**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/>

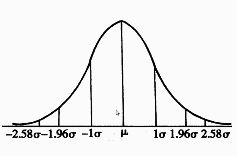
image.net.org/index

**单因子探索分析与可视化**

1、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

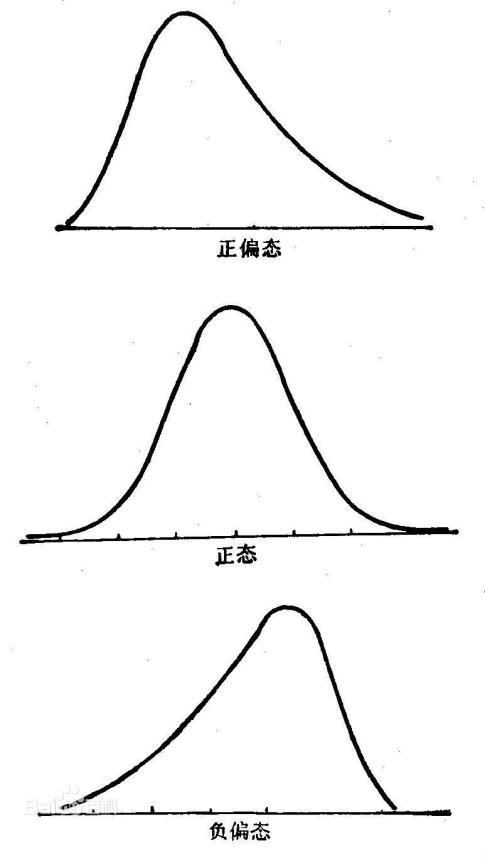
Q3位置 = (n+1) \* 0.75

离中趋势：标准差，方差

σ =



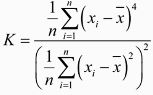
正太分布方差特性

2、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)

3、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）

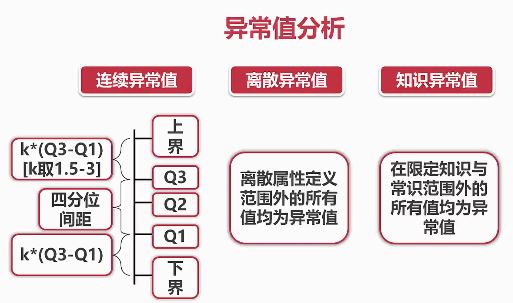
单属性分析

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



pd聚合

def fun(attr):

return attr.max()-attr.min()

Pdg.agg(fun)

number\_pro = df['number\_project']

number\_pro[number\_pro.isnull()]

number\_pro.value\_counts()

number\_pro.value\_counts(normalize=True).sort\_index()

salary = df['salary']

salary2 = salary.where(salary=="nme",other=False)

salary[salary2.where(salary2==False,other=True)]

Np.histogram(salary,bins=np.arrage(0.0,1.1,0.1)

可视化：

For x,y in zip(zuobiao,zhi):

Plt.text(x,y,y)

Import seaborn as sns

Sns.set\_style(style=”wihtegrid”)

Sns.set\_context(context=”poster”)

Sns.set\_palette（[sns.color\_palette(“RdBu”,n\_colors=7)[5]]） #调色板

Sns.countplot(x=”salary”,hue=”department”,,data=df) 柱状图

直方图有意义的是面积，x轴是区间

F = plt.figure() #定义画布

F.add\_subplot(1,3,1)#获取第一个画板

Sns.distplot(df[“satisfaction\_level”],bins=10]，kde=False(分布图),hist=False(直方图)) 直方图

Sns.pointplot(salary.index,salary) #折线图

Lbs = department.index

Explodes=[0.1 if i=”sales”else 0 for i in lbs]

Plt.pie(df[“department”].value\_coutns(normalize=True),explode=explodes,labels=lbs,autopct=”%1.1f%%”，colors=sns.color\_palette(“Reds) #饼图

折线图 sub\_d

**多因子分析**

1、假设检验：建立原假设H0和反命题H1

选择检验统计量

根据显著性水平（一般为0.05，人为设定），确定拒绝域

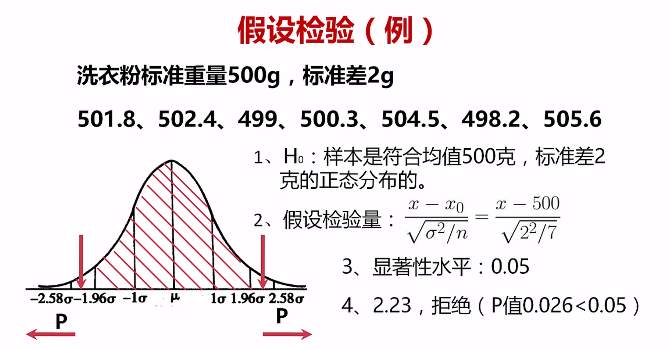
计算样本统计值，作出判断

例子：洗衣粉标准重量500g，标准差2g

抽检 501.8，502.4,499,500.3,504.5,498.2,505.6

假设符合均值500，标准差2的正态分布

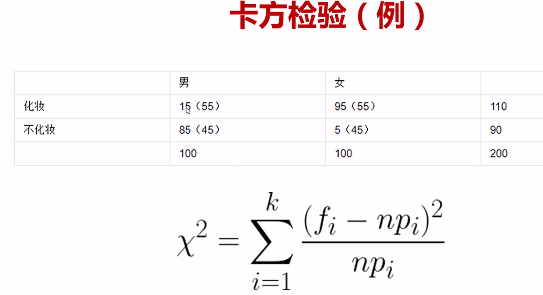
捕获假设检验量 ，显著性水平0.05（值越低越契合假设）

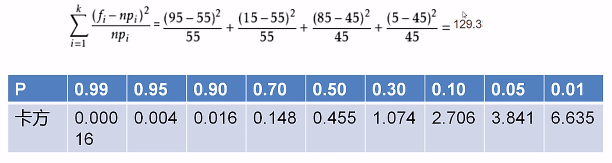


图中阴影部分为接受域，非阴影部分为拒绝域

2、卡方检验：

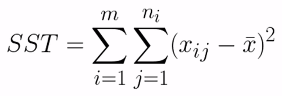
化妆与性别的关系，假设与性别无关



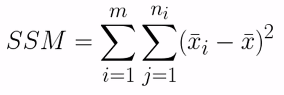


结果：0.05对应卡方值为3.841，计算值大于3.841，所以假设不成立

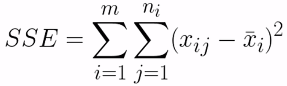
3、方差检验（因用到F分布，也叫F检验）：检验样本两两之间关系



总变差平方和：

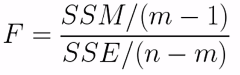


平均平方和，组间平方和

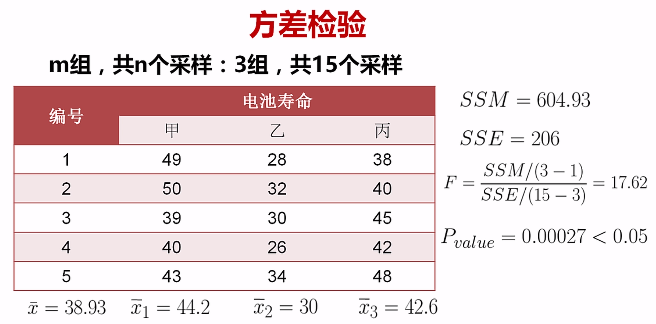


残差平方和，组内平方和

方差检验统计量F，做假设检验（F满足自由度(m-1,n-m)的F分布）



例子：三种电池寿命均值是否有差别



p值小于0.05，所以拒绝原假设，寿命均值是有差别的

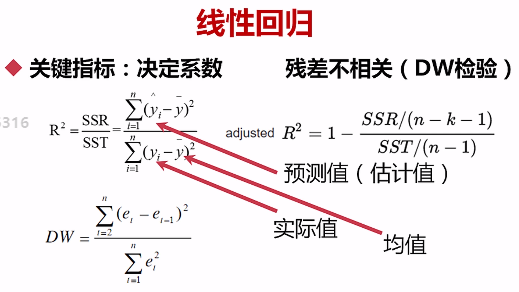
4、相关系数：

皮尔逊相关系数、Spearman相关系数（只跟排序有关，和具体差值无关，比较适用于相对比较）

5、线性回归：

常用解法：最小二乘法。最小化误差平方

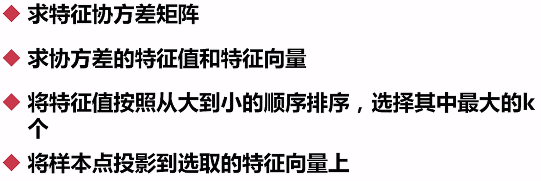
关键指标：决定系数，残差不相关（DW检验）



决定系数：接近1，回归效果好，接近0，效果不好

残差不相关：取值范围0~4，接近2不相关，4正相关，0负相关

6、主成分分析（PCA）



7、交叉分析

stats.ttest\_ind

sns.heatmap

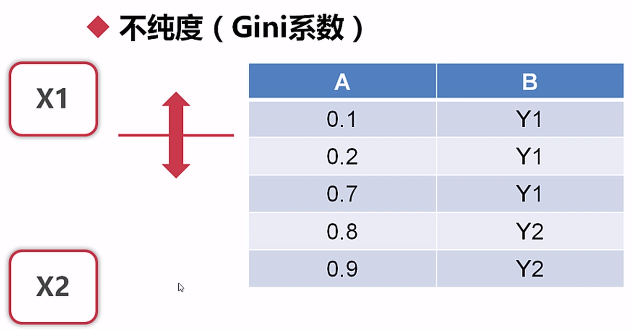
8、分组分析

分组与钻取

钻取：改变维的层次，变换分析力度

向下钻取：展开数据，查看细节，一门考试每班一个分组，知道每个班平均成绩，男生女生各自平均分就是向下钻取

向下钻取：汇总分组数据

连续属性分组：需要首先离散化，相同分组聚拢，不同分组分离，可以用聚类分组，不纯度计算分组

不纯度Gini系数：

A是属性，B是分类

连续属性通过排序属性值，滑动

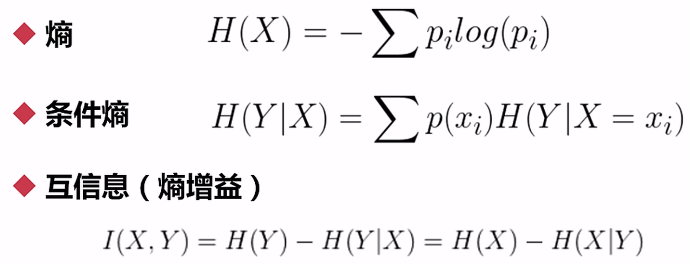
属性分割点计算对应基尼系数。

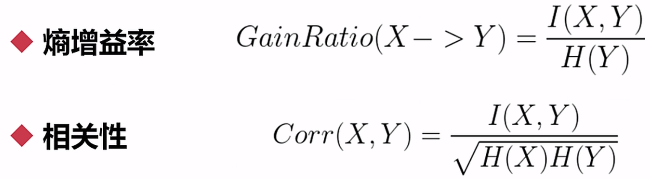
9、相关性分析

连续属性使用相关系数、Gini系数

二分属性可以使用Gini系数

多类离散属性：熵，编码转换为值使用皮尔逊相关系数（会有部分失真）





10、因子分析

**特征工程：**

1、特征使用：数据选择（sql表、文件、抓取获得）、数据可用性（数据需要实时更新，但是现实没有那么高的时效性）

特征获取：特征来源（多源、单源）、特征存储（将多源特征整合存储，如两张表整合成一张表，一个文件一个表整合成一张表等）

特征处理：数据清洗、特征预处理

特征监控：随时间推移数据越来越多，特征可能会变化，模型也需要进化。

数据清洗：数据样本采集（抽样） 样本要有代表性

样本比例不平衡时如何处理

尽量使用全量数据

异常值（空值处理）： isnull()/duplicated()

Drop()/dropna()/drop\_duplicated() 异常值不多时

Fillna() 异常值较多时

Series.intercept()

特征选择：

剔除与标注不相关或冗余的特征

数据规约的思路：

1、过滤思想--判断特征与标签的关联度，过滤掉关联性小的

连续-》连续 相关系数、假设检验

连续-》离散（二值） 相关系数、连续二值化（最小gini，最大熵切分）

连续-》离散（非二值） 相关系数（定序）

离散（二值）-》离散（二值） 相关系数，熵相关，F分值

离散-》离散（非二值） 熵相关，Gini，相关系数（定序）

2、包裹思想（REF）

特征集和---->构造简单模型，根据系数去掉弱特征---->余下特征重复过程，直到评价指标下降较大或低于阈值。

1. 嵌入思想

建立简单模型，如线性回归模型，通过正则化去掉系数较小的特征

特征变换：对指化、离散化、数据平滑、归一化（标准化）、数值化、正规化

特征降维：

特征衍生：