# 统计学中数据分析方法汇总

## Part1 描述统计

描述统计是通过图表或数学方法，对数据资料进行整理、分析，并对数据的分布状态、数字特征和随机变量之间关系进行估计和描述的方法。描述统计分为集中趋势分析和离中趋势分析和相关分析三大部分。

### 集中趋势分析

集中趋势分析主要靠平均数、中数、众数等统计指标来表示数据的集中趋势。例如被试的平均成绩多少？是正偏分布还是负偏分布？

### 离中趋势分析

离中趋势分析主要靠全距、四分差、平均差、方差（协方差：用来度量两个随机变量关系的统计量）、标准差等统计指标来研究数据的离中趋势。例如，我们想知道两个教学班的语文成绩中，哪个班级内的成绩分布更分散，就可以用两个班级的四分差或百分点来比较。

### 相关分析

相关分析探讨数据之间是否具有统计学上的关联性。这种关系既包括两个数据之间的单一相关关系——如年龄与个人领域空间之间的关系，也包括多个数据之间的多重相关关系——如年龄、抑郁症发生率、个人领域空间之间的关系；

既包括A大B就大(小)，A小B就小(大)的直线相关关系，也可以是复杂相关关系（A=Y-B*X）；既可以是A、B变量同时增大这种正相关关系，也可以是A变量增大时B变量减小这种负相关，还包括两变量共同变化的紧密程度——即相关系数。

实际上，相关关系唯一不研究的数据关系，就是数据协同变化的内在根据——即因果关系。获得相关系数有什么用呢？

简而言之，有了相关系数，就可以根据回归方程，进行A变量到B变量的估算，这就是所谓的回归分析，因此，相关分析是一种完整的统计研究方法，它贯穿于提出假设，数据研究，数据分析，数据研究的始终。

例如，我们想知道对监狱情景进行什么改造，可以降低囚徒的暴力倾向。我们就需要将不同的囚舍颜色基调、囚舍绿化程度、囚室人口密度、放风时间、探视时间进行排列组合，然后让每个囚室一种实验处理，然后用因素分析法找出与囚徒暴力倾向的相关系数最高的因素。假定这一因素为囚室人口密度，我们又要将被试随机分入不同人口密度的十几个囚室中生活，继而得到人口密度和暴力倾向两组变量（即我们讨论过的A、B两列变量）。

然后，我们将人口密度排入X轴，将暴力倾向分排入Y轴，获得了一个很有价值的图表，当某典狱长想知道，某囚舍扩建到N人/间囚室，暴力倾向能降低多少。我们可以当前人口密度和改建后人口密度带入相应的回归方程，算出扩建前的预期暴力倾向和扩建后的预期暴力倾向，两数据之差即典狱长想知道的结果。

## 推论统计

推论统计是统计学乃至于心理统计学中较为年轻的一部分内容。它以统计结果为依据，来证明或推翻某个命题。具体来说，就是通过分析样本与样本分布的差异，来估算样本与总体、同一样本的前后测成绩差异，样本与样本的成绩差距、总体与总体的成绩差距是否具有显著性差异。

例如，我们想研究教育背景是否会影响人的智力测验成绩。可以找100名24岁大学毕业生和100名24岁初中毕业生。采集他们的一些智力测验成绩。用推论统计方法进行数据处理，最后会得出类似这样儿的结论：“研究发现，大学毕业生组的成绩显著高于初中毕业生组的成绩，二者在0.01水

平上具有显著性差异，说明大学毕业生的一些智力测验成绩优于中学毕业生组。”

其中，如果用EXCEL 来求描述统计。其方法是：工具-加载宏-勾选"分析工具库"，然后关闭Excel然后重新打开，工具菜单就会出现"数据分析"。描述统计是“数据分析”内一个子菜单，在做的时候，记得要把方格输入正确。最好直接点选。

## 正态性检验

很多统计方法都要求数值服从或近似服从正态分布，所以之前需要进行正态性检验。常用方法：非参数检验的K-量检验、P-P图、Q-Q图、W检验、动差法。

# Part2 假设检验

## 参数检验

参数检验是在已知总体分布的条件下（一股要求总体服从正态分布）对一些主要的参数(如均值、百分数、方差、相关系数等）进行的检验。

### U检验

**使用条件：** v当样本含量n较大时，样本值符合正态分布。

### T检验

**使用条件：** 当样本含量n较小时，样本值符合正态分布。

**① 单样本t检验**

推断该样本来自的总体均数μ与已知的某一总体均数μ0 (常为理论值或标准值)有无差别；

**② 配对样本t检验**

当总体均数未知时，且两个样本可以配对，同对中的两者在可能会影响处理效果的各种条件方面扳为相似。

**③ 两独立样本t检验**

无法找到在各方面极为相似的两样本作配对比较时使用。

## 非参数检验

非参数检验则不考虑总体分布是否已知，常常也不是针对总体参数，而是针对总体的某些一股性假设（如总体分布的位置是否相同，总体分布是否正态）进行检验。

### 适用情况

顺序类型的数据资料，这类数据的分布形态一般是未知的。

- 虽然是连续数据，但总体分布形态未知或者非正态；
- 体分布虽然正态，数据也是连续类型，但样本容量极小，如10以下。

### 主要方法包括

卡方检验、秩和检验、二项检验、游程检验、K-量检验等。

# Part3 信度分析

## 介绍

信度（Reliability）即可靠性，它是指采用同样的方法对同一对象重复测量时所得结果的一致性程度。信度指标多以相关系数表示，大致可分为三类：

- 稳定系数（跨时间的一致性）
- 等值系数（跨形式的一致性）
- 内在一致性系数（跨项目的一致性）。

信度分析的方法主要有以下四种：**重测信度法、复本信度法、折半信度法、α信度系数法。**

## 方法

### （1）重测信度法

这一方法是用同样的问卷对同一组被调查者间隔一定时间重复施测，计算两次施测结果的相关系数。显然，重测信度属于稳定系数。重测信度法特别适用于事实式问卷，如性别、出生年月等在两次施测中不应有任何差异，大多数被调查者的兴趣、爱好、习惯等在短时间内也不会有十分明显的变化。

如果没有突发事件导致被调查者的态度、意见突变，这种方法也适用于态度、意见式问卷。由于重测信度法需要对同一样本试测两次，被调查者容易受到各种事件、活动和他人的影响，而且间隔时间长短也有一定限制，因此在实施中有一定困难。

## （2）复本信度法

让同一组被调查者一次填答两份问卷复本，计算两个复本的相关系数。复本信度属于等值系数。复本信度法要求两个复本除表述方式不同外，在内容、格式、难度和对应题项的提问方向等方面要完全一致，而在实际调查中，很难使调查问卷达到这种要求，因此采用这种方法者较少。

## （3）折半信度法

折半信度法是将调查项目分为两半，计算两半得分的相关系数，进而估计整个量表的信度。折半信度属于内在一致性系数，测量的是两半题项得分间的一致性。

这种方法一般不适用于事实式问卷（如年龄与性别无法相比），常用于态度、意见式问卷的信度分析。在问卷调查中，态度测量最常见的形式是5级李克特（Likert）量表（李克特量表(Likert scale)是属评分加总式量表最常用的一种，属同一构念的这些项目是用加总方式来计分，单独或个别项目是无意义的。

它是由美国社会心理学家李克特于1932年在原有的总加量表基础上改进而成的。该量表由一组陈述组成，每一陈述有**"非常同意"、"同意"、"不一定"、"不同意"、"非常不同意"五种回答，分别记为5、4、3、2、1，**每个被调查者的态度总分就是他对各道题的回答所得分数的加总，这一总分可说明他的态度强弱或他在这一量表上的不同状态。）。

进行折半信度分析时，如果量表中含有反意题项，应先将反意题项的得分作逆向处理，以保证各题项得分方向的一致性，然后将全部题项按奇偶或前后分为尽可能相等的两半，计算二者的相关系数（rhh，即半个量表的信度系数），最后用斯皮尔曼-布朗（Spearman-Brown）公式：求出整个量表的信度系数（ru）。

## （4）α信度系数法：Cronbach

α信度系数是目前最常用的信度系数，其公式为：

$$\alpha = (k/(k-1)) * (1 - (\sum Si^2)/ST^2)$$

其中，K为量表中题项的总数， $Si^2$为第i题得分的题内方差， $ST^2$为全部题项总得分的方差。从公式中可以看出，α系数评价的是量表中各题项得分间的一致性，属于内在一致性系数。这种方法适用于态度、意见式问卷（量表）的信度分析。

总量表的信度系数最好在 0.8以上 ， 0.7-0.8 之间可以接受；分量表的信度系数最好在0.7以上， 0.6-0.7 还可以接受。Cronbach's alpha系数如果在 0.6以下 就要考虑重新编问卷。

检查测量的可信度，例如调查问卷的真实性。

## 分类

### （1）外在信度

不同时间测量时量表的一致性程度，常用方法重测信度。

### （2）内在信度

每个量表是否测量到单一的概念，同时组成两表的内在体项一致性如何，常用方法分半信度。

# Part4 列联表分析

列联表是观测数据按两个或更多属性（定性变量）分类时所列出的频数表。

## 简介

一般，若总体中的个体可按两个属性A、B分类，A有r个等级 $A1,A2,\cdots,Ar$ ，B有c个等级 $B1,B2,\cdots,Bc$ ，从总体中抽取大小为n的样本，设其中有nij个个体的属性属于等级Ai和Bj，nij称为频数，将r×c个nij排列为一个r行c列的二维列联表，简称r×c表。**若所考虑的属性多于两个，也可按类似的方式作出列联表，称为多维列联表。**

**列联表又称交互分类表**，所谓交互分类，是指同时依据两个变量的值，将所研究的个案分类。交互分类的目的是将两变量分组，然后比较各组的分布状况，以寻找变量间的关系。

用于分析离散变量或定型变量之间是否存在相关。

**列联表分析的基本问题是，判明所考察的各属性之间有无关联，即是否独立。**

如在前例中，问题是：一个人是否色盲与其性别是否有关？

在r×c表中，若以pi、pj和pij分别表示总体中的个体属于等级Ai，属于等级Bj和同时属于Ai、Bj的概率（pi，pj称边缘概率，pij称格概率），“A、B两属性无关联”的假设可以表述为 $H0:pij=pi·pj,(i=1,2,\cdots,r;j=1,2,\cdots,c)$ ，未知参数pij、pi、pj的最大似然估计（见点估计）分别为行和及列和（统称边缘和）。

根据K.皮尔森(1904)的拟合优度检验或似然比检验（见假设检验），当h0成立，且一切pi>0和pj>0时，统计量的渐近分布是自由度为 $(r-1)(c-1)$ 的Χ分布，式中 $Eij=(ni·nj)/n$ 称为期望频数。n为样本大小，当n足够大，且表中各格的Eij都不太小时，可以据此对h0作检验：若Χ值足够大，就拒绝假设h0，即认为A与B有关联。在前面的色觉问题中，曾按此检验，判定出性别与色觉之间存在某种关联。

## 需要注意

若样本大小n不是很大，则上述基于渐近分布的方法就不适用。对此，在四格表情形，R.A.费希尔(1935)提出了一种适用于所有n的精确检验法。其思想是在固定各边缘和的条件下，根据超几何分布（见概率分布），可以计算观测频数出现任意一种特定排列的条件概率。把实际出现的观测频数排列，以及比它呈现更多关联迹象的所有可能排列的条件概率都算出来并相加，若所得结果小于给定的显著性水平，则判定所考虑的两个属性存在关联，从而拒绝h0。

对于二维表，可进行卡方检验，对于三维表，可作Mentel-Hanszel分层分析。

列联表分析还包括配对计数资料的卡方检验、行列均为顺序变量的相关检验。

# Part5 相关分析

研究现象之间是否存在某种依存关系，对具体有依存关系的现象探讨相关方向及相关程度。

## 单相关

两个因素之间的相关关系叫单相关，即研究时只涉及一个自变量和一个因变量。

## 复相关

三个或三个以上因素的相关关系叫复相关，即研究时涉及两个或两个以上的自变量和因变量相关。

## 偏相关

在某一现象与多种现象相关的场合，当假定其他变量不变时，其中两个变量之间的相关关系称为偏相关。

# Part6 方差分析

**使用条件：** 各样本须是相互独立的随机样本；各样本来自正态分布总体；各总体方差相等。

## 分类

### （1）单因素方差分析

一项试验只有一个影响因素，或者存在多个影响因素时，只分析一个因素与响应变量的关系。

### （2）多因素有交互方差分析

一项实验有多个影响因素，分析多个影响因素与响应变量的关系，同时考虑多个影响因素之间的关系。

### （3）多因素无交互方差分析

分析多个影响因素与响应变量的关系，但是影响因素之间没有影响关系或忽略影响关系。

### （4）协方差分析

传统的方差分析存在明显的弊端，无法控制分析中存在的某些随机因素，使之影响了分析结果的准确度。协方差分析主要是在排除了协变量的影响后再对修正后的主效应进行方差分析，是将线性回归与方差分析结合起来的一种分析方法。

# Part7 回归分析

## 分类

### 一元线性回归分析

只有一个自变量X与因变量Y有关，X与Y都必须是连续型变量，因变量y或其残差必须服从正态分布。

### 多元线性回归分析

**使用条件：** 分析多个自变量与因变量Y的关系，X与Y都必须是连续型变量，因变量y或其残差必须服从正态分布。

**（1）变呈筛选方式**

选择最优回归方程的变里筛选法包括全横型法（CP法）、逐步回归法，向前引入法和向后剔除法。

**（2）横型诊断方法**

① **残差检验：** 观测值与估计值的差值要艰从正态分布。
② **强影响点判断：** 寻找方式一般分为标准误差法、Mahalanobis距离法。

**③ 共线性诊断：**

- **诊断方式：** 容忍度、方差扩大因子法(又称膨胀系数VIF)、特征根判定法、条件指针CI、方差比例。
- **处理方法：** 增加样本容量或选取另外的回归如主成分回归、岭回归等。

## Logistic回归分析

线性回归模型要求因变量是连续的正态分布变里，且自变量和因变量呈线性关系，而Logistic回归模型对因变量的分布没有要求，一般用于因变量是离散时的情况。

**分类：** Logistic回归模型有条件与非条件之分，条件Logistic回归模型和非条件Logistic回归模型的区别在于参数的估计是否用到了条件概率。

## 其他回归方法

非线性回归、有序回归、Probit回归、加权回归等

# Part8 聚类分析

聚类与分类的不同在于，聚类所要求划分的类是未知的。

聚类是将数据分类到不同的类或者簇这样的一个过程，所以同一个簇中的对象有很大的相似性，而不同簇间的对象有很大的相异性。

从统计学的观点看，聚类分析是通过数据建模简化数据的一种方法。传统的统计聚类分析方法包括系统聚类法、分解法、加入法、动态聚类法、有序样品聚类、有重叠聚类和模糊聚类等。采用k-均值、k-中心点等算法的聚类分析工具已被加入到许多著名的统计分析软件包中，如SPSS、SAS

等。

从机器学习的角度讲，簇相当于隐藏模式。聚类是搜索簇的无监督学习过程。与分类不同，无监督学习不依赖预先定义的类或带类标记的训练实例，需要由聚类学习算法自动确定标记，而分类学习的实例或数据对象有类别标记。聚类是观察式学习，而不是示例式的学习。

聚类分析是一种探索性的分析，在分类的过程中，人们不必事先给出一个分类的标准，聚类分析能够从样本数据出发，自动进行分类。聚类分析所使用方法的不同，常常会得到不同的结论。不同研究者对于同一组数据进行聚类分析，所得到的聚类数未必一致。

从实际应用的角度看，聚类分析是数据挖掘的主要任务之一。而且聚类能够作为一个独立的工具获得数据的分布状况，观察每一簇数据的特征，集中对特定的聚簇集合作进一步地分析。聚类分析还可以作为其他算法（如分类和定性归纳算法）的预处理步骤。

## 定义

依据研究对象（样品或指标）的特征，对其进行分类的方法，减少研究对象的数目。各类事物缺乏可靠的历史资料，无法确定共有多少类别，目的是将性质相近事物归入一类。

各指标之间具有一定的相关关系。

聚类分析(cluster analysis)是一组将研究对象分为相对同质的群组(clusters)的统计分析技术。聚类分析区别于分类分析(classification analysis)，后者是有监督的学习。

## 变量类型

定类变量、定量（离散和连续）变量

样本个体或指标变量按其具有的特性进行分类，寻找合理的度量事物相似性的统计量。

### （1）性质分类

- **Q型聚类分析：** 对样本进行分类处理，又称样本聚类分析使用距离系数作为统计量衡量相似度，如欧式距离、极端距离、绝对距离等。
- **R型聚类分析：** 指标进行分类处理，又称指标聚类分析使用相似系数作为统计量衡量相似度，相关系数、列联系数等。

### （2）方法分类

- **系统聚类法：** 适用于小样本的样本聚类或指标聚类，一般用系统聚类法来聚类指标，又称分层聚类。
- **逐步聚类法：** 适用于大样本的样本聚类
- **其他聚类法：** 两步聚类、K均值聚类等

# Part9 判别分析

## 判别分析

根据已掌握的一批分类明确的样品建立判别函数，使产生错判的事例最少，进而对给定的一个新样品，判断它来自哪个总体。

## 与聚类分析区别

- 聚类分析可以对样本进行分类，也可以对指标进行分类；而判别分析只能对样本。
- 聚类分析事先不知道事物的类别，也不知道分几类；而判别分析必须事先知道事物的类别，也知道分几类。
- 聚类分析不需要分类的历史资料，而直接对样本进行分类；而判别分析需要分类历史资料去建立判别函数，然后才能对样本进行分类

## 分类

### （1）Fisher判别分析法

以距离为判别准则来分类，即样本与哪个类的距离最短就分到哪一类，适用于两类判别；

以概率为判别准则来分类，即样本属于哪一类的概率最大就分到哪一类，适用于多类判别。

### （2）BAYES判别分析法

BAYES判别分析法比FISHER判别分析法更加完善和先进，它不仅能解决多类判别分析，而且分析时考虑了数据的分布状态，所以一般较多使用。

# Part10 主成分分析

## 介绍

主成分分析（Principal Component Analysis，PCA），是一种统计方法。通过正交变换将一组可能存在相关性的变量转换为一组线性不相关的变量，转换后的这组变量叫主成分。

在实际课题中，为了全面分析问题，往往提出很多与此有关的变量（或因素），因为每个变量都在不同程度上反映这个课题的某些信息。

主成分分析首先是由K.皮尔森（Karl Pearson）对非随机变量引入的，尔后H.霍特林将此方法推广到随机向量的情形。信息的大小通常用离差平方和或方差来衡量。

将彼此相关的一组指标变适转化为彼此独立的一组新的指标变量，并用其中较少的几个新指标变量就能综合反应原多个指标变量中所包含的主要信息。

## 原理

在用统计分析方法研究多变量的课题时，变量个数太多就会增加课题的复杂性。人们自然希望变量个数较少而得到的信息较多。在很多情形，变量之间是有一定的相关关系的，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此课题的信息有一定的重叠。

主成分分析是对于原先提出的所有变量，将重复的变量（关系紧密的变量）删去多余，建立尽可能少的新变量，使得这些新变量是两两不相关的，而且这些新变量在反映课题的信息方面尽可能保持原有的信息。

设法将原来变量重新组合成一组新的互相无关的几个综合变量，同时根据实际需要从中可以取出几个较少的综合变量尽可能多地反映原来变量的信息的统计方法叫做主成分分析或称主分量分析，也是数学上用来降维的一种方法。

## 缺点

在主成分分析中，我们首先应保证所提取的前几个主成分的累计贡献率达到一个较高的水平（即变量降维后的信息量须保持在一个较高水平上），其次对这些被提取的主成分必须都能够给出符合实际背景和意义的解释（否则主成分将空有信息量而无实际含义）。

主成分的解释其含义一般多少带有点模糊性，不像原始变量的含义那么清楚、确切，这是变量降维过程中不得不付出的代价。因此，提取的主成分个数m通常应明显小于原始变量个数p（除非p本身较小），否则维数降低的“利”可能抵不过主成分含义不如原始变量清楚的“弊”。

# Part11 因子分析

一种旨在寻找隐藏在多变量数据中、无法直接观察到却影响或支配可测变量的潜在因子、并估计潜在因子对可测变量的影响程度以及潜在因子之间的相关性的一种多元统计分析方法。

## 与主成分分析比较

- **相同**： 都能够起到治理多个原始变量内在结构关系的作用。
- **不同**： 主成分分析重在综合原始变适的信息.而因子分析重在解释原始变量间的关系，是比主成分分析更深入的一种多元统计方法

## 用途

- 减少分析变量个数。
- 通过对变量间相关关系探测，将原始变量进行分类。

# Part12 时间序列分析

动态数据处理的统计方法，研究随机数据序列所遵从的统计规律，以用于解决实际问题；时间序列通常由4种要素组成：**趋势、季节变动、循环波动和不规则波动。**

## 主要方法

**移动平均滤波与指数平滑法、ARIMA横型、量ARIMA横型、ARIMAX模型、向呈自回归横型、ARCH族模型。**

时间序列是指同一变量按事件发生的先后顺序排列起来的一组观察值或记录值。构成时间序列的要素有两个：**其一是时间，其二是与时间相对应的变量水平。**

实际数据的时间序列能够展示研究对象在一定时期内的发展变化趋势与规律，因而可以从时间序列中找出变量变化的特征、趋势以及发展规律，从而对变量的未来变化进行有效地预测。

时间序列的变动形态一般分为四种：**长期趋势变动，季节变动，循环变动，不规则变动。**

## 时间序列预测法的应用

### 系统描述

根据对系统进行观测得到的时间序列数据，用曲线拟合方法对系统进行客观的描述。

### 系统分析

当观测值取自两个以上变量时，可用一个时间序列中的变化去说明另一个时间序列中的变化，从而深入了解给定时间序列产生的机理。

### 预测未来

一般用ARMA模型拟合时间序列，预测该时间序列未来值。

### 决策和控制

根据时间序列模型可调整输入变量使系统发展过程保持在目标值上，即预测到过程要偏离目标时便可进行必要的控制。

## 特点

- 假定事物的过去趋势会延伸到未来；
- 预测所依据的数据具有不规则性；
- 撇开了市场发展之间的因果关系。

## （1）

时间序列分析预测法是根据市场过去的变化趋势预测未来的发展，它的前提是假定事物的过去会同样延续到未来。事物的现实是历史发展的结果，而事物的未来又是现实的延伸，事物的过去和未来是有联系的。

市场预测的时间序列分析法，正是根据客观事物发展的这种连续规律性，运用过去的历史数据，通过统计分析，进一步推测市场未来的发展趋势。市场预测中，事物的过去会同样延续到未来，其意思是说，市场未来不会发生突然跳跃式变化，而是渐进变化的。

时间序列分析预测法的哲学依据，是唯物辩证法中的基本观点，即认为一切事物都是发展变化的，事物的发展变化在时间上具有连续性，市场现象也是这样。市场现象过去和现在的发展变化规律和发展水平，会影响到市场现象未来的发展变化规律和规模水平；市场现象未来的变化规律和水平，是市场现象过去和现在变化规律和发展水平的结果。

需要指出，由于事物的发展不仅有连续性的特点，而且又是复杂多样的。因此，在应用时间序列分析法进行市场预测时应注意市场现象未来发展变化规律和发展水平，不一定与其历史和现在的发展变化规律完全一致。

随着市场现象的发展，它还会出现一些新的特点。因此，在时间序列分析预测中，决不能机械地按市场现象过去和现在的规律向外延伸。必须要研究分析市场现象变化的新特点，新表现，并且将这些新特点和新表现充分考虑在预测值内。这样才能对市场现象做出既延续其历史变化规律，又符合其现实表现的可靠的预测结果。

（2）

时间序列分析预测法突出了时间因素在预测中的作用，暂不考虑外界具体因素的影响。时间序列在时间序列分析预测法处于核心位置，没有时间序列，就没有这一方法的存在。虽然，预测对象的发展变化是受很多因素影响的。但是，运用时间序列分析进行量的预测，实际上将所有的影响因素归结到时间这一因素上，只承认所有影响因素的综合作用，并在未来对预测对象仍然起作用，并未去分析探讨预测对象和影响因素之间的因果关系。

因此，为了求得能反映市场未来发展变化的精确预测值，在运用时间序列分析法进行预测时，必须将量的分析方法和质的分析方法结合起来，从质的方面充分研究各种因素与市场的关系，在充分分析研究影响市场变化的各种因素的基础上确定预测值。

需要指出的是，时间序列预测法因突出时间序列暂不考虑外界因素影响，因而存在着预测误差的缺陷，当遇到外界发生较大变化，往往会有较大偏差，时间序列预测法对于中短期预测的效果要比长期预测的效果好。因为客观事物，尤其是经济现象，在一个较长时间内发生外界因素变化的可能性加大，它们对市场经济现象必定要产生重大影响。如果出现这种情况，进行预测时，只考虑时间因素不考虑外界因素对预测对象的影响，其预测结果就会与实际状况严重不符。

## Part13 生存分析

用来研究生存时间的分布规律以及生存时间和相关因索之间关系的一种统计分析方法

### 包含内容

- 描述生存过程，即研究生存时间的分布规律
- 比较生存过程，即研究两组或多组生存时间的分布规律，并进行比较。
- 分析危险因素，即研究危险因素对生存过程的影响。

- 建立数学模型，即将生存时间与相关危险因素的依存关系用一个数学式子表示出来。

## 方法

### （1）统计描述

包括求生存时间的分位数、中数生存期、平均数、生存函数的估计、判断生存时间的图示法，不对所分析的数据作出任何统计推断结论。

### （2）非参数检验

检验分组变量各水平所对应的生存曲线是否一致，对生存时间的分布没有要求，并且检验危险因素对生存时间的影响。

- 乘积极限法（PL法）
- 寿命表法（LT法）

### （3）半参数横型回归分析

在特定的假设之下，建立生存时间随多个危险因素变化的回归方程，这种方法的代表是Cox比例风险回归分析法。

### （4）参数模型回归分析

已知生存时间服从特定的参数横型时，拟合相应的参数模型，更准确地分析确定变量之间的变化规律。

# Part14 典型相关分析

相关分析一般分析两个变量之间的关系，而典型相关分析是分析两组变量（如3个学术能力指标与5个在校成绩表现指标）之间相关性的一种统计分析方法。

典型相关分析的基本思想和主成分分析的基本思想相似，它将一组变量与另一组变量之间单变量的多重线性相关性研究转化为对少数几对综合变量之间的简单线性相关性的研究，并且这少数几对变量所包含的线性相关性的信息几乎覆盖了原变量组所包含的全部相应信息。

## Part15 R0C分析

R0C曲线是根据一系列不同的二分类方式(分界值或决定阈）.以真阳性率（灵敏度)为纵坐标，假阳性率（1-特异度)为横坐标绘制的曲线。

### 用途

1. R0C曲线能很容易地查出任意界限值时的对疾病的识别能力；
2. 选择最佳的诊断界限值。R0C曲线越靠近左上角，试验的准确性就越高；
3. 两种或两种以上不同诊断试验对疾病识别能力的比较，一股用R0C曲线下面积反映诊断系统的准确性。

## Part16 其他分析方法

多重响应分析、距离分祈、项目分祈、对应分祈、决策树分析、神经网络、系统方程、蒙特卡洛模拟等。

### 决策树分析与随机森林

尽管有剪枝等等方法，一棵树的生成肯定还是不如多棵树，因此就有了随机森林，解决决策树泛化能力弱的缺点。（可以理解成三个臭皮匠顶过诸葛亮）

## 决策树(Decision Tree）：

是在已知各种情况发生概率的基础上，通过构成决策树来求取净现值的期望值大于等于零的概率，评价项目风险，判断其可行性的决策分析方法，是直观运用概率分析的一种图解法。由于这种决策分支画成图形很像一棵树的枝干，故称决策树。

在机器学习中，决策树是一个预测模型，他代表的是对象属性与对象值之间的一种映射关系。Entropy = 系统的凌乱程度，使用算法ID3, C4.5和C5.0生成树算法使用熵。这一度量是基于信息学理论中熵的概念。

决策树是一种树形结构，其中每个内部节点表示一个属性上的测试，每个分支代表一个测试输出，每个叶节点代表一种类别。

分类树（决策树）是一种十分常用的分类方法。他是一种监督学习，所谓监督学习就是给定一堆样本，每个样本都有一组属性和一个类别，这些类别是事先确定的，那么通过学习得到一个分类器，这个分类器能够对新出现的对象给出正确的分类。这样的机器学习就被称之为监督学习。

## 优点

决策树易于理解和实现，人们在在学习过程中不需要使用者了解很多的背景知识，这同时是它的能够直接体现数据的特点，只要通过解释后都有能力去理解决策树所表达的意义。

对于决策树，数据的准备往往是简单或者是不必要的，而且能够同时处理数据型和常规型属性，在相对短的时间内能够对大型数据源做出可行且效果良好的结果。易于通过静态测试来对模型进行评测，可以测定模型可信度；如果给定一个观察的模型，那么根据所产生的决策树很容易推出相应的逻辑表达式。

## 缺点

对连续性的字段比较难预测；对有时间顺序的数据，需要很多预处理的工作；当类别太多时，错误可能就会增加的比较快；一般的算法分类的时候，只是根据一个字段来分类。

> 来源：https://zhuanlan.zhihu.com/p/39214084

- EOF -