## 数据探索分析-单因子分析

### 1.集中趋势、离中趋势

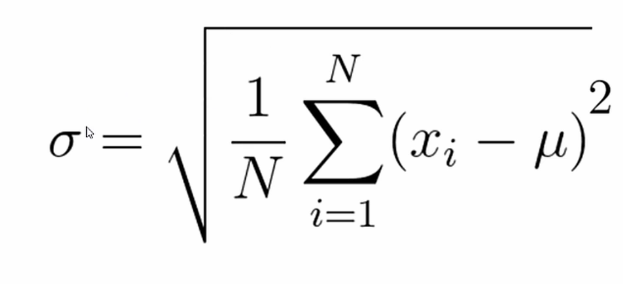
#### 1.1集中趋势：

均值、中位数、众数、分位数（将数据从小到大排序，按照等间隔切分）

常见的四分位计数方法：如分位（n+1）\*0.25、（n+1）\*0.5、（n+1）\*0.75

#### 1.2 离中趋势：数据离散程度的衡量

标准差（方差开平方）、方差：



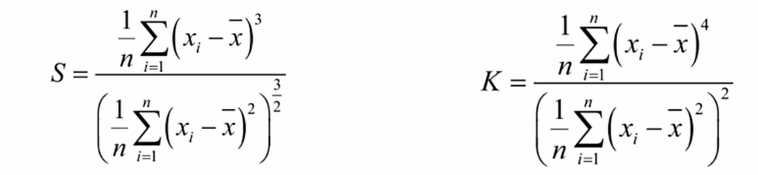
值越大，说明数据越离散。

注意：正态分布的离中趋势

### 2.数据分布：

#### 2.1偏态与峰度

偏态系数S与峰态系数K：

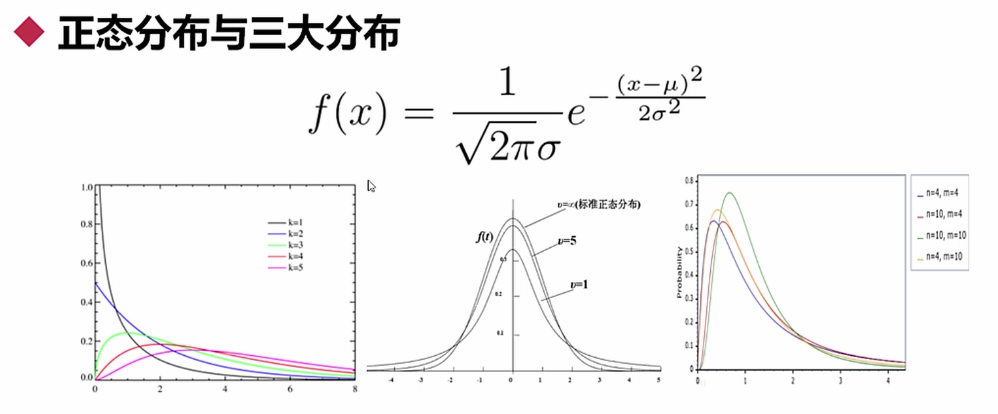


偏态系数S：平均值的偏差；小于0，平均值偏小，即大部分数据的值小于平均值，称为负偏。

峰态系数K：

注意：正态分布的峰态系数为3，如果一个特征数据的峰态系数与3相差2以上（小于1或大于5），那么可以认为该数据不具备正态分布特点。

#### 2.2分布概率：



T分布，卡方分布，F分布

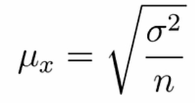
（注意应用）

### 3.抽样理论www

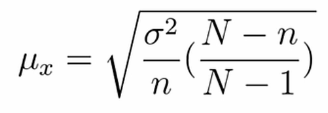
#### 3.1抽样误差与精度

##### 3.1.1抽样平均误差计算公式：

重复抽样：



不重复抽样：



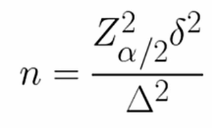
n：抽样数量

N:总体数量

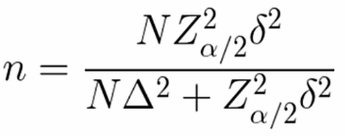
总体方差；

##### 3.1.2估计总体时抽样数目的确定：

重复抽样：



不重复抽样：

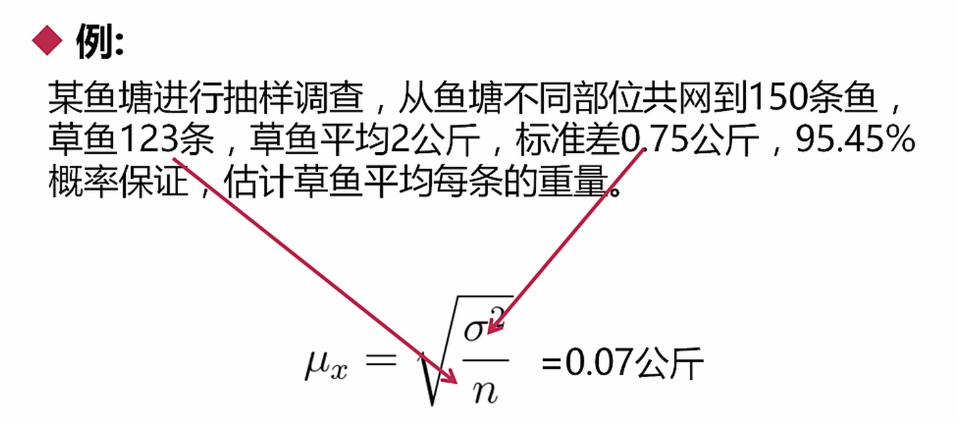


Z:概率保证

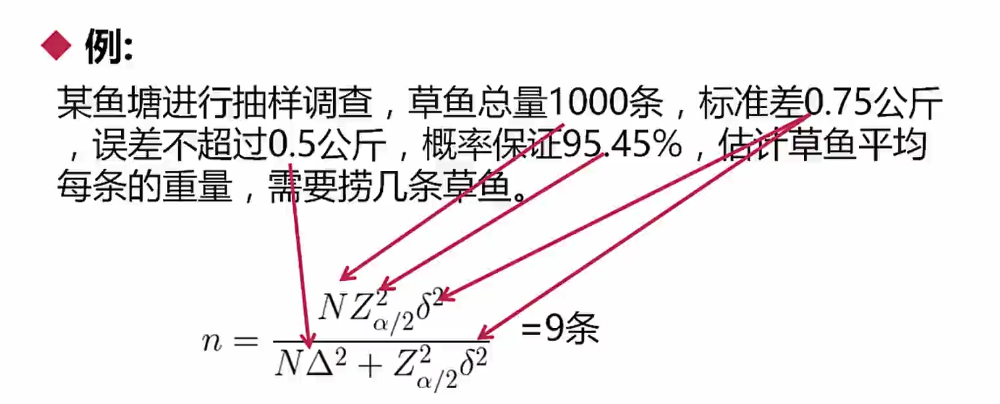
△：总体方差

需控制的方差

##### 举例说明：



故草鱼平均每条的重量为：2+\_2\*0.07公斤



### 4.Python实现数据探索分析，参考test1\_data\_explore.ipy

### 5.数据分类

定类（类别）：根据事务离散、无差别属性进行的分类；如性别，民族，没有哪两个民族直接距离更大或者更小。

定序（顺序）：可以界定数据的大小，但是不能测试差值；如工资的高，中，低，工资具体差值无法用具体值衡量。

定距（间隔）：可以界定数据大小的同时，可测定差值，但无绝对零点；如摄氏温度，没有20度比10度热一倍的说法。

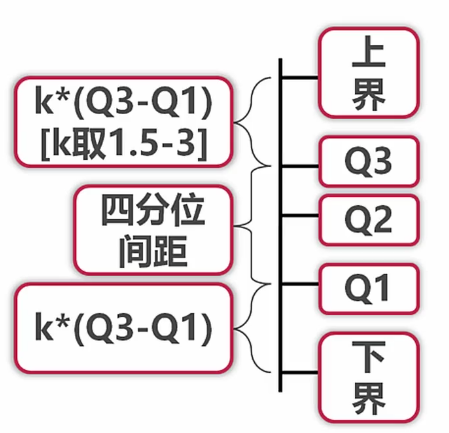
定比（比率）：可以界定数据大小，可测定差值，有绝对零点；如身高体重，做乘除比率是有一定意义的。

### 6.单属性分析

#### 6.1异常值分析：

连续异常值：

取值超出界定范围的异常值；（如工资，大多数人工资水平不高，少数人工资非常高，会导致平均工资偏高）



离散异常值：

离散属性定义范围外的所有值均为异常值；（性别中除了男女之外的值，如空值。这样的值通常会作为一个新的类别进行特殊处理）

知识异常值：

在限定知识与常识范围外的所有值均为异常值；（如身高10米）

#### 6.2对比分析：

绝对数比较：如身高，工资

相对数比较：结构（考试通过率，），比例（产业所占比例），比较（不同互联网公司的薪资水平），动态（有时间概念，如用户数量的增速），强度（gdp靠前，人均gdp靠后，人口密度，粮食亩产）

对比分析维度：

时间：同比（同一个时间：和去年同期（六月份）比较），环比（前后一段时间，和今年五月份比较）

空间：公司中不同部门进行比较

经验与计划：（经验）失业率的比较（失业率过高可能会产生暴乱），（计划）工作排期与实际工作对比

#### 6.3结构分析：

部分：总体

静态：当前产业结构，产业结构分布是否均衡；

动态：分析产业结构变化趋势，反应国家性质变化，产业转型；

#### 6.4分布分析：

直接获得概率分布：

是不是正态分布：生活中最常见的分布

怎么判断一个数据是不是服从正态分布：偏态（1）和峰态（3）系数来确定

检验估计：？？？

极大似然：？？？

哪个分布下极大似然越大，则越接近这个分布

### 单属性分析python实现：test2\_data\_explore.ipy

### 可视化分析：test3\_data\_explore.ipy

## 数据探索分析-多因子与复合分析

属性与属性之间的关联关系分析探索

### 1.假设检验

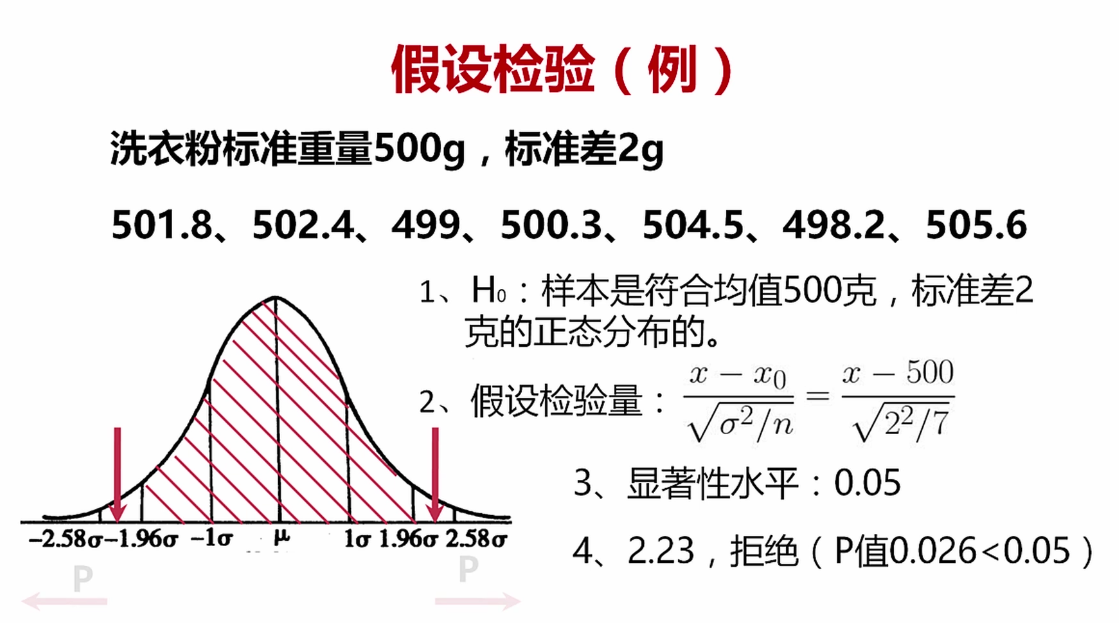
1.1作出假设：根据假设进行检验，进而推断。

建立原假设与反命题进行对比检验。

1.2选择检验统计量：

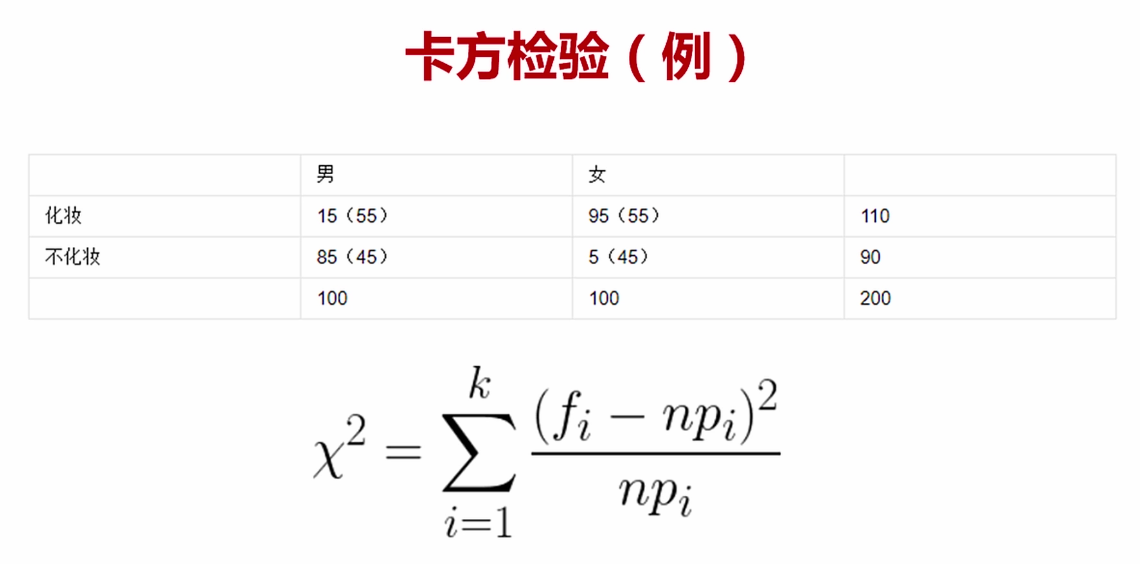
1.3根据显著性水平（一般为0.05），确定拒绝域（对立为：接受域）。

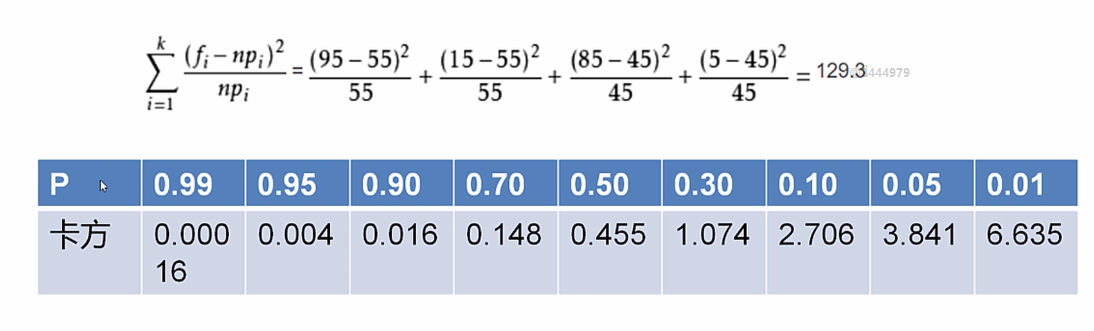
1.4计算p值或样本统计值，作出判断。



### 2.卡方检验

检验：性别与化妆是否有关系





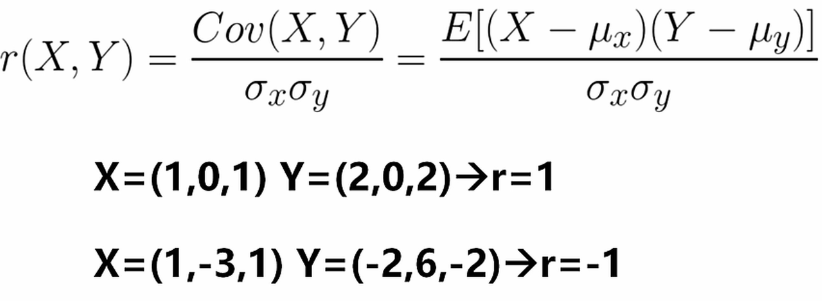
拒绝原假设，性别与化妆与否是有较强的关系的。

### 3.方差检验（F检验）

### 4.相关系数

正相关，负相关，无相关。

#### Pearson相关系数：

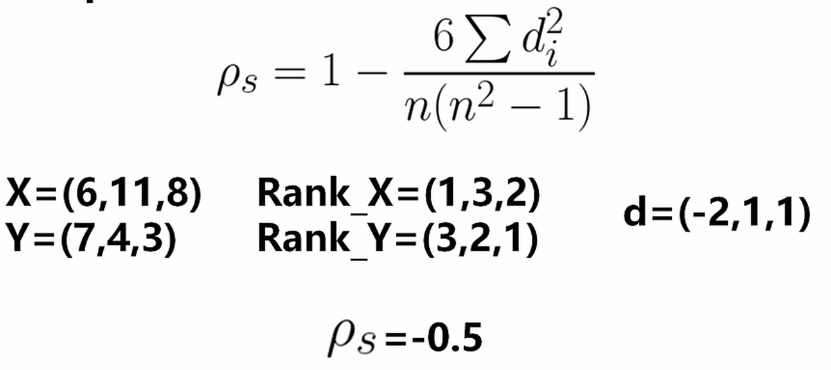


分子：两组数据的协方差；

分母：两组数据的标准差的积（归一化作用）。

#### Spearman相关系数：

跟名次差有关，跟具体的数值差相关不大。



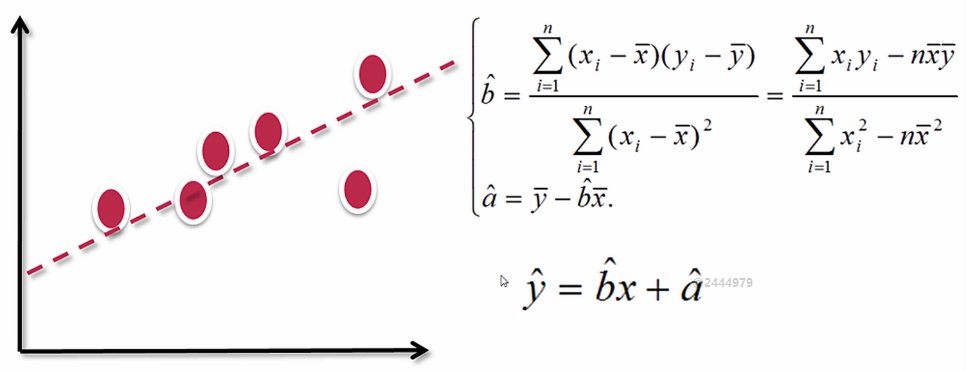
n:每组数据的数量；

d:每组数据的名次差：

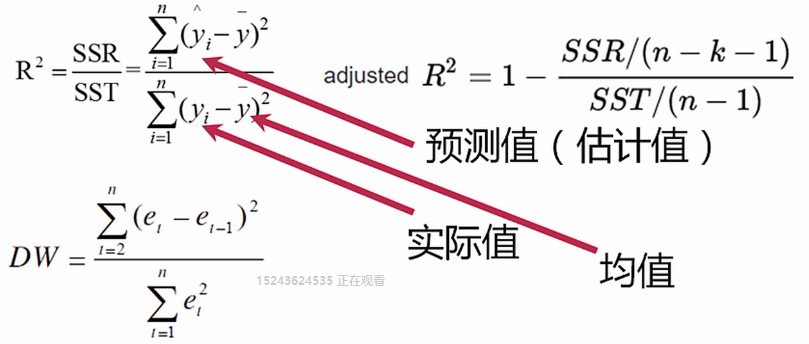
### 5.线性回归

回归：确定两种或两种以上变量间相互依赖的定量关系的一种统计分析方法。

最常用的线性回归问题解决方法：最小二乘法：最小化误差的平方。



关键指标：决定系数、残差不相关（DW检验）



决定系数：越接近1，回归效果越好；越接近0，回归效果越差；

多个变量相关性：DW:取值0-4

好的线性回归，残差接近2，残差不相关；接近4，为残差正相关；接近0，残差负相关。

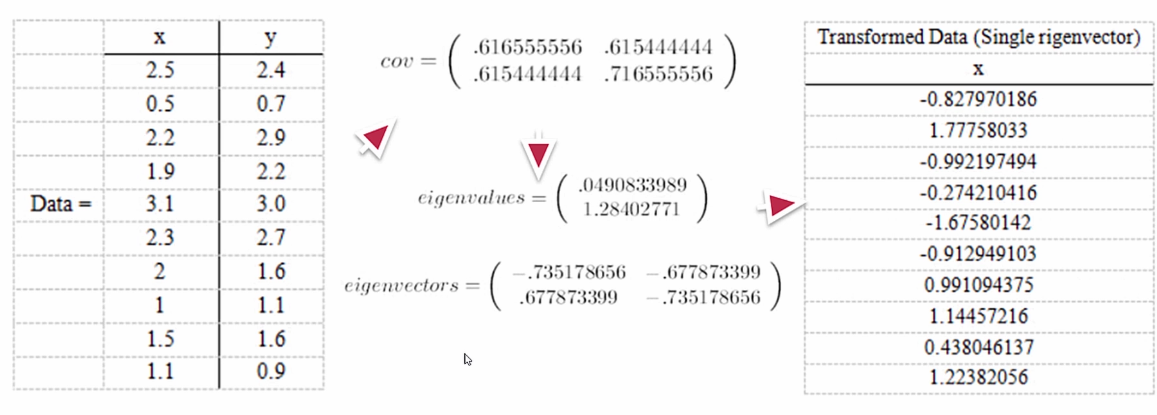
### 6.主成分分析（PCA）

6.1求特征协方差矩阵；

6.2求协方差的特征值和特征向量；

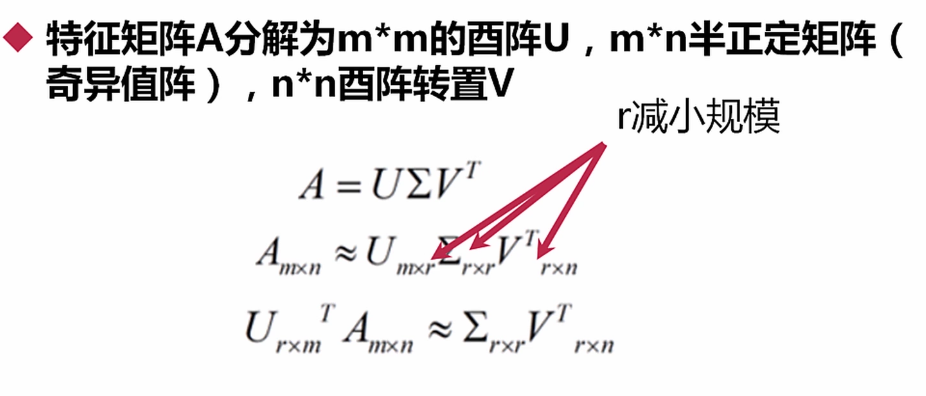
6.3将特征值按照从大到小的顺序排列，选择其中最大的k个；

6.4将样本点投影到选取的特征向量上。



达到降维的功能。

奇异值分解（SVD）:



### 7.编码实现，参考test4\_data\_explore.ipy

### 8.价差分析方法与实现，参考

### 9.分组分析方法与实现，

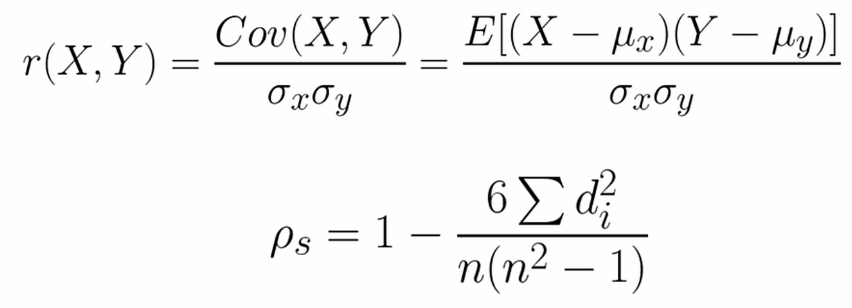
钻取：改变维的层次，变换分析的粒度。

分为：向上钻取（汇总），向下钻取（细分）

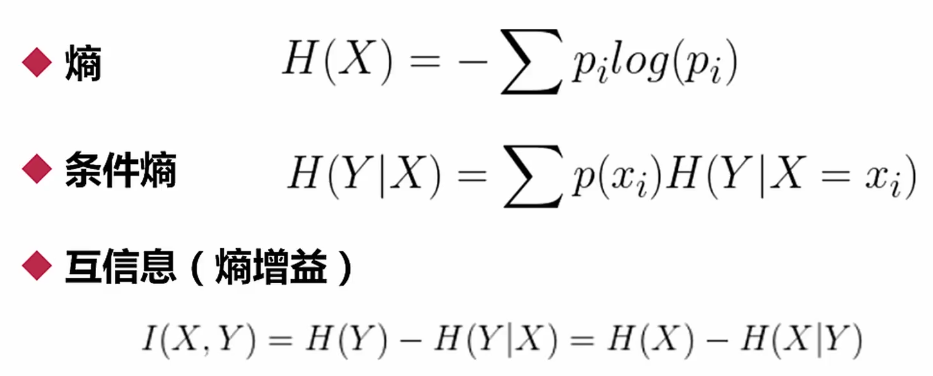
连续型属性需进行分组：



### 10.相关分析与实现，



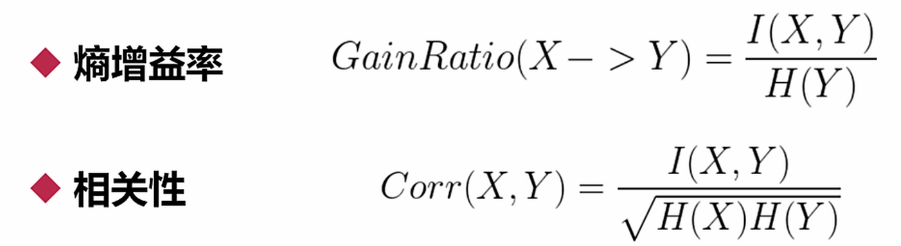
离散属性相关性分析：



熵：衡量不确定性的一个值。越接近0，说明不确定性越小。（注意：样本越多，分布越均匀，则样本该属性的熵越大）

条件熵：在x条件下，y的不确定性。（相对熵是有减少的）

互信息（熵增益）、信息增益：类别越多的值，信息增益的值的范围无法界定，难以参考对比。



熵增益率：具有不对称性。

相关性：解决不对称问题。

### 11.因子分析与实现（成分分析），

探索性因子分析：主成分分析

验证性因子分析：假设检验、卡方检验……

### 12.总结：

