

多元统计分析方法简介

参考文献

- ❖ 数据挖掘方法与模型, Daniel T. Larose, 高等教育出版社
- ❖ 数据分析方法五种, 尤恩.苏尔李, 上海人民出版社(缺失数据插补的方法)
- ❖ 问卷统计分析实务——SPSS操作与应用, 吴明隆, 重庆大学出版社
- ❖ 应用商务统计分析, 王汉生, 北京大学出版社
(介绍了各种不同的回归模型)
- ❖ 应用多元统计分析, 沃尔夫冈等, 陈诗一译, 北京大学出版社 (各种多元统计模型, 比较全)

如何快速学习多元统计方法

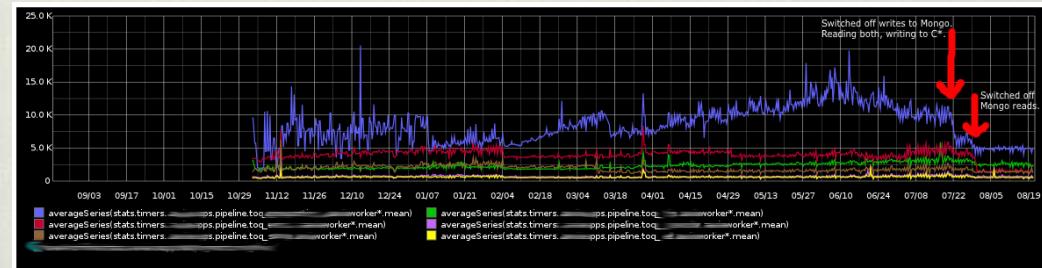
- ❖ 了解具体方法用于解决什么问题;
- ❖ 该方法用于什么样的数据类型;
- ❖ 使用该方法的基本前提是什么, 即数据应该满足的基本条件, 需要检验;
- ❖ 在统计软件上试运行(R, SPSS, Matlab)

数据（变量类型）

- ❖ 定类数据（属性数据）
- ❖ 定序数据（非参数检验）
- ❖ 定距数据
- ❖ 定比数据（身高、体重、时间）

❖ 时间序列

❖ 面板数据



- ❖ 定序数据可以通过求和、加权求和等方法换算成可按定距数据处理的数据，从而可以用各种统计分析方法。

——量表的转换

(一) 方差分析

- ❖ 什么是方差分析(ANOVA)?(analysis of variance)

1. 检验多个总体均值是否相等

通过分析数据的误差判断各总体均值是否相等

2. 研究分类型自变量对数值型因变量的影响

一个或多个分类尺度的自变量

两个或多个 (k 个) 处理水平或分类

一个间隔或比率尺度的因变量

3. 有单因素方差分析和双因素方差分析

- ❖ 单因素方差分析：涉及一个分类的自变量
- ❖ 双因素方差分析：涉及两个分类的自变量

【例】为了对几个行业的服务质量进行评价，消费者协会在四个行业分别抽取了不同的企业作为样本。最近一年中消费者对总共23家企业投诉的次数如下表

消费者对四个行业的投诉次数				
	行业			
观测值	零售业	旅游业	航空公司	家电制造业
1	57	68	31	44
2	66	39	49	51
3	49	29	21	65
4	40	45	34	77
5	34	56	40	58
6	53	51		
7	44			



(例题分析)

1. 分析四个行业之间的服务质量是否有显著差异，也就是要判断“行业”对“投诉次数”是否有显著影响；
2. 作出这种判断最终被归结为检验这四个行业被投诉次数的均值是否相等；
3. 若它们的均值相等，则意味着“行业”对投诉次数是没有影响的，即它们之间的服务质量没有显著差异；若均值不全相等，则意味着“行业”对投诉次数是有影响的，它们之间的服务质量有显著差异。

双因素方差分析

(two-way analysis of variance)

1. 分析两个因素(行因素Row和列因素Column)对试验结果的影响
2. 如果两个因素对试验结果的影响是相互独立的，分别判断行因素和列因素对试验数据的影响，这时的双因素方差分析称为**无交互作用的双因素方差分析或无重复双因素方差分析**(Two-factor without replication)
3. 如果除了行因素和列因素对试验数据的单独影响外，两个因素的搭配还会对结果产生一种新的影响，这时的双因素方差分析称为**有交互作用的双因素方差分析或可重复双因素方差分析**(Two-factor with replication)

双因素方差分析

(例题分析)

【例】有4个品牌的彩电在5个地区销售，为分析彩电的品牌(品牌因素)和销售地区(地区因素)对销售量是否有影响，对每种品牌在各地区的销售量取得以下数据。试分析品牌和销售地区对彩电的销售量是否有显著影响？($\alpha=0.05$)

不同品牌的彩电在各地区的销售量数据					
品牌因素	地区因素				
	地区1	地区2	地区3	地区4	地区5
品牌1	365	350	343	340	323
品牌2	345	368	363	330	333
品牌3	358	323	353	343	308
品牌4	288	280	298	260	298

方差分析的基本假定

1. 每个总体都服从正态分布

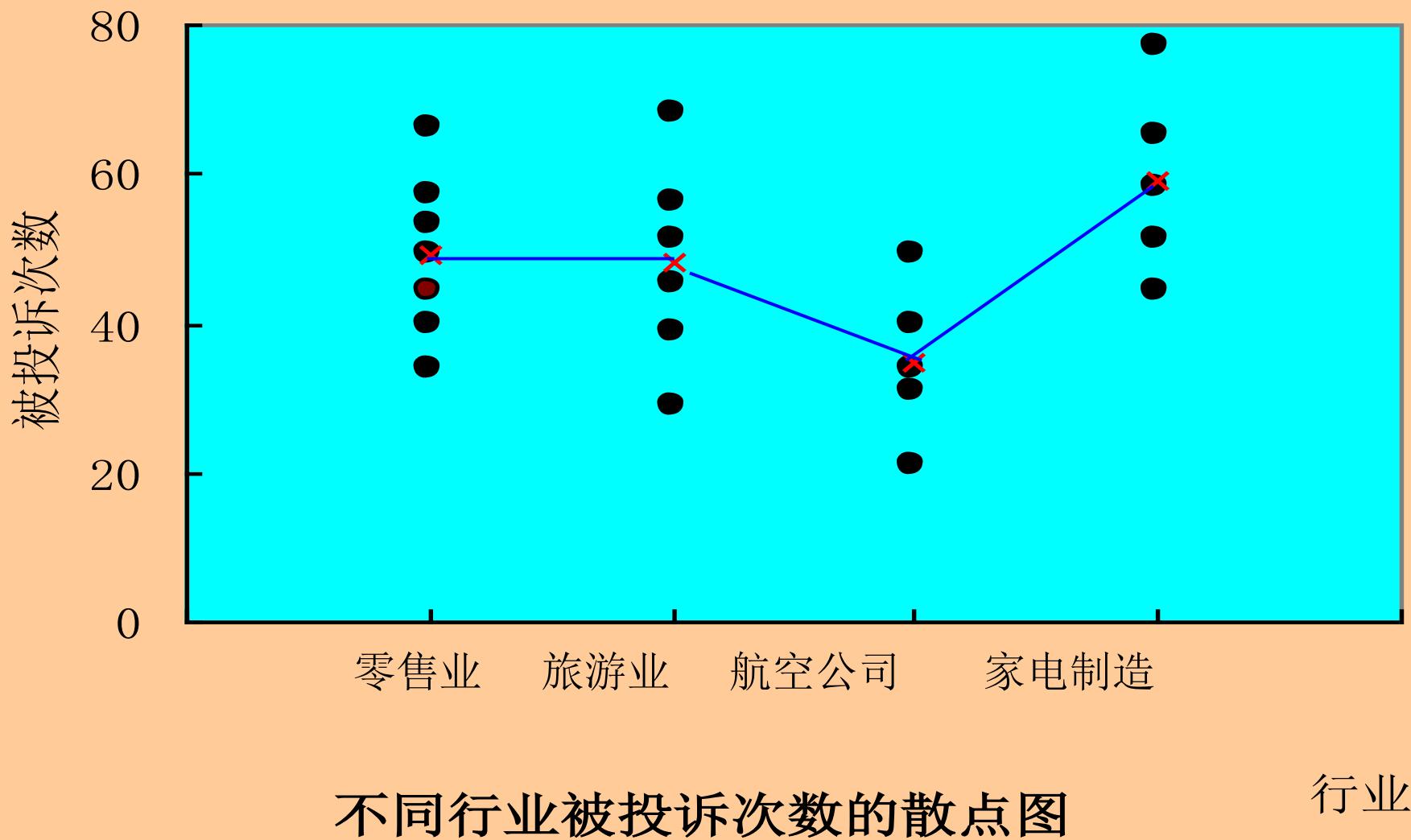
- ❖ 对于因素的每一个水平，其观察值是来自正态分布总体的简单随机样本

2. 各个总体的方差必须相同

- ❖ 对于各组观察数据，是从具有相同方差的总体中抽取的

3. 观察值是独立的

方差分析的基本思想和原理 (图形分析)

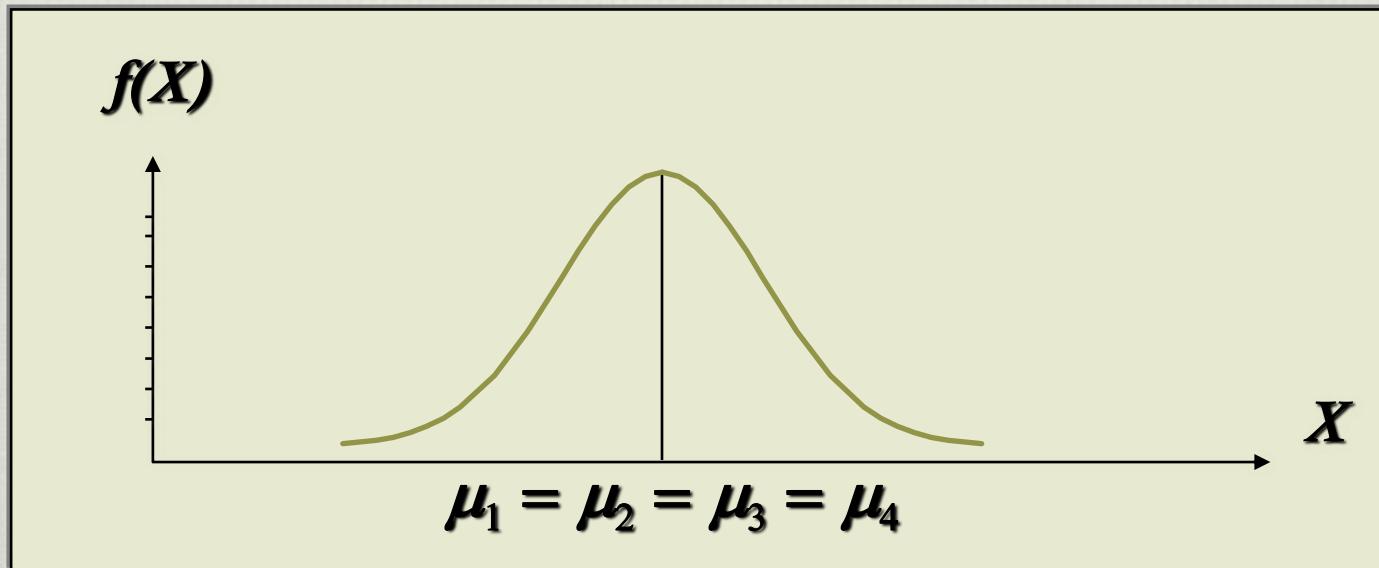


方差分析的基本思想和原理

1. 仅从散点图上观察还不能提供充分的证据证明不同行业被投诉的次数之间有显著差异
 - ❖ 这种差异也可能是由于抽样的随机性所造成的
2. 需要有更准确的方法来检验这种差异是否显著，也就是进行方差分析
 - ❖ 所以叫方差分析，因为虽然我们感兴趣的是均值，但在判断均值之间是否有差异时则需要借助于方差
 - ❖ 这个名字也表示：它是通过对数据误差来源的分析判断不同总体的均值是否相等。因此，进行方差分析时，需要考察数据误差的来源

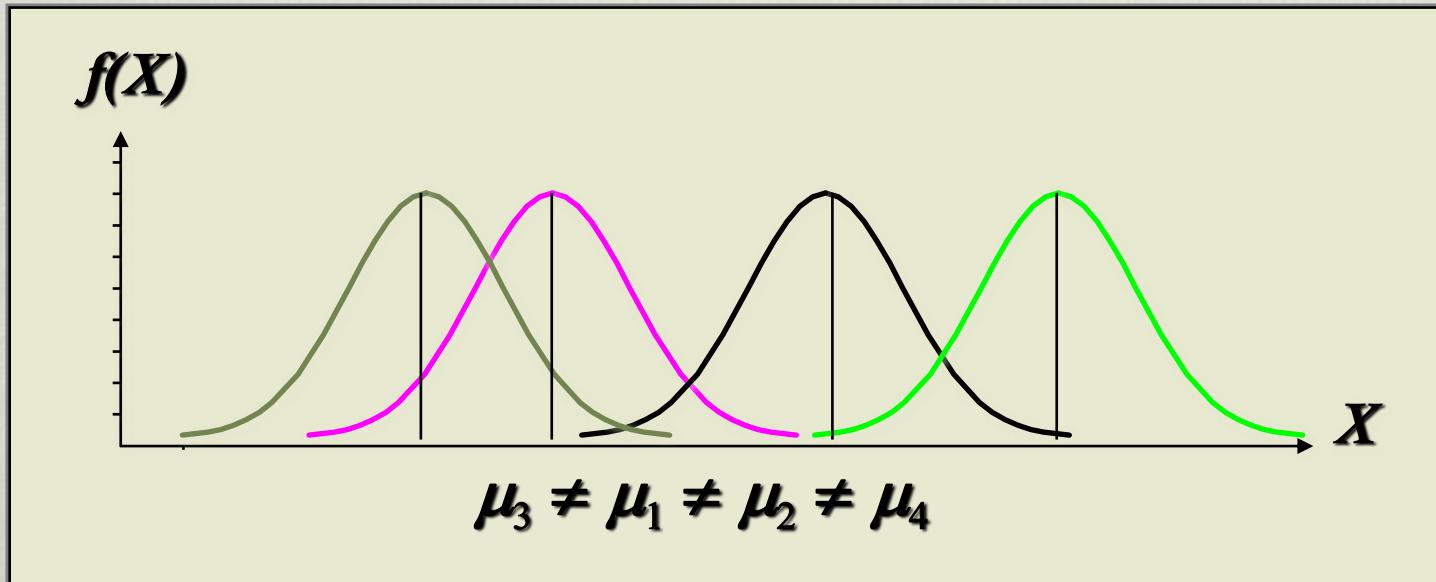
方差分析中基本假定

- ❖ ➔ 如果原假设成立，即 $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$
 - ❖ 四个行业被投诉次数的均值都相等
 - ❖ 意味着每个样本都来自均值为 μ 、方差为 σ^2 的同一正态总体



方差分析中基本假定

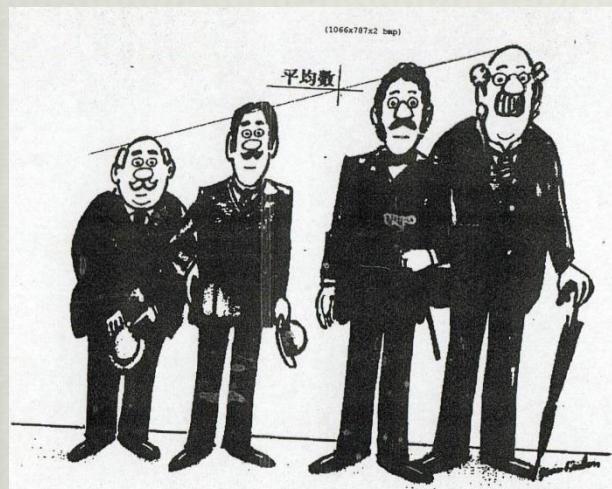
- ◆ → 若备择假设成立，即 $H_1: \mu_i (i=1,2,3,4)$ 不全相等
 - ❖ 至少有一个总体的均值是不同的
 - ❖ 四个样本分别来自均值不同的四个正态总体



(二) 相关分析与回归分析

回归的古典意义：

高尔顿遗传学的回归概念



父母身高与子女身高的关系：
无论高个子或低个子的子女
都有向人的平均身高回归的
趋势

$$y=0.516x+33.73 \text{ (in)}$$

实例1： 中国妇女生育水平的决定因素是什么？

妇女生育水平除了受计划生育政策影响以外，还可能与社会、经济、文化等多种因素有关。

1. 影响中国妇女生育率变动的因素有哪些？
2. 各种因素对生育率的作用方向和作用程度如何？
3. 哪些因素是影响妇女生育率主要的决定性因素？
4. 如何评价计划生育政策在生育水平变动中的作用？
5. 计划生育政策与经济因素比较，什么是影响生育率的决定因素？
6. 如果某些地区的计划生育政策及社会、经济、文化等因素发生重大变化，预期对这些地区的妇女生育水平会产生怎样的影响？

实例2:

全球吃死的人比饿死的人多？

据世界卫生组织统计，全球肥胖症患者达3亿人，其中儿童占2200万人，11亿人体重过重。肥胖症和体重超常早已不是发达国家的“专利”，已遍及五大洲。目前，全球因”吃”致病乃至死亡的人数已高于因饥饿死亡的人数。（引自《光明日报》刘军/文）

问题：肥胖症和体重超常与死亡人数真有显著的数量关系吗？

这些类型的问题可以运用相关分析与回归分析的方法去解决。

相关与回归的基本概念

一、变量间的相互关系

二、相关关系的类型

三、相关分析与回归分析

变量间的相互关系

◆ 确定性的函数关系 $Y=f(X)$

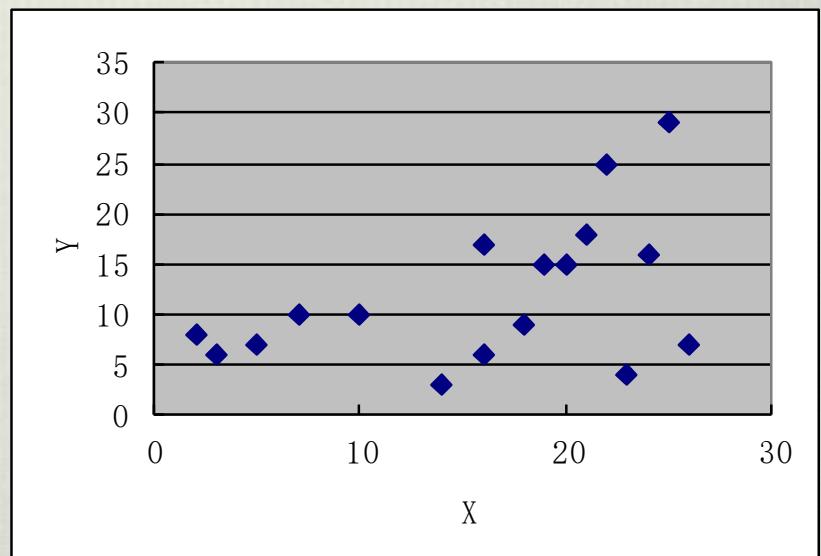
◆ 不确定性的统计关系—相关关系

$$Y = f(X) + \varepsilon \quad (\varepsilon \text{为随机变量})$$

◆ 没有关系

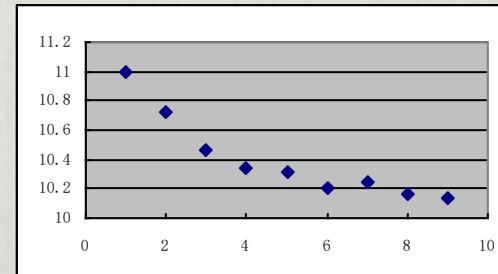
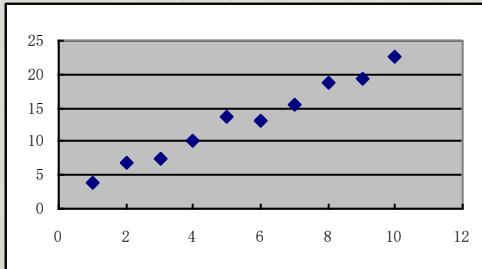
变量间关系的图形描述：

坐标图(散点图)



相关关系的类型

- 从涉及的变量数量看
 - 简单相关
 - 多重相关（复相关）
- 从变量相关关系的表现形式看
 - 线性相关——散布图接近一条直线（左图）
 - 非线性相关——散布图接近一条曲线（右图）



相关关系的类型

● 从变量相关关系变化的方向看

正相关——变量同方向变化

同增同减 (A)

负相关——变量反方向变化

一增一减 (B)

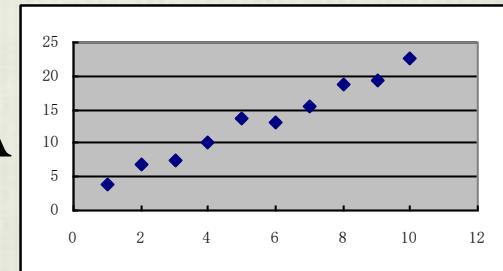
● 从变量相关的程度看

完全相关 (B)

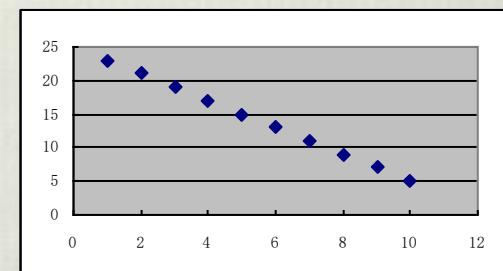
不完全相关 (A)

不相关 (C)

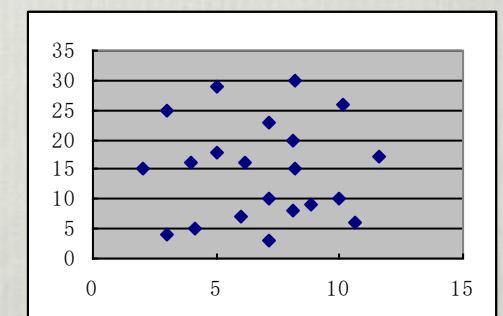
A



B



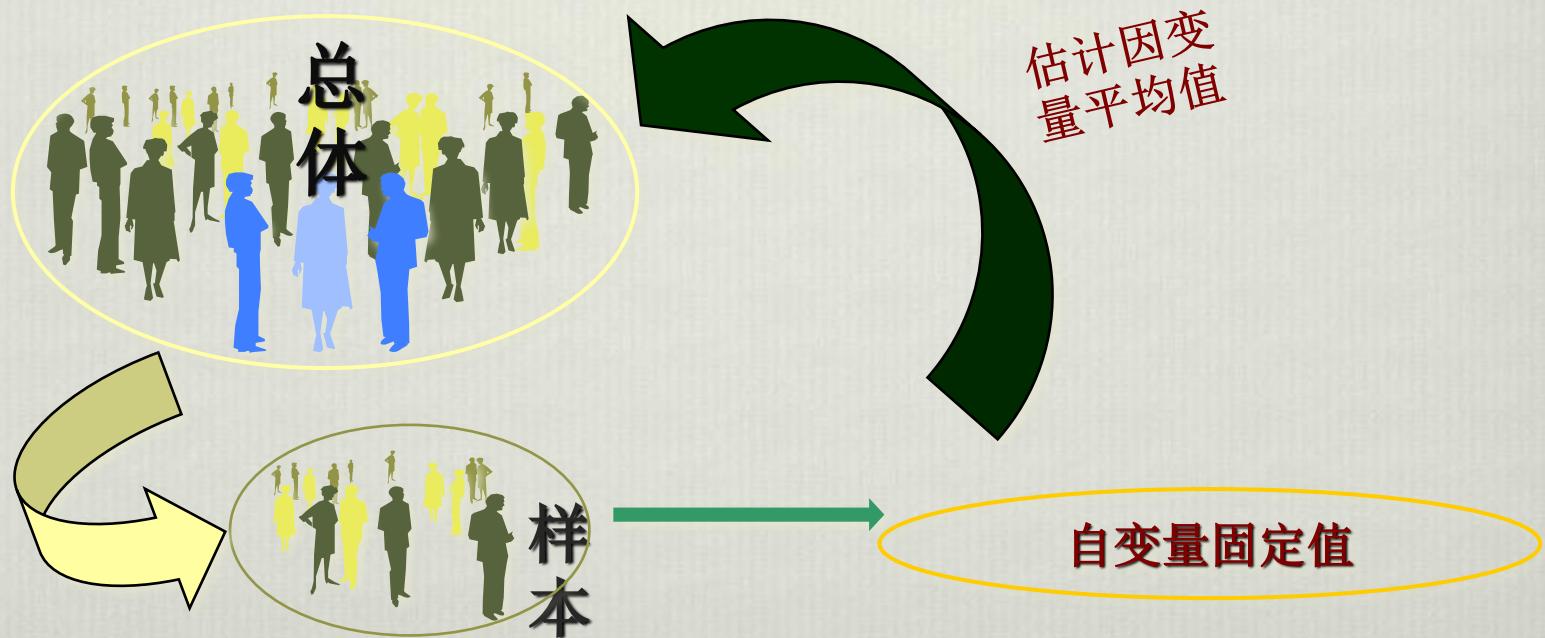
C



回归的现代意义

一个因变量对若干解释变量依存关系的研究
回归的目的（实质）：

由固定的自变量去估计因变量的平均值



相关分析与回归分析的联系

- 共同的研究对象：都是对变量间相关关系的分析
- 只有当变量间存在相关关系时，用回归分析去寻求相关的具体数学形式才有实际意义
- 相关分析只表明变量间相关关系的性质和程度，要确定变量间相关的具体数学形式依赖于回归分析
- 相关分析中相关系数的确定建立在回归分析的基础上

多元线性相关与回归分析

- ❖ 一、多元线性回归模型及假定
- ❖ 二、多元线性回归模型的估计
- ❖ 三、多元线性回归模型的检验
- ❖ 四、多元线性回归模型的预测
- ❖ 五、复相关系数和偏相关系数
- ❖ 六、残差分析

多元总体线性回归函数

一般形式

$$Y_i = b_1 + b_2 X_{2i} + b_3 X_{3i} + \cdots + b_k X_{ki} + u_i$$

多元总体线性回归模型的矩阵表示

$$\begin{matrix} Y_1 & 1 & X_{21} & X_{31} & \cdots & X_{k1} & b_1 & u_1 \\ Y_2 & 1 & X_{22} & X_{32} & \cdots & X_{k2} & b_2 & u_2 \\ \vdots & \cdots & \cdots & \cdots & \cdots & \vdots & \vdots & \vdots \\ Y_n & 1 & X_{2n} & X_{3n} & \cdots & X_{kn} & b_k & u_n \end{matrix}$$

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{U}$$

多元线性回归模型的估计

多元回归模型的假定

假定1：零均值、同方差、无自相关、随机扰动项与自变量不相关、U正态性

假定2：各自变量之间不存在线性关系，在此条件下，自变量观测值矩阵X列满秩

$$\text{Rank}(X) = k$$

方阵 $X'X$ 满秩 $\text{Rank}(X'X) = k$

意义： $X'X$ 可逆， $(X'X)^{-1}$ 存在

1. 多元线性回归的最小二乘估计式

回归系数 β 的最小二乘估计为

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{Y}$$

2. 多元线性回归模型的检验

3. 多元线性回归模型的预测

- ❖ 【例1】 1991年浙江省城乡广播电视台听众观众调查中，
- ❖ $Y=$ 平时看报的频度；
- ❖ $X_1=$ 是否订阅报纸（是为1，否为0）
- ❖ $X_2=$ 文化程度
- ❖ 利用对农村调查511人的数据，得到如下回归方程：

$$Y = 0.54 + 0.31X_1 + 0.36X_2$$

- ❖ 【例2】在1990年10月份的《亚运会广播电视台宣传效果》的调查问卷的调查结果中，以对亚运会的态度总分作为因变量Y，其他有9个相关的自变量，得到如下回归方程：
$$Y' = 24.74 + 2.12 U + 1.06 R + 0.92 S + 0.52 A3 - 0.96 T + 0.58 F + 0.28 A2 + 0.26 P + 0.30 B$$

P, R, Y: 相应问题的累加李克量表度量的变量

T: 取值0或1（根据什么评价亚运会成功与否，自己的判断是1，其他为0）

A2, A3: 性别，年龄，教育程度等

B: 体育锻炼情况

U : 1~5 (希望在我国举办奥运会吗)

F, S: 1~5 (听广播时间，评价亚运会是否成功)

非线性回归

常用的可以转换为线性的非线性函数形式
幂函数

$$Y_i = \beta_1 X_i^{\beta_2} e^{u_i}$$

参数度量了变量Y对变量X的弹性，即X的单位百分比变动引起Y变动的百分比

对数函数

$$Y_i = a + b \ln X_i + u_i$$

参数说明当变量X每变动一个百分点，引起因变量Y绝对量的变动量

非线性回归

指数函数

$$Y_i = ab^{X_i} e^{u_i}$$

转换为线性函数

$$\ln Y_i = \ln a + X_i \ln b + u_i$$

双曲函数

$$Y_i = a + b \frac{1}{X_i} + u_i$$

多项式函数

$$Y_i = b_0 + b_1 X_i + b_2 X_i^2 + \cdots + b_k X_i^k + u_i$$

共同特点：虽然对于变量而言都是非线性的，但对

于参数而言却是线性的，可以转换为线性回归去估计其参数。

典型相关分析

- ❖ 典型相关分析有助于综合地描述两组变量之间的典型的相关关系。
- ❖ 其条件是：两组变量都是连续变量，都必须服从多元正态分布。

(三) 聚类分析

- ❖ 聚类分析是将数据进行分类，同一个类中的对象有很大的相似性，不同的类间有很大的相异性。
- ❖ 聚类分析是一种比较直观的探索性的图解方法。

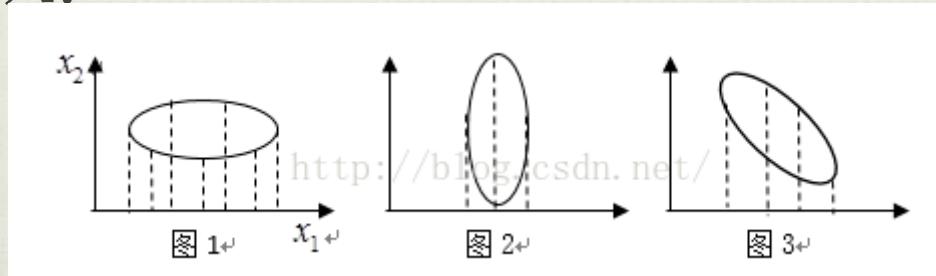
❖ 【例1】将下面5个人进行分类：

No	身高 (cm)	体重 (kg)	性别	职业
1	170	61	w	工人
2	168	60	w	干部
3	173	65	m	工人
4	175	64	m	干部
5	169	62	m	教授

- ❖ 【例2】1992年《小学生与电视》的调查中，对“你认为自己应该具有哪些品质和性格”的15个变量进行了聚类分析，它们都是0, 1变量。聚类分析的不同方法可以将其分成若干类，例如若聚为5类，则分别是：
 - ❖ 1. 勇敢，幽默，机智，坚韧，敏捷，独立，稳健
 - ❖ 2. 守纪律，有礼貌，听话
 - ❖ 3. 富有同情心
 - ❖ 4. 乐于请教人
 - ❖ 5. 认真，勤奋，虚心

(四) 主成分分析

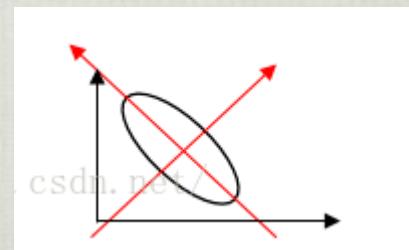
❖ 问题的引入:



- ❖ 如果将3个图的数据点投影到 x_1 轴上，图1的数据离散度最高，图3其次，图2最小。
- ❖ 数据离散性越大，代表数据在所投影的维度上具有越高的区分度，这个区分度就是信息量。
- ❖ 如果用方差来形容数据的离散性，就是数据方差越大，表示数据的区分度越高，也就是蕴含的信息量是越大的。

方法：主轴投影

目的：选择一个离散度最大的轴进行投影，在尽量保留最多信息量的情况下，进行了数据降维。



- ❖ 一种简化数据结构的方法，把多个变量简化为少数的几个综合变量，而这几个综合变量可以反映出原来多个变量的大部分信息。
- ❖ 主成分分析浓缩了众多指标的信息，降低了指标的维度，从而简化指标的结构，深刻反映问题的内在规律。

- ❖ 在大部分实际问题中，需要考察的变量多，变量之间是有一定的相关性的，主成分分析就是以损失很少部分信息为代价，保留绝大部分信息的前提下，将原来众多具有一定线性相关性的多个指标压缩成少数几个互不相关的综合指标（主成分），并通过原来变量的少数几个的线性组合来给出各个主成分的具有实际背景和意义的解释。

- ❖ 【例】评价企业优劣等级，要考虑很多指标：固定资产、流动资金、产值、利润、管理水平、技术力量、工人素质、厂区环境、节能环保等等上百个变量。如何寻找少量的几个综合变量代替原来的众多的变量？
- ❖ 主成分分析的目的：
- ❖ 一是简化数据；
- ❖ 二是揭示变量间的关系。

❖ 主成分分析的步骤

- ❖ 1. 将原始的 p 个变量进行标准化处理；
- ❖ 2. 计算标准化指标的相关系数矩阵
- ❖ 3. 求解相关系数矩阵的特征向量和特征值；
- ❖ 4. 计算各个主成分的方差贡献率及累积贡献率；
- ❖ 5. 确定主成分的个数；
- ❖ 通常根据实际问题的需要由累计贡献率大于85%的前k个成分来代替原来 p 个变量的信息，或选取所有特征值大于1的成分作为主成分，也可根据特征值的变化来确定，即根据统计软件输出的碎石图的转折点来决定选取主成分的个数。

- ❖ 6. 对确定出的主成分作出实际意义的解释；
- ❖ 7. 利用所确定出的主成分的方差贡献率计算综合评价值，从而对被评价对象进行排名和比较。
- ❖ 综合得分= \hat{a} (各主成分得分×各主成分方差贡献率)

- ❖ 主成分分析的数学模型
- ❖ 设对某一事物的研究涉及指标(变量): X_1, X_2, \dots, X_p
- ❖ 对 X_1, X_2, \dots, X_p 进行线性变换, 形成新的综合变量 Y:
 - ❖
$$\begin{array}{l} Y_1 = u_{11}X_1 + u_{12}X_2 + \cdots + u_{1p}X_p \\ Y_2 = u_{21}X_1 + u_{22}X_2 + \cdots + u_{2p}X_p \\ \vdots \dots \dots \\ Y_p = u_{p1}X_1 + u_{p2}X_2 + \cdots + u_{pp}X_p \end{array}$$
 - ❖ 希望 Y_i 的方差尽可能大且新的综合变量 Y_i 之间相互独立。

- ❖ 在实际应用时，通常挑选前几个方差比较大的主成分，虽然这样做会丢失一部分信息，但它使我们抓住了主要矛盾进行深入分析，并从原始数据中进一步提出了某些新的信息，因而在某些实际问题的研究中得益比较大，这种既减少了变量的个数又抓住了主要矛盾的做法有利于问题的分析和处理。

- ❖ **注意：**

主成分分析对单位变化敏感

(如果把一个变量乘以一个标量，会得到不同的特征值和特征向量，因为特征值分解是基于协方差矩阵而不是相关系数矩阵。)

解决方法：

1. 主成分应该被应用到具有相似单位的数据上
2. 变量标准化后再做分析，标准化主成分 (NPC) , $\widetilde{X_i} = \frac{X_i - E(X_i)}{\sqrt{D(X_i)}}$

(五) 因子分析

- ❖ 因子分析是主成分分析的推广，因子分析的主要目的是用几个因子（不可观测的隐性变量）来描述可以观测的显在变量。
- ❖ 比如说学生的学习成绩可以表示成学生的态度、认识、爱好、能力、智力等不可直接观测的潜在变量的线性组合。
- ❖ 因子分析是根据相关性大小把原始变量进行分组，使得同组内的变量之间相关性高，而不同组的变量之间的相关性低。每组变量代表一个基本结构（即**公共因子**），并用一个不可观测的综合变量来表示。
- ❖ 对于所研究的某一具体问题，原始变量分解为两部分之和。一部分是少数几个**不可观测的公共因子的线性函数**，另一部分是不能被公共因子包括的**特殊因子**。

因子分析的数学模型

❖ 设有 p 个指标，则因子分析数学模型为：

$$\begin{aligned} X_1 &= m_1 + r_{11}Y_1 + r_{12}Y_2 + \cdots + r_{1k}Y_k + u_1 \\ X_2 &= m_2 + r_{21}Y_1 + r_{22}Y_2 + \cdots + r_{2k}Y_k + u_2 \\ &\vdots \dots \dots \\ X_p &= m_p + r_{p1}Y_1 + r_{p2}Y_2 + \cdots + r_{pk}Y_k + u_p \end{aligned}$$

X_1, X_2, \dots, X_p 是已标准化的可观测的评价指标。

Y_1, Y_2, \dots, Y_k 是不可观测的公因子，其含义要根据具体问题来解释。

u_i 特殊因子，它与公共因子相互独立，服从正态分布

m_i 总平均 r_{ij} 因子负荷，反映了因子对指标的贡献程度

- ❖ 因子分析的首要目的是估计因子负荷 r_{ij} 和特殊因子 u_i 的方差，然后给公因子 Y_1, Y_2, \dots, Y_k 一个合理的解释，若难以找到合理的解释，需进一步作因子旋转，以求旋转后能得到合理的解释。

因子分析模型有两个特点：

1. 模型不受量纲影响；
2. 模型中因子负荷是不唯一的，通过因子旋转可以使新的因子具有更鲜明的实际意义。

- ❖ 【例1】 调查青年对婚姻的态度，抽取250名青年，问了50个问题，包括以下几个方面：对文化和职业的要求，对经济收入的态度，对老人的责任，对相貌的重视，对孩子的观点等。每个方面就是一个因子。
- ❖ 【例2】 公司老板与48名申请工作的人进行面谈，然后就申请人15个方面进行打分，15个变量包括：申请信的形式、外貌、专业能力、讨人喜欢的能力、自信心、洞察力、诚实、推销本领、经验、驾驶汽车的本领、志向、领会能力、潜在能力、对工作要求强烈程度、工作是否合适。许多变量高度相关，通过因子分析，找到公共因子，减少变量个数，从而对申请者进行正确判断。

❖ 因子分析的目的是研究一种假设的结构，用 k 个假设的公共因子来说明 p 个变量之间的相互依赖结构。因子分析带有很大的主观性，不但需要客观的复杂的计算，还需要经过主观思维的加工补充，是一种探索性研究，与以往的研究相结合，则有可能确认或确立某种理论模式。

因子分析与回归分析的区别

	回归模型	因子模型
要估计的主要参数	回归系数， 随机误差的方差	因子负荷系数 特殊因子的方差
自变量的性质	自变量是可观测的 显在变量	自变量是不可观测 的潜在变量
自变量的个数	已知	未知，需要估计
自变量之间的关系	可能是相关的	相互独立

谢谢！