02

经典统计学基础

描述性统计
度量指标总述

描述性统计
（Descriptive statistical analysis）
指运用制表和分类，图形以及计算概括性数据来描述数据特征的各项活动。

数据集中趋势度量

- 均值（**算数**、加权算数、几何、调和）
- **中位数**（分位数）
- **众数**

数据离散趋势度量

- **极差**（四分距、百分位距）
- 平均偏差（**标准差、方差**）
- **变异系数**

数据分布形态度量

- **偏度**
- **峰度**

补充：*中心矩（Central Moment）* *是关于某一随机变量平均值构成随机变量的概率分布的矩。*
1阶中心矩=期望 | 2阶中心矩=方差 | 3阶中心矩=偏度 | 4阶中心矩=峰度

描述性统计
集中趋势-均值

均值是描述数据集的最有用的单个量，是集中趋势的最主要测度值。但是但是它并非总是度量数据中心的最佳方法。均值对极端值很敏感。

01.算数平均

N个数$x_1, x_2, \ldots, x_N$的算术平均，简称均值（Mean），用$\bar{x}$表示:

$$\overline{X} = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1 + x_2 + \ldots + x_N}{N}$$

02.加权算数平均

有时，我们需要在$x_1, x_2, \ldots, x_N$上加某些加权因子（或权）$w_1, w_2, \ldots, w_N$来反映数字的重要性。此时$\bar{x}$称做加权算术平均:

$$\bar{x} = \frac{\sum_{i=1}^{N} w_i x_i}{\sum_{i=1}^{N} w_i} = \frac{w_1 x_1 + w_2 x_2 + \ldots + w_N x_N}{w_1 + w_2 + \ldots + w_N}$$

03.几何平均

N个正数$x_1, x_2, \ldots, x_N$的几何平均G等于这些数乘积的N次方根，计算公式为:

$$G = \sqrt[N]{\prod_{i=1}^{N} x_i} = \sqrt[N]{x_1 \bullet x_2 \cdots x_N}$$

04.调和平均

N个数$x_1, x_2, \ldots, x_N$的调和平均H等于这些数的乘倒数的算数平方方根，计算公式为:

$$H = \frac{N}{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_N}} = \frac{N}{\sum_{i=1}^{N} \frac{1}{x_i}}$$

描述性统计

均值适用性与缺陷

几何平均

受极端值的影响较算术平均数小，它仅适用于具有等比或近似等比关系的数据

调和平均

1. 调和平均数易受极端值的影响，且受极小值的影响比受极大值的影响更大；
2. 只要有一个标志值为0，就不能计算调和平均数；
3. 当组距数列有开口组时，其组中值即使按相邻组距计算，假定性也很大，这时的调和平均数的代表性很不可靠。调和平均数应用的范围较小。

描述性统计

H、G、$\overline{X}$ 的关系

一组正数$x_1, x_2, \ldots, x_N$的几何平均G小于等于它们的算数平均$\overline{X}$，但大于等于它们的调和平均G，用符号表示即为： $\boldsymbol{H \leq G \leq \overline{X}}$。当所有的数都相等时，等号成立。

Q&A

例2.1 假设我们有12个区域的物种数如下（单位：种），按递增次序显示：750，800,860,1000,1100,1100,1250,1300,1300,1360,1540,1620。求12个区域的平均物种数。

使用算术平均计算公式，我们有

$$\overline{X}=\frac{750+800+860+1000+1100+1100+1250+1300+1300+1360+1540+1620}{12}=1165$$

因此，12 个区域的平均物种数为 1165 种。

例2.2 假定某地储蓄年利率（按复利计算）：5%持续2年，3%持续1.5年，2.2%持续1.5年，求此5年内该地平均储蓄年利率。

解：由几何平均公式得到该地平均储蓄年利率：

$$G=\sqrt[2+1.5+1.5]{1.05^{2}\times1.03^{1.5}\times1.022^{1.5}}-1=3.54\%$$

例2.3 5名学生分别在一个小时内解题3、4、6、8、9，问平均解题速度是多少?

解：由调和平均公式得到 5 名学生的平均解题速度：

$$H=5/\left(1/3+1/4+1/6+1/8+1/9\right)\approx4.06$$

描述性统计
中位数与分位数

01 中位数

- 一组数按照数量大小排列，如果中间的数或两个中间数的算术平均把这组数分成了2个相等的部分，那么这样的数称为***中位数（Median）***。

02 分位数

- 同样的如果我们将那些把一组数分成4个相等部分的数用Q_1、Q_2、Q_3表示，分别称为第一个、第二个、第三个***四分位数***，其中Q_2等于中位数；
- 而把一组数分为10个相等部分的数称为***十分位数***，并且用D_1、D_2、…、D_9表示；
- 把一组数分为100个相等部分的数称为***百分位数***，用P_1、P_2、…、P_{99}表示。
- 四分位数、十分位数、百分位数及其他这些通过等分数据而得到的数统称为***分位数***。

Q&A

例2.4 12个区域的物种数如下(单位：种)，按递增次序显示：750，800,860,1000,1100,1100,1250,1300,1300,1360,1540,1620。求12个区域物种数的中位数。

该数据已经按递增序排序。有偶数个观测（即 12 个观测），因此中位数不唯一。它可以是最中间两个值 1100 和 1250（即列表中的第 6 个和第 7 个值）中的任意数。根据约定，我们指定这两个最中间的值的平均值为中位数。即 $\frac{1100+1250}{2}=\frac{2350}{2}=1175$。于是，中位数为 1175（种）。

描述性统计

众数 ***不一定 存在***，即使存在也 ***不一定唯一*** 。一般地，只有一个众数的分布称为 ***单峰*** 的（Unimodal）；具有两个或更多众数的数据集合是 ***多峰*** 的（Multimodal），具有两个、三个众数的数据集合分别称为 ***双峰*** 的（Bimodal）和 ***三峰*** 的（Trimodal）；在另一种极端情况下，如果每个数据值仅出现一次，则没有众数。

众数

Mode

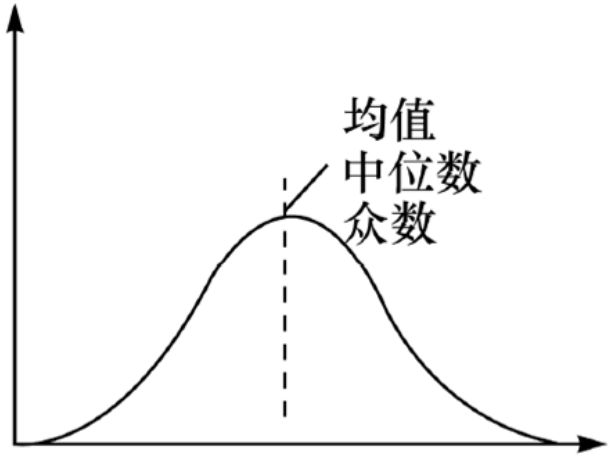
一组数的众数（Mode）是集合中出现次数最多的那个数。

描述性统计

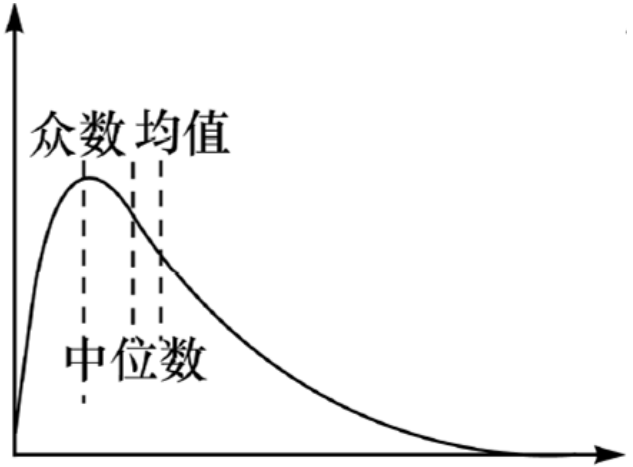
对于适度倾斜（非对称）的单峰频数曲线，我们有以下经验关系:

$$均值 - 众数 \approx 3 \times (均值 - 中位数)$$

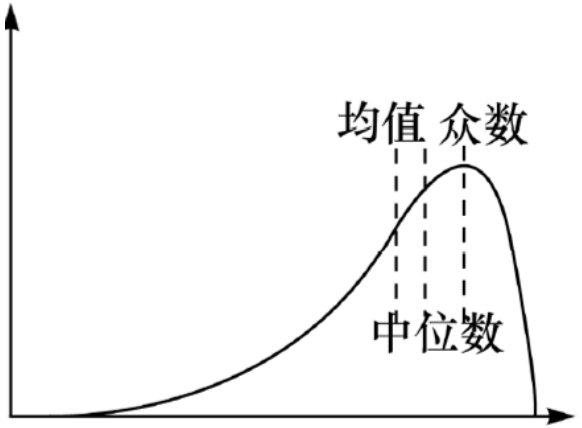
这意味：如果均值和中位数已知，则适度倾斜的单峰频率曲线的众数容易近似计算。



对称分布：均值=中位数=众数



右倾分布：均值 > 中位数 > 众数



左倾分布：众数 > 中位数 > 均值

描述性统计

中列数

中列数（Midrange）也可以用来评估数值数据的中心趋势。中列数是数据集最大值与最小值的平均值。

描述性统计
离散趋势度量

极差

一组数的极差（Range）或全距是这组数中最大的数（Max）与最小的数（Min）的差。

半内四分位距

一组数的半内四分位数（Q）间距或半内四分距用为第三个四分位数(Q3)与第一个(Q1)四分位数之差的一半。

平均偏差

一组数的平均偏差(MD)是各数与算数平均之差的绝对值之和的平均

10~90百分位距

一组数的10~90百分位 距(P)定义为第90个(P90)与第10个(P10)百分位数之差。

Q&A

例2.7&2.8 假设我们有12个区域的物种数如下(单位：种)，按递增次序显示：750，800,860,1000,1100,1100,1250,1300,1300,1360,1540,1620。求12个区域的物种数的极差与半内四分位距。

上述数据包含 12 个观测值，已经按递增序排序。极差为 1620-750=870。有时，极差（全距）也可简单的用最大的数与最小的数来表示。此例中全距可表示为 750 到 1620，或 750~1620。另一方面，该数据集的四分位数分别是该有序表的第 3、第 6 和第 9 个值。因此，Q_1=860 种，而 Q_3=1300 种。于是，四分位数极差为 Q=(1300-860)/2=220 种。

描述性统计
离散趋势度量

方差和标准差

标准差

标准差（Standard deviation）是各变量值与其平均数离差平方平均数的算数平方根，用s表示。

$$s=\sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{x})^2}$$

分母修正

在定义一个样本数据的标准差时，分母常用N-1代替N，这样产生的值是总体标准差的较好估计。当N较大时，通常当N>30时，这两种定义的区别不大。同样，我们可以根据以上定义得到的标准差乘以$\sqrt{\mathrm{N}/(N-1)}$得到这个较好的估计值。

方差

方差（Variance）是各变量值与其平均数离差平方的平均数，用s^2表示。

$$s^2=\frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{x})^2$$

描述性统计
离散趋势度量

相对离差

- 从标准差或其他离差得到的真实离差称为**绝对离差**；
- **相对离差**定义为绝对离差与均值的比值；
- 10英寸的离差在测量距离分别1000英尺时和20英寸时，产生的影响程度是有很大区别的。这种影响的程度可用相对离差来度量。

变异系数

- 绝对离差是标准差s，平均值是算数平均$\overline{X}$时的相对离差称为变异系数，用CV(COfficient of Viriation)表示。
- CV没有量纲，通常表示为百分数，可以用于不同单位的分布的客观比较。但均值接近零时，CV无效。

$$CV=\frac{s}{\overline{X}}$$

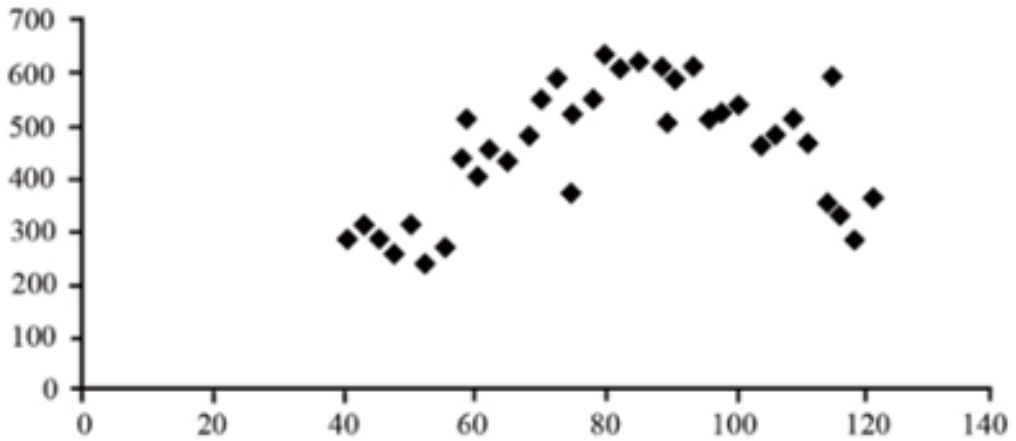
描述性统计
常用统计图

散点图

散点图（Scatter plot）是确定两个数值变量之间是否存在关联的最有效的图形方法之一。常用于识别离群点及判断相关关系。

作图方法

为构造散点图，每个值对视为一个**代数坐标对**，并作为一个点画在平面上。



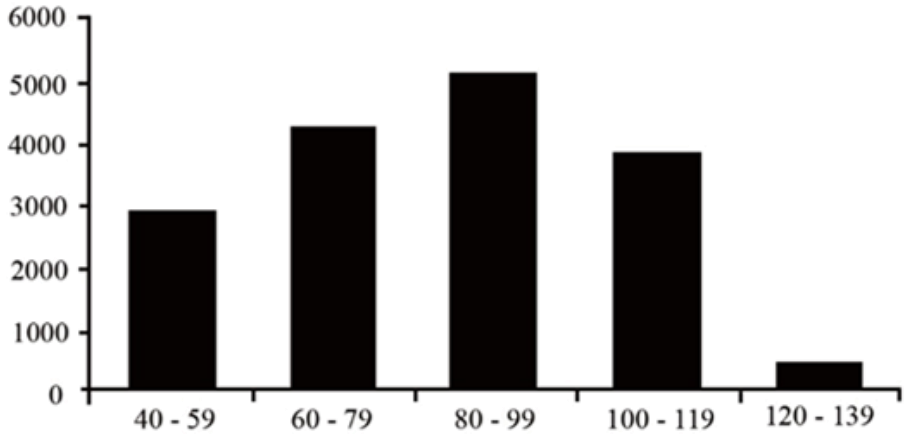
描述性统计
常用统计图

直方图

直方图是一种概括给定属性X分布的图形方法，常用于判断数据分布模式，初步确定数据是否符合正态分布。

作图方法

- 如果X是**分类变量**，则对于X的每个已知值，画一个柱或竖直条。条的高度表示该X值出现的频率（即计数）。结果图更多地称做条形图（Bar chart）。
- 如果X是**数值变量**，则更多使用术语直方图。X的值域被划分成不相交的连续子域。子域称做桶（Bucket）或箱（Bin），是X的数据分布的不相交子集。桶的范围称做宽度。通常，诸桶是等宽的。



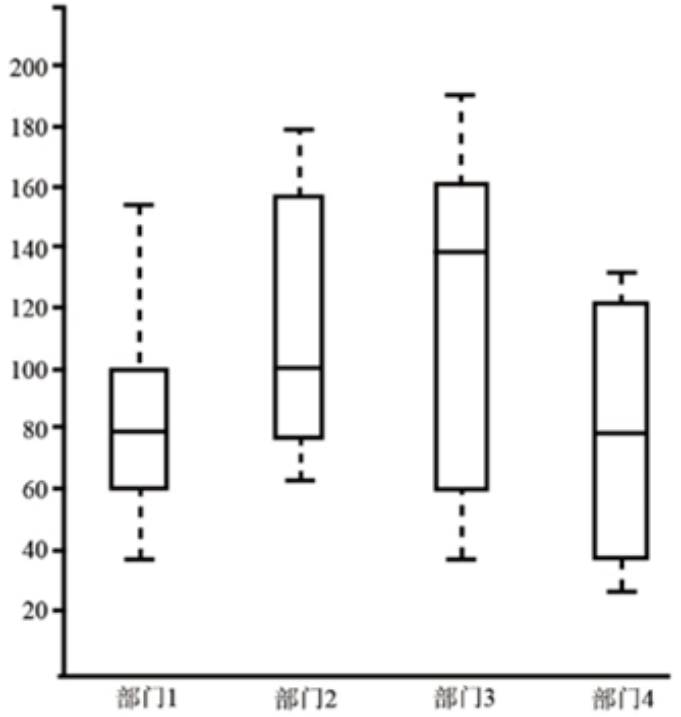
描述性统计
常用统计图

盒图

盒图（Boxplot）是一种流行的分布的直观表示。体现了五数概括：中位数、上四分位数、下四分位数、最小值、最大取。

作图方法

盒的端点一般在**四分位数**上，使得盒的长度是**四分位数极差IQR**。
(1)**中位数**用盒内的线标记。
(2)盒外的两条线（称做胡须）延伸到**最小**（Minimum）和**最大**（Maximum）观测值。



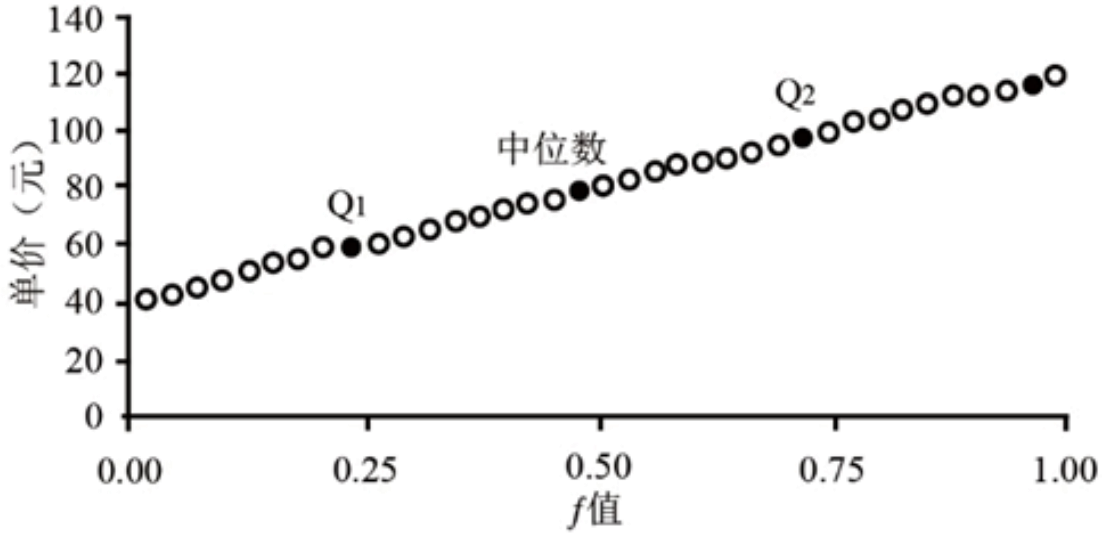
描述性统计
常用统计图

分位数图

分位数图（Quantile plot）是一种观察单变量数据分布的简单有效方法。首先，它显示给定属性的所有数据。其次，它绘出了分位数信息。

作图方法

在分位数图中，**x_i对应于百分位数f_i画出。**这使得我们可以基于分位数比较不同的分布。例如，给定两个不同时间段的销售数据的分位数图，我们一眼就可以比较它们的Q1、中位数、Q3以及其他 值。



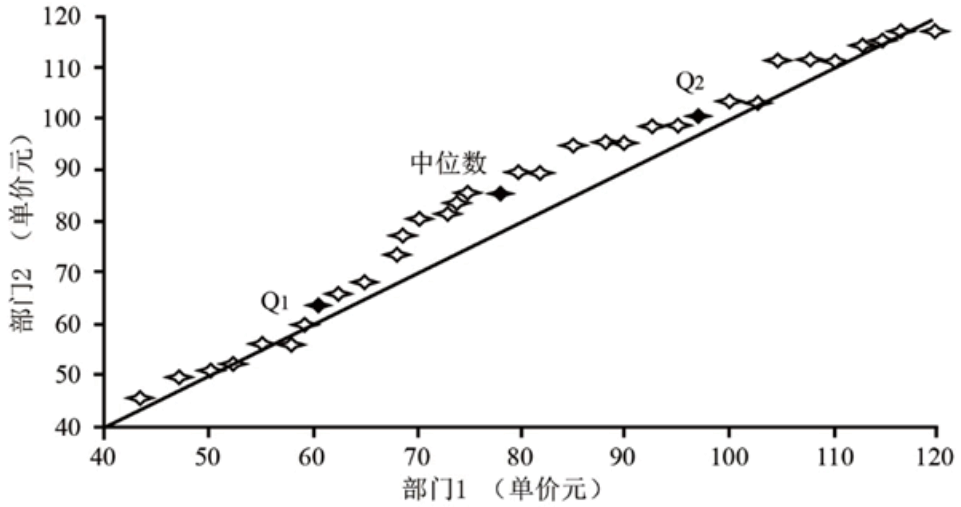
描述性统计
常用统计图

分位数-分位数图

分位数-分位数图（Quantile-quantile plot）或q-q图对着另一个对应的分位数，绘制一个单变量分布的分位数，使得用户可以观察从一个分布到另一个分布是否有漂移。

作图方法

假定对于属性或变量unitprice（单价），我们有两个观测集，取自两个不同的部门。设$x_1,\ldots,x_N$是取自第一个部门的数据，$y_1,\ldots,y_M$是取自第二个部门的数据，其中每组数据都已按递增序排序。如果**M=N**（即每个集合中的点数相等），则我们简单地**对着x_i画y_i**，其中x_i和y_i都是它们的对应数据集的第(i-0.5)/N个分位数。如果**M<N**(即第二个部门的观测值比第一个少)，则可能只有M个点在q-q图中。这里，**y_i是y数据的第(i-0.5)/M个分位数，对着x 数据的第(i-0.5)/M个分位数画。**在典型情况下，该计算涉及插值。



正态分布

时间简史

棣莫弗(1730~1733)

利用二项分布逼近的方式，得到正态密度函数的形式。

伯努利(1705)

提出伯努利大数定律。

伽利略

正态分布起源于误差分析。伽利略第一个提出随机误差。

高斯(1809)

引入正态误差理论，补充了最小二乘法，首次导出正态分布。

勒让德(1805)

发明最小二乘法。

拉普拉斯(1780)

建立中心极限定理的一般形式。

19世纪末

数据与正态拟合欠佳的情况日渐为人们重视。

拉普拉斯(1811)

利用中心极限定理论证正态分布。

海根(1837)

提出元误差学说，逐步确立误差服从正态分布。

20世纪30年代

最终完成独立和中心极限定理的一般形式，发现当样本量趋近于无穷时，一系列重要统计量的极限分布都构成正态形式，这构成了大样本方法的基础。

20世纪初

学者们提出了统计学三大分布：卡方分布、t分布、F分布，掀起了小样本理论革命。

20世纪初

依赖于正态分布的传统方法开始招致质疑。

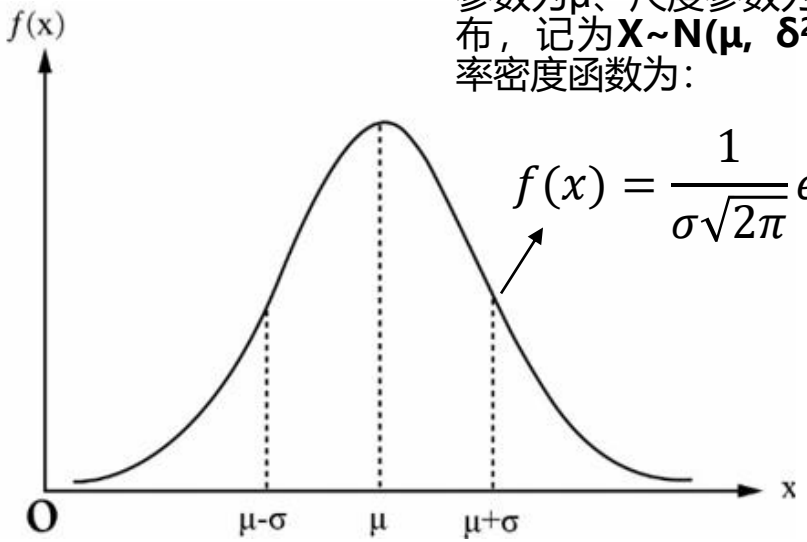
推荐阅读：《正态分布的前世今生》 https://cosx.org/2013/01/story-of-normal-distribution-1/

正态分布

正态分布又叫高斯分布(Gaussian distribution)，是一个在数学、物理及工程等领域都非常重要的概率分布。身高、寿命、成绩、测量误差等都符合正态分布。

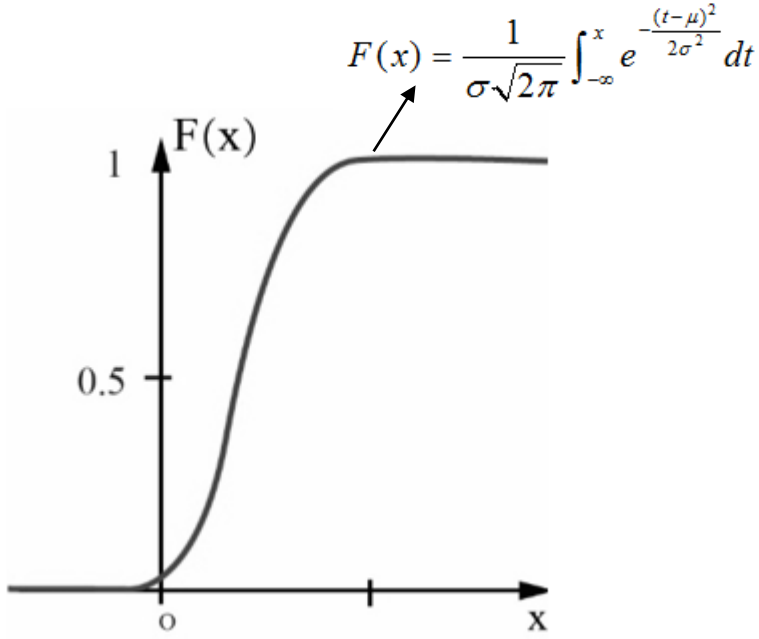
若随机变量X服从一个位置参数为μ、尺度参数为δ的正态分布，记为**X~N(μ，δ²)**，则其概率密度函数为：

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



正态分布的概率密度曲线

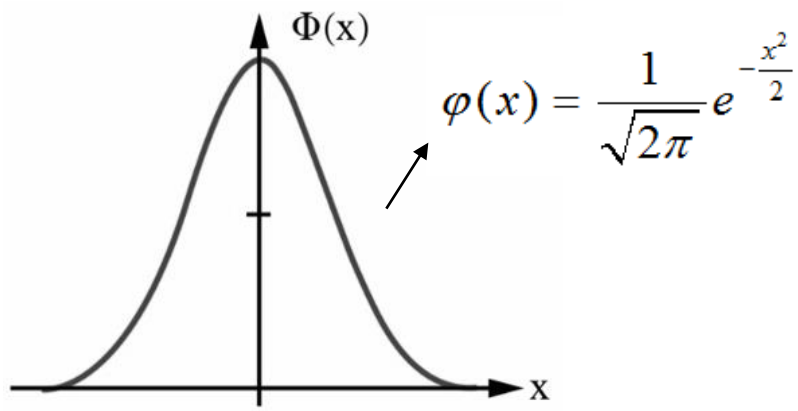
$$F(x)=\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{x}e^{-\frac{(t-\mu)^2}{2\sigma^2}}dt$$



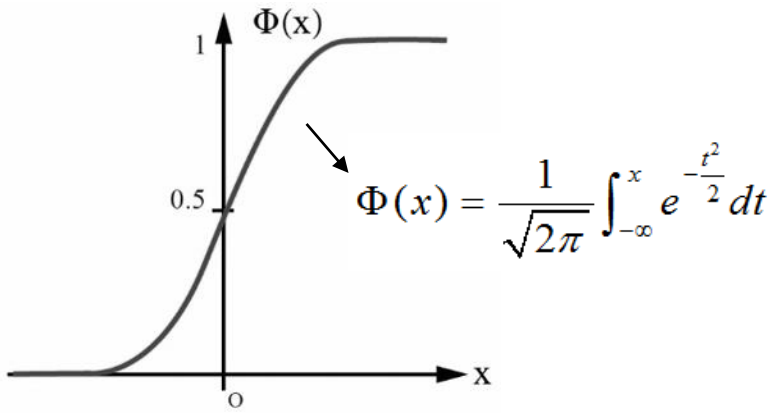
正态分布的分布函数

正态分布

μ=0，σ=1的正态分布称为标准正态分布。实践中常将一般正态分布化为标准正态分布。



标准正态分布的概率密度曲线



标准正态分布的分布函数

正态分布函数密度曲线特征

- 正态分布由μ和δ决定位置和形状
- 正态分布曲线以均值为中心，左右对称
- 正态分布分曲线以X轴为渐进线
- 正态分布曲线的拐点为x= μ± δ
- 3δ准则（0.6826/0.9544/0.9974）

正态分布

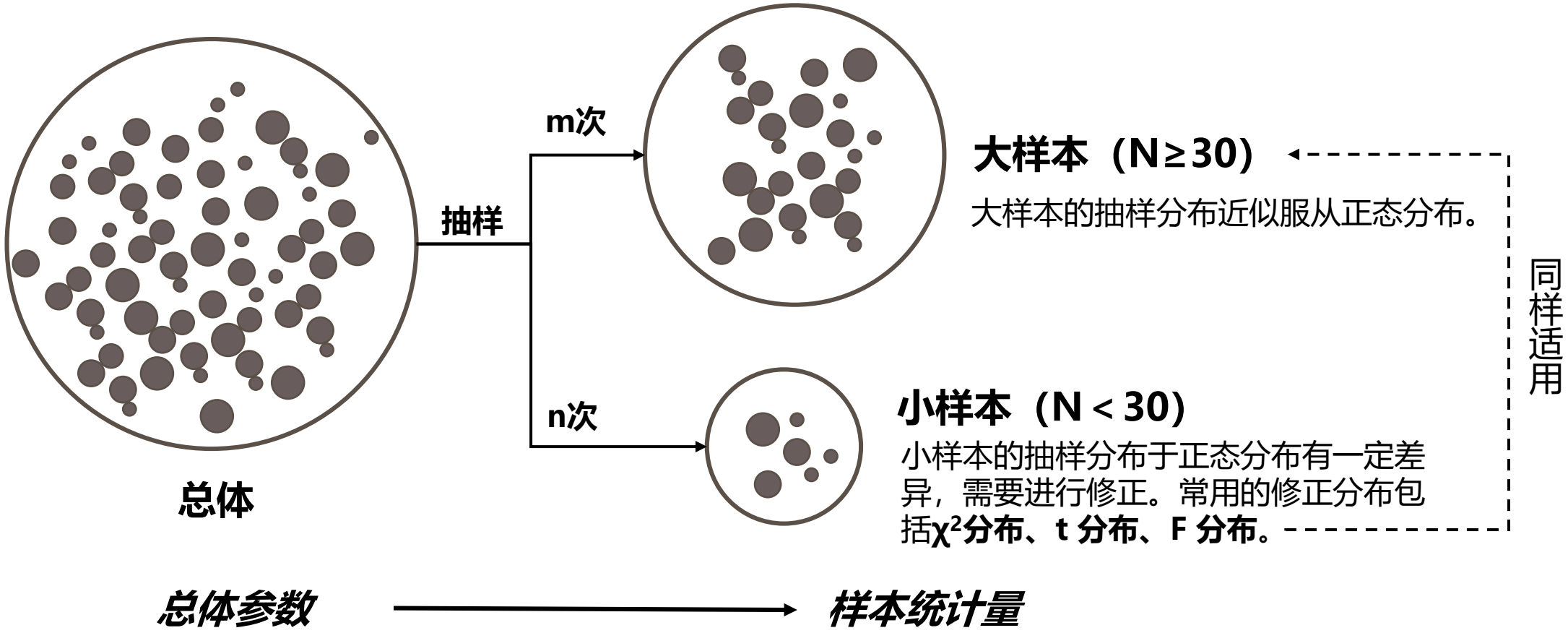
精确抽样理论

基础概念

- 多次抽样获得不同样本，每个样本都有一个统计量。多次抽样获得的统计量服从的分布即为**抽样分布**。
- 对小样本统计量的抽样分布研究称为**小样本理论**，然而因所得结论不仅适用小样本，也同样适用于大样本问题，故也称之为**精确抽样理论**。
- 精确抽样理论中常用的三种分布为：**χ2分布、t 分布、F 分布。**

正态分布

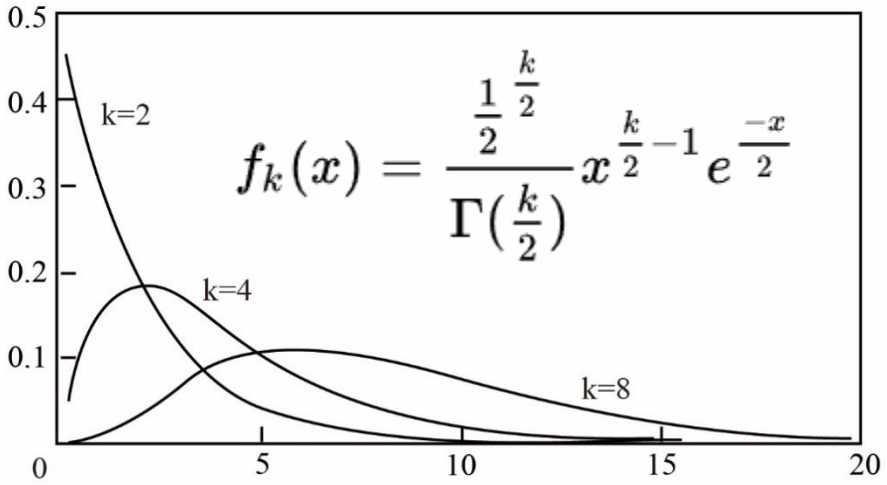
精确抽样理论
概念图示



总体参数 → ***样本统计量***

精确抽样理论
卡方分布

若随机变量U1与U2服从自由度为d1、d2的卡方分布，则两个随机变量除以各自自由度的比值服从F分布。即$(U_1/d_1)/(U_2/d_2)$服从**F分布。**



应用 皮尔森卡方检验

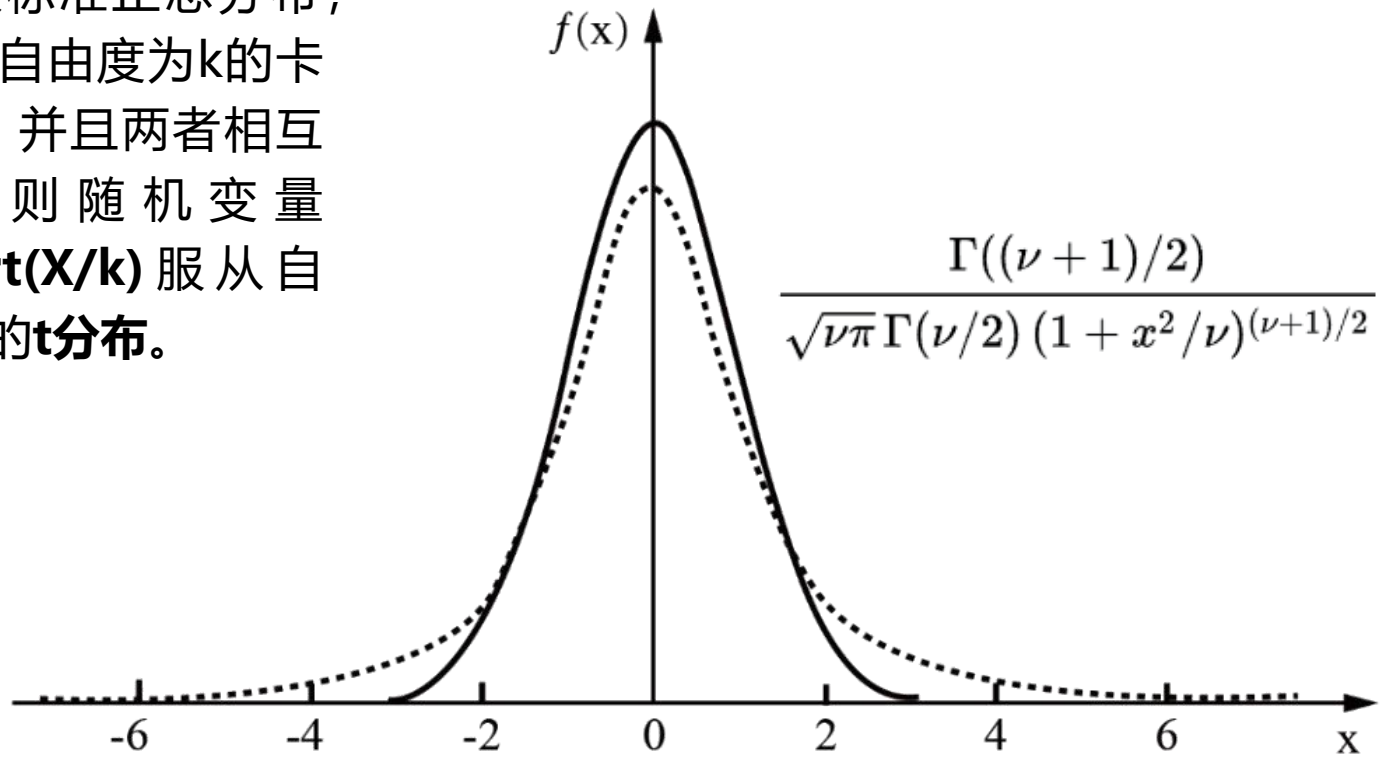
- 样本某性质的比例分布与总体理论分布的拟合优度（例如某行政机关男女比是否符合该机关所在城镇的男女比）；
- 同一总体的两个随机变量是否独立（例如人的身高与交通违规的关联性）；
- 二或多个总体同一属性的同素性检定（意大利面店和寿司店的营业额有没有差距）。

正态分布

精确抽样理论

t分布

若Z服从标准正态分布，而X服从自由度为k的卡方分布，并且两者相互独立，则随机变量**t=Z/sqrt(X/k)**服从自由度为k的**t分布**。



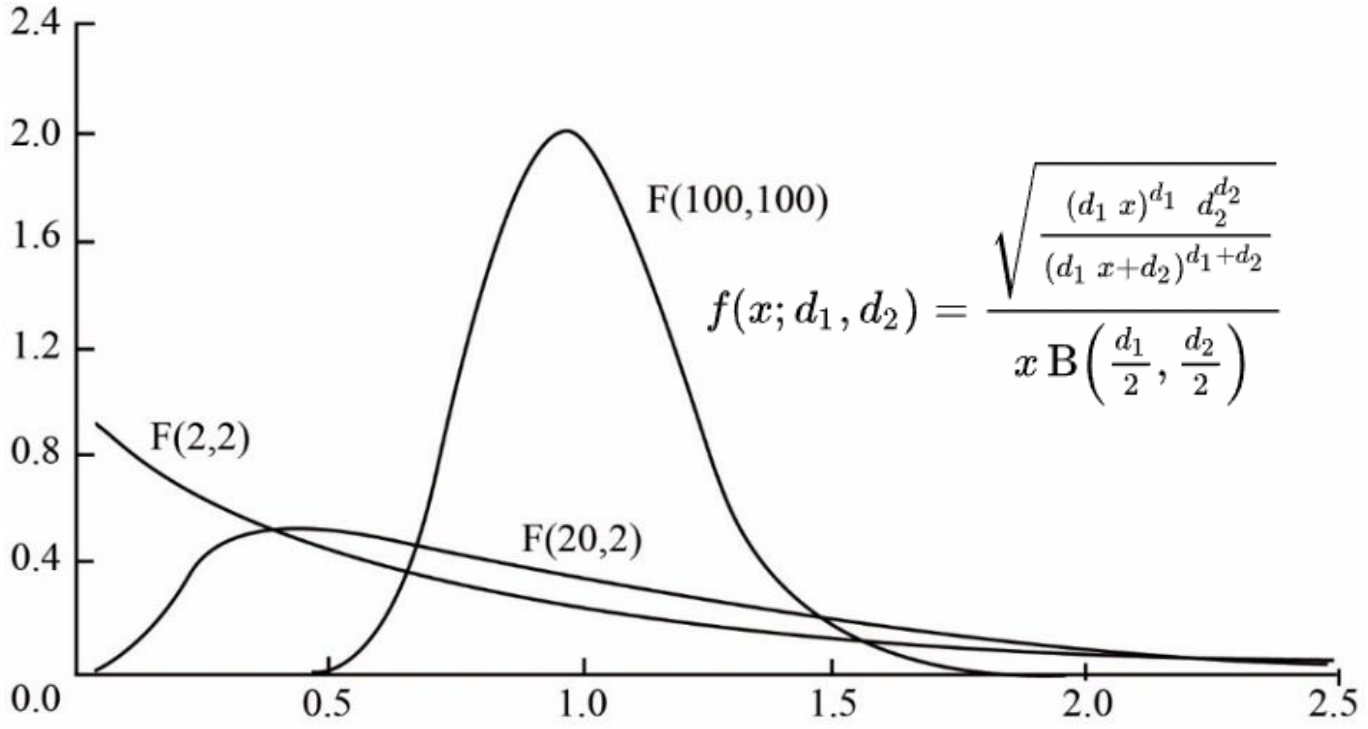
$$\frac{\Gamma((\nu+1)/2)}{\sqrt{\nu\pi}\,\Gamma(\nu/2)\,(1+x^2/\nu)^{(\nu+1)/2}}$$

应用 **t检验**

- 根据小样本来估计呈正态分布且方差未知的总体的均值

精确抽样理论
F分布

k个独立的标准正态分布变量的平方和服从自由度为k的卡方分布：**$X \sim \chi^2(k)$**



***应用* F检验**

- 比较两正态总体的方差
- 方差分析(ANOVA)

参数估计

描述性统计

精确抽样理论

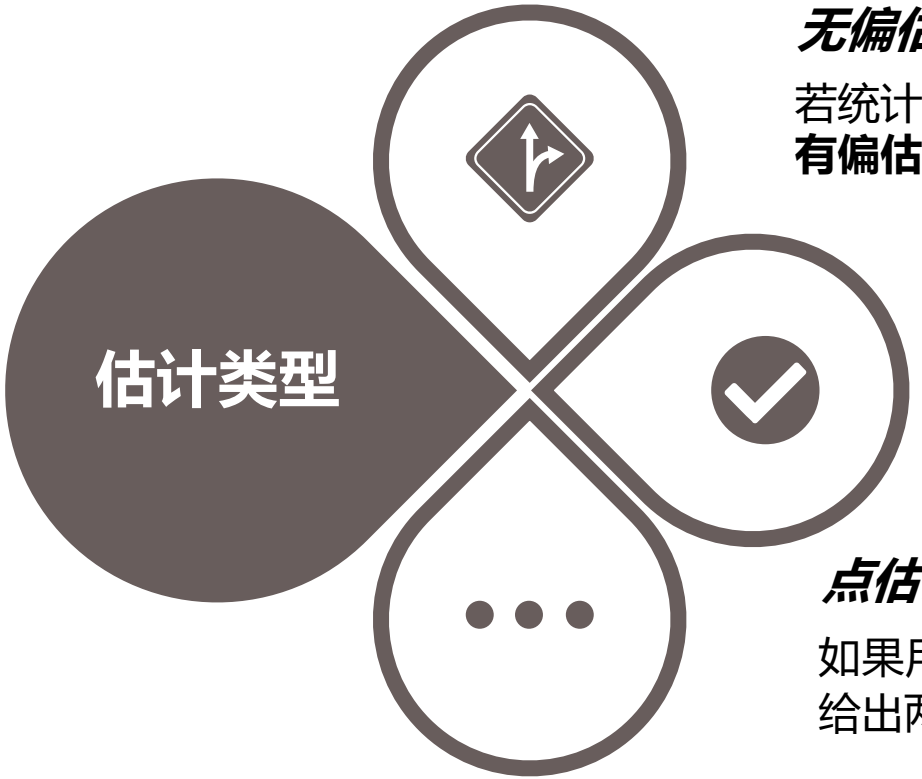
统计推断

参数估计

假设检验

参数估计
基础概念

参数估计是从总体中抽取的随机样本来估计总体分布中未知参数的过程。



无偏估计与有偏估计

若统计量抽样分布的均值与总体参数一致，则该统计量为**无偏估计**，否则为**有偏估计。**

有效估计与无效估计

若两个统计量的抽样分布有相同的均值，那么方差较小的那个统计量是此均值的**有效估计**，另一个是**无效估计。**

点估计与区间估计

如果用一个数来估计总体的参数，那么这种估计叫作参数的**点估计**，如果给出两个数，指出参数位于其间，那么这种估计叫作参数的**区间估计。**

在参数估计中，用来估计总体参数的统计量称为 **估计量**，*如样本均值、样本方差。根据一个具体样本计算出来的估计的量的数值称为* **估计值** 。

参数估计

置信区间

- 在区间估计中，由样本统计量所构造的总体参数的估计区间称为**置信区间。**
- 将构造置信区间的步骤重复多次，置信区间中包含总体参数真值的次数所占的比例称为**置信水平**也称为**置信度**或**置信系数。**

常用置信水平的$Z_{\alpha/2}$值

置信水平	α	$\alpha/2$	$Z_{\alpha/2}$
90%	0.10	0.05	1.645
95%	0.05	0.025	1.96
99%	0.01	0.005	2.58

假设检验

统计决策与假设

- 在实际问题中，常常需要根据样本的信息对总体情况作出决策，这些决策称为统计决策。
- 要做出某些决策，常常要对总体先做出某些假定或猜测，这些假定可能正确也可能不正确，统称为统计假设

需要进行决策 —*提出*→ **统计假设**

对总体做出某些假定或猜测统称为统计假设，可能正确也可能错误。

统计假设 —*分为*→

原建设（零假设H_0）

多数情况下，我们给出一个统计假设是为了拒绝它，这种假设称之为原假设。如判断某方法是否优于其他方法时，先假设差异不存在。

备择假设（研究假设H_1）

任何不同于零假设的假设都称为备择假设，即这种假设是供拒绝原建设后选择的一种假设。

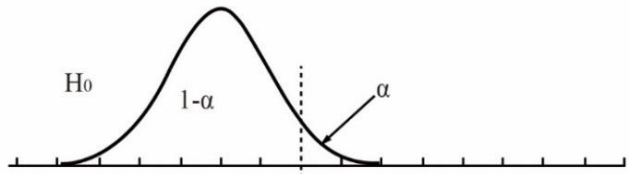
假设检验
两类错误

先对总体参数μ提出一个**假设**，然后利用样本信息检验假设**是否成立**的过程称为**假设检验。**

决策 —— *错误*

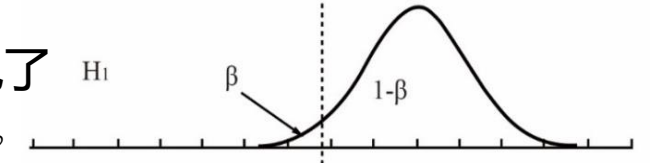
Ⅰ类错误

当我们拒绝一个本该接受的假设时，就犯了Ⅰ类错误，又称为弃真错误，其概率记为α。



Ⅱ类错误

当我们接受了一个本该拒绝的假设时，就犯了Ⅱ类错误，又称为存伪错误，其概率记为β。



开发某新药，需要判断该药是否有效

原假设(H_0): 新药无效

研究假设(H_1): 新药有效

决策 \ 事实	新药无效	新药有效
接受 H_0	正确判断(不使用无效药)	取伪错误(不使用有效药)
拒绝 H_1	弃真错误(使用无效药)	正确判断(使用有效药)

假设检验
显著性水平

对于任何给定的样本容量，在减少一种类型错误的同时往往会使另一种类型的错误增加，同时减少两种类型错误的唯一方法是增加样本容量。在实际问题中，第一类错误的影响可能比第二类错误影响更严重，这时往往需要作出一些妥协来限制更为严重的那类错误。愿意犯Ⅰ类错误的最大概率，称为**显著性水平**，一般记α。

真实情况	样本假设检验的结论	
	拒绝 H_0	不拒绝 H_0
H_0 正确	Ⅰ类错误	推断正确
	犯错误的概率为 α	正确结论的概率为（1-α）
	即**检验水准**	又称为**置信度**
H_0 不正确	推断正确	Ⅱ类错误
	正确结论的概率为（1-β）	犯错误的概率为 β
	又称为**检验效能**	

常用检验量

假设检验

根据假设检验的不同内容和进行检验的不同条件，需要采用不同的检验统计量，常用的检验统计量包括包括χ^2**统计量、t 统计量、F 统计量**

选择检验统计量主要考虑的因素包括

- **样本量n的大小**
- **总体标准差σ是否已知**
- **是独立样本检验还是配对样本检验**

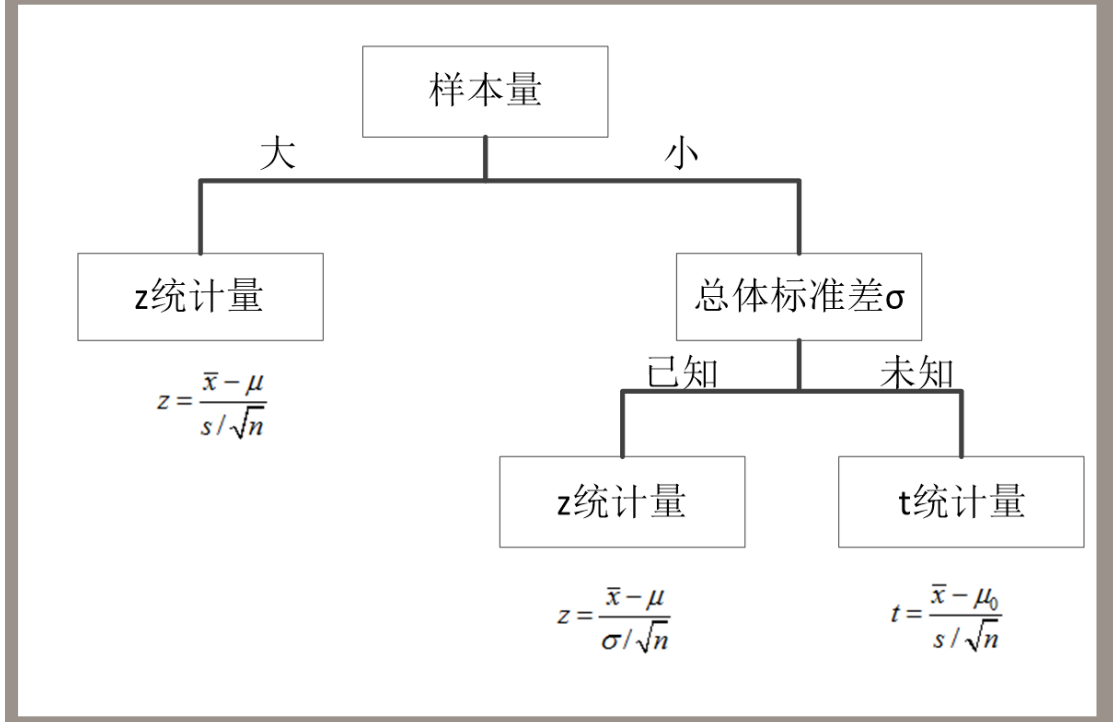
常用检验模式

- **一个总体均值的检验**
- **两个总体均值之差的检验**
- **一个总体比例的检验**
- **两个总体比例之差的检验**
- **一个总体方差的检验**
- **两个总体方差之比的检验**

总体均值的检验

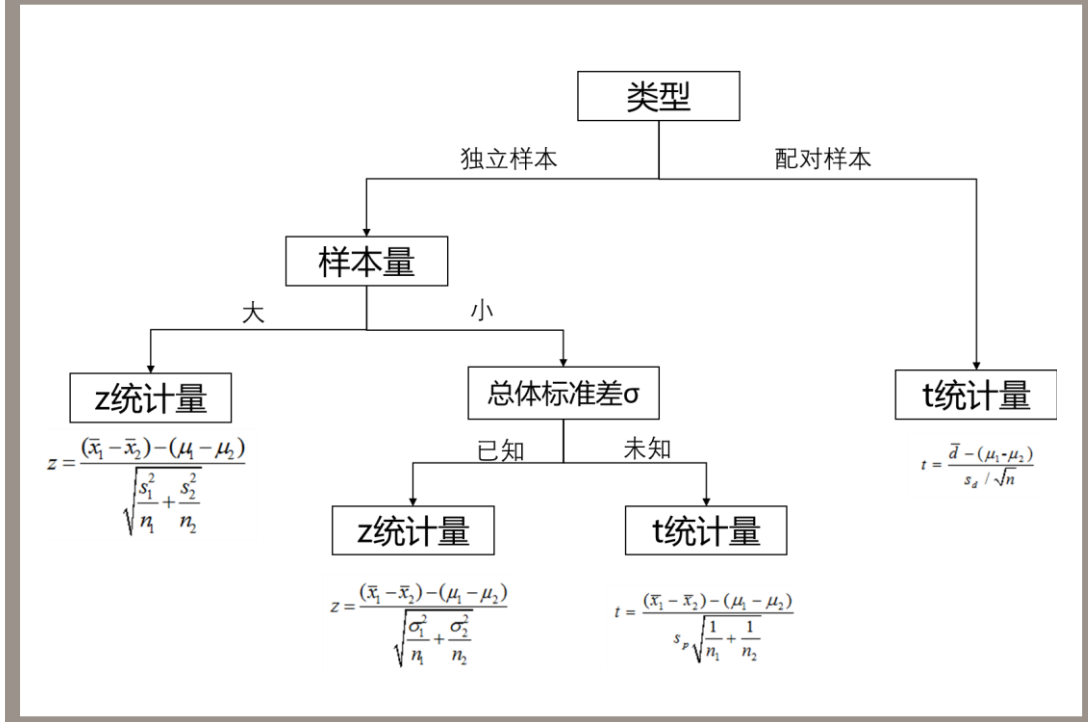
01

一个总体均值的检验



02

两个总体均值之差的检验



其他常用检验模式

总体比例的检验

一个总体比例的检验

大样本比例近似服从正态分布，使用z检验统计量。

$$z=\frac{p-\pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}$$

两个总体比例之差的检验

两大样本比例近似服从正态分布，使用z检验统计量。

$$z=\frac{(p_1-p_2)-(\pi_1-\pi_2)}{\sigma_{p_1-p_2}}$$

总体方差的检验

一个总体方差的检验

要求总体服从正态分布，使用卡方检验统计量。

$$\chi^2=\frac{(n-1)s^2}{\sigma_0^2}$$

两个总体方差之比的检验

要求两总体服从正态分布，使用F检验统计量。

$$F=s_1^2/s_2^2$$

以z分数为例的决策过程

适用于大样本单个总体均值检验

决策依据

有95%的把握断定：如果假设是正确的，则一个实际的样本统计量S的z分数在-1.96到1.96之间（正态曲线其积分面积为0.95）。阴影部分的总面积0.95就是检验的**显著性水平。**在（-1.96,1.96）之外的z分数的集合构成了所谓的假设的临界区域，也称假设的**拒绝域**

决策法则

如果统计量S的z分数在（-1.96,1.96）之外，即z>1.96或z<-1.96，则以0.05的显著性水平拒绝假设，否则接受假设。

Q&A

例2.13 某机床厂加工一种零件，根据经验知道，该厂加工零件的椭圆度渐近服从正态分布，其总体均值为0.081mm，今另换一种新机床进行加工，取200个零件进行检验，得到椭圆度均值为0.076mm，样本标准差为0.025mm，问新机床加工零件的椭圆度总体均值与以前有无显著差别。

解：在这个例题中，我们所关心的是新机床加工零件的椭圆度总体均值与老机床加工零件的椭圆度均值0.081mm是否有所不同，于是可以假设：

$H_0: \mu = 0.081mm$ 没有显著差别

$H_1: \mu \neq 0.081mm$ 有显著差别

这是一个双侧检验问题，所以只要 $\mu > \mu_0$ 或 $\mu < \mu_0$ 二者之中一个成立，就可以拒绝原假设。

由提议可知，$\mu_0 = 0.081mm$，$s = 0.025mm$，$\bar{x} = 0.076mm$。因为 $n>30$，故选用 z 统计量。

$$z = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{0.076 - 0.081}{0.025/\sqrt{200}} = -2.83$$

规定显著性水平 α=0.05，查表可以得出临界值：$z_{\alpha/2} = \pm 1.96$，z 的下标 $\alpha/2$ 表示双侧检验。因为 $|z| > |z_{\alpha/2}|$，根据决策标准，拒绝 H_0，可以认为新老机床加工零件椭圆度的均值有显著差别。

相关分析原理

1.相关关系

相关关系，又叫非确定性关系，它的特点是给定了一个变量值后，另一个变量值可以在一定的范围内变化，它必须借助于统计手段才能加以研究，故又称为统计相关。

2.相关分析

相关分析是研究变量之间相关关系的数理统计方法，它可以从影响某个变量的诸多变量中判断哪些是显著的，哪些是不显著的。而且，在得到相关分析的结果后，还可以用其他统计分析方法对其做更进一步的分析、预测或控制，比如回归分析、因子分析等。

相关关系类型	变量X	变量Y	影响因素
强正相关关系	增加	明显增加	X是影响Y的主要因素
弱正相关关系	增加	增加，但增加幅度不明显	X不是唯一影响因素
强负相关关系	增加	明显减少	X是影响Y的主要因素
弱负相关关系	增加	减少，但减少幅度不明显	X不是唯一的影响因素
非线性关系	X、Y之间没有明显的线性关系，却存在着某种非线性关系		X是影响Y的因素
不相关	X、Y之间不存在相关关系		X不是影响Y的因素

表1 相关关系的类型

相关分析
计算

Pearson相关系数

线性相关（Linear correlation）又称简单相关。用来度量具有线性关系的两个变量之间相关关系的密切程度及其相关方向，适用于双变量正态分布资料。线性相关系数又称为简单相关系数、Pearson相关系数或相关系数，有时也称为积差相关系数。

其中$Cov(X,Y)$是随机变量X、Y的协方差，$Var(X)$和$Var(Y)$分别代表X和Y的方差。总体相关系数是反映两变量之间线性相关程度的一种特征值。

样本相关系数是根据样本观测值计算的，抽取的样本不同，其具体的数值也会有所差异。可以证明，样本相关系数是总体相关系数的一致估计量。

r：样本相关系数；ρ：总体相关系数

公式

- 总体相关系数定义公式：

$$\rho_{XY} = Corr(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}}$$

- 样本相关系数的定义公式：

$$r_{XY} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

Pearson相关系数性质

01

若$0<r\le1$，表明x与y之间存在正线性相关关系；若 $-1\le r<0$,表明x与y之间存在负线性相关关系；若$r=0$，表明x与y之间无线性关系；若$r=+1$，表明x与y之间为完全正线性相关关系；若$r=-1$，表明x与y之间为完全负线性相关关系。

02

R具有对称性，x与y之间的相关系数 r_{xy}，和y与x之间的相关系数r_{yx}相等，即 $r_{xy}=r_{yx}$。

03

r的数值大小与x和y的原点及尺度无关。改变x和y的数据原点并不改变 r的数值大小。

04

r仅仅是x与y之间线性关系的一个度量, $r=0$ 只表示两个变量之间不存在线性相关关系，并不说明变量之间没有任何关系，它们之间可能存在非线性相关关系。

05

$|r|\to1$说明两个变量之间的线性关系越强；$|r|\to0$说明两个变量之间的线性关系越弱。当$r\ge0.75$ 时可视为高度相关；$0.50\le|r|<0.75$时，可视为中度相关；当 $0.30\le|r|<0.50$ 时，视为低度相关；当$|r|<0.30$，说明两个变量之间的相关程度极弱。但这种解释必须建立在对相关系数的显著性进行检验的基础之上。

06

判断样本相关系数 r 是否来自 $\rho\neq0$ 的总体，需要对它进行显著性检验，此处可以采用 t 检验或F 检验，此时的零假设和被择假设分别为 $H_0:\rho=0,H_A:\rho\neq0$。

t 检验统计量 $t=r/S_r,df=n-2;S_r=\sqrt{(1-r^2)/(n-2)}$ 称为相关系数的标准误差。

F 检验统计量 $F=\dfrac{r^2}{(1-r^2)/(n-2)},df_1=1,df_2=n-2$

Spearman等级相关系数

相关分析
计算

01

取值范围为[-1，1]

Spearman相关系数的取值范围在-1到1之间，绝对值越大相关性越强，取值符号也表示相关的方向。

02

计算公式

随机变量X、Y之间的Spearman相关系数记为 r_S ，其计算公式为 $r_S = 1 - \frac{6\sum d^2}{n(n^2-1)}$ 其中 d 为分别对X、Y去秩之后每对观察值 (x, y) 的秩之差，n 为所有观察对的个数。

03

P值

Spearman相关系数 r_S 假设检验的零假设为 r_S 是来自 $\rho_S = 0$ 的总体（即X与Y独立）。以显著性水平 $\alpha = 0.05$ 为例，当 $n \le 30$ 或50时，可以查Spearman’s相关系数表来确定 P 值，此时有：当 $P \le 0.05$ 时，拒绝零假设，说明X与Y之间存在着较为显著的相关关系；当时 $P > 0.05$ ，接受零假设。

Kendall等级相关系数

相关分析
计算

Kendall相关系数是对两个有序变量或两个秩变量之间相关程度的度量统计量。

Kendall’ s tau(nonparametric correlations algorithms) 算法:

两个随机变量X、Y共有t组观测对(x, y)对任意第(i, j)个观测数据，若满足$i < j$，就计算$d_{ij} = [R(X_j) - R(X_i)][R(Y_j) - R(Y_i)]$

令$S = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} sign(d_{ij})$，则Kendall’ s tau$(\tau)$按如下公式计算:

$$\tau = \frac{S}{\sqrt{\frac{N^2 - N - \tau_x}{2}} \sqrt{\frac{N^2 - N - \tau_y}{2}}}$$

Kendall’ s tau相关系数的显著性检验通过统计量$Z = \frac{S}{\sqrt{\mathrm{d}}}$进行，在零假设（X、Y不相关）成立的条件下，它近似服从正态分布。

Kendall 的 tau-b (Kendall's tau-b)是将结果考虑在内的有序变量或排序变量的非参数相关性测量。系数的符号指示关系的方向，绝对值指示强度，绝对值越大则表示关系强度越高，可能的取值范围是从-1到1。

1.数学模型

相关分析
偏相关分析

变量之间的相关关系

- 线性相关分析计算的是两个变量间的相关系数。
- 在多变量的情况下，只有除去其他变量影响后再计算相关系数，才能真正反映它们之间的相关关系

简单相关

- 在其他变量固定不变的情况下，计算两个指定变量之间的相关系数，这样的相关分析就是偏相关分析，经此得出的相关系数叫做偏相关系数。
- 根据固定变量个数的多少，偏相关分析可分为零阶偏相关、一阶偏相关和(p-1)阶偏相关，其中零阶偏相关就是简单相关。

偏相关系数

- 设随机变量X,Y,Z之间彼此存在着相关关系，为了研究Z和Y之间的关系，就必须在假定Z不变的条件下，计算X和Y 的偏相关系数，记为r_{xyz}
- 偏相关系数是由简单相关系数决定的，但是在计算偏相关系数时要考虑其他自变量对指定变量的影响。

相关分析
偏相关分析

以下标0代表X，下标1代表Y，下标2代表Z，则Z和Y间的一阶偏相关系数定义为 $r_{01\cdot 2}=\frac{r_{01}-r_{02}r_{12}}{\sqrt{1-r_{02}^2}\sqrt{1-r_{12}^2}}$,其中 $r_{01\cdot 2}$ 是剔除Z的影响之后Z和Y的偏相关系数， r_{01},r_{02},r_{12} 分别是X，Y，Z之间的两两简单相关系数。如果增加个变量T(以下标3表示)，则Z和Y的二阶偏相关系数定义为 $r_{01\cdot 23}=\frac{r_{02}-r_{03\cdot 2}r_{13\cdot 2}}{\sqrt{1-r_{03\cdot 2}^2}\sqrt{1-r_{13\cdot 2}^2}}$

一般，考察多个变量时，Y与 $x_i\,(\mathrm{i}=1,2,\ldots,\mathrm{p})$之间的 $p-1$ 阶偏相关系数可由如下的递推式定义：

$$r_{0i\cdot 12\ldots(\mathrm{i}-1)(\mathrm{i}+1)\cdots\mathrm{p}}=\frac{r_{0i\cdot 12\ldots(\mathrm{i}-1)(\mathrm{i}+1)\cdots(\mathrm{p}-1)}-r_{0ip\cdot 12\ldots(p-1)}r_{ip\cdot 12\ldots(\mathrm{i}-1)(\mathrm{i}+1)\cdots(\mathrm{p}-1)}}{\sqrt{1-r_{0ip\cdot 12\ldots(p-1)}^2}\sqrt{1-r_{ip\cdot 12\ldots(\mathrm{i}-1)(\mathrm{i}+1)\cdots(\mathrm{p}-1)}^2}}$$

相关分析
偏相关分析

2.偏相关系数的检验

偏相关系数检验的零假设为：总体中两个变量间的偏相关系数为0。使用t检验方法。

公式：

$$t = \frac{\sqrt{b^2 - 4ac} * r}{\sqrt{1 - r^2}}$$

自由度

其中，r是相应的偏相关系数，n是样本观测数，k是可控制变量的数目，n-k-2是自由度。当$t>t_{0.05}(n-k-2)$或$p<0.05$时，拒绝原假设。

主要应用

偏相关分析的一个主要应用是根据观测资料计算样本的偏相关系数

分析

经过这样的选择，在进行多元回归等分析时，就可以只考虑那些起主要作用的因素。

应用

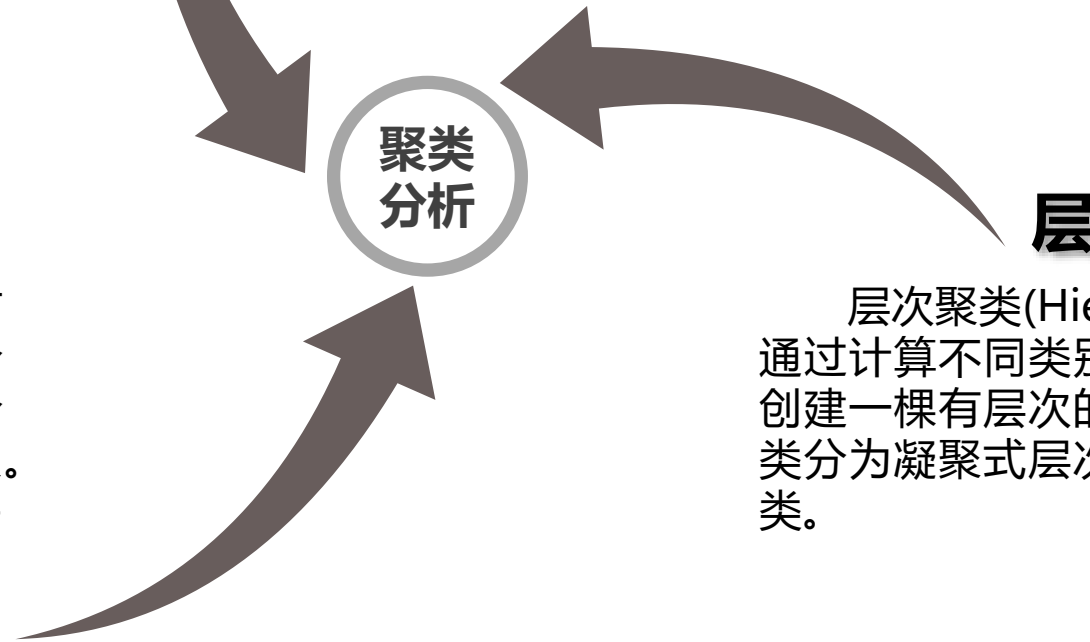
偏相关分析的应用也非常广泛，涉及自然科学和社会科学的各个方面。

聚类分析

概念

概念：

聚类（Clustering）是将某个对象集划分为若干组或簇（Class或Cluster）的过程，使得同一个组内的数据对象具有较高的相似度，而不同组中的数据对象是不相似的。相似或不相似的定义基于属性变量的取值确定，一般就采用各对象间的距离来表示。



划分方法：

给定一个n个样本的数据集，划分方法构建数据的k个分区(k≤n)，其中每个分区表示一个簇。也就是说，它把数据划分为k个组，使得每个组至少包含一个对象。常用划分方法有k-means算法和k-medoids算法。

层次方法：

层次聚类(Hierarchical Clustering)通过计算不同类别数据点间的相似度来创建一棵有层次的嵌套聚类树。层次聚类分为凝聚式层次聚类和分裂式层次聚类。

聚类分析
划分方法

1.k-means算法(k-均值算法)

简介

(1)误差平方和达到最优（小）时，可以使各聚类的类内尽可能紧凑，使各聚类之间尽可能分开。
(2)对于同一个数据集，可用该准则评价聚类结果的优劣。
(3) 对于任意一个数据集，k-means算法无法达到全局最优，只能达到局部最优。

算法优点

(1)算法快速、简单；

(2)对大样本量数据有较高的效率并且具有可伸缩性；

(3)时间复杂度近于线性。

算法缺点

(1)在 K-means 算法中 K 是事先给定的，K 值的选定难以估计
(2)初始聚类中心的选择对聚类结果有较大的影响，一旦初始值选择的不好，可能无法得到有效的聚类结果。
(3) 当样本量非常大时，算法的时间开销非常大。
(4)对噪声和离群点数据敏感。

聚类分析
划分方法

1.k-means算法(k-均值算法)

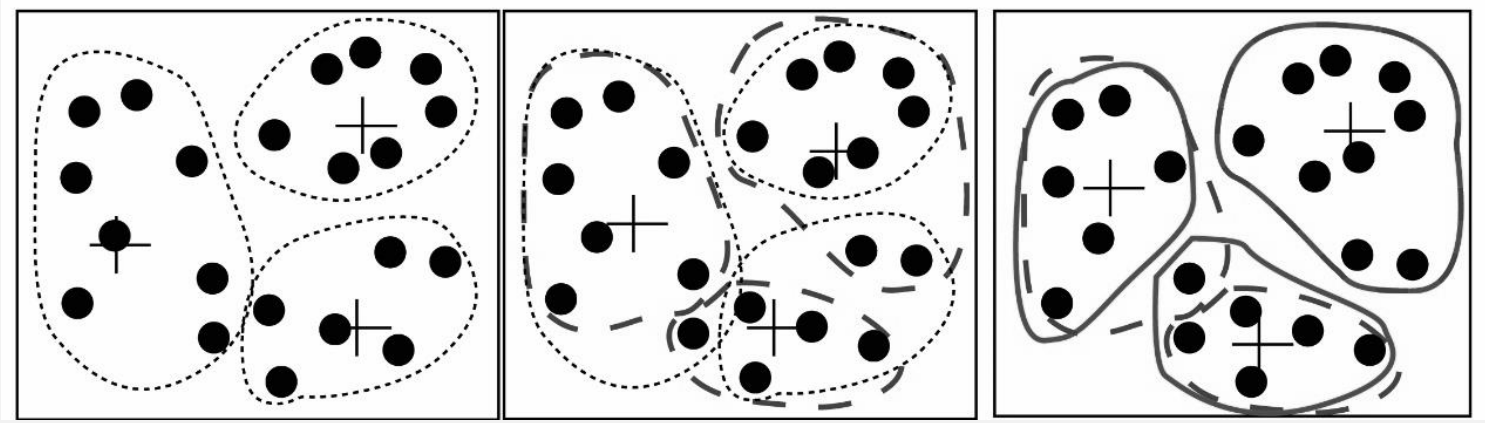
例2.19 使用K-means聚类算法

考虑二维空间的对象集合，如图2-22a所示。令k =3，即用户要求将这些对象划分成3个簇。

我们任意选择3个对象作为3个初始的簇中心，其中簇中心用”+”标记。根据与簇中心的距离，每个对象被分配到最近的一个簇。这种分配形成了如图2-22a中虚线所描绘的轮廓。

下一步，更新簇中心。也就是说，根据簇中的当前对象，重新计算每个簇的均值。使用这些新的簇中心，把对象重新分布到离簇中心最近的簇中。这样的重新分布形成了图2-22b中虚线所描绘的轮廓。

重复这一过程，形成图2-22c所示结果。这种迭代地将对象重新分配到各个簇，以改进划分的过程被称为迭代的重定位(Iterative relocation)。最终，对象的重新分配不再发生，处理过程结束，聚类过程返回结果簇。



a)初始聚类　　b)迭代　　c)最终的聚类

图2-22 使用k一均值方法聚类对象集；更新簇中心，并相应地重新分配诸对象 (每个簇的均值都用“+”标注)

聚类分析
划分方法

2. k-medoids算法(k-中心点算法)

1) K-medoids算法基本思想

为了降低聚类算法对离群点的敏感度，可以挑选实际对象来代表簇，即：选取最靠近中心点（medoid）的那个对象来代表整个簇。每个簇使用一个代表对象。其余的每个对象被分配到与其最为相似的代表性对象所在的簇中。定义为：

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} dist(p, o_i)$$

这里 E 是数据集中所有对象 p 与 C_i 的代表对象 o_i 的绝对误差之和。

2. k-medoids算法(k-中心点算法)

2）例子

例2.20 使用**k-medoids算法。**设 $o_{1,}\cdots,o_k$ 是当前代表对象（即中心点）的集合，为了决定一个非代表对象 O_{random} 是否是一个当前中心点$o_j(1\le j\le k)$的好替代，我们计算每个对象P到集合$\{o_{1,}\cdots,o_{j-1},o_{random},o_{j+1,}\cdots,o_k\}$中最近对象的距离，并使用该距离更新代价函数。对象重新分配到 $\{o_{1,}\cdots,o_{j-1},o_{random},o_{j+1,}\cdots,o_k\}$ 中是简单的。假设对象P当前被分配到中心点 o_j 代表的簇中(见图2-23a或图2-23b)。当o_{random}替代代表对象o_j后，对于数据集中的每一个对象P，它所属的簇的类别将有以下四种可能的变化：

四种可能变化

(1)对象 p 属于代表对象 o_j。替代后 p 最接近于 o_i，因此 p 被分配为 $o_i\ (i\neq j)$ 。

(1)对象 p 属于代表对象 o_j。替代后 p 最接近于 o_{random}，因此 p 被分配为 o_{random}。

(1)对象 p 属于代表对象 o_j。替代后 p 最接近于 o_i，因此 p 无变化。

(1)对象 p 属于代表对象 o_j。替代后 p 最接近于 o_{random}，因此 p 被分配为 o_{random} 。

聚类分析
划分方法

2. k-medoids算法(k-中心点算法)

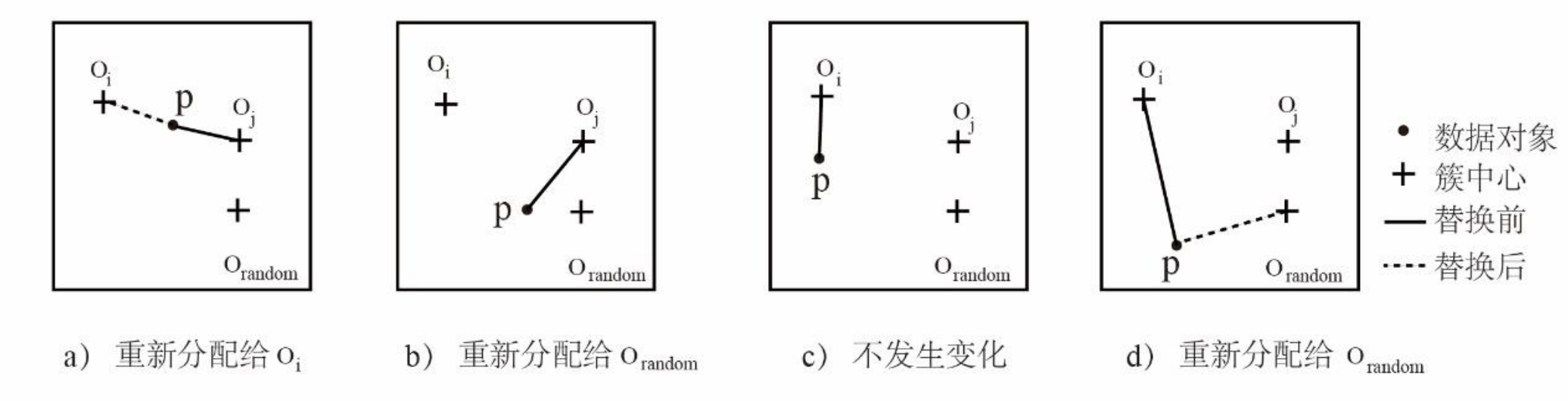


图2-23 k-medoids聚类代价函数的4种情况

k-means

（1）当存在噪声和离群点时，k-medoids算法比k-means算法更加棒。

VS

k-medoids

（2）时间复杂度为 $O(k(n-k)^2)$，k-medoids算法执行代价比k-means算法要高 k 。

聚类分析
层次方法

层次聚类(Hierarchical Clustering)是聚类算法的一种，通过计算不同类别数据点间的相似度来创建一棵有层次的嵌套聚类树。

1. 凝聚式层次聚类算法

凝聚式层次聚类的合并算法通过计算两类数据点间的相似性，对所有数据点中最为相似的两个数据点进行组合，并反复迭代这一过程。并将距离最近的两个数据点或类别进行组合，生成聚类树。

所谓从下而上地合并cluster，具体而言，就是每次找到距离最短的两个cluster，然后进行合并成一个大的cluster，直到全部合并为一个cluster。整个过程就是建立一个树结构，类似于图2-24。

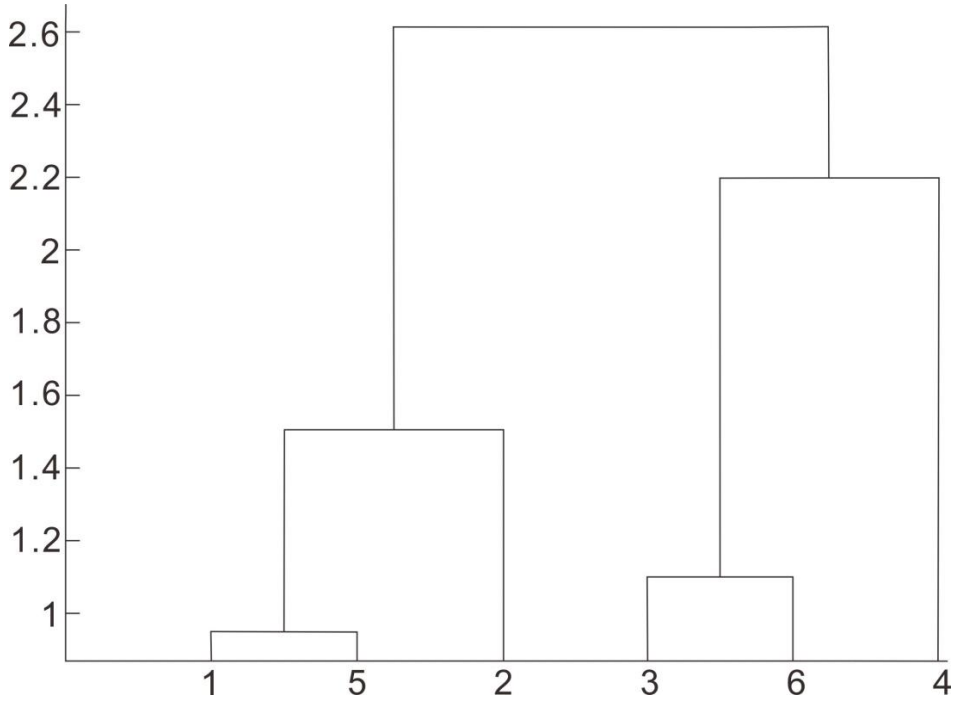


图2-24 层次的嵌套聚类树示意图

聚类分析
层次方法

1. 凝聚式层次聚类算法

如图2-25为示例数据，我们通过欧氏距离计算下面P1到P5的欧式距离矩阵，并通过合并的方法将相似度最高的数据点进行组合，并创建聚类树。

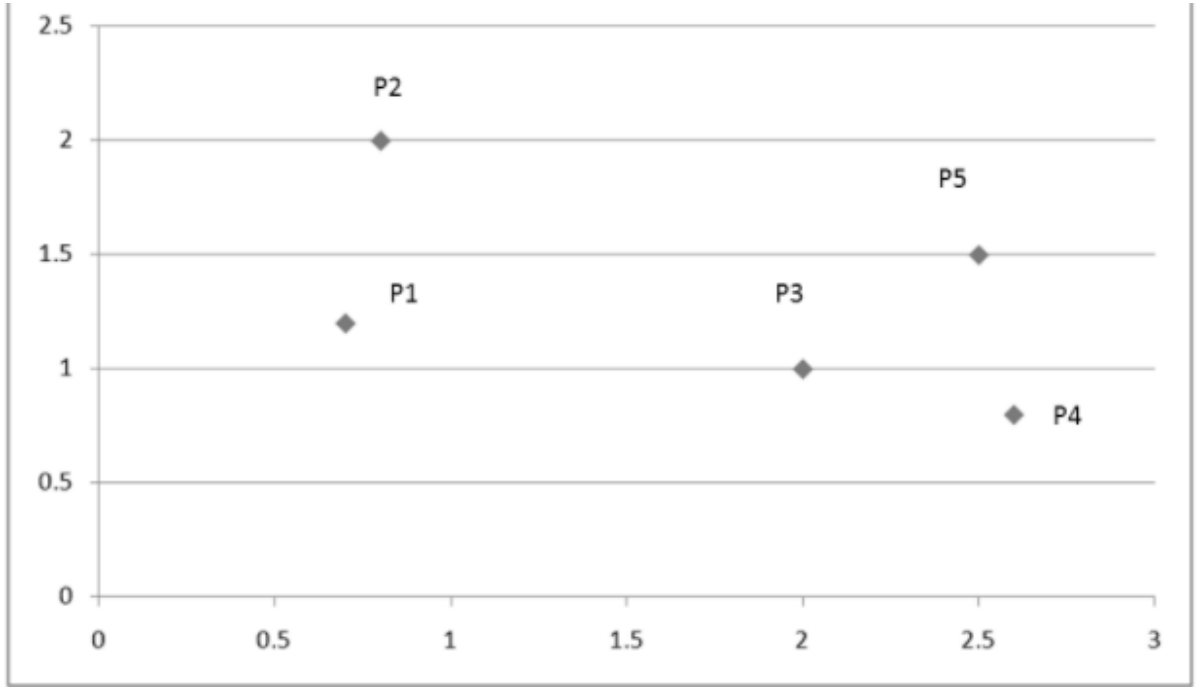


图2-25 层次聚类举例

	P1	P2	P3	P4	P5
P1	0	0.81	1.32	1.94	1.82
P2	0.81	0	1.56	2.16	1.77
P3	1.32	1.56	0	0.63	0.71
P4	1.94	2.16	0.63	0	0.71
P5	1.82	1.77	0.71	0.71	0

图2-26 欧式距离原始矩阵

聚类分析
层次方法

1. 凝聚式层次聚类算法

凝聚式层次聚类的合并算法通过计算两类数据点间的相似性，对所有数据点中最为相似的两个数据点进行组合，并反复迭代这一过程。并将距离最近的两个数据点或类别进行组合，生成聚类树。

所谓从下而上地合并cluster，具体而言，就是每次找到距离最短的两个cluster，然后进行合并成一个大的cluster，直到全部合并为一个cluster。整个过程就是建立一个树结构，类似于图2-24。

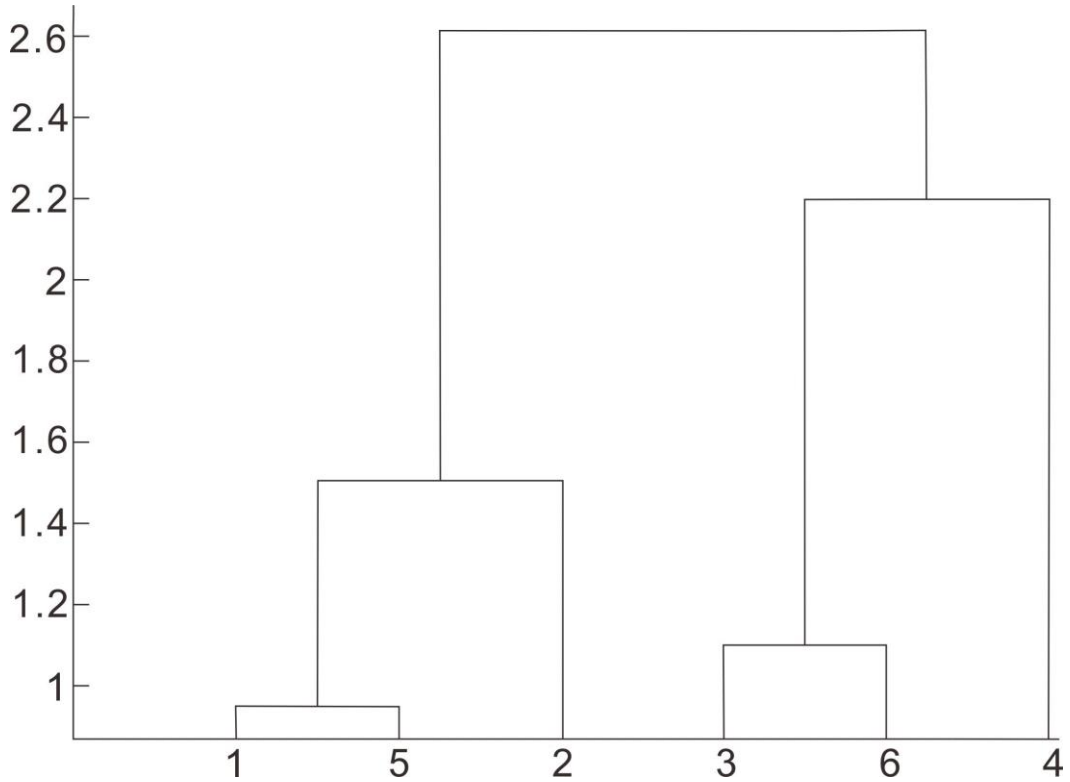


图2-24 层次的嵌套聚类树示意图

聚类分析
层次方法

1. 凝聚式层次聚类算法

01 根据算法流程，我们先找出距离最近的两个簇，P3，P4。合并 P3，P4 为 {P3, P4}，根据 MIN 原则更新矩阵如下：
MIN.distance({P3, P4}, P1) = 1.32 ; MIN.distance({P3, P4}, P2) = 1.56;
MIN.distance({P3, P4}, P5) = 0.70

	P1	P2	{P3, P4}	P5
P1	0	0.81	1.32	1.82
P2	0.81	0	1.56	1.77
{P3, P4}	1.32	1.56	0	0.71
P5	1.82	1.77	0.71	0

图2-27 欧式距离更新矩阵1

02 接着继续找出距离最近的两个簇，{P3, P4}, P5。合并 {P3, P4}, P5 为 {P3, P4, P5}，根据 MIN 原则继续更新矩阵：
MIN.distance(P1, {P3, P4, P5}) = 1.32;
MIN.distance(P2, {P3, P4, P5}) = 1.56;

	P1	P2	{P3, P4, P5}
P1	0	0.81	1.32
P2	0.81	0	1.56
{P3, P4, P5}	1.32	1.56	0

图2-28 欧式距离更新矩阵2

03 接着继续找出距离最近的两个簇，P1, P2。 合并 P1, P2 为 {P1, P2}，根据 MIN 原则继续更新矩阵：
MIN.distance({P1,P2}, {P3, P4, P5}) = 1.32

	{P1, P2}	{P3, P4, P5}
{P1, P2}	0	1.32
{P3, P4, P5}	1.32	0

图2-29 欧式距离更新矩阵3

04 最终合并剩下的这两个簇即可获得最终结果，如图2-30：

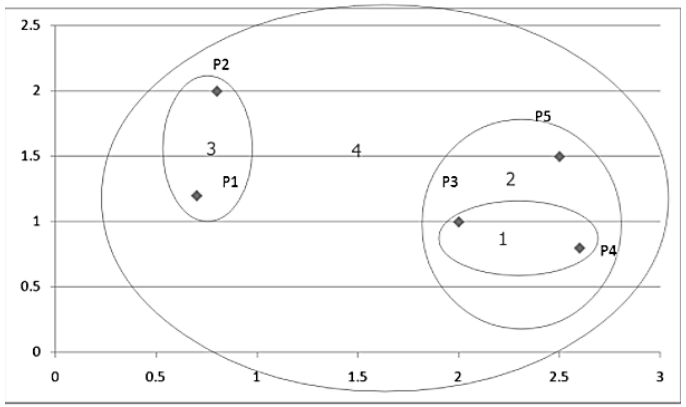


图2-30 层次聚类举例结果

聚类分析
层次方法

2. 分裂式层次聚类算法

分裂方法DIANA(Divisive Analysis)将所有的对象形成一个初始簇，根据某种原则(如簇中最近的相邻对象的最大欧氏距离)，将该簇分裂。簇的分裂过程反复进行，直到最终每个新的簇只包含一个对象。

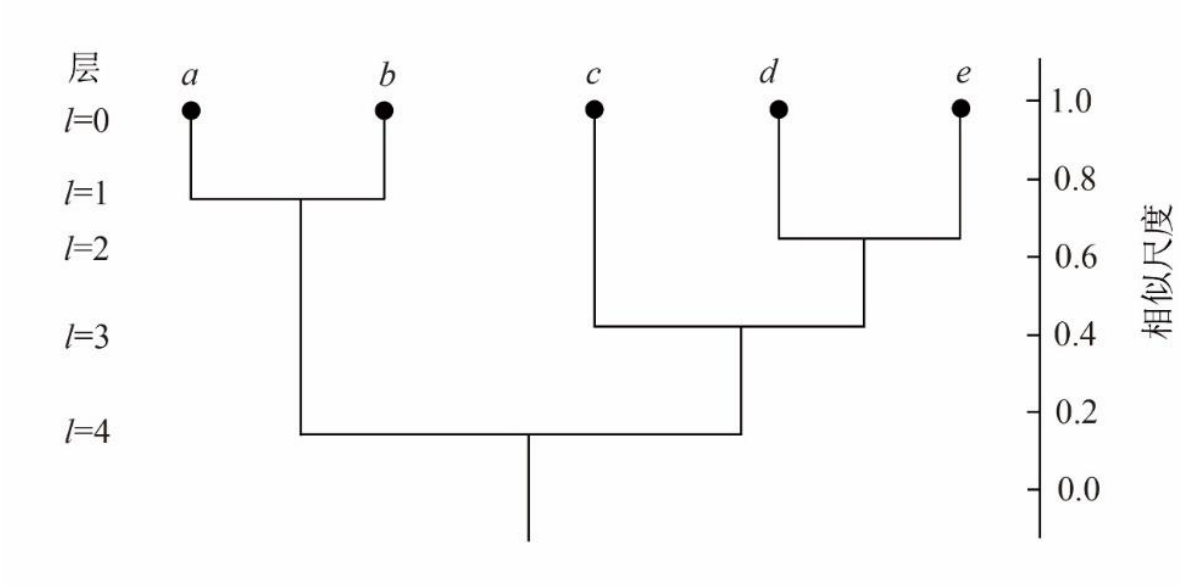


图2-31 数据对象 $\{a,b,c,d,e\}$ 层次凝聚的树状图表示

判别分析

判别分析是判别样品所属类型的一种统计方法，根据观测到的数据资料，对所研究的对象进行分类。

判别分析与聚类分析不同

- 判别分析是在已知研究对象分成若干类型（或组别）并已取得各种类型的一批已知样品的观测数据，在此基础上根据某些准则建立判别式，然后对未知类型的样品进行判别分类。
- 对于聚类分析来说，一批给定样品要划分的类型事先并不知道，正需要通过聚类分析来给以确定类型的。
- 因此，判别分析和聚类分析往往联合起来使用。

判别分析
距离判别

1) 两个总体的距离判别方法（方差相等）

首先根据已知分类的数据，分别计算各类的重心即各组（类）的均值，判别的准则是对任给样品，计算它到各类平均数的距离，哪个距离最小就将它判归哪个类。

$y \in G_1$，如 $d^2(y,G_1) < d^2(y,G_2)$

$y \in G_2$，如 $d^2(y,G_2) < d^2(y,G_1)$

待判，如 $d^2(y,G_1) = d^2(y,G_2)$

$$
\begin{aligned}
&d^2(y,G_2) - d^2(y,G_1) \\
&= (y-\mu_2)'\sum{}^{-1}(y-\mu_2) - (y-\mu_1)'\sum{}^{-1}(y-\mu_1) \\
&= y'\sum{}^{-1}y - 2y'\sum{}^{-1}\mu_2 + \mu'_2\sum{}^{-1}\mu_2 - (y'\sum{}^{-1}y - 2y'\sum{}^{-1}\mu_1) + \mu'_1\sum{}^{-1}\mu_1) \\
&= 2y'\sum{}^{-1}(\mu_1-\mu_2) - (\mu_1+\mu_2)'\sum{}^{-1}(\mu_1-\mu_2) \\
&= 2[y - \frac{(\mu_1+\mu_2)}{2}]'\sum{}^{-1}(\mu_1-\mu_2)
\end{aligned}
$$

令 $\bar{\mu} = \frac{\mu_1+\mu_2}{2}$　　$\alpha = \sum{}^{-1}(\mu_1-\mu_2) = (a_1,a_2,...,a_p)'$

$$
\begin{aligned}
W(y) &= (y-\bar{\mu})'\alpha = \alpha'(y-\bar{\mu}) \\
&= a_1(y_1-\bar{\mu}_1) + ... + a_p(y_p-\bar{\mu}_p) \\
&= \alpha' y - \alpha'\bar{\mu}
\end{aligned}
$$

$y \in G_1$，如 $W(y) > 0$

$y \in G_2$，如 $W(y) < 0$

待判，如 $W(Y) = 0$

2) 多总体的距离判别方法（协方差相等）

设有个K总体，分别有均值向量$u_1(i = 1,2,\ldots,k)$和协方差阵∑i=∑,又设Y是一个待判样品。则Y与各总体的距离为（即判别函数）：

$$
\begin{aligned}
d^2(y,G_i) &= (y-\mu_i)'\sum{}^{-1}(y-\mu_i) \\
&= y'\sum{}^{-1}y - 2y'\sum{}^{-1}\mu_i + \mu'_i\sum{}^{-1}\mu'_i \\
&= y'\sum{}^{-1}y - 2(y'\sum{}^{-1}\mu_i - 0.5\mu'_i\sum{}^{-1}\mu'_i)
\end{aligned}
$$

令 $f_i(y) = (y'\sum{}^{-1}\mu_i - 0.5\mu'_i\sum{}^{-1}\mu'_i)$

则距离判别法的判别函数为：

$f_i(y) = (y'\sum{}^{-1}\mu_i - 0.5\mu'_i\sum{}^{-1}\mu'_i)$

判别规则为 $f_i(y) = \max\limits_{1\le i\le k} f_i(y)$，则 $y \in G_l$

$f_i(y) = \max\limits_{1\le i\le k} f_i(y)$，意味着 $d^2(y,G_l) = \min\limits_{1\le i\le k} d^2(y,G_i)$

判别分析
贝叶斯判别

贝叶斯判别法是通过计算被判样本x属于k个总体的条件概率 $P(n \mid x), n = 1, 2, \ldots, k$ 。比较k个概率的大小，将样本判归为来自出现概率最大的总体（或归属于错判概率最小的总体）的判别方法。

1）最大后验概率准则

设有 k 个总体 $G_1, G_2, G_3, \ldots, G_k$，且总体 G_i 的概率密度为 $f_i(x)$，样本 x 来自 G_i 的先验概率为 $q_i, i = 1, 2, \ldots, k$，满足 $q_1 + q_2 + \ldots + q_k = 1$。利用贝叶斯理论，x 属于 G_i 的后验概率，即当样本 x 已知时，它属于 G_i 的概率为：

$$P(G_i \mid x) = \frac{q_i f_i(x)}{\sum_{i=1}^{k} q_i f_i(x)} \quad i = 1, 2, \ldots, k$$

最大后验概率判别准则： $x \in G_l$， 若 $P(G_l \mid x) = \max_{1 \le i \le k} P(G_l \mid x)$。

判别分析
贝叶斯判别

2）最小平均误判准则

定义（平均错判损失）：

用 $P(j|i)$ 表示将来自总体 G_i 的样品错判到总体 G_j 的条件概率。

$$P(j|i)=P(X\in D_i|G_i)=\int_{D_j} f_i(x)dx \qquad i\neq j$$

$C(j|i)$ 表示相应错判所造成的损失。

则平均错判损失为：$ECM=\sum_{i=1}^{k} q_i \sum_{j\neq i} C(j|i)P(j|i)$，是使 ECM 最小的分划，是贝叶斯判别分析的解。

判别分析
费歇尔判别

费歇尔判别法，就是用投影的方法将k个不同总体在p维空间上的点尽可能分散，同一总体内的各样本点尽可能的集中。用方差分析的思想则可构建一个较好区分各个总体的线性判别法。

两个总体的费歇尔判别法

基本思想

设有A、B两个总体，分别有n_1和n_2个历史样本数据，每个样本有P个观测指标，每个样本可看作P维空间中的一点。费歇尔借助于方差分析的思想构造一个线性判别函数：

$$y = c_1 x_1 + c_2 x_2 + \ldots + c_p x_p$$

其中，判别系数 $c_1, c_2, \cdots c_p$ 的选择应使得y值满足：

(1)A类和B类的样本点群尽可能远离；

(2)同一类的样本点尽可能集中。

判别分析
费歇尔判别

两个总体的费歇尔判别法

费歇尔思想：不同总体A、B的投影点应尽量分散，即 $\max(\bar{y}_a-\bar{y}_b)^2$；而同一总体的投影点应该尽量集中，即 $\min(y_{bi}-\bar{y}_b)^2$ ，$(i=1,2,...,n_a)$ 和 $\min(y_{bi}-\bar{y}_b)^2$ ，$(i=1,2,...,n_b)$。

构造统计量

$$F(c_1,c_2,...,c_m)=\frac{(\bar{y}_a-\bar{y}_b)^2}{\sum_{i=1}^{n_a}(y_{ai}-\bar{y}_a)^2+\sum_{i=1}^{n_b}(y_{bi}-\bar{y}_b)^2}$$

则只要F取最大值就满足费歇尔准则要求。

求 $F(c_1,c_2,...,c_m)$ 的最大值点确定出参数 $c_1,c_2,...,c_m$ 的值，从而可以得出判别函数：$y(X)=c_1x_1+c_2x_2+...+c_mx_m$

为此，只需要令 $\frac{\partial F}{\partial c_i}=0$ ，$(i=1,2,...,m)$

$$v\begin{cases}s_{11}c_1+s_{12}c_2+\cdots+s_{1m}c_m=d_1\\ s_{21}c_1+s_{22}c_2+\cdots+s_{2m}c_m=d_2\\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\\ s_{m1}c_1+s_{m2}c_2+\cdots+s_{mm}c_m=d_m\end{cases}$$

其中 $d_i=\bar{x}_{Ai}-\bar{x}_{Bi}$

$$\bar{x}_{Ai}=\frac{1}{n_a}(x_{A1i}+x_{A2i}+\cdots+x_{An_ai})$$

$$\bar{x}_{Bi}=\frac{1}{n_b}(x_{B1i}+x_{B2i}+\cdots+x_{Bn_bi})$$

$$s_{jl}=\sum_{i=1}^{n_a}(x_{Aji}-\bar{x}_{Aj})(x_{Ali}-\bar{x}_{Al})+\sum_{i=1}^{n_b}(x_{Bji}-\bar{x}_{Bj})(x_{Bli}-\bar{x}_{Bl})$$

判别分析
费歇尔判别

两个总体的费歇尔判别法

费歇尔判别法是致力于寻找一个最能反映组和组之间差异的投影方向，即寻找线性判别函数$Y(x)=c_1x_1+...+c_px_p$。设有k个总体$G_1,G_2,...,G_k$,分别有均值向量$\vec{\mu}_1,\vec{\mu}_2,...,\vec{\mu}_k$和协方差阵$\Sigma_1,...,\Sigma_k$,分别各总体中得到样品：

公式

$$X_1^{(1)},...,X_{n_1}^{(1)}$$
$$X_1^{(2)},...,X_{n_2}^{(2)}$$
$$...$$
$$n_1+n_2+...+n_k=n$$

第i个总体样本组内离差平方和 $V_i=\sum_{t=1}^{n_i}(X_t^{(i)}-\bar{X}_i)(X_t^{(i)}-\bar{X}_i)'$

综合的组内离差平方和 $E=V_1+V_2+...+V_k$

组间离差平方和 $B=\sum_{i=1}^{k}n_i(\bar{X}_i-\bar{X})(\bar{X}_i-\bar{X})'$

因为 $Y(x)=c_1x_1+...+c_px_p$

$$V_{iy}=\sum_{t=1}^{n_i}(Y_t^{(i)}-\bar{Y}_i)^2=V_i=\sum_{t=1}^{n_i}(Y_t^{(i)}-\bar{Y}_i)(Y_t^{(i)}-\bar{Y})'=C'V_iC$$

$$E_0=\sum_{i=1}^{k}V_{iy}=\sum_{i=1}^{k}C'V_iC=C'EC$$

$$B_0=\sum_{i=1}^{k}n_i(\bar{Y}_i-\bar{Y})^2=\sum_{i=1}^{k}n_i(\bar{Y}_i-\bar{Y})(\bar{Y}_i-\bar{Y})'=C'BC$$

判别分析
费歇尔判别

两个总体的费歇尔判别法

如果判别分析是有效的，则所有的样品的线性组合 $Y(x)=c_1x_1+...+c_px_p$ 满足组内离差平方和小，而组间离差平方和大。则:

公式

$$\Delta^2(C)=\frac{B_0}{E_0}=\frac{C'BC}{C'EC}=\max$$

$$\Delta^2(C)=\frac{B_0}{E_0}=\frac{C'BC}{C'EC}$$ 的最大值Δ是最大的特征根 λ_1。

而 λ_1 所对应的特征向量即 $C_1=(c_{11},...,c_{p1})'$

判别规则：设 $y_i(X)$ 为第i个线性判别函数，$(i=1,2,...,m)$，

$$d(x,G_k)=\sum_{i=1}^{r}(y_i(x)-y_i(\overline{x}_k))^2$$

则 $d(x,G_t)=\min_{1\le j\le k}d(x,G_k)$

判别分析
费歇尔判别

例2.21 试用各种判别方法对样本进行判类。

1990年联合国开发计划署公布的《人类发展报告》，用出生时的预期寿命（x1）、成人识字率（x2）、实际的人均GDP（x3）等三个变量衡量人类发展状况，现从高发展水平国家和中等发展水平国家中各选了5个样本，另选M、N两国作为待判样本（如表2-8所示）。要求：

（1）作距离判别分析（假定两总体协方差阵相等）；

（2）作Fisher判别分析。

类别	序号	国家	出生时的预期寿命(x_1)	成人识字率(x_2)	实际人均GDP(x_3)
第一类（高发展水平国家）	1	A	76	99	5374
	2	B	79.5	99	5359
	3	C	78	99	5372
	4	D	72.1	95.9	5242
	5	E	73.8	77.7	5370
第二类（中等发展水平国家）	6	F	71.2	93	4250
	7	G	75.3	94.9	3412
	8	H	70	91.2	3390
	9	I	72.8	99	2300
	10	J	62.9	80.6	3799
待判样本	11	M	68.5	79.3	1950
	12	N	77.6	93.8	5233

相关关系的类型

判别分析
费歇尔判别

(1)距离判别

1 计算两类样本均值：

$$\bar{X}_1=\begin{pmatrix}75.88\\94.08\\5343.4\end{pmatrix}\qquad \bar{X}_2=\begin{pmatrix}70.44\\91.74\\3430.2\end{pmatrix}$$

2 计算样本协方差和总体协方差：

$$S_1=\sum_{a=1}^{n_i}(X_{a1}-\bar{X}_1)(X_{a1}-\bar{X}_1)'=\begin{pmatrix}36.228 & 56.022 & 448.74\\56.022 & 344.228 & -252.24\\448.74 & -252.24 & 12987.2\end{pmatrix}$$

$$S_2=\sum_{a=1}^{n_i}(X_{a2}-\bar{X}_2)(X_{a2}-\bar{X}_2)'=\begin{pmatrix}86.812 & 117.682 & -4895.74\\117.682 & 188.672 & -11316.54\\-4895.74 & -11316.54 & 2087384.8\end{pmatrix}$$

$$\hat{\Sigma}=\frac{1}{n_1+n_2-2}(S_1+S_2)=\frac{1}{8}\begin{pmatrix}123.04 & 173.704 & -4447\\173.704 & 532.9 & -11568.78\\-4447 & -11568.78 & 2100372\end{pmatrix}$$

$$=\begin{pmatrix}15.38 & 21.713 & -555.875\\21.713 & 66.6125 & -14446.0975\\-555.875 & -1446.0975 & 262546.5\end{pmatrix}$$

$$\hat{\Sigma}^{-1}=\begin{pmatrix}0.120896 & -0.03845 & 0.0000442\\-0.03845 & 0.029278 & 0.0000799\\0.0000442 & 0.0000799 & 0.0000434\end{pmatrix}$$

3 求线性判别函数

令$a=\hat{\Sigma}^{-1}(\bar{X}_1-\bar{X}_2)$

$$W(X)=(X-\bar{X})'\hat{\Sigma}^{-1}(\bar{X}_1-\bar{X}_2)=a'(X-\bar{X})=0.6523x_1+0.0122x_2+0.00873x_3-87.1525$$

4 待判样本归类：

M国：

$$W(X)=0.6523\times68.5+0.0122\times79.3+0.00873\times1950-87.1525=-24.47899$$

判别到第二类。

N国：

$$W(X)=0.6523\times77.6+0.0122\times93.8+0.00873\times5233-87.1525=4.47899$$

判别到第一类。

判别分析
费歇尔判别

(2) Fisher判别：

1 建立判别函数：

$$\begin{pmatrix} c_1 \\ c_2 \\ c_3 \end{pmatrix} = S^{-1}\begin{pmatrix} d_1 \\ d_2 \\ d_3 \end{pmatrix} = \frac{1}{8}\hat{\Sigma}^{-1}\left(\bar{X}_1 - \bar{X}_2\right) = \frac{1}{8}a = \begin{pmatrix} 0.0815375 \\ 0.001525 \\ 0.00109125 \end{pmatrix}$$

$$y = 0.0815375x_1 + 0.001525x_2 + 0.00109125x_3$$

2 计算临界值y_0：

$\bar{y}_1 = 12.1615$　$\bar{y}_2 = 9.6266$　　临界值$y_0 = \dfrac{n_1\bar{y}_1 + n_2\bar{y}_2}{n_1 + n_2} = 10.8941$

3 判别分析：

因$\bar{y}_1 > \bar{y}_2$　　$y_M = 7.8342 < y_0$

将M国判别到第二类。

$y_N = 12.1809 > y_0$

将N国判别到第一类。

主成分分析

提出问题
Question

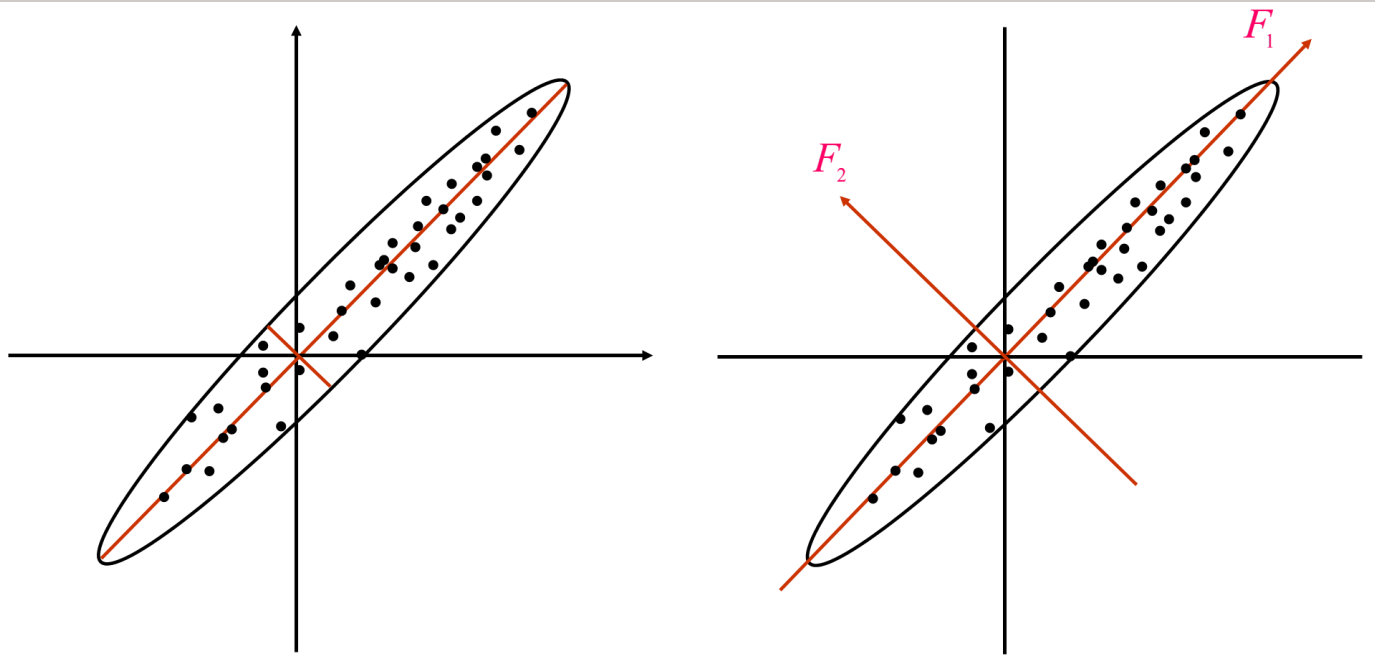
在处理信息时，当两个变量之间有一定相关关系时，可以解释为这两个变量反映出的信息有一定的重叠。而变量之间信息的高度重叠和高度相关会给统计方法的应用带来许多障碍。

解决问题
Answer

为了解决这些问题，最简单和最直接的解决方案是削减变量的个数，但这必然又会导致信息丢失和信息不完整等问题的产生。为此，人们希望探索。主成分分析，既能大大减少参与数据建模的变量个数，同时也不会造成信息的大量丢失。是一种能够有效降低变量维数，并已得到广泛应用的分析方法。

主成分分析
基本原理

在主成分计算时要进行旋转变换（如图2-32所示），其目的是为了使得n个样本点在F1轴方向上的离散程度最大，即F1的方差最大，变量F1代表了原始数据的绝大部分信息，也就是主成分方向。然后在二维空间中取和F1方向正交的方向，就是F2的方向。数据在F1上的投影代表了原始数据的绝大部分信息，即使不考虑F2，信息损失也不多。F1与F2除起了浓缩作用外，还具有不相关性。椭圆的长短轴相差得越大，降维也越有意义。



(a) 原始数据分布　　(b) 旋转坐标轴

图2-32 主成分分析旋转变换

主成分分析
计算步骤

(1)计算协方差矩阵
计算样本数据的协方差矩阵:

$$\Sigma = (S_{ij})p \times p$$

其中:

$$s_{ij} = \frac{1}{n-1}\sum_{k=1}^{n}(x_{ki} - \overline{x}_i)(x_{kj} - \overline{x}_j)$$

$$i, j = 1, 2, \ldots, p$$

01

(2)求出的特征值 λ_i 及相应的正交化单位特征向量 a_i

λ_i 的前m个较大的特征值 $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_m > 0$,就是前m个主成分对应的方差， λ_i 对应的单位特征向量 a_i 就是主成分 F_i 的关于原变量的系数，则原变量的第 i 个主成分F_i 为: $F_i = a_i' X$

主成分的方差（信息）贡献率用来反映信息量的大小， α_i 为:

$$\alpha_i = \lambda_i / \sum_{i=1}^{m}\lambda_i$$

02

（3）选择主成分
最终要选择几个主成分，即 $F_1, F_2, \ldots, F_m$ 中m的确定是通过方差（信息）累计贡献率 $G(m)$ 来确定

$$G(m) = \sum_{i=1}^{m}\lambda_i / \sum_{k=1}^{p}\lambda_k$$

当累积贡献率大于80%时，就认为能足够反映原来变量的信息了，对应的m就是抽取的前m个主成分。

03

主成分分析
计算步骤

(4)计算主成分载荷

主成分载荷是反映主成分F_i与原变量X_j之间的相互关联程度，原来变量$X_j\ (j=1,2,...,p)$在诸主成分$F_i\ (i=1,2,...,m)$上的荷载

$$l_{ij}\ (i=1,2,...,m;\ j=1,2,...,p)$$

$$l(Z_i,X_j)=\sqrt{\lambda_i}a_{ij}(i=1,2,\cdots,m;\ j=1,2,\cdots,p)$$

(5)计算主成分得分

计算样品在m个主成分上的得分：

$$F_i=a_{1i}X_1+a_{2i}X_2+...+a_{pi}X_p\ ,\ i=1,2,...,m$$

实际应用时，指标的量纲往往不同，所以在主成分计算之前应先消除量纲的影响。消除数据的量纲有很多方法，常用方法是将原始数据标准化，即做如下数据变换：

$$x_{ij}^{*}=\frac{x_{ij}-\overline{x}_j}{s_j}\qquad i=1,2,...,n;\ j=1,2,...,p$$

其中：$\overline{x}_j=\frac{1}{n}\sum_{i=1}^{n}x_{ij}$ ，$s_j^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_{ij}-\overline{x}_j)^2$

主成分分析
计算步骤

① 计算相关系数矩阵:

$$R = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1p} \\ r_{21} & r_{22} & \cdots & r_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ r_{p1} & r_{p2} & \cdots & r_{pp} \end{bmatrix}$$

r_{ij}（i，j=1，2，…，p）为原变量x_i与x_j的相关系数，$r_{ij}=r_{ji}$，其计算公式为：

$$R = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1p} \\ r_{21} & r_{22} & \cdots & r_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ r_{p1} & r_{p2} & \cdots & r_{pp} \end{bmatrix}$$

1

2

② 求出相关系数矩阵的特征值 λ_i 及相应的正交化单位特征向量 ：

首先解特征方程 $|\lambda I - R| = 0$ ，求出特征值，并使其按大小顺序排列；再分别求出对应于特征值 λ_i 的特征向量 $\sim -\begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1p} \\ r_{21} & r_{22} & \cdots & r_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ r_{p1} & r_{p2} & \cdots & r_{pp} \end{bmatrix}$，要求，$\sum_{j=1}^{p} l_{ij}^2 = 1$ 其中 l_{ij} 表示向量 [illegible] 的第 j 个分量。

③ 计算主成分贡献率及累计贡献率:

贡献率：$\dfrac{\lambda_i}{\sum_{k=1}^{p} \lambda_k} \quad (i = 1, 2, \cdots, p)$

累计贡献率：$\dfrac{\sum_{k=1}^{i} \lambda_k}{\sum_{k=1}^{p} \lambda_k} \quad (i = 1, 2, \cdots, p)$

一般取累计贡献率达80—95%的特征值所对应的第一、第二•••第m（m≤p）个主成分。

3

4

④ 计算主成分得分：

$$Z = \begin{bmatrix} l_{11} & l_{12} & \cdots & l_{1p} \\ l_{21} & l_{22} & \cdots & l_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ l_{n1} & l_{n2} & \cdots & l_{np} \end{bmatrix} \begin{bmatrix} X_1 \\ x_2 \\ . \\ x_p \end{bmatrix}$$

主成分分析实例

例:右表是某农业生态经济系统各区域单元的有关数据，对此进行主成分分析。

x_1：人口密度(人/)km^2)	x_2：人均耕地面积(ha)	x_3：森林覆盖率(%)	x_4：农民人均纯收入(元/人)	x_5：人均粮食产量 (kg/人)	x_6：经济作物占农作物播面比例(%)	x_7：耕地占土地面积比率(%)	x_8：果园与林地面积之比(%)	x_9：灌溉田占耕地面积之比(%)
363.91	0.35	16.10	192.11	295.34	26.72	18.49	2.23	26.26
141.50	1.68	24.30	1752.35	452.26	32.31	14.46	1.46	27.07
100.70	1.07	65.60	1181.54	270.12	18.27	0.16	7.47	12.49
143.74	1.34	33.21	1436.12	354.26	17.49	11.81	1.89	17.53
131.41	1.62	16.61	1405.09	586.59	40.68	14.40	0.30	22.93
68.34	2.03	76.20	1540.29	216.39	8.13	4.07	0.01	4.86
95.42	0.80	71.11	926.35	291.52	8.14	4.06	0.01	4.86
62.90	1.65	73.31	1501.24	225.25	18.35	2.65	0.03	3.20
86.62	0.84	68.90	897.36	196.37	16.86	5.18	0.06	6.17
91.39	0.81	66.50	911.24	226.51	18.28	5.64	0.08	4.48
76.91	0.86	50.30	103.52	217.09	19.79	4.88	0.00	6.17
51.27	1.04	64.61	968.33	181.38	4.01	4.07	0.02	5.40
68.83	0.84	62.80	957.14	194.04	9.11	4.48	0.00	5.79
77.30	0.62	60.10	824.37	188.09	19.41	5.72	5.06	8.41
76.95	1.02	68.00	1255.42	211.55	11.10	3.13	0.01	3.43
99.27	0.65	60.70	1251.03	220.91	4.38	4.62	0.01	5.59
118.51	0.66	63.30	1246.47	242.16	10.71	6.05	0.15	8.70
141.47	0.74	54.21	814.21	193.46	11.42	6.44	0.01	12.95
137.76	0.60	55.90	1124.05	228.44	9.52	7.88	0.07	12.65
117.61	1.25	54.50	805.67	175.23	18.11	5.79	0.05	8.46
122.78	0.73	49.10	1313.11	236.29	26.72	7.16	0.09	10.08

主成分分析
实例

步骤如下：

01 将表2-0中的数据作标准差标准化处理，计算相关系数，得到相关系数矩阵（见表2-10）。

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
x_1	1	-0.327	-0.714	-0.336	0.309	0.408	0.79	0.156	0.744
x_2	-0.327	1	-0.035	0.644	0.42	0.255	0.009	-0.078	0.094
x_3	-0.714	-0.035	1	0.07	-0.74	-0.755	-0.93	-0.109	-0.924
x_4	-0.336	0.644	0.07	1	0.383	0.069	-0.046	-0.031	0.073
x_5	0.309	0.42	-0.74	0.383	1	0.734	0.672	0.098	0.747
x_6	0.408	0.255	-0.755	0.069	0.734	1	0.658	0.222	0.707
x_7	0.79	0.009	-0.93	-0.046	0.672	0.658	1	-0.03	0.89
x_8	0.156	-0.078	-0.109	-0.031	0.098	0.222	-0.03	1	0.29
x_9	0.744	0.094	-0.924	0.073	0.747	0.707	0.89	0.29	1

表2-10 相关系数矩阵

主成分分析
实例

步骤如下：

02 由相关系数矩阵计算特征值，以及各个主成分的贡献率与累计贡献率，由表2-11可知，第一，第二，第三主成分的累计贡献率已高达86.596%（大于85%），故只需要求出第一、第二、第三主成分z1，z2，z3即可。

主成分	特征值	贡献率(%)	累积贡献率(%)
z_1	4.661	51.791	51.791
z_2	2.089	23.216	75.007
z_3	1.043	11.589	86.596
z_4	0.507	5.638	92.234
z_5	0.315	3.502	95.736
z_6	0.193	2.14	97.876
z_7	0.114	1.271	99.147
z_8	0.0453	0.504	99.65
z_9	0.0315	0.35	100

特征值及主成分贡献率

主成分分析
实例

步骤如下：

03 对于特征值为4.6610、2.0890、1.0430的主成分，分别求出其特征向量l1，l2，l3。

	l_1	l_2	l_3	占方差的百分数（%）
x_1	0.739	-0.532	-0.0061	82.918
x_2	0.123	0.887	-0.0028	80.191
x_3	-0.964	0.0096	0.0095	92.948
x_4	0.0042	0.868	0.0037	75.346
x_5	0.813	0.444	-0.0011	85.811
x_6	0.819	0.179	0.125	71.843
x_7	0.933	-0.133	-0.251	95.118
x_8	0.197	-0.1	0.97	98.971
x_9	0.964	-0.0025	0.0092	92.939

表2-12　单位特征向量

主成分分析
实例

步骤如下：

04 计算主成分得分： 例如第一主成分z_1的得分：

$F_1=0.739*X_1+0.123*X_2-0.964*X_3+0.0042*X_4+0.813*X_5+0.819*X_6+0.933*X_7+0.197*X_8+0.964*X_9$。

得到主成分得分结果：

F_1	F_2	F_3
559.27	106.78	-0.5773
522.74	1651.9	7.41833
264.5	1096	13.6878
411.03	1330.2	5.5781
633.02	1417	5.8808
174.82	1400.1	5.81351
258.1	885.04	3.25148
186.04	1374.7	7.30744
185.71	823.73	4.14659
216.08	846.77	4.17781
212.09	149.45	1.45721
139.55	895.47	3.22571
169.43	882.7	3.57088
185.98	761.26	8.91186
184.02	1145.8	5.22869
213.06	1132.5	3.80677
251.58	1128.7	4.2706
240.93	719.69	2.39085
265.8	1005.5	2.99208
208.77	718.63	3.5197
279.41	1184.3	6.03884

主成分分析
实例

结果分析：

1. 第一主成分z1与x1，x5，x6，x7，x9呈显出较强的正相关，与x3呈显出较强的负相关，而这几个变量则综合反映了生态经济结构状况，因此可以认为第一主成分z1是生态经济结构的代表。

2. 第三主成分z3，与x8呈显出的正相关程度最高，其次是x6，而与x7呈负相关，因此可以认为第三主成分在一定程度上代表了农业经济结构。

3. 第二主成分z_2与x_2，x_4，x_5呈显出较强的正相关，与x_1呈显出较强的负相关，其中，除了x_1为人口总数外，x_2，x_4，x_5都反映了人均占有资源量的情况，因此可以认为第二主成分z_2代表了人均资源量。

4. 显然，用三个主成分z1、z2、z3代替原来9个变量（x1，x2，…，x9），描述农业生态经济系统，可以使问题更进一步简化、明了。

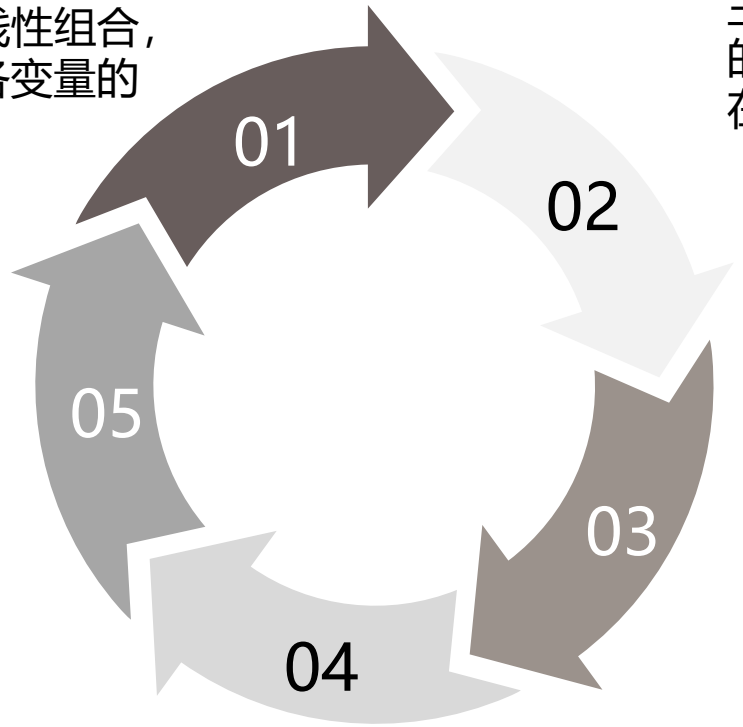
主成分分析与因子分析的区别

主成分分析
区别

因子分析是研究如何以最少的信息丢失，将众多原始变量浓缩成少数几个因子变量，以及如何使因子变量具有较强的可解释性的一种多元统计分析方法。有时易与主成分分析方法混淆，二者的主要区别

01 因子分析中是把变量表示成各因子的线性组合，而主成分分析中则是把主成分表示成各变量的线性组合。

02 主成分分析的重点在于解释各变量的总方差，而因子分析则把重点放在解释各变量之间的协方差。



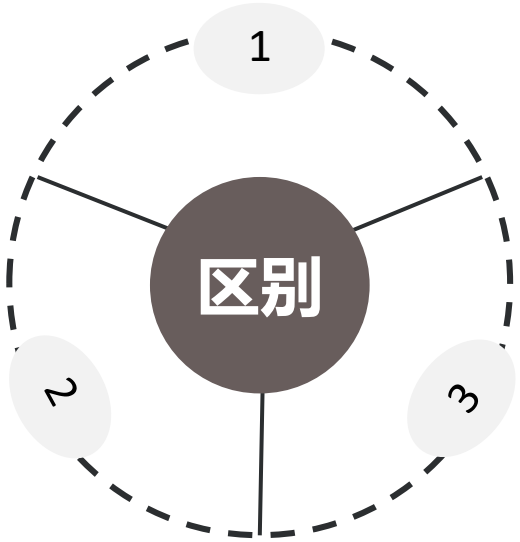
03 主成分分析中不需要有假设（assumptions），因子分析则需要一些假设。

04 主成分分析中，当给定的协方差矩阵或者相关矩阵的特征值唯一，主成分一般是独特的；而因子分析中因子不是独特的，可以旋转得到不同的因子。

05 在因子分析中，因子个数需要分析者指定，而指定的因子数量不同而结果不同。在主成分分析中，成分的数量是一定的，一般有几个变量就有几个主成分。

主成分分析
区别

大致说来，当需要寻找潜在的因子，并对这些因子进行解释的时候，更加倾向于使用因子分析，并且借助旋转技术帮助更好解释。



而如果想把现有的变量变成少数几个新的变量（新的变量几乎带有原来所有变量的信息）来进入后续的分析，则可以使用主成分分析。。

在算法上，主成分分析和因子分析很类似，不过在因子分析中所采用的协方差矩阵的对角元素不再是变量的方差，而是和变量对应的共同度（变量方差中被各因子所解释的部分）。

Thank you