**数据获取-->探索分析与可视化-->预处理理论-->分析建模-->模型评估**

**数据获取**：1数据仓库-->每次用户交互信息，文件和日志也可以，但是不方便查找

2解析网页，接口，文件信息（爬取网页，解析页面等）

3填写、日志、埋点（用户填写信息，特定流程的信息记录，记录叫交互信息）

4计算（通过已有数据计算生成衍生数据，如统计报表某些项目，投入产出比等）

数据库vs仓库

1数据库面向业务存储，仓库面向主题存储（主题：较高层次上对分析对象数据的一个完整且一致的描述），如购买图书就划为购买主题，什么时候买的，买的什么等等

2数据库针对应用-->OLTP，数据仓库针对分析-->OLAP

3数据库组织规范，仓库可能冗余，数据量大，变化大

数据集获取

数据竞赛网站（kaggle & 天池）

数据集网站（imageNet / Open images）

各领域统计信息（统计局、政府机构、公司财报等）

[www.kaggle.com](http://www.kaggle.com)

<https://tianchi.aliyun.com/>

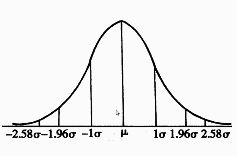
image.net.org/index

**探索分析与可视化**

**集中趋势**：均值、中位数、众数、分为数

四分位数计算方法：Q1位置 = (n+1) \* 0.25

Q2位置 = (n+1) \* 0.5

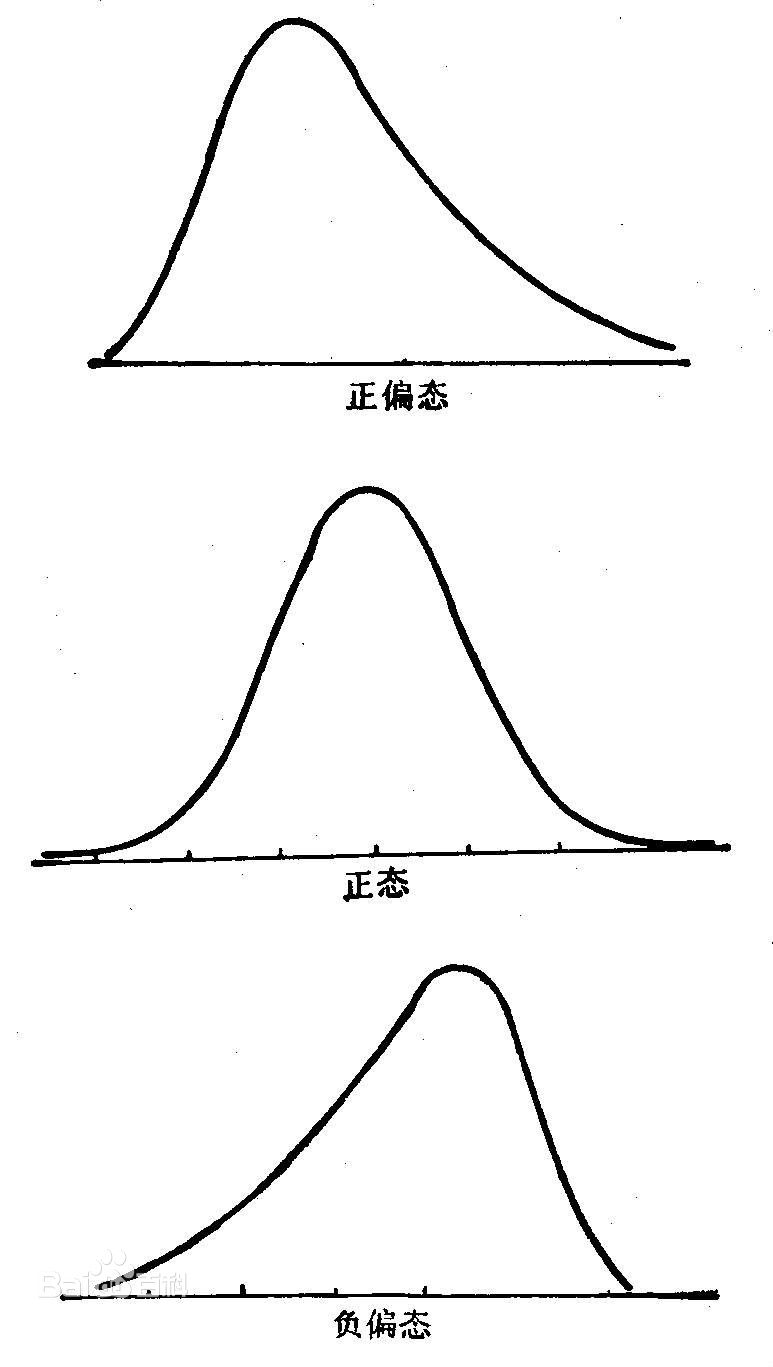
Q3位置 = (n+1) \* 0.75

**离中趋势**：标准差，方差

σ =



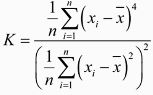
正太分布方差特性

**数据分布**：偏态与峰态、正态分布与三大分布

IMG_256

偏态系数

例：1,2,3,4,10 均值4>中位数3，正偏态



峰态分布峰态系数一般为3

若峰态系数与3相差大于2一般认为不是正态分布

峰度系数是用来反映[频数分布](https://baike.baidu.com/item/%E9%A2%91%E6%95%B0%E5%88%86%E5%B8%83" \t "https://baike.baidu.com/item/%E5%B3%B0%E5%BA%A6%E7%B3%BB%E6%95%B0/_blank)曲线顶端尖峭或扁平程度

的指标在金融中，若某分布与正态分布有相同的方差，

但是峰度系数大于3，则呈现“尖峰厚尾”形态，

峰度更高，两段的尾部更厚，也就是极值更多的意思。

卡方分布：几个变量都是标准正态分布，其平方和组成的分布

T分布：正态分布变量/卡方分布变量，常用来根据小样本来估计成正态分布，且方差未知的总体的均值

F分布： 卡方分布变量/卡方分布变量

独立同分部的中心极限定理：设随机变量X1，X2，......Xn，......独立同分布，并且具有有限的数学期望和[方差](https://baike.baidu.com/item/%E6%96%B9%E5%B7%AE" \t "https://baike.baidu.com/item/%E4%B8%AD%E5%BF%83%E6%9E%81%E9%99%90%E5%AE%9A%E7%90%86/_blank)：E(Xi)=μ，D(Xi)=σ20(k=1,2....)，则对任意x，分布函数

IMG_256

该定理说明，当n很大时，随机变量

IMG_256

近似地服从标准正态分布N(0，1)

**抽样理论**：抽样误差、抽样精度

分类分层抽样

重复抽样：误差 = n是抽样数量



不重复抽样：误差=

指定概率下抽样数目的确定：

数据分类

定类（类别）：根据事物离散、无差别属性进行分类（类别间无差距，如男、女）

定序（顺序）：可以界定数据大小，不能测定插值（有差值，如高、中、低，不能衡量差值多少）

定距（间隔）：可以界定数据大小的同时，可测差值，但无绝对零点（乘除无意义，如20度温度是10度的二倍，是不对的）

定比（比率）：可界定数据大小，可测定差值，有绝对零点（如身高，体重）

单属性分析

异常值分析：连续异常值（四分位差值的1.5~3倍，对应一般异常和重大异常），离散异常值（如空值，离散范围外的值），常识异常值（如身高5米）

对比分析：绝对数与相对数，时间、空间、理论维度比较

结构分析：各组成部分的分布与规律

分布分析：数据分布频率的显式分析

