多元统计分析

- 石坚
- 办公室：数学研究院（中关村）南楼510
- 电话：　13521367153，82541602
- Email：　jshi@iss.ac.cn

- 助教：
- 薛原：　13120373836
- Email：　xueyuan115@mails.ucas.ac.cn

- 教　材：

  王静龙：多元统计分析. 科学出版社，2008.

- 参考书：

1. 张尧庭, 方开泰. 多元统计分析引论.
   科学出版社, 1982.
2. Anderson, T.W.. An Introduction to
   Multivariate Statistical Analysis, 3rd ed.
   New York: John Wiley & Sons, 2003.
3. 白志东, 郑术蓉, 姜丹丹. 大维统计分析.
   高等教育出版社，2012.

课程简介

- 多重（维）观测数据：在观测或者设计研究中，每个试验单元有 p 个指标被同时观测收集，记为

$$Y_i = (Y_{i1}, \ldots, Y_{ip})', \quad i = 1, \ldots, n.$$

- 多元分析是一类用于分析多重（维）观测数据的统计学方法。

- 基本想法是利用多重（维）观测之间的潜在相关性来提升推断效率。

- 一些多元分析的技术基于特定的概率模型，特别是**多元正态分布**以及由其导出的分布；不依赖于特定分布的方法称为“分布自由”方法(distribution-free)。

- 多元分析方法的应用

- 维数压缩：通过对大量观测变量的部分组合来降低变量的维数，同时不损失重要的信息，降低数据存储量。

  - 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。

  - 主成分分析、因子分析

例子：如何制定成年男子上衣服装号码的案例

成年男子上衣的8个人体部位尺寸的均值与标准差

(样本量: 5115, 单位: cm)

部位	均值	标准差
身高	167.48	6.09
颈椎点高	142.91	5.60
腰围高	100.58	4.44
坐姿颈椎点高	65.61	2.67
颈围	36.83	2.11
胸围	87.53	5.55
后肩横弧	43.24	2.75
臂全长	54.53	3.04

成年男子上衣的8个人体部位尺寸的协方差阵

	身高	颈椎点高	腰围高	坐姿颈椎点高	颈围	胸围	后肩横弧	臂全长
身高	37.115							
颈椎点高	33.069	31.314						
腰围高	24.631	22.624	19.739					
坐姿颈椎点高	12.364	11.506	7.119	7.131				
颈围	2.695	2.593	1.217	1.575	4.437			
胸围	11.155	11.177	6.163	5.334	7.013	30.784		
后肩横弧	7.367	7.075	4.030	3.229	2.084	7.472	7.554	
臂全长	12.597	11.911	9.322	3.573	0.577	4.049	2.340	9.246

目标：找出一个或几个指标来制定成年男子上衣的号型。

可行方法：主成分分析

例子：分析我国各省市自治区的农业生产情况案例

从农业生产条件、生产结果和效益出发，选取六项指标，分别为：

X_1：乡村劳动力人口（万人）；
X_2：人均经营耕地面积（亩）；
X_3：户均生产性固定资产原值（元）；
X_4：家庭基本纯收入（元）；
X_5：人均农业总产值（千元/人）；
X_6：增加值占总产值比重（%）。

序号	地 区	X_1	X_2	X_3	X_4	X_5	X_6
1	北 京	66.9	0.93	2972.41	3290.73	2.525	49.7
2	天 津	80.2	1.64	4803.54	2871.62	1.774	49.6
3	河 北	1621.8	2.03	4803.54	2871.81	0.8004	54
4	山 西	635.4	2.76	2257.66	1499.14	0.555	56.2
5	内蒙古	514.1	10.17	5834.94	1550.15	0.9051	66.4
6	辽 宁	605.1	2.96	3108.86	2059.35	1.4752	53.1
7	吉 林	534.2	4.73	4767.51	1940.46	1.1154	63.1
8	黑龙江	494.8	8.24	5573.02	2075.42	1.6283	57.8
9	上 海	66	1.02	1660.03	4571.81	3.0448	35.6
10	江 苏	1530.2	1.26	2826.86	2868.33	1.1921	50.6
11	浙 江	1123.1	0.94	5494.23	3289.07	0.8565	63.3
12	安 徽	1953.6	1.44	3573.62	1508.24	0.5756	59.2
13	福 建	775.8	0.82	2410.05	2295.19	1.1496	62.8
14	江 西	1103.2	1.3	2310.98	1804.93	0.6649	59.9
15	山 东	2475.1	1.44	3109.11	1989.53	0.8809	55
16	河 南	2815.8	1.5	3782.26	1508.36	0.5823	58.5
17	湖 北	1296.5	1.6	2291.6	1754.13	0.8799	62.8
18	湖 南	2089.3	1.42	2348.72	1719.18	0.587	64.7
19	广 东	1439.8	0.88	3249.61	2928.24	1.096	59.7
20	广 西	1579.9	1.43	3090.17	1590.9	0.5694	64.5
21	海 南	165.9	1.35	4454.77	1575.49	0.3535	65.2
22	四 川	3903.7	1.08	2870.45	1340.61	0.4443	64.1
23	贵 州	1376.6	1.18	2282.27	1206.25	0.2892	65.4
24	云 南	1642.2	2.42	4025.06	1096.73	0.3456	64.2
25	西 藏	88.6	2.51	11559.83	1257.71	0.4349	70.4
26	陕 西	1046.1	2.6	2228.55	1091.96	0.4383	59.7
27	甘 肃	672	5.86	2879.36	1037.12	0.4883	57.2
28	青 海	137.1	2.62	6725.11	1133.06	0.4096	70.3
29	宁 夏	139.1	4.01	5607.97	1346.89	0.4973	62.5
30	新 疆	288.5	3.96	7438.13	1161.71	1.4939	57.8

目标：分析

1）六个指标可以归为几类？
2）影响每类指标的潜在因素是什么？
3）30个省区市可分为几个农业生产类型？

可行方法：因子分析

- **聚类**：识别观测单元中“相似”的单元 （**无监督或半监督学习**）
  - 电子商务通过分组聚类出具有相似浏览行为的客户，分析客户的共同特征，以便向客户提供更合适的服务。
  - 通过对设备的性能检测信号进行分析，将设备的非正常状态聚类成多个故障模式，对设备进行故障诊断。
  - 聚类分析

例子：体育彩票投注策略案例

盘口：竞彩彩票的价格变化曲线。

目标：如何基于“盘口”确定投注策略，如:“胜、平、负”等，
以期获得正收益。

可行方法：聚类分析

- **分类**：利用特定指标集将观测单元归于事先指定的类（**有监督学习**）
  - 通过医检指标判别检测对象为“正常”或“患病”类。
  - 判别分析

例子：儿童阻塞性睡眠呼吸暂停综合症判别分析案例

阻塞性睡眠呼吸暂停综合症（OSAS）是以睡眠时上气道阻塞为特征，通常伴有血氧饱和度下降和（或）高碳酸血症。

OSAS在儿童中是一种较常见的疾病，如果不及时治疗，可能导致生长障碍等多种生理缺陷。

OSAS最常见的临床表现是打鼾，但并不是打鼾的儿童就有OSAS。有些儿童虽然夜间打鼾但无生理异常，医学上称为原发鼾症（PS）。

因此，把OSAS和PS区分开来非常重要，因为前者应及早诊断和治疗，而后者无需处理。

可行方法：判别分析

- **相关性分析**：研究变量之间的关联性

  - 用户通过搜索引擎，检索跟网站相关的内容以找到该网站， 搜索引擎通常使用**相关性**规则来展示搜索结果
  - 典型相关分析

例子：家庭特征与家庭消费之间的关系案例

为了解家庭的特征与其消费模式之间的关系，调查70个家庭的下面两组变量

消费模型变量

$$\begin{cases} X_1: \text{每年去餐馆就餐的频率} \\ X_2: \text{每年外出看电影频率} \end{cases}$$

家庭特征变量

$$\begin{cases} Y_1: \text{户主的年龄} \\ Y_2: \text{家庭的年收入} \\ Y_3: \text{户主受教育程度} \end{cases}$$

目的：分析两组变量之间的关系。

可行方法：典型相关分析。

- **预测**：变量之间若有关联，则可以通过部分变量的信息来预测其它变量

  - 利用高中成绩变量与大学成绩变量之间的联系，构造用于预测在大学里会成功与否的指标

  - 线性模型、相关分析

- **假设检验**：检验两组或多组响应变量之间的差异

  - 某类癌症患者在不同治疗方案下的生存寿命差异。

  - 方差分析、统计检验

例子：有4种新研制的镇痛药. 为检验新药相对于原有药物的镇痛效果, 选取情况相似的9个病人作试验. 对每个病人先后分别按随机排列的次序使用这4种新药以及原有药物镇痛, 镇痛时间(min)见下表：

		病　人								
		1	2	3	4	5	6	7	8	9
原有药物		15.8	16.7	15.7	14.0	16.2	13.7	15.9	17.9	15.8
新药	A	17.8	15.9	17.7	17.4	19.2	17.6	16.7	17.4	17.6
	B	19.1	20.0	18.0	19.3	20.0	19.1	19.0	20.4	19.4
	C	16.8	14.9	16.9	15.8	14.4	14.8	16.2	17.6	16.6
	D	21.4	20.4	20.1	21.3	19.4	20.2	21.1	21.2	20.3

问题：这4种新药的镇痛效果是否与原有药物等效.

方法：多重比较检验.

- 多元分析方法的数学工具

- **数学分析（微积分）：**
  微分（偏导），积分

- **线性代数：**
  矩阵的基本性质（秩、迹、行列式、特征分解）与运算

- **概率统计：**
  概率分布、概率密度、特征函数、数学期望、方差（协方差）