## 概率函数P(x)、概率分布函数F(x)、概率密度函数f(x)

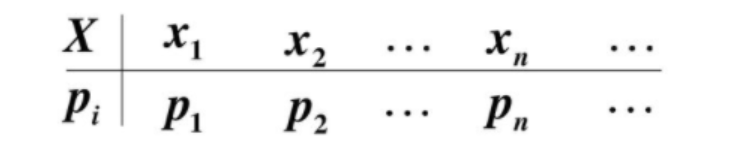
进入主题前，先明确几个概念：  
**离散型变量（或取值个数有限的变量）**：取值可一一列举，且总数是确定的，如投骰子出现的点数（1点、2点、3点、4点、5点、6点）。  
**连续型变量（或取值个数无限的变量）**：取值无法一一列举，且总数是不确定的，如所有的自然数（0、1、2、3……）。

**离散型变量取某个值xi的概率P(xi)是个确定的值（虽然很多时候我们不知道这个值是多少），即P(xi)≠0**：例如，投一次骰子出现2点的概率是P(2)=1/6。

**连续型变量取某个值xi的概率P(xi)=0**：对于连续型变量而言，**“取某个具体值的概率”的说法是无意义的，因为取任何单个值的概率都等于0**，只能说**“取值落在某个区间内的概率”**，或**“取值落在某个值邻域内的概率”**，

即**只能说P(a<xi≤b)，而不能说P(xi)。** 为什么是这样？且看下例：  
  例如，从所有自然数中任取一个数，问这个数等于5的概率是多少？从所有的自然数中取一个，当然是有可能取到5的，但是自然数有无穷多个，因此取到5的概率是1/∞，也就是0。先说离散型随机变量的分布情况

**概率分布**：给出了所有取值及其对应的概率（少一个也不行），**只对离散型变量有意义**。例如



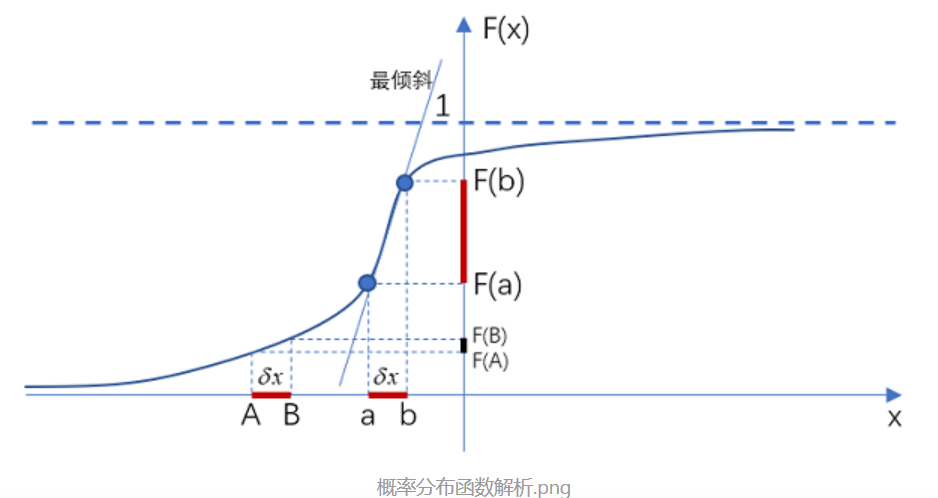
**概率函数**：用函数形式给出每个取值发生的概率，P(x)（x=x1，x2，x3，……），**只对离散型变量有意义**，实际上是对概率分布的数学描述。

### 概率分布和概率函数只对离散型变量有意义，那如何描述连续型变量呢？

答案就是“**概率分布函数F(x)**”和“**概率密度函数f(x)**”，当然这两者也是可以描述离散型变量的。

**概率分布函数F(x)：**

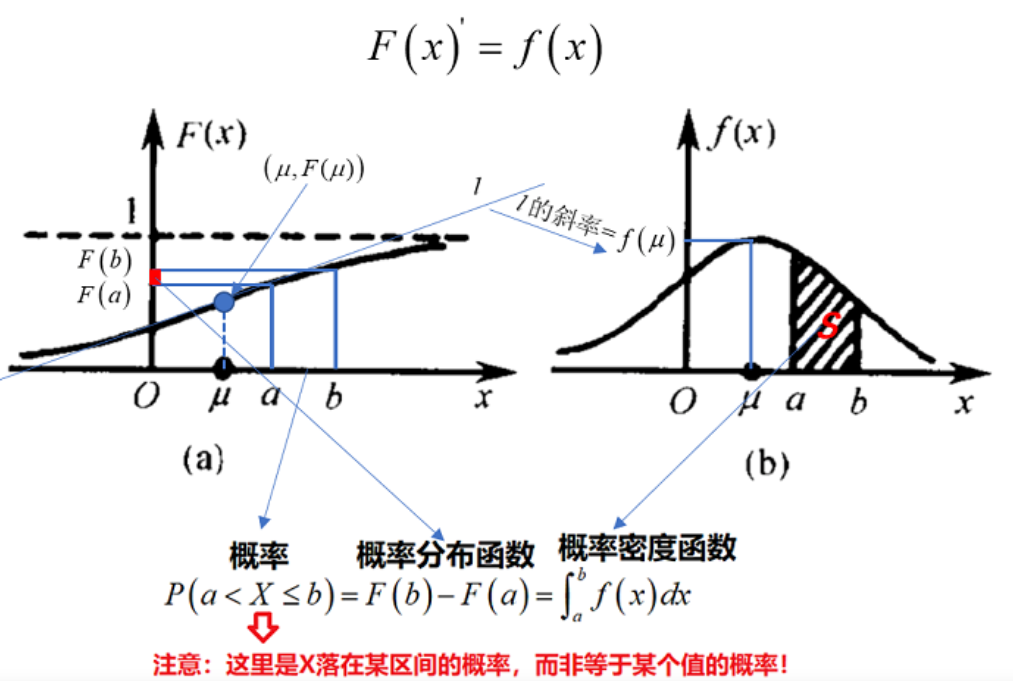
**给出取值小于某个值的概率，是概率的累加形式**，即：F(xi)=P(x<xi)=sum(P(x1),P(x2),……,P(xi))（对于离散型变量）或求积分（对于连续型变量）。

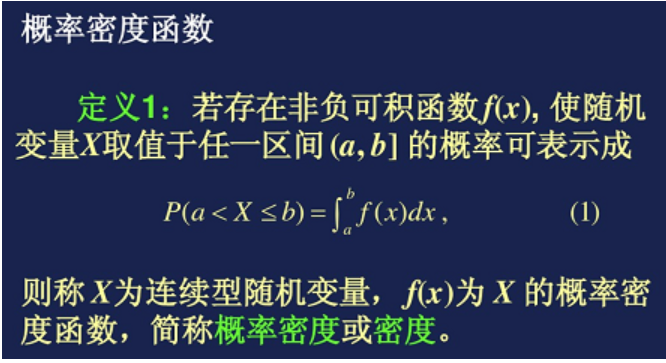


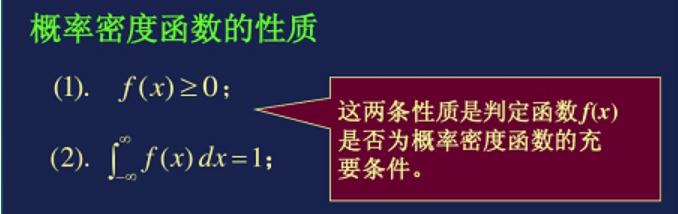
概率分布函数是一个累加函数，递增的函数，对于一个连续型变量，无法给出某一点的发生概率，对于某一个范围内、区间内事件发生的概率P(a<x<b)=F(b)-F(a)，从上图可以看出，

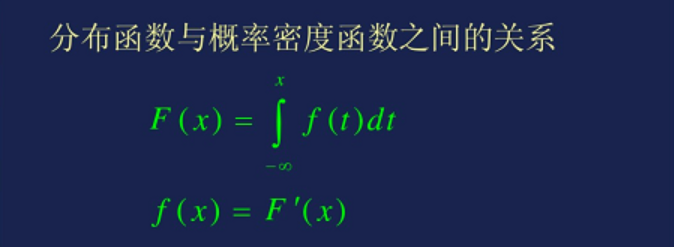
[A, B]区间出现的概率远小于[a, b]区间出现的概率，故概率分布函数的斜率好像可以来评估某一事件发生概率的大小。故概率密度函数出现了。

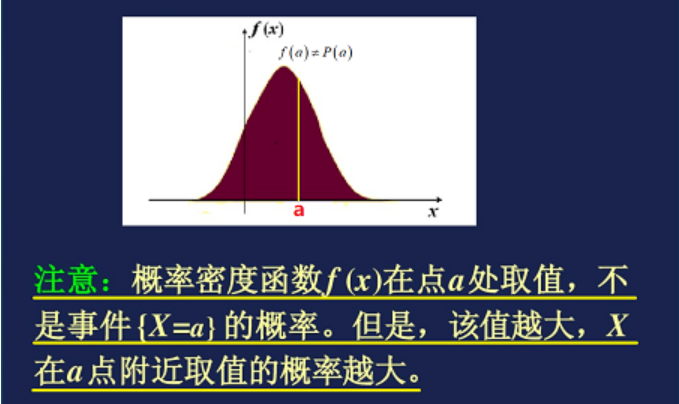
**概率密度函数f(x)**：给出了变量落在某值xi邻域内（或者某个区间内）的**概率变化快慢**，**概率密度函数的值不是概率，而是概率的变化率，概率密度函数下面的面积才是概率**。可以说概率密度函数是分布函数的导函数，从概率密度函数的y值我们可以直接看出某值xi邻域内事件发生概率的大小



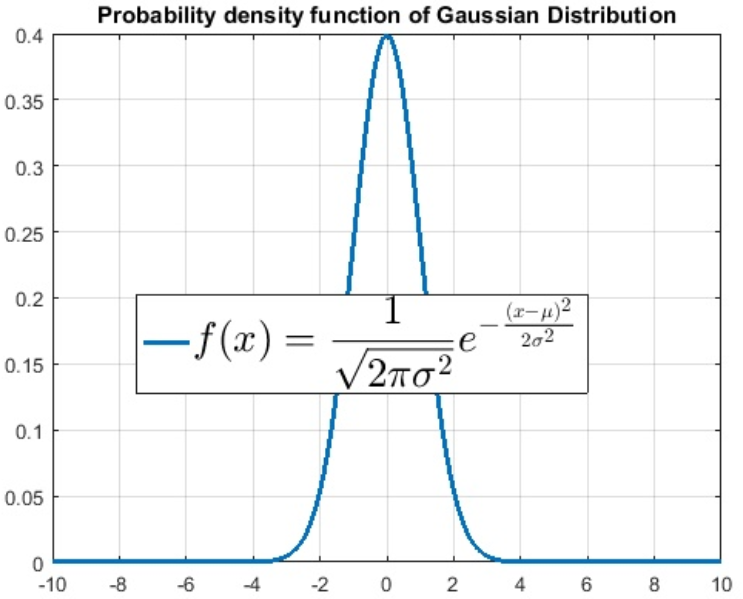
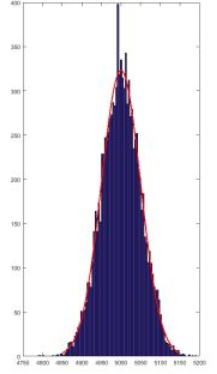








在日常的数据分析任务中，有时候我们用polt将样本某些特征比如年纪用条形图或是柱状图画出来，看起来挺像正太分布的密度函数的，但是他并不是，polt画的结果是x是特征的不同区间，y是各个区间出现的次数，这其实暗合了密度函数



Polt的结果给出了各个区间出现次数、频率的高低，其实就是可以理解为该区间事件发生的概率的高低，这与概率密度函数暗合，可以作为一个变量分布的一个参考。

# 中心极限定理

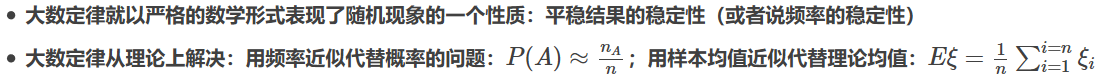
中心极限定理指的是给定一个任意分布的总体。我每次从这些总体中随机抽取 n 个抽样，一共抽 m 次，一共抽mxn次，分m组。 然后把这 m 组抽样分别求出平均值。 这些平均值的分布**接近正态分布**。

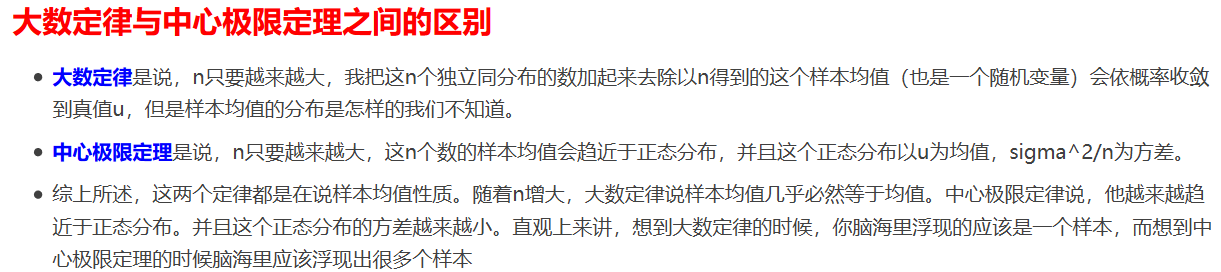
**注意** ：不一定是平均值，可以是任何统计量。

意思就是当满足某些条件的时候，比如Sample Size比较大，采样次数区域无穷大的时候，就越接近正态分布。**而这个定理神奇的地方在于，无论是什么分布的随机变量，都满足这个定理。自然界与生产中，一些现象受到许多相互独立的随机因素的影响，如果每个因素所产生的影响都很微小时，总的影响可以看作是服从正态分布的.在独立同分布的情况下，无论 的分布函数为何，它们的平均数，当充分大的时候总是近似地服从正态分布。**

# 大数定律

**大数定律**是说，n只要越来越大，我把这n个独立同分布的数加起来去除以n得到的这个样本均值（也是一个随机变量）会依概率收敛到真值u，**但是样本均值的分布是怎样的我们不知道**。**大数定律从理论上解决：用频率近似代替概率的问题**





# 高斯分布

# 

正态分布的概率密度函数曲线呈钟形，因此人们又经常称之为**钟形曲线**（类似于寺庙里的大钟，因此得名）。我们通常所说的**标准正态分布**是位置参数，



正态分布的定义

有几种不同的方法用来说明一个随机变量。最直观的方法是概率密度函数，这种方法能够表示随机变量每个取值有多大的可能性。累积分布函数是一种概率上更加清楚的方法，请看下边的例子。还有一些其他的等价方法，例如cumulant、特征函数、动差生成函数以及cumulant-生成函数。这些方法中有一些对于理论工作非常有用，但是不够直观。请参考关于概率分布的讨论。

下图分别是高斯分布的概率密度函数和概率分布函数

