概率论基础

本章主要内容

1. 随机事件及概率
2. 条件概率、乘法公式
3. 全概率公式、贝叶斯公式
4. 随机变量及其分布函数：离散和连续
5. 常用分布：正态分布、均分分布、二项分布、泊松分布
6. 随机变量的数字特征：期望、方差及标准差
7. 二维随机变量： 联合分布律、协方差及相关系数、矩

本章主要内容

1. 随机事件及概率
2. 条件概率、乘法公式
3. 全概率公式、贝叶斯公式
4. 随机变量及其分布函数：离散和连续
5. 常用分布：正态分布、均分分布、二项分布、泊松分布
6. 随机变量的数字特征：期望、方差及标准差
7. 二维随机变量： 联合分布律、协方差及相关系数、矩

试验和事件

- 例如，在相同的条件下，多次抛掷一个色子（骰子），观察它的点数。
- 抛掷色子这个行为就叫作试验，每次抛掷出现的点数叫作事件。

随机事件

- 如果在相同条件下，每次试验可能出现也可能不出现的事件，叫作随机事件，或者偶然事件。
- 如果每次试验都必然出现，则是必然事件。
- 如果每次试验都不可能出现，则是不可能事件。

例如：

- 事件1：出现3点
- 事件2：出现偶数点
- 事件3：出现的点数小于7
- 事件4：出现的点数大于7

概率

- 虽然，随机事件在一次试验中是否发生不能事先确定，但是在大量重复试验的情况下，它的发生会呈现出一定的规律性。
- 例如，抛硬币这个试验，每一次出现正面或者反面无法确定，但是大量进行这个试验后，发现出现正面或者反面的次数几乎相等。

实验者	总次数	正面朝上次数	概率
德摩根	2048	1061	0.5181
浦　丰	4040	2048	0.5069
K.皮尔逊	12000	6019	0.5016
K.皮尔逊	24000	12012	0.5005

概率

- 在相同条件下，随机试验n次，其中事件A出现了m次($m \le n$)，比值m/n称为事件A发生的频率。
- 随着n的增大，这个频率会趋于一个稳定的值p，即事件A的概率。
- 所以，则事件A发生的概率为：

$$P(A) = \frac{m}{n} = p$$

- 对于日常生活中的很多现象，无法进行大量重复试验，只能按照有限的已知信息计算出一个主观概率。

概率例题

【例】某电子公司所属企业职工人数如下表所示，从该公司随机抽取一人，问：

1、该职工为男性的概率是多少？

2、该职工为手机公司职工的概率是多少？

分公司	男职工	女职工	合计
电脑公司	4400	1800	6200
手机公司	3200	1600	4800
半导体公司	900	600	1500
合计	8500	4000	12500

解：

1、事件A：抽中的职工为男性

$$P(A) = \frac{8500}{12500} = 0.68$$

2、事件B：抽中的职工为手机公司职工

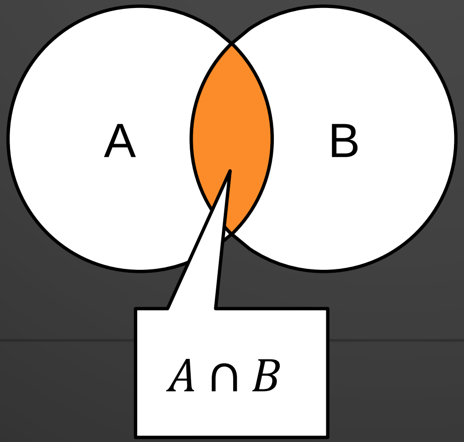
$$P(B) = \frac{4800}{12500} = 0.384$$

概率的基本性质

1. 对任一随机事件，有$0 \leq P(A) \leq 1$
2. 必然事件的概率为1，不可能事件的概率为0，即$P(\emptyset) = 0, P(\Omega) = 1$
3. 如果事件A与B互斥，则$P(A \cup B) = P(A) + P(B)$

   A：抽中的为手机公司职工，B：抽中的为半导体公司职工
4. 对于任意两个随机事件A和B，$P(A \cup B) = P(A) + P(B) - P(A \cap B)$

   特别地，若A与B互斥，则$P(A \cup B) = P(A) + P(B)$



概率的其他性质

1. 设$\bar{A}$是A的对立事件，则$P(A) = 1 - P(\bar{A})$

   A：抽中的为手机公司职工，$\bar{A}$：抽中的为非手机公司职工

2. 若事件B包含事件A，即$A \subset B$，则$P(A) \le P(B)$

3. 若$A \subset B$，则$P(B - A) = P(B) - P(A)$

掷色子，A：出现2点，B：出现偶数点，B-A：出现非2偶数点

$$P(B) = \frac{1}{2}, P(A) = \frac{1}{6}$$

$$P(B - A) = \frac{2}{6} = \frac{1}{3}$$

例题

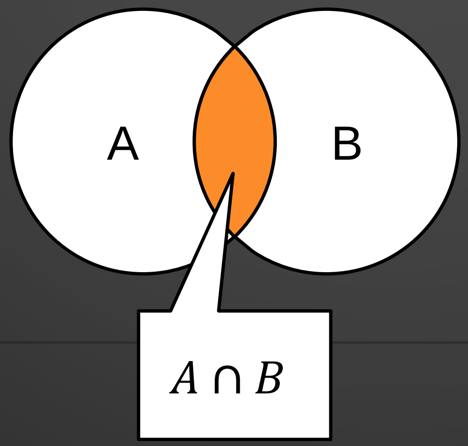
【例】设某地有甲、乙两种报纸，该地成年人中有30%读甲报纸，15%读乙报纸，10%两种报纸都读，成年人中有百分之几至少读一种报纸?

解:

A={读甲报纸}，B={读乙报纸}，C={至少读一种报纸}，

则$P(C) = P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0.3 + 0.15 - 0.1 = 0.35$

即有35%的成年人至少读一种报纸。



本章主要内容

1. 随机事件及概率
2. 条件概率、乘法公式
3. 全概率公式、贝叶斯公式
4. 随机变量及其分布函数：离散和连续
5. 常用分布：正态分布、均分分布、二项分布、泊松分布
6. 随机变量的数字特征：期望、方差及标准差
7. 二维随机变量： 联合分布律、协方差及相关系数、矩

条件概率

当某一事件B已经发生时，事件A发生的概率，称这种概率为事件B发生条件下事件A发生的条件概率（conditional probability），记为P(A|B)。

$$P(A|B) = \frac{P(AB)}{P(B)}, P(B) > 0$$

一个例子：

掷色子：B：出现偶数点，A：出现2点，A|B：已知是偶数点，出现2点

$$P(A|B) = \frac{P(AB)}{P(B)} = \frac{\frac{1}{6}}{\frac{1}{2}} = \frac{1}{3}$$

条件概率：例题

【例】100件产品中，有80件正品，20件次品；而80件正品中有50件一等品，30件二等品。现从这100件产品中任取1件，用A表示“取到一等品”，B表示“取到正品”，求P(A)及P(A|B)。

解：

$$P(A)=\frac{50}{100}=0.5$$

$$P(A|B)=\frac{50}{80}=0.625$$

或者按照条件概率公式，有

$$P(A|B)=\frac{P(AB)}{P(B)}=\frac{50/100}{80/100}=0.625$$

乘法公式

$$P(AB) = P(B)P(A \mid B)$$

将A,B的位置对换，有

$$P(BA) = P(A)P(B \mid A)$$

而

$$P(BA) = P(AB)$$

所以

$$P(AB) = P(A)P(B \mid A)$$

乘法公式：例题

【例】设有1000件产品，其中850件是正品，150件是次品，从中依次抽取2件，2件都是次品的概率是多少?

解:

设A_i={第i次抽到的是次品}（i=1,2），所求概率为$P(A_1A_2)$

$$P(A_1)=\frac{150}{1000}$$

$$P(A_2|A_1)=\frac{149}{999}$$

根据概率的乘法公式，有

$$P(A_1A_2)=P(A_1)P(A_2|A_1)=\frac{150}{1000}\times\frac{149}{999}=0.0224$$

乘法公式与独立事件

前面已得到，

$$P(AB) = P(A)P(B|A)$$

特别地，当两个事件相互独立时，其乘法公式则可以简化为：

$$P(AB) = P(A)P(B)$$

因为$P(B \mid A) = P(B)$（独立）

本章主要内容

1. 随机事件及概率
2. 条件概率、乘法公式
3. 全概率公式、贝叶斯公式
4. 随机变量及其分布函数：离散和连续
5. 常用分布：正态分布、均分分布、二项分布、泊松分布
6. 随机变量的数字特征：期望、方差及标准差
7. 二维随机变量： 联合分布律、协方差及相关系数、矩

全概率公式：划分

例如，掷色子这个试验记为E，S为其样本空间，即

- A_1：出现1点、A_2：出现2点、A_3：出现3点、A_4：出现4点、A_5：出现5点、A_6：出现6点
- 其中，$A_iA_j=\emptyset$，$A_1\cup A_2\cup A_3\cup A_4\cup A_5\cup A_6=S$

则称A_1,A_2,A_3,A_4,A_5,A_6为样本空间S的一个划分。

定义：设S为试验E的样本空间，$B_1,B_2,\cdots,B_n$为E的一组事件，若

1. $B_iB_j=\emptyset, i\neq j, i,j=1,2,\cdots,n$
2. $B_1\cup B_2\cup\cdots\cup B_n=S$

则称$B_1,B_2,\cdots,B_n$为样本空间S的一个划分。

全概率公式

定理：设试验E的样本空间为S，A为E的事件，$B_1, B_2, \ldots, B_n$为S的一个划分，且$P(B_i) > 0(i = 1, 2, \ldots, n)$，则

$$P(A) = P(A \mid B_1)P(B_1) + P(A \mid B_2)P(B_2) + \cdots + P(A \mid B_n)P(B_n)$$

上式称为全概率公式。

用求和符号，可以将全概率公式表示为：

$$P(A) = \sum_{i=1}^{n} P(A \mid B_i)P(B_i)$$

贝叶斯公式

定理：设试验E的样本空间为S。A为E的事件，$B_1, B_2, \ldots, B_n$为S的一个划分，且

$P(A) > 0, P(B_i) > 0(i = 1, 2, \ldots, n)$，则

$$P(B_i \mid A) = \frac{P(A|B_i)P(B_i)}{\sum_{j=1}^{n} P(A|B_j)P(B_j)}, i = 1, 2, \ldots, n$$

以上公式称为贝叶斯公式。

证明：

$$P(B_i|A) = \frac{P(B_iA)}{P(A)} = \frac{P(A|B_i)P(B_i)}{\sum_{i=1}^{n} P(A \mid B_i)P(B_i)}$$

贝叶斯公式：例题

【例】对以往数据分析结果表明，当机器调整得良好时，产品的合格率为98%，而当机器发生某种故障时，其合格率为55%。每天早上机器开动时，机器调整良好的概率为95%。试求已知某日早上第一件产品是合格品时，机器调整良好的概率是多少?

解：事件A：产品合格，事件B：机器调整良好

根据已知，有:

- 当机器调整得良好时，产品的合格率为98%，这是一个条件概率：$P(A|B) = 0.98$
- 当机器发生某种故障时，其合格率为55%，这也是一个条件概率：$P(A|\bar{B}) = 0.55$
- 每天早上机器开动时，机器调整良好的概率为95%：$P(B) = 0.95$
- 每天早上机器开动时，机器调整不良好的概率为5%：$P(\bar{B}) = 0.05$

现在要求的是：

已知某日早上第一件产品是合格品时，机器调整良好的概率是多少？相当于求$P(B|A)$

根据贝叶斯公式，有

$$
\begin{aligned}
P(B|A) &= \frac{P(A|B)P(B)}{P(A|B)P(B)+P(A|\overline{B})P(\overline{B})} \\
&= \frac{0.98\times 0.95}{0.98\times 0.95+0.55\times 0.05} \\
&= 0.97
\end{aligned}
$$

在上面这例子中，涉及到以下两个概念：

先验概率：由以往的数据分析得到

后验概率：在得到信息之后再重新修正的概率

本章主要内容

1. 随机事件及概率
2. 条件概率、乘法公式
3. 全概率公式、贝叶斯公式
4. 随机变量及其分布函数：离散和连续
5. 常用分布：正态分布、均分分布、二项分布、泊松分布
6. 随机变量的数字特征：期望、方差及标准差
7. 二维随机变量： 联合分布律、协方差及相关系数、矩

随机变量及分布函数

以掷色子为例：

- 事件A：点数为2
- 事件A的概率记为$P(A)$，也可以记为$P(X=2)$

由于X可以取不同的值，$X=1,2,3,4,5,6$，称X为随机变量

分布函数：对于任意实数x，记函数$F(x)=P\{X\le x\}$，$F(x)$为随机变量X的分布函数。

$F(x)$的值等于随机变量X在区间$(-\infty,x]$内取值的概率，即事件"$X\le x$"的概率。

按照随机变量的特性，可以把随机变量分为两类，即离散型随机变量和连续型随机变量。

离散型随机变量及其分布

- 随机变量X的所有取值都可以逐个列举出来，则称X为离散型随机变量。
- 例如，掷色子出现的点数、一批产品抽样抽取到的次品的个数。

均匀分布

X	1	2	3	4	5	6
P(x)	1/6	1/6	1/6	1/6	1/6	1/6

离散型随机变量X的概率分布(probability distribution)

离散型随机变量及其分布：0-1分布

已知一批产品的废品率为$p = 0.05$，合格率为$q = 1 - p = 0.95$。指定废品用1代表，合格品用0代表，考察任抽取一件废品或合格品，即1或0，这一离散型随机变量的概率分布如下表。

X	1	0
P(x)	0.05	0.95

0-1分布可以写成如下形式：

X	1	0
P(x)	p	q

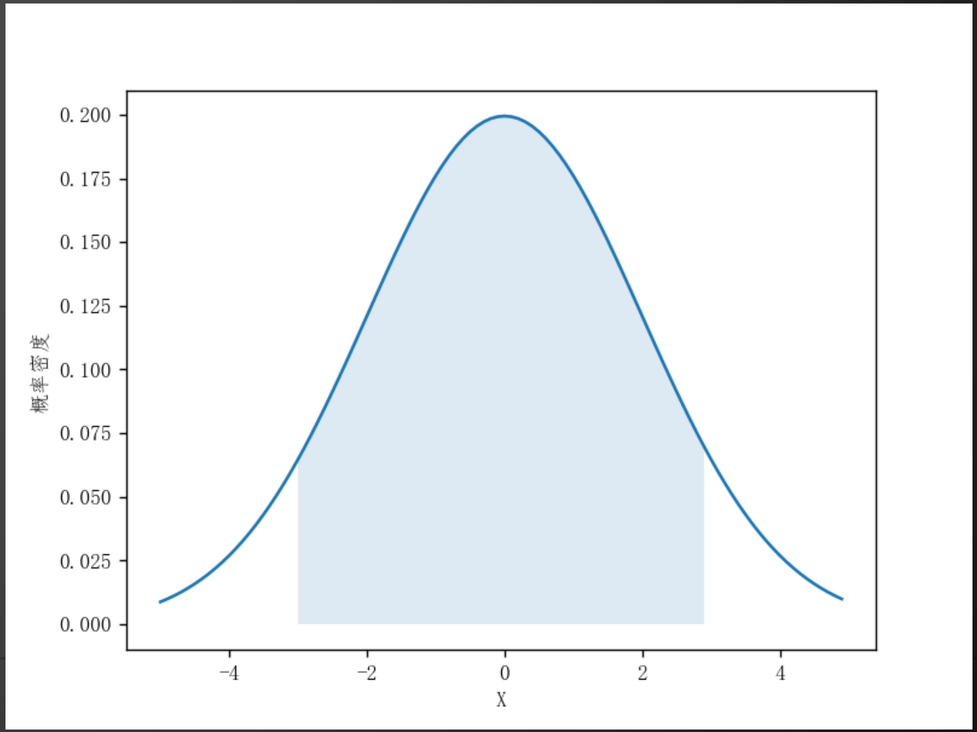
连续型随机变量及其分布

- 连续型随机变量可以取某一区间或整个实数轴上的任意一个值，必须用数学函数的形式来描述。例如，一批产品的寿命、某公司的年销售量。
- 对于连续型随机变量，用函数来表示其分布，有两个函数：概率密度函数$f(x)$和分布函数$F(x)$。
- 概率密度函数$f(x)$满足以下两个条件：

1. $f(x) \geq 0$
2. $\int_{-\infty}^{+\infty} f(x)dx = 1$

- 分布函数为：$F(x) = P(a \leq X \leq b) = \int_a^b f(t)dt$

说明：分布函数表示连续型随机变量的概率。



本章主要内容

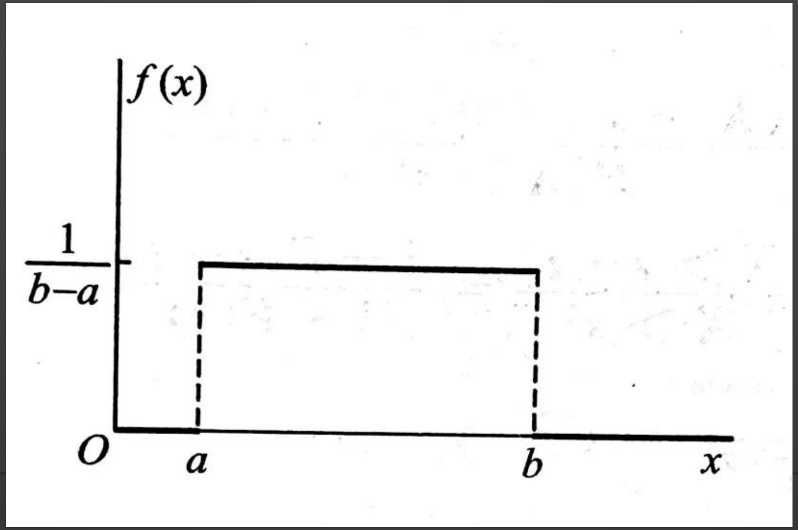
1. 随机事件及概率
2. 条件概率、乘法公式
3. 全概率公式、贝叶斯公式
4. 随机变量及其分布函数：离散和连续
5. 常用分布：均分分布、正态分布、二项分布、泊松分布
6. 随机变量的数字特征：期望、方差及标准差
7. 二维随机变量： 联合分布律、协方差及相关系数、矩

均匀分布

连续型随机变量X 服从均匀分布，其概率密度为：

$$f(x) = \begin{cases} \dfrac{1}{b-a}, a < x < b \\ 0, \text{其他} \end{cases}$$

记为$X \sim U(a, b)$



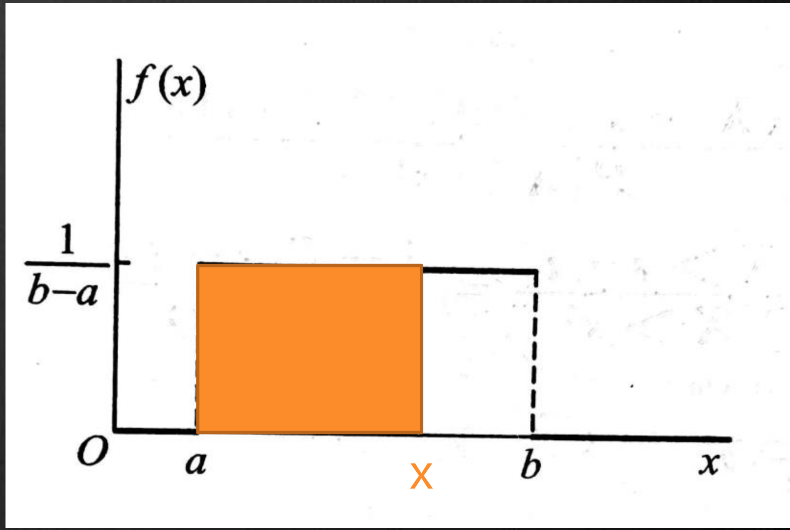
均匀分布

解：根据分布函数的定义，有

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)dt = \int_{a}^{x} \frac{1}{b-a} dt = \frac{x-a}{b-a}$$

所以，

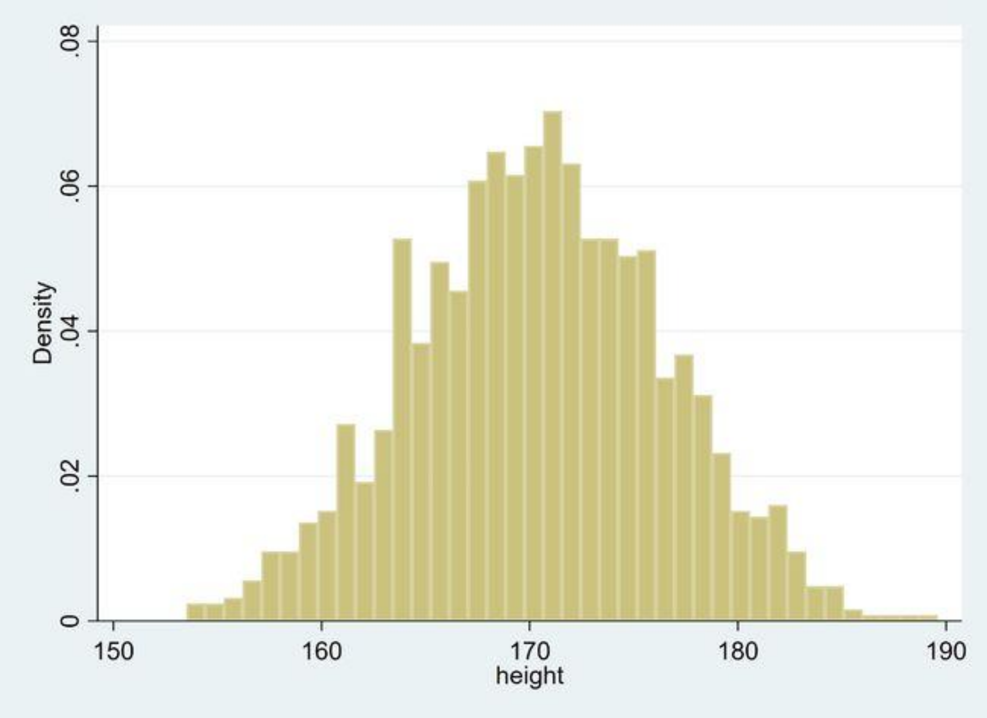
$$F(x) = \begin{cases} 0, & x \leq a \\ \frac{x-a}{b-a}, & a < x < b \\ 1, & b \leq x \end{cases}$$

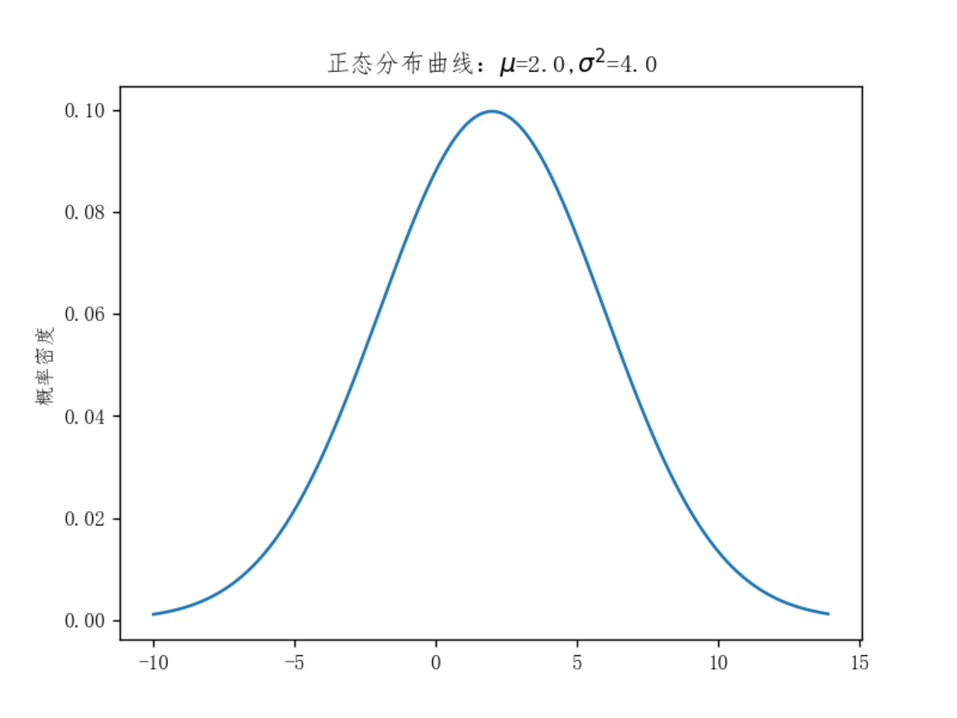


正态分布

正态分布，英文为：Normal distribution，即常态分布、典型分布

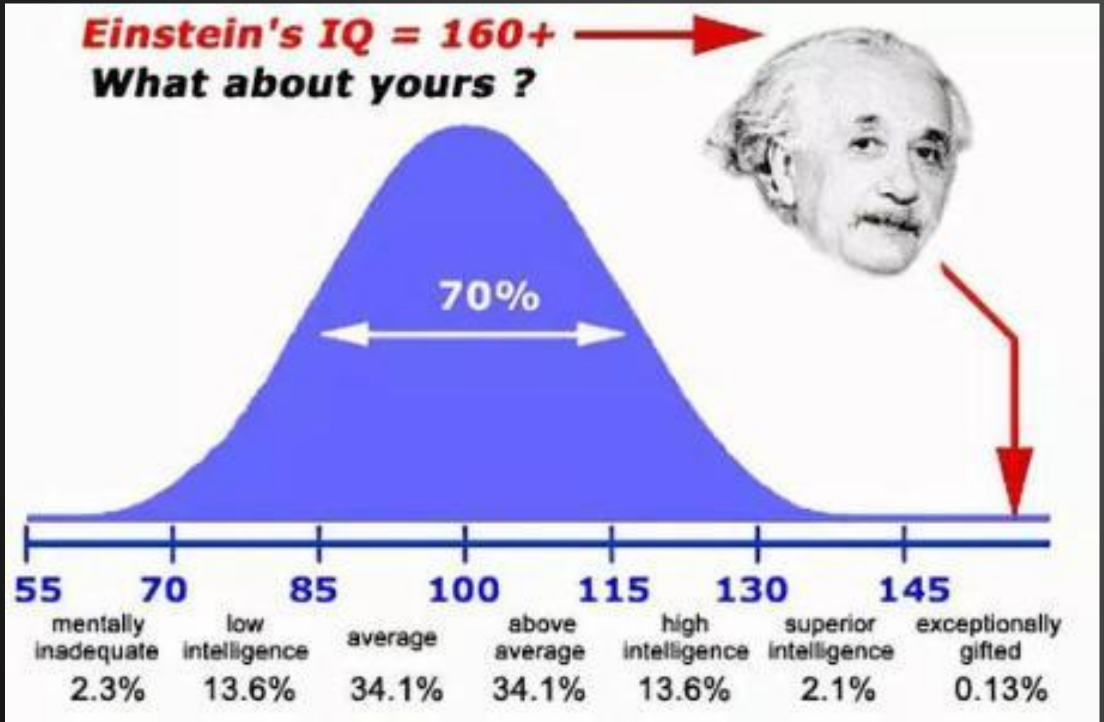
生活中很多现象都服从正态分布，例如身高，大部分人的身份都是不高也不矮，特别高的非常少，特别矮的也不多见。

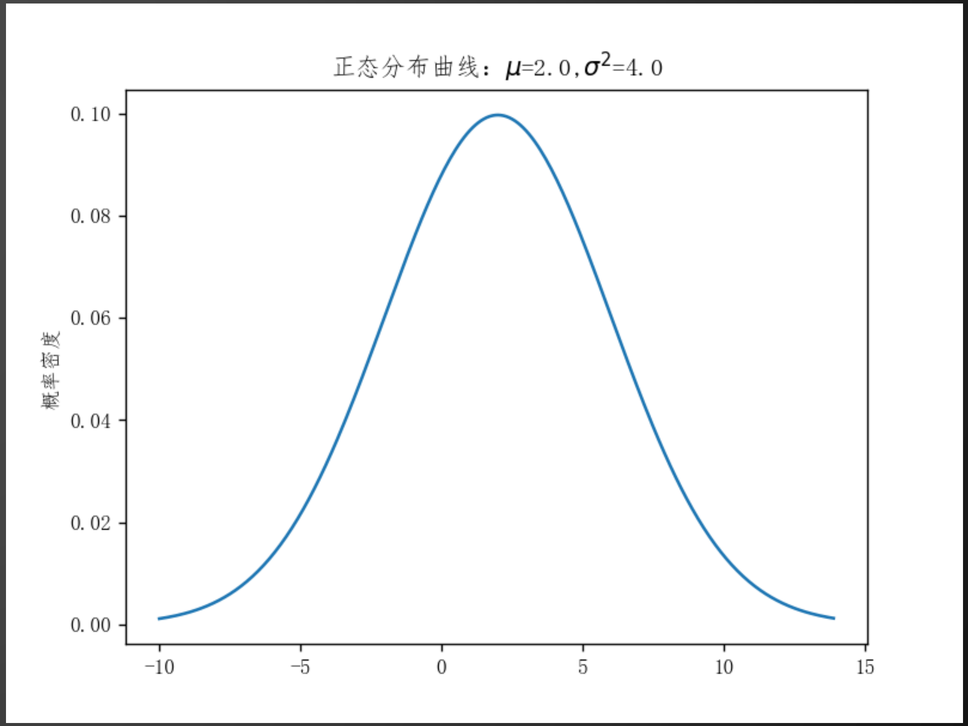




正态分布

- 智商：大部分人的智商都是正常的，特别聪明的人非常少，但是特别笨的人也不多见。





正态分布

生活中还有很多其他现象也都符合正态分布。

- 业绩：大部分员工的业绩，都是一般的，做得特别好的非常少，做得特别差的也不多见。
- 考试成绩：大部分学生的考试成绩都是正常，特别高的非常少，特别低的也不多见。

由于高斯深入研究了正态分布的性质，所以正态分布也叫高斯分布（Gaussian distribution）

- 定义：X是连续型随机变量，则X的概率密度函数

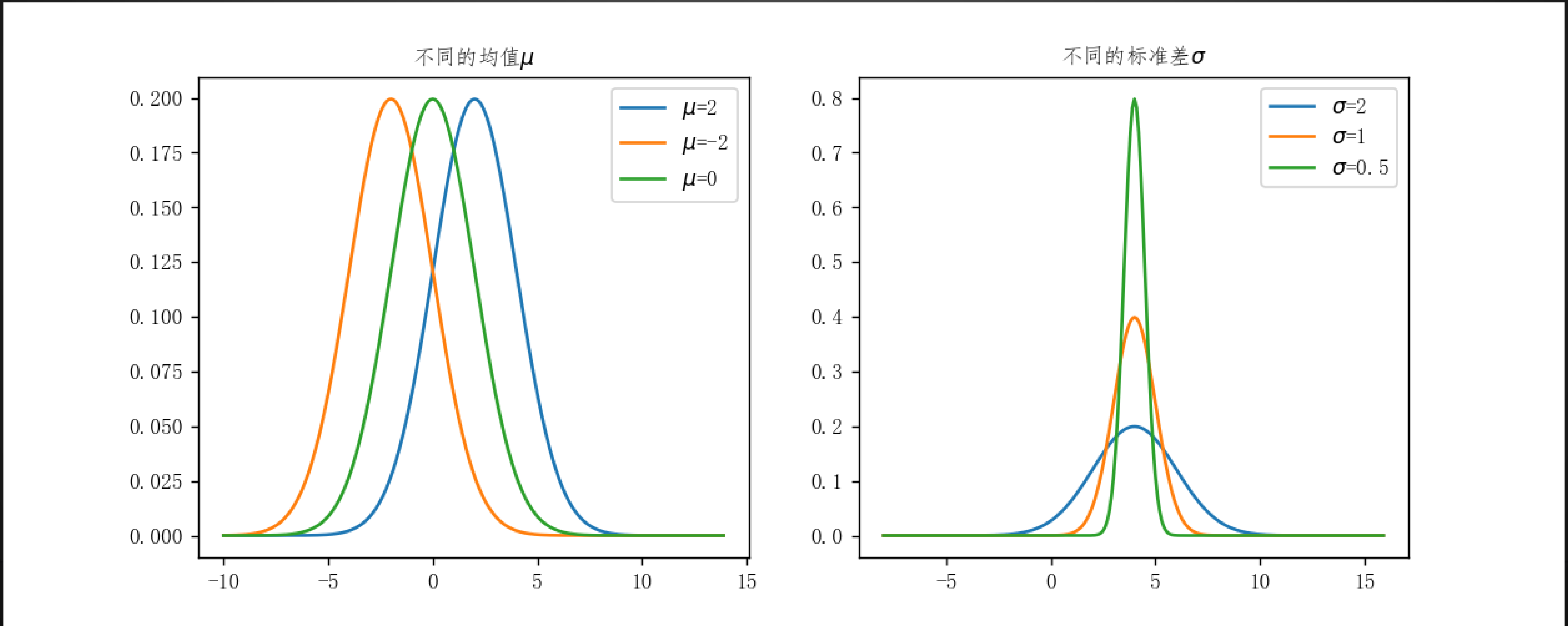
$$f(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}},-\infty<x<\infty$$

记为$X\sim N(\mu,\sigma^2)$，称X服从正态分布。

- 其中，μ为随机变量X的均值，σ为随机变量X的标准差，是正态分布中两个重要的参数。

正态分布

- 下图说明了参数μ, σ对曲线位置及形状的影响。

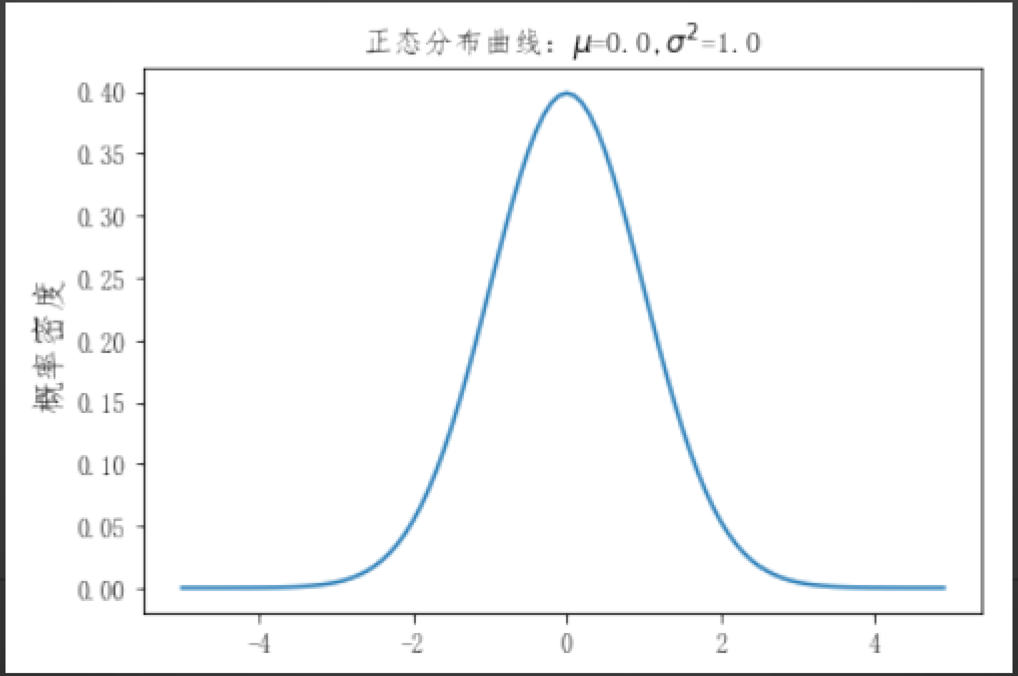


标准正态分布

当$\mu = 0, \sigma = 1$时，有

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, -\infty < x < \infty$$

此时，正态分布称为标准正态分布(standard normal distribution)，记作$X \sim N(0,1)$。



正态分布

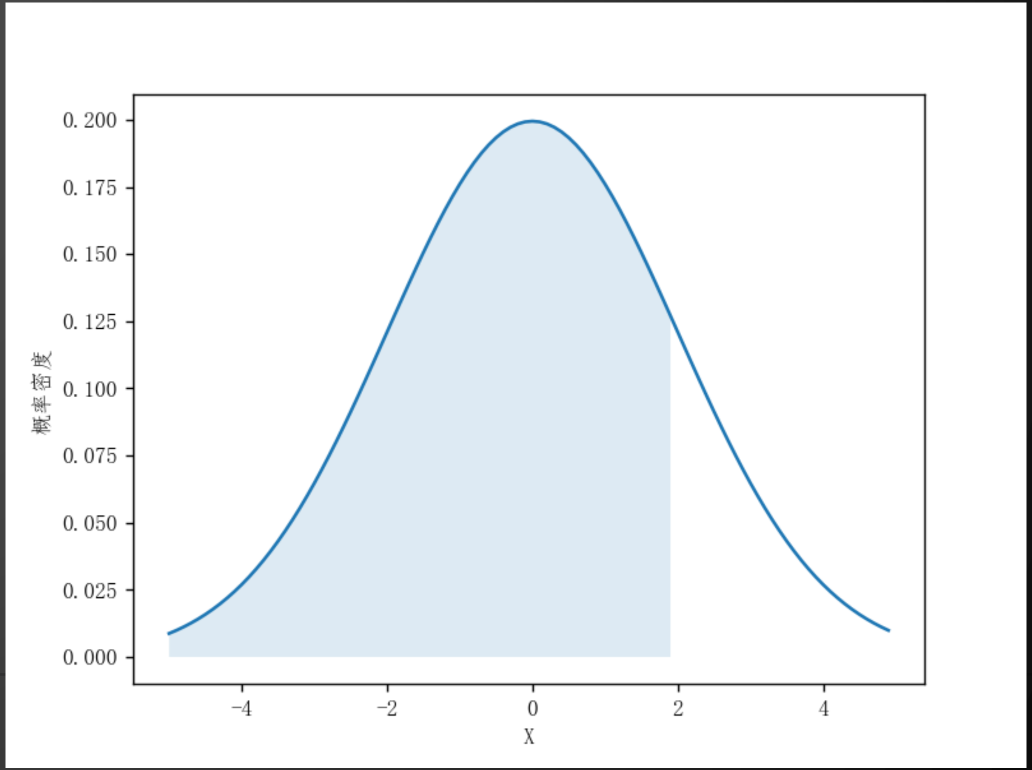
- 对于标准正态分布来说，通常用$\varphi(x)$表示概率密度函数，即

$$\varphi(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$$

- 用$\Phi(x)$表示分布函数，即

$$\Phi(x)=\int_{-\infty}^{x}\varphi(t)dt=\int_{-\infty}^{x}\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}dt=P(X\le x)$$

- 对于标准正态分布，给定x的范围，通过分布函数$\Phi(x)$就可以计算出概率值，一般通过查表可以轻松获得对应的概率值。
- 对于正态分布，先将其转化为标准正态分布，再查表。



正态分布：例题

【例】设$X \sim N(0,1)$，求以下概率：

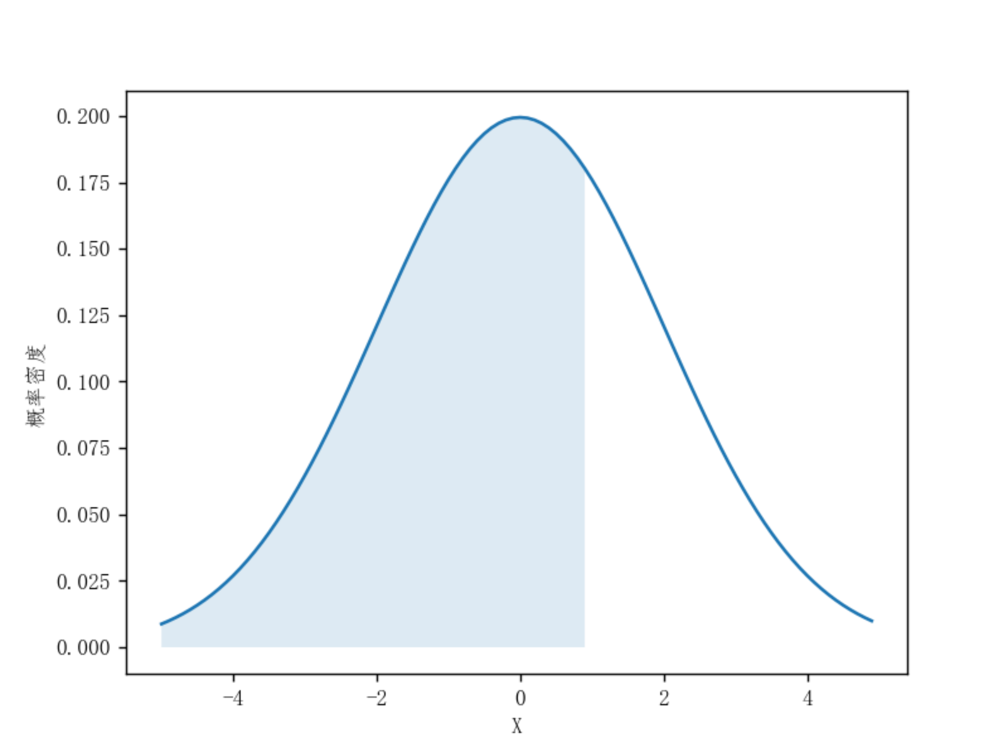
标准正态分布概率密度函数：$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}$

1、$P(X < 1.5)$

2、 $P(X > 2)$

3、 $P(-1 < X \leq 3)$

4、 $P(|X| \leq 2)$



解：

1、 $P(X < 1.5) = \int_{-\infty}^{1.5} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt = \Phi(1.5) = 0.933$

计算方法：查表、Excel公式

正态分布：例题

【例】设$X \sim N(0,1)$，求以下概率：

1、$P(X < 1.5)$

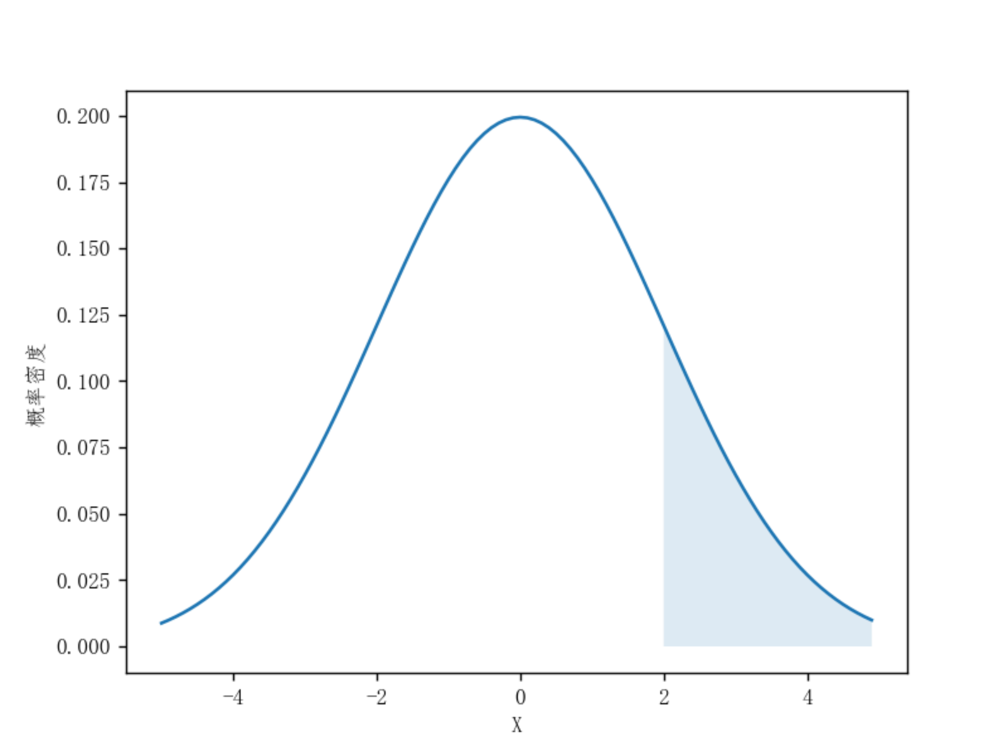
2、 $P(X > 2)$

3、 $P(-1 < X \leq 3)$

4、 $P(|X| \leq 2)$

解：

2、 $P(X > 2) = 1 - P(X \leq 2) = 1 - \Phi(2)$
$= 1 - 0.977 = 0.023$



正态分布：例题

【例】设$X \sim N(0,1)$，求以下概率：

1、$P(X < 1.5)$

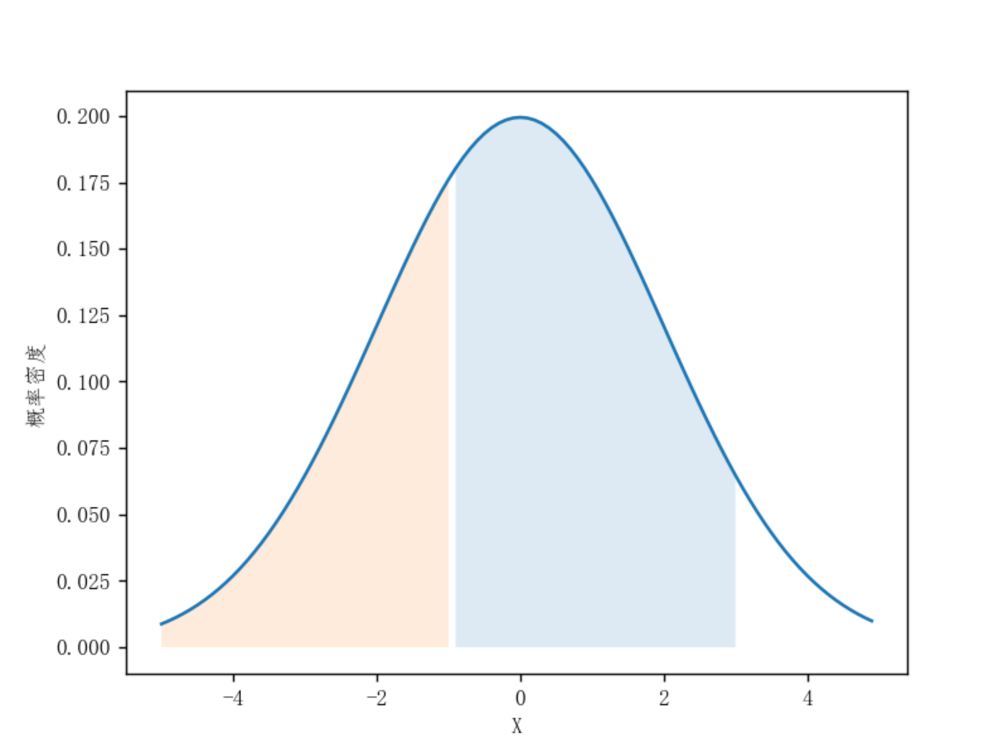
2、 $P(X > 2)$

3、 $P(-1 < X \leq 3)$

4、 $P(|X| \leq 2)$

解：

3、$P(-1 < X \leq 3) = P(X \leq 3) - P(X \leq -1)$
$= \Phi(3) - \Phi(-1) = \Phi(3) - [1 - \Phi(1)]$
$= 0.9987 - (1 - 0.8413) = 0.84$



正态分布：例题

【例】设$X \sim N(0,1)$，求以下概率：

1、$P(X < 1.5)$

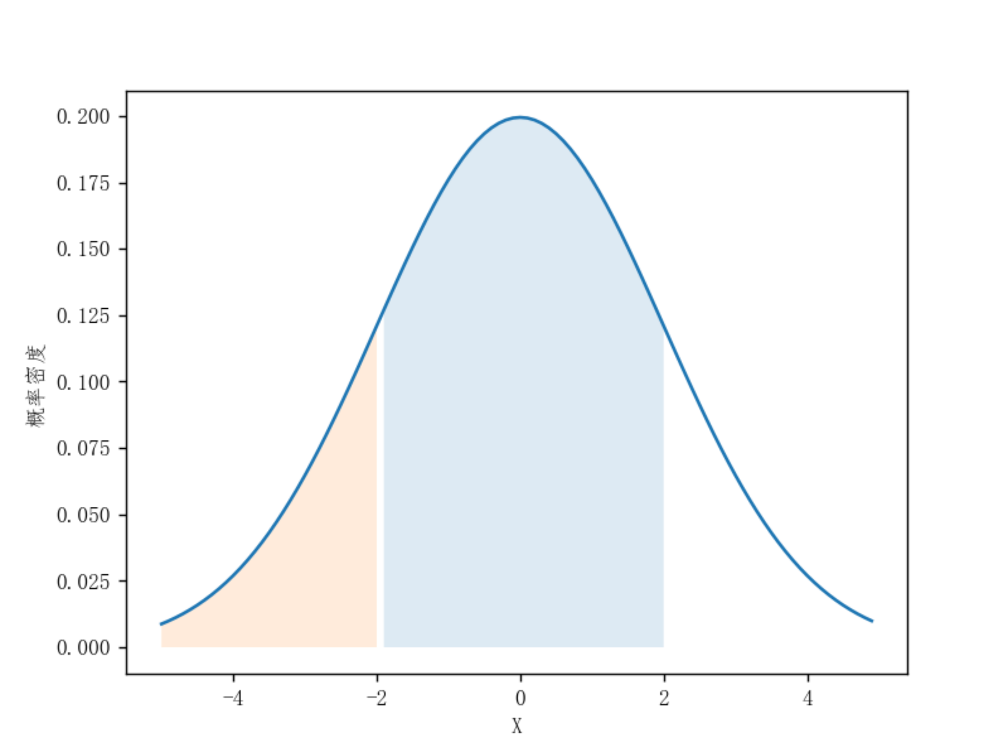
2、 $P(X > 2)$

3、 $P(-1 < X \leq 3)$

4、 $P(|X| \leq 2)$

解：

4、$P(|X| \leq 2) = P(-2 \leq X \leq 2)$
$= P(X \leq 2) - P(X \leq -2) = \Phi(2) - \Phi(-2)$
$= \Phi(2) - [1 - \Phi(2)] = 2\Phi(2) - 1$
$= 2 \times 0.9772 - 1 = 0.95$



正态分布：例题

【例】设$X \sim N(5,3^2)$，求以下概率：

1. $P(X \leq 10)$
2. $P(2 < X < 10)$

解：

已知：$X \sim N(5,3^2)$，则$\frac{X-5}{3} \sim N(0,1)$

1. $P(X \leq 10)$等价于$P(X \leq 1.67) = \Phi(1.67) = 0.95$
2. $P(2 < X < 10)$等价于$P(-1 < X < 1.67) = \Phi(1.67) - \Phi(-1) = 0.95 - [1 - \Phi(1)] = 0.79$

二项分布

- 实际问题中，有许多现象与抛硬币这个试验相似，只包含两个结果。例如：市场调查中考虑消费者对某个产品的喜好，调查一个人对该产品的喜好相当于抛一次硬币，这个人喜欢该产品相当于出现正面，不喜欢该产品相当于出现反面，则在调查中的样本中喜欢该产品的人数是一个随机变量，这种随机变量所服从的概率分布称为二项分布（binomial distribution）。
- 二项分布的特点如下：

1. 包含了n个相同的试验，且相互独立，即在相同条件下，进行n次抛硬币。
2. 每次试验只有两个可能的结果："成功"或者"失败"，即每次抛硬币，出现正面或者反面。
3. 每次实验，"成功"的概率都为p，"失败"的概率都为q，且$p+q=1$。
4. n次试验，"成功"或者"失败"的次数是一个离散型随机变量。

二项分布

随机变量X服从二项分布，记作$X \sim B(n, p)$，若进行n次独立重复试验，事件A出现的次数X是一个随机变量，则X的概率为：

$$P\{X = x\} = C_n^x p^x q^{n-x}, x = 0,1,2,\cdots,n$$

- 说明：

1. p表示每次试验事件A出现的概率，q表示不出现的概率，且$p + q = 1$
2. $P\{X = x\} \geq 0, x = 0,1,2,\cdots,n$
3. $\sum_{x=0}^{n} C_n^x p^x q^{n-x} = (p + q)^n = 1$

二项分布：例题

【例】已知100件产品中有5件次品，现从中任取1件，有放回地取3次，求在所取的3件中恰有2件次品的概率。

解：

根据题意，每次试验取到次品的概率为：$p=\frac{5}{100}=0.05$

设X为所取的3件中的次品数，则$X\sim B(3,0.05)$

所以，$P\{X=2\}=C_3^2(0.05)^2(0.95)^1=0.007125$

泊松分布

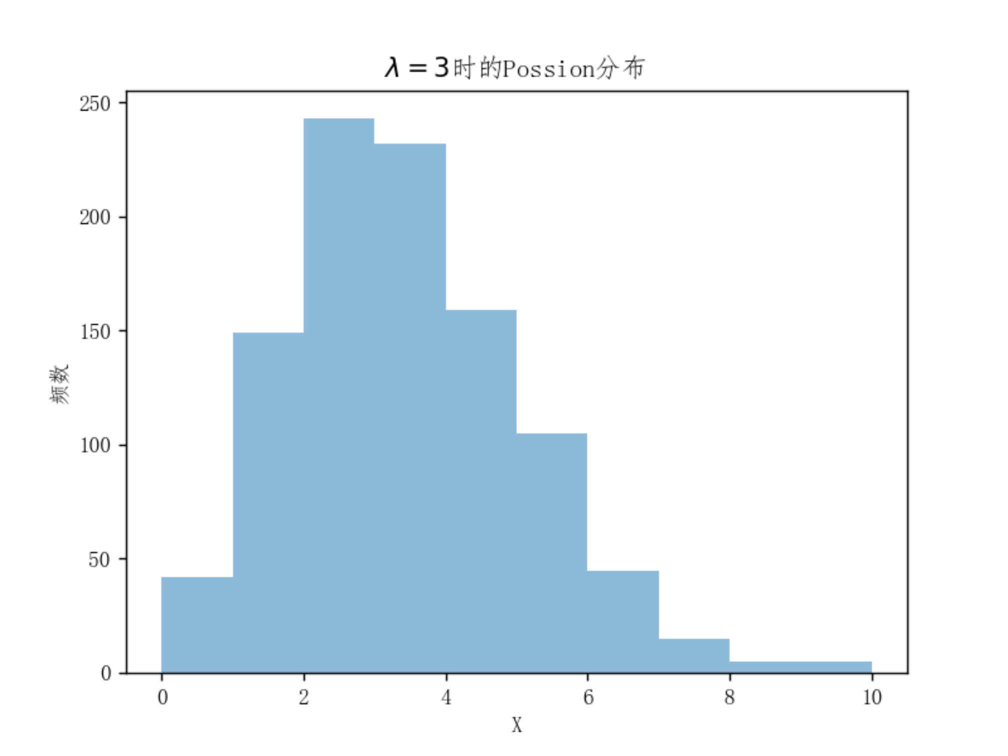
泊松分布（Poisson distribution）是用来描述在一定时间范围内或在指定的面积或体积内某一事件出现的次数的分布。例如：

1. 某种仪器每月出现故障的次数
2. 人寿保险公司每天收到的死亡声明的个数
3. 一本书一页中的印刷错误数
4. 某地区一天内遗失的信件数

泊松分布的概率分布函数为：

$$P(X=k)=\frac{\lambda^k e^{-\lambda}}{k!}, k=0,1,2,\cdots$$

其中，λ为给定的时间间隔内事件的平均数。



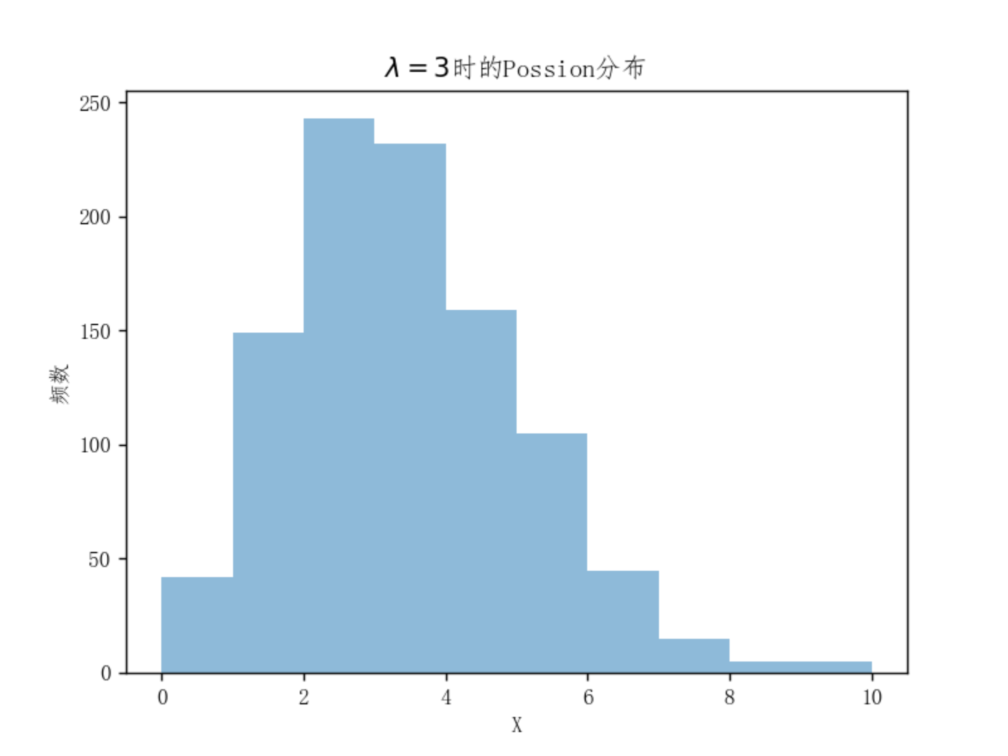
泊松分布

【例】假定某企业职工在周一请事假的人数X近似服从泊松分布，且设周一请事假的平均数为2.5人。要求计算：在给定的某周一正好请事假是5人的概率。

解：

$$P(X=5)=\frac{\lambda^5 e^{-\lambda}}{5!}=\frac{2.5^5 e^{-2.5}}{5!}=0.067$$

$$P(X=3)=\frac{\lambda^3 e^{-\lambda}}{3!}=\frac{2.5^3 e^{-2.5}}{3!}=0.21$$



常用分布总结

连续型随机变量

- 均匀分布（连续型）
- 正态分布

离散型随机变量

- 二项分布
- 泊松分布
- 均匀分布（离散型）
- 0-1分布

本章主要内容

1. 随机事件及概率
2. 条件概率、乘法公式
3. 全概率公式、贝叶斯公式
4. 随机变量及其分布函数：离散和连续
5. 常用分布：正态分布、均分分布、二项分布、泊松分布
6. 随机变量的数字特征：期望、方差及标准差
7. 二维随机变量： 联合分布律、协方差及相关系数、矩

期望

期望是概率论中的一个概念，又称为均值。

- 离散：$E(X) = x_1 p_1 + x_2 p_2 + \cdots + x_n p_n = \sum_{i=1}^{n} x_i\, p_i$
- 连续：$E(X) = \int_{-\infty}^{+\infty} x f(x) dx = \mu$

例如，掷色子出现的点数X这个随机变量，其概率分布为：

X	1	2	3	4	5	6
P(x)	1/6	1/6	1/6	1/6	1/6	1/6

$$E(X) = \sum_{i=1}^{6} x_i\, p_i = 1 \times \frac{1}{6} + 2 \times \frac{1}{6} + 3 \times \frac{1}{6} + 4 \times \frac{1}{6} + 5 \times \frac{1}{6} + 6 \times \frac{1}{6} = 3.5$$

说明：对于均匀分布来说，期望其实就是求随机变量取值的均值。

方差

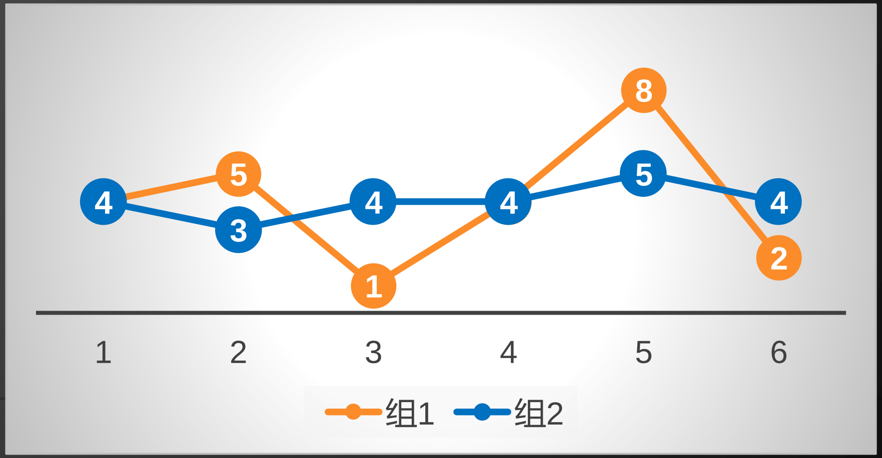
方差用于描述随机变量的波动性，也就是所谓的离散程度。

方差用σ^2或$D(X)$表示，有$\sigma^2 = D(X) = E[X - E(X)]^2$

- 离散：$D(X) = \sum_{i=1}^{\infty}[x_i - E(X)]^2 p_i$，其中，$p_i = P\{X = x_i\}(i = 1,2,\cdots)$
- 连续：$D(X) = \int_{-\infty}^{+\infty}[x - E(x)]^2 f(x)dx = \sigma^2$

例如，随机变量X分别取以下两组数据，试比较这两数据的离散程度。

随机变量X	x_1	x_2	x_3	x_4	x_5	x_6	均值
组1	4	5	1	4	8	2	4
组2	4	3	4	4	5	4	4
$P(X)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	



方差

随机变量X	x_1	x_2	x_3	x_4	x_5	x_6	均值
组1	4	5	1	4	8	2	4
组2	4	3	4	4	5	4	4
$P(X)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	

- 离散：$D(X)=\sum_{i=1}^{\infty}[x_i-E(X)]^2p_i$
- 组1：$D(X)=(4-4)^2\times\frac{1}{6}+(5-4)^2\times\frac{1}{6}+(1-4)^2\times\frac{1}{6}+(4-4)^2\times\frac{1}{6}+(8-4)^2\times\frac{1}{6}+(2-4)^2\times\frac{1}{6}=5$
- 组2：$D(X)=(4-4)^2\times\frac{1}{6}+(3-4)^2\times\frac{1}{6}+(4-4)^2\times\frac{1}{6}+(4-4)^2\times\frac{1}{6}+(5-4)^2\times\frac{1}{6}+(4-4)^2\times\frac{1}{6}=0.33$
- 结论：数据组1的离散程度要大于数据组2。

方差及标准差

- 方差： $D(X) = E[X - E(X)]^2$
- 为了方便计算，方差的简化形式：$D(X) = E(X^2) - [E(X)]^2$
- 推导： $D(X) = E[X - E(X)]^2 = E[X^2 - 2XE(X) + [E(X)]^2] = E(X^2) - 2E(X)E(X) + [E(X)]^2$

- 标准差：方差的平方根，记为$\sigma = \sqrt{D(X)}$
- 由于标准差与随机变量X有相同的度量单位，所以在实际中经常使用。
- 说明：在Excel中，通过公式可以很容易计算出期望（均值）、方差及标准差。

泊松分布的期望和方差

$$P(X=k)=\frac{\lambda^k e^{-\lambda}}{k!}, k=0,1,2,\cdots$$

泊松分布的期望和方差分别为：

$E(X)=\lambda,\ D(X)=\lambda$

证明：

$$E(X)=\sum_{k=0}^{\infty}k\frac{\lambda^k e^{(-\lambda)}}{k!}=\lambda e^{-\lambda}\sum_{k=0}^{\infty}\frac{\lambda^{k-1}}{(k-1)!}=\lambda e^{-\lambda}.e^{\lambda}=\lambda$$

$$D(X)=E(X^2)-[E(X)]^2=\lambda$$

$$E(X^2)=E[X(X-1)+X]=E[X(X-1)]+E(X)$$

$$=\sum_{k=0}^{\infty}k(k-1)\frac{\lambda^k e^{-\lambda}}{k!}+\lambda=\lambda^2 e^{-\lambda}\sum_{k=0}^{\infty}\frac{\lambda^{k-2}}{(k-2)!}+\lambda=\lambda^2 e^{-\lambda}.e^{\lambda}+\lambda=\lambda^2+\lambda$$

本章主要内容

1. 随机事件及概率
2. 条件概率、乘法公式
3. 全概率公式、贝叶斯公式
4. 随机变量及其分布函数：离散和连续
5. 常用分布：正态分布、均分分布、二项分布、泊松分布
6. 随机变量的数字特征：期望、方差及标准差
7. 二维随机变量： 联合分布律、协方差及相关系数、矩

二维随机变量及其分布

- 之前说的随机变量只有一个变量，例如掷色子出现的点数随机变量X有6种不同的值，$X = 1,2,3,4,5,6$。又比如，抛硬币，一个硬币，结果有正面和反面两种情况，随机变量X有两种不同的取值。
- 设想，如果每次试验抛掷两枚硬币，第一枚硬币代表的随机变量记为X，第二枚硬币代表的随机变量记为Y，则(X,Y)构成一个二维随机变量。
- 与一维随机变量类似，二维随机变量也可以分为离散型和连续型。

联合分布律

- 例如，每次试验抛掷两枚硬币，第一枚硬币代表的随机变量记为X，第二枚硬币代表的随机变量记为Y，可以通过表格来描述X和Y的概率分布，这种分布叫作联合分布律。

Y \ X	0	1
0	$\frac{1}{4}$	$\frac{1}{4}$
1	$\frac{1}{4}$	$\frac{1}{4}$

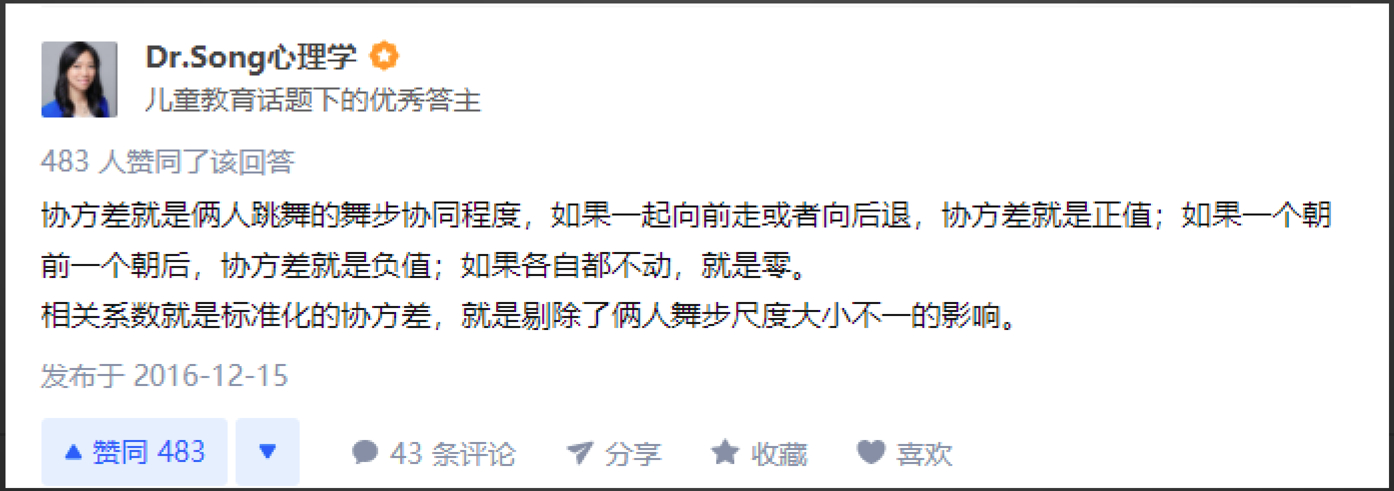
- 说明：这里以0代表正面，1代表反面

协方差及相关系数

协方差就是两人跳舞的舞步协同程度：

- 如果一起向前走或者向后退，协方差为正；
- 如果一个朝前，一个朝后，协方差为负；
- 如果都不动，协方差等于零。

相关系数就是标准后的协方差，即剔除了两人舞步尺度大小不一的影响。


Dr.Song心理学
儿童教育话题下的优秀答主

483 人赞同了该回答

协方差就是俩人跳舞的舞步协同程度，如果一起向前走或者向后退，协方差就是正值；如果一个朝前一个朝后，协方差就是负值；如果各自都不动，就是零。
相关系数就是标准化的协方差，就是剔除了俩人舞步尺度大小不一的影响。

发布于 2016-12-15

▲ 赞同 483 ▼ 43 条评论 分享 收藏 喜欢

协方差及相关系数

对于二维随机变量(X,Y)，除了讨论X与Y的期望和方差外，还需要讨论X与Y之间的相关关系。

X与Y之间的相关关系用协方差来描述。

定义：随机变量(X,Y)的协方差记为$Cov(X,Y)$

$$Cov(X,Y) = E\{[X - E(X)][Y - E(Y)]\}$$

将$Cov(X,Y)$展开，可得

$$\begin{aligned} Cov(X,Y) &= E\{XY - XE(Y) - YE(X) + E(X)E(Y)\} \\ &= E(XY) - E(X)E(Y) - E(Y)E(X) + E(X)E(Y) \\ &= E(XY) - E(X)E(Y) \end{aligned}$$

当随机变量X和Y相互独立时，有$Cov(X,Y) = 0$

协方差及相关系数

相关系数定义如下：

$$\rho_{XY} = \frac{Cov(X,Y)}{\sqrt{D(X)}\sqrt{D(Y)}}$$

$$\rho_{XY} = \frac{E\{[X - E(X)][Y - E(Y)]\}}{\sqrt{D(X)}\sqrt{D(Y)}}$$

从上面可以看出：

- 相关系数其实就是剔除了这两个随机变量变化幅度的影响。
- 所以，相关系数其实就是将随机向量X和Y标准化后的协方差。

矩

定义：设X是随机变量，如果$E(X^k), k = 1, 2, \ldots$存在，则称之为X的k阶原点矩。

说明：期望是X的1阶原点矩。

定义：设X是随机变量，如果$E[X - E(X)]^k, k = 1, 2, \ldots$存在，则称之为$X$的$k$阶中心矩。

说明：方差是X的2阶中心矩。

定义：设X和Y是两个随机变量，如果$E(X^kY^l), k, l = 1, 2, \ldots$存在，则称之为$X$和$Y$的$k + l$阶混合矩。

定义：设X和Y是两个随机变量，如果$E[X - E(X)]^k[Y - E(Y)]^l, k, l = 1, 2, \ldots$存在，则称之为$X$和$Y$的$k + l$阶混合中心矩。

说明：协方差是X和Y的2阶混合中心矩。

说明：矩是一个高度抽象的概念，期望、方差、协方差等都是特殊的矩。