**地理数据的类型**

（1）根据地理数据本身性质不同：定性数据、定量数据。

（2）根据地理数据来源及表征系统的特征不同：社会－经济数据、环境与自然资源数据；

（3）空间数据与属性数据

* **空间数据：**用于描述地理实体、地理要素、地理现象、地理事件及地理过程产生、存在和发展的地理位置、区域范围及空间联系。(点线面及其拓扑关系)
* **属性数据：**用于描述地理实体、地理要素、地理现象、地理事件、地理过程的有关属性特征。
* **数量标志数据**

① **间隔尺度数据:**以有量纲的数据形式表示测度对象在某种单位(量纲)下的绝对量。

② **比例尺度数据:**以无量纲的数据形式表示测度对象的相对量。

* **品质标志数据**

1. **有序尺度数据:**当测度标准不是连续的量，只是表示其顺序关系的数据。
2. **二元数据:**用0、1表示地理事物、地理现象或地理事件的是非判断问题。
3. **名义尺度数据:**数字表示地理实体、地理要素、地理现象或地理事件的状态类型。

**地理数据的表达方式**：（1）表格（2）地理数据矩阵

**地理数据的基本特征**：1数量化、形式化与逻辑化2不确定性 3多种时空尺度 4多维性

**形式化、逻辑化与数量化**，是所有地理数据的共同特征。

* **定量化的地理数据**是建立地理数学模型的基础，其作用为：确定模型的**参数**、给定模型运行的初值条件;检验模型的**有效性**。
* 地理计算学，对于地理数据的形式化、逻辑化提出了更高的要求，要求“整体”和“大容量”的地理数据具有**统一的数据形式和交换标准**。

**不确定性**是地理数据的**基本特征**之一。地理数据不确定性的来源：

* 地理系统**本身的复杂性**从本质上决定着地理数据的不确定性。
* 各种原因所导致的**数据误差**。

**多种时空尺度**

* **从空间尺度上来看**：全球尺度的、洲际尺度的、国家尺度的，也有流域尺度的、地区尺度的、城市尺度的、社区尺度的。
* **从时间尺度上来看**，如历史年代、天、月、季度、年等。

**多维性**：地理数据的这种多维性，被人们描述为**地理数据立方体**

对于一个地理对象的具体意义要从**空间、属性、时间**三个方面综合描述:

* **空间方面，**描述该地理对象所处的地理位置和空间范围，一般需要2～3个变量；
* **属性方面，**描述该地理对象的具体内容，至少需要1个以上；
* **时间方面，**描述该地理对象产生、发展和存在的时间范围 ，需要1个变量

**地理数据的变换**：

1定性数据转换成定量数据

1有序数据转换：不用连续的量，而是用表示其次序或等级关系的数据来作为测度标准。

2二元数据转换：即0、1转换: 1－是、有；0－否、无

2定量数据本身转换－数值标准化方法

对数变换、开（立）平方变换、取倒数变换、概率变换、模数变换、指数变换、滑动平均法、差分算子法等

**地理数据采集与处理：**

地理数据的采集：

渠道来源：野外调查、室内化验分析、定位或半定位观测、从地形图、航片、卫片上提取地理信息、从有关部门收集观测或统计资料

采集地理数据过程中需要注意的问题…

**地理数据处理**，是所有地理问题研究的**核心环节**。

* 从理论上讲，在地理学中，数学方法的运用主要有两个目的：

（1）运用数学语言对地理问题的描述，建立地理数学模型，从**更高、更深层次**上**揭示地理问题的机理**；

（2）运用有关数学方法，通过**定量化的计算和分析**，对地理数据进行处理，从而**揭示有关地理现象的内在规律**。因此，从一定意义上来说，地理数据处理也是计量地理学的任务之一。

* 地理信息系统的**核心功能**是地理数据处理，它实现了空间数据与属性数据的完美结合。数学方法确实是其强有力的支撑。

**地理计算学**（Geocomputation）的实质是借助于现代化的计算理论、计算方法和计算技术，通过对“整体”和“大容量”的地理数据进行处理，揭示复杂地理系统的运行机制，探索和寻求新的地理系统理论。

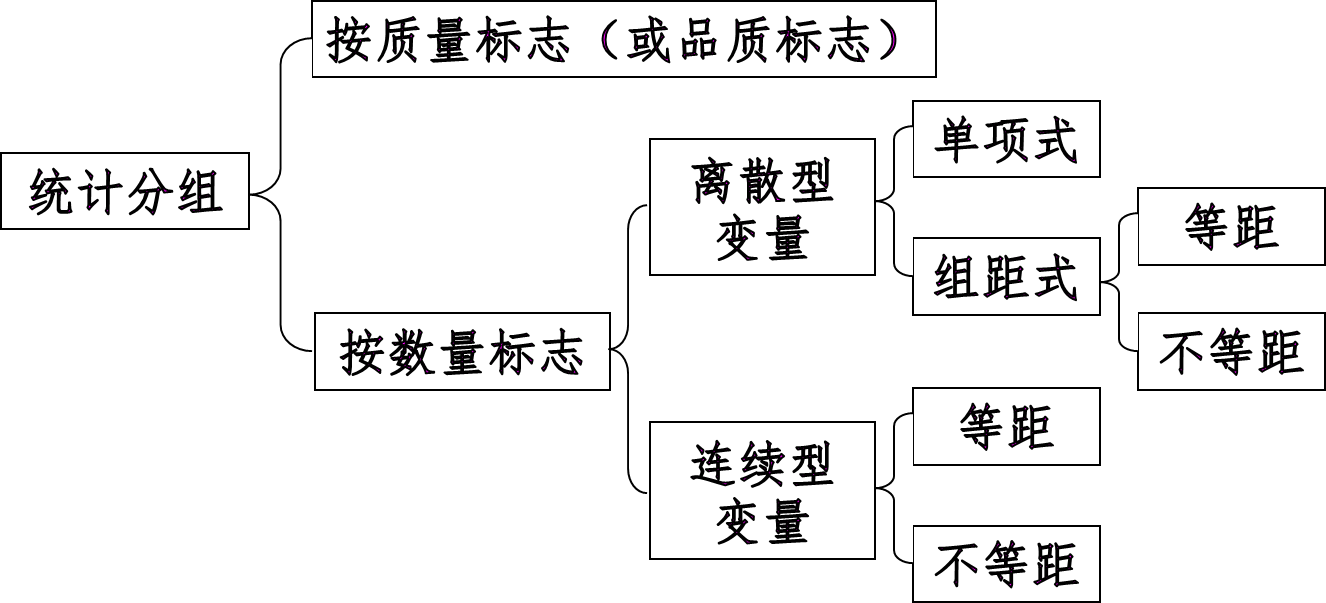
地理数据整理

检查资料； 统计分组。

分组：通过对地理数据分组来研究各组数据出理的频数(次数)和频率，以便概括数据总体结构及其分布特征。

**统计指标:总体，**可量性，无论是数量指标还是质量指标，都能用数值表示

**统计标志：**数量标志具有可量性，品质标志不具有可量性。



**统计分组——统计标志分类：**

1、**按变异情况**，可分为：

不变标志：国籍（人口普查）

变异标志：性别、年龄、民族、职业等

2、**按性质，**可分为：

质量标志：表明总体单位属性方面的特征，只能用文字来表现

数量标志：表明总体单位数量方面的特征，其标志表现可以用数值表示，即标志值

**-离散型变量：**是指其数值只能用自然数或整数单位计算。

**-连续型变量：**指在一定区间内可以任意取值的变量。

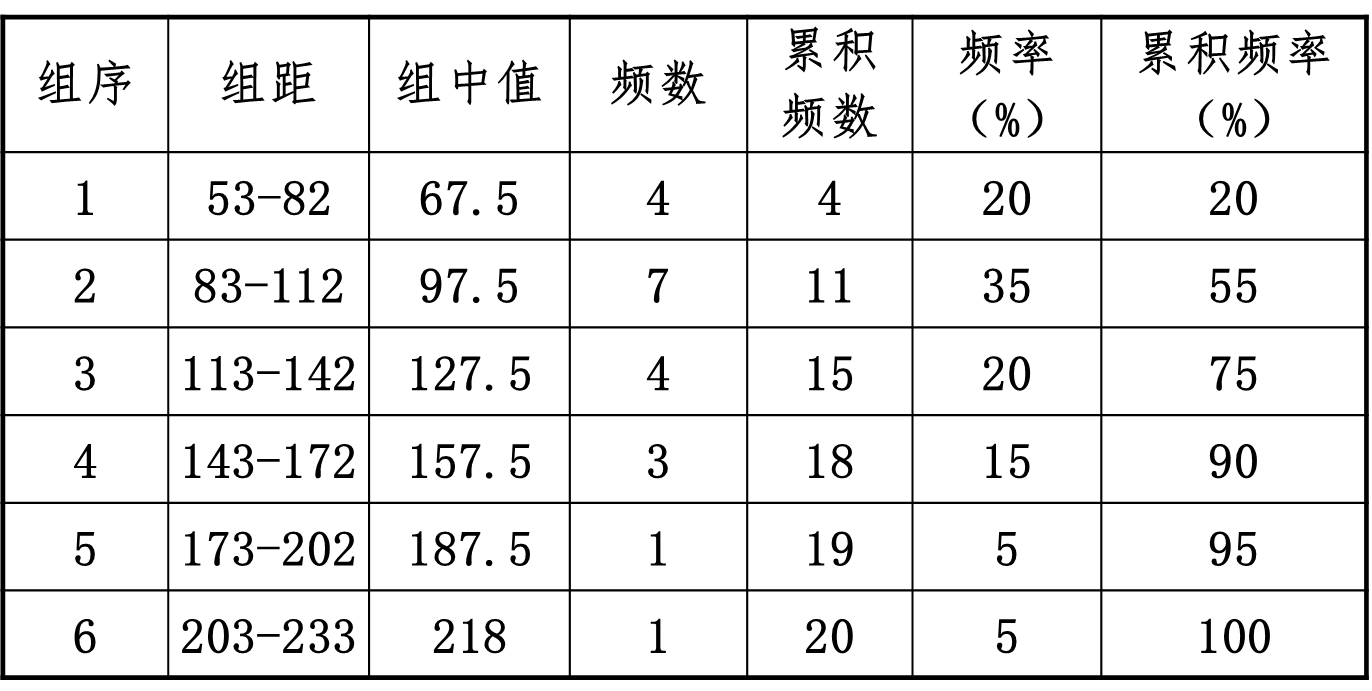
**按组距分组**

[1] 求变数的全距（R=max{X}-min{X}）

[2] 确定组数（n=**1+3.32lgN**,经验参考Sturges公式）

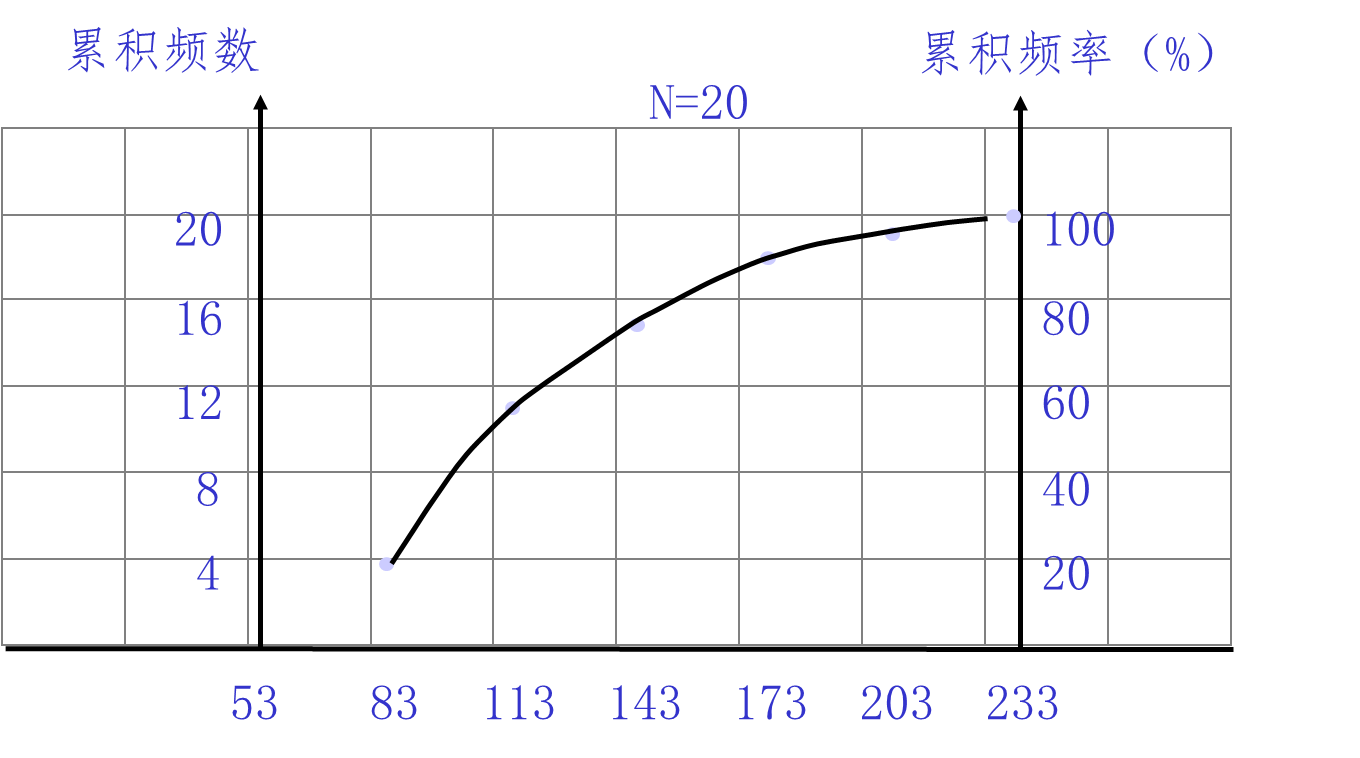
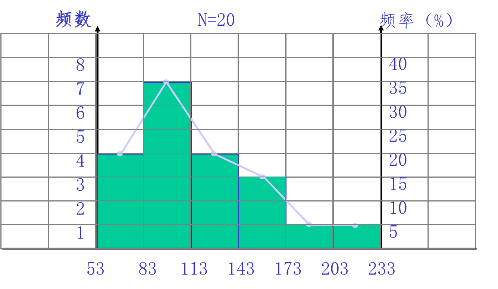
[3] 计算组距h=R/n

[4] 确定组限，组限修正（边界点），其中**第一组的下限=最小值-1/2组距**

[5] 计算组中值**m =（下限+上限）/ 2**

[6] 频数分布图表的绘制

频数分布表、频数分布图（直方图、多边形图、累积频数）



（1） “不重不漏” （2） “组距合适”

**（一）集中性的代表值**

**1.****平均数：**

**算术平均数**



**几何平均数**



**环比公式（平均发展速度）**

**平均增长速度 = 平均发展速度 — 1**

与**历史同时期**比较，例如2005年7月份与2004年7月份相比称其为**同比；**与**上一统计段**比较，例如2005年7月份与2005年6月份相比较称其为**环比。**

**2.中位数**

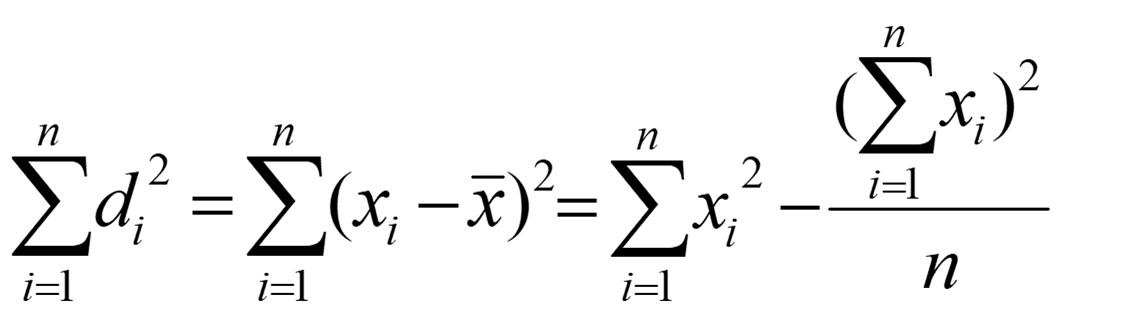


**3.众数**

单峰正偏态分配上M0<Me<avg(X)

**（二）离散性的代表值**

**a.绝对离散度**

****1.**离差**

2.**离差平方和**



3.**方差:总体方差 样本方差**



4.**标准差**

**b.相对离散度**

* 两系列数据单位不同、均值相差较大、标准差相同时

**变差系数Cv：离差系数、变异系数**(ki为模比系数)

**偏度系数g1：**测度地理数据分布的不对称性情况

**峰度系数g2：**它测度了地理数据在均值附近的集中程度

地理数据库：

组成：计算机硬件设备、地理数据库软件（DBMS）、地理数据文件、地理数据库应用程序

建立步骤：（1）用户需求调查（2）系统分析（3）系统设计（4）系统调试和运行

空间分布的测度：(1)点状分布类型(2)线状分布类型：分支、回路、区划

(3)面状分布类型：离散区域分布类型、连续区域分布类型

**点状分布的测度**

**三种模式：凝聚、均等（离散）**

**随机：**均匀平面的泊松随机过程（每个位置上出现点的概率相同；点出现的位置相互独立；）

**凝聚：**出现一些点分布的高密度区域，点密度的空间变化大；

**均等（离散）：**比完全随机分布更少聚集性（散布、均一、等距；排斥、竞争；低离差）

**邻近平均距离：最邻近距离越小，说明点在空间分布越密集;**

**1.顺序法**(点型分布为**凝聚型**时适用)

1.各点依次作为基准点：

计算该点dih（到其它全部点距离）和dib（到区域边界最短距离）;

找出满足dih≤ dib的距离；若有p个，按顺序排列：di1,di2,…dip,p=0,1,2,..,n-1

2.n个点得到n个序列，取每个序列最小值求均值，第k级最邻近距离则取每个序列第k小

**2.区域法**(点型分布为**随机型、均等型**时适用)

1.各点依次作为基准点，以之为中心将研究区分割成k个等大的齿轮状区域（一般6个）

量度dih（区内中点到最邻近点的距离），dib（到区域边界的最短距离）;

找出满足dih≤ dib的距离；若有m个，按顺序排列：di1,di2,…dim,m<=k

2. n个点得到n个序列，取每个序列最小值求均值，第k级最邻近距离则取每个序列第k小



**邻近指数（R）**

为理论的随机分布型的最邻近平均距离（泊松分布时理想平均距离）。

D=n/A，为点的密度，其中A为区域面积，n为区域内点的个数。

**R=1，随机型分布； R<1，趋向于凝集型分布； R>1，趋向于离散型的均匀分布。**

1.可以把要讨论的点的空间分布图式放在一个从凝集的、通过随机的一直到均匀分布的**连续的定量范围**之内，此尺度范围为：0-2.149。一般在0.33-1.67之间。

2.对于一个固定地域，点的空间分布随时间而变化，可通过R尺度分析去判断：

1其空间分布**比原先的是更凝集还是更趋于分散**；2定量的表达出其**凝集或分散的程度**。

**对中心位置的测度**

**中项中心：**作一对经纬线使得划分的四部分点数一致，交点即中项中心

中项中心总是偏向**分布点密度较大的一侧**，选择这样的中心，可以使中心与多数分布点之间取得较好的联系。寻找中项中心的过程比较简便，应用也较广。

**平均中心（分布重心）：**作x，y轴； 确定每一点的坐标； 计算坐标均值。

当某一空间现象的空间均值**显著区别于区域几何中心**，就说明这一空间现象的不均衡分布，或称“**重心偏离**”。偏离方向指示了空间现象的“高密度”部位，偏离的距离则指示了**均衡程度**。

**离散程度的测度**

1. **对于平均中心（中项中心）的离散程度**

1计算标准距离：离平方的均值开根号

2划分小区，计算小区内标准距离dω，再计算全区域中小区间标准距离db,d2=两者平方和

3中项中心四个区域面积分别与总面积之比

4按点与选择的中心之间的距离进行分组，画出频率累积曲线，读出占50%的累积频率半径。

**②对于任意指定中心的离散程度**

**③各点之间离散程度的测定：邻近指数R**

**线状分布的测度**

网络图：是仅由一些点以及点之间的连线所组成的图形。(**顶点、边**)

有向图：如果G的每条边给定了方向，即： （v，e)≠（e，v)，则称G为有向图；

无向图：如果G的每条边都没有方向，即： （v，e)=（e，v)，则G称为无向图。

赋权图：如果图G＝（V，E）中的每一条边（vi，vj）都相应地赋有一个数值wij，则称G为赋权图，其中wij称为边（vi，vj）的**权值**。除了可以给图的边赋权外，也可以给图的顶点赋权。即，对图G中的每一顶点vj，也可以赋予一个**载荷**a（vj）。

图的定义只关注点之间是否连通，而不关注点之间的连结方式。

对于许多地理问题，当它们被抽象为图论意义下的网络图时，问题的核心就变成了**网络图上的优化计算**问题。其中，最为常见的是关于**路径**和**顶点**的优选计算问题。

在路径的优选计算问题中，最常见的是最短路径问题；

而在顶点的优选计算问题中，最为常见的是中心点和中位点选址问题。

**最短路径问题**

权值：空间距离、经济距离、时间距离

**标号法（Dijkstar算法）**

不仅可以求出起点到终点的最短路径及其长度

而且可以求出起点到其他任何一个顶点的最短路径及其长度；

同时适用于求解有向图或无向图上的最短路径问题。

1.画表初始化0或正无穷， 取第一行最小值（出发点为0），该店为当前点

2.更新一步最短值比较与原来的值，取最小值，该点最短路径确定

3.以2中确定点为当前点，重复2

画表时注意： 1.某点最短路径确定时，应圈起来，下一行时不再补格子，若多行同值，圈出所有同值的格子以便回溯2.选点时还要记录该点最短距离

**服务点的最优区位**

1.求G的最短距离矩阵，2.求每一顶点最大服务距离（1中行最大值）3.求2中值最小的点

2中可加入顶点载荷

**面状分布的测度**

**1） 离散区域分布测度**

空间罗伦兹曲线：分组（区域），两列属性，按属性比值降序排列，做累积百分比图

基尼系数：洛伦兹曲线与对角线所夹面积

< 0.2 ：收入绝对平均；0.2-0.3：收入比较平均；0.3-0.4：收入相对合理；

0.4-0.5：收入差距较大；>0.5 ：收入差距悬殊

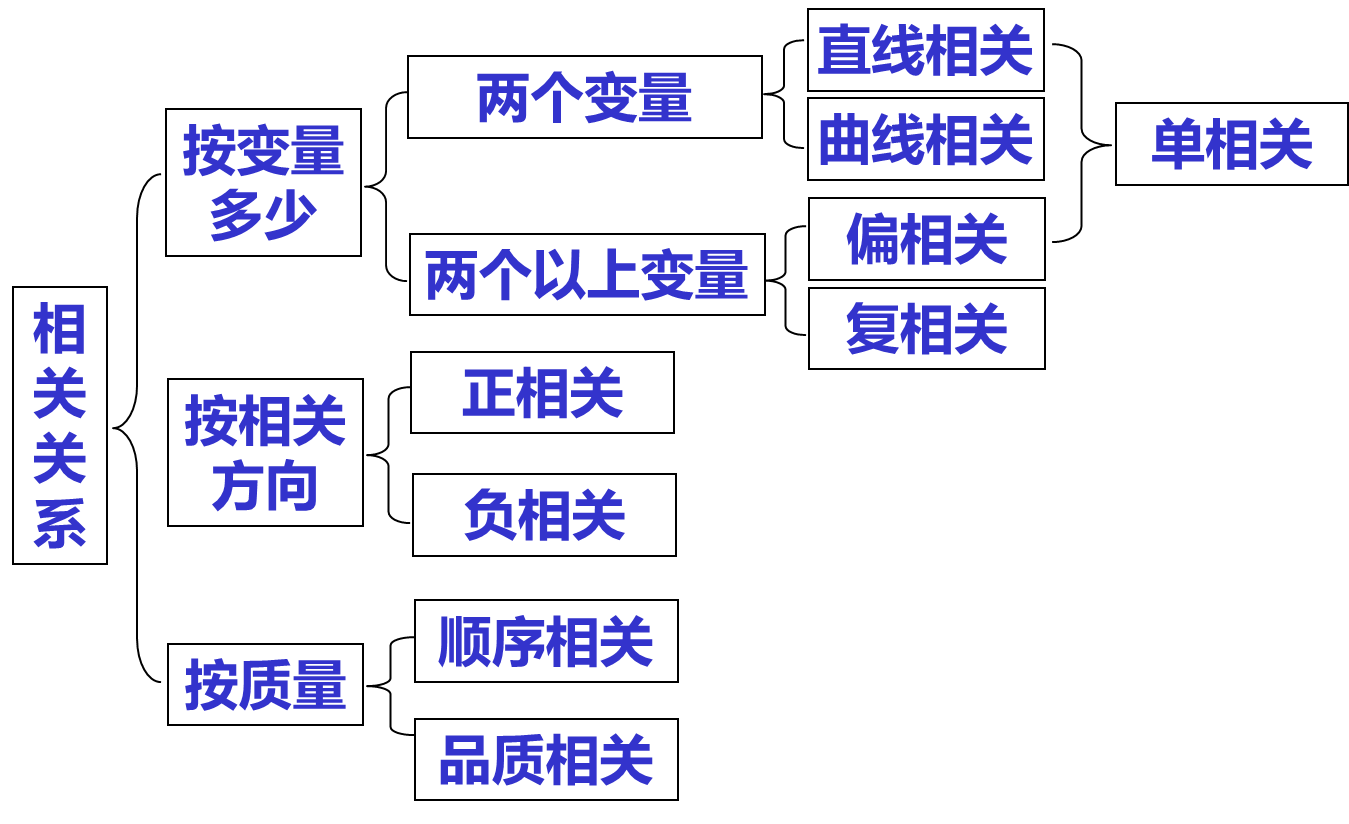
集中化指数

洛伦兹曲线中：C:各点x轴（总产值）之和，R:各点y轴（部门产值）之和,

M=n(最不均匀时的C)

**2） 连续区域分布测度**

高程曲线：相对高度=（绝对高度—高程最低值）/（高程最高值—高程最低值）

相关分析与回归分析

1.相关分析；2.回归分析；3.趋势面分析

1.相关分析：**简单线性相关程度的测度**

自变量和因变量的判定

地理相关

地理要素之间关系的类型：函数关系（完全相关）、相关关系（统计相关）、独立

1.1相关表

某一变量按其取值大小顺序排列

1.2相关图（散点图）

1.3相关系数

A）Pearson’s R積差相關： 离差积和与离差平方和积开方之比





0-0.3-0.5-0.8-0.8-1完全不-微弱-低度-显著-高度-完全

显著性检验，f=n-2

B）Spearman’s R等级相关：

1. 未有相同等級者： （D為二變數對稱之等級差）
2. 有相同等級者：

   t：表示得到相同等第的人數。

C）多要素相关：相关矩阵，偏相关与复相关

1.两两算相关（Pearson’s R单相关）得到相关矩阵

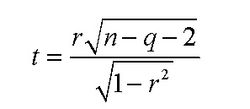
2.一阶偏相关（控制z后的XY相关性）：

3.多阶偏相关（M可迭代）

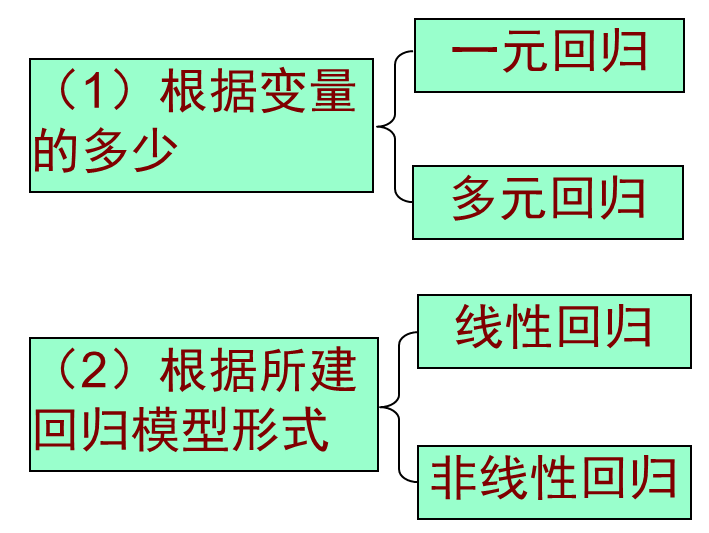
**①**偏相关系数分布的范围在-1到1之间；

**②**偏相关系数的绝对值越大，表示其偏相关程度越大；

**③**偏相关系数的绝对值必小于或最多等于由同一系列资料所求得的复相关系数。

显著性检验

r为偏相关系数，n为样本数，q为阶数，统计量服从n-q-2个自由度的t分布

2.回归分析

作用：确定数学表达式（建立回归模型）、预测

2.1.线性回归与非线性回归

通过变换坐标轴（取ln）将线性模型转化为幂指对模型

2.2一元线性回归：最小二乘法

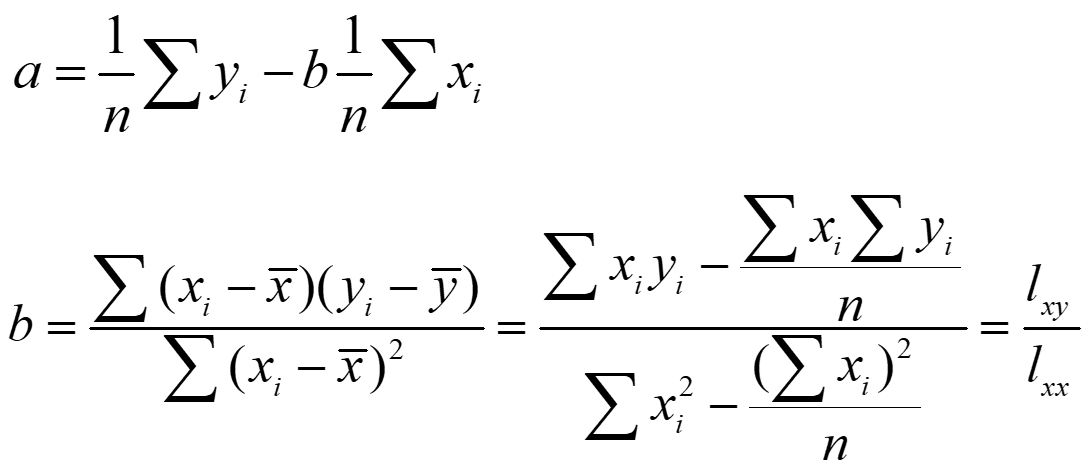
回归值

Min

2.2.1 解：



即



2.2.2 效果检验

A标准估计误差（标准标准差）



 B显著性检验 (方差分析)1.离差平方和

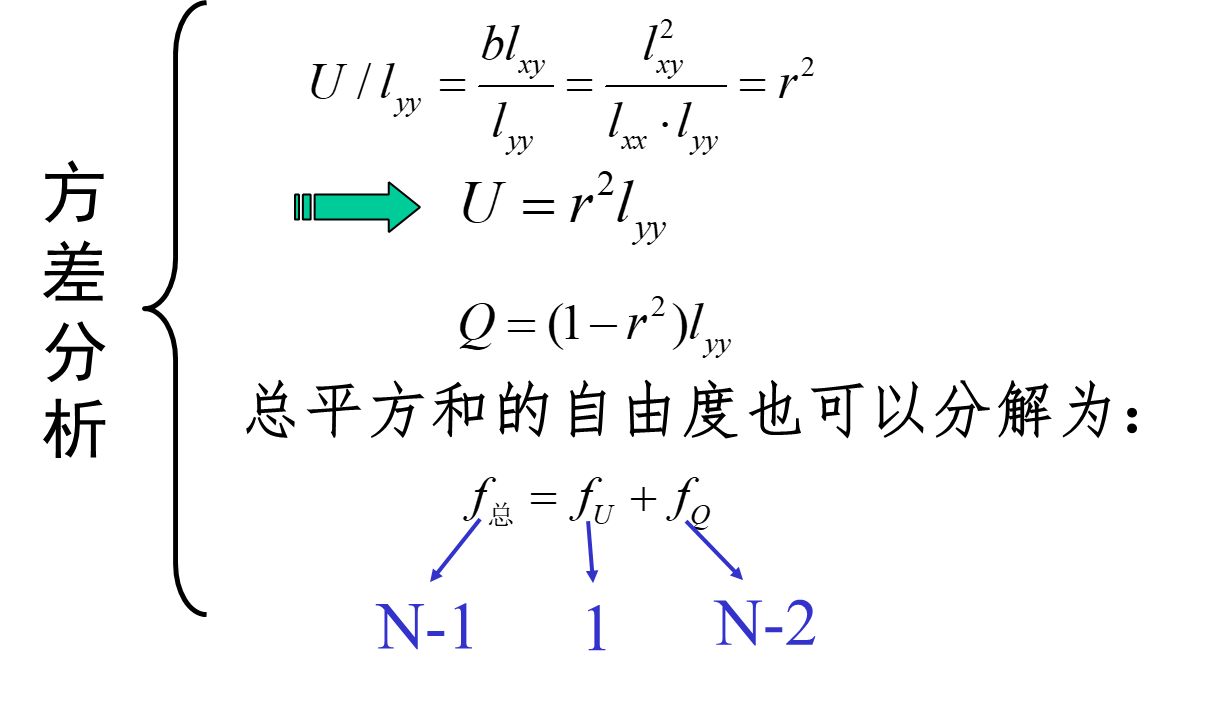
又有

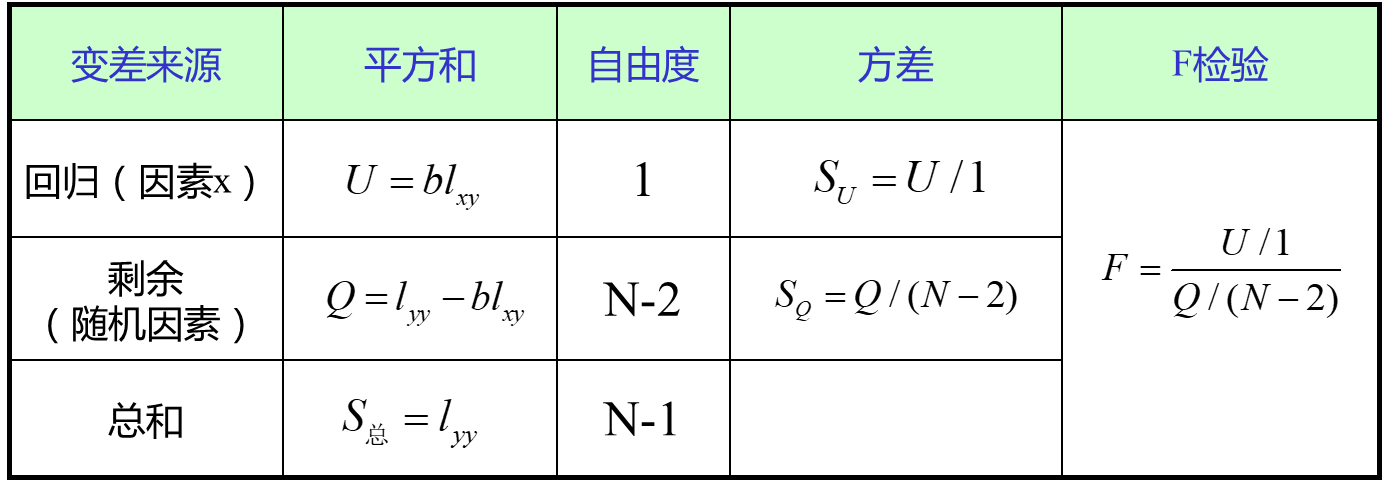
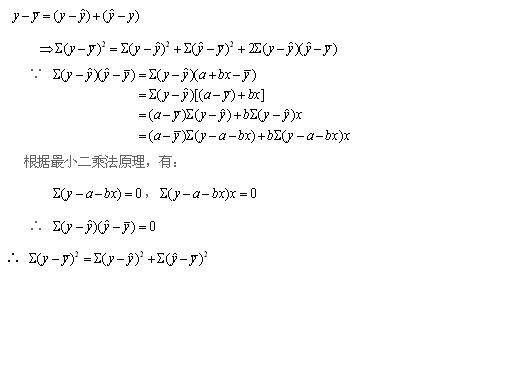
 

得

记为 S总=U+Q 或 SST=SSR+SSE

即 **离差平方和=回归平方和+剩余（残差）平方和**





2.4多元回归模型

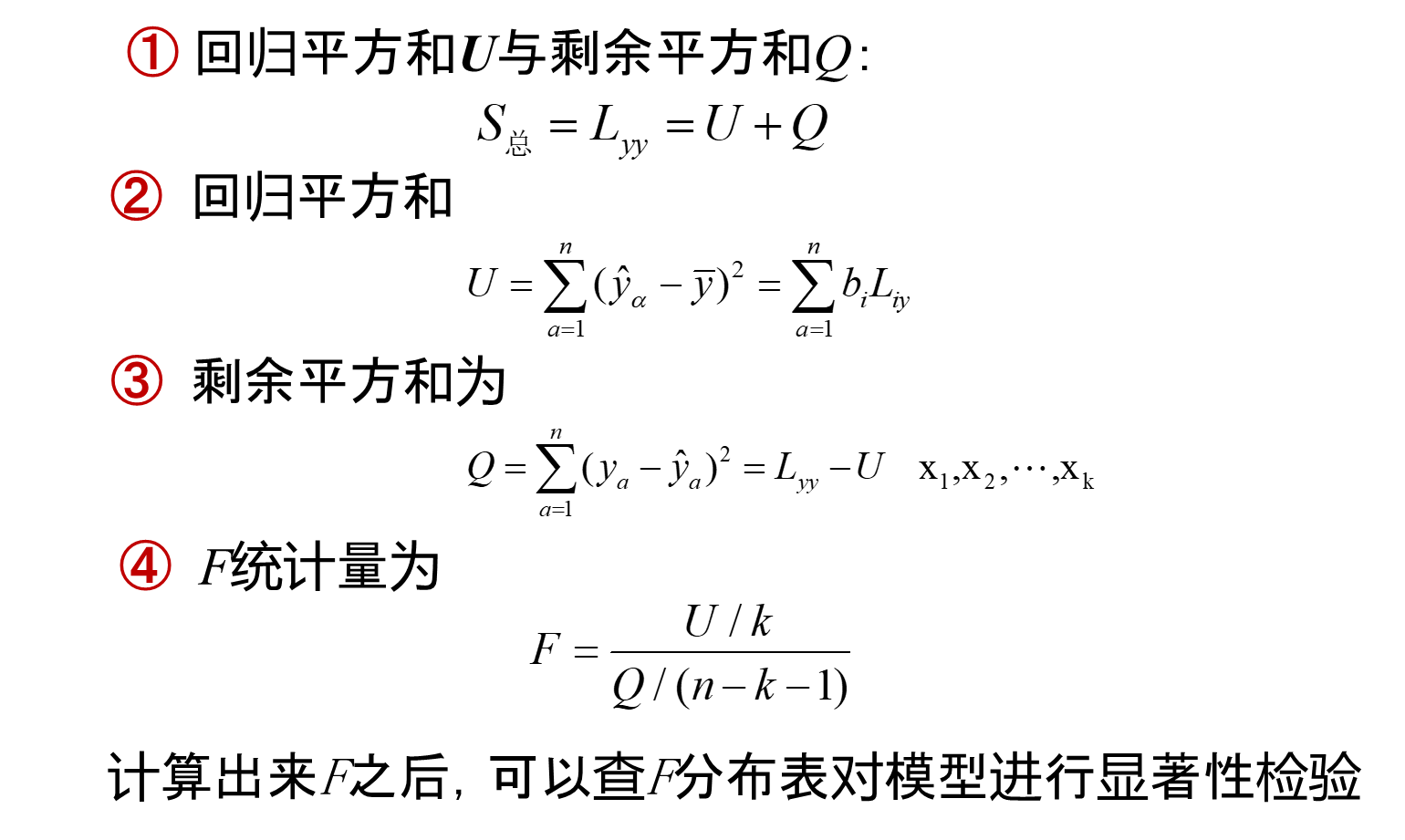
为 拟合值，称偏回归系数



由最小二乘法：



展开得正规方程



2.4一元非线性回归模

通过变换坐标轴（取ln）将线性模型转化为幂指对模型（换元为线性问题）





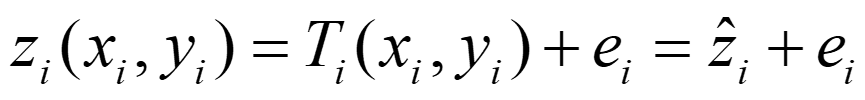
相关指数（与剩余平方和相似）

根号得到相关系数R

2.5 趋势面分析

趋势面时一种代表地理数据在大范围内得空间变化趋势的光滑曲面，是实际曲面的近似。

即：**实际曲面=趋势面+剩余曲面**

**某观测点的观测值=确定性函数值+随机性函数值=趋势值+剩余值**









将多项式回归（非线性模型）模型转化为多元线性回归模型。



F检验 ：U为回归平方和; Q为残差平方和（剩余平方和）;

p为多项式的项数（不包括b0）; n为使用的资料数目。



拟合程度指数： I≥75%时，拟合程度较为满意

**时间序列**

**时间序列分析的基本原理**

两种形式：**增长和下降** 四种序列：**线性、对数、指数、波动**

时间序列概念：是要素（变量）的数据按照**时间顺序**排列而形成的一种数列，反映**变量时间变化的发展过程**。地理过程的时间序列分析，就是通过分析地理要素（变量）随时间变化历史过程，揭示其发展**变化规律**，并对其未来状态进行**预测**。

长期趋势（T），是时间序列随时间的变化而逐渐增加或减少的**长期变化**之趋势。

季节变动（S），是时间序列在**一年中或固定时间**内，呈现出的**固定规则**的变动。

循环变动（C），是指沿着**趋势线如钟摆般地循环变动**，又称景气循环变动。

不规则变动（I），是指在时间序列中由于**随机因素**影响所引起的变动。

时间序列的分解模型

* 时间序列y可以表示为以上四个因素的函数，即：yt=f(Tt,St,Ct,It)
* 常用的模型有：加法模型：Y=T+S+C+I乘法模型：Y=T×S×C×I

分解步骤：

（1）运用**移动平均法剔除长期趋势和周期变化**，得到序列*TC*。然后再用按**月（季）平均法**求出季节指数*S*。

（2）做散点图，选择适合的**曲线模型拟合序列的长期趋势**，得到长期趋势T。

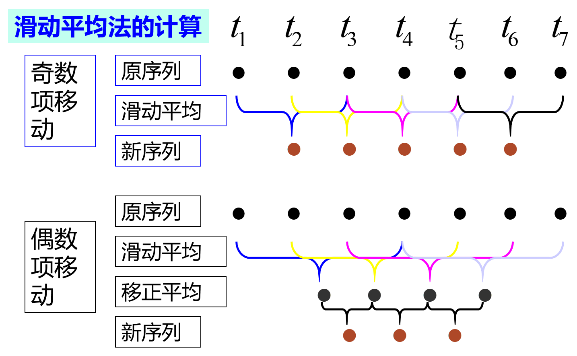
（3）计算周期因素C。用序列TC除以T即可得到周期变动因素C。

（4）将时间序列的T、S、C分解出来后，剩余的即为不规则变动，即：I=Y/TSC

**长期趋势分析：平滑法、趋势线法、自回归分析法**

时间序列分析的平滑法主要有三类 ：

**移动平均法**：移动平均=最近N期之和/N，N为移动时距

当数据的**随机因素较大时，宜选用较大的*N*，**这样有利于较大限度地平滑由随机性所带来的严重偏差；反之，当数据的**随机因素较小时，宜选用较小的*N***，这有利于跟踪数据的变化，并且预测值滞后的期数也少。

**滑动平均法：**单侧滑动时距l（2l+1）点滑动平均

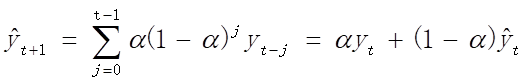
使用滑动平均法应注意的问题：

* 滑动平均法可以平滑修匀序列；
* 对于**季节性序列，要采用 4 项或 12 项滑动平均**，方可平滑掉其季节波动；
* **一般的滑动平均方法使原序列首尾各去除了若干项，因此不能用于外推预测**；
* 当序列没有明显的长期趋势、季节变动和循环变动时，可以用移动平均法进行预测，但要进行特别的计算处理。

**指数平滑法：**时间序列的态势具有**稳定性或规则性**，所以时间序列可被合理地顺势推延；他认为**最近的过去态势，在某种程度上会持续的未来**，所以将较大的权数放在最近的资料。

**任一期的指数平滑值都是本期实际观察值与前一期指数平滑值的加权平均。**

一次指数平滑法：利用前一期的预测值，订正得到预测通式（平滑系数α平稳取小波动取大）

变化趋势较稳定，以直接用第一个数据作为初始值；变动趋势有起伏**波动**时，则应以n个数据的**平均值**为初始值，以减少初始值对平滑值的影响。

高次指数平滑：平滑后平滑

**趋势线拟合法**：用某种趋势线（直线或曲线）来对原序列的**长期趋势进行拟合**。其主要作用是进行外推预测。

直线趋势方程、曲线趋势方程（指数函数、二次函数）

**1判断趋势类型2计算待定参数3利用方程预测**

**1判断趋势类型**

散点图法

差分法

当数据的一阶差分趋近于一常数时，可以配合直线方程

当数据的二阶差分趋近于一常数时，可以配合二次曲线方程。

当数据的环比趋近于一常数时，可配合指数曲线方程。yi/yi+1=const

**2计算待定参数**

线性回归（最小二乘）

**3利用方程预测**

**自回归模型：**时间序列的自相关，是指序列前后期数值之间的相关关系，对这种相关关系程度的测定便是自相关系数

K阶自相关系数：y1,y2,…,yt-1,…,yn-1与 yk+1,yk+2,…,yk+t, …,yk+n的相关系数

 p阶线性自回归模型

**季节变动分析：**是对包含**季节波动**的**时间序列**进行**预测**的方法

经济现象由于自然条件、生产条件和生活习惯等因素的影响，随着季节的转变而呈现的**周期性**变动。这种周期通常为1年。

**规律性**。每年重复出现，其表现为逐年同月(或季)有相同的变化方向和大致相同的变化幅度。

**季节变动：**一年之内因**纯季节原因**造成的序列的波动，以及**与季节无关的类似**的变动。

要点：

首先，利用**统计**方法计算出预测目标的季节指数，以测定季节变动的规律性；

然后，在已知季节的平均值的条件下，预测未来某个月（季）的预测值。

1）对原时间序列求移动平均，以消除季节变动和不规则变动，保留长期趋势；

2）将原序列*y*除以其对应的趋势方程值（或平滑值），分离出季节变动（含不规则变动），即SI=TSCI/TS,TS=平滑值（**S=各季平均**/**历年季度总平均数**）

3）将月度（或季度）的季节指标加总，以由计算误差导致的值去除理论加总值，得到一个**校正系数**，并以**该校正系数乘以季节性指标**从而获得调整后季节性指标。

4）求预测模型，若求下一年度的预测值，延长趋势线即可；若求各月（季）的预测值，需以**趋势值乘以各月份（季度）的季节性指标**。