016). 多元统计分析. 北京: 科学出版社.
- Härdle, W. K. and Simar, L. (2012). *Applied Multivariate Statistical Analysis*(Third Edition). Springer.

课程要求

- **课程要求：**
  - 多元统计分析的基础课程：高等代数、矩阵论、数理统计
  - 理解各种分析方法的思想、原理、理论、应用
  - 认真完成课后作业，至少能掌握1种统计软件进行多元统计分析(如R语言、Matlab、SPSS等)

- **考试：**
  1. 期末总评=期末考试+平时成绩(课堂表现、出勤率、作业等)
  2. 加权计算

课程介绍

- **多元统计分析：** 是应用数理统计学来研究多变量(多指标)问题的理论和方法，它是一元统计学的推广和发展。
  - 是统计学的一个重要分支
  - 是一门具有很强应用性的课程
  - 在自然科学和社会科学等各个领域中得到广泛的应用
  - 包括了很多非常有用的数据处理方法

例子：

1. **地区经济发展的指标：** 总产值、利润、效益、劳动生产率、固定资产、物价、信贷、税收等
2. **医学诊断：** 血压、脉搏、白血球、体温等

- 多重观测数据：许多观测或设计研究中，每个试验单元的多个指标被同时观测或收集

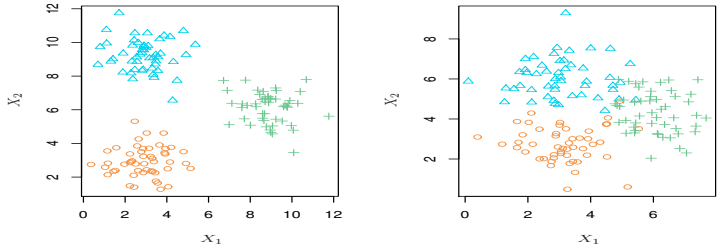
$$\boldsymbol{X}_i = (X_{i1}, \cdots, X_{ip})', \qquad i = 1, \cdots, n$$

- 多元统计分析是一类用于分析多重观测数据的方法
- 基本想法是利用多重观测之间的潜在相关性来提升推断效率
- 一些多元技术基于特定的概率模型，特别是多元正态分布，其他不依赖于特定分布的方法称为“模型自由的" (model-free)

多元统计方法的应用

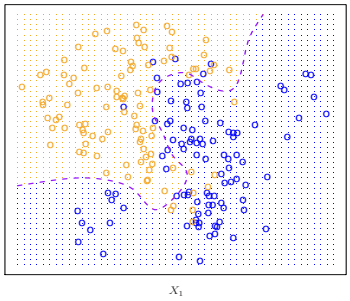
- 维数缩减或降维: 通过考虑大量测量变量的少部分组合来降低维数, 同时不损失重要的信息。用途: 多元数据可视化, 发现重要特征(变量)。
  - 消费者价格指数(CPI) 通过组合一大类商品价格来得到。
  - 体脂肪健康指数(BMI) 通过测量并组合身高和体重观测。值来得到，$\mathrm{BMI} = w/h^2$，其中w是体重(单位：kg)，h 表示身高(单位：m)
  - MDS 通过研究对象之间某种亲近关系为依据（如距离、相似系数等）, 将研究对象（样品或变量）在低维空间中给出标度或位置, 以便全面而又直观地再现原始各研究对象之间的关系，同时在此基础上也可按对象点之间距离的远近实现对样品的分类。

- 聚类(clustering): 识别观测单元中“相似" 的单元
  - 电子商务通过分组聚类出具有相似浏览行为的客户，并分析客户的共同特征，可以更好的帮助电子商务的用户了解自己的客户，向客户提供更合适的服务。
  - 基因表达数据: 包含了64个癌症患者细胞系中6830 个基因表达的测量数据。研究兴趣: 基因表达测量上的细胞系数据中是否有集群存在。
- 聚类目标: 是基于观测值$\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n$，将观测值归入不同的群



多元统计方法的应用

- 分类: 使用特定的指标集将观测单元分为事先指定的类
  - 美国国税局使用退税信息(收入, 扣缴税款, 捐款, 年龄等) 将纳税人分为两组: 需要审查和不需要审查
  - 通过检测铅合金中元素(铜, 银, 锡, 锑) 的含量, 公安机构可以判断一些子弹是否来自同一批次



多元统计方法的应用

- 相关性分析: 变量之间的关联性是什么?
  - 搜索引擎与使用它的人之间的桥梁就是网站的相关性, 用户通过搜索引擎检索跟网站相关的内容找到该网站, 而搜索引擎通常使用相关性规则, 来展示搜索结果。一个有极高相关性的匹配是对那个搜索请求排名第一的候选结果
- 预测: 若变量之间是有关联的, 则可以通过给定的信息来预测另一些变量
  - 用历史天气数据预测未来几天的天气情况
  - 基于用户移动通信记录数据, 对用户流失进行预测
- 假设检验: 可否发现两组或多组响应变量之间的差异?
  - 测量一些与污染有关的变量, 以研究一个城市地区的污染程度是在一周中大致保持不变, 还是在工作日和周末之间会有明显的不同
  - 利用观测数据来研究职业结构的差异, 以决定支持两个对立的社会理论中的哪一个

一些记号

- $\mathbf{X}$表示一个$n \times p$矩阵，表示为

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$

- $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \cdots, x_{ip})'$表示长度为$p$的列向量
- $\mathbf{x}_j = (x_{1j}, x_{2j}, \cdots, x_{nj})'$表示长度为$n$的列向量
- $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_p) = \begin{pmatrix} x_1' \\ x_2' \\ \vdots \\ x_n' \end{pmatrix}$
- $\mathbf{y} = (y_1, y_2, \cdots, y_n)'$


多元统计分析
STATISTICS DATA
统计与数据科学丛书 4
STATISTICS AND DATA SCIENCE SERIES 4
多元统计分析
李高荣 吴密霞 编著
Multivariate Statistical Analysis
科学出版社

谢谢，请多提宝贵意见！