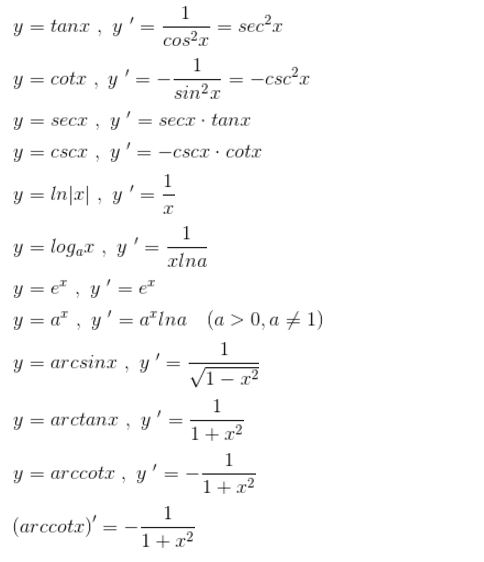
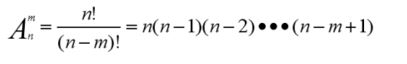
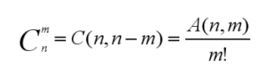
**常用导数:**



**排列组合**：





1、0-1分布:E(X)=p ,D(X)=p(1-p)  
2、二项分布B(n,p)：P(X=k)=C(k\n)p^k·(1-p)^(n-k),E(X)=np,D(X)=np(1-p)  
3、泊松分布X~P(X=k)=(λ^k/k!)·e^-λ,E(X)=λ,D(X)=λ  
4、均匀分布U(a,b):f(x)=1/(b-a),a

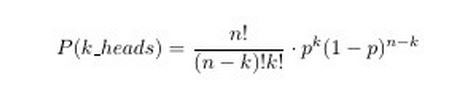
正态分布:钟形曲线 9345d688d43f8794895ee1ead51b0ef41ad53a54

二项分布:其中P称为成功概率。记作ξ~B(n,p)

**正态分布是所有分布趋于极限大样本的分布，属于连续分布。二项分布与泊松分布，则都是离散分布，二项分布的极限分布是泊松分布、泊松分布的极限分布是正态分布，即np=λ，当n很大时，可以近似相等。当n很大时（还没达到连续的程度），可以用泊松分布近似代替二项分布；当n再变大，几乎可以看成连续时，二项分布和泊松分布都可以用正态分布来代替！**

**泊松分布**

泊松分布由二项分布演进而来。二项分布十分好理解，给你n次机会抛硬币，硬币正面向上的概率为p，问在这n次机会中有k次（k<=n）硬币朝上的概率为多少？



在这n次抛硬币中，硬币朝上的次数的期望有多少？

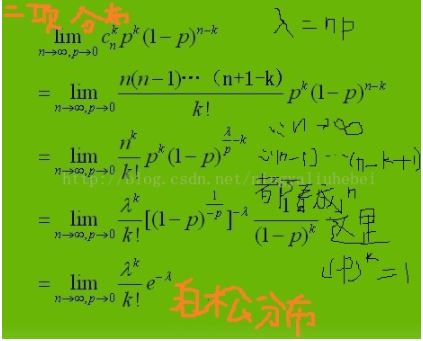
E = p n

如果现在我能根据n的大小来控制p，从而控制这个期望，即无论n为多大，硬币朝上的次数的期望不变(恒为lambda）：



那么当n趋于无穷的时候，P(K\_heads)将趋于泊松分布，即：

捕获

****

所以，实验结果**满足泊松分布的实验即为泊松过程**。**泊松过程把离散的伯努利过程变得连续化了**：原来是抛n次硬币，现在变成了无穷多次抛硬币；原来某次抛硬币得到正面的概率是p，而现在p无限接近于0（p=lambda/n），即：非常难抛出正面朝上的硬币；**但是n次实验中硬币朝上的次数的期望不变，即lambda恒定**。在泊松过程中，我们把**抛出硬币正面这样的事件叫做到达（Arrival）**。把**单位时间内到达的数量，叫做到达率（Arrival Rate）**。

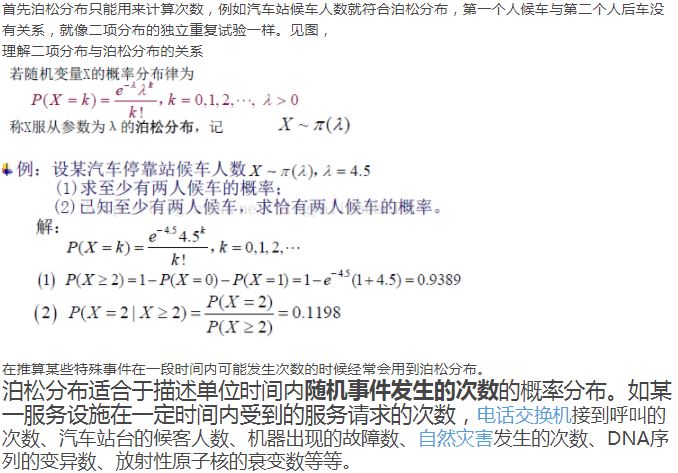
故，泊松过程需要满足以下三个性质：

1. 在任意单位时间长度内，到达率是稳定的。对应于无穷次抛硬币的例子，我们相当于把一个单位时间分割成了无穷次抛硬币的实验，每次实验产生正面的概率都是一样的（为lambda/n），而在这无穷个抛硬币实验之后（即一个单位时间之后）我们期望能抛出lambda个正面的硬币。这个性质类比于在有限次抛硬币（二次分布）的例子中保证了每次掷出硬币为正面的概率都为p。

2. 未来的实验结果与过去的实验结果无关。对应于无穷次抛硬币的例子，之前不管抛出了多少个正面和反面的硬币，都不会影响之后硬币出现的结果。

3. 在极小的一段时间内，有1次到达的概率非常小，没有到达的概率非常大。对应于无穷次抛硬币的例子，我们发现硬币朝上的概率p=lambda/n趋向于0。

**正态分布是所有分布趋于极限大样本的分布，属于连续分布。二项分布与泊松分布，则都是离散分布，二项分布的极限分布是泊松分布、泊松分布的极限分布是正态分布，即np=λ，当n很大时，可以近似相等。当n很大时（还没达到连续的程度），可以用泊松分布近似代替二项分布；当n再变大，几乎可以看成连续时，二项分布和泊松分布都可以用正态分布来代替 ????**

****

**[最大熵原理](https://baike.baidu.com/item/%E6%9C%80%E5%A4%A7%E7%86%B5%E5%8E%9F%E7%90%86" \t "https://baike.baidu.com/item/%E6%9C%80%E5%A4%A7%E7%86%B5%E5%8E%9F%E7%90%86/_blank)**

是一种选择随机变量统计特性最符合客观情况的准则，也称为最大信息原理。随机量的概率分布是很难测定的，一般只能测得其各种均值（如数学期望、方差等）或已知某些限定条件下的值（如峰值、取值个数等），符合测得这些值的分布可有多种、以至无穷多种，通常，其中有一种分布的熵最大。

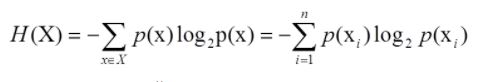
符合已知知识的[概率分布](https://baike.baidu.com/item/%E6%A6%82%E7%8E%87%E5%88%86%E5%B8%83" \t "https://baike.baidu.com/item/%E6%9C%80%E5%A4%A7%E7%86%B5%E5%8E%9F%E7%90%86/_blank)可能不止一个。我们知道，熵定义的实际上是一个[随机变量](https://baike.baidu.com/item/%E9%9A%8F%E6%9C%BA%E5%8F%98%E9%87%8F" \t "https://baike.baidu.com/item/%E6%9C%80%E5%A4%A7%E7%86%B5%E5%8E%9F%E7%90%86/_blank)的不确定性，熵最大的时候，说明随机变量最不确定，换句话说，也就是随机变量最随机，对其行为做准确预测最困难。

最大熵的连续分布：  
已知区间==>均匀分布  
已知均值==>指数分布  
已知均值和标准差（方差）==>正态分布

即在均值和标准差已知的情况下，正态分布是最大熵分布。

# **最大熵模型(Maximum Entropy Model, ME)**

熵的计算公式：(为什么要将熵计算公式定义成这样？ [香农](http://baike.baidu.com/view/607030.htm" \t "http://blog.csdn.net/qianwenhong/article/details/_blank)这样定义肯定有他的道理哈。在后面推导以及应用的时候就能感受到香农这么定义的强大

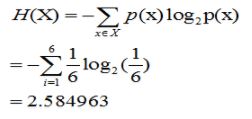


也相当于随机变量X每一个取值的概率乘以对应的概率的对数。

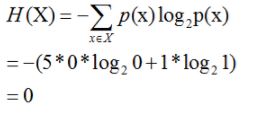
其中，x表示随机变量，与之相对应的是所有可能输出的集合，定义为符号集,随机变量的输出用x表示。P(x)表示输出概率函数。变量的不确定性越大，熵也就越大，把它搞清楚所需要的信息量也就越大.

现在我们可以用熵的公式来比较图4与5中到底哪个更随机了。

* X均匀分布，正常骰子投掷时的熵：

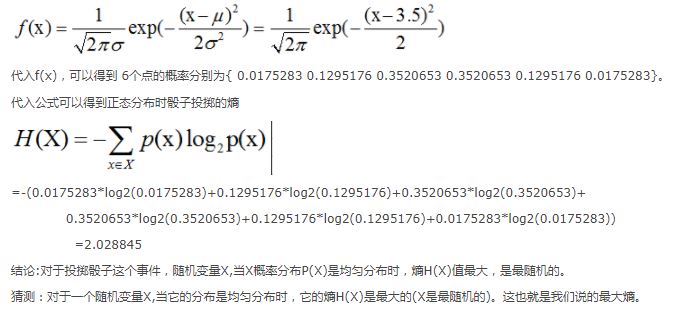


* X特殊分布，韦小宝骰子投掷时的熵：



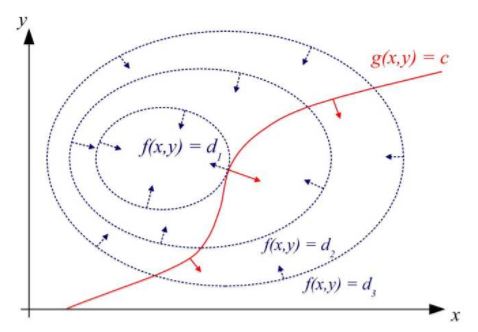
* X服从正态分布X~N(3.5,1)，正态分布时骰子投掷的熵： H(X)=2.028845

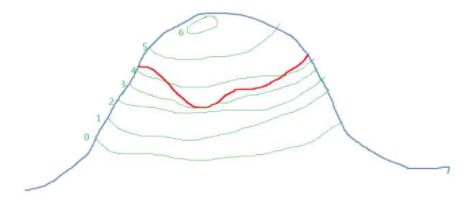
由X~N(3.5,1)可得随机变量X的概率分布为

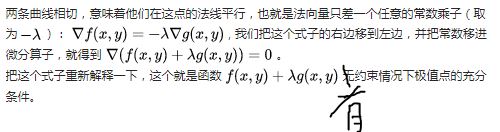


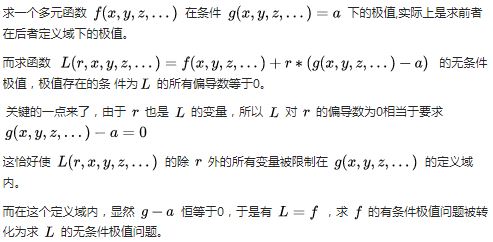
**拉格朗日乘数法**（Lagrange multiplier）有很直观的几何意义。  
举个2维的例子来说明：  
假设有自变量x和y，给定约束条件g(x,y)=c，要求f(x,y)在约束g下的极值。

我们可以画出f的等高线图，如下图。此时，约束g=c由于只有一个自由度，因此也是图中的一条曲线（红色曲线所示）。显然地，当约束曲线g=c与某一条等高线f=d1相切时，函数f取得极值。  
两曲线相切等价于两曲线在切点处拥有共线的法向量。因此可得函数f(x,y)与g(x,y)在切点处的梯度（gradient）成正比。  
于是我们便可以列出方程组求解切点的坐标(x,y)，进而得到函数f的极值。



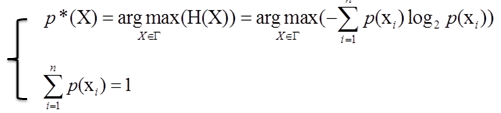




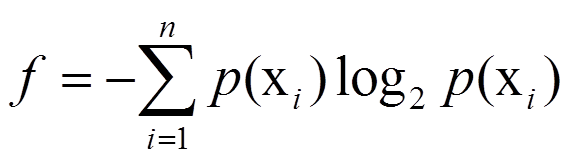
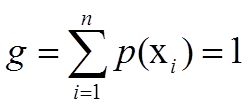


## 2.2 单约束 最大熵推导

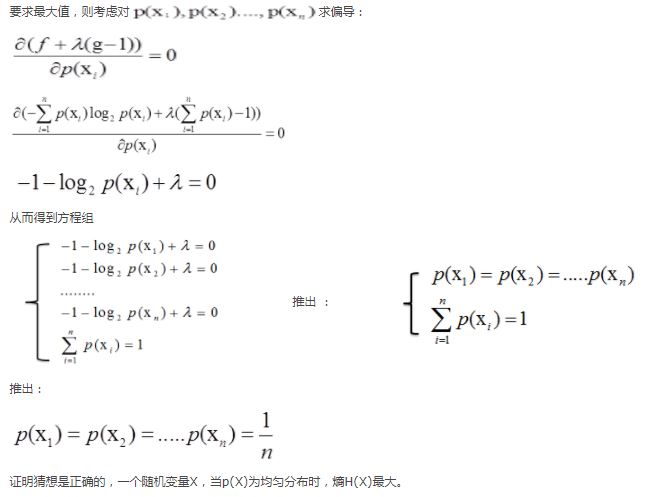
单约束最大熵的基本想法就是在一定条件下(概率和为1)，找到一个分布p\*(X),使熵H(X)的值达到最大。可以写成：



在约束下求最大值，使用**拉格朗日乘子法**。设

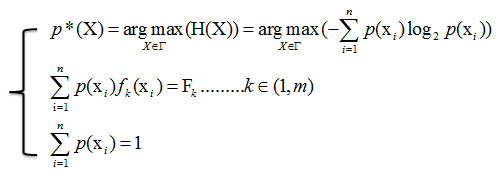
 

得到拉格朗日方程 f + λ(g-1)



## 2.3 多约束 最大熵推导

多约束最大熵的基本想法就是在多个条件下(共m+1个约束)，找到一个分布p\*(X),使熵H(X)的值达到最大。可以写成：



上面讲的都是关于一个随机变量的熵H(X)，对于条件概率模型P(Y|X)的条件熵H(Y|X)也常常被用到，定义为

081422001714768