Contents

[LaTeX 2](#_Toc86697767)

[数论 3](#_Toc86697768)

[整除理论 3](#_Toc86697769)

[同余（余数相同） 9](#_Toc86697770)

[连分数 13](#_Toc86697771)

[微积分 15](#_Toc86697772)

[一元导数和积分 15](#_Toc86697773)

[向量代数与空间解析几何 19](#_Toc86697774)

[多元导数和积分 19](#_Toc86697775)

[无穷级数 20](#_Toc86697776)

[应用： 23](#_Toc86697777)

[线性代数 24](#_Toc86697778)

[矩阵 24](#_Toc86697779)

[线性方程组 （研究线性方程 对应的矩阵和向量） 25](#_Toc86697780)

[二次型 (研究二次型 对应的对称矩阵) 26](#_Toc86697781)

[运筹学 27](#_Toc86697782)

[线性规划：目标和约束方程均是线性的 28](#_Toc86697783)

[<Statistics>统计学 28](#_Toc86697784)

[基本概念 29](#_Toc86697785)

[统计推断 29](#_Toc86697786)

[变量间关系 31](#_Toc86697787)

[多元分析 32](#_Toc86697788)

[如何生成概率分布？ 33](#_Toc86697789)

[可视化 34](#_Toc86697790)

[《概率统计》 34](#_Toc86697791)

[基本概念 34](#_Toc86697792)

[生成模型：产生指定分布数据的模型 （均匀分布 -> 指定分布） 35](#_Toc86697793)

[独立性检验，卡方独立性检验 36](#_Toc86697794)

[假设检验 38](#_Toc86697795)

# LaTeX

同时按Alt + ，弹出公式输入框

输入LaTeX, 右下角箭头选择Professional，转换为公式；选择Linear,转换为LaTeX

|  |  |
| --- | --- |
| LaTeX Examples | Built-up format |
|  |  |
|  |  |
|  |  |
| \sqrt[5]{a^2} |  |
|  |  |
|  |  |
| \sum\limits\_{i=0}^{n}{x^i} |  |
|  |  |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| + | - | < | > | = | , | ; | ( | ) | [ |
| / |  | f’ |  |  |  | \| | | | \vec{abc} |  |
| \{ |  | \} |  | \lfloor |  | \langle |  | \hat{abc} |  |
| \alpha |  | \theta |  | o |  | \tau |  | \beta |  |
| \pi |  | \upsilon |  | \phi |  | \gamma |  | \delta |  |
| \kappa |  | \rho |  | \varphi |  | \epsilon |  | \lambda |  |
| \chi |  | \psi |  | \mu |  | \sigma |  | \varepsilon |  |
| \zeta |  | \nu |  | \omega |  | \eta |  | \xi |  |
| \Gamma |  | \Lambda |  | \Sigma |  | \Psi |  | \Delta |  |
| \Upsilon |  | \Omega |  | \Theta |  | \Pi |  | \Phi |  |
| \colon |  | \cdots |  | \times |  | \div |  | \ast |  |
| \pm |  | \cap |  | \cup |  | \vee |  | \wedge |  |
| \leq |  | \geq |  | \equiv |  | \ll |  | \gg |  |
| \sim |  | \simeq |  | \approx |  | \cong |  | \perp |  |
| \subset |  | \subseteq |  | \supset |  | \supseteq |  | \leftarrow |  |
| \in |  | \ni |  | \propto |  | \leftrightarrow |  | \longleftarrow |  |
| \hat{a} |  | \bar{x} |  | \vec{x} |  | \Longrightarrow |  | \Longleftrightarrow |  |
| a^b |  | a\_b |  | \dot{x} |  | \iint |  | \oint |  |
| \exists |  | \forall |  |  |  |  |  | \prod |  |

# 数论

整除理论

整除:

性质：

1. 若
2. 若
3. 若

证明：

必要性，

充分性，

1. 若
2. 多项式，

* 最大公约数理论

最大公约数：

**Euclid算法(欧几里得)求最大公约数：**

假定a > b, 令a = r0, b = r1

r2 = r0 (mod r1) 求余运算：q = int(r0/r1), r2 = r0 – r1\*q

r3 = r1 (mod r2)

…

直至整除为止，最后一个非零余数即为最大公因数

性质：

2. 两个整数的最大公约数一定可以表示成这两个整数的整系数线性组合

252 = 1\*198 + 54 18 = 4\*252 + (-5)\*198

198 = 3\*54 + 36 18 = -1 \* 198 + 4\*54

54 = 1\*36 + 18 18 = 1\*54 + (-1)\*36

35 = 2\*18

证明：不妨设, 由带余数除法得

则 = +

其中 |

由且

继续使用Euclid算法

最大公倍数：

带余数除法：

整数分类：

正整数的a进位表示：

最大公约数理论

1. 公倍数一定是最小公倍数的倍数
2. 公约数一定是最大公约数的约数
3. 若干数乘以的最大公约数等于它们最大公约数乘以

1. 求若干数的最大公约数：将这些数分组，分别求出各组数最大公约数，然后再求这些最大公约数的最大公约数
2. 求与另一个数的最大公约数时，可以把另一个数中与互素的因数去掉
3. 若一个数被整除，则把这个数中与互素的因数去掉后仍被整除
4. 两个数的最小公倍数乘以它们的最大公约数就等于这两个数的乘积的绝对值
5. 若干数的最大公约数一定等于这些数的整系数线性组合，且是这些数的所有整系数线性组合里的最小正整数

推论：

2. 若, 则

* 算术基本定理: 正整数N (>1) 可以**唯一分解**成有限个质数的乘积

标准素因子分解

正除数个数：

所有正除数之和：

= 函数f在a的所有不同的正除数上的值之和

= 函数f在a的所有不同的素除数上的值之和

的素因数分解： , while

: 表示不超过x的最大整数

: 表示x的小数部分

如：20!的标准素因数分解式

解：20以内的素数有{2, 3, 5, 7, 11, 13, 17, 19}

,

K元一次不定方程:

变量个数多于方程个数，且变量取整数值的方程（或方程组）称为不定方程（或组）

求

解：

1. 判断是否有解：
2. 转成k-1个二元不定方程组

其中

1. 用Euclid算法求二元不定方程

或者用观察法求特解，然后求二元不定方程通解：

1. 从而求k元一次不定方程解

如：

求的全部整数解

由(18, 24) = 6 not | 9, 所以无解

求 的全部整数解 （观察法）

由(10, 7) | 17 说明有解

观察容易得 是一组特解

因此全部解是

求的全部整数解 （Euclid算法）

由(907, 731) | 2107 说明有解

若有解，则731必整除 （**取系数绝对值较小的变量**）, 所以

= 从而原方程等价于如下不定方程

=

=

（**直到不定方程变元系数为为止，从而可以直接解出**）

逆推：

求的全部整数解

, 从而转化为方程组

由的通解是

由的通解是

消去, 得原不定方程通解：

非负解或正整数解

令通解, 从而求出自由变量t的取值范围，从而确定t的整数解

例：1只公鸡15元，1只母鸡9元，1只小鸡1元，想300元买一百只鸡，怎么买？

解：令 分别代表买公鸡，母鸡，小鸡的数目，列出不定方程组

由Euclid方法可以求出通解

由

最终解为：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 0 | 1 | 2 | 3 |
| 公鸡 | 0 | 4 | 8 | 12 |
| 母鸡 | 25 | 18 | 11 | 4 |
| 小鸡 | 75 | 78 | 81 | 84 |

## 同余（余数相同）

**同余式**，模m的同余式，a同余式b模m：

* 同余式性质：

1. 同余式可以相加

1. 同余式可以相乘

特别有：

1. , ,

1. ，
2. if , then ，称为对模的逆

容易得,

1. 同余式组

举例：

求写成十进位数时的个位数

解：等价于被10除后的最小非负余数, 即

由

求写成十进位数时的最后两位数

解：等价于被100除后的最小非负余数, 即

想办法先求

逐个计算对模4的剩余，直至模剩余为1 or -1

|  |  |  |
| --- | --- | --- |
| x | 1 | 2 |
| 模4的剩余 | 3 | 1 |

逐个计算对模25的剩余，直至模剩余为1 or -1

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| x | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| 模25的剩余 | 3 | 9 | 2 | 6 | -7 | 4 | 12 | 11 | 8 | -1 |

**同余类**与剩余系

模m的同余类：

同余类性质：

模m的完全剩余系： 同余类：

则整数

模m的最小非负完全剩余系：

模m的绝对最小（完全）剩余系:

模m互素剩余系： 满足

模m的互素剩余系个数： Euler函数

Fermat定理：

Euler定理：

Wilson 定理： 其中是模p的互素剩余系

特别有：

**同余方程**

模m的同余方程： , 其中 是 整系数多项式

1. 在模m的一组完全剩余 系中来解模m的同余方程

如：

解：取模15的绝对最小完全剩余系：-7, …, -1, 0, 1, …7，分别代入计算得

1. 的解
2. 等价变形

while 整系数多项式

如：

while 整系数多项式

常用的是m为素为p及

1. 判断同余方程有解的必要条件： 其中

举例：

求同余方程

解：

其中,

所以原同余方程等价于

一元一次同余方程：

1. 有解的必要条件：
2. 与不定方程 等价

若有解，同余方程解数

若是特解，则它的个解是：

1. 直接解法：

取

由等价变形**同余方程等价**

由同余方程与不定方程等价从而可以转为较小模的新同余方程

即

同余方程与**新的同余方程等价**

重复以上过程，直至能求解模很小的同余方程

求同余方程

解：由 说明存在19个解

说明

一元一次同余方程组：

若是两两互素，则方程组有唯一解

中国剩余定理 （孙子定理）：

同余方程组的解 : , 其中

若不两两互素，将大模拆分成其它模(或模因子)乘积，从而化简方程组

如同余方程组：,

不两两既约，拆分模20=4\*5，15=3\*5，从而等价于

,

由包含， , 故方程组简约为：

, ， 用孙子定理可以求出

例：今有物，不知其数，三三数之剩二，五五数之剩三，七七数之剩二，问物几何？

解：等价于求正整数x满足同余方程组

关键如何求逆？

因此可取

因此可取

因此可取

故满正整数解是

求相邻的四个整数，它们依次可被整除

解：设相邻四个整数依次为, 据题意应满足

故最小的相邻四个正整数是：29348, 29349, 29350, 29351

## 连分数

有限简单连分数：

有理分数可以表示为有限简单连分数

无限简单连分数：

无理数可以唯一表示为无限简单连分数，若令,

则

求连分数表示算法：

在<>，小数写成分数表示

如：求13/5的有限简单连分数

连分数求值

有 限简单连分数，往回推，如

无限简单连分数，寻找最大纯循环部分，列方程求出纯循环部分，然后再求剩下的

欲求， 先求

解方程，由

从而

应用：

求平方根

将大分母分子转化为小分母分子

Euclid算法：

# 微积分

## 一元导数和积分

导数：函数变化率的极限， 函数的斜率

函数可导可微蕴涵着连续性和光滑性

连续函数的复合也是连续的

所有初等函数，初等函数四则运算和复合运算，都可导可微

求导运算法则：

复杂函数都可以由初等函数组合而成，使用如下的求导法则：

隐函数f(x, y) = 0 求

微分法：将x, y看成独立变量，方程两边微分，解出

链式法：将y看成x函数，方程两边同时对x求导，若有y的函数，先对y求导然后y对x求导，解出

参数式函数求：

常见初等函数导数：

证明：

其中均可从上式简单推导得出

函数极值点：只可能是导数为0或没有定义的点，或函数定义域端点

有界区间的连续函数存在最大值和最小值

函数图形的形状

**一阶导数判断增减性** , 增函数

**二阶导数判断凹凸性** , 凹函数

积分：函数曲线下的面积

所有连续函数都是可积分的

定积分性质

若, 则 （定积分不等式）

（积分中值定理）

由导数获得基本积分公式：

积分方法：

若有解析式解，用积分表（常用技术：**变量替换，分部积分**）

常见的换元法：

1. *,*

*含，令*

*含，令*

*含，令*

1. *,*

*令*

令

常见的分部积分：

若无解析式解，用**级数展开**求这类积分

数值积分：梯形法，Simpson法，或Monte Carle积分

导数与求和次序可以调整

积分与求和次序可以调整

曲线非负且连续，则曲线下面积

连续曲线之间的面积

笛卡尔坐标：

极坐标： 小圆扇近似为三角形面积：

求圆面积

取小角度，扇形约等于等腰三角形，底边约等于弧长, 则

取小圆环，面积约等于矩形, 则

连续曲线长度

笛卡尔坐标：

极坐标：

(用切片法)求立体体积 while is 横截面积

## 向量代数与空间解析几何

向量**点积**：

向量在的投影：

Where is the 在的投影分量，方向是方向

向量分解成正交向量：

向量**叉积**： where

平行向量叉积等于零向量

向量叉积表示平行四边形的面积：

所以三角形面积：

平行六面体的体积：

空间直线向量方程： 过点沿方向运动,为位置向量

空间平面向量方程： 过点且垂直于

空间曲线向量形式：

向量函数微分与积分：分别应用于各分量的微分与积分

## 多元导数和积分

多变量微积分是同时在各个方向运用单变量微积分

偏导数：令一个自变量之外的自变量固定，而对这一个变量求导

如=固定, 求对的导数

若多元函数及偏导数连续，则，即偏导结果与次序无关

注：偏导数是有顺序的，是先对求偏导，然后对求偏导

是先对求偏导，然后对求偏导

若, 且是的函数，则

的梯度：

同一元函数一样，多元函数的极值只可能发生在偏导数为0或偏导数不存在的点，或函数定义域端点

鞍点：可微函数在偏导数=0或不存在的点, 附近领域内即存在点, 又存在点, 则称为鞍点

求带约束条件下的函数极值

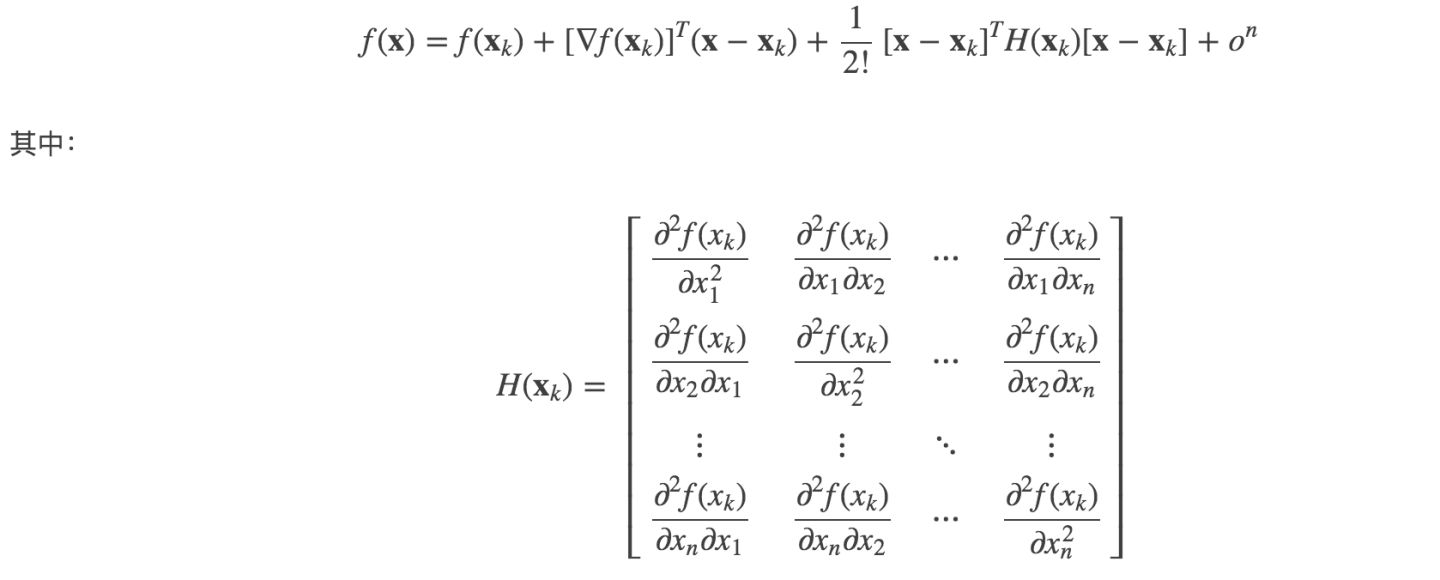
函数， 约束条件,

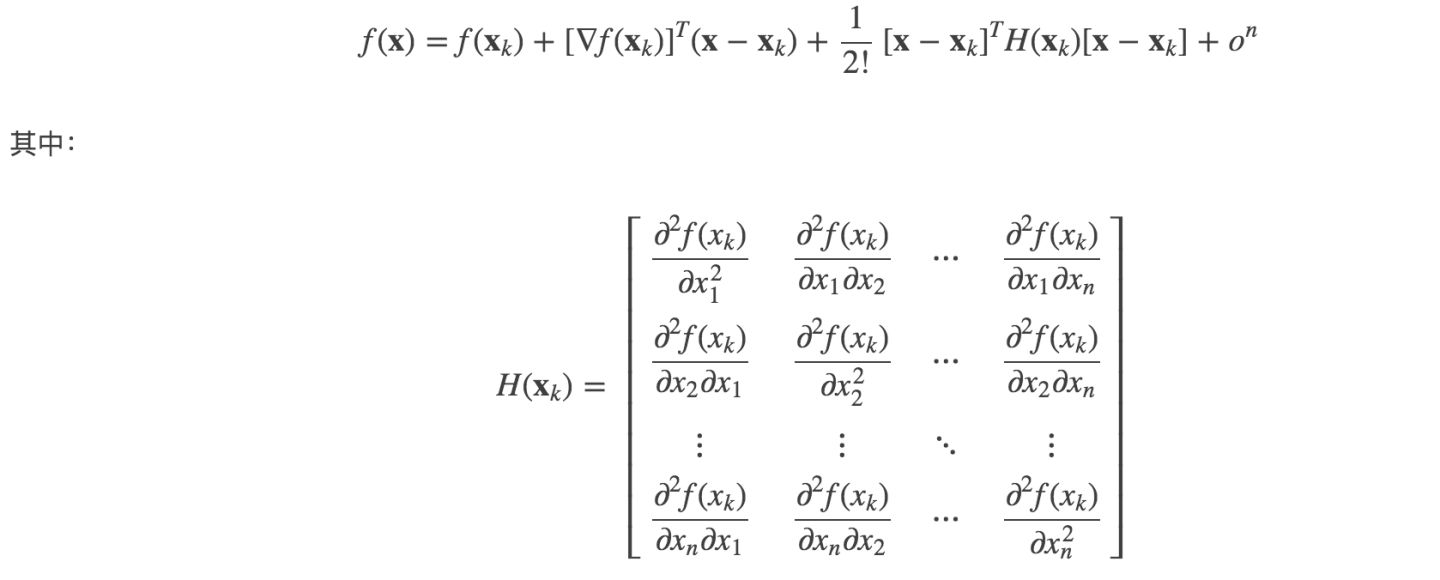
Lagrange乘子法：求解方程组

求体积的二重积分

1. 画草图，画出积分区域的草图并标示出边界曲线
2. 确定y积分限。设想一条竖起线L按y增加的方向纵穿区域R.标好L穿进与穿出的y值，通常是x的函数。
3. 确定x积分限

Tayor Series for Multivariable Functions



Where 

## 无穷级数

如

把循环小数5.232323…表示成两个整数之比

如何判断无穷级数收敛和发散？流程图见托马斯微积分p681

若不存在或，则无穷级数发散

若, 两个无穷级数同收敛或同发散

如

与调和级数同发散

**Taylor定理**

对函数进行局部化处理时，常用**切线近似曲线**： **(柯西中值定理**)， 这种近似是比较粗糙的，而且只在点的附近才有意义

拓展为Taylor定理：若在包含的开区间(a, b)内是n+1阶可微（即函数是连续）的，则对任意 有

特别注意：误差近似n阶多项式 = , 为与之间的某个值。

几何意义：**在局部范围内**用多项式函数近似替代复杂函数。

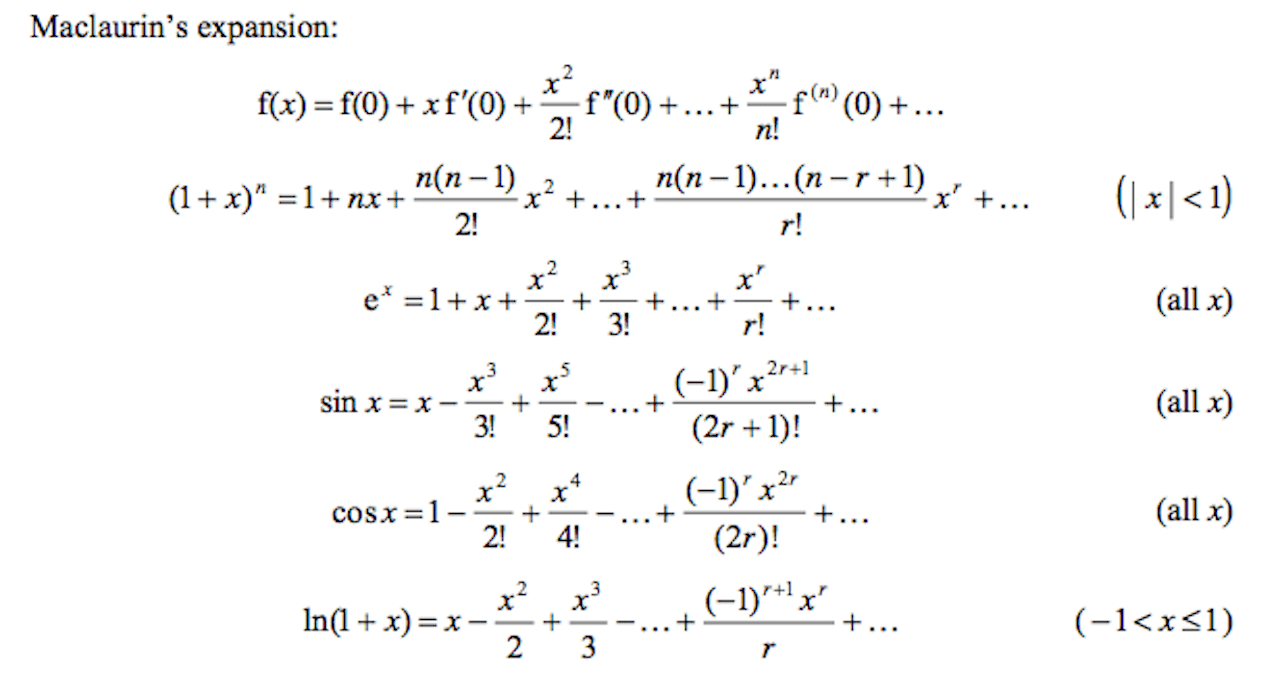
在点上，复杂函数等于近似函数

在点附近，即很小，则复杂函数约等于近似函数

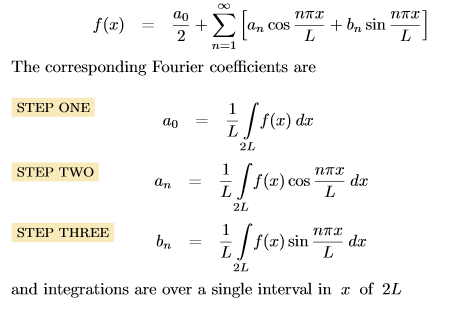
离点很远，二者近似程度难以保证，除非n非常大

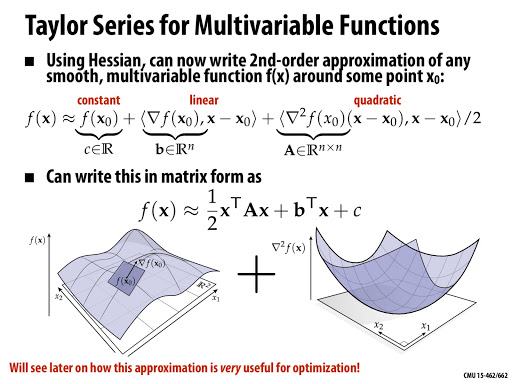
麦克劳林公式： 复杂函数在处的多项式近似

为0与之间的某个值。



**Fourier级数**





## 应用：

证明不等式:

1. 单调性证

函数单调递增，左侧大于等于0

或函数单调递减，右侧小于等于0

令

如果 且当 时，有

1. 最值证
2. 拉格朗日中值公式证

欲证

由拉格朗日中值公式：

只要证明 其中

1. 拉格朗日余项泰勒公式证

求连续曲线与x轴之间的面积？

1. 用的零点分割[a,b]
2. 在每个子区间上积分
3. 把积分的绝对值相加

# 线性代数

## 矩阵

矩阵乘法不满足交换律, 也不满足平方差和完美平方公式，即

矩阵乘法满足结合律和分配律：即

矩阵乘法转置和逆阵：

初等变换法求逆阵：

初等行(列)变换：倍乘，互换，倍加

1. 常数k乘A的某行(列)
2. 互换A的某两行(列）
3. 将A的某行(列) k倍加到另一行(列)

分块矩阵乘法：

,

即B的列向量是齐次方程组的解

,

即矩阵AB的行向量可以由B的行向量线性表示

,

即矩阵AB的列向量可以由A的列向量线性表示

向量组等价： 两个向量组和互相线性表示

## 线性方程组 （研究线性方程 对应的矩阵和向量）

齐次线性方程组

矩阵形式：

(矩阵A的列)向量形式： 即

齐次线性方程组有非零解

（注：齐次线性方程组的解是使矩阵A的列向量线性组合为零的线性组合系数）

线性相关

秩

基础解系向量个数：n(未知量个数) -

通解： 其中是方程组的基础解系

非齐次线性方程组

矩阵形式：

(矩阵A的列)向量形式： 即

非齐次线性方程组有非零解

可以由A的列向量线性表示

秩

通解：

其中是对应齐次线性方程组的基础解系, 是的特解

特别地：

若 A的列向量线性无关，b可由A的列向量线性表示，且表示唯一 有唯一解

若 A的列向量线性相关，b可由A的列向量线性表示，且表示不唯一 有无穷多解

若 b无法由A的列向量线性表示 无解

## 二次型 (研究二次型 对应的对称矩阵)

如：

若二次型只有平方项，没有混合项，称二次型为标准形（又称平方和）

若二次型是标准形，且平方项系数为1, -1, 0时，则二次型也称为规范形

任意二次型 为标准型

* 正交变换法：

注意n阶实对称阵A必存在**正交阵Q**,使得

所以必存在正交变换, 化二次型为标准型，即

令

其中 是A的n 个特征值， A必既相似又 合同于对角阵

如

对应矩阵

对应的特征向量：

对特征向量作标准正交化（用施密特正交化方法）

再单位化，得正交阵

* 配方法 (可逆线性变换)

注意n阶实对称阵A必存在可逆阵C,使得， A必合同于对角阵

令

方法：

1. 若二次型中有平方项，对所有含的项配完全平方（经配方后，使得所余各项中不再含有, 再配第二个平方项， …, 令
2. 若二次型中没有平方项，只有混合项，令, 然后按1)进行配方，连续使1), 2)

如

令 即

即存在可逆线性变换,

特别注意：二次型变换成标准型或规范形，可能不唯一，但正平方项的项数p, 负平方项的项数q是唯一的

正定二次型f，正定矩阵A

*,* 恒有

特别地：标准型系数全为正数，即为正定二次型

# 运筹学

运筹学模型：决策方案 + 目标评判 + 限制条件

max or min 目标函数

s.t. 约束条件

## 线性规划：目标和约束方程均是线性的

图解法：（一般仅用于二元，对应平面坐标系）

1. 确定可行解空间

每个约束条件对应平面坐标系的区域或直线

将每个不等式用等式方程替换，不等式的作用就是将平面(x1, x2)分成两个半空间，为了确定正确的一侧，选择参考点如(0, 0), 如果它满足不等式，则参考点所在的一侧是可行的半空间，否则另一侧是可行的。

1. 最优解的确定

线性规划最优解总是发生在解空间的角点

所以最优争通过枚举有限个角点来找到

代数法：（可以用于多元）

1. 转化为标准型线性规划模型

把≤不等式约束转换成等式约束，在约束的左端，加上非负的松弛变量

把≥不等式约束转换成等式约束，在约束的左端，减去非负的剩余变量

如 转换为

如 转换为

1. n个变量的m个方程, m<n

穷举n-m个变量为非基变量，余下的m个变量为基变量，基本解为基变量的解（由解m个方程得到），验证是否可行（是否满足所有的约束条件），计算目标值

注意：穷举数量 = 对应角点数目 =

单纯形法：通过借助考查解空间中所有可能的基本可行解（角点）的一小部分，极大地减轻了计算负担

# <Statistics>统计学

统计描述　＋　统计推断

Statistical inference

Descriptive statistics is solely concerned with properties of the observed data, descriptive statistics are typically used as a preliminary step before more formal inferences are drawn

Statistical inference is the process of using data analysis to deduce properties of an underlying probability distribution. The majority of the problems in statistical inference: How the translation from subject-matter problem to statistical model

Some common forms of statistical proposition

* point estimate,
* interval estimate,
* credible interval,
* rejection of a hypothesis,
* clustering or classification of data points into groups.

Any statistical inference requires some assumptions. A statistical model is a set of assumptions concerning the generation of the observed data and similar data

three levels of modeling assumptions:

* Fully parametric

The probability distributions describing the data-generation process are assumed to be fully described by a family of probability distributions involving only a finite number of unknown parameters

* Non-parametric:
* Semi-parametric:

Importance of valid models/assumptions

## 基本概念

统计学定义为收集数据，分析数据，由数据得出结论而组成的概念，原则和方法

统计可被定义为在随机性中寻找规律性

经验变量：特征，字段，属性

**理论变量：经验变量→变换**，如z, t, gamma, F变量

**抽样误差**：如果研究被再做一遍，结果未必会和上次一模一样。抽样误差告诉我们，样本离总体的实际值可能有多远

抽样误差的大小依赖于得到样本的方式和样本中包含的观测的个数。

在公布任何一次抽样调查的结果时都应说明抽样误差的大小，不管是比例，均值还是其它形式

标准得分：(观察值-均值)/标准差

不同的变量一般有不同的均值和标准差，统计上，均值和标准差不同时，一个变量的值不能与另一变量的值相比较。将变量值转换化标准得分，从而一个变量的任何值都可以和任何其它变量的值相比较，因为我们知道任何一个得分与均值的相对距离

mean-2\*std mean-std mean mean+std mean-2\*std

-2 -1 0 1 2

任何变量的标准得分的值大部分落在[-2, 2]，2/3的标准得分值落在[-1, 1]。如果一个观察值的标准得分>2 or <-2，那么这个观察值就不寻常的大或小。标准得分常称为t值

## 统计推断

从样本数据得出与总体参数值有关的结论　＝　估计＋假设检验

尽管样本中的信息不完全，而且来自样本的结果一般者不等于总体真值，为了弥补样本结果的不准确，研究者们计算抽样误差－－这个数能使95%的抽样结果都位于由总体真值加减样本误差所得到的区间内

点估计：是一个用来估计总体参数的数，一般用样本统计量作为总体参数的点估计

最大似然法，Bayes估计法

注意：特定数据训练得到的估计方法的准确性不能通过相同的数据进行测试，（举个例子，考试题目如果与课堂习题完全相同，对知识的理解效果就得不到检验，因为仅靠死记硬背就能得到最佳答案。所谓的实力，是解决之前未曾遇到过的问题的能力）这涉及到模型选择的方法：一般采用交叉检验法

置信区间：包含总体参数真值的区间　＝　点估计＋抽样误差

无法回答某个特定的置信区间是否包含总体真值，我们知道的仅仅是－－**在多次抽样中有95%的样本得到的区间包含真值，95%称为置信水平**

**计算方法(bootstrap)：**

**Bootstrap -> large samples -> [metric1, metric2, …] -> sort -> [2.5% percentage, 97.5% percentage] as confidence level**

置信区间是一个随机区间，因为置信区间会因样本的不同而不同，而且不是所有的区间都包含参数真值。

一次抽样无法回答，因为每次抽样的样本集不同，导致点估计和区间估计均不同，但无数次抽样，得到的无数个点估计和区间估计，

无数个点估计的平均为总体真值；

无数个区间估计中，至少有95%的区间估计中包含真值，则95%称为置信水平

换句话，95%置信水平是指多次抽样中有95%的置信区间包含未知的总体参数值，至于在一次抽样得到的置信区间是包含总体参数的众多区间中的一员呢，还是属于个别不含参数值的区间就永远不得而知了。

置信区间特点：

大的样本产生较短的置信区间

低的置信水平产生较短的区间

按照惯例，抽样误差默认是基于95%置信水平的，所有好的调查都应该提供抽样误差，这样读者自己就可以构造置信区间了

多个总体，抽取各自的样本，若样本均值有差异，能否推出总体均值有差异？若有差异，则称为统计显著

因为样本均值的差异，有可能仅仅归于随机性

零假设：假设的内容总是没有差异，或没有改变，或变量间没有关系等等

零假设所提问题的答案是以样本数据为基础的，值得强调的是零假设总是一个与总体参数有关的问题，关于样本统计量如样本均值或方差的零假设是没有意义的，因为样本统计量是已知的

零假设：两人总体均值无差异

假设检验机制：

将样本均值之差1.3，标准差及样本数　→ 　标准得分(t变量)　→ 查t分布表，得出样本均值之差>1.3的概率，即p-value, → 若p-value < 0.05

两种解释：

* 零假设是正确的，观测到的数据恰好是不常发生的那一类；
* 数据倒是常见的那一类，只是零假设是错的

**若零假设为真，那么样本均值差异1.3应很常见，而实际计算出基本样本均值差异1.3的概率小于0.05, 说明不常见，那么选择第二种解释，即认为导致这个小概率的假设－－两总体均值无差异是错的**，即应认为两人总体均值差异不为0.

总结：首先要对我们研究的事物作出某种假设，然后收集数据，并在假设的基础上计算得到该数据的概率，如果这个概率非常小，如小于0.05，则一开始的假设就是错误的

假设检验只考虑到一个可能值，置信区间给了我们一个参数值的可能范围。即使当我们拒绝了零假设，并得到均值之差不为零的结论之后，紧接着的一个问题就是该差异有多大，这个问题是由置信区间来负责回答

## 变量间关系

许多变量可能有统计关系但没有因果关系

比如冰淇淋销量与车祸中儿童受伤数，只是统计关系，背后的第三个变量，温度则是与这个变量有因果关系

用一个变量的值预测另一个变量的值时，它们之间不必非得有因果关系

自变量：先发生的变量

因变量：受影响的变量

两个分类变量的相关分析　－－卡方分析

列联表：描述两个分类变量分布的频率表　contingency table

若自左上角至右下角的对角线上有更多的观测，这种频率结构表明在这些数据中这两个变量是相关的

若两列的百分比分布不同，则两变量存在着关系

若相关系数<0.3，认为关系比较弱; 若[0.3 0.7],则关系适中；若>0.7,则关系强

相关系数计算见公式

两个数值型变量的回归分析和相关分析

离群点对相关系数影响大

列差变量：除自变量外其他所有变量对因变量的效应 = sum(yi - a-bxi)2 观测值－估计值

总的平方和：度量所有的变量对因变量的效应　＝ sum(yi – u)2 观测值－均值

回归平方和：度量自变量对因变量的效应 ＝sum( a+bxi - u)2 估计值－均值

r = sqrt(回归平方和/总的平方和)

相关系数的平方　＝　因变量取值变化的效应可归于自变量的比例

分类自变量和数值因变量的方差分析

若分类自变量只有两种取值（如西海岸和东海岸），那么可以定义虚拟变量，西海岸映射为0, 东海岸映射为1，从而做回归分析

若分类自变量有多种取值，采用方差分析anova

方差分析：用来对比因变量在不同组中的平均值的统计方法

常用盒子图来对比各组中因变量的不同

假定自变量区域，因变量暴力率

方差分析：先根据自变量（地区）分组，再求出每一组的因变量的平均值，我们感兴趣的是因变量的均值在各组之间是否有差异，需要借助方差

自变量平方和：　sum(组均值－总均值)2 度量地区变量的效应

残差平方和：sum(观测－组均值)2

总平方和：sum(观测－总均值)2 = 自变量平方和+残差平方和

r = sqrt(自变量平方和/总平方和)

地区变量与暴力犯罪率变化的R2有关联，而残差变量与剩下的1-R2有关联

类别型变量~类别型变量

独立检验: 卡方检验 （p值：独立概率,若p值小，则拒绝二者独立）

Library(vcd)

Mytable <- xtabs(~Treatment+Improved, data=Arthritis)

Chisq.test(mytable)

相关性度量:相关性强弱系数Phi

Assocstats(mytable)

数值型变量~二元类别型变量 （即样本分成两个组）

用于检验两个总体的均值是否相等？（若p值小，拒绝两组相同）

两组样本独立的t检验，且从正态总体中抽样

t.test(数值变量~二元变量,data=..)

## 多元分析

在多元分析中，残差变量对因变量的影响减少了，这是因为我们将所有自变量的作用同时从残差变量中分离出来，而非逐一分离

统计方法的选择总是决定于所考虑的变量的性质

自变量 因变量 方法

分类型 分类型 偏phi法

数值，哑元 数值型 多元回归

回归模型无截距项，有m个特征，设置

偏phi法：

控制其余变量，计算单变量和因变量的phi

然后取各组中的phi的均值

回归，分类器

若自变量是数值型和分类型，对分类型变量引入哑元

若某分类变量level=5, 则引入5个哑元变量，每个哑元变量取值0/1，每个二进制特征对应分类数据的一个等级

categorical\_feature = pd.Series(['sunny', 'cloudy', 'snowy', 'rainy', 'foggy'])

dummies = pd.get\_dummies(categorical\_feature)

数值型变量~多元类别型变量(即样本分成多个组)

方差分析：名字令人误解，更适合的名字是均值分析：先根据自变量分组，再求出每一组因变量的均值，目的是讨论各组的均值是否不同。但这里判断均值之间是否有差异是借助于方差.

Fit <- aov(数值型变量~类别型变量)

## 如何生成概率分布？

伪随机序列：由随机的种子确定的长序列，只有种子是随机的

真随机：长序列的每个元素都是随机的

代码库一般提供[0 1]范围内的均匀分布函数，对生成的随机序列进行变换，使其遵从特定的分布

0-1分布： y = 正面朝上　if x < 0.5

均匀分布：y = (b-a)\*x + a

正态分布：y = 0.5\*exp(-x\*x/2) and z = std\*y+mean

多元正态分布：N(mean\_vec, cov\_mat)

随机生成变量(x1, x2, ...xn), 从而求f(x1, x2, ...xn)

采用scipy.stats工具包

from scipy import stats

# 二项分布binomial distribution

n次事件中，发生概率为0.5, 恰好发生x次的概率

#计算概率值p(x=650), p(x=651), ...p(x=1200), n=1200, p=0.5

stats.binom.pmf(range(650, 1201), 1200, 0.5)

#生成随机变量x, 服从二项分布n=1200, p=0.5

data = stats.binom.rvs(n=1200, p=0.5, size=1000)

plt.hist(data, bins=24, normed=True) #可视化随机变量pmf

# 泊松分布 Poisson distribution

某路口事件发生的比率是ramda(每天发生次数), 求一段固定时间间隔内(这里是1天）发生x次事件的概率

stats.poisson.pmf(range(0, 10), 2)

#生成随机变量x, 服从泊松分布ramda=2

data = stats.poisson.rvs(mu=2, size=1000)

# 正态分布

stats.norm.pdf(np.arange(-5, 5, 0.1), loc=0, scale=1)

data = stats.norm.rvs(loc=0, scale=1, size=1000)

## 可视化

可视化：盒子图boxplot(数值型~类别型)

两组样本不独立的t检验，组间差异呈正态分布

可视化：带有置信区间的组均值图形

Library(gplots)

Plotmeans(数值型变量~类型别变量)

boxplot

# 《概率统计》

## 基本概念

PDF (probability density function) : 连续型随机变量X的概率密度函数 = CDF的导数

描述随机变量的输出值，在某个确定的取值点附近的可能性的函数

CDF (cumulative distribution function): 分布函数 = PDF的积分

描述随机变量X的概率分布: F(x) = P(X<=x)

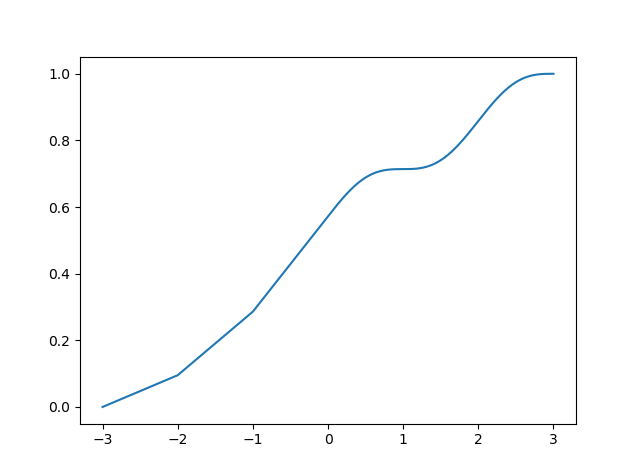
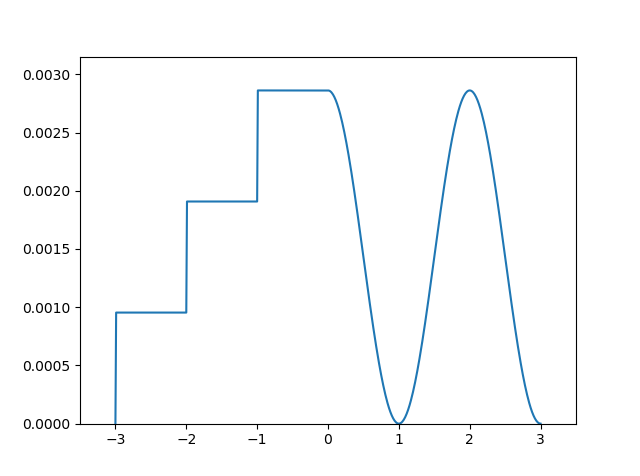
X落在区间(a, b)之内的概率为：P(a < X <= b) = F(b) – F(a)

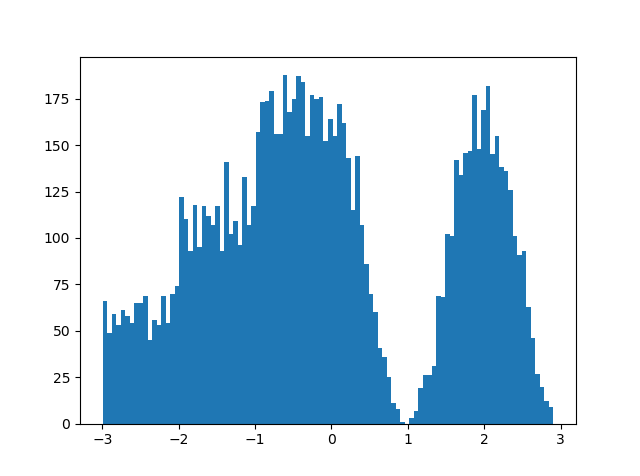
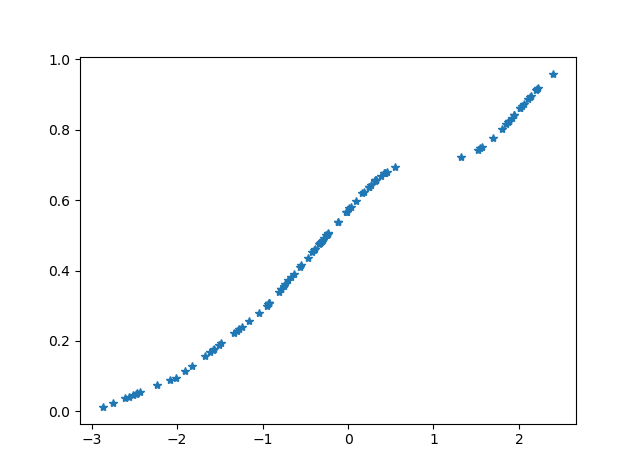
例：

## 生成模型：产生指定分布数据的模型 （均匀分布 -> 指定分布）

生成指定分布PDF （见左上图） -> CDF {x, y=[0, 1]} （见右上图）

生成均匀分布，作为CDF y轴，进行插值(见左下图)，得{x=指定分布, y=均匀分布} （见右下图）





Python 代码：

x\_samples = np.arange(-3, 3.01, 0.01)

pdf = define\_pdf(x\_samples)

cdf = calc\_cdf(pdf)

generate = partial(np.interp, xp=cdf, fp=x\_samples)

u\_rv = np.random.random(10000) #均匀分布[0, 1]

x = generate(u\_rv) #插值成指定分布

## 独立性检验，卡方独立性检验

频率稳定于概率原理：利用随机抽样获得一定数量的样本数据，再利用随机事件发生的概率稳定于概率的原理对问题答案作出推断

比如：为比较两所学校学生的数学水平，采用简单随机抽样抽取88名学生，甲校43名学生10名数学优秀，乙校45名学生7名数学优秀

则甲校数学优秀频率 = 10/43 = 0.23

乙校数学优秀频率 = 7/45 = 0.16

认为甲校数学成绩高于乙校

但是，二者之间确定有差异吗？**样本的随机性可能导致差异**，那么犯样本随机性导致差异的概率大小是多少呢？

也就是说，对于随机样本而言，因为频率具有随机性，频率与概率之间存在误差，所以我们的推断可能犯错误，而且在样本容量较小时，犯错误的可能性会较大

零假设，原假设(null hypothesis): 分类变量X和Y独立

即

|  |  |  |  |
| --- | --- | --- | --- |
| X | Y | | total |
| Y=0 | Y=1 |
| X=0 | a | b | a+b |
| X=1 | c | d | c+d |
| total | a+c | b+d | a+b+c+d |

由独立性可知

=》 -> 0

=》 -> 0

=》 -> 0

=》 -> 0

若以上右边的取值较大时，就可以推断原假设不成立

设计随机变量

随机变量取值大小作为判断原假设是否成立的依据，如何确定大小？

反证法：**根据小概率事件**在一次试验中不大可能发生的规律，假定小概率值，即事件是不太可能发生的

* 若计算出的 时，则可以推断原假设不成立，即X and Y不独立，该推断犯错误的概率不超过小概率值
* 若时，则没有充分证据推断原假设不成立，即X and Y独立

一般我们取小概率值， 那么对应的临界=3.841

对甲乙学校学生成绩优秀是否有差异，采用独立检验求解如下：

解： 定义分类变量，

|  |  |  |  |
| --- | --- | --- | --- |
| X（学校） | Y（数学成绩） | | total |
| Y=0 （不优秀） | Y=1 （优秀） |
| X=0 （甲校） | 33 | 10 | 43 |
| X=1 （乙校） | 38 | 7 | 45 |
| total | 71 | 17 | 88 |

零假设为:学校与数学成绩相互独立，即两校学生的数学成绩优秀率无显著差异

计算= 0.837 < 3.841 =

根据小概率值0.05的卡方检验，没有充分证据推断原假设不成立 ，因此可以认为两校数学成绩优秀率没有显著差异

独立性检验的本质是比较观测值与期望值之间的差异，由所代表的这种差异的大小是通过确定适当的小概率值进行判断的

## 假设检验

P-value:

1. P值一种概率，一种在原假设为真的前提下，出现观察样本以及更极端情况的概率。

2. 拒绝原假设的最小显著性水平。

3. 观察到的(实例的) 显著性水平。

4. 表示对原假设的支持程度，是用于确定是否应该拒绝原假设的另一种方法。