# 统计学学习笔记

* **黎建文，2018-03-03**

# 概述

## 对概率的理解

### 概率的5种解释

#### 古典解释

，古典概率的一个前提条件是等可能性

#### 频率解释

不等可能的情况下，古典概率无法解释。为了近似计算，人们想出用频率来代替。对于一个可在同样条件下重复进行的试验，如果事件A在所有N次试验中共发生M次，则它的概率可以用其发生的频率来近似：P(A)≈M/N。这个近似的理论支持是大数定律：当N趋于无穷大时，频率几乎处处等于概率。即当N较大时，频率经常稳定地出现在概率附近，而当N越大时，越是更经常地稳定于概率。

按照频率解释，概率只是当试验可以在同等条件下无限次重复时才有意义。当研究一些不可重复事件发生的概率如当次总统竞选，频率解释无法处理。

#### 主观解释

贝叶斯。

#### 特性解释

#### 逻辑解释

# 随机变量分布

## 离散型随机变量分布

### 伯努利分布

#### 定义

一次试验，要么成功，要么不成功。成功的概率为，失败的概率为。

#### 伯努利分布

概率：

期望：

方差：

#### 应用

### 二项分布

#### 定义

在次独立重复试验中，每次试验可能的结果只有两种，发生和不发生，一次试验中发生的概率为。在次独立重复试验中，事件恰好发生了次。

#### 二项分布

概率：

期望：

方差：

#### 应用

### 多项分布

#### 定义

在次独立重复试验中，每次试验可能的结果有个，概率分别为，满足，在次独立重复试验中，第种结果恰好发生了次，满足。

#### 多项分布

概率：

#### 性质

多项分布与伯努利分布和二项分布的关系。伯努利分布是一个离散型的随机分布，做一次实验，实验的结果只有两个，而二项分布则将实验次数扩展到了多个，而多项式分布则将实验的结果也扩展到了多个。

#### 应用

### 范畴分布

### 泊松分布

#### 定义

#### 性质

#### 应用

泊松分布适合于描述单位时间内**随机事件发生的次数**的概率分布

### 负二项分布

### 几何分布

### 超几何分布

## 连续型随机变量分布

### 均匀分布

### 指数分布

指数分布的重要特性是**无记忆性**。它表示随机变量的概率只与时间间隔有关，而与时间起点无

### 单正态分布

#### 定义

设是正态分布，是均值，是方差。密度函数

#### 性质

，，则

### Z分布(标准正态分布)

#### 定义

设是正态分布，是均值，是方差。当，时，为Z分布。

#### 性质

#### 应用

总体分布为正态分布，总体方差已知，检验总体均值。

### 多元正态分布

参照https://www.zhihu.com/question/36339816

先从均值为0，方差为1的一元正态分布开始，概率密度函数为：



在均值为，方差为的情况下，需要标准化一下: ，标准化之后方差变为1，标准化的意义在于将数据点到均值的距离转化为数据点到均值0的距离等于多少个总体的标准差，这样，就消除了数据分布差异和量纲对概率计算的影响，此时的概率密度函数为：



可见，**高斯分布的概率密度计算核心在于计算数据点到中心的距离，并且除以标准差将这个绝对距离转化为相对距离，然后通过距离平方的指数衰减计算概率密度。**

回到多元正态分布，先从**各维度不相关**的多元正态分布入手，数据点通过维的列向量描述，各个维度的均值方差分别为, 来描述，高斯概率密度函数可以表示为：



前面多出的项是为了让概率之和为1，其实这个方程可以这样子去解读：



这样，各个维度之间不相关的多元正态分布概率密度其实就是各个维度的正态分布概率密度函数的乘积，其实是因为各变量之间互不相关，因此联合概率密度等于各自概率密度的乘积，我们来把她写的漂亮点：





其中，是协方差矩阵，里面的第行第列元素表示第个变量第个变量的协方差，由于假设了各个维度之间不相关，因此协方差矩阵只有在对角线的位置有值，代表不同变量的方差大小。这里面用到了（行列式的计算，对角行列式等于对角线上元素的乘积）。

**在维度之间互相关的多元高斯分布中**，需要借助一下主成分分析的一点思路。将原来维的列向量通过一个坐标转换，使得，转换之后，各个成分相互独立(请了解主成分分析的原理)。

现在数据的各个维度已经去相关，那么我们可以拿前面的各维度不相关的多元正态分布来计算啦！当前在计算之前还需要将数据标准化一下，消除一下量纲的影响：



相对距离为



而，是去相关后数据的协方差矩阵，因为是对角阵，所以逆等于对角元素取倒数方程。

由于变换后各个维度不相关，也就是变换后的协方差矩阵是

从定义看，

而，是正交矩阵，所以

，又

有



在计算的过程中，得到的最终零均值，方差为1的，相当于对原坐标做了一次变换,使成为去相关的零均值，方差为1的正态分布，因此概率密度函数在源空间做全空间积分的时候需要做换元变换，整体减小了

由于，有

所以为了保证概率密度函数全空间积分为1，需要乘上，还需要除以

这一项是在计算的积分时引入的，每个维度都会有，所以是次方

因此，整体的概率为：



### 分布/卡方分布

#### 定义

设来自总体分布是正态分布的一个样本，称统计量

服从自由度为的分布，记作。

#### 性质

概率密度函数：

期望：

方差：

分布在第一象限内，卡方值都是正值，呈正偏态（右偏态），随着参数的增大；分布趋近于正态分布；随着自由度的增大，分布向正无穷方向延伸（因为均值越来越大），分布曲线也越来越低阔（因为方差越来越大）。

#### 应用

卡方分布经常用于我们常见的卡方检验中。卡方检验一方面可以用来衡量观测分布和理论分布之间的拟合程度，另一方面也可以

测量定性数据两个分类标准间的独立性。事实上，卡方检验还有很多其它的作用。

总体的均值和总体的方差都未知，检验总体方差。

#### 在参数估计中的应用

设总体，统计量服从自由度的分布，则，

##### 在参数假设检验中的应用

##### 在分布拟合检验中的应用

### 分布/Gammar分布/伽马分布

#### 函数



#### *分布*

概率密度函数：

### *Beta分布*

#### *函数*



#### *Beta分布*

概率密度函数：

期望：

方差：

#### *应用场景*

Beta分布可以看作一个概率的概率分布，当不知道一个东西的具体概率是多少时，它可以给出了所有概率出现的可能性大小。

### *t分布*

#### 定义

在统计应用中，可以把任何一个均值为，标准差为的正态分布转变为， 的标准正态分布，即将正态变量值用来代替。由于服从正态分布，故服从标准正态分布，但由于实际资料中，往往未知，故标准化演变为，服从的分布

#### 性质

t分布以0为中心，左右对称的单峰分布；t分布是一簇曲线，其形态变化与n（确切地说与自由度ν）大小有关。自由度ν越小，t分布曲线越低平；自由度ν越大，t分布曲线越接近标准正态分布（u分布）曲线。

#### 应用

小样本情况下，总体分布为正态分布，总体方差未知，用样本方差代替总体方差，检验总体均值。

### *F分布*

#### 定义

X和Y为两个独立的随机变量，X服从自由度为n的卡方分布，Y服从自由度为m的卡方分布，这两个独立的卡方分布除以各自的自由度以后的比率服从F分布，即

#### 性质

F分布是一种非对称分布；它有两个自由度，即n-1和m-1，相应的分布记为F（ n–1，m-1）， n-1通常称为分子自由度， m-1通常称为分母自由度；F分布是一个以自由度（n-1）和（m-1）为参数的分布族，不同的自由度决定了F 分布的形状。

#### 应用

两个正态分布样本的均值方差都未知情况下求两个总体的方差比值。

### *拉普拉斯分布*

### *帕雷托分布*

### *柯西分布*

### *瑞利分布*

### [*狄利克雷*](http://www.baidu.com/link?url=XnIOlr8PK5PF70TMifJIKx3uYpTFVT4N2tLzAIE-Xz8aWUPjJIqKkK_4xmW-CHC666rjCwqGMoLIqH0_yg-w2HT75paNdhbovrBiJh936Zt_flItoWKTwEUMbI1FQHEJk-RkGWa_1JoPRF-7rb3OWK)*/ Dirichlet分布*

#### 定义

满足,和，

，，Dirichlet定义在单纯型上：







#### Dirichlet分布

概率密度函数：

# 抽样方法

# 样本的数字特征



对于随机变量，假若存在，则称它为随机变量的阶原点矩。若存在，则称它为随机变量的阶中心矩。

随机变量的数字特征与样本的数字特征还是不同的。譬如拿均值来说，随机变量的均值往往是指期望，是每一个可能取值与可能取值的概率相乘之和。而样本的均值就是指算术平均，是值之和除以值的数目。

假设样本为，举例：1，2，3，4

## 位置统计量：样本均值

均值是描述数据取值中心位置的一个度量。

样本值的算术平均值：

## 位置统计量：样本中位数

一组样本中，中位数,就是这些数据从小到大排列好了以后中间的那个数字。如果是奇数个数，中位数就是中间那个数。如果是偶数个数，中位数就是中间两个数的算术平均值。

中位数的一个优点是不受个别极端数据的影响，具有稳健性。



## 位置统计量：样本众数

一组样本中，出现次数最多的数据。

## 位置统计量：样本四分位间距

四分位数（Quartile）是统计学中分位数的一种，即把所有数值由小到大排列并分成四等份，处于三个分割点位置的数值就是四分位数。第三四分位数与第一四分位数的差距又称四分位距。

## 离散程度统计量：样本方差

，

## 离散程度统计量：样本标准差

标准差是表示个体间变异大小的指标,反映了整个样本对样本平均数的离散程度,是数据精密度的衡量指标。



## 离散程度统计量：样本极差

一组样本中最大的值减去最小的值。

## 离散程度统计量：未校平方和



## 离散程度统计量：校正平方和



## 离散程度统计量：变异系数

当需要比较两组数据[离散程度](https://baike.baidu.com/item/%E7%A6%BB%E6%95%A3%E7%A8%8B%E5%BA%A6/6775049)大小的时候，如果两组数据的测量尺度相差太大，或者数据[量纲](https://baike.baidu.com/item/%E9%87%8F%E7%BA%B2/100412)的不同，直接使用[标准差](https://baike.baidu.com/item/%E6%A0%87%E5%87%86%E5%B7%AE/1415772)来进行比较不合适，此时就应当消除测量尺度和量纲的影响，而变异系数可以做到这一点，它是原始数据标准差与原始数据[平均数](https://baike.baidu.com/item/%E5%B9%B3%E5%9D%87%E6%95%B0/11031224)的比。

，在SAS里用的是百分比。

## 离散程度统计量：样本的均值的标准误差/标准误

标准误反映样本平均数对总体平均数的变异程度,从而反映抽样误差的大小 ,是量度结果精密度的指标。



## 分布形状统计量：样本偏度

偏度是描述数据分布形态的统计量，其描述的是某总体取值分布的对称性。这个统计量同样需要与正态分布相比较，偏度为0表示其数据分布形态与正态分布的偏斜程度相同；偏度大于0表示其数据分布形态与正态分布相比为正偏或右偏，即有一条长尾巴拖在右边，数据右端有较多的极端值；偏度小于0表示其数据分布形态与正态分布相比为负偏或左偏，即有一条长尾拖在左边，数据左端有较多的极端值。偏度的绝对值数值越大表示其分布形态的偏斜程度越大。

，

## 分布形状统计量：样本峰度

峰度是描述总体中所有取值分布形态陡缓程度的统计量。这个统计量需要与正态分布相比较，峰度为0表示该总体数据分布与正态分布的陡缓程度相同；峰度大于0表示该总体数据分布与正态分布相比较为陡峭，为尖顶峰，成为轻尾；峰度小于0表示该总体数据分布与正态分布相比较为平坦，为平顶峰，称为厚尾。峰度的绝对值数值越大表示其分布形态的陡缓程度与正态分布的差异程度越大。

，注意，不同的统计软件计算峰度的公式会略有差别。

## 分布形状统计量：样本的分位数

对于有限的数集，可以通过把所有观察值高低排序后找出正中间的一个作为[中位数](https://baike.baidu.com/item/%E4%B8%AD%E4%BD%8D%E6%95%B0)。如果观察值有偶数个，则[中位数](https://baike.baidu.com/item/%E4%B8%AD%E4%BD%8D%E6%95%B0)不唯一，通常取最中间的两个数值的平均数作为[中位数](https://baike.baidu.com/item/%E4%B8%AD%E4%BD%8D%E6%95%B0)，即二分位数。

四分位数是统计学中分位数的一种，即把所有数值由小到大排列并分成四等份，处于三个分割点位置的数值就是四分位数。

(1) 第一四分位数(Q1)，又称“较小四分位数”，等于该样本中所有数值由小到大排列后第25%的数字。

(2) 第二四分位数(Q2)，又称“中位数”，等于该样本中所有数值由小到大排列后第50%的数字。

(3) 第三四分位数(Q3)，又称“较大四分位数”，等于该样本中所有数值由小到大排列后第75%的数字。

第三四分位数与第一四分位数的差距又称四分位距。

类似的，还有百分位数。

## 样本的协方差

协方差计算的是不同维度之间的协方差，而不是不同样本之间的协方差。协方差反映一组数据中两个不同属性间变化方向一致性。



## 样本的皮尔森相关系数

皮尔森相关系数，就是用协方差除以两个变量的标准差。相关系数也可以看成协方差：一种剔除了两个变量量纲影响、标准化后的特殊协方差。



作为一种特殊的协方差，它可以反映两个变量变化时是同向还是反向，如果同向变化就为正，反向变化就为负。由于它是标准化后的协方差：它消除了两个变量变化幅度的影响，而只是单纯反映两个变量每单位变化时的相似程度。

# 参数估计

参数估计和假设检验是统计推断的基本内容。**总体的分布类型已知，但其中一个或多个参数未知**，需要对这些未知的参数做出合理的估计，并且对所做出的估计进行评价。这样的过程就叫做参数估计。

## 点估计

点估计就是用一个单一的取值去近似地作为未知参数的估计值。一般会根据样本适当构造统计量作为未知参数的估计量。当样本的取值为时，则以为未知参数的一个估计值。

如用样本的均值来估算总体的均值，用样本的标准差来估算总体的标准差。

### 点估计法的评估标准

对于同一个未知参数，可以提出不同的估计量，即同一批样本，也许既可采用矩估计法估计，也可以用极大似然估计。用无偏性、有效性和相合性来衡量一个参数估计的好坏。

一个估计量的数学期望等于被估计参数的真实值时，这个估计量是无偏的。一个无偏的估计量是指反复抽取样本的次数足够大时，由这些样本计算出来的该估计量的均值可以无限接近被估计参数的真实值。无偏估计的实际意义就说无系统误差。

假设比较参数的两个无偏估计为和，反复抽取样本足够多次时，如果根据计算的取值较更加集中在被估计参数的真实值附近，也就是估计量比的方差更小，则比更有效。

当样本容量增大时，它将越来越接近于被估计参数的真实值，这个标准就叫做相合性。因为当越大时，得到的关于总体的信息就越多。若估计量不具备相合性，那么不论将样本容量如何扩大，都不能将参数估计得足够准确。这样的估计量是不可取的。

在大样本场合下，极大似然法一般都具备这三个性质。

### 矩估计

用样本矩作为总体相关矩的估计，特别是不知道总体分布的情况下，也可以对总体数学期望和方差进行估算。只要知道总体随机变量的一些矩存在(如何证明这些矩是存在的呢？)，就可以做相应的矩估计。但当总体的参数不能表示为矩的函数时，就不能用矩估计。矩估计常常没有利用总体分布函数所提供的信息，因此很难保证它有优良的性质。矩估计法由皮尔逊提出。

### 最小二乘法

### 极大似然估计

极大似然估计的基本思想是，以使样本的出现获得最大概率的参数值作为未知参数的估计值。极大似然法可以简单地理解为在所有可能的选择中选择“看起来最像”的值作为参数的估计，是参数估计用得最多的方法。最早由高斯提出，并由R.A.Fisher发扬光大。

由于独立抽样，构造参数的似然函数为样本值和参数的函数。似然函数是一个事件发生的概率值。或者经过数学处理与概率值成正比的值。



当似然函数有极值，极值点就表明了事情发生了。取到的值，导致了事情以极大似然发生了。既然事情实际上发生了，那么确实该取那个值。

接下来就是构造似然函数的一阶导数。求在条件下的的取值。

### 贝叶斯估计

#### 贝叶斯法则



是先验概率，表示在观察样本之前，依照经验觉得符合某种概率分布

是似然函数，表示在给定模型参数的条件下，样本数据服从这一概率模型的相似程度。

是样本概率

是后验概率，表示在观察一系列样本数据后，模型参数服从的概率分布。

#### 共轭先验分布

共轭分布是一种极大简化贝叶斯分析的方法。其作用是，在贝叶斯公式包括多种概率分布的情况下。使这些分布的未知参数在试验前被赋予的物理意义，延续到试验后，便于分析。

在贝叶斯概率论中，如果后验概率和先验概率满足同样的分布律，那么，先验分布和后验分布被叫做共轭分布，同时，先验分布叫做似然函数的共轭先验分布。

Beta分布是二项式分布的共轭先验分布：观测到的数据符合二项分布，参数的先验分布和后验分布属于Beta分布的情况，就是Beta-Binomial共轭。换言之，Beta分布是二项式分布的共轭先验概率分布。

狄利克雷分布式多项式分布的共轭先验分布：观测到的数据符合多项式分布，参数的先验分布和后验分布属于狄利克雷分布的情况，就是Dirichlet-Multinomial共轭。换言之，狄利克雷分布式多项式分布的共轭先验分布。

#### 常用共轭先验分布

|  |  |  |
| --- | --- | --- |
| 总体分布 | 待估计参数 | 共轭先验分布 |
| 二项分布 | 成功概率 |  |
| 泊松分布 | 均值 |  |
| 指数分布 | 均值的倒数 |  |
| 正态分布(方差已知) | 均值 |  |
| 多项分布 |  | Dirichlet分布 |
| 正态分布(方差未知) | 方差 |  |

## 区间估计

对一个未知量，人们在测量和计算时，常不会满足于仅得到一个近似值，还需估计误差，即要求知道近似值的精确程度。类似地，对于未知参数来说，我们除了关心它的点估计以外，还希望估计出一个范围，并希望知道参数落在这个范围内的可信程度，这就是区间估计。这样的范围通常以区间的形式给出，这样的区间即参数的置信区间。

置信区间，就是一个包含统计量值的取值范围，并且这个范围在一定置信水平下包含参数的真实值：假设某一个分布，含有一个未知参数，对于给定值，若由样本计算出的两个统计量和（），满足,称是的置信水平为的置信区间，称为置信水平，为显著性水平，和分别成为置信水平为的双侧置信区间的置信下限和置信上限。

置信水平为的置信区间的含义是：对于一个确定的置信水平，若反复抽样足够多次，并且每次抽样的样本容量相等，**每一个样本都确定一个区间**，这类区间要么包含的真值，要么不包含的真值。根据大数定律，在这么多的区间中，包含的真值的区间约占，不包含的真值的区间约仅占。

### 利用已知的抽样分布

### 利用区间估计与假设检验的联系

### 利用大样本理论

## 递推参数估计

## 非参数方法

### 条形密度图估计

### 核密度估计

### 近邻方法

## 参数选择规则

### 范数

### AIC

AIC是衡量统计模型拟合优良性的一种标准，由日本统计学家赤池弘次在1974年提出，它建立在熵的概念上，提供了权衡估计模型复杂度和拟合数据优良性的标准。

通常情况下，AIC定义为：



其中k是模型参数个数，L是似然函数。从一组可供选择的模型中选择最佳模型时，通常选择AIC最小的模型。

当两个模型之间存在较大差异时，差异主要体现在似然函数项，当似然函数差异不显著时，上式第一项，即模型复杂度则起作用，从而参数个数少的模型是较好的选择。

一般而言，当模型复杂度提高（k增大）时，似然函数L也会增大，从而使AIC变小，但是k过大时，似然函数增速减缓，导致AIC增大，模型过于复杂容易造成过拟合现象。目标是选取AIC最小的模型，AIC不仅要提高模型拟合度（极大似然），而且引入了惩罚项，使模型参数尽可能少，有助于降低过拟合的可能性。

### BIC

BIC（Bayesian InformationCriterion）贝叶斯信息准则与AIC相似，用于模型选择，1978年由Schwarz提出。训练模型时，增加参数数量，也就是增加模型复杂度，会增大似然函数，但是也会导致过拟合现象，针对该问题，AIC和BIC均引入了与模型参数个数相关的惩罚项，BIC的惩罚项比AIC的大，考虑了样本数量，样本数量过多时，可有效防止模型精度过高造成的模型复杂度过高。



其中，*k*为模型参数个数，*n*为样本数量，*L*为似然函数。*k*ln(*n*)惩罚项在维数过大且训练样本数据相对较少的情况下，可以有效避免出现维度灾难现象。

# 非参数估计

1. 检验均值是否等于某个数。

参数：Z检验，t检验

非参：符号检验（Sign Test)

1. 成对观测是否相等

参数：配对t检验 （Paired t Test）

非参；威尔科克森符号秩检验（Wilcoxon Sign Rank Test）

1. 检验两个样本的均值是否相等

参数：Z检验，t检验

非参：Wilcoxon秩和检验 （Wilcoxon Rank Sum Test）或者 Mann-Whitney U Test

1. 检验单因素影响下，多个样本均值是否相等

参数：单因子方差分析（One-Way ANOVA）

非参：Kruskal–Wallis 方差分析

1. 检验两个因素影响下，多个样本均值是否相等

参数：双因子方差分析（Two-Way ANOVA）

非参：Friedman检验 （Friedman Test）

1. 检验两个样本方差是否相等

参数：F检验

非参：Levene 检验 （Levene Test）

1. 检验多个样本方差是否相等

参数：Bartlett检验 （Bartlett‘s Test）

非参：Levene检验（Levene Test）以上，参数检验对应的假设条件是要总体服从或近似服从正态分布，非参检验没有这个要求。

# 假设检验

## 假设检验和区间估计的联系与区别

假设检验与抽样估计都是统计推断的重要内容。

参数估计是根据样本统计量估计总体参数的真值；假设检验是根据样本统计量来检验对总体参数的先验假设是否成立。

区间估计与假设检验都是根据样本信息对总体参数进行推断，都是以抽样分布为理论依据，都是建立在概率基础上的推断，推断结果都有一定的可信程度或风险。

对同一问题的参数进行推断，二者使用同一样本、同一统计量、同一分布，因而二者可以互相转换。区间估计问题可以转换为假设问题，假设问题也可以转换成区间估计问题。区间估计中的置信区域对应于假设检验中的接受区域，置信区间以外的区域就是假设检验中的拒绝域。

区间估计通常求得的是以样本估计值为中心的双侧置信区间，而假设检验以假设总体参数值为基准，不仅有双侧检验也有单侧检验。

区间估计立足于大概率，通常以较大的把握程度(置信水平)去保证总体参数的置信区间。而假设检验立足于小概率，通常是给定很小的显著性水平去检验对总体参数的先验假设是否成立。

## 假设检验和显著性检验的关系

在统计学领域不是一个概念，常被混用。

1、显著性检验是费希尔提出的。目前使用的多种显著性检验方法都可以在其专著《研究工作者的统计方法》中找到。其中的核心概念是值（判断显著性的概率），通过显著性检验可以获得三个结论：小于通常0.01，宣布检验出一个影响因素；大于通常0.2，影响因素即使存在也微小，不能通过当前实验检测出来；介于两者之间，需进一步设计实验验证。当然，任何统计方法几乎都有不适用的情况。

2、假设检验是内曼-皮尔逊提出的。内曼认为，要想让显著性检验有意义，至少要有两个可能的假设。被检验的假设为“零假设”，其他假设为“备择假设”。此处p用于检测零假设是否成立。

如果原假设为真，而检验的结论却劝你放弃原假设。此时，我们把这种错误称之为第一类错误。通常把第一类错误出现的概率记为。**放弃真。**

  如果原假设不真，而检验的结论却劝你不放弃原假设。此时，我们把这种错误称之为第二类错误。通常把第二类错误出现的概率记为。**接受假。**

通常只限定犯第一类错误的最大概率， 不考虑犯第二类错误的概率。我们把这样的假设检验称为显著性检验，概率称为显著性水平。显著性水平是数学界约定俗成的，一般有 =0.05,0.025.0.01这三种情况。代表着显著性检验的结论错误率必须低于5%或2.5%或1%（统计学中，通常把在现实世界中发生几率小于5%的事件称之为“不可能事件”）。

# 统计量使用

## 总体分布已知(单总体)

## 正态分布

是正态分布，是均值，是方差。

|  |  |  |
| --- | --- | --- |
|  | 未知 | 已知 |
| 未知 | 构造统计量，的置信区间为 |  |
| 已知 | 构造统计量，的置信区间为 | —— |

## 总体分布未知

# 样本分布检验

<https://blog.csdn.net/qq_20207459/article/details/80331805>

## 夏皮罗维尔克检验(Shapiro-Wilk test)

检验小样本数据(数据量)是否服从正态分布。越接近1就越表名数据和正态分布拟合得越好。是国家标准推荐使用的犯第二类错误最小的检验。

将样本数据按数值大小重新排序，使得，计算，其中，是一个常数向量，已经做成表格。，是由正态分布标准化(样本数据减去均值然后除以标准差)后的分位数的期望值组成的向量。注意是从小到大排过序的。

假设：

：样本数据与正态分布没有显著区别

：样本数据与正态分布存在显著区别

有了统计量，我们就可以设定一显著性水平（常见的是0.05），然后获得它的分位数或者临界值，如果则拒绝，否则接受。如果使用，如果 小于显著性水平**.，**则拒绝。

## 柯尔莫哥洛夫KS检验(Kolmogorov-Smirnov)

Kolmogorov-Smirnov检验是基于累计分布函数的，用于检验一个分布是否符合某种理论分布或比较两个经验分布是否有显著差异。

单样本K-S检验是用来检验一个数据的观测经验分布是否符合已知的理论分布。

两样本K-S检验由于对两样本的经验分布函数的位置和形状参数的差异都敏感，所以成为比较两样本的最有用且最常用的非参数方法之一。

构造统计量，其中为观测序列值，为理论序列值或者另一观测序列值。

## 单样本T检验

作用：样本均数和总体均数的比较。用于正态分布的总体中获取个样本，判断总体均数是否和

## 成对T检验

## 独立样本T检验

## 配对样本T检验

## DW检验

## 皮尔逊检验

## 对数似然比检验

## 达戈斯提诺D检验

## Ryan-Joiner检验

## Lilliefors检验

## Anderson–Darling AD检验

# 时间序列分析

## 经典时间序列分析和现代时间序列分析

经典时间序列分析包括：MA、AR、ARMA等平稳时间序列模型，分析时间序列自身的变化规律。现代时间序列分析模型，主要分析时间序列之间的结构关系，单位根检验、协整检验是核心内容，也是现代宏观计量经济学的主要内容。

## 时间序列分析方法

### 描述性时间序列分析

### 统计时间序列分析

#### 频域分析方法

假设任何一种无趋势的时间序列都可以分解成若干不同频率的周期波动

早期的频域分析方法借助富里埃分析从频率的角度揭示时间序列的规律，后来借助了傅立叶变换，用正弦、余弦项之和来逼近某个函数。

20世纪60年代，引入最大熵谱估计理论，进入现代谱分析阶段

特点是：非常有用过的动态数据分析方法，但由于分析方法复杂，结果抽象，具有一定的使用局限性。

#### 时域分析方法

事件的发展通常都具有一定的惯性，这种惯性用统计语言来描述就是序列值之间存在一定的相关关系，这种相关关系通常具有某种统计规律。

寻找出序列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预测序列未来的走势。

时域分析方法的发展过程：

**基础阶段——**

G.U.Yule：1927年，AR模型

G.T.Walker：1931年，MA模型，ARMA模型

**核心阶段——**

G.E.P.Box和G.M.Jenkins，1970年，出版《Time Series Analysis Forecasting

and Control》提出ARIMA模型(Box-Jenkins模型)，实际上是主要运用于单变量、同方差场合的线性模型

**完善阶段——**

异方差场合：Robert F.Engle，1982年，ARCH模型

Bollerslov，1985年GARCH模型

多变量场合：C.Granger，1987年，提出了协整(co-integration)理论

非线性场景：汤家豪等，1980年，门限自回归模型

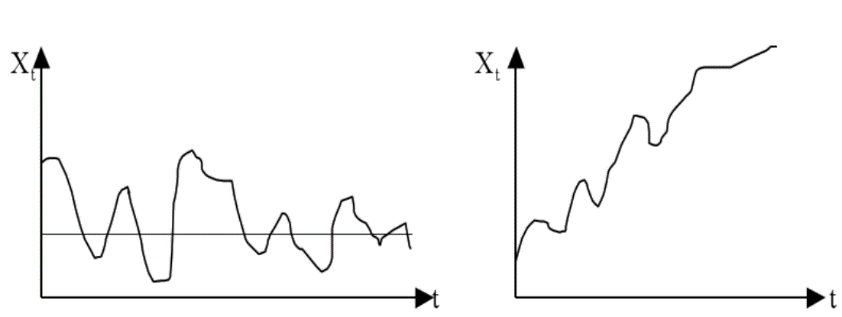
## 虚假回归问题

在线性回归模型中，我们总是以样本决定系数作为回归方程对解释变量与被解释变量样本变化关系的拟合程度的度量。**然而变量之间的样本相关与总体相关是两个概念**，虽然经济变量的样本之间的关系在一定程度上可以说明变量总体之间的关系，但也有例外，这主要取决于经济变量总体分布的性质。有研究表明，**当用两个相互独立的非平稳时间序列建立回归模型时，常常会得到一个在统计意义上显著的回归方程**。我们称之为虚假回归(Spurious Regression)或伪回归

## 平稳性

### 为何要平稳？

每一个统计学问题，都需要对其做一些基本假设。如在一元线性回归中()，我们要假设不相关且非随机(是固定值或当做已知)；独立同分布服从正态分布。**在时间序列分析中，我们考虑很多合理且可以简化的假设，其中最重要的假设是平稳。如果平稳性不满足，很多使用平稳性假设的分析都是错的。所以第一步，首先要对平稳性进行检测。下图左边是平稳序列，右图是非平稳序列。**



之所以要求平稳，从物理的角度来看，就是要求系统是非时变的，这样平稳时间序列才有统计意义，才能用时间序列分析的方法与预测未来。如果系统是时变的，基于过去的数据来预测未来就很荒谬了！

### 严平稳VS宽平稳

**说到平稳，其实有两种平稳——宽平稳、严平稳。**

严平稳相较于宽平稳来说，条件更多更严格，而我们时常运用的时间序列，大多宽平稳就够了。

**什么是严平稳：**是在固定时间和位置的概率分布与所有时间和位置的概率分布相同的随机过程。这样，数学期望和方差这些参数也不随时间和位置变化。（比如白噪声）

**什么是宽平稳：**宽平稳是使用序列的特征统计量来定义的一种平稳性。它认为序列的统计性质主要由它的低阶矩决定，所以只要保证序列低阶矩平稳（二阶），就能保证序列的主要性质近似稳定。

**两者关系：**

一般关系：严平稳条件比宽平稳条件苛刻，通常情况下，严平稳（低阶矩存在）能推出宽平稳成立，而宽平稳序列不能反推严平稳成立。

特例：不存在低阶矩的严平稳序列不满足宽平稳条件，例如服从柯西分布的严平稳序列就不是宽平稳序列。当序列服从多元正态分布时，宽平稳可以推出严平稳。

## 趋势

识别趋势是很重要的。因为传统的时间序列分析方法，都是假设序列是平稳的。但是有趋势存在，序列就不可能平稳。因此，在做时间序列分析时，先要分离出趋势和平稳序列，再分别分析。

### 随机趋势

#### 随机游走

随机游走，，随机游走的一阶差分

最终，，



(注意方差的性质：，当和互相独立时，)

随机游走有常数的期望，有不断增大的方差。就是随机趋势。

### 确定性趋势

#### 常数

，

#### 线性趋势

，是线性趋势

#### 二次趋势

，是二次趋势

#### 周期性或季节性趋势



，为周期，是周期性趋势

#### 余弦趋势



 是余弦趋势

### 混合的趋势

#### 随机游走+漂移

，一阶差分，

## 线性时间序列模型：AR、MA、ARMA、ARIMA模型

### 自回归过程：AR(p)模型

剔除了均值和确定性成分，

,其中是白噪声，所谓白噪声，是指具有有限均值和有限方差的独立同分布随机变量序列。

### 移动平均过程：MA(q)模型

剔除了均值和确定性成分，

,其中都是白噪声

### 自回归移动平均过程：*ARMA(p,q)*模型

剔除了均值和确定性成分，



### 自回归积分滑动平均过程：*ARIMA(p,d,q)*模型

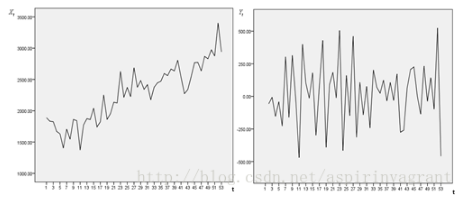
如果序列进行次差分后可以转换为平稳的、可逆的过程，而当进行次差分后仍是非平稳过程，则称此过程具有阶单整性，也称单位根过程。

给定序列，，，……，，当是一个自回归移动平均过程，那么就是自回归积分滑动平均过程。*ARIMA*模型的构建在于：寻找差分次数*d*；估计*ARMA*模型参数。差分次数*d*不宜过大，否则波动过大。

### 给定一个样本时间序列进行ARIMA预测的步骤：

**并不是所有时间序列都适合转为ARIMA处理。ARIMA可以用来对付“随机过程的特征随着时间变化而非固定”且“导致时间序列非平稳的原因是随机而非确定”的问题。**

步骤一：根据时间序列的散点图、自相关函数和偏自相关函数图以*ADF*[单位根检验](https://baike.baidu.com/item/%E5%8D%95%E4%BD%8D%E6%A0%B9%E6%A3%80%E9%AA%8C)其方差、趋势及其季节性变化规律，对序列的平稳性进行识别。



步骤二：非平稳时间序列分析时，若导致非平稳的原因是确定的，可以用的方法主要有趋势拟合模型、季节调整模型、移动平均、指数平滑等方法。若导致非平稳的原因是随机的，方法主要有ARIMA及自回归条件异方差模型等。

对非平稳序列进行平稳化处理。如果数据序列是非平稳的，并存在一定的增长或下降趋势，则需要对数据进行差分处理，如果数据存在异方差，则需对数据进行技术处理，直到处理后的数据的自相关函数值和偏相关函数值无显著地异于零。

步骤三：对序列进行白噪声检验，如果序列是一个白噪声，就没有做分析的必要。

步骤四：当确定使用*ARIMA*模型而不是平滑法来做分析，使用样本时间序列的*ACF、PACF*和*EACF*等，确定是、、和的哪一个并确定阶数。**可以使用信息准则，来确定*p、q、d*的一个组合。**

步骤五：对、、和进行参数估计

步骤六：对估计的参数进行检验，诊断残差序列是否为白噪声。

步骤七：使用模型进行预测。

### 平稳性定义

假定时间序列是由某一个随机过程产生，即的每个值都是从一个概率分布中随机得到，如果满足：

是与时间无关的常数。

是与时间无关的常数

协方差是与时间无关，只与时间间隔相关的常数。则称是平稳的，该随机过程为平稳随机过程。

白噪声过程是平稳的；随机游走过程是非平稳的；随机游走的一阶差分是平稳的。

### 平稳性检测

要使用*AR(p)、MA(q)、ARMA(p,q)*模型，首先要满足的一个条件就是，序列是平稳的。如果不是平稳的，要使用差分或对数的方法，转换为平稳序列再处理。

#### 时间路线图判断

通过画出时间序列图，肉眼观察趋势。平稳序列在图形上往往表现出一种围绕其均值在有限范围内不断波动的过程。非平稳序列在不同时段可能有不同均值，或者不同的时段有大相径庭的波动范围。

#### 样本相关函数判断

区分自相关函数和样本自相关函数，样本自相关函数是自相关函数的一个估计。对滞后*k*期，

自相关函数

样本自相关函数

随着*k*的增加，样本自相关函数下降且趋于零。但从下降速度来看，平稳序列要比非平稳序列快得多。

#### Q统计量判断

#### 单位根检测(DF检验/ADF检验)

### 检验方差？

### 平稳化处理

#### 对数变换

如果有指数趋势，可以先对序列进行对数处理，再进行差分。ARIMA模型不建议用对数变换。因为如果用差分的话，ARIMA可以直接做预测。如果用对数的话，ARIMA之后还要手工做一次对数还原。

#### 平滑

#### 差分

对于随机趋势，可以使用差分的方法处理。但确定性趋势无法通过差分的方法消除，只能通过除去趋势项消除。

#### 分解

   时间序列分为平稳性的与非平稳性的。非平稳时间序列可能包含四种趋势：长期趋势、循环趋势、季节趋势和不规则趋势。

一、趋势分解的一般理解

   可以用移动平均法（季度数据的移动平均项数为4，月度数据的移动平均项数为12）或者模型拟合法（寻找拟合的回归模型）测定长期趋势，然后用原序列除以趋势值以剔除长期趋势；以某月（季）的平均数除以总平均数得到的季节指数测定季节趋势，用上一步的得到的序列除以季节指数以剔除季节趋势；对上一步所得的序列进行移动平均以得到循环趋势，最后跟原序列相比，剩余的就是不规则趋势。

二、深入分析

1、中心化移动平均

    当移动平均项数为奇数时，只用移动平均一次即可；当移动平均项数为偶数时，必须在第一次移动平均的基础上在进行一次同样项数的移动平均。3\*3项移动平均表示对3项移动平均值再进行3项移动平均，也称为5项加权平均。5\*5项移动平均与此类似，也称为9项加权平均。

2、Henderson加权移动平均

    Henderson加权移动平均不同之处在于加权系数，它的加权系数名为Henderson加权移动平均系数。Henderson加权移动平均有5、9、13和23项之别，不规则循环越明显，需要的项数越多。

3、最常用的季节调整方法：X-11和X-12

    这两种方法的基本思路就是采用中心化移动加权平均法测定趋势并逐项剔除，不同之处在于它的最终趋势确定是通过多次迭代和分解完成的。

    X12季节调整方法主要有四种趋势模型：加法模型、乘法模型、对数加法模型和伪加法模型。加法模型主要适用于呈线性增长的数据序列，或者是围绕某一个中值波动的数据序列，如 PMI数据序列等；乘法模型主要适用于呈指数级数增长的序列，如 GDP、工业增加值、投资数据的名义值、实际值及物价的指数序列等；对数加法模型主要适用于同比增速呈线性增长的数据序列，如GDP、工业、投资及 CPI 的同比增速数据；伪加法模型则主要是对某些非负时间序列进行季节调整，它们具有这样的性质：在每一年中的相同月份出现接近于0的正值，在这些月份含有接近于0的季节因子，受这些小因子的影响，季节调整结果将出现偏差。在一年的特定时期，农产品产量就是这样的数据序列。

4、H-P滤波方法

    假定一个时间序列只包含长期趋势和循环趋势（这可以通过前面的X12季节调整方法得到），H-P滤波就是将长期趋势项分离出来的方法。

    平滑系数Lambda：年度数据100，季度数据1600，月度数据14400，当然这个eviews会自动提供。

5、B-P滤波方法

    把时间序列看成不同谐波的叠加，研究时间序列在频率域（frequency domain）里的结构特征。谱分析的基本思想是：把时间序列看作是互不相关的周期（频率）分量的叠加，通过研究和比较各分量的周期变化，以充分揭示时间序列的频域结构，掌握其主要波动特征。时间序列X的变动可以分解成各种不同频率波动的叠加和，根据哪种频率的波动具有更大的贡献率来解释X的周期波动的成分，这就是谱分析（频率分析）名称的缘由。这就是说当具有各种周期的无数个波包含于景气变动中时，看看哪个周期(频率)的波强烈地表现现实景气变动。谱分析中的核心概念是功率谱密度函数（简称功率谱），它集中反映了时间序列中不同频率分量对功率或方差的贡献程度。

    频率和周期的乘积等于2π。功率谱图像横坐标为频率，由左到右频率扩大，由于周期与频率呈反比例关系，周期缩小；纵坐标为功率谱值。

三、指数平滑

1、单指数平滑（一个参数：加权因子）

    实质上就是自适应预期模型，适用于序列值在一个常数均值上下随机波动的情况，无趋势及季节要素的情况，单指数平滑的预测对所有未来的观测值都是常数。

2、双指数平滑（一个参数：加权因子）

    使用相同的参数将单指数平滑进行两次，适用于有线性趋势的序列。

3、Holt-Winters — 无季节趋势（两个参数）

    适用于具有线性时间趋势无季节变差的情形。这种方法与双指数平滑法一样以线性趋势无季节成分进行预测。双指数平滑法只用了一个参数，这种方法用两个参数，形式上与简单一元线性回归模型相同。

4、 Holt-Winters — 加法模型（三个参数）

    适用于具有线性时间趋势和加法季节变化，其实就是在3的基础上加上一个季节因子。需要用简单的方法给出季节因子的第一年的初值，以及截距和斜率的初值。

5、 Holt-Winters — 乘法模型（三个参数）

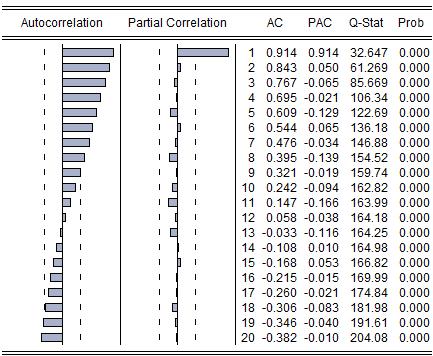
    适用于具有线性时间趋势和乘法季节变化，其实就是在3的基础上乘上一个季节因子。需要用简单的方法给出季节因子的第一年的初值，以及截距和斜率的初值。

### 白噪声的检验

Barlett定理：如果一个时间序列是纯随机的，得到一个观察期数为n的观察序列，那么该序列的延迟非零期的样本自相关系数将近似服从均值为零，方差为序列观察期数倒数的正态分布。

做平稳化处理后，得到的平稳序列，还要检验是否为白噪声。因为白噪声没有分析的意义。

### 模型识别(包括阶数)



分析：平稳的序列的自相关图和偏相关图不是拖尾就是截尾。截尾就是在某阶之后，系数都为 0 ，怎么理解呢，看上面偏相关的图，当阶数为 1 的时候，系数值还是很大， 0.914. 二阶长的时候突然就变成了 0.050. 后面的值都很小，认为是趋于 0 ，这种状况就是截尾。再就是拖尾，拖尾就是有一个衰减的趋势，但是不都为 0 。

如果自相关是拖尾，偏相关截尾，则用 AR 算法  
如果自相关截尾，偏相关拖尾，则用 MA 算法  
如果自相关和偏相关都是拖尾，则用 ARMA 算法， ARIMA 是 ARMA 算法的扩展版。

### 参数估计

### 假设检验

诊断残差序列是否为白噪声。如果残差是白噪声，认为我们构造的ARIMA模型已经尽可能地提取出原序列的信息。当白噪声检验的P值大于0.5，可以认为模型可用。

时间序列的自回归检验通常有两种：Box-Pierce 与 Box-Ljung。两个大致一样，唯一的区别就是后者更加适合小样本。如果你的样本比较少，那么用后面一个比较好。

### 预测

## 线性时间序列模型：ARCH、GARCH模型

## 非线性时间序列模型：双线性模型

## 非线性时间序列模型：门限自回归模型

## 非线性时间序列模型：状态相依模型

## 非线性时间序列模型：马尔科夫转移模型

## 协整

对于不平稳的时间序列，我们可以通过差分的方法使它平稳，但是差分之后的问题是有的经济意义就无法直观解释了，所以我们又有了构建协整关系这一方法。

EG检验和CRDW检验

## [格兰杰因果关系检验](https://blog.csdn.net/qtlyx/article/details/53466437)

## 单位根检验、协整检验和格兰杰因果关系检验三者之间的关系

实证检验步骤：先做单位根检验，看变量序列是否平稳序列，若平稳，可构造[回归模型](http://wenwen.soso.com/z/Search.e?sp=S%E5%9B%9E%E5%BD%92%E6%A8%A1%E5%9E%8B&ch=w.search.yjjlink&cid=w.search.yjjlink)等经典[计量经济学](http://wenwen.soso.com/z/Search.e?sp=S%E8%AE%A1%E9%87%8F%E7%BB%8F%E6%B5%8E%E5%AD%A6&ch=w.search.yjjlink&cid=w.search.yjjlink)模型；若非平稳，进行差分，当进行到第i次差分时序列平稳，则服从i阶单整（注意趋势、[截距](http://wenwen.soso.com/z/Search.e?sp=S%E6%88%AA%E8%B7%9D&ch=w.search.yjjlink&cid=w.search.yjjlink)不同情况选择，根据P值和原假设判定）。若所有检验序列均服从同阶单整，可构造VAR模型，做协整检验（注意滞后期的选择），判断模型内部变量间是否存在协整关系，即是否存在长期均衡关系。如果有，则可以构造VEC模型或者进行Granger因果检验，检验变量之间“谁引起谁变化”，即因果关系。

一、讨论一

1、单位根检验是序列的平稳性检验，如果不检验序列的平稳性直接OLS容易导致伪回归。

2、当检验的数据是平稳的（即不存在单位根），要想进一步考察变量的因果联系，可以采用格兰杰因果检验，但要做格兰杰检验的前提是数据必须是平稳的，否则不能做。

3、当检验的数据是非平稳（即存在单位根），并且各个序列是同阶单整（协整检验的前提），想进一步确定变量之间是否存在协整关系，可以进行协整检验，协整检验主要有EG两[步法](http://wenwen.soso.com/z/Search.e?sp=S%E6%AD%A5%E6%B3%95&ch=w.search.yjjlink&cid=w.search.yjjlink)和JJ检验

A、EG两步法是基于回归[残差](http://wenwen.soso.com/z/Search.e?sp=S%E6%AE%8B%E5%B7%AE&ch=w.search.yjjlink&cid=w.search.yjjlink)的检验，可以通过建立OLS模型检验其残差平稳性

B、JJ检验是基于[回归系数](http://wenwen.soso.com/z/Search.e?sp=S%E5%9B%9E%E5%BD%92%E7%B3%BB%E6%95%B0&ch=w.search.yjjlink&cid=w.search.yjjlink)的检验，前提是建立VAR模型（即模型符合ADL模式）

4、当变量之间存在协整关系时，可以建立ECM进一步考察短期关系，Eviews这里还提供了一个Wald－Granger检验，但此时的格兰杰已经不是因果关系检验，而是变量外生性检验，请注意识别

二、讨论二

1、格兰杰检验只能用于平稳序列！这是格兰杰检验的前提，而其因果关系并非我们通常理解的因与果的关系，而是说x的前期变化能有效地解释y的变化，所以称其为“格兰杰原因”。

2、非平稳序列很可能出现伪回归，协整的意义就是检验它们的[回归方程](http://wenwen.soso.com/z/Search.e?sp=S%E5%9B%9E%E5%BD%92%E6%96%B9%E7%A8%8B&ch=w.search.yjjlink&cid=w.search.yjjlink)所描述的因果关系是否是伪回归，即检验变量之间是否存在稳定的关系。所以，非平稳序列的因果关系检验就是协整检验。

3、平稳性检验有3个作用：1）检验平稳性，若平稳，做格兰杰检验，非平稳，作协正检验。2）协整检验中要用到每个序列的单整阶数。3）判断[时间学](http://wenwen.soso.com/z/Search.e?sp=S%E6%97%B6%E9%97%B4%E5%AD%A6&ch=w.search.yjjlink&cid=w.search.yjjlink)列的数据生成过程。

三、讨论三

其实很多人存在误解。有如下几点，需要澄清：

第一，格兰杰因果检验是检验统计上的时间先后顺序，并不表示而这真正存在因果关系，是否呈因果关系需要根据理论、经验和模型来判定。

第二，格兰杰因果检验的变量应是平稳的，如果单位根检验发现两个变量是不稳定的，那么，不能直接进行格兰杰因果检验，所以，很多人对不平稳的变量进行格兰杰因果检验，这是错误的。

第三，协整结果仅表示变量间存在长期均衡关系，那么，到底是先做格兰杰还是先做协整呢？因为变量不平稳才需要协整，所以，首先因对变量进行差分，平稳后，可以用差分项进行格兰杰因果检验，来判定变量变化的先后时序，之后，进行协整，看变量是否存在长期均衡。

第四，长期均衡并不意味着分析的结束，还应考虑短期波动，要做误差修正检验。

# 实用算法

## 采样算法

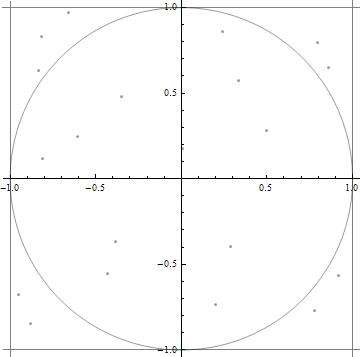
### 蒙特卡罗用于计算积分值

如，形式已知，，积分的显式表达式找不出来。如何计算呢？



使用随机数的方式，在矩形区域产生大量的点对，统计矩形区域落在下方的点的数目，假如一共产生的点对有个，落在下方的点有个，则的值为。数据越接近随机，产生的点对越多，统计的结果越接近真实值。

### 蒙特卡罗用于计算圆面积



使用随机数的方式，在变长为产生大量的点对，统计落在圆里的点对的数目，假如一共产生的点对有个，落在圆里的点有个，那么圆的面积近似为

### 吉布斯采样(Gibbs Sampling)

#### 用途

对于一组属性，已知道条件概率分布、、、，求联合分布

参照http://blog.sina.com.cn/s/blog\_5e9e98210102v1wb.html

#### 预定义