[1. Term 2](#_Toc468261818)

[2. 概念 2](#_Toc468261819)

[2-1. 随机变量 2](#_Toc468261820)

[2-1-1. 随机变量可以是离散型或者连续型的 2](#_Toc468261821)

[2-2. 概率分布 2](#_Toc468261822)

[2-2-1. 离散概率分布 3](#_Toc468261823)

[2-2-2. 连续概率分布 3](#_Toc468261824)

[2-2-3. 为什么概率密度函数线下的面积等于1 4](#_Toc468261825)

[2-2-4. 概率密度函数对特定x取值无意义 4](#_Toc468261826)

[2-3. 累计分布函数 4](#_Toc468261827)

[2-3-1. 离散随机变量和连续随机变量都有累计分布函数 4](#_Toc468261828)

[3. 正态分布 4](#_Toc468261829)

[3-1. 正态分布为什么能变换成标准正态分布来计算 4](#_Toc468261830)

# Term

随机变量：random variable

离散随机变量：discrete random variable

连续随机变量： continuous random variable

概率： probability

概率分布： probability distribution

离散随机变量的概率分布函数： probability mass function

概率密度函数： probability density function

累计分布函数： cumulative distribution function CDF

分布函数： distribution function

置信区间： confidence interval CI

置信水平： level of confidence

# 概念

## 随机变量

有些实验如掷骰子或道琼斯指数，他们的结果本身就是数量型的。但是有些实验，如天气，其样本空间是分类的，例如{晴朗，阴天，多云}。 是非数值型的。

为了有一致的数据基础来计算概率，人们希望实验的结果都是数值型的。随机变量是一个为样本空间中的每个元素赋予一个实数的函数。如果结果是分类的，可以为每个结果任意分配一个实数。 如0=晴天 1=多云 2 = 阴天等。

但是这些数字没有实际意义，只是为了将所有的实验结果都统一为数值的变现形式。

随机变量通常以大写的罗马字母来表示，如X 或者Y

### 随机变量可以是离散型或者连续型的

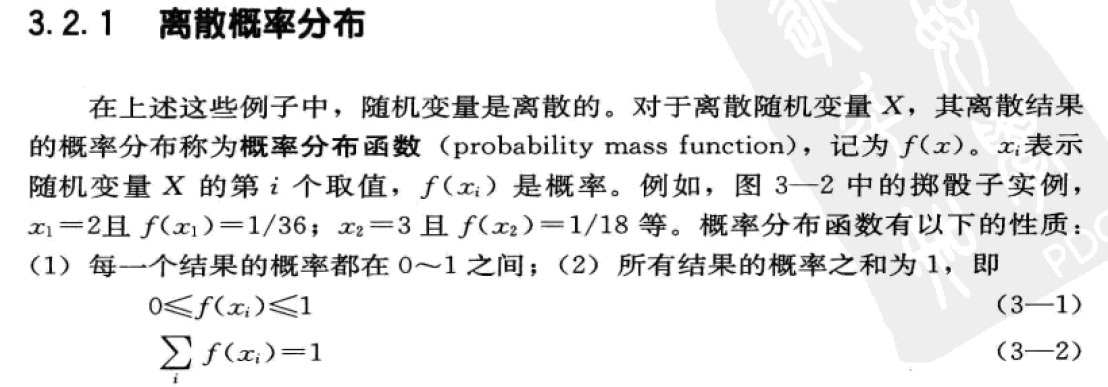
离散随机变量的结果时可以计数的，如掷骰子的各种可能结果

连续随机变量假设其结果时一些列连续区域的数

## 概率分布

是对随机变量的可能取值及其发生概率的描述

### 离散概率分布

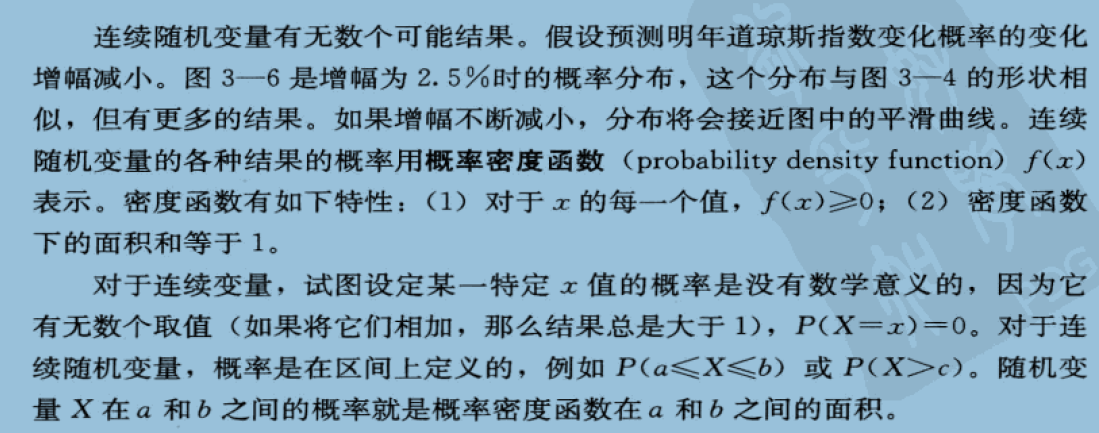


离散随机变量一般不说它的密度函数，只说它的分布函数。（其实这里的分布函数就相当于是连续随机变量的概率密度函数）

它的分布函数是可以对某一个特定的x取值的，而且取的值就是该x锁对应的概率

而且对于离散随机变量来说，既有分布函数，又有累计分布函数

### 连续概率分布



连续随机变量没有像离散随机变量那样的probability **mass function, 描述它的函数有两个： 概率密度函数f(x)和累计分布函数F(x)。**

**f(x) 不像离散随机变量那样可以对特定x求值，连续随机变量对任意特定x求出的值没有意义（虽然它的确能求出一个值）。 f(x) 值能用在积分公式中，连续性随机变量只能求某个某个区间的值，即f(x)dx 的值，表示x 位于该dx 区间内的概率，其数值等于概率密度函数线下的面积。**

### 为什么概率密度函数线下的面积等于1

线下的面积其实是对f(x) 求积分，统计术语里表示 P(a<=X<=b) 的概率。当a和b扩展到正无穷和负无穷时，表示的是x位于负无穷到正无穷的概率，当然其值为1了。这也就是为什么概率密度函数线下的面积等于1

注意，只有连续性随机变量才有概率密度函数的概念。

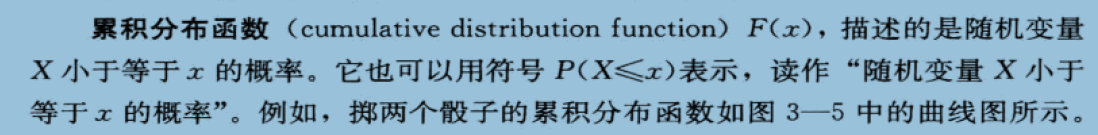
而对于离散型随机变量，其概率分布函数可以对各个特定的x 取值，然后所有的值加起来等于1

### 概率密度函数对特定x取值无意义

虽然它作为函数，给定一个x确实可以求出一个值，但是这个值在统计学里是没有意义的，也不能反映任何的概率取值情况。

概率密度函数只能用在微积分公式中，来求一个区间的概率取值。

## 累计分布函数



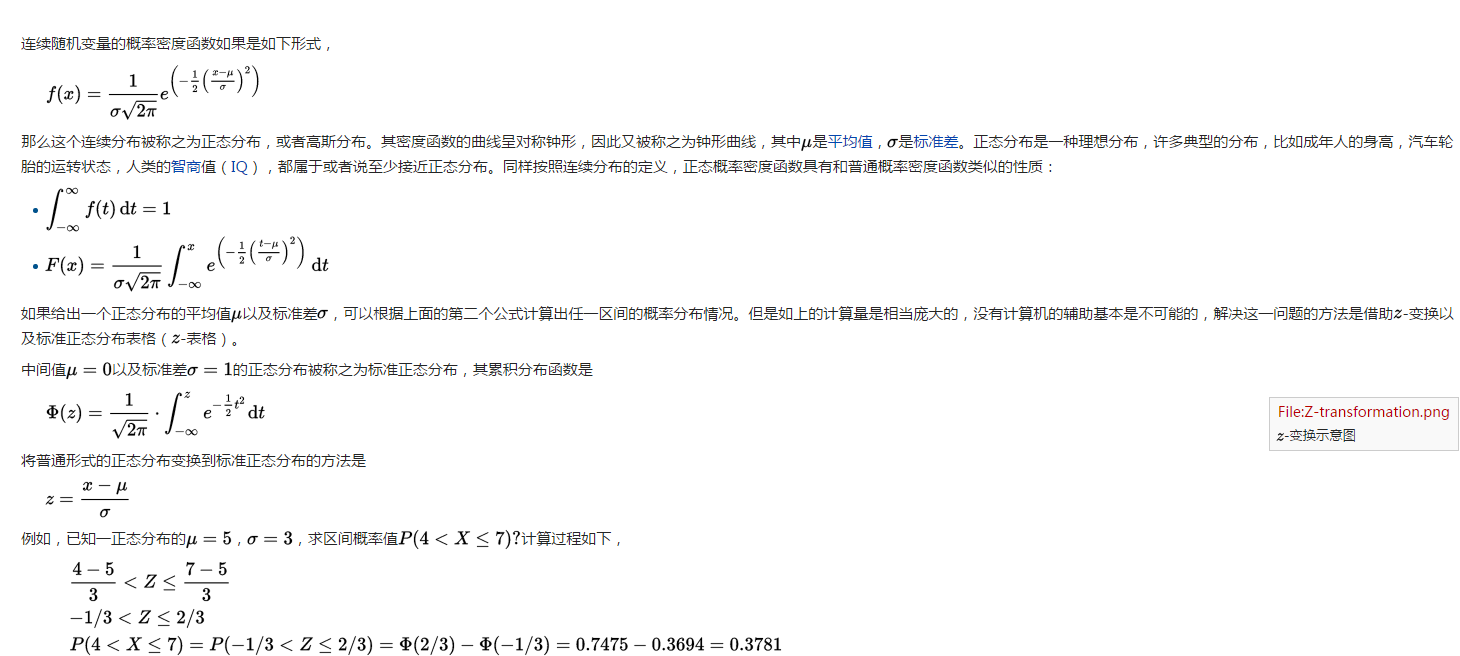
### 离散随机变量和连续随机变量都有累计分布函数

# 正态分布

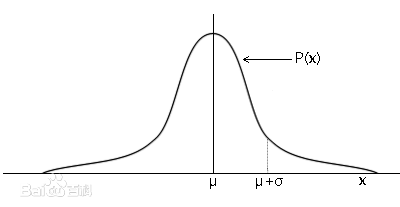
## 正态分布为什么能变换成标准正态分布来计算

对于分布来说，我们关心的是其概率，而不是其形态。

概率从数值上等于概率密度函数线下的面积。考虑 u 不同



## 正态分布下的一些特殊值



标准正态分布又称为u分布，是以0为均数、以1为标准差的正态分布，记为N（0，1）

标准正态分布曲线下面积分布规律是：**在-1.96～+1.96范围内曲线下的面积等于0.9500，在-2.58～+2.58范围内曲线下面积为0.9900。**统计学家还制定了一张统计用表（自由度为∞时），借助该表就可以估计出某些特殊u1和u2值范围内的曲线下面积。

这两个值只所以重要是因为在算置信区间时，常用的置信水平就是95% 和99%

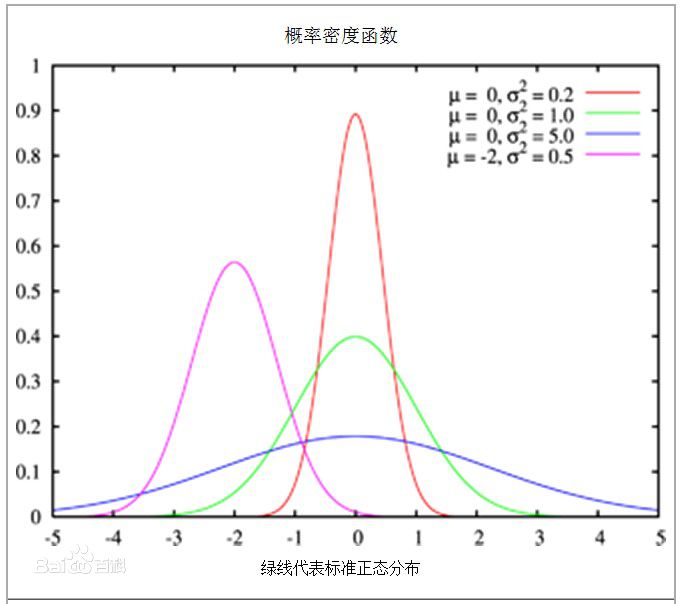
函数曲线下68.268949%的面积在平均数左右的一个标准差范围内。

95.449974%的面积在平均数左右两个标准差的范围内。 （[-2, 2] 比[-1.96, 1.96] 范围大一些）

99.730020%的面积在平均数左右三个标准差的范围内。

99.993666%的面积在平均数左右四个标准差的范围内。

### 方差越小，分布的越集中，图形越陡峭



# 抽样 Sampling

## 为什么要抽样

## 抽样的设计--抽样计划

### 抽样目的

抽样的目的决定了选择合理的抽样框和选择何种抽样方法。

抽样目的是进行抽样分析的第一步

### 抽样框

确定待抽样的总体

抽样框的选择一定要力求全面客观，如果抽样框本身的选择是偏薄的，那根据抽样框得到抽样样本，然后进行分析得出的结论肯定也是不准确的。

### 抽样方法是抽样计划的一项内容

## 抽样方法

### 没有固定的抽样方法

抽样是为了能更好的模拟整体，所以选择合众抽样方法要以整体数据的分布情况来定。否则，可能会产生抽样得到的样本无法代表整体的情况

## 抽样误差 sampling varance

### 非抽样误差

### 抽样（统计性）误差

## 基于样本分析结果进行整体估计

### 点估计 point estimates

### 置信区间 confidence interval estimates

## 样本方差公式 (n-1)

<https://www.zhihu.com/question/20099757>

<https://en.wikipedia.org/wiki/Bias_of_an_estimator#Sample_variance>

# 置信区间

## 置信区间并不是概率区间

比如我们说95%的置信区间是[a,b], 只是说明整体的参数（均值或者方差）有95%的可能性位于a 到b之间，至于是靠近a多一点，还是不多一点，还是 (a+b)/2 多一点，这个不是置信区间解决的问题，置信区间也没告诉我们这个

## 不同的置信级别或者置信水平对应不同的置信区间

<http://liusimin122.blog.sohu.com/149608382.html>

https://zh.wikipedia.org/wiki/%E7%BD%AE%E4%BF%A1%E5%8C%BA%E9%97%B4

## 置信区间依据样本和要求的置信水平来确定

置信区间的基本值如 均值和方差 都是通过样本计算得来的

根据不同的置信水平， 均值 +- 系数\* 方差 来得到置信区间

系统是和置信水平相关的。

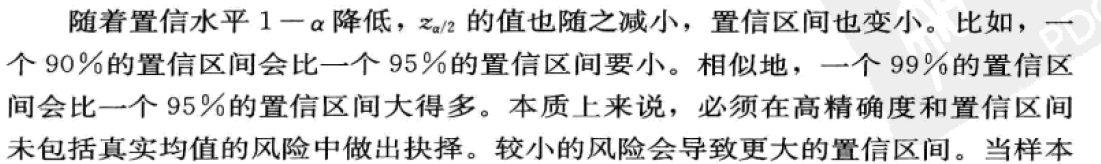
所以，置信区间依据要求的置信水平和样本空间求得，是为了对整体的参数（均值）进行估计。

## 置信水平是个概率

这个概率指的是整体参数位于置信区间内的概率。而不是说整体参数更接近于置信区间的那个值。

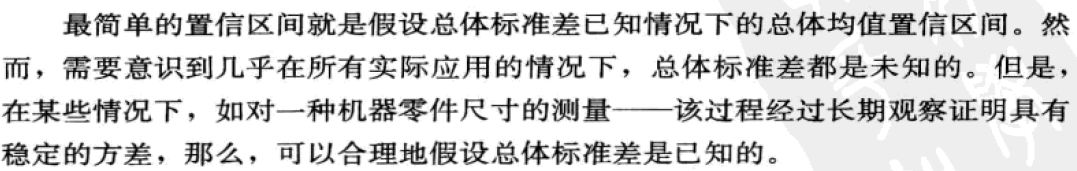
95%是实际运用中最常使用的置信水平数值

本质上来说，必须在高精确度和置信区间未包括真是均值的风险中做出抉择



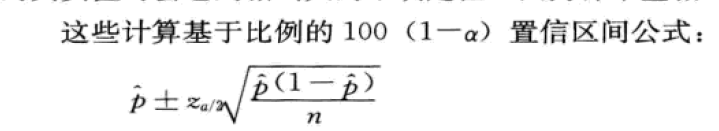
## 计算置信区间的方法

### 总体标准差已知的均值置信区间



### 总体标准差未知的均值置信区间

### 比例的置信区间



### 方差和标准差的置信区间