# 数据分析方法


北京华软新元信息技术有限公司

2009-03-18

# 目　　录

# 数据分析方法

## 第一节　数据分析指标

数据可分为定性数据（Qualitative Data）和定量数据（Quantitative Data）。这里讨论范围着眼于定量数据。对于定量数据，从数据的时间属性来看，可以被分为截面数据（Cross-sectional Data）和时间序列数据（Time series Data）。截面数据是在同一时点或近似同一时点上搜集到的数据，例如 2008 年 6 月北京市 18 个区县的收入数据；时间序列数据是在一系列时间段采集的数据，例如从 1998-2008 年北京市历年的收入数据。对于截面数据一般着重于不同群体之间的差异分析以及群体内各要素的结构分析；对于时间序列数据一般着重于观测数据的趋势分析。

数理指标：纯数理的，表征一个群体特征的指标，包括平均值，最大值，最小值，方差，集中度，基尼系数等

业务指标：在数理指标的基础之上结合业务产生的，反应业务特点的指标。如宏观税负率，边际税负率，税收弹性，GDP 等。

## 一　常用数理指标

体现特征的指标主要有平均数（mean）、标准差（standard deviation）与变异系数（variation coefficient）三个常用统计量，前者用于反映资料的集中性，即观测值以某一数值为中心而分布的性质；后两者用于反映资料的离散性，即观测值离中分散变异的性质。

### （一）平均数指标

平均数是统计学中最常用的统计量，用来表明资料中各观测值相对集中较多的中心位置。在财政收支分析中，平均数被广泛用来描述或比较各地区或部门的平均水平等，如平均赋税水平，人均教育支出水平等。平均数主要包括有算术平均数（arithmetic mean）、中位数（median）、众数（mode）、几何平均数（geometric mean）及调和平均数（harmonic mean），现分别介绍如下。

#### 1.　算术平均数

算术平均数是指资料中各观测值的总和除以观测值个数所得的商，简称平均数或均数，记为 $\bar{x}$。算术平均数可根据样本大小及分组情况而采用直接法或加权法计算。

设某一资料包含 $n$ 个观测值：$x_1$、$x_2$、…、$x_n$，则样本平均数 $\bar{x}$ 可通过下式计算：

$$\bar{x}=\frac{x_1+x_2+\cdots+x_n}{n}=\frac{\sum_{i=1}^{n}x_i}{n}$$

其中，Σ 为总和符号；$\sum_{i=1}^{n} x_i$ 表示从第一个观测值 $x_1$ 累加到第 $n$ 个观测值 $x_n$。当 $\sum_{i=1}^{n} x_i$ 在意义上已明确时，可简写为 Σ $x$，上式即可改写为：

$$\bar{x} = \frac{\sum x}{n}$$

算术平均数对个别极值反应比较灵敏，因而在某些情况下可能具有一定欺骗性，这时它有可能走样。对于严重偏态的分布，算术平均数会失去它所应有的代表性。

**2. 截尾均数**

因为算术平均数较易受极端值的影响，因此可以考虑将数据进行排序后，按照一定比例去掉最两端的数据，包括中部的数据来求平均数。如果截尾均数和原来平均数差异不大，说明数据不存在极端值，或者两侧极端值的影响正好抵消，反之，则说明数据存在极端值，截尾均数能更好的反映数据的集中趋势。常用的截尾均数有 5%截尾均数，即两端各去掉 5%的数据。

**3. 中位数**

将资料内所有观测值从小到大依次排列，位于中间的那个观测值，称为中位数，记为 $M_d$。当观测值的个数是偶数时，则以中间两个观测值的平均数作为中位数。中位数简称中数。当所获得的数据资料呈偏态分布时，中位数的代表性优于算术平均数。

中位数的计算方法：先将各观测值由小到大依次排列。

(1) 当观测值个数 n 为奇数时，(n+1)/2 位置的观测值，即 x(n+1)/2 为中位数；

$$M_d = x_{(n+1)/2}$$

(2) 当观测值个数为偶数时，n/2 和（n/2+1）位置的两个观测值之和的 1/2 为中位数，即：

$$M_d = \frac{x_{n/2} + x_{(n/2+1)}}{2}$$

**4. 几何平均数**

$n$ 个观测值相乘之积开 $n$ 次方所得的方根，称为几何平均数，记为 $G$。它主要应用于动态分析，用几何平均数比用算术平均数更能代表其平均水平。其计算公式如下：

$$G = \sqrt[n]{x_1 \cdot x_2 \cdot x_3 \cdots x_n} = (x_1 \cdot x_2 \cdot x_3 \cdots x_n)^{\frac{1}{n}}$$

为了计算方便，可将各观测值取对数后相加除以 $n$，得 $lgG$，再求 $lgG$ 的反对数，即得 $G$ 值，即

$$G = \lg^{-1}[\frac{1}{n}(\lg x_1 + \lg x_2 + \cdots + \lg x_n)]$$

**5. 众数**

资料中出现次数最多的那个观测值或次数最多一组的组中值，称为众数，记为 $M_0$。一组数据中的众数可能不止一个，也可能没有。

**6. 调和平均数**

资料中各观测值倒数的算术平均数的倒数，称为调和平均数，记为 $H$，即

$$H=\frac{1}{\frac{1}{n}(\frac{1}{x_1}+\frac{1}{x_2}+\cdots\frac{1}{x_n})}=\frac{1}{\frac{1}{n}\sum\frac{1}{x}}$$

对于同一资料，算术平均数>几何平均数>调和平均数。

上述五种平均数，最常用的是算术平均数。

（算术、几何、调和）平均数，中位数，众数都是描述数据的集中趋势的特征数，它们各自特点如下；

（1）用平均数作为一组数据的代表，比较可靠和稳定，它与这组数据中的每一个数都有关系，对这些数据所包含的信息的反映最为充分，因而应用最为广泛，特别是在进行统计推断时有重要作用，但易受极端数据的影响.

（2）用众数作为一组数据的代表，可靠性较差，但众数不受极端数据的影响，并且求法简便，当一组数据中个别数据变动较大时，适宜选择众数来表示这组数据的“集中趋势”。

（3）用中位数作为一组数据的代表，可靠性也较差，中位数也不受极端数据的影响，也可选择中位数来表示这组数据的“集中趋势”。用中位数来描述连续变量会损失很多信息对于样本较小的时，中位数会不太稳定，并不是一个很好的选择。

平均数的最灿烂之处在于它高度浓缩了数据的精华，使大量观测数据转变为一个代表性数值。数据中任何频次、次序和数值大小的变化都会引起平均数的改变。但平均数在高度概括数据从而使问题简单化的同时，却失去了某些有用的信息，比如它把各个数据间的差异性掩盖起来，这样仅仅依靠平均数就不能完整反映数据特征。

## （二）波动差异指标

用平均数作为样本的代表，其代表性的强弱受样本资料中各观测值变异程度的影响。如果各观测值变异小，则平均数对样本的代表性强；如果各观测值变异大，则平均数代表性弱。因而仅用平均数对一个资料的特征作统计描述是不全面的，还需引入一个表示资料中观测值变异程度大小的统计量。

**1. 全距（极差）**

表示资料中各观测值变异程度大小最简便的统计量。全距大，则资料中各观测值变异程度大，全距小，则资料中各观测值变异程度小。但是全距只利用了资料中的最大值和最小值，并不能准确表达资料中各观测值的变异程度，比较粗略。当资料很多而又要迅速对资料的变异程度作出判断时，可以利用全距这个统计量。

**2. 标准差（方差）**

为了准确地表示样本内各个观测值的变异程度，首先会考虑到以平均数为标准，求出各个观测值与平均数的离差，即（ $x-\bar{x}$ ），称为离均差。虽然离均差能表达一个观测值偏离平

均数的性质和程度，但因为离均差有正、有负，离均差之和为零，即 Σ （$x-\bar{x}$）=0，因而不能用离均差之和 Σ （$x-\bar{x}$）来表示资料中所有观测值的总偏离程度。为了解决离均差有正、有负，离均差之和为零的问题，可先求离均差的绝对值并将各离均差绝对值之和除以观测值 $n$ 求得平均绝对离差，即 Σ $|x-\bar{x}|/n$。虽然平均绝对离差可以表示资料中各观测值的变异程度，但由于平均绝对离差包含绝对值符号，使用很不方便，在统计学中未被采用。我们还可以采用将离均差平方的办法来解决离均差有正、有负，离均差之和为零的问题。先将各个离均差平方，即 $(x-\bar{x})^2$，再求离均差平方和，即 Σ $(x-\bar{x})^2$，简称平方和，记为 $SS$；由于离差平方和常随样本大小而改变，为了消除样本大小的影响，用平方和除以样本大小，即 Σ $(x-\bar{x})^2/n$，求出离均差平方和的平均数；为了使所得的统计量是相应总体参数的无偏估计量，统计学证明，在求离均差平方和的平均数时，分母不用样本含量 $n$，而用自由度 $n-1$，于是，我们采用统计量 Σ $(x-\bar{x})^2/n-1$ 表示资料的变异程度。统计量 Σ $(x-\bar{x})^2/n-1$ 称为均方差（mean square 缩写为 MS），又称样本方差，记为 $S^2$，即

$$S^2=\sum(x-\bar{x})^2/n-1$$

相应的总体参数叫总体方差，记为 σ $^2$。对于有限总体而言，σ $^2$ 的计算公式为：

$$\sigma^2=\sum(x-\mu)^2/N$$

由于样本方差带有原观测单位的平方单位，在仅表示一个资料中各观测值的变异程度而不作其它分析时，常需要与平均数配合使用，这时应将平方单位还原，即应求出样本方差的平方根。统计学上把样本方差 $S^2$ 的平方根叫做样本标准差，记为 $S$，即：

$$S=\sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$$

由于 $\sum(x-\bar{x})^2=\sum(x^2-2x\bar{x}+\bar{x}^2)$

$$=\sum x^2-2\bar{x}\sum x+n\bar{x}^2$$

$$=\sum x^2-2\frac{(\sum x)^2}{n}+n(\frac{\sum x}{n})^2$$

$$=\sum x^2-\frac{(\sum x)^2}{n}$$

所以（3-11）式可改写为：

$$S=\sqrt{\frac{\sum x^2-\frac{(\sum x)^2}{n}}{n-1}}$$

相应的总体参数叫总体标准差，记为 σ 。对于有限总体而言，σ 的计算公式为：

$$\sigma = \sqrt{\sum (x-\mu)^2 / N}$$

在统计学中，常用样本标准差 S 估计总体标准差 σ 。

标准差的特性：

（1）标准差的大小，受资料中每个观测值的影响，如观测值间变异大，求得的标准差也大，反之则小。

（2）在计算标准差时，在各观测值加上或减去一个常数，其数值不变。

（3）当每个观测值乘以或除以一个常数 *a*，则所得的标准差是原来标准差的 *a* 倍或 1/*a* 倍。

（4）在资料服从正态分布的条件下，资料中约有 68.26%的观测值在平均数左右一倍标准差（$\bar{x}$ ±S）范围内；约有 95.43%的观测值在平均数左右两倍标准差（$\bar{x}$ ±2S）范围内；约有 99.73%的观测值在平均数左右三倍标准差（$\bar{x}$ ±3S）范围内。也就是说全距近似地等于 6 倍标准差，可用（全距/6）来粗略估计标准差。

Z 计分：Z-Score 就是正态得分，先将变量标准化（减去均值除以标准差），根据上面推导，约有 99.73%的个体 Z 计分会小于 3，所以当我们想寻求离异点时，一般认为 Z 计分大于 3 的个体就属于离异点。

## （三）变异系数

变异系数是衡量资料中各观测值变异程度的另一个统计量。当进行两个或多个资料变异程度的比较时，如果度量单位与平均数相同，可以直接利用标准差来比较。如果单位和（或）平均数不同时，比较其变异程度就不能采用标准差，而需采用标准差与平均数的比值（相对值）来比较。标准差与平均数的比值称为变异系数，记为 $C \cdot V$。变异系数可以消除单位和（或）平均数不同对两个或多个资料变异程度比较的影响。

变异系数的计算公式为：

$$C \cdot V = \frac{S}{\bar{x}} \times 100\%$$

变异系数的大小，同时受平均数和标准差两个统计量的影响，因而在利用变异系数表示资料的变异程度时，最好将平均数和标准差也列出。

## （四）集中度

集中度最早起源于市场研究，市场集中度是指特定产业或市场的集中程度，一般用该产业或市场中较大企业、消费者占有的市场份额的大小来表示，它主要反映市场垄断程度的高低，根据贝恩的市场结构分析法，从产业内企业组合的角度对市场结构进行计量。

贝恩根据产业内前四位和前八位的产业集中度指标，对不同垄断，竞争结合程度的产业进行如下分类：

表 2—1 贝恩的产业集中度分类方法

| 集中度市场结构 | C4 值(%) | C8 值(%) |
|---|---|---|
| 寡占Ⅰ | 85%≤C4 | — |

| | | |
|---|---|---|
| 寡占Ⅱ | 75%≤C4<85% | 75%≤C8<85% |
| 寡占Ⅲ | 50%≤C4<75% | 75%≤C8<85% |
| 寡占Ⅳ | 35%≤C4<50% | 45%≤C8<75% |
| 寡占Ⅴ | 30%≤C4<35% | 40%≤C8<45% |
| 竞争型 | C4<30% | C8<40% |

表 2—2 植草益市场分类方法

| 市场结构 | | C8 值(%) |
|---|---|---|
| 粗分 | 细分 | |
| 寡占型 | 极高寡占型<br>高中寡占型 | 70%<C8<br>40%<C8<70% |
| 竞争型 | 低集中竞争型<br>分散竞争型 | 20%<C8<40%<br>C8<20% |

集中度的分析可以反映群体的集中程度，当不同群体进行比较时，就可以发现不同群体之间的几种差异情况，进而反映出群体的竞争的激烈程度，对于产业政策的制定者来说有较强的借鉴意义。当动态观测一个群体不同时期的集中度的变化时，就可以观测出这个群体的进化轨迹，对于研究这个群体的生命周期有一定的帮助。

# 二 常用业务指标

## （一）增长指标

### 1. 增长量

增长量是说明所分析的业务在一定时期内增长的绝对量的指标,是分析期与基期发展水平的差额。同比增长量是指当期值与去年同期值之间的差值，同比增长量消除了季节变动的影响。环比增长量是指当期值与上一期值之间的差值。

衍生指标：年均增长量。

n 年的年均增长量计算公式为：（第 n 年数-第 1 年数）/n-1

### 2. 增长速度，增幅

增长速度是用来反映业务的成长性相对指标，可以分为同比增长速度、环比增长速度。其中同比增长速度是当期增长量与去年同期值之比，说明当期业务水平对去年同期业务水平增长的相对程度；环比增长速度是当期增长量与前一期水平之比，说明业务分析期与相邻前期业务水平的相对增长程度。

衍生指标：年均增长速度

n 年的年均增长速度计算公式为：（第 n 年数-第 1 年数）$^{(1/n-1)}$ -1

### 3. 年增 1%对应的增长量

增长 1%对应的增长量，这个指标主要用来描述不同阶段下，保持相同增长速度的前提下应该绝对增长的数量。一般来说，基数越大，年增 1%对应的增长量也越高，所以在发展到一定阶段后，想保持绝对稳定的增长速度一般来说是不现实的。

**4. 增收贡献率**

某项目增收贡献率计算公式为：某项目增长量/所有项目总的增长量

对于增长的部分进行的进一步的结构分析，一般来讲，增收贡献率可能会与静态的结构比重有较大差异。所以这个指标一般用来寻求真正对增长起拉动作用的主导因素。

## （二）进度指标

**1. 收入进度**

收入进度=累计收入/预算，反映财政收入的预算执行情况

**2. 支出进度**

支出进度=累计支出/预算，反映财政支出的预算执行情况

**3. 时间进度**

时间进度是收入或支出进度的时间标准。对于季度数据，时间进度=季度/4，对于月度数据，时间进度=月度/12，对于日数据，时间进度=所处月度/12+所处日期/30（31）。

收入进度、支出进度通过和时间进度对比，反映了收入或支出预算执行的快慢程度。通常来讲，收入（支出）进度应和时间进度接近，以此反映预算执行较为正常。太快或太慢的进度都在一定程度上反映了预算执行不够正常。

# 三 财政经济指标

**1. 宏观税负率**

宏观税负率是从宏观经济的角度来衡量税收收入的一个静态指标。它是指某一时期国内生产总值（GDP）中税收收入所占的比率。 计算方法就是：

宏观税负率= （税收收入/GDP）*100%

宏观税负率高，说明国民收入分配中政府部门所占的份额越高，则政府发挥宏观调控制能的空间就越大。

**2. 边际税负率**

边际税负率是衡量 GDP 增长量与税收收入增长量之间的关系的一个动态指标。 它是指每新增一单位的 GDP 中所含的新增税收所占的比例。计算方法是：

边际税负率=（税收收入增长量/GDP 增长量）*100%

边际税负率和宏观税负率关系密切，当边际税负率大于同期宏观税负率时，宏观税负率就提高，反之，当边际税负率小于同期宏观税负率时，宏观税负率就下降。

**3. 收入（支出）弹性**

经济学中常用弹性概念来反映一个自变量的相对变化对于另一个因变量的相对变化的反应程度，即反映两个变量之间的相对变动。如果弹性大于零，说明两变量变化方向相同，反之则相反；如果弹性大于 1，说明因变量变化幅度大于自变量，反之则相反。

收入（支出）弹性是衡量 GDP 增长率与财政收入（支出）增长率之间的关系的一个动态指标。它反映的是 GDP 每增长 1%时，财政收入（支出）增长率的情况。计算方法是：

收入（支出）弹性=GDP 增长率/财政收入（支出）增长率

收入（支出）弹性为 1 时，收入（支出）增长与经济增长同步；收入（支出）弹性小于 1 时，收入（支出）增长慢于经济增长；收入（支出）弹性大于 1 时，收入（支出）增长快于经济增长，也就是超经济增长。短时间来看，收入（支出）增长可能会出现慢于或快于经济增长的现象，但从长期来看，最终应该稳定在同步增长的状态下，这样，经济发展也才趋于健康。

一般认为合理地收入弹性应该介于 0.8-1.2 之间。

**4. 收入（支出）的波动系数**

财政收入（支出）在时间上的变动状况是财政收入（支出）中的一个重要方面，反映了财政收入（支出）受外在政策体制等因素影响变化的情况。它的计算方法就是前面提到的变异系数，这里我们称为波动系数

波动系数越大，说明经济发展以外的影响因素不稳定，也反映了税制的发展不完善。实际经验的结果显示，国际上发达国家的财政收入（支出）波动系数普遍比发展中国家的波动系数小很多。

## 第二节 统计分析方法

数据分析的方法从技术来分大致可以分为三种，统计分析类，以基础的统计分析为主；高级分析方法，以计量经济建模理论为主；数据挖掘类，以机器学习，数据仓库等复合技术为主。

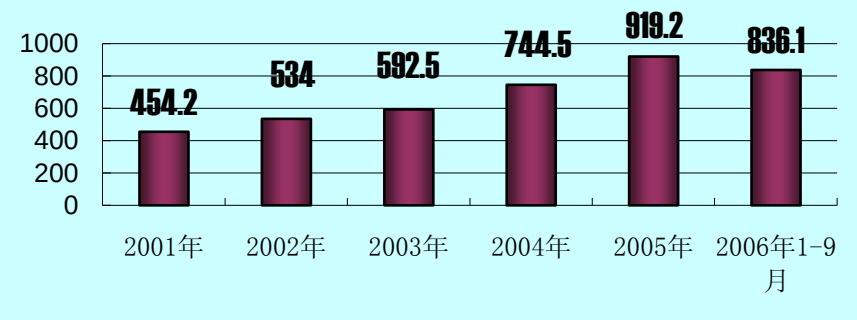
### 一 对比分析

对比分析法是把客观事物加以比较，以达到认识事物的本质和规律并做出正确的评价。对比分析法通常是把两个相互联系的指标数据进行比较，从数量上展示和说明研究对象规模的大小，水平的高低，速度的快慢，以及各种关系是否协调。

对比分析是财政收支分析中经常用到的方法，一般来说有以下几种对比方法：

#### （一）纵向对比

同一指标，不同时间下进行比较。最常用的是当期与上年同期比较，如收入同比增长，还可以与前一时期比较，如环比增长即“环比”，此外还可以与达到历史最好水平的时期或历史上一些关键时期进行比较。

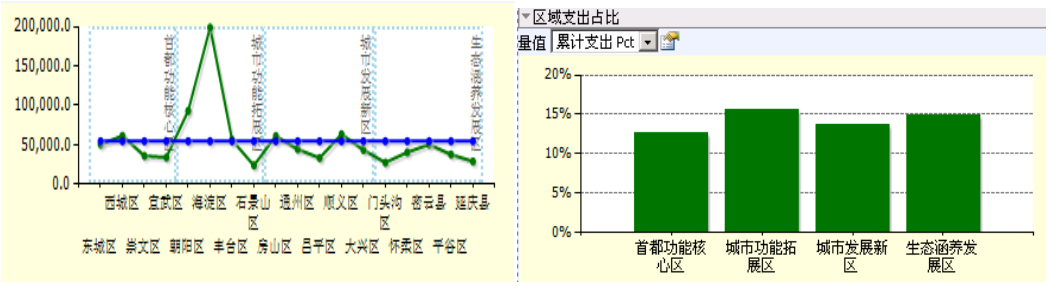



纵向对比分析示例图

如：上图分析 2006 年 9 月财政收入，纵向对比分析发现，该收入已经超过 2004 年全年收入水平。

## （二）横向对比

在同一时间下，部分与总体的对比，即比重，或是部分之间的对比。如区县支出水平与北京十八区县平均水平的对比，不同区域支出比重之间的差异等。

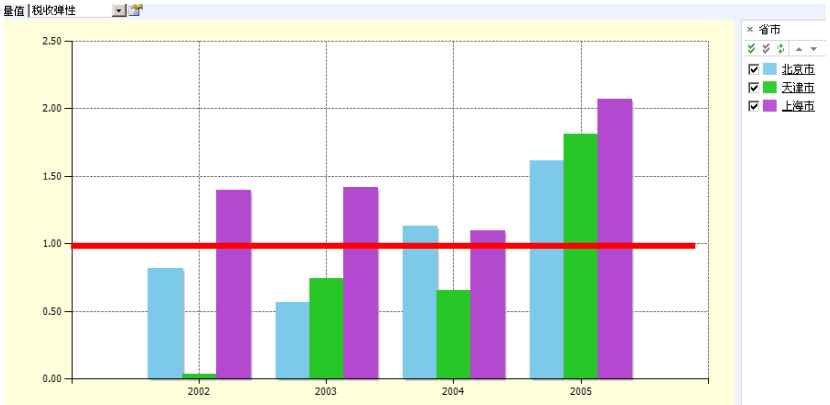



横向对比分析示例图

如上图，对某支出科目进行北京市十八区县横向对比分析，可以发现海淀区该科目支出大大超过全市平均水平，而石景山等区则低于平均水平。若分析该科目占总支出比重，则可发现四大功能区中城市功能拓展区支出比重最高，首都功能核心区比重最低。

## （三）标准对比

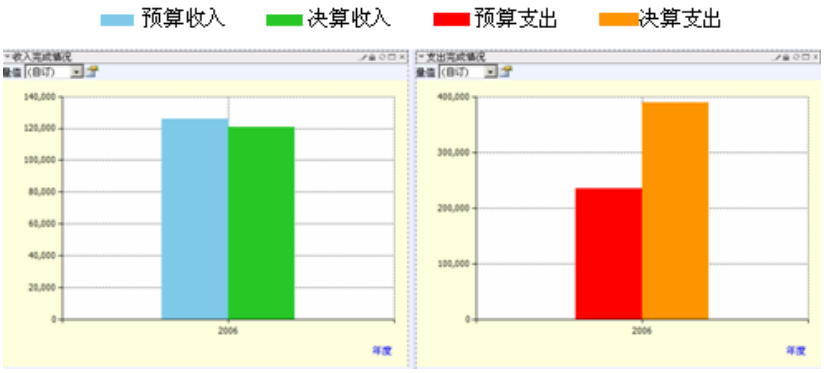
通过和由经验或理论而得出来的标准水平进行对比，了解当前的指标和标准的差异。如下图，税收经济弹性等于 1 是正常合理标准，大于或小于 1 都显示税收和经济的变化不正常。




标准对比分析示例图

### （四）实际与计划对比

当前实际值与计划数、预算数、指标数等对比，反映实际与目标值的差异。如下图财政收支决算分析中，用收支的实际数和年初预算计划值对比，可以反映预算执行的差异情况。




实际与计划对比分析示例图

对比分析对于了解财政收支数据特征，进行差异分析非常有用，可以是单指标对比，也可以是多指标的综合评价。需要注意的是对比的对象、指标、和方法上必须有可比性，根据分析目的选择合适的对比标准。适用于能够进行分组对比，具有统一的数据口径的数据。

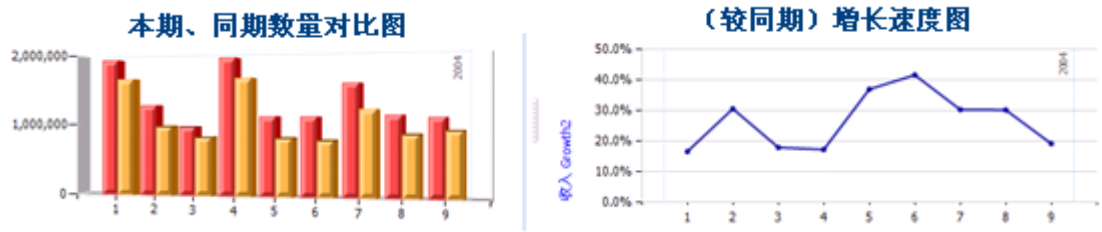
## 二　同比分析

按照时间即年度、季度、月份、日期等进行扩展，用本期实际发生数与同口径历史数字相比，产生动态相对指标，用以揭示发展水平以及增长速度。由于采用基期的不同，可分为同比、环比和定基三种分析方法，均用百分数或倍数表示。

同比分析主要是为了消除季节变动的影响，用以说明本期水平与去年同期水平对比而达到的相对值。如，本期 2 月比去年 2 月，本期 6 月比去年 6 月等。在实际工作中，经常使用这个指标，如某年、某季、某月与上年同期相比计算的发展速度，就是同比增长速度。

同比增长速度=（本期-同期）/同期×100%

同比分析可以消除季节影响带来的税收波动，但不能反映两个时间段内具体有哪些波动。




同比分析示例图

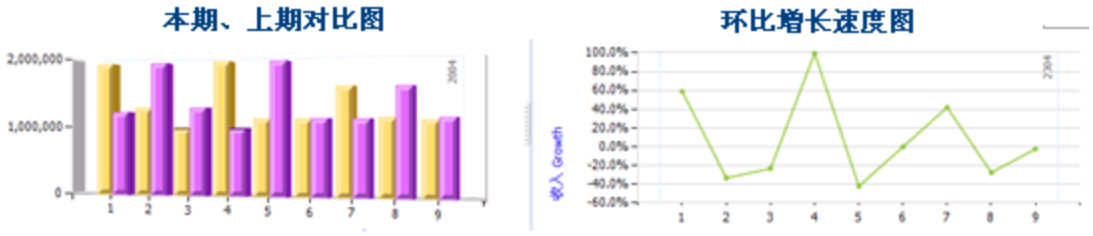
## 三　环比分析

环比分析是报告期水平与前一时期水平之比，表明现象逐期的变化趋势。如果计算一年

内各月与前一个月对比，即 2 月比 1 月，3 月比 2 月，4 月比 3 月 ……12 月比 11 月，说明逐月的变化程度。本期数据与上期数据比较，形成时间序列图。例如，按月进行环比分析，系统自动获得按月采样的指标趋势图。

环比增长速度=（本期-上期）/ 上期×100%

环比分析能够反映逐期的变化情况，但受季节影响，会出现收支的大幅度波动，不能真实反映收支增长的长期趋势。同比分析适用于受季节影响明显的时间序列数据，而环比分析适用于没有季节因素的时间序列数据。

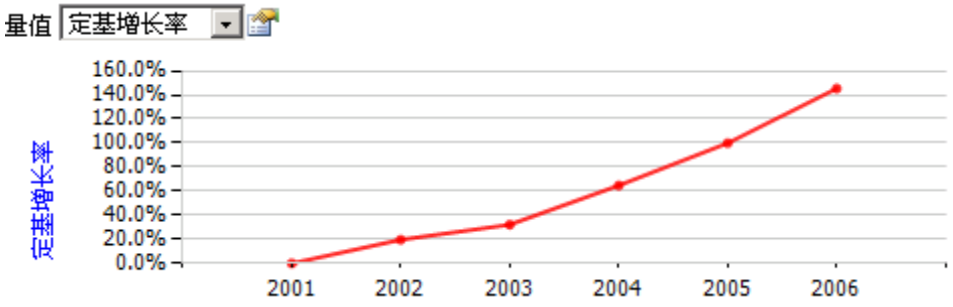



本期、环比分析示例图

## 四　定比分析

定基比分析是报告期水平与某一固定时期水平之比，表明这种现象在较长时期内总的变化水平。如，“十五”期间各年水平都以 2001 年水平为基期进行对比，在实际工作中多用于分析年度发展变化的总速度。

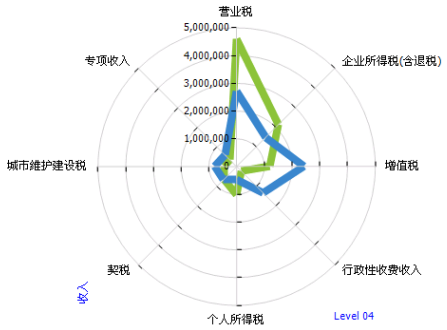
定基比增长速度=（本期-基期）/ 基期×100%




定比分析示例图

## 五　差异分析

分析两个样本之间的差异程度，雷达图分析是进行差异分析的有效手段。如：对比分析北京和收入总量较为接近的山东省收入结构差异，可以看出，北京主要是营业税和企业所得税为主，而山东则以营业税和增值税为主。

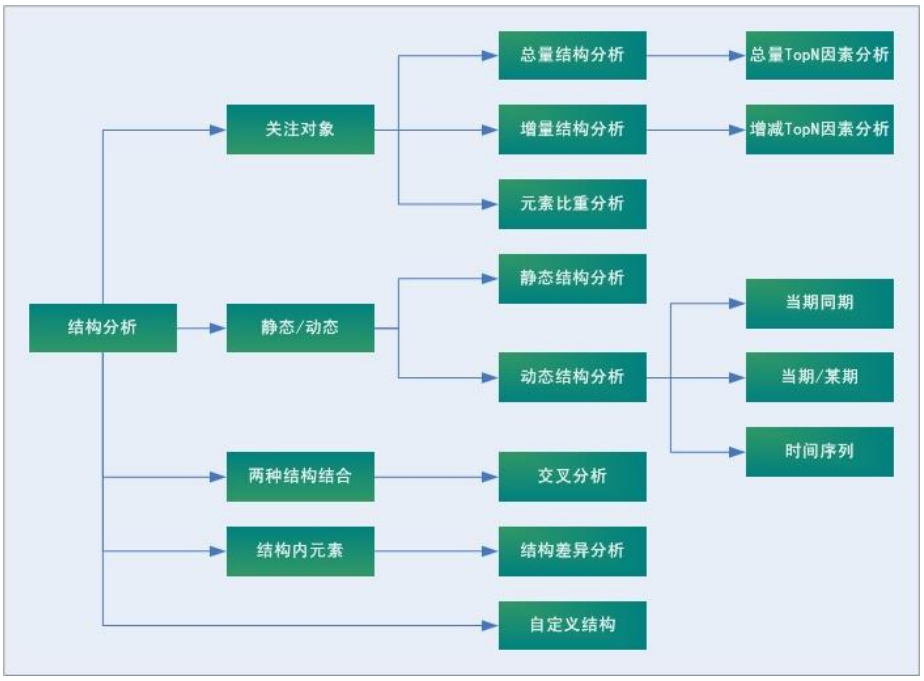



差异分析示例图

# 六　结构分析

财政收支的结构，可从多个维度进行结构分析，如区域结构、科目结构等。饼图、圆锥图和金字塔图都是开展结构分析的有效工具。

从分析方法方面来说，财政收支结构分析按所关注的对象可分为总量结构分析、增量结构分析以及所关注对象中的元素的比重分析；按所关注的时间可分为静态结构分析和动态结构分析，其中动态结构分析中又可以按当期与同期，当期与某一历史时期以及时间序列等进行结构对比分析；结构分析还可以在不同分类间进行交叉结构分析。




结构分析分类图

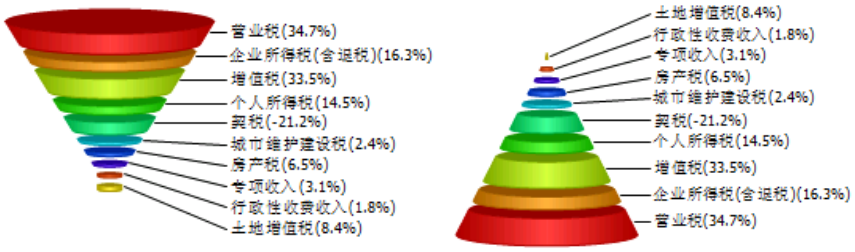
从分析内容上看，结构分析通常关注贡献情况。贡献是指某一因素的影响占总变动的比重。贡献分析在财政收支分析中的应用主要是分科目贡献率，分区域贡献率等。按实际内容又分为两种指标，总量贡献率和增长贡献率。增长贡献率有时还会出现负值，称为负贡献。

## （一）　总量贡献结构

总量贡献率是指某一因素对总变动的绝对贡献。

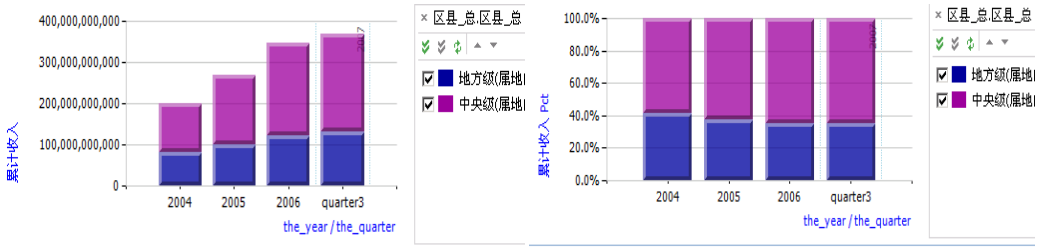
总量贡献率=分项总量/总量

如用圆锥图或金字塔图分析某市收入科目结构。




总量贡献结构示例图 1

还可以用堆积柱形图分析在某个时间段、或者某些区域之间的结构对比。在这种情况下，结构分析不仅可以做到单一结构的分析，还可以和时间序列分析或者横向的对比分析相结合。如分析某市近几年财政收入级次结构，既可以看到时间序列上的总量变化情况，还可以看到每一时间上的结构因素。



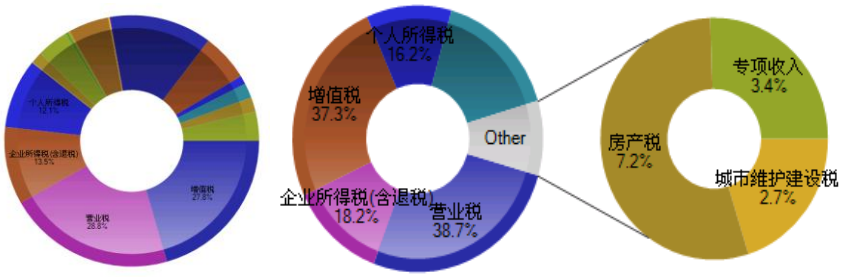

总量贡献结构示例图 2

## （二） 增长贡献结构

增长贡献率是指由于某一因素的影响使总变动增长的份额占总变动的比重。

增长贡献率=分项增量/总增量

如用饼图分析某市收入增长科目结构。




增量贡献结构示例图

# 七 因素分析

因素分析法是依据分析指标与其影响因素的关系，从数量上确定各因素对分析指标影响方向和影响程度的一种方法。因素分析法既可以全面分析各因素对某一经济指标的影响，又可以单独分析某个因素对经济指标的影响，在财政收支分析使用中颇为广泛。

因素分析的方法常见的有以下几种：（1）连环替代法。它是将分析指标分解为各个可以计量的因素，并根据各个因素之间的依存关系，顺次用各因素的比较值（通常即实际值）替代基准值（通常为标准值或计划值），据以测定各因素对分析指标的影响。（2）差额分析法。它是连环替代法的一种简化形式，是利用各个因素的比较值与基准值之间的差额，来计算各因素对分析指标的影响。（3）定基替代法。分别用分析值替代标准值，测定各因素对指标的影响。

例如，对支出进度偏慢进行因素分析，分解影响支出进度的若干因素。假设影响支出进度的因素由如下因子构成：实际指标：Po=Ao×Bo×Co；标准指标：Ps=As×Bs×Cs；实际与标准的总差异为 Po-Ps，这一总差异同时受到 A、B、C 三个因素的影响，它们各自的影响程度可分别由以下式子计算求得：

A 因素变动的影响：（Ao-As）×Bs×Cs；

B 因素变动的影响；Ao×（Bo-Bs）×Cs；

C 因素变动的影响：Ao×Bo×（Co-Cs）。

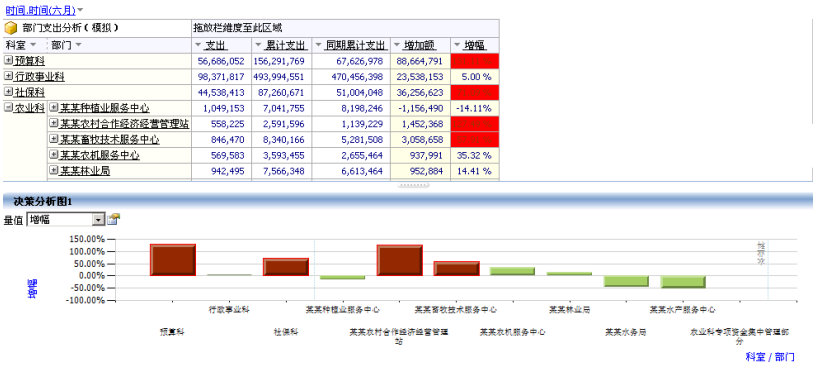
最后，可以将以上三大因素各自的影响数相加就等于总差异 Po-Ps。

因素分析通过分析财政收支变动的影响因素，从中找出主要的原因，也可借助因子分析法，将多个影响因素浓缩成较少的因素，使信息更加集中。

## 八　预警分析

根据预警条件，可以生成预警分析图，进而实现对关注条件的监控，及时发现关键点。而预警条件的设置则可以视实际情况需要而定。

例如：对某区支出单位进行预警分析，监控增幅过高单位，可以将预警条件设置为监控增幅大于 50%的单位，并对其进行重点跟踪关注。




预警分析示例图

## 九　80/20 分析（二八分析）

80/20 效率法则（the 80/20 principle），又称为帕累托法则、帕累托定律、最省力法则或不平衡原则。此法则是由意大利经济学家帕累托提出的。80/20 法则认为：原因和结果、

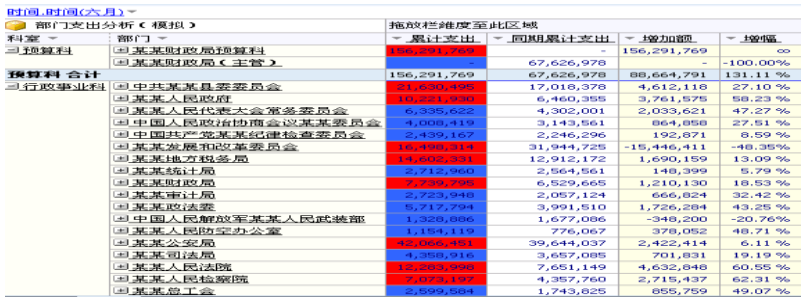
投入和产出、努力和报酬之间本来存在着无法解释的不平衡。

一般情形下，产出或报酬是由少数的原因、投入和努力所产生的。原因与结果、投入与产出、努力与报酬之间的关系往往是不平衡的。若以数学方程式测量这个不平衡，得到的基准线是一个 80/20 关系；结果、产出或报酬的 80%取决于 20%的原因、投入或努力。例如，80%的财政支出用于 20%的重点单位，80%的税收收入来源于 20%的重点税源企业。

80/20 原则包含在任何时候对原因的静态分析，而不是动态的。使用 80/20 原则的艺术在于确认哪些现实中的因素正在起作用并尽可能地被利用。 80/20 这一数据仅仅是一个比喻和实用基准。真正的比例未必正好是 80%：20%。 80/20 原则表明在多数情况下该关系很可能是不平衡的，并且接近于 80/20。

将 80/20 原则应用于财政收支分析的主要思想就是怎样以最少的代价来获取最大的利益和价值—对 20%重点税源户的关注可能带来 80%的税收收入；对 20%重点支出单位的监督可能影响 80%的支出进度。

例如：图中用红色监控某区 20%的重点支出单位，对这些单位重点关注，分析其支出构成，及时跟踪其支出进度，从而对全区 80%的总支出有较好的把握。



时间.时间(六月)

部门支出分析（模拟）　　拖放栏维度至此区域

| 科室 | 部门 | 累计支出 | 同期累计支出 | 增加额 | 增幅 |
|---|---|---|---|---|---|
| 己预算科 | 某某财政局预算科 | 156,291,769 | - | 156,291,769 | ∞ |
| | 某某财政局（主管） | - | 67,626,978 | - | -100.00% |
| 预算科 合计 | | 156,291,769 | 67,626,978 | 88,664,791 | 131.11 % |
| 己行政事业科 | 中共某某县委委员会 | 21,630,495 | 17,018,378 | 4,612,118 | 27.10 % |
| | 某某人民政府 | 10,221,930 | 6,460,355 | 3,761,575 | 58.23 % |
| | 某某人民代表大会常务委员会 | 6,335,622 | 4,302,001 | 2,033,621 | 47.27 % |
| | 中国人民政治协商会议某某委员会 | 4,008,419 | 3,143,561 | 864,858 | 27.51 % |
| | 中国共产党某某纪律检查委员会 | 2,439,167 | 2,246,296 | 192,871 | 8.59 % |
| | 某某发展和改革委员会 | 16,498,314 | 31,944,725 | -15,446,411 | -48.35% |
| | 某某地方税务局 | 14,602,331 | 12,912,172 | 1,690,159 | 13.09 % |
| | 某某统计局 | 2,712,960 | 2,564,561 | 148,399 | 5.79 % |
| | 某某财政局 | 7,739,795 | 6,529,665 | 1,210,130 | 18.53 % |
| | 某某审计局 | 2,723,948 | 2,057,124 | 666,824 | 32.42 % |
| | 某某政法委 | 5,717,794 | 3,991,510 | 1,726,284 | 43.25 % |
| | 中国人民解放军某某人民武装部 | 1,328,886 | 1,677,086 | -348,200 | -20.76% |
| | 某某人民防空办公室 | 1,154,119 | 776,067 | 378,052 | 48.71 % |
| | 某某公安局 | 42,066,451 | 39,644,037 | 2,422,414 | 6.11 % |
| | 某某司法局 | 4,358,916 | 3,657,085 | 701,831 | 19.19 % |
| | 某某人民法院 | 12,283,298 | 7,651,149 | 4,632,148 | 60.55 % |
| | 某某人民检察院 | 7,073,197 | 4,357,760 | 2,715,437 | 62.31 % |
| | 某某总工会 | 2,599,584 | 1,743,825 | 855,759 | 49.07 % |

二八分析示例图

## 第三节　高级分析方法

高级分析方法以计量经济建模理论为主，参考借鉴了前人的许多优秀的研究思想和先进的分析方法。采用的技术包括：

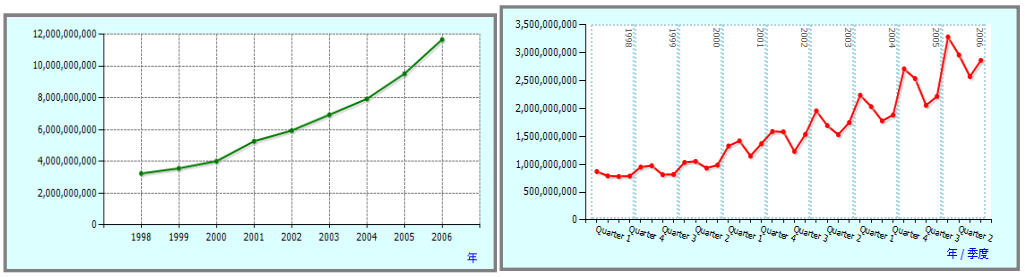
### 一　时间序列分析

按照时间的顺序把随机事件变化发展的过程记录下来就构成一个时间序列。时间序列分析就是对时间序列进行观察、研究、找寻它变化发展的规律，预示它将来的走势。

时间序列分析方法可分为描述性时序分析和统计时序分析。

描述性时序分析是通过直观的数据比较或绘图观测，寻找序列中蕴含的发展规律。如财政收入的历年增长趋势，季节波动趋势等。年度财政收入数据呈现持续稳定增长。季度财政

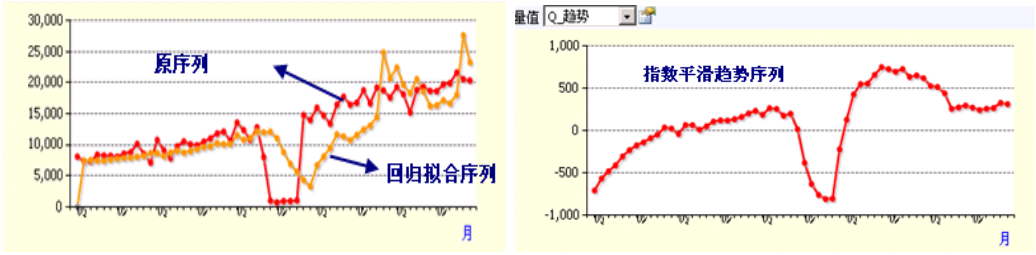
收入数据的季节性波动比较明显。




时间序列数据图示

统计时序分析的原理是：事件的发展通常都具有一定的惯性，这种惯性用统计的语言来描述就是序列值之间存在一定的相关关系，这种相关通常具有某种统计规律。时序分析方法的目的是找出时序值之间相关关系的统计规律，并拟合出适当的数学模型来描述这种规律，进而利用这个拟合模型来预测未来的走势。

对于时间序列数据（按年/按月）的柱型图、折线图，形成趋势变化，可以利用指数平滑等技术对折线图进行趋势拟合。




时间序列数据趋势拟合图

**时间序列的季节性**

传统的时间序列分析把时间序列的波动归结为四大因素，趋势变动（T），季节变动（S），循环变动（C），不规则变动（I）。

长期趋势因素（T）反映了经济现象在一个较长时间内的发展方向，它可以在一个相当长的时间内表现为一种近似直线的持续向上或持续向下或平稳的趋势，可能含转折点。

季节变动因素（S）是经济现象受季节变动影响所形成的一种长度和幅度固定的周期波动。它存在的主要原因是自然因素，另外还有行政或法律规定以及社会、文化、宗教等传统因素。

循环变动因素（C）也称周期变动因素，它是受各种经济因素影响形成的上下起伏不定的波动。通常是指周期为数年的经济周期变动。

不规则变动（I）又称随机变动，它是受各种偶然因素影响所形成的不规则变动，是一种随机变动。不规则因素在什么时间出现、影响程度和持续时间都不可预测。

对于具有季节变动规律的时间序列，如季度财政收入或月度财政收入，在进行分析时除了进行惯常的趋势因素分析外，还必须考虑其季节因素，并需要对数据作相应季节调整后再进行进一步的分析。

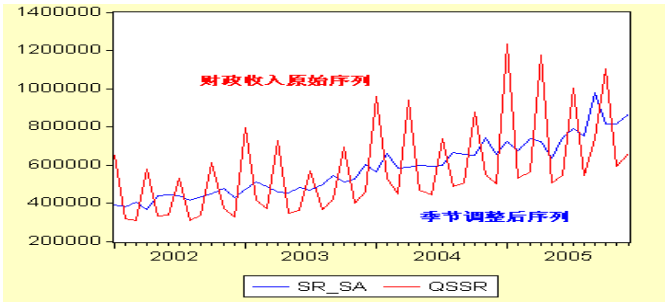
季节调整常用的是乘法模型和加法模型。

乘法模型 Y=T×S×C×I

加法模型 Y=T+S+C+I

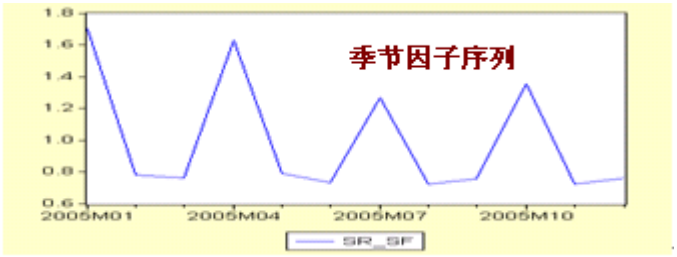
其中乘法模型适用于 T，S，C 相关的情形，比如，季节变动的幅度随趋势上升而增大。加法模型在适用于 T，S，C 相互独立的情形。要想获得合适的季节调整结果，通常需要利用连续 3-5 年的月度数据或季度数据。

SPSS、Eviews 等主流软件都可以把季节性质的时间序列分解，得到季节因子序列和季节调整后序列。例如：用 Eviews 软件附带的季节调整方法来分析受季节影响较大的月度财政收入数据。原始数据波动比较大，趋势不明显，而季节调整后的数据去除了季节效应，显示出了明显的增长趋势。



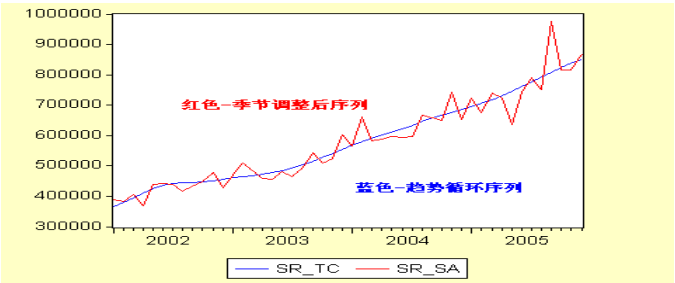

财政收入数据原始序列图

经过季节模型分解的季节因子可以看到一年的不同月份财政收入的差异。一年中 1 月最高，4 月、7 月和 10 月的收入也高于其它月份。

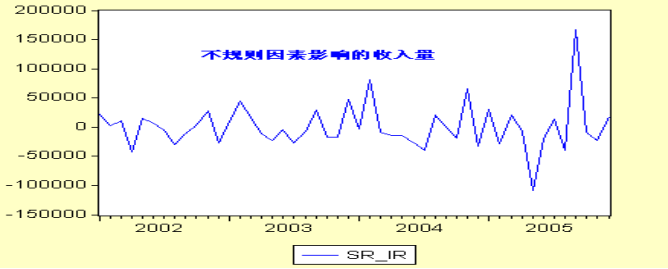



季节因子序列图

影响收入的四个因素中，一般来说对长期趋势和循环变动的分析都是以季节调整后的序列为基础的。趋势和循环因素一般难以严格区分开，常放在一起进行分析，图中较光滑的蓝色曲线即代表了趋势和循环因素影响的财政收入，反映收入的基本水平，包括长于一年的变动和循环，以及可能含有的转折点。波动较大的是季节调整后的收入序列，中间相差的数量是代表不规则因素影响的收入量。




季节调整后序列图




季节调整后不规则因素影响的收入图

## 二　聚类分析

聚类分析，是数理统计的重要方法，它是研究多要素事物分类问题的数量方法。其基本原理是，根据样本自身的属性，用数学方法按照某些相似性或差异性指标，定量地确定样本之间的亲疏关系，并按这种亲疏关系程度对样本进行聚类，通过“物以类聚”进行数据处理，为事物的分类管理提供数据支持的一种分析方法。

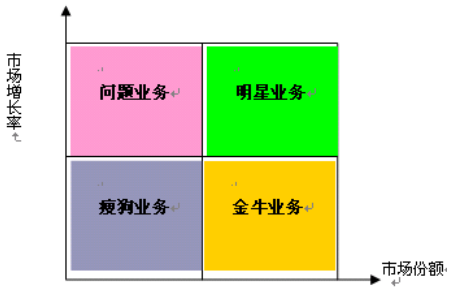
对样本进行分类，需要定义一个衡量样本之间接近程度的量，一般用距离和相似系数来衡量。在数理统计的角度上，距离和相似系数针对不同的问题有比较详细地分类，这里不作详细说明。

将聚类分析技术应用财政收支分析工作中，可以有效地利用多项指标以量化的形式对收支数据分类评价考核，避免了单一指标评价考核财政收支的局限性，从而使财政收支评价考核更具有综合性和合理性，使相关政策制定更具有针对性。

在财政收支分析中，需要聚类的目标通常是各个区县、各种行业以及预算单位等，考察这些目标的众多指标往往不能够十分清晰的做出分类，这时需要用聚类的方法进行分类，再根据每一类的特点分别展开分析。

## 三　波士顿矩阵分析

波士顿矩阵是由波士顿咨询公司在上个世纪 70 年代开发的，BCG 矩阵将组织的每一个战略事业单位标在一种二维的矩阵图上，从而显示出组织的若干产品中哪一个提供高额的潜在收益，以及哪个是组织资源的漏斗。波士顿矩阵又称市场增长率—相对市场份额矩阵、四象限分析法等。市场增长率—相对市场份额矩阵分为四个方格，每个方格代表不同类型的业务：（1）问题业务：问题业务是指高市场增长率、低市场份额的业务。（2）明星业务：明星业务是指高速增长市场中的具有高市场份额业务。(3)现金牛业务：当市场的年增长率不高，而它的市场份额却很高的业务（4）瘦狗业务：瘦狗业务是指市场增长率低缓、市场份额也低的业务。

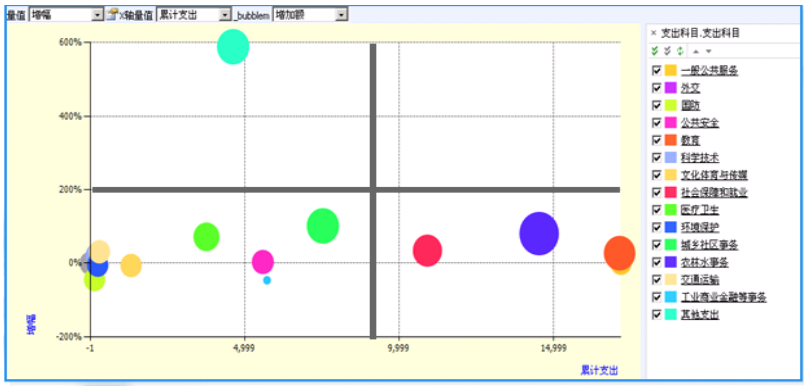



波士顿矩阵图

举例说明：波士顿矩阵法因其评估的有效性，逐渐被引入各行业的分析领域，扩大了评估对象的范围。 在财政收支分析领域里的应用可以用下面一个简单的例子说明。

一般如果按一个指标很容易能够进行分类，但是对于两个或以上指标的分类就相对复杂。波士顿矩阵就提供了一种比较快速而又直观的分类。如分析财政支出，把支出科目按照支出总量和增长速度两个指标进行分类，利用波士顿矩阵法得到这张图。把科目划分在四个象限内，

位于左上角的科目需要引起关注，总量虽然不高，但增长过快；而右下角的科目是重点科目，总量一直较大并且保持着平稳增长节奏；大部分科目则集中在总量低、平稳增长的领域。




波士顿矩阵分析财政支出图示

## 四　相关分析

对两个不同的经济变量进行相关性判断，确定经济变量之间是否存在相关关系。相关分析是进行因果分析的基本工具，通过相关分析可以判断经济指标之间的替代关系和关联度。

相关分析用来研究两个变量（x，y）的相互关系，测定它们联系的紧密程度。测定的方法可以从散点图直观的进行观察，也可以通过计算相关系数得到较为精确的判断。计算公式如下，也可以通过相应软件，如 SPSS、Eviews 软件等直接得到。

$$\gamma = \frac{n\sum xy - \sum x \sum y}{\sqrt{n\sum x^2 - \left(\sum x\right)^2} \times \sqrt{n\sum y^2 - \left(\sum y\right)^2}}$$

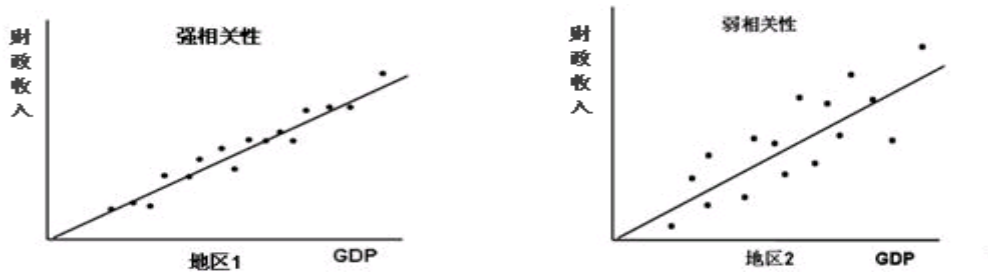
相关系数 r 的取值范围是|r|≤1，当 r=1 时完全正相关，r=-1 时完全负相关，r=0 时完全不相关。一般来说两个变量的相关系数不会出现这种极值，而是位于中间的某个区间，通常我们有这样的区间划分：

如果 |r|> 0.80 时具有强的相关关系

如果 0.3 <|r|< 0.80 时具有弱的相关关系.

如果 |r|< 0.30 时认为没有有效的相关关系.

例：考察区域的经济发展和财政收入的相互关系，为方便比较，选取两个区域的 GDP 和财政收入指标进行分析。先用散点图进行观察，可以看到地区 1 的财政收入和 GDP 存在较强的相关关系，说明该地区的财政收入和经济是密切相关的。而地区 2 的财政收入和 GDP 存在较弱的相关关系，可能因为税源结构的差异、政策等因素造成财政收入和经济的发展不是密切相关。进一步通过计算相关系数，可以看到地区 1 的财政收入和 GDP 的相关系数达到 0.953，而地区 2 的相关系数是 0.782。




强相关性&弱相关性

此外，值得注意的是，相关分析不是因果分析，没有对两个变量的因果关系进行判断，在回归分析中更强调自变量和随之而变的因变量。相关系数的计算方法是以直线关系为前提的，如果是曲线关系，则相关系数方法计算时会出现错误的结果

## 五 回归分析

回归分析是研究变量之间相关关系以及相互影响程度的一种统计推断法。通过建立自变量和因变量的方程，研究某个因素受其它因素影响的程度或用来预测。回归分析有线性和非线性回归、一元和多元回归之分。常用的回归有一元线性和多元线性回归方程。

一元线性回归方程在财政收支分析中的主要应用是建立以财政收入（或财政支出）为因变量，以国内经济总量（GDP）为自变量的一元线性方程。用以分析 GDP 对财政收入（或财政支出）的影响程度，或使用对未来 GDP 增长的判断，一方面可以通过方程预测未来的财政收入大小；另一方面可以通过方程把握财政支出与经济增长的制约关系。

多元线性回归方程在财政收支分析中的应用主要是考虑了更多其它影响财政收支的因素，如建立以财政收入为因变量，以国内经济总量（GDP）、财政支出、资本形成总额、货物或服务净出口等其它若干经济指标为自变量的多元回归方程。

建立一个回归分析一般要经历这样一个过程：先收集数据、再用散点图确认关系，利用最小二乘法或其他方法建立回归方程，检验统计参数是否合适，进行方差分析或残差分析，

优化回归方程。此外，还应考虑变量的多重共线和自相关性，以及是否有必要加入虚拟变量等。具体的方法涉及到的经济学和统计学的知识比较多，这里不作详细介绍。

例：收集某区域的财政收入、GDP 的历年数据，建立收入与 GDP 的一元线性回归分析。经过多次拟合，最终确立的回归方程是

T= 0.140× GDP + 1.762×AR(1)-0.938×AR(2)

(15.320) (11.640) (-5.097)

$R^2$=0.997 DW=2.587

方程中的 AR(1)和 AR(2)是为了消除残差自相关因素引入的变量，该方程 t 检验指标都显著，$R^2$ 达到较高水平，DW 值显示不存在自相关。因此可以确立该区域的财政收入和 GDP 的回归方程。并可以得到 GDP 每增加一元，财政收入大约增加 0.14 元。若已知 GDP，则代入方程可预测财政收入。

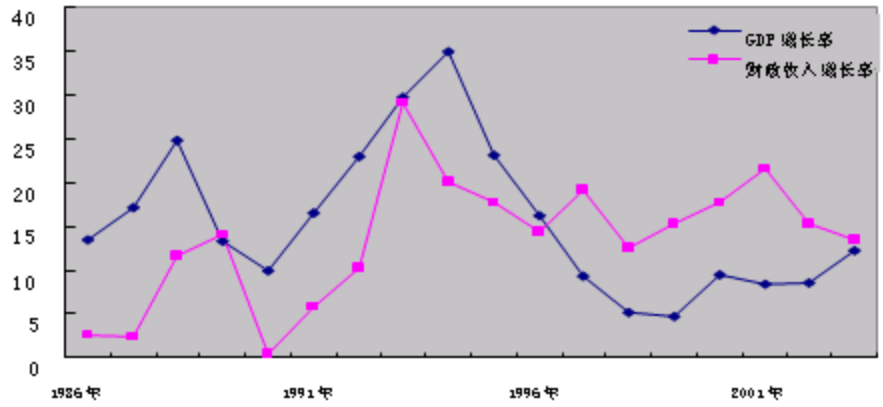
## 六 协整分析

协整分析是计量经济学里常用的分析方法，通常用于处理非平稳时间序列的关系。考察序列之间是否存在一种长期的均衡关系，如收入增长和经济增长是否在长期内保持协调一致的发展趋势。协整分析一般是要在较长时间内对时间序列进行分析，因为经济指标的变化往往在长期内才能看出一定的趋势和一致性，短期则更多地表现为失衡，不协调的形态。

假设两个非平稳时间序列 $X_t$，$Y_t$，存在一个非零向量 A=($A_1$，$A_2$)，使得 $A_1*X_t+ A_2* Y_t$ 是一个平稳序列，我们就称时间序 $X_t$，$Y_t$ 之间存在协整关系。A 就称为协整向量。存在协整关系的两个时间序列具有某种长期均衡的关系，因此即使它们在短期内由于某种原因而偏离了均衡状态，但这种偏离是暂时性的，随着时间的推移，这种偏离的趋势将会消失，序列间的关系又会回到均衡状态。此外，配合协整分析的还有误差修正模型（ECM），用来反映具有协整关系的两个时间序列在短期波动中偏离他们长期均衡关系的程度。

协整分析在财政收支分析中用于分析财政收入增长和经济增长在长期是否存在均衡关系，并以此建立误差修正模型，研究经济增长率的短期波动对财政收入增长所造成的影响。这种方法在当前的收入计量分析中被越来越多的采用，尤其是针对收入增长速度极大快于经济增长速度的情况下，考察二者的变动对协调收入与经济的关系具有重要意义。

例：某地区 GDP 增长率与财政收入增长率的走势图如下，两条曲线有着大致相同的变化趋势，说明二者可能存在协整关系。




GDP 增长率和财政收入增长率协整关系图示

利用协整检验，我们确立了 GDP 增长率与财政收入增长率确实存在协整关系，并得到协整回归方程：

财政收入增长率=10.43+0.21×GDP 增长率

这说明该地区的财政收入增长和经济增长在长期内存在均衡一致的关系，虽然短期内有失衡现象，但总能在未来时间进行修正，使两条曲线之间的分离不会太大。从这条曲线来看，虽然财政收入增长率近几年来高于 GDP 增长率，但财政收入增长率已经开始趋于下降，而 GDP 增长率也开始上升，正逐渐趋于一致。

## 七　支出偏好分析

“偏好”（Preference）一词源自经济学术语，反映用户对某种物品或劳务的喜爱或不喜爱程度，这种喜爱或不喜爱与物品或劳务的价格及用户收入无关。

该分析借用了经济学中的这一术语，并将其延伸作“支出偏好”，用于分析区域间的支出结构特点。假设共有 m 个区域， n 个支出科目，支出偏好模型如下：

若 $S$ 为支出矩阵，$s_{ij}$ 为 i 区域 j 科目支出；$P$ 为支出偏好矩阵，$p_{ij}$ 为 i 区域 j 科目支出。则有下式成立：

$$P \times u = X$$

其中，$X$ 为支出比重矩阵，$x_{ij}$ 为 i 区域 j 科目支出比重；$w$ 为平均支出行向量，$w_j$ 为 j 科目平均支出；$u$ 为平均支出比重行向量，$u_j$ 为 j 科目平均支出比重，并有如下计算公式：

$$x_{ij} = \frac{s_{ij}}{\sum_{j=1}^{n} s_{ij}}; \quad w_j = \frac{\sum_{i=1}^{m} s_{ij}}{m}; \quad u_j = \frac{w_{ij}}{\sum_{j=1}^{n} w_{ij}}$$

支出偏好分析一方面可以避免区域间差异造成的绝对额和增长额的不可比性，另一方面，可以避免忽略那些支出总量和比重都较小的科目。若支出偏好>>1，说明在该科目上支出大大高于平均水平，属于区域特色支出，可能与该区域特定政策倾向有关。

通过建立一个区域的支出偏好模型，可以跟踪分析各区域支出偏好，了解财政政策导向；此外，还可以跟踪某些项目的偏好集中在哪些区域，从而有效指导区域支出。如：某市将下属区县分为若干功能区域，某些区县重点以发展生态环保为主，市里对其进行政策和相应拨款支持。通过支出偏好模型分析，市里可以及时了解其支出发展是否符合宏观发展意图。

## 八　支出甩尾评价模型

预算执行的水平主要体现在计划与实际执行的吻合程度上，吻合度越高，说明预算水平以及执行控制水平就越高；反之亦然。但是有些地区往往不做月度计划，即使有预算的月度计划数，数据可能比较粗糙，或者难以获取。因此，我们通过财政支出的月度数据建立支出波动评价模型，以进行财政支出管理评价。

财政支出存在一个甩尾现象，即年尾支出突然增大，这种现象背后的原因比较多，不能一概认为是年终突击花钱，但是这种现象毕竟很是普遍，而且在其它条件基本相似的情况下，能在一定程度上反映预算管理和执行水平的高低。

通过定义月度甩尾系数和季度甩尾系数两个评价指标，来衡量这种甩尾现象的程度。

首先定义，M1 为 1 月支出额，M2 为 2 月支出额…..M12 为 12 月支出额；

Q1 为 1 季度支出额，Q2 为 2 季度支出额，Q3 为 3 季度支出额，Q4 为 4 季度支出额。

指标一：　F_M 月度甩尾系数

定义 F_M = Max(M1,M2…,M12)/Min(M1,M2…,M12)

指标二：　F_Q 季度甩尾系数

定义 F_Q = Max(Q1,Q2,Q3,Q4)/Min(Q1,Q2,Q3,Q4)

在指标实际使用中，由于有些科目月度支出可能会非常小，甚至可以是 0，这样的话 F_M 就会非常大，造成其值域非常广，不利于分析，此时可能需要对指标做一定的修正。

可以修改公式为：F_M=Max(M10,M11,M12)/Ave(M1,M2, …, M12)

这种修改的原因是由于我们通常更多的关注是年尾的几个月（10，11，12）相对于月平均支出总额的波动，即年尾波动性衡量。

# 第四节　数据挖掘方法

## 一　数据挖掘定义与商业应用

### （一）数据挖掘的技术定义

从技术角度来看，数据挖掘是从大量的，不完全的，有噪声的，模糊的，随机的实际数据中，提取隐含在其中的，人们不知道的，但又是潜在有用的信息和知识的过程。

### （二）数据挖掘的商业定义

从商业角度来看，数据挖掘是一种商业信息处理技术，主要特点是对商业数据库的大量业务数据进行抽取，转化，分析和模式化处理，从中提取辅助商业决策的关键知识。

数据挖掘方法可以依据其功能被分成四组：分类预测、聚类、关联规则和时间序列预测。为实现数据挖掘的每一项功能，有许多不同的方法或算法可以使用。例如：分类模型可以通过使用决策树、神经网络分类、逻辑回归或概率回归等算法来建立。而这些可归属于统计类或非统计类方法。统计类方法是建立在传统统计理论上的，而非统计类方法则是主要基于神经网络技术上的。每一项功能都可以被开发和修改成为适应银行业务的应用。比如：分类模型可以被运用到建立目标模型、预选模型、风险模型、流失模型、欺诈预测模型和破产模型。而这些模型可以用来提高市场销售效率、增强风险管理和提高客户管理的效率和效力。但是，这类细节将不在本文中涉及。下表总结了有关数据挖掘方法、技巧和典型应用。

表 2—3 数据挖掘方法分类

| 种类 | 功能 | 算法 | 典型应用 |
|---|---|---|---|
| 分类预测 | 分类 | 决策树、神经网络分类、区别分析、逻辑回归、概率回归 | 风险分析、客户挽留分析、欺诈探测 |
| | 预测 | 线性回归、非线性回归 | 收益率分析，收入预测，信用价值预测，客户潜在价值预测 |
| 聚类 | 集群分析 | K-平均值，神经网络聚类 | 客户分割 |
| 关联规则 | 关联分析 | 统计学， 集合理论 | 交叉销售，捆绑销售 |
| | 序列关联分析 | 统计学， 集合理论 | 交叉销售 |
| | 相似时间序列分析 | 统计学， 集合理论 | 产品生命周期 |
| 预测 | 时间序列预测 | 统计时间序列模型、神经网络 | 销售预测、利率预测、损失预测 |

# 二 数据挖掘常用模型

## （一）决策树模型

### 1. 什么是决策树

决策树是一种分类形式，回答一系列的是/否问题，直到事例能够归到某个特定的类别中。由于大多数的问题具有复杂的多面性，决策不是一次性做出，而是分层级分类别的递归过程，因此形成了决策树分叉形式。每个决策或回答都可能引出两个或多个事件，导致不同的结果，把这种决策分支画成图形很像一棵树的枝干，故称决策树。

### 2. 决策树的应用

决策树用来解决分类预测的问题。如解释为什么某些人被排除在某些范围之外，为什么拒绝某些人贷款等。根据分类结果在某些方面表现出的共性特征，它还可以做出预测，例如某人是否喜欢一场电影，是否会买某样东西，是否会参加某俱乐部等。

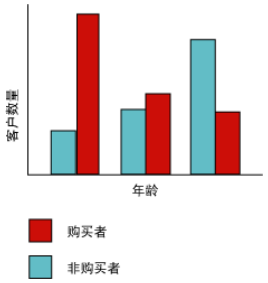
决策树算法识别大部分的特征，然后创建一组规则，规定了新的事物属于某一模式的概率。并且用图形的方式形象而又清晰地表达决策过程和结果，为决策者的每个决策过程提供依据。

### 3. 决策树的算法原理

决策树通过在树中创建一系列拆分（也称为节点）来生成数据挖掘模型。每当发现输入列与可预测列密切相关时，算法便会向该模型中添加一个节点。该算法确定拆分的方式不同，主要取决于它预测的是连续列还是离散列。

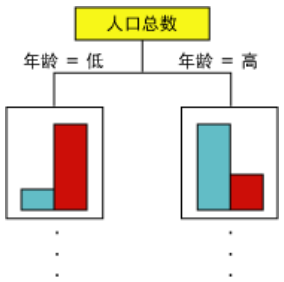
### 4. 预测离散列

通过柱状图可以演示决策树算法为可预测的离散列生成树的方式。下面的关系图显示了一个根据输入列“年龄”绘出可预测列“客户数量”的柱状图。该柱状图显示了客户的年龄可帮助判断该客户是否将会购买自行车。




决策树模型示例图 1

该关系图中显示的关联将会使决策树算法在模型中创建一个新节点。

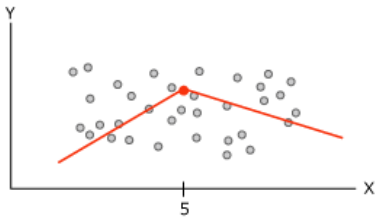



决策树模型示例图 2

随着算法不断向模型中添加新节点，便形成了树结构。该树的顶端节点描述了客户总体可预测列的分解。随着模型的不断增大，该算法将考虑所有列。

### 5. 预测连续列

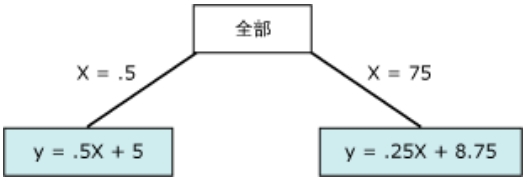
当 Microsoft 决策树算法根据可预测的连续列生成树时，每个节点都包含一个回归公式。拆分出现在回归公式的每个非线性点处。例如，请看下面的关系图。




决策树模型示例图 3

该关系图包含可通过使用一条或两条连线建模的数据。不过，一条连线将使得模型表示数据的效果较差。相反，如果使用两条连线，则模型可以更精确地逼近数据。两条连线的相交点是非线性点，并且是决策树模型中的节点将拆分的点。例如，与上图中的非线性点相对

应的节点可以由以下关系图表示。两个等式表示两条连线的回归等式。




决策树模型示例图 4

## （二）群集算法

### 1. 什么是群集算法

俗语说，物以类聚、人以群分。当有一个分类指标时，分类比较容易。但是当有多个指标，要进行分类就不是很容易了。比如，要想把中国的县分成若干类，可以按照自然条件来分：考虑降水、土地、日照、湿度等各方面；也可以考虑收入、教育水准、医疗条件、基础设施等指标；对于多指标分类，由于不同的指标项对重要程度或依赖关系是相互不同的，所以也不能用平均的方法，因为这样会忽视相对重要程度的问题。所以需要进行多元分类，即群集分析。

### 2. 群集算法的应用

在市场研究领域，群集算法主要应用方面是帮助我们寻找目标消费群体，进行市场分隔。运用这项研究技术，我们可以划分出产品的细分市场，并且可以描述出各细分市场的人群特征，以便于客户可以有针对性的对目标消费群体施加影响，合理地开展工作。

拥有顾客信息资料数据的企业，可以通过群集算法对顾客群体进行划分，预防欺诈，侦测呆账，或者进行个人化产品设计和顾客价值管理等。

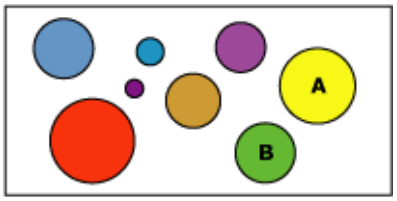
此外，在日常业务中进行文件分类、以关键字查询文章，客服申诉处理，病人病例分析等都离不开群集算法的应用。

### 3. 群集算法的原理

和决策树不同，群集算法没有事先已定的任何规则，是在对结果没有任何预先的概念下使用的，算法基于对一串连续的值进行划分，严格地根据数据以及该算法所标识的分类中存在的关系定型。

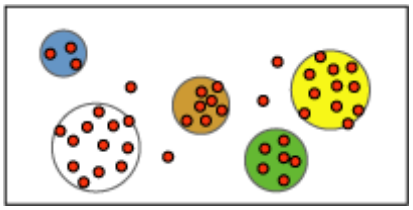
在最初定义分类后，算法将通过计算确定分类表示点分组情况的适合程度，然后尝试重新定义这些分组以创建可以更好地表示数据的分类。该算法将循环执行此过程，直到它不能再通过重新定义分类来改进结果为止。

例如，在逻辑上可以得知，骑自行车上下班的人的居住地点通常离其工作地点不远。但该算法可以找出有关骑自行车上下班人员的其他并不明显的特征。在下面的关系图中，分类 A 表示有关通常开车上班人员的数据，而分类 B 表示通常骑自行车上班人员的数据。




群集算法示例图 1

算法最初定义这样的分类后，尝试重新定义，调整群集的中心，从而找到一组最能描述相似事物的特征，形成下面散点图中的较为稳定合适的群集。如骑自行车上班的人员在年龄上、职业上可能有存在更为明显的群集特征。



群集算法示例图 2

群集分析算法提供下列两种方法来计算点在分类中的适合程度：Expectation Maximization (EM) 和 K-Means。对于 EM 聚类分析，该算法使用一种统计方法来确定分类中存在数据点的概率。对于 K-Means，该算法使用距离度量值将数据点分配给其最接近的分类。

## （三）类神经网络

### 1. 什么是类神经网络

类神经网络是一种处理复杂资料的方法，它包含了一组运算单元以及将这些单元连结在一起的计算路径。人类大脑的结构是类神经网络发明的灵感来源，因此，它就像人脑一样，具备学习和分析庞大资料的能力，这是普通演算法无法达到的。

### 2. 类神经网络的应用

类神经网络已被证明对处理真实世界的问题非常管用，这类问题的特性是拥有非常复杂而不完整的资料，它最初被应用在视觉辨认及语音辨识等问题上，此外，近来还有人用类神经网络来进行文字转语音的研究。而在 PDA 上，常见的手写软件也免不了使用类神经网络。

类神经网络的商业用途目前已经引起人们越来越多的关注。一些大型金融机构已在使用类神经网络来增进某些特殊功能，如评定顾客的信用、评判抵押品的价值、目标行销以及贷款风险评估等，虽然这些系统实际运行时只比传统演算法的精确度高出几个百分点，但因为这些评估涉及了巨大的资金交易，因此它具有突出的实用价值。现在还有些机构使用类神经网络来分析信用卡的交易情况，以判断是否被盗刷。

此外，类神经网络也可用来侦测犯罪行为，美国大多数机场都用类神经网络来侦测旅客的行李箱里是否夹藏炸弹或其他爆裂物品，而芝加哥警察局的风纪处则用类神经网络来"过滤"受贿警官，看来，类神经网络也可以成为"廉政"的利器。

### 3. 类神经网络的算法原理

类神经网络的关键，就在于每个处理单元之间建立适当的连结，这就像脑神经细胞间彼此用突触做沟通的渠道一样。我们可用一般的电脑来模拟人工的神经元，每个神经元都可同时接受许多输入讯号，根据内部简单的运算，送出一个单一的讯号，作为另一神经元的输入讯号。 每个神经元都收到由前一层发出的讯号，然后，将各个讯号经过加权处理，归纳出最后的结果，再传给上一层的神经元。于是，类神经网络的关键就在于如何决定每一神经元的权值。

类神经网络都需要经过训练来调整这个比值，在训练开始前，每一比值都用随机方法决定，这种类神经网络可被解读为人脑最初的浑沌状态。然后可以有两种训练方法：第一种方法是自我组织类神经网络，它通常是处理大量资料，并且要从资料里自动归纳出特定的模式和资料彼此间的关联，这种方法常被研究者用来分析研究资料。

相对地，第二种方法是逆向传播类神经网络，它需要操作者介入训练，在输入每笔资料的过程中，操作者观察类神经网络的输出结果是否正确，如果正确，那么就加强产生这个结果的权重比值，反之，则降低那些权重的比值，这种类神经网络通常被用于认知学习的研究以及处理特定问题上。

## （四）关联规则

### 1. 什么是关联规则

关联规则分析主要用于发现不同事件之间的关联性，即一事物发生时，另一事物也经常发生。关联规则分析的重点在于快速发现那些有实用价值的关联发生的事件。

### 2. 关联规则的应用

关联规则挖掘的典型例子是购物篮分析。例如超级市场利用前端收款机收集存储了大量的售货数据，这些数据是一条条的购买事务记录，每条记录存储了事务处理时间，顾客购买的物品、物品的数量及金额等。这些数据中常常隐含形式如下的关联规则：在购买铁锤的顾客当中，有 70 %的人同时购买了铁钉。这些关联规则很有价值，商场管理人员可以根据这些关联规则更好地规划商场，如把铁锤和铁钉这样的商品摆放在一起，能够促进销售。

有些数据不像售货数据那样很容易就能看出一个事务是许多物品的集合，但稍微转换一下思考角度，仍然可以像售货数据一样处理。比如人寿保险，一份保单就是一个事务。保单上记录有投保人的年龄、性别、健康状况、工作单位、工作地址、工资水平等。这些投保人的个人信息就可以看作事务中的物品。通过分析这些数据，可以得到类似以下这样的关联规则：年龄在 40 岁以上，工作在 A 区的投保人当中，有 45 %的人曾经向保险公司索赔过。在这条规则中，“年龄在 40 岁以上”是物品甲，“工作在 A 区”是物品乙，“向保险公司索赔过”则是物品丙。可以看出来，A 区可能污染比较严重，环境比较差，导致工作在该区的人健康状况不好，索赔率也相对比较高。

### 3. 关联规则的算法原理

一个关联规则可以特征化为两个参数：支持度(support)和置信度(confidence)。其主要依据是：事件发生的概率和条件概率应该符合一定的统计意义。

可信度：一个商品的购买暗示着另一个的购买。例如，对所有购买牙刷的人来说，48% 的人同时也购买牙膏。48% 就是规则的可信度。

支持度：同时购买两件商品。例如，所有购买者的 0.135% 同时购买牙刷与牙膏。0.135% 就是这个规则的支持度。某一特定情况下的支持度通常很低，由于要根据所有的购买者及所有的事务来计算。

可信度是对关联规则的准确度的衡量，支持度是对关联规则重要性的衡量。支持度说明了这条规则在所有事务中有多大的代表性，显然支持度越大，关联规则越重要。有些关联规则可信度虽然很高，但支持度却很低，说明该关联规则实用的机会很小，因此也不重要。

## （五）时序群集

### 1. 什么是时序群集

时序群集类似于群集演算法，都是对事件进行分类分组，不同的是，时序群集演算法会寻找时序中路径相似的案例群集，而非包含相似属性的案例群集。

### 2. 时序群集的应用

网上购物的行销部门最经常用到时序群集。客户在网上通过点击随意浏览，网站公司猜想客户会依某种次序模式，将产品放入购物篮中，他们利用时序群集演算法来寻找客户将商品加入购物篮中的顺序。之后，他们利用这项资讯来简化网站的流程，以便引导客户购买其他产品。

### 3. 时序群集的算法原理

时序群集演算法会透过分组或群集相同的顺序来寻找最常见的顺序，这些顺序有许多种形式，包括上面案例中的：1）描述客户在整个网站依序按下的路径资料；2）描述客户在线购物过程中将商品放入购物篮的顺序资料。

演算法建立的挖掘模式中包含了对资料中最常见时序的描述，这样就可以使用描述来预测新时序中下一个可能的步骤。当演算法将记录群集时，它也可以计算资料中与顺序没有直接关联的资料行。因为此演算法包含不相关的资料行，可以使用产生的相关模型，来识别顺序资料和顺序中未出现的资料之间的关联性，进而对分析不同时序下的群集关联。

## （六）贝氏机率分类

贝氏机率分类演算法是用于预测机率的分类演算法，它会计算输入资料行和可预测资料行之间的条件式机率（即贝氏机率），并假设资料行是独立的。这种独立性假设产生了贝氏机率分类这个名称，透过这样的假设，演算法不会考虑可能存在的相倚性。

此演算法比其他演算法更少计算，因此对于快速产生挖掘模型来探索输入资料行和可预测资料行之间的关联很有用，通过贝氏机率分类，得到初始结果，然后根据这个结果，利用其他更为精确的演算法建立数据挖掘模型。

贝氏机率分类演算法的一个典型应用是目标邮寄分类。行销部门决定邮寄广告传单来锁定目标潜在客户。为了减少成本，他们想要将广告单只寄给那些有可能回应的客户。公司会将有关人口统计资料和对邮件回应等资讯存在资料库中。他们相应使用此资料来了解人口统计资料（如年龄和职业）如何协助预测促销的回应，将潜在客户与具有类似特性而且过去曾

向公司购买产品的客户做比较。尤其想要看看那些购买产品和没有购买产品的客户之间的差异。使用贝氏机率分类，行销部门可以快速判断那些客户最有可能对广告单做出回应，这些结果还可以以视觉化的方式显示出来。

# 三　数据挖掘在财政收支分析中的应用

数据挖掘在财政收支分析中的应用领域主要有财政收入的变化分析、税源管理分析以及单位支出分析等综合的专题分析中。在实践应用中具体涉及到的问题可以是：

## （一）预测问题

如财政收入的预测。可以尝试不同的预测模型。回归模型将各产业的 GDP 值、各行业发展的增长指标引入预测模型；时间序列模型强调的是数据的历史变化特征和变化趋势，应用时间序列模型需要注意数据的使用范围，不同地区的误差情况等问题。

此外，还可以用数据挖掘开拓新的思路，如对某一税种的序列进行预测，可以考虑引进新的预测变量，如营业税是财政收入中占据较大份额的税种，考虑对营业税单独预测。预测中引入宏观经济、重点产业经济发展指标以及税收政策等外部预测变量，力求尽可能的涵盖营业税收入变化的各种影响因素，对未来的预测达到较高的准确度。

## （二）群体特征问题

简单的群体特征可以从统计学的描述方法入手，如描述某区域单位的支出特征，运用平均值、标准差、最大最小值等统计量，或者描述某单位的支出特征，如所属部门、支出类别支出规模等特征。利用数据挖掘进行的群体特征分析往往针对的是非既定的特征，需要模型挖掘的新特征。一般需要借助聚类或主成分等分析方法，如把单位支出的所有相关信息等放入模型，让模型自动归类，对每一类别的特征具体分析，往往能够发现一些意想不到的群体。或者把重点税源的各项财务指标放入模型，能够把这些指标适当归类，降维，得到一些综合指标，这样更能够清晰的把握税源，利于进行税源管理。

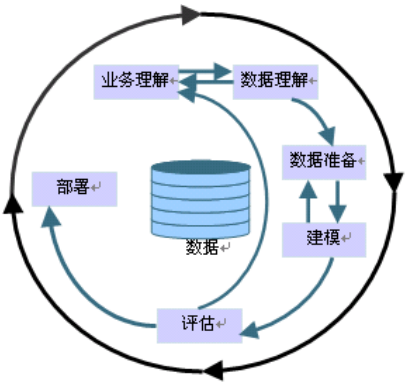
## （三）关联分析问题

财政收支行为并不是孤立的，取决于多种因素的影响。比如，财政收入除了受宏观经济因素的影响，还和产业发展、税收政策、税源属性等相关；财政支出更是受政府财力、财政政策、经济发展等多种因素影响。利用数据挖掘方法进行财政收支的关联分析就是尝试寻找地方经济发展的关键影响因素，研究财政支出和财政收入以及经济发展的关联关系，以使财政更好地实现资源配置和调控经济的管理职能。

# 四　数据挖掘的处理过程

整个数据挖掘的流程按顺序经历了由业务理解到数据理解，在经过数据准备后建模，对模型评估，选择效果较优的模型，最好把模型的结果部署到客户端，便于应用。这是一个闭

合的循环过程，在不同阶段可能会遇到各种问题，需要返还到上一阶段或重新开始。具体每一步的要做的内容如下。




数据挖掘的处理过程

**业务理解**

- 找问题——确定目标
- 对现有资源的评估
- 确定问题是否能够通过数据挖掘来解决
- 确定数据挖掘的目标
- 制定数据挖掘计划

**数据理解**

- 确定数据挖掘所需要的数据
- 对数据进行描述
- 数据的初步探索
- 检查数据的质量

**数据准备**

- 选择数据
- 清理数据
- 对数据进行重建
- 调整数据格式使之适合建模

**建立模型**

- 对各个模型进行评价
- 选择数据挖掘模型
- 建立模型

**模型评估**

- 评估数据挖掘的结果
- 对整个数据挖掘过程的前面步骤进行评估
- 确定下一步怎么办？是发布模型？还是对数据挖掘过程做进一步的调整，产生新的模型

**模型发布**

- 把数据挖掘模型的结果送到相应的管理人员手中
- 对模型进行日常的监测和维护
- 定期更新数据挖掘模型

## 五　数据挖掘实践中的问题

在进行数据挖掘的实践工作中，影响模型和实际结果的往往很少是技术上的问题，有些看似简单的步骤往往会影响到整个数据挖掘流程的进行，影响到模型的准确度和置信度，进而影响到结果的正确与否。这些问题主要表现在两个方面：

首先是数据问题。一方面是数据本身存在的质量问题，如基础信息缺失或资料填报不实等，有时候由于缺乏对数据质量的控制导致数据信息不真实，不可用。另一方面是数据源问题，由于系统错误、采集错误或管理错误导致的数据质量问题。这种问题在数据理解阶段要尽可能发现并纠正。

其次是挖掘主题理解方面的问题。由于不懂业务或没有深刻理解业务，把挖掘的目标、主题和核心都理解错了，第一步的方向就错了，即使以下的步骤再完美，结果业没有实际应用的意义。这需要分析人员在业务理解阶段，将挖掘的思想和实际业务紧密结合起来，深刻理解各项业务逻辑关系。

在以上所介绍的各种分析方法中，每种方法的侧重点各不相同。例如：

对比分析与同比分析是分析总体之间的比较分析，得出总体之间的差异；

时间序列分析是一种动态的分析，能够得出分析对象在不同发展时段的发展变化状况；

聚类分析和波士顿矩阵分析是一种分类思想，先分类，再分析；

相关性分析、回归分析以及协整分析是对多指标关联影响的分析，研究几种相互存在关联关系的业务互相作用结果；

数据挖掘的方法是一种微观、细致的研究，针对个体的行为特征，研究个体变化对总体的影响。

在实际分析中应该根据具体情况选择不同的分析方法或多种分析方法并用定性或定量的分析，使分析结论能够确定问题的基本性质或确切的量化指标。

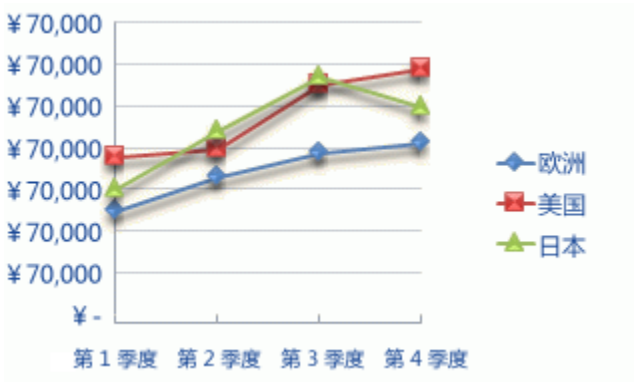
不论是采取哪种方法，被分析的对象标样应该是具有典型意义的，并且在数据量方面也要求有足够的历史资料以确保分析的一般性和规律性。

# 第五节　常用展现图形

## 一　折线图

折线图是用直线段将各数据点连接起来而组成的图形，以折线方式显示数据的变化趋势。折线图可以显示随时间（根据常用比例设置）而变化的连续数据，因此非常适用于显示在相

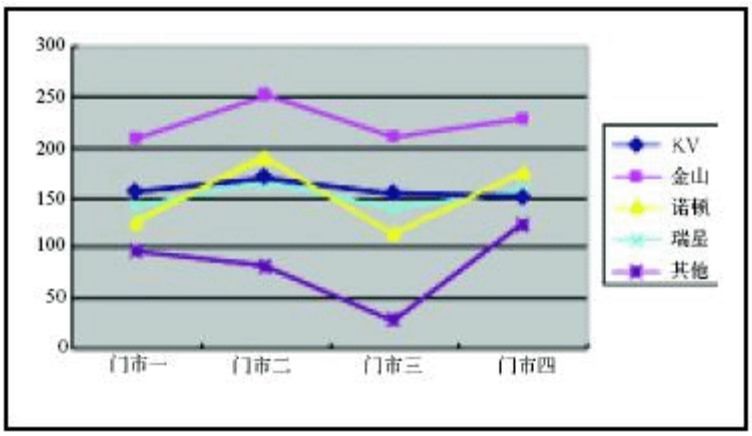
等时间间隔下数据的趋势。在折线图中，类别数据沿水平轴均匀分布，所有值数据沿垂直轴均匀分布。




另外，在折线图中，数据是递增还是递减、增减的速率、增减的规律（周期性、螺旋性等）、峰值等特征都可以清晰地反映出来。所以，折线图常用来分析数据随时间的变化趋势，也可用来分析多组数据随时间变化的相互作用和相互影响。例如可用来分析某类商品或是某几类相关的商品随时间变化的销售情况，从而进一步预测未来的销售情况。在折线图中，一般水平轴（X 轴）用来表示时间的推移，并且间隔相同；而垂直轴（Y 轴）代表不同时刻的数据的大小。

**折线图的特点**

折线图的特点是反映事物在一段时间内的趋势。如：速度——时间曲线，推力——耗油量，升力系数——马赫数，压力——温度。




**折线图的类型**

**1、折线图和带数据标记的折线图**

折线图用于显示随时间或有序类别而变化的趋势，可能显示数据点以表示单个数据值，也可能不显示这些数据点。在有很多数据点并且它们的显示顺序很重要时，折线图尤其有用。如果有很多类别或者数值是近似的，则应该使用不带数据标记的折线图。

**2、堆积折线图和带数据标记的堆积折线图**

堆积折线图用于显示每一数值所占大小随时间或有序类别而变化的趋势，可能显示数据点以表示单个数据值，也可能不显示这些数据点。如果有很多类别或者数值是近似的，则应该使用无数据点堆积折线图。

**3、百分比堆积折线图和带数据标记的百分比堆积折线图**

百分比堆积折线图用于显示每一数值所占百分比随时间或有序类别而变化的趋势，可能显示数据点以表示单个数据值，也可能不显示这些数据点。如果有很多类别或者数值是近似的，则应该使用无数据点百分比堆积折线图。

**4、三维折线图**

三维折线图将每一行或列的数据显示为三维标记。三维折线图具有可修改的水平轴、垂直轴和深度轴。

## 二　圆饼图

圆饼图也称馅饼图，用扇形的面积，也就是圆心角的度数来表示数量。圆饼图主要用来表示组数不多的品质资料或间断性数量资料的内部构成，且各部份百分比之和必须是 100%。圆饼图可以使企业根据圆中各个扇形面积的大小，判断某一部分在总体中所占比例的多少。

表2-9 1995工业企业单位数和工业总产值

| 类别 | 企业单位数（万个） |
|---|---|
| 国有经济 | 11.80 |
| 集体经济 | 147.50 |
| 村办工业 | 68.99 |
| 城乡联营工业 | 37.16 |
| 城乡个体工业 | 568.82 |
| 其它经济类型 | 6.03 |
| 总计 | 734.15 |


图2-3 国民生产总值与居民消费水平增长速度饼图

## 三 直条图

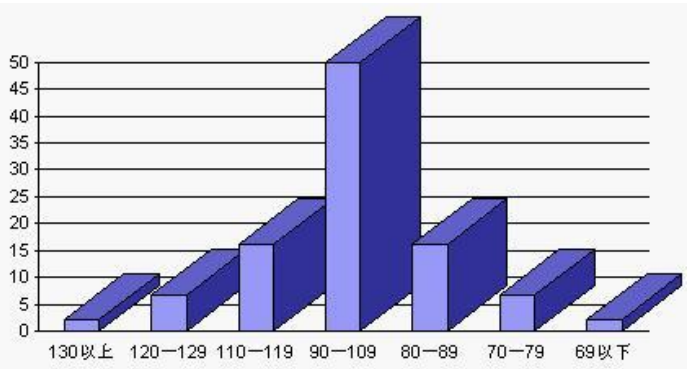
直条图又称“条形图”，是在直角坐标系中，用相同宽度长条的不同长短来表示数量资料的多少，还可在同一张图表中用不同颜色或阴影的条形表示研究对象中不同的各组，能直观地进行数量多少的对比。如果用柱形代替条形就得到柱形图，其原理与直条图相同。统计数量刻度比例要合适，并在适当位置作必要说明，如图例、单位等。

直条图一般适用于内容较为独立，缺乏连续性的数量资料，用来表示有关数量的多少，特别适合于对各数量进行对比。例如，某小组对地铁二号线运营初期，一号线和二号线的客流量进行了统计。

**直条图分为单式和复式两种：**

**1、单式直条图**

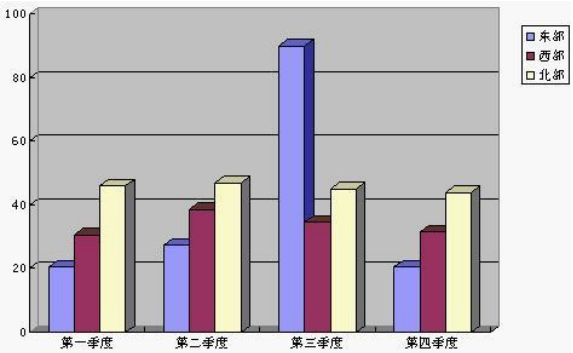
单式直条图是用同类的直方长条来比较若干统计事项之间数量关系的一种图示方法。适用于统计事项仅按一种特征进行分类的情况。




图一 智商分布理论百分比

**2、复式直条图**

复式直条图由两个或多个直条组构成，同组的直条间不留间隙，每组直条排列的次序要前后一致。




图二 不同地区四季销售百分比表

**直条图的绘制**

1、坐标轴：横轴为观察项目，纵轴为数值，纵轴坐标一定要从 0 开始。

2、直条的宽度：各直条应等宽，等间距，间距宽度和直条相等或为其一半。复式直条图在同一观察项目的各组之间无间距。

3、排列顺序：可以根据数值从大到小，从小到大，或按时间顺序排列。

**直条图和直方图的区别**

1、条形图是用条形的长度表示各类别频数的多少，其宽度（表示类别）则是固定的。

2、直方图是用面积表示各组频数的多少，矩形的高度表示每一组的频数或频率，宽度则表示各组的组距，因此其高度与宽度均有意义。

3、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，而条形图则是分开排列。

4、条形图主要用于展示分类数据，而直方图则主要用于展示数据型数据。

## 四　绩效考核五星图

绩效考核五星图模型是根据 360 度绩效评价法的基本原理，通过利用数学调整函数、几何图形和借鉴关键绩效指标考核法（Key Performance Index,KPI）和图尺度评价法（Graphic Rating Scale,GRS），全方位考核指标、全方位考核渠道对被考评者进行考评的方法。它包含绩效考核五星图、绩效分析五星图两部分的内容。如下图:

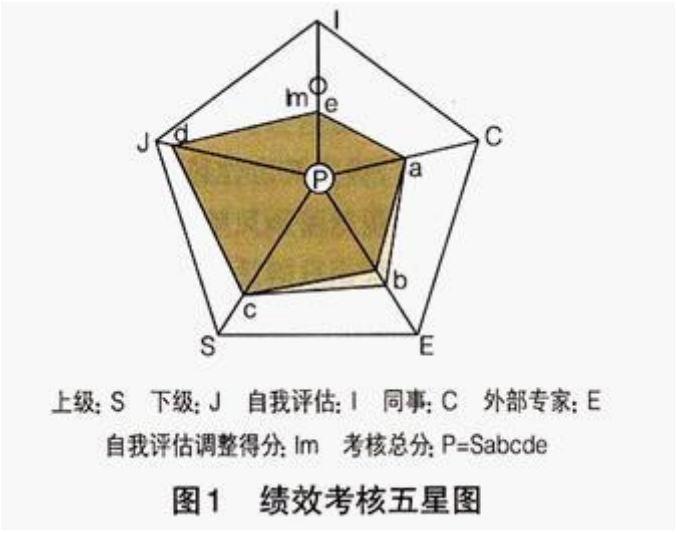
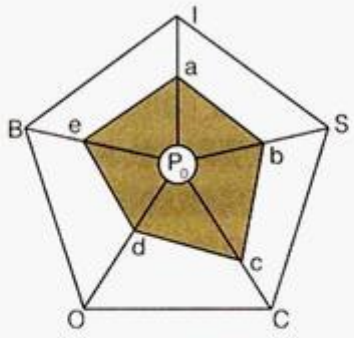



图1　绩效考核五星图

五个考评渠道即上级（Superior，S)、同事（Colleague，C)、下级（Junior,J)、外部专家（Expert,E）和被考评者本人（I）分处正五边形的五个顶点。小圆代表自我评分 I，自评调整得分为 Im。绩效五星图绘制程序如下：按 S、C、J、E、Im 分值大小排序，先将 Im 值赋予 e 点，然后以 e 点为参考，再将剩余四个分值分别赋予 a、b、c、d 点，使五星图内轴轴长依次递减。

绩效分析五星图(如下图 2)由斯蒂芬·P·罗宾斯教授提出，绩效受能力、激励和机会三因素共同影响。另外，一些绩效方面的专家将行为和环境因素也纳入绩效影响范畴。综合各种观点，绩效是能力（Skill,S)，激励（Inspire,I)，机会（Opportunity,O)，环境（Condition,C)，行为（Behavior,B）五变量的函数，即绩效＝f(S，I，O，C，B)。显然，无论忽略哪一个变量，绩效分析都不可能全面反映被考评者的真实情况。绩效分析五星图绘制与考核图相仿，即按大小顺序排列的各变量分值依次赋予五星图 abcde 各点。




能力：S　行为：B　环境：C　机会：O　激励：I

绩效潜质：$P_0$=Sabcde

图2　绩效分析五星图

**1. 绩效潜质分析**

从分析图中的面积计算，可以得出一个绩效潜质评价。所谓绩效潜质（设为 P0＝Sabcde），即被考评者的绩效增长空间。具有优秀潜质的员工不是各项分值很高的员工，而是总分中等偏上，但在能力、行为方面表现不凡的员工。由于改变系统绩效因素（环境、机会、激励）比改变个人绩效因素（行为、能力）要来得容易得多，对于这些员工，组织可以用更小的代价激发员工绩效创造。

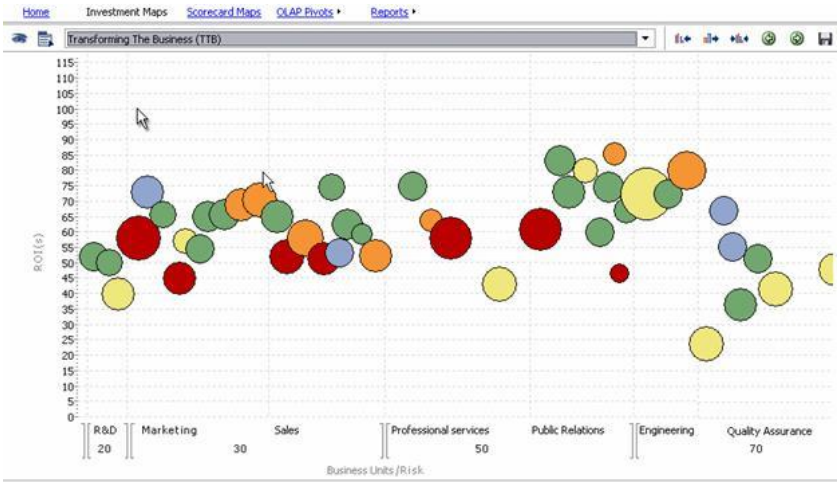
**2. 优秀绩效的典型模式**

优秀绩效模式有三种典型的情况。内力推动型。很多员工具有出色的素质能力，组织只要给他们设立恰当的激励模式，他们就能够主动整合组织资源，创造优秀绩效。外力牵引型。在一个出色的团队环境里，虽然成员的能力不是特别突出，但他迫于整体进步的压力而产生了优秀行为，从而创造出优秀绩效。在五星图模型中前者的表现对应的是行为和能力的分值很高，但是其他三个方面可能略低，而后者却恰恰相反。全力维持型。员工在五个方面都有很高的分数，就其岗位来说，这时已经达到了员工绩效创造的极点。企业应针对不同情况进行绩效沟通。

## 五　气泡图

气泡图用来为项目组合中各个因素的比较情况提供一个快速的可见的图，并可以深入观察项目组合管理整体实施的情况，为做出权衡和重新平衡项目组合的决定提供有利的支持。气泡图拥有将五个因数结合起来组成一个抽象的画面。

如图所示，它将我们选择的多个项目或者项目组合包含的所有项目组合元素显示在一张 4 维的图中进行比较，每个气泡代表一个项目组合元素。




**气泡图具有下列图表子类型：**气泡图和三维气泡图。

气泡图与 XY 散点图类似，但是它们对成组的三个数值而非两个数值进行比较。第三个数值确定气泡数据点的大小。您可以选择气泡图或者三维气泡图子类型。

气泡图的四个维度

气泡图的维度分别是：气泡的大小、气泡的颜色、纵坐标、以及两个横坐标。每个维度代表的内容可以根据需要，设置成项目组合管理者关心的各种业务决策准则。

（1） 气泡的大小：可以代表项目组合元素的成本大小、利润大小、平衡记分卡得分的高低等；

（2） 气泡的颜色：代表项目组合元素的健康状况；而健康代表的具体含义也是可以定制的，它可以代表平衡记分卡的评分等级、进度或成本的偏差范围、当前出现问题的严重程度等等；

（3） 纵坐标：可以是投资回报率（ROI）、项目组合元素优先级、成本或利润的高低等等；

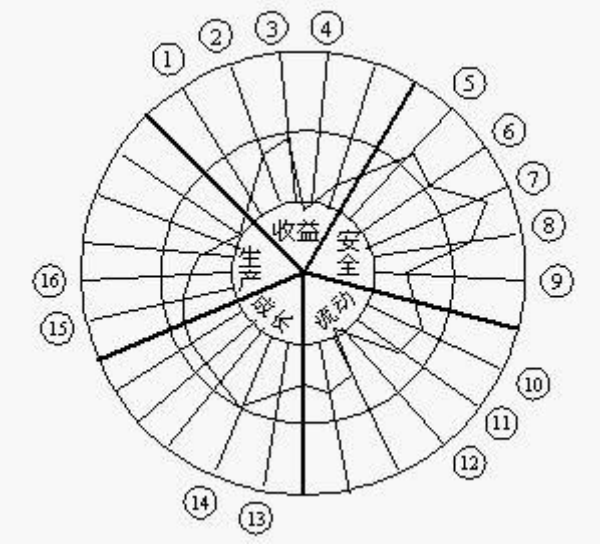
（4） 横坐标：如项目组合元素的状态、类别、当前出现问题的数量、项目组合元素属性等等。

## 六 雷达图

雷达图法是日本企业界的综合实力进行评估而采用的一种财务状况综合评价方法。按这种方法所绘制的财务比率综合图状似雷达，故得此名。

雷达图是对客户财务能力分析的重要工具，从动态和静态两个方面分析客户的财务状况。静态分析将客户的各种财务比率与其他相似客户或整个行业的财务比率作横向比较；动态分

析把客户现时的财务比率与先前的财务比率作纵向比较，就可以发现客户财务及经营情况的发展变化方向。雷达图把纵向和横向的分析比较方法结合起来，计算综合客户的收益性、成长性、安全性、流动性及生产性这五类指标。




依据上图我们可以看出，当指标值处于标准线以内时，说明该指标低于同行业水平，需要加以改进；若接近最小圆圈或处于其内，说明该指标处于极差状态，是客户经营的危险标志；若处于标准线外侧，说明该指标处于较理想状态，是客户的优势所在。当然，并不是所有指标都处于标准线外侧就是最好，还要具体指标具体分析。

## 七　面积图

排列在工作表的列或行中的数据可以绘制到面积图中。面积图强调数量随时间而变化的程度，也可用于引起人们对总值趋势的注意。例如，表示随时间而变化的利润的数据可以绘制在面积图中以强调总利润。

通过显示所绘制的值的总和，面积图还可以显示部分与整体的关系。

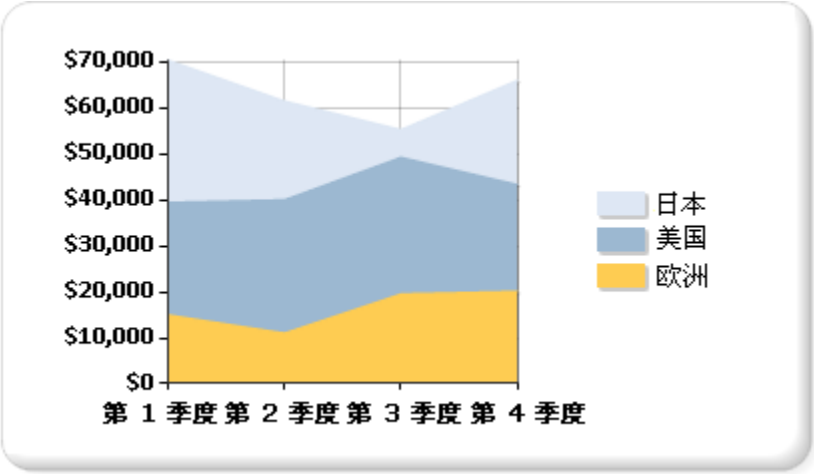
面积图具有下列图表子类型：

面积图和三维面积图 面积图显示数值随时间或类别而变化的趋势。三维面积图显示相同的内容，但以三维格式显示面积图；它不以三维格式显示数据。要以使用可修改的三个轴（水平轴、垂直轴和深度轴）的三维格式显示数据，应该使用三维面积图子类型。通常情况下，应考虑使用折线图而不是非堆积面积图。

堆积面积图和三维堆积面积图 堆积面积图显示各个数值所占大小随时间或类别而变化的趋势。三维堆积面积图显示相同的内容，但以三维格式显示面积图；它不以三维格式显示数据。要以使用可修改的三个轴（水平轴、垂直轴和深度轴）的三维格式显示数据，应该使用三维面积图子类型。

百分比堆积面积图和三维百分比堆积面积图 百分比堆积面积图显示各个数值所占百分比随时间或类别变化的趋势。三维百分比堆积面积图显示相同的内容，但以三维格式显示面积图；它不以三维格式显示数据。要以使用可修改的三个轴（水平轴、垂直轴和深度轴）的三维格式显示数据，应该使用三维面积图子类型。

三维面积图 三维面积图通过使用可修改的三个轴（水平轴、垂直轴和深度轴）显示各个数值随时间或类别变化的趋势。




## 八　散点图

散点图将序列显示为一组点。值由点在图表中的位置表示。类别由图表中的不同标记表示。散点图通常用于比较跨类别的聚合数据。

在回归分析中，数据点在直角坐系平面上的分布图。

散点图表示因变量随自变量而变化的大致趋势，据此可以选择合适的函数对数据点进行拟合。

散点图将序列显示为一组点。值由点在图表中的位置表示。类别由图表中的不同标记表示。散点图通常用于比较跨类别的聚合数据。

散点图的数据注意事项

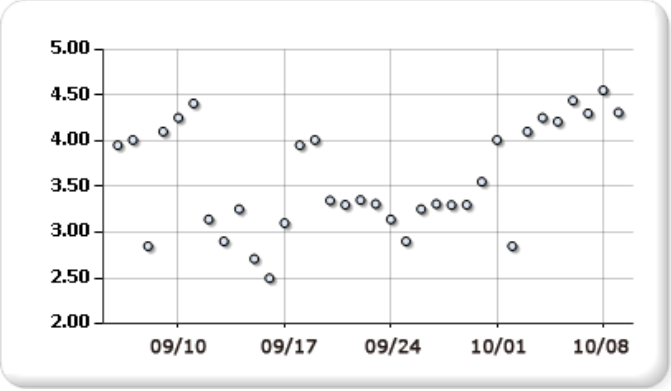
散点图通常用于显示和比较数值，例如科学数据、统计数据和工程数据。

当要在不考虑时间的情况下比较大量数据点时，请使用散点图。散点图中包含的数据越多，比较的效果就越好。

气泡图要求每个数据点具有两个值（探顶值和探底值）。

对于处理值的分布和数据点的分簇，散点图都很理想。如果数据集中包含非常多的点（例如，几千个点），那么散点图便是最佳图表类型。在点状图中显示多个序列看上去非常混乱，这种情况下，应避免使用点状图，而应考虑使用折线图。

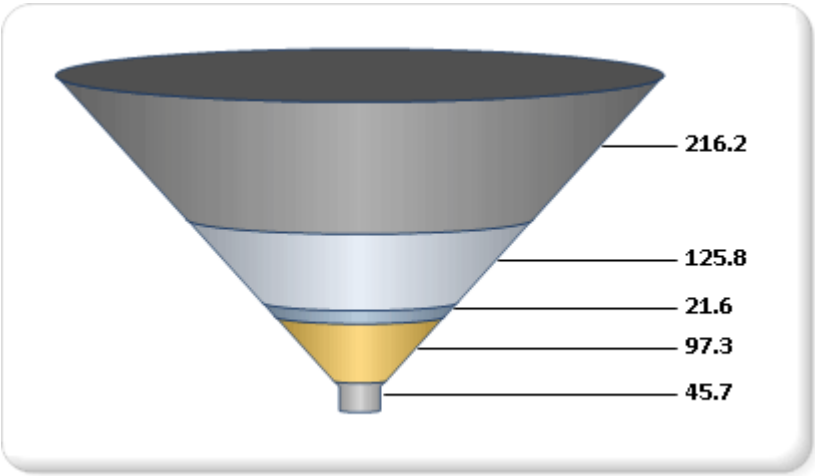
默认情况下，散点图以圆圈显示数据点。如果在散点图中有多个序列，请考虑将每个点的标记形状更改为方形、三角形、菱形或其他形状。




## 九　漏斗图

形状图将值数据显示为整体的百分比。形状图通常用于显示数据集中不同值之间的比例比较结果。类别由各个形状段来表示。形状段的大小由值来决定。形状图与饼图的用法类似，但前者是按从最大到最小的顺序来排列类别。

漏斗图按逐渐递减的比例来显示值。漏斗区的大小由序列值在所有值总计中所占的百分比来确定。例如，您可能使用漏斗图来显示网站访问者的趋势。漏斗图可能会在顶部显示较宽的区域，表明访问者对主页的点击率；而其他区域将成比例缩小。



# 十　圆环图

仅排列在工作表的列或行中的数据可以绘制到圆环图中。像饼图一样，圆环图显示各个部分与整体之间的关系，但是它可以包含多个数据系列 （数据系列：在图表中绘制的相关数据点，这些数据源自数据表的行或列。图表中的每个数据系列具有唯一的颜色或图案并且在图表的图例中表示。可以在图表中绘制一个或多个数据系列。饼图只有一个数据系列。）。

注释 圆环图不易于理解。您可能需要改用堆积柱形图或者堆积条形图。

圆环图具有下列图表子类型：

圆环图 圆环图在圆环中显示数据，其中每个圆环代表一个数据系列。例如在上图中，内环代表汽油税的收入，外环代表财产税的收入。

分离型圆环图 分离型圆环图显示每一数值相对于总数值的大小，同时强调每个单独的数值。它与分离型饼图很相似，但是可以包含多个数据系列。