Smith Duan

Zhi-peng.duan@hp.com

摘要

通过对《统计学》（贾俊平）第五版这本书的学习，   
系统整理统计的主要用途和各种方法。

**统计学**

学习笔记

# 统计导论

统计是处理数据的一门科学。统计学主要是依靠对大量数据的分析而得到结论的科学。【统计学是关于数据的科学，它所提供的是一套有关数据收集、处理、分析、解释，最终从数据中得到结论的一种方法。数据收集就是取得统计数据；数据处理是将数据用图表等形式展示出来；数据分析则是选择适当的统计方法研究数据，并从数据中提取有用信息进而得到结论。】统计分析的直接载体就是数据，而数据也同样具有以下的特点。

1. 现实世界中“大量”的实际意思只能是指“样本（部分）”，因为数据是在无时无刻不在增加，而且没有穷尽，所以只能无限接近全体数据但是永远不会达到全体的程度。
2. 数据同样具有两面性。正确的理解数据意义是可以起到一定的指导生产生活的作用，但如果不能正确的理解数据的意义，同时将会给生产生活带来不可估计的灾难。

由于上面的特点所以决定了统计结果只是通过数据对现实生活的一种猜测。猜测的结果永远不可能是完全正确的。猜测的结果只是对未来的一种预见，指导我们规避一些风险和意外。

统计分析步骤：

1. 收集统计数据
2. 展示数据
3. 数据分析（分为描述统计方法和推断统计方法）

* 描述统计(descriptive statistics)：研究的是数据收集、处理、汇总、图表描述、概括与分析等统计方法
* 推断统计(inferential statistics)：研究如何利用样本数据来推断总体特征的统计方法

1. 得到结论

## 统计数据的类型

统计数据是对现象进行测量的结果。

* **分类数据、顺序数据和数值型数据**

按照所采用的计量尺度不同，可以将统计数据分为分类数据、顺序数据和数值型数据。

**分类数据**是只能归于某一类别的非数字型数据。例子，人口按照性别分为男、女两类；企业按照行业属性分为医药企业、家电企业、纺织品企业等。对于分类数据可以用数字代码来表示各个类别。

**顺序数据**是只能归于某一有序类别的非数字型数据。例子，产品分为一等品、二等品三等品、次品等；考试成绩分为优、良、中、及格、不及格等。同样对于顺序数据也可以用数字代码来表示各个类别。

**数值型数据**是按数字尺度测量的观察值。

分类数据、顺序数据是定性数据（品质数据），数值型数据是定量数据（质量数据）。

* **观测数据和实验数据**

按照统计数据的收集方法，可将数据分为观测数据和实验数据。**观测数据**是通过调查或观测而收集到的数据，这类数据是在没有对事物人为控制条件下得到的，有关社会经济现象的统计数据几乎都是观测数据。**实验数据**则是在实验中控制实验对象而收集到的数据，比如对一种新药疗效的实验数据，自然科学领域的大多数数据都是实验数据。

* **截面数据和时间序列数据**

按照被描述的现象与时间的关系，可以将统计数据分为截面数据和时间序列数据。**截面数据**是在相同或近似相同的时间点上收集的数据，这类数据通常是在不同空间获得的，用于描述现象在某一时刻的变化情况。**时间序列数据**是在不同时间上收集到的数据，用以描述现象随时间变化的情况。

图1给出了统计数据分类的框图。

统计数据的类型

按计量尺度

按收集方法

按时间状况

分类数据

顺序数据

数值型数据

观测数据

实验数据

截面数据

时序数据

计算各组的频数和频率，进行列联表分析和x2检验

计算中位数和四分位差，计算等级相关系数

更多统计方法，参数估计和假设检验等

图 1 统计数据的分类

## 统计中的几个基本概念

* **总体和样本**

**总体：**包含所研究全部个体（数据）的集合，称为总体。总体只是一种理想化的概念，因为世界是无时无刻不在发展变化的，所以不可能获得总体。

**样本：**总体中抽样的一部分元素成为样本。我们研究更多的是样本。构成样本元素的数量称为样本容量或是样本量。

* **参数和统计量**

**参数：用来描述总体特征的概括性数字度量**（研究者想要了解的总体的某种特征值），参数是反映了总体的某些特征。我们所关心的参数通常有总体平均数、标准差、总体比例等。

**统计量：用来描述样本特征的概括性数字度量**（统计量是根据样本数据计算出来的一个量，统计量是样本函数）。样本我们可以抽出来，所以统计量是可以确定知道的。而抽样的目的就是通过样本统计量去估计总体参数。

图2给出总体、样本，参数与统计量之间的关系。

图 2 总体和样本，统计量和参数的关系

**统计量**

**参数**

# 用图表展示数据-【描述性统计】

## 用图表描述定性数据

定性数据包括分类数据和顺序数据，它们的图表展示方法基本相同。通常可以用频数分布表和图形来描述。另外适用于定性数据的图形还有条形图（Bar chart）、帕累托图、饼图等。

### 频率（frequency）

定性数据本身是对事物的一种分类，因此，只要先把所有的类别都列出来，然后统计出每个类别的频数，就是一张**频数分布**（frequency distribution）表。通过频数分布可以观察不同类型数据的分布状况。网状网中历史交易数据就可以生成频数分布表，如表1各省交易量统计数据，省代码可以理解为分类数据，而交易数则可理解为省业务交易出现的频率。

|  |  |
| --- | --- |
| **省代码** | **交易数** |
| 1000 | 2458807 |
| 2100 | 1243174 |
| 2200 | 2275882 |
| 2500 | 2023574 |
| 2700 | 2064099 |
| 3110 | 1181768 |
| 4710 | 2091169 |
| 5510 | 1847723 |
| 7310 | 112580228 |
| 7710 | 2492102 |
| 9310 | 2159397 |

表- 1各省交易量统计

注意：当涉及两个分类变量时，可以生成两张频数分布表来观察它们各自的分布状况，也可以将一个变量放在行的位置，另一个变量放在列的位置，这种由两个变量交叉分类的频数分布表成为**列联表**（contingency table），也称**交叉表**（cross table）。频数分布表和列联表中的变量都只能是定性数据。

下面学习如何利用SPSS生成频数分布表或列联表。



生成频数分布图的方法：运行SPSS，在菜单栏中选择“分析”-“描述统计”-“频率”。

在弹出的“频率”界面中，将要统计的变量添加进变量框（如图2频率），最后点击“确定”，SPSS会自动生成统计结果分析（如图4频率统计分析结果）。



图 2 频率

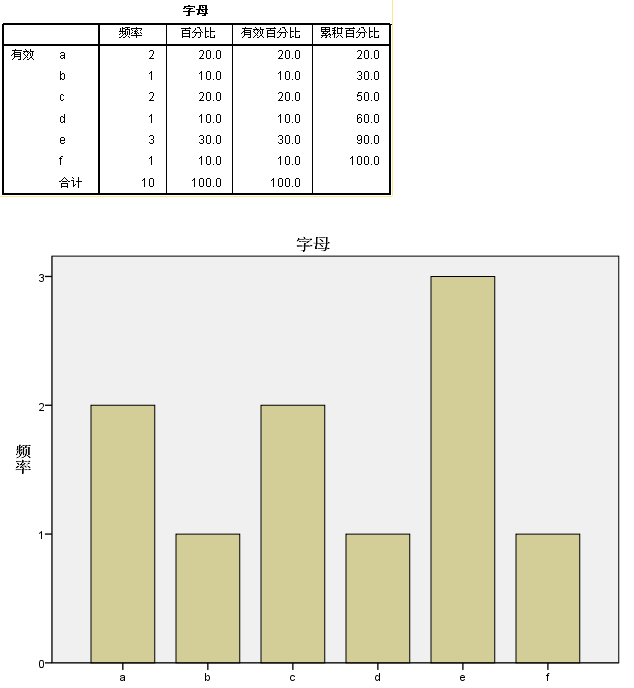


图 3 频数统计分析结果

生成列联表地方法：运行SPSS，在菜单栏中选择“分析”-“描述统计”-“交叉表”。

在弹出的“交叉表”界面中，将要统计的变量添加行、列框（如图4交叉表），最后点击“确定”，SPSS会自动生成统计结果分析（如图5交叉表统计分析结果）。



图 4 交叉表

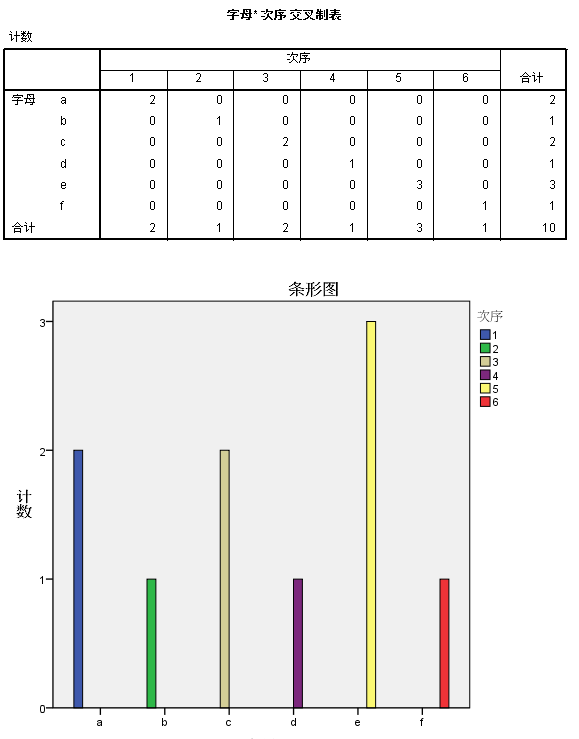


图 5 交叉表统计分析结果

### 条形图（Bar chart）

条形图（Bar chart）是用宽度形同的条形来表示数据多少的图形，用于观察不同的类别频数的多少或分布情况。



生成条形图的方法：运行SPSS，在菜单栏中选择“分析”-“图形”-“旧对话框”-“条形图”。

在“条形图”界面中可以根据需要选择图形样式（包括简单箱图、复式条形图、堆积面积图三种样式），在此选择简单箱图并点击“定义”，弹出“定义简单条形图”对话框（图6）。“条的表征”中选择“其他统计量（例如均值）”，才能在“变量”中增加变量。点击“确定”生成条形图（如图7）。



图 6 定义简单条形图

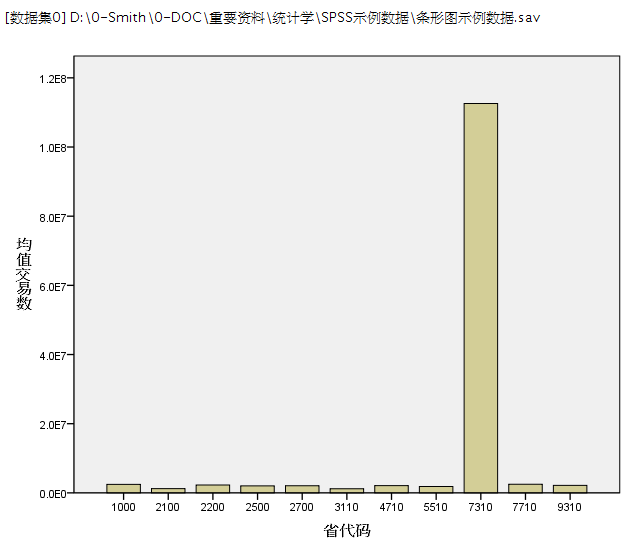


图 7 条形图

|  |
| --- |
| 例如当变量为字符型（性别），描述频率的时候如果采用直方图，SPSS会出现以下提示“性别为一字符串，因此无法生成直方图。” |

### 帕累托图

帕累托图是按类别出现的频数多少排序后绘制的条形图。

### 饼图

饼图的应用场景同条形图很类似。用圆形以及圆内扇区的角度来表示数值大小的图形，主要用来表示一个样本（或总体）中各类别的频数占总部频数的比例。

## 用图表描述定量数据

定性数据的图示方法都适用于定量数据。因为定量数据是表示考察的某种特性是具体的数值，概率研究的是频率，所以要考虑将定量数据展示为定性数据。主要的方法的分组，将原始数据按照某种标准（宽度）进行分组，再统计每个分组的数据频数分布。

### 直方图（分组数据看分布）

在统计学中，直方图（英语：Histogram）是一种对数据分布情况的图形表示，是一种二维[统计图表](http://www.bing.com/knows/%E7%BB%9F%E8%AE%A1%E5%9B%BE%E8%A1%A8)，它的两个坐标分别是统计样本和该样本对应的某个属性的度量。

**条形图与直方图的区别**：

由于分组数据具有连续性，[直方图](http://www.bing.com/knows/%E7%9B%B4%E6%96%B9%E5%9B%BE)中的各[矩形](http://www.bing.com/knows/%E7%9F%A9%E5%BD%A2)通常是连续排列，直方图的高度表示各分组内数据个数，在直方图中X轴是连续的，宽度是分组间隔，高度是分组数据出现的频数。而条形图则是分开排列，此外，而条形图高度表示某项目内的数据个数。

* 条形图是用条形的长度表示各类别频数的多少，其宽度则是固定的，没有实际意义；直方图是用面积表示各组频数的多少，矩形的高度表示各分组的频数或是频率，宽度则表示各组的组距，因此其高度宽度均有意义。
* 由于分组数据具有连续性，直方图的各矩形通常是连续排列，而条形图是分开排列。
* 条形图主要用于展示分类数据；直方图主要用于展示（数值型）分组数据。

|  |
| --- |
| 用图形描述数值型数据的频率时可用条形图也可以用直方图。 |

### 茎叶图（未分组数据看分布）

茎叶图不仅可以看出数据分布，同时可以看出原始数据。但是茎叶图也有不适用的场景，当数据量过大时，不适宜采用茎叶图。

### 箱型图（未分组数据看分布）

五个数据位置准确的描述数据的分布情况以及数据的波动。更直观的体现不同考察量数据的差异特点。

# 统计量

## 协方差

标准差和方差一般是用来描述一维数据的，但现实生活我们常常遇到含有多维数据的数据集，面对多维数据集，我们当然可以按照每一维独立的计算其方差，但是通常我们还想了解更多，比如，多维数据之间是否存在一些联系，协方差就是这样一种用来度量两个随机变量关系的统计量，我们可以仿照方差的定义：

http://album.sina.com.cn/pic/001T8VKLgy6OI74QVD016

来度量各个维度偏离其均值的程度，标准差可以这么来定义：

http://album.sina.com.cn/pic/001T8VKLgy6OI75EG0tf5

协方差的结果有什么意义呢？如果结果为正值，则说明两者是正相关的(从协方差可以引出“相关系数”的定义)，也就是说一个人越猥琐就越受女孩子欢迎，嘿嘿，那必须的~结果为负值就说明负相关的，越猥琐女孩子越讨厌，可能吗？如果为0，也是就是统计上说的“相互独立”。

从协方差的定义上我们也可以看出一些显而易见的性质，如：

http://album.sina.com.cn/pic/001T8VKLgy6OI77suBI4a

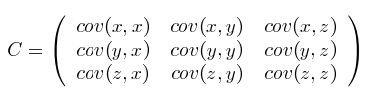
http://album.sina.com.cn/pic/001T8VKLgy6OI798oUg70

**协方差多了就是协方差矩阵**

上一节提到的猥琐和受欢迎的问题是典型二维问题，而协方差也只能处理二维问题，那维数多了自然就需要计算多个协方差，比如n维的数据集就需要计算[关于协方差矩阵的概念及意义](http://photo.blog.sina.com.cn/showpic.html#blogid=672c5a470100miqq&url=http://album.sina.com.cn/pic/001T8VKLgy6OI7aocP18b)个协方差，那自然而然的我们会想到使用矩阵来组织这些数据。给出协方差矩阵的定义：

[关于协方差矩阵的概念及意义](http://photo.blog.sina.com.cn/showpic.html#blogid=672c5a470100miqq&url=http://album.sina.com.cn/pic/001T8VKLgy6OI7dfzLn7d)

这个定义还是很容易理解的，我们可以举一个简单的三维的例子，假设数据集有[关于协方差矩阵的概念及意义](http://photo.blog.sina.com.cn/showpic.html#blogid=672c5a470100miqq&url=http://album.sina.com.cn/pic/001T8VKLgy6OI7dUdPNef)三个维度，则协方差矩阵为

[](http://photo.blog.sina.com.cn/showpic.html#blogid=672c5a470100miqq&url=http://album.sina.com.cn/pic/001T8VKLgy6OI7eVgrE2a)

可见，协方差矩阵是一个对称的矩阵，而且对角线是各个维度上的方差。

**协方差矩阵计算的是不同维度之间的协方差，而不是不同样本之间的。**

# 统计中的检验法

## 单样本的正态分布检验（K-S检验）

正态分布在统计学中是比较重要的概念，因为很多统计分析的前提是要符合正态分布。所以在最开始学习的时候就想搞清楚正态分布检验的方法。例如一拿到数据的时候是不确定数据是否符合正态分布，那么如何检验数据符合正态分布呢？用假设验证的思路进行验证，由于原假设要确定相等关系，所以先确定原假设是样本来自的总体数据同理论无显著性差异，再通过数据解释拒绝或是接收原假设就可以得到答案了。其中有一个比较重要的概念是显著性水平（可以根据实际情况设计，一般显著性水平α=0.05）。

显著性水平（Significance [Level](http://baike.baidu.com/view/1539485.htm)）是估计总体参数落在某一区间内，可能犯错误的概率，用α表示。显著性是对差异的程度而言的，程度不同说明引起变动的原因也有不同：一类是条件差异，一类是随机差异。它是在进行[假设检验](http://baike.baidu.com/view/1445854.htm)时事先确定一个可允许的作为判断界限的小概率标准。

假设检验运用了[小概率原理](http://baike.baidu.com/view/640673.htm)，事先确定的作为判断的界限，即允许的小概率的标准，称为显著性水平。如果根据[命题](http://baike.baidu.com/view/2050.htm)的原假设所计算出来的[概率](http://baike.baidu.com/view/45320.htm)小于这个标准，就拒绝原假设；大于这个标准则不拒绝原假设。这样显著性水平把[概率分布](http://baike.baidu.com/view/45323.htm)分为两个[区间](http://baike.baidu.com/view/70334.htm)：拒绝区间，不拒绝区间。

显著性水平不是一个固定不变的数字，其越大，则原假设被拒绝的可能性愈大，原假设为真而被否定的风险也愈大。显著性水平应根据所研究的的性质和我们对结论准确性所持的要求而定。

|  |
| --- |
| 原假设：H0为样本来自的总体与指定理论分布无[显著性差异](https://www.baidu.com/s?wd=%E6%98%BE%E8%91%97%E6%80%A7%E5%B7%AE%E5%BC%82&hl_tag=textlink&tn=SE_hldp01350_v6v6zkg6)。  一般假设你的[显著性水平](https://www.baidu.com/s?wd=%E6%98%BE%E8%91%97%E6%80%A7%E6%B0%B4%E5%B9%B3&hl_tag=textlink&tn=SE_hldp01350_v6v6zkg6" \t "_blank)为α=0.05。即置信区间为95%  如果概率小于或等于用户的[显著性水平](https://www.baidu.com/s?wd=%E6%98%BE%E8%91%97%E6%80%A7%E6%B0%B4%E5%B9%B3&hl_tag=textlink&tn=SE_hldp01350_v6v6zkg6)α，则应拒绝[原假设](https://www.baidu.com/s?wd=%E9%9B%B6%E5%81%87%E8%AE%BE&hl_tag=textlink&tn=SE_hldp01350_v6v6zkg6)，认为样本来自的总体与指定理论分布的总体有显著差异（就是[小概率事件](https://www.baidu.com/s?wd=%E5%B0%8F%E6%A6%82%E7%8E%87%E4%BA%8B%E4%BB%B6&hl_tag=textlink&tn=SE_hldp01350_v6v6zkg6)发生了，拒绝假设，就可以理解样本来自的总体和[正态分布](https://www.baidu.com/s?wd=%E6%AD%A3%E6%80%81%E5%88%86%E5%B8%83&hl_tag=textlink&tn=SE_hldp01350_v6v6zkg6)之间的相似性不存在）  如果概率大于α就是具有相似性，可以理解为服从[正态分布](https://www.baidu.com/s?wd=%E6%AD%A3%E6%80%81%E5%88%86%E5%B8%83&hl_tag=textlink&tn=SE_hldp01350_v6v6zkg6)。  一般用起来的时候【渐进显著性（双侧）值（P值）】小于0.05就认为两者有显著差异，就是两者不相似；反之大于0.05可认为近似服从正态分布。 |

## F检验（方差齐性检验）

从两研究总体中随机抽取样本，要对这两个样本进行比较的时候，首先要判断两总体方差是否相同，即方差齐性。若两总体方差相等，则直接用t检验，若不等，可采用t'检验或变量变换或[秩和检验](http://baike.baidu.com/view/1889901.htm)等方法。

F检验是通过比较两组数据的反方差，来判断两组数据是否存在较大的偶然误差，是精密度检验。

而t检验是与标准值比较，用于判断某一分析方法或操作过程是否存在较大的误差。

显著性检验的顺序应该为先进行F检验，确认两组数据没有显著性差异之后，在进行两组数据均值是否存在系统误差的t检验。

## 相关性检验

### Pearson（线性相关检验）

在对两个变量之间关系有了线性猜想后，应该对两个变量进行线性相关性检验。

操作步骤：“分析”-“相关”-“双变量”，Pearson（皮尔逊）是用来检验两个变量之间是否存在线性相关关系的。

例子：

检验落地交易数量与系统错误数量之间是否存在线性相关关系

| **描述性统计量** | | | |
| --- | --- | --- | --- |
|  | 均值 | 标准差 | N |
| 落地交易数 | 40744732.82 | 66540761.132 | 33 |
| 系统错误数 | 146666.82 | 370278.110 | 33 |

描述统计量是对考察变量的统计量进行描述，均值、标准差、样本数

| **相关性** | | | |
| --- | --- | --- | --- |
|  | | 落地交易数 | 系统错误数 |
| 落地交易数 | Pearson 相关性 | 1 | .823\*\* |
| 显著性（双侧） |  | .000 |
| N | 33 | 33 |
| 系统错误数 | Pearson 相关性 | .823\*\* | 1 |
| 显著性（双侧） | .000 |  |
| N | 33 | 33 |
| \*\*. 在 .01 水平（双侧）上显著相关。 | | | |

原假设H0 : 落地交易数与系统错误数不存在显著的线性相关关系

在结果中”1”表示各变量同自己的相关关系为1；重点是显著性（双侧）【P】= 0.000<0.05，小概率事件发生，拒绝原假设。所以得到结论是落地交易数与系统错误数存在明显的线性相关关系。

注意：0.823（相关系数）只能表明落地交易数与系统错误数之间的相关系数为正，落地交易数变大系统错误数也将增大。0.823并不能表明两个变量间的因果关系或关系系数的具体值，这些是要由线性回归模型来确定的。另外相关系数是属于(-1，1)之间，如果为负数表明变量之间成负相关，一个变量增加，另外一个变量就会随之减小；反之正相关就是一个变量增加，另一个变量也增加。 越接近1表明相关性越强。

### Spearman（等级相关检验）

操作步骤：“分析”-“相关”-“双变量”，Spearman是用来检验两个变量之间是否存在等级相关关系的。

注意：线性相关是研究两类数值型变量之间是否存在线性关系。而spearman是研究其中有一类变量是等级变量（顺序变量）。

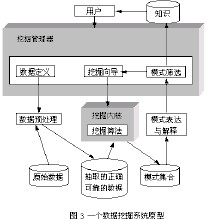
Spearman的结果解读与pearson的结果解读方法很相似。

# 数据挖掘与统计分析

## 数据挖掘

数据挖掘：（Data mining），又译为资料探勘、数据采矿。它是数据库知识发现（英语：Knowledge-Discovery in Databases，简称：KDD)中的一个步骤。数据挖掘一般是指从大量的数据中通过算法搜索隐藏于其中信息的过程。数据挖掘通常与计算机科学有关，并通过统计、在线分析处理、情报检索、机器学习、专家系统（依靠过去的经验法则）和模式识别等诸多方法来实现上述目标。可以理解为从数据中发现现象，进而对现象产生的原因进行假设，通过数据以及数学方法对自己的假设进行论证。

**分析方法：**

[](http://baike.baidu.com/picture/7893/7893/0/718e25c70206f79ed10060da.html?fr=lemma&ct=single)

数据挖掘的方法：

· 分类 （Classification）

· 估计（Estimation）

· 预测（Prediction）

· 相关性分组或[关联规则](http://baike.baidu.com/view/1076817.htm)（Affinity grouping or association rules）

· 聚类（Clustering）

· 复杂数据类型挖掘(Text， Web ，图形图像，视频，音频等)

**方法简介：**

·**分类 （Classification）**

首先从数据中选出已经分好类的训练集，在该训练集上运用数据挖掘分类的技术，建立分类模型，对于没有分类的数据进行分类。

例子：

a 信用卡申请者，分类为低、中、高风险

b 故障诊断：[中国](http://baike.baidu.com/view/61891.htm)宝钢集团与上海天律信息技术有限公司合作，采用数据挖掘技术对钢材生产的全流程进行质量监控和分析，构建故障地图，实时分析产品出现瑕疵的原因，有效提高了产品的优良率。

注意： 类的个数是确定的，预先定义好的

**· 估计（Estimation）**

估计与分类类似，不同之处在于，分类描述的是离散型变量的输出，而估值处理连续值的输出；

分类的类别是确定数目的，估值的量是不确定的。

例子：

a 根据购买模式，估计一个家庭的孩子个数

b 根据购买模式，估计一个家庭的收入

c 估计real estate的价值

一般来说，估值可以作为分类的前一步工作。给定一些输入数据，通过估值，得到未知的[连续变量](http://baike.baidu.com/view/618491.htm)的值，然后，根据预先设定的阈值，进行分类。例如：银行对家庭贷款业务，运用估值，给各个[客户](http://baike.baidu.com/view/88584.htm)记分（Score 0~1）。然后，根据[阈值](http://baike.baidu.com/view/409216.htm)，将贷款级别分类。

**· 预测（Prediction）**

通常，预测是通过分类或估值起作用的，也就是说，通过分类或估值得出模型，该模型用于对未知变量的预言。从这种意义上说，预言其实没有必要分为一个单独的类。预言其目的是对未来未知变量的预测，这种预测是需要时间来验证的，即必须经过一定时间后，才知道预言准确性是多少。

**· 相关性分组或关联规则（Affinity grouping or association rules）**

决定哪些事情将一起发生。

例子：

a 超市中客户在购买A的同时，经常会购买B，即A => B(关联规则)

b 客户在购买A后，隔一段时间，会购买B （序列分析）

**· 聚类（Clustering）**

聚类是对记录分组，把相似的记录在一个聚集里。聚类和分类的区别是聚集不依赖于预先定义好的类，不需要训练集。

例子：

a 一些特定症状的聚集可能预示了一个特定的疾病

b 租VCD类型不相似的客户聚集，可能暗示成员属于不同的亚文化群

聚集通常作为数据挖掘的第一步。例如，"哪一种类的促销对客户响应最好？"，对于这一 类问题，首先对整个客户做聚集，将客户分组在各自的聚集里，然后对每个不同的聚集，回答问题，可能效果更好。

**· 描述和可视化（Description and Visualization）**

是对数据挖掘结果的表示方式。一般只是指数据可视化工具，包含报表工具和商业智能分析产品（BI）的统称。譬如通过Yonghong Z-Suite等工具进行数据的展现，分析，钻取，将数据挖掘的分析结果更形象，深刻的展现出来。

数据挖掘的几个规律：

第一，目标律：业务目标是所有数据解决方案的源头。

第二，知识律：业务知识是数据挖掘过程每一步的核心。

第三，准备律：数据预处理比数据挖掘其他任何一个过程都重要。

第四，试验律(NFL律：No Free Lunch)：对于数据挖掘者来说，天下没有免费的午餐，一个正确的模型只有通过试验(experiment)才能被发现。

第五，模式律(大卫律)：数据中总含有模式。

第六，洞察律：数据挖掘增大对业务的认知。

第七，预测律：预测提高了信息泛化能力。

第八，价值律：数据挖掘的结果的价值不取决于模型的稳定性或预测的准确性。

第九，变化律：所有的模式因业务变化而变化。

**NO.1 Data Mining 和统计分析有什么不同？**

硬要去区分Data Mining和Statistics的差异其实是没有太大意义的。一般将之定义为Data Mining技术的CART、CHAID或[模糊计算](http://baike.baidu.com/view/1183818.htm)等等理论方法，也都是由统计学者根据统计理论所发展衍生，换另一个角度看，Data Mining有相当大的比重是由高等统计学中的多变量分析所支撑。但是为什么Data Mining的出现会引发各领域的广泛注意呢？主要原因在相较于传统统计分析而言，Data Mining有下列几项特性：

1处理大量实际数据更强势，且无须太专业的统计背景去使用Data Mining的工具；

2数据分析趋势为从大型数据库抓取所需数据并使用专属计算机分析软件，Data Mining的工具更符合[企业需求](http://baike.baidu.com/view/177964.htm)；

3纯就理论的基础点来看，Data Mining和统计分析有应用上的差别，毕竟Data Mining目的是方便企业终端用户使用而非给统计学家检测用的。

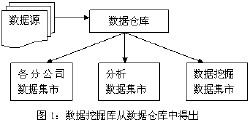
**NO.2 数据仓库和数据挖掘的关系为何？**

若将Data Warehousing（数据仓库）比喻作矿坑，Data Mining就是深入矿坑采 矿的工作。毕竟Data Mining不是一种无中生有的[魔术](http://baike.baidu.com/view/4503.htm)，也不是点石成金的炼金术，若没有够丰富完整的数据，是很难期待Data Mining能挖掘出什么有意义的信息的。

要将庞大的数据转换成为有用的信息，必须先有效率地收集信息。随着科技的进步，功能完善的数据库系统就成了最好的收集数据的工具。数据仓库，简单地说，就是搜集来自其它系统的有用数据，存放在一整合的储存区内。所以其实就是一个经过处理整合，且容量特别大的关系型数据库，用以储存[决策支持系统](http://baike.baidu.com/view/8048.htm)（Decision Support System）所需的数据，供决策支持或数据分析使用。从[信息技术](http://baike.baidu.com/view/3226.htm)的角度来看，数据仓库的目标是在组织中，在正确的时间，将正确的数据交给正确的人。

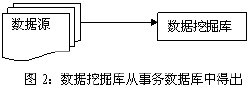
许多人对于Data Warehousing和Data Mining时常混淆，不知如何分辨。其实，数据仓库是数据库技术的一个新主题，利用[计算机系统](http://baike.baidu.com/view/1130583.htm)帮助我们操作、计算和思考，让作业方式改变，决策方式也跟着改变。

数据仓库本身是一个非常大的数据库，它储存着由组织作业数据库中整合而来的数据，特别是指[事务处理系统](http://baike.baidu.com/view/1363167.htm)OLTP（On-Line Transactional Processing）所得来的数据。将这些整合过的数据置放于数据仓库中，而公司的决策者则利用这些数据作决策；但是，这个转换及整合数据的过程，是建立一个数据仓库最大的挑战。因为将作业中的[数据转换](http://baike.baidu.com/view/286832.htm)成有用的的策略性信息是整个数据仓库的重点。综上所述，数据仓库应该具有这些数据：整合性数据（integrated data）、详细和汇总性的数据(detailed and summarized data)、历史数据、解释数据的数据。从数据仓库挖掘出对决策有用的信息与知识，是建立数据仓库与使用Data Mining的最大目的，两者的本质与过程是两回事。换句话说，数据仓库应先行建立完成，Data mining才能有效率的进行，因为数据仓库本身所含数据是干净(不会有错误的数据参杂其中)、完备，且经过整合的。因此两者关系或许可解读为Data Mining是从巨大数据仓库中找出有用信息的一种过程与技术。

[](http://baike.baidu.com/picture/7893/7893/0/8c1001e93901213f8eb83c9f54e736d12e2e95ed.html?fr=lemma&ct=single)

大部分情况下，数据挖掘都要先把数据从数据仓库中拿到数据挖掘库或数据集市中（见图1）。

从数据仓库中直接得到进行数据挖掘的数据有许多好处。就如我们后面会讲到的，数据仓库的数据清理和数据挖掘的数据清理差不多，如果数据在导入数据仓库时已经清理过，那很可能在做数据挖掘时就没必要在清理一次了，而且所有的数据不一致的问题都已经被你解决了。

[](http://baike.baidu.com/picture/7893/7893/0/18d8bc3eb13533faea43e16da8d3fd1f40345bff.html?fr=lemma&ct=single)

数据挖掘库可能是你的数据仓库的一个逻辑上的子集，而不一定非得是物理上单独的数据库。但如果你的数据仓库的计算资源已经很紧张，那你最好还是建立一个单独的数据挖掘库。

当然为了数据挖掘你也不必非得建立一个数据仓库，数据仓库不是必需的。建立一个巨大的数据仓库，把各个不同源的数据统一在一起，解决所有的数据冲突问题，然后把所有的数据导到一个数据仓库内，是一项巨大的工程，可能要用几年的时间花上百万的钱才能完成。只是为了数据挖掘，你可以把一个或几个事务数据库导到一个只读的数据库中，就把它当作[数据集市](http://baike.baidu.com/view/32574.htm)，然后在他上面进行数据挖掘。

**NO.3 OLAP 能不能代替 Data Mining？**

所谓OLAP（Online Analytical Process）意指由数据库所连结出来的在线分析处理程序。有些人会说：「我已经有OLAP的工具了，所以我不需要Data Mining。」事实上两者间是截然不同的，主要差异在于Data Mining用在产生假设，OLAP则用于查证假设。简单来说，OLAP是由使用者所主导，使用者先有一些假设，然后利用OLAP来查证假设是否成立；而Data Mining则是用来帮助使用者产生假设。所以在使用OLAP或其它Query的工具时，使用者是自己在做探索（Exploration），但Data Mining是用工具在帮助做探索。

举个例子来看，一市场分析师在为超市规划货品架柜摆设时，可能会先假设婴儿尿布和婴儿奶粉会是常被一起购买的产品，接着便可利用OLAP的工具去验证此假设是否为真，又成立的证据有多明显；但Data Mining则不然，执行Data Mining的人将庞大的结帐[数据整理](http://baike.baidu.com/view/582722.htm)后，并不需要假设或期待可能的结果，透过Mining技术可找出存在于数据中的潜在规则，于是我们可能得到例如尿布和啤酒常被同时购买的意料外之发现，这是OLAP所做不到的。

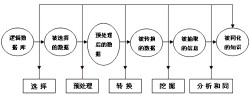
Data Mining常能挖掘出超越归纳范围的关系，但OLAP仅能利用人工查询及可视化的报表来确认某些关系，是以Data Mining此种自动找出甚至不会被怀疑过的[数据模型](http://baike.baidu.com/view/72142.htm)与关系的特性，事实上已超越了我们经验、教育、想象力的限制，OLAP可以和Data Mining互补，但这项特性是Data Mining无法被OLAP取代的。

**NO.4 完整的Data Mining 包含哪些步骤？**

1、数据挖掘环境

数据挖掘是指一个完整的过程，该过程从大型数据库中挖掘先前未知的，有效的，可实用的信息，并使用这些信息做出决策或丰富知识。

数据挖掘环境可示意如下图:

[](http://baike.baidu.com/picture/7893/7893/0/b3fb43166d224f4ad26e015009f790529922d1fa.html?fr=lemma&ct=single)

数据挖掘环境框图

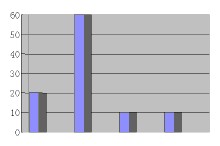
2、数据挖掘过程图

下图描述了数据挖掘的基本过程和主要步骤

数据挖掘的基本过程和主要步骤

3、数据挖掘过程工作量

在数据挖掘中被研究的业务对象是整个过程的基础，它驱动了整个数据挖掘过程，也是检验最后结果和指引分析人员完成数据挖掘的依据和顾问。图2各步骤是按一定顺序完成的，当然整个过程中还会存在步骤间的反馈。数据挖掘的过程并不是自动的，绝大多数的工作需要人工完成。图3给出了各步骤在整个过程中的工作量之比。可以看到，60%的时间用在数据准备上，这说明了数据挖掘对数据的严格要求，而后挖掘工作仅占总工作量的10%。

[](http://baike.baidu.com/picture/7893/7893/0/267f9e2f07082838cdd1ae57b899a9014d08f191.html?fr=lemma&ct=single)

数据挖掘过程工作量比例

4、数据挖掘过程简介

过程中各步骤的大体内容如下:

(1) 确定业务对象

清晰地定义出业务问题，认清数据挖掘的目的是数据挖掘的重要一步。挖掘的最后结构是不可预测的，但要探索的问题应是有预见的，为了数据挖掘而数据挖掘则带有盲目性，是不会成功的。

(2) 数据准备

1)、数据的选择

搜索所有与业务对象有关的内部和外部数据信息，并从中选择出适用于数据挖掘应用的数据。

2)、数据的预处理

研究数据的质量，为进一步的分析作准备。并确定将要进行的挖掘操作的类型。

3)、数据的转换

将数据转换成一个分析模型。这个分析模型是针对挖掘算法建立的。建立一个真正适合挖掘算法的分析模型是数据挖掘成功的关键。

(3) 数据挖掘

对所得到的经过转换的数据进行挖掘。除了完善从选择合适的挖掘算法外，其余一切工作都能自动地完成。

(4) 结果分析

解释并评估结果。其使用的分析方法一般应作数据挖掘操作而定，通常会用到可视化技术。

(5) 知识的同化

将分析所得到的知识集成到业务信息系统的[组织结构](http://baike.baidu.com/view/543252.htm)中去。

5、数据挖掘需要的人员

数据挖掘过程的分步实现，不同的步会需要是有不同专长的人员，他们大体可以分为三类。

业务分析人员:要求精通业务，能够解释业务对象，并根据各业务对象确定出用于数据定义和挖掘算法的业务需求。

数据分析人员:精通数据分析技术，并对统计学有较熟练的掌握，有能力把业务需求转化为数据挖掘的各步操作，并为每步操作选择合适的技术。

数据管理人员:精通数据管理技术，并从数据库或[数据仓库](http://baike.baidu.com/view/19711.htm)中收集数据。

从上可见，数据挖掘是一个多种专家合作的过程，也是一个在资金上和技术上高投入的过程。这一过程要反复进行牞在反复过程中，不断地趋近事物的本质，不断地优先问题的解决方案。数据重组和细分添加和拆分记录选取数据样本可视化数据探索[聚类分析](http://baike.baidu.com/view/903740.htm)神经网络、决策树数理统计、时间序列结论综合解释评价数据知识数据取样数据探索数据调整模型化评价。

**NO.5 Data Mining 运用了哪些理论与技术？**

Data Mining是近年来数据库应用技术中相当热门的议题，看似神奇、听来时髦，实际上却也不是什么新东西，因其所用之诸如预测模型、[数据分割](http://baike.baidu.com/view/4466818.htm)，连结分析（Link Analysis）、偏差侦测（Deviation Detection）等，美国早在二次世界大战前就已应用运用在人口普查及军事等方面。

随着信息科技超乎想象的进展，许多新的计算机分析工具问世，例如关系型数据库、模糊计算理论、基因算法则以及类[神经网络](http://baike.baidu.com/view/5348.htm)等，使得从数据中发掘宝藏成为一种系统性且可实行的程序。

一般而言，Data Mining的理论技术可分为传统技术与改良技术两支。传统技术以统计分析为代表，统计学内所含序列统计、[概率论](http://baike.baidu.com/view/45337.htm)、回归分析、类别数据分析等都属于传统数据挖掘技术，尤其 Data Mining 对象多为变量繁多且样本数庞大的数据，是以高等统计学里所含括之多变量分析中用来精简变量的因素分析（Factor Analysis）、用来分类的[判别分析](http://baike.baidu.com/view/1394284.htm)（Discriminant Analysis），以及用来区隔群体的分群分析（Cluster Analysis）等，在Data Mining过程中特别常用。

在改良技术方面，应用较普遍的有[决策树](http://baike.baidu.com/view/589872.htm)理论（Decision Trees）、类神经网络（Neural Network）以及规则归纳法（Rules Induction）等。决策树是一种用树枝状展现数据受各变量的影响情形之[预测模型](http://baike.baidu.com/view/1590251.htm)，根据对目标变量产生之效应的不同而建构分类的规则，一般多运用在对客户数据的分析上，例如针对有回函与未回含的邮寄对象找出影响其分类结果的变量组合，常用[分类方法](http://baike.baidu.com/view/8348989.htm)为CART（Classification and Regression Trees）及CHAID（Chi-Square Automatic Interaction Detector）两种。

类神经网络是一种仿真[人脑](http://baike.baidu.com/view/79294.htm)思考结构的数据分析模式，由输入之变量与数值中自我学习并根据学习经验所得之知识不断调整参数以期建构数据的型样(patterns)。类神经网络为非线性的设计，与传统回归分析相比，好处是在进行分析时无须限定模式，特别当数据变量间存有[交互效应](http://baike.baidu.com/view/3816523.htm)时可自动侦测出；缺点则在于其分析过程为一[黑盒子](http://baike.baidu.com/view/111845.htm)，故常无法以可读之模型格式展现，每阶段的加权与转换亦不明确，是故类神经网络[多利](http://baike.baidu.com/view/321386.htm)用于数据属于高度非线性且带有相当程度的变量交感效应时。

规则归纳法是知识发掘的领域中最常用的格式，这是一种由一连串的「如果…/则…（If / Then）」之逻辑规则对数据进行细分的技术，在实际运用时如何界定规则为有效是最大的问题，通常需先将数据中发生数太少的项目先剔除，以避免产生无意义的逻辑规则。

**NO.6 Data Mining包含哪些主要功能？**

Data Mining实际应用功能可分为三大类六分项来说明：Classification和Clustering属于分类区隔类；Regression和Time-series属于推算预测类；Association和Sequence则属于序列规则类。

Classification是根据一些变量的数值做计算，再依照结果作分类。（计算的结果最后会被分类为几个少数的离散数值，例如将一组数据分为 "可能会响应" 或是 "可能不会响应" 两类）。Classification常被用来处理如前所述之邮寄对象筛选的问题。我们会用一些根据历史经验已经分类好的数据来研究它们的特征，然后再根据这些特征对其他未经分类或是新的数据做预测。这些我们用来寻找特征的已[分类数据](http://baike.baidu.com/view/1072506.htm)可能是来自我们的现有的客户数据，或是将一个完整数据库做部份取样，再经由实际的运作来测试；譬如利用一个大型邮寄对象数据库的部份取样来建立一个Classification Model，再利用这个Model来对数据库的其它数据或是新的数据作分类预测。

Clustering用在将数据分群，其目的在于将群间的差异找出来，同时也将群内成员的相似性找出来。Clustering与Classification不同的是，在分析前并不知道会以何种方式或根据来分类。所以必须要配合专业领域知识来解读这些分群的意义。

Regression是使用一系列的现有数值来预测一个连续数值的可能值。若将范围扩大亦可利用Logistic Regression来预测类别变量，特别在广泛运用现代分析技术如类神经网络或决策树理论等分析工具，推估预测的模式已不在止于传统线性的局限，在预测的功能上大大增加了选择工具的弹性与应用范围的广度。

Time-Series Forecasting与Regression功能类似，只是它是用现有的数值来预测未来的数值。两者最大差异在于Time-Series所分析的数值都与时间有关。Time-Series Forecasting的工具可以处理有关时间的一些特性，譬如时间的周期性、阶层性、季节性以及其它的一些特别因素（如过去与未来的关连性）。

Association是要找出在某一事件或是数据中会同时出现的东西。举例而言，如果A是某一事件的一种选择，则B也出现在该事件中的[机率](http://baike.baidu.com/view/3204218.htm)有多少。（例如：如果顾客买了火腿和柳橙汁，那么这个顾客同时也会买牛奶的机率是85%。）

Sequence Discovery与Association关系很密切，所不同的是Sequence Discovery中事件的相关是以时间因素来作区隔（例如：如果A股票在某一天上涨12%，而且当天股市加权指数下降，则B股票在两天之内上涨的机率是 68%）。

**NO.7 Data Mining在各领域的应用情形为何？**

Data Mining在各领域的应用非常广泛，只要该产业拥有具分析价值与需求的数据仓储或数据库，皆可利用Mining工具进行有目的的挖掘分析。一般较常见的应用案例多发生在[零售业](http://baike.baidu.com/view/175601.htm)、直效行销界、制造业、财务金融保险、通讯业以及医疗服务等。

于销售数据中发掘顾客的消费习性，并可藉由交易纪录找出顾客偏好的产品组合，其它包括找出流失顾客的特征与推出新产品的时机点等等都是零售业常见的实例；直效行销强调的分众概念与数据库行销方式在导入Data Mining的技术后，使直效行销的发展性更为强大，例如利用Data Mining分析顾客群之消费行为与交易纪录，结合基本数据，并依其对品牌价值等级的高低来区隔顾客，进而达到差异化行销的目的；制造业对Data Mining的需求多运用在品质控管方面，由制造过程中找出影响产品品质最重要的因素，以期提高作业流程的效率。

近来电话公司、信用卡公司、保险公司以及股票交易商对于诈欺行为的侦测（Fraud Detection）都很有兴趣，这些行业每年因为诈欺行为而造成的损失都非常可观，Data Mining可以从一些信用不良的客户数据中找出相似特征并预测可能的诈欺交易，达到减少损失的目的。财务金融业可以利用 Data Mining来分析市场动向，并预测个别公司的营运以及股价走向。Data Mining的另一个独特的用法是在医疗业，用来预测手术、用药、诊断、或是[流程控制](http://baike.baidu.com/view/2846066.htm)的效率。

**NO.8 Web Mining 和数据挖掘有什么不同？**

如果将Web视为CRM的一个新的Channel，则Web Mining便可单纯看做Data Mining应用在[网络数据](http://baike.baidu.com/view/727321.htm)的泛称。

该如何测量一个网站是否成功？哪些内容、优惠、广告是人气最旺的？主要访客是哪些人？什么原因吸引他们前来？如何从堆积如山之大量由网络所得数据中找出让网站运作更有效率的操作因素？以上种种皆属Web Mining 分析之范畴。Web Mining 不仅只限于一般较为人所知的log file分析，除了计算网页浏览率以及访客人次外，举凡网络上的零售、财务服务、通讯服务、政府机关、医疗咨询、远距教学等等，只要由网络连结出的数据库够大够完整，所有Off-Line可进行的分析，Web Mining都可以做，甚或更可整合Off-Line及On-Line的数据库，实施更大规模的模型预测与推估，毕竟凭借[网际网络](http://baike.baidu.com/view/3964529.htm)的便利性与渗透力再配合网络行为的可追踪性与高互动特质，一对一行销的理念是最有机会在网络世界里完全落实的。

整体而言，Web Mining具有以下特性：1。[数据收集](http://baike.baidu.com/view/2922355.htm)容易且不引人注意，所谓凡走过必留下痕迹，当访客进入网站后的一切浏览行为与历程都是可以立即被纪录的；2。 以交互式个人化服务为终极目标，除了因应不同访客呈现专属设计的网页之外，不同的访客也会有不同的服务；3。 可整合外部来源数据让分析功能发挥地更深更广，除了log file、cookies、会员填表数据、线上调查数据、线上交易数据等由网络直接取得的资源外，结合实体世界累积时间更久、范围更广的资源，将使分析的结果更准确也更深入。

利用Data Mining技术建立更深入的访客数据剖析，并赖以架构精准的预测模式，以期呈现真正智能型个人化的网络服务，是Web Mining努力的方向。

**NO.9 数据挖掘在 CRM 中扮演的角色为何？**

CRM（Customer Relationship Management）是近来引起热烈讨论与高度关切的议题，尤其在直效行销的崛起与网络的快速发展带动下，跟不上CRM的脚步如同跟不上时代。事实上CRM并不算新发明，奥美直效行销推动十数年的CO（Customer Ownership）就是现在大家谈的CRM—[客户关系管理](http://baike.baidu.com/view/10090.htm)。

Data Mining应用在CRM的主要方式可对应在Gap Analysis之三个部分：

针对Acquisition Gap，可利用Customer Profiling找出客户的一些共同的特征，希望能藉此深入了解客户，藉由Cluster Analysis对客户进行分群后再通过Pattern Analysis预测哪些人可能成为我们的客户，以帮助行销人员找到正确的行销对象，进而降低成本，也提高行销的成功率。

针对Sales Gap，可利用Basket Analysis帮助了解客户的产品消费模式，找出哪些产品客户最容易一起购买，或是利用Sequence Discovery预测客户在买了某一样产品之后，在多久之内会买另一样产品等等。利用 Data Mining可以更有效的决定产品组合、产品推荐、进货量或库存量，甚或是在店里要如何摆设货品等，同时也可以用来评估促销活动的成效。

针对Retention Gap，可以由原客户后来却转成竞争对手的客户群中，分析其特征，再根据分析结果到现有客户数据中找出可能转向的客户，然后设计一些方法预防[客户流失](http://baike.baidu.com/view/3013797.htm)；更有系统的做法是藉由Neural Network根据客户的消费行为与交易纪录对[客户忠诚度](http://baike.baidu.com/view/341649.htm)进行Scoring的排序，如此则可区隔流失率的等级进而配合不同的策略。

CRM不是设一个（800）客服专线就算了，更不仅只是把一堆客户基本数据输入计算机就够，完整的CRM运作机制在相关的硬[软件系统](http://baike.baidu.com/view/8343.htm)能健全的支持之前，有太多的数据准备工作与分析需要推动。企业透过Data Mining可以分别针对策略、目标定位、操作效能与测量评估等四个切面之相关问题，有效率地从市场与顾客所搜集累积之大量数据中挖掘出对消费者而言最关键、最重要的答案，并赖以建立真正由客户需求点出发的客户关系管理。

**NO.10 目前业界常用的数据挖掘分析工具？**

Data Mining工具市场大致可分为三类：

1. 一般分析目的用的[软件](http://baike.baidu.com/view/37.htm)包

K-Miner（神通数据挖掘分析系统，MPP+SMP并行计算架构）

AlpineMiner(AlpineDataLabs)

TipDM(顶尖数据挖掘平台)

GDM（Geni-Sage Data Mining Analysis System，博通数据挖掘分析系统）

SAS Enterprise Miner

KXEN(凯森)

IBM Intelligent Miner

Unica PRW

SPSS Clementine

SGI MineSet

Oracle Darwin

Angoss KnowledgeSeeker

1. 针对特定功能或产业而研发的软件

KD1（针对零售业）

Options & Choices（针对保险业）

HNC（针对信用卡诈欺或呆帐侦测）

Unica Model 1（针对行销业）

iEM System （针对流程行业的实时历史数据）

1. 整合DSS（Decision Support Systems）/OLAP/Data Mining的大型分析系统

Cognos Scenario and Business Objects

国际相关

**[Journals]**

1．ACM Transactions on Knowledge Discovery from Data (TKDD)

2．IEEE Transactions on Knowledge and Data Engineering (TKDE)

3．Data Mining and Knowledge Discovery

4．Knowledge and Information Systems

5．Data & Knowledge Engineering

**[Conferences]**

1．SIGMOD:ACM Conference on Management of Data (ACM)

2．VLDB:International Conference on Very Large Data Bases (Morgan Kaufmann/ACM)

3．ICDE:IEEE International Conference on Data Engineering (IEEE Computer Society)

4．SIGKDD:ACM Knowledge Discovery and Data Mining (ACM)

5．WWW:International World Wide Web Conferences (W3C)

6．CIKM:ACM International Conference on Information and Knowledge Management (ACM)

7．PKDD:European Conference on Principles and Practice of Knowledge Discovery in Databases ([Springer-Verlag](http://baike.baidu.com/view/3663784.htm)LNAI)

# 附录

## SPSS如何检验正态分布

一、图示法

1、P-P图

以样本的累计频率作为横坐标，以安装正态分布计算的相应累计概率作为纵坐标，把样本值表现为直角坐标系中的散点。如果资料服从整体分布，则样本点应围绕第一象限的对角线分布。

2、Q-Q图

以样本的分位数作为横坐标，以按照正态分布计算的相应分位点作为纵坐标，把样本表现为指教坐标系的散点。如果资料服从正态分布，则样本点应该呈一条围绕第一象限对角线的直线。

以上两种方法以Q-Q图为佳，效率较高。

3、直方图

判断方法：是否以钟形分布，同时可以选择输出正态性曲线。

4、箱式图

判断方法：观测离群值和中位数。

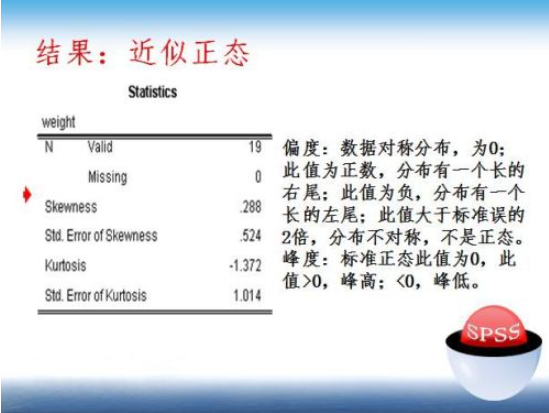
5、茎叶图

类似与直方图，但实质不同。

二、计算法

1、偏度系数（Skewness）和峰度系数（Kurtosis） 计算公式：

g1表示偏度，g2表示峰度，通过计算g1和g2及其标准误σg1及σg2然后作U检验。两种检验同时得出U<U0.05=1.96，即p>0.05的结论时，才可以认为该组资料服从正态分布。由公式可见，部分文献中所说的“偏度和峰度都接近0……可以认为……近似服从正态分布”并不严谨。



2、非参数检验方法

非参数检验方法包括Kolmogorov-Smirnov检验（D检验）和Shapiro- Wilk （W 检验）。

SAS中规定：当样本含量n ≤2000时，结果以Shapiro – Wilk（W 检验）为准，当样本含量n >2000 时，结果以Kolmogorov – Smirnov（D 检验）为准。

SPSS中则这样规定：

（1）如果指定的是非整数权重，则在加权样本大小位于3和50之间时，计算 Shapiro-Wilk 统计量。对于无权重或整数权重，在加权样本大小位于3 和 5000 之间时，计算该统计量。由此可见，部分SPSS教材里面关于“Shapiro – Wilk适用于样本量3-50之间的数据”的说法是在是理解片面，误人子弟。

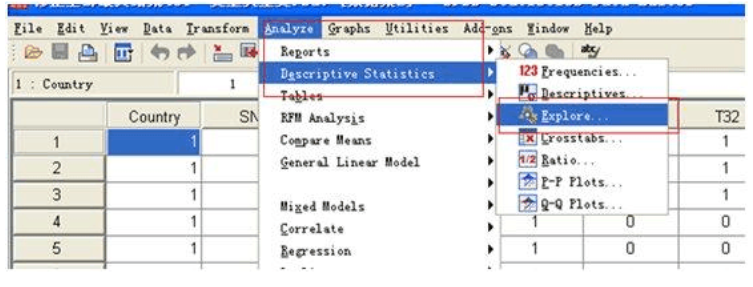
（2）单样本 Kolmogorov-Smirnov 检验可用于检验变量（例如income）是否为正态分布。

对于此两种检验，如果P值大于0.05，表明资料服从正态分布。

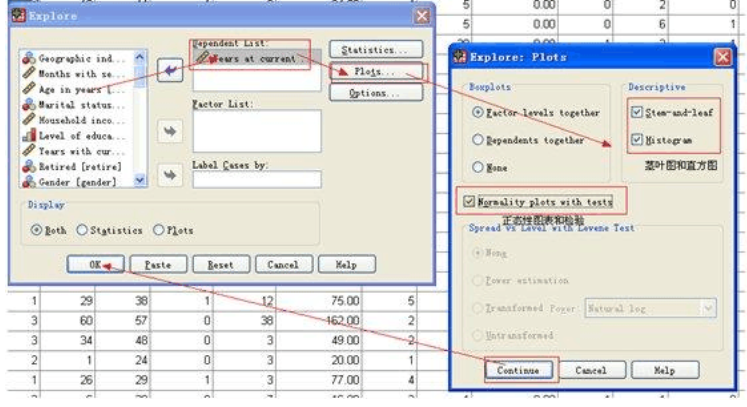
三、SPSS操作示例

SPSS中有很多操作可以进行正态检验，在此只介绍最主要和最全面最方便的操作：

1、工具栏--分析—描述性统计—探索性



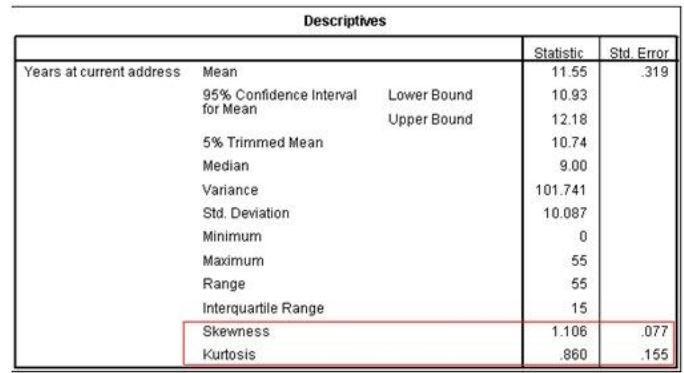
2、选择要分析的变量，选入因变量框内，然后点选图表，设置输出茎叶图和直方图，选择输出正态性检验图表，注意显示（Display）要选择双项（Both）。



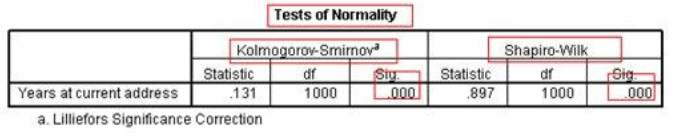
3、Output结果

（1）Descriptives：描述中有峰度系数和偏度系数，根据上述判断标准，数据不符合正态分布。

Sk=0，Ku=0时，分布呈正态，Sk>0时，分布呈正偏态，Sk<0时，分布呈负偏态，时，Ku>0曲线比较陡峭，Ku<0时曲线比较平坦。由此可判断本数据分布为正偏态（朝左偏），较陡峭。

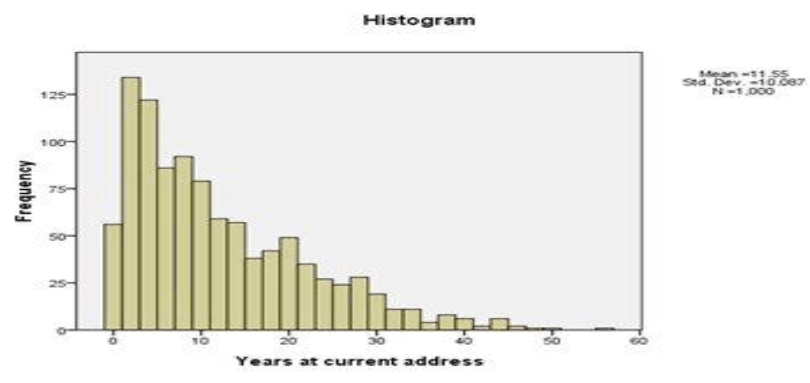


（2）Tests of Normality：D检验和W 检验均显示数据不服从正态分布，当然在此，数据样本量为1000，应以W检验为准。



（3）直方图

直方图验证了上述检验结果。



（4）此外还有茎叶图、P-P图、Q-Q图、箱式图等输出结果，不再赘述。结果同样验证数据不符合正态分布。

## 为什么要做正态分布检验

**什么是正态分布？**

正态分布是在统计分析最广泛应用的一类分布，自然界、社会、科研、生活、生产中的很多现象都被发现近似地服从正态分布，它无处不在，让你在纷繁芜杂的数据背后看到隐隐的秩序。主要指变量的频数或频率呈中间最多，两端逐渐对称地减少，表现为钟形的一种概率分布，具体的数学公式就不再提了。

**为什么要进行正态分布检验？**

假设检验可分为正态分布检验、正态总体均值分布检验、非参数检验三类。正态分布检验，即判断一样本所代表的背景总体与理论正态分布是否没有显著差异的检验，具有最重要的意义，也是应用最为广泛的检验方法。许多统计过程均依赖于总体正态性，是参数统计分析的前提。

提醒大家：对数据进行正态性检验时，大部分数据都会拒绝正态分布假设，只要数据样本大，数据接近正态分布即可接受。

**不完全符合正态分布应该怎么办？**

如果深究下去，你会发觉正态分布是最不讲理的分布，因为许多统计分析方法，都是基于正态分布的假设条件下的，而现实中完全的正态分布存在较少，大多是有偏态的，这时候怎么办，不符合正态分布，意味着不符合许多统计分析方法的适用条件，不少人在这个地方感动无奈和沮丧。

我的建议，深入研究你期望使用的统计分析方法，查看它适用条件的严苛程度，比如方差分析，它有“样本独立”“正态”“方差齐”的条件，首先，我们判断出方差分析对“样本独立”的条件是最为严苛的，但这个条件基本都满足，同时，我们了解到方差分析模型本身的稳定性比较好，所以，可以认为，在样本充足的状况下，因变量不过分的偏态（非完全正态），是可以接受的。

总结一句：对于正态分布检验的使用，应结合具体的统计分析方法，判断它对于正态检验的严苛程度，如果这个条件不允许“近似”，必须严谨对待，则考虑使用非参数检验。如果统计分析模型建模相对稳定，则可以考虑适度的降低对正态分布的严苛程度。供参考。

## spss用单样本k-s检验验证正态分布的方法

 k-s检验是用来检验数据是否符合正态分布的，因为在检验之前我们并不知道该数据是否符合正态分布，所以这种检验属于非参数检验。对要检验的数据也没有限制，只要数据量够大就可以，数据量偏少会造成检验不准确。下面来具体介绍k-s检验的过程和结果分析。

1. 在spss中打开数据，怎么打开就不介绍了，打开数据以后，在菜单栏上执行：analyze--非参检验--legacy disalogs（旧对话框）--1 sample k-s（1-样本 K-S）。
2. 将你要检验正态的变量放到test variables list，勾选下面的normal，这是正态的意思。
3. 点击exact（精确），进入选择检验方法的对话框。
4. 这里有三个选项可以选择，第一个asymptotic的意思是基于渐进分布的显著性水平的检验指标，适于大样本，如果样本过小或者分布不好，就会影响检验的效力；monte carlo适用于精确显著性水平的无偏估计，如果样本过大，数据处理过程太长，就应该使用这个选项；exact精确计算概率值，可以设定数据处理的时间，如果数据处理时间超过了你设定时间30分钟，就应该使用monte carlo。
5. 回到k-s检验对话框，点击options按钮，设置输出的参数​。
6. 勾选descriptive和quartiles，这两个的意思分别是输出描述性统计和四分位数，点击continue按钮。
7. 我们可以看到对数据的检验结果，最后的显著性检验值为0.000，非常显著。我们首先要知道在做显著性检验的时候虚无假设是数据符合正态分布，因为检验显著，所以否定虚无假设，所以数据是不符合显著性的要求的。​

辽宁省交易数的正态分布检验结果（非参数检验：单样本 K-S检验）

| **单样本 Kolmogorov-Smirnov 检验** | | |
| --- | --- | --- |
|  | | VAR00001 |
| N | | 12 |
| 正态参数a,b | 均值 | 1142776.1667 |
| 标准差 | 311173.24057 |
| 最极端差别 | 绝对值 | .299 |
| 正 | .299 |
| 负 | -.209 |
| Kolmogorov-Smirnov Z | | 1.037 |
| 渐近显著性(双侧) | | .233 |
| a. 检验分布为正态分布。  b. 根据数据计算得到。 | | |

K-S Z:1.037

渐进显著性（双侧）[P值]：0.23>0.05 可认为近似正态分布。

## SPSS T检验的应用场景与结果解读

## 如何利用SPSS对数据进行分组？