

# 数据科学基础知识 Fundamentals

## 第一部分:概率论 Probability theory

ISEP第二年  
2023-2024

基于 Nathalie Colin 和 Jean-Claude Guillerot 教授的课程

# 概率论

---

第二次会议（2023 年 10 月 6 日）：

第3章:实值随机变量

- 3.1 随机变量的定义 – 3.2 累积分布函数 (CDF)
- 3.3 概率密度函数 – 3.4 连续随机变量 – 3.5 离散随机变量

## 实值随机变量简介

---

到目前为止：

实验  $\Rightarrow$  所有可能结果的  $\Omega$  空间

硬币： $\Omega = \{\text{正面}, \text{反面}\}$

芯片： $\Omega = \{1, 2, 3, 4, 5, 6\}$

} 结果的识别

现在：考虑实数集

确定每个结果 $\omega_i$ 并将其与一个实数相关联：

$$x_i : x_i = X(\omega_i)$$

兴趣：描述各种随机实验的独特支持。 \_\_\_\_\_

\_\_\_\_\_

示例1:\_\_\_\_\_

实验:掷骰子: $\Omega = \{\text{Ace}, \text{King}, \text{Queen}, \text{Jack}, 10, 9\}$

对于每个结果,我们都会关联一个数字。

结果:无线	实变量: $x_i = X(\omega_i)$
高手	1
国王	2
女王	3
杰克	4
十	5
九	6

$$\omega_i = \{\text{Queen}\} \Leftrightarrow x_i = 3$$

示例2:\_\_\_\_\_

实验:抛硬币3次

随机变量:结果出现尾部的次数 (3 次试验中)

结果: $w_i$ 实数变量: $x_i = X(w_i)$	
$w_1 = \{ TTT \}$	3
$w_2 = \{ TTH \}$	2
$= \{ THT \} w_3$	2
$= \{ THH \} w_4$	1
$w_5 = \{ HTT \}$	2
$= \{ HTH \} w_6$	1
$= \{ HHT \} w_7$	1
$= \{ HHH \} w_8$	0

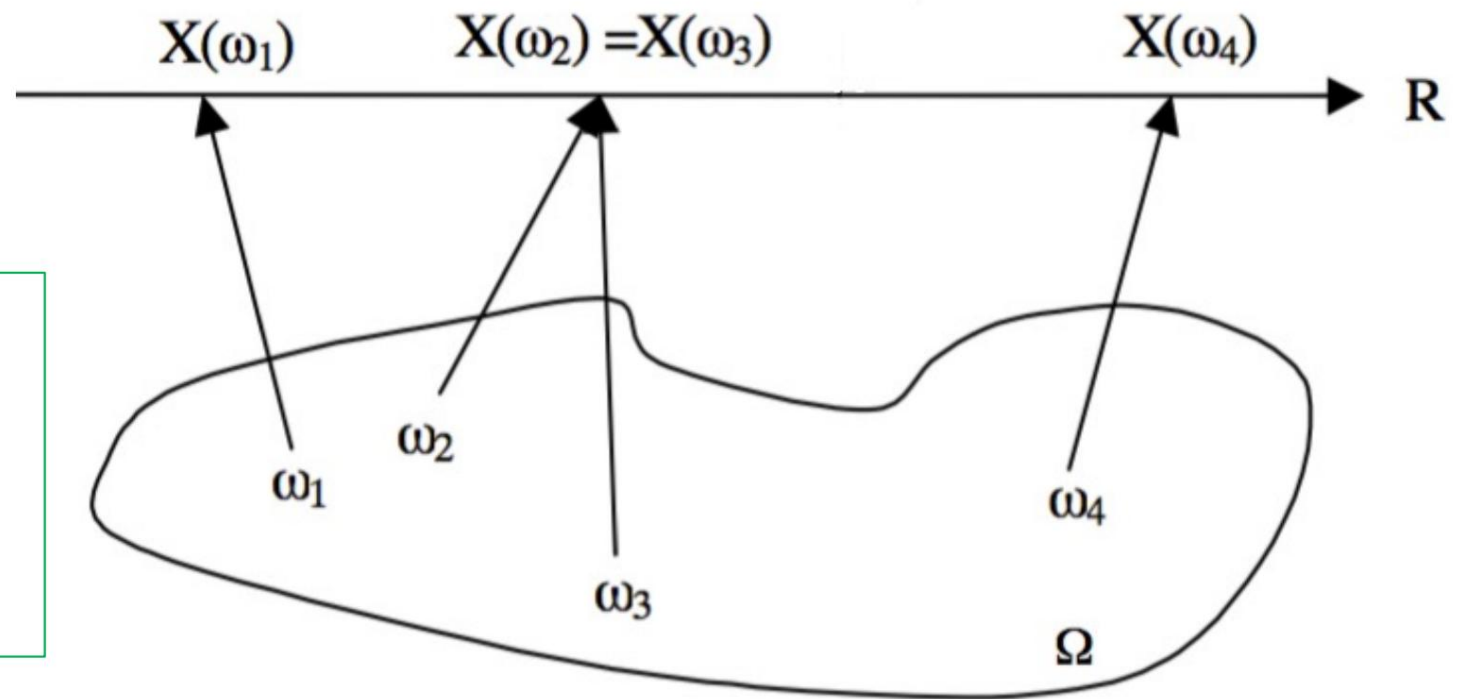
许多结果  
结果相同  
函数图像  
 $X(\cdot)$

## 实值随机变量的定义\*：

在从样本映射的概率空间上定义的函数  
空间  $\Omega$  到实数集。

函数  $X(.) : \Omega \rightarrow$

我们通过函数  $X(.)$  将  $\Omega$  的元素投影到  $R$  上：



提醒：

实际功能：

所有结果都只有一个图

像。2个结果可以有相同的  
图像,反之则不然。

\*通常缩写为 rv

## 如何将概率与随机变量关联起来？

鉴于  $X(\cdot)$  是从  $\Omega$  到  $\mathcal{X}$  的映射,我们可以将  $\mathcal{X}$  的子集与  $\Omega$  的子集匹配来找出概率。

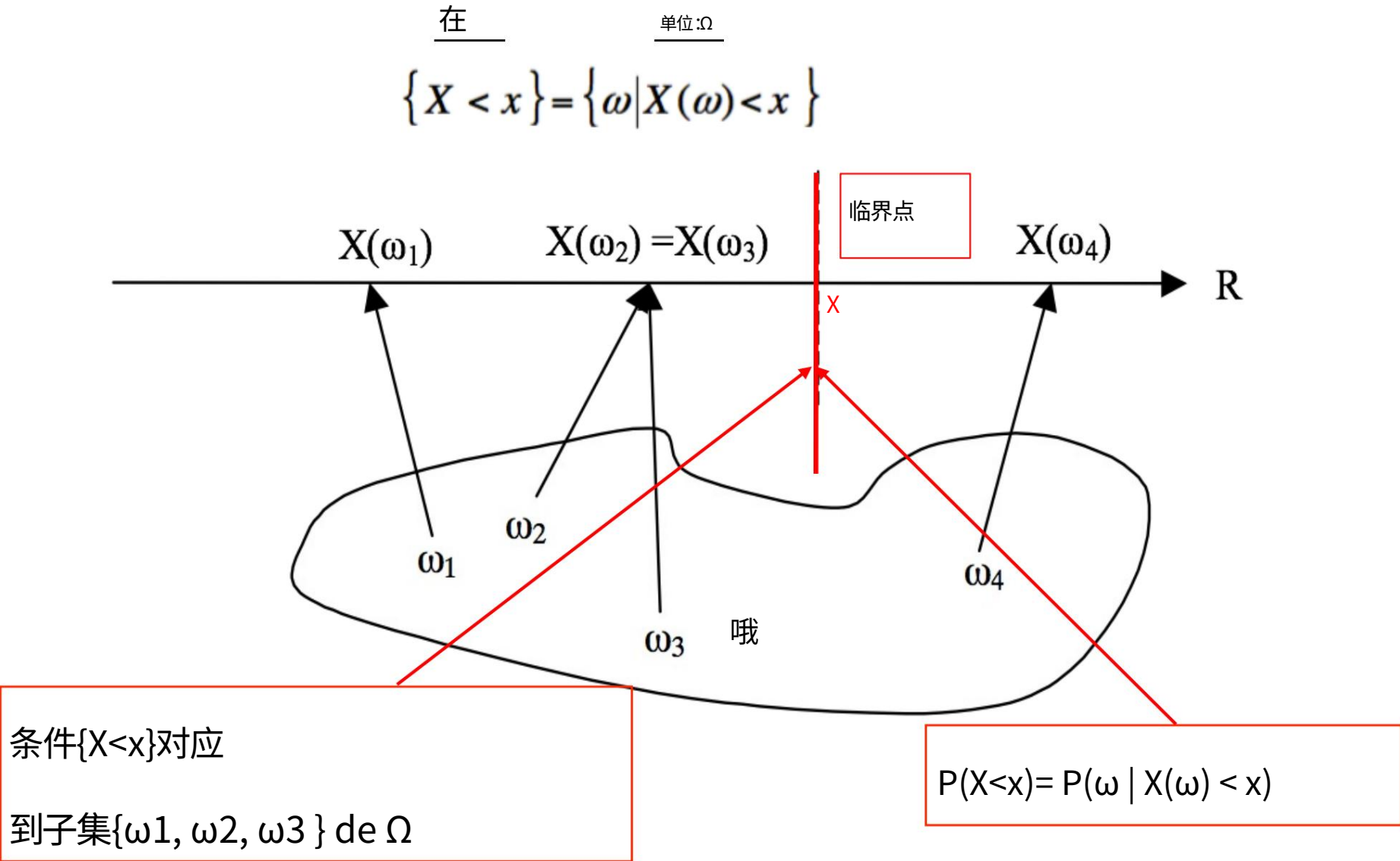
例如：

- 令  $A$  为  $\Omega$  的子集( $A \subset \Omega$ ), 概率为  $P(A)$ ,
- 令  $X(A)$  为  $x$  的集合,  $x$  是  $A$  通过函数的图像中的  $X(\cdot)$ 。
- 那么,  $X(A)$  的概率为：

$$P(X(A)) = P(\{\omega \text{ 使得 } X(\omega) \in X(A)\})$$

插图

在 中,令子集  $\{X < x\}$  对应于所有结果  $\omega$  的图像,使得它们的  $X(\cdot)$  图像小于阈值  $x$ :





### 实值随机变量的一般定义：

随机变量  $X$  是函数  $X(\cdot)$ ：

$$X : \Omega \rightarrow \mathbb{R} \quad \text{这样}$$

a) 满足条件  $\{X(\omega) < x\}$  并表示为  $\{X < x\}$  的点集  $\omega$  构成所有  $x$  的事件。

b) 事件  $\{X = +\infty\}$  和  $\{X = -\infty\}$  的概率为空。

在实践中： $X$ 是一个随机变量

如果  $P(X < x) \forall x$  实数已知

备注:某个区间的概率

$$C = \{x_1 \leq X < x_2\}$$

$X$  是一个真正的随机变量,考虑事件  $A$  和  $B$ :

$$A = \{X < x_1\} \qquad B = \{X < x_2\}$$

$$B = A \cup C = \{X < x_1\} \cup \{x_1 \leq X < x_2\}$$

$A$  和  $C$  不相交,根据公理 3 (可加性)我们得到:

$$P(\{x_1 \leq X < x_2\}) = P(B) - P(A) = P(\{X < x_2\}) - P(\{X < x_1\})$$

## 累积分布函数 (CDF)

---

累积分布函数 (CDF)的定义:

---

实数随机变量的累积分布函数 (CDF)

X是事件{  $X \leq x$  }的概率,表示为:

$$F_X(x) = P(\{X \leq x\})$$

符号:

F : 累积分布函数

X : 随机变量 ( F 的下标)

x : 实际阈值

P : 概率

例子：

实验：抛硬币3次。

随机变量 (rv)  $X$ ：3 次试验中结果出现尾部的次数。

$$P(TTT)=P(THT)=\cdots=P(HHH) = 1/8$$

$X$ 取的值和相应的概率：

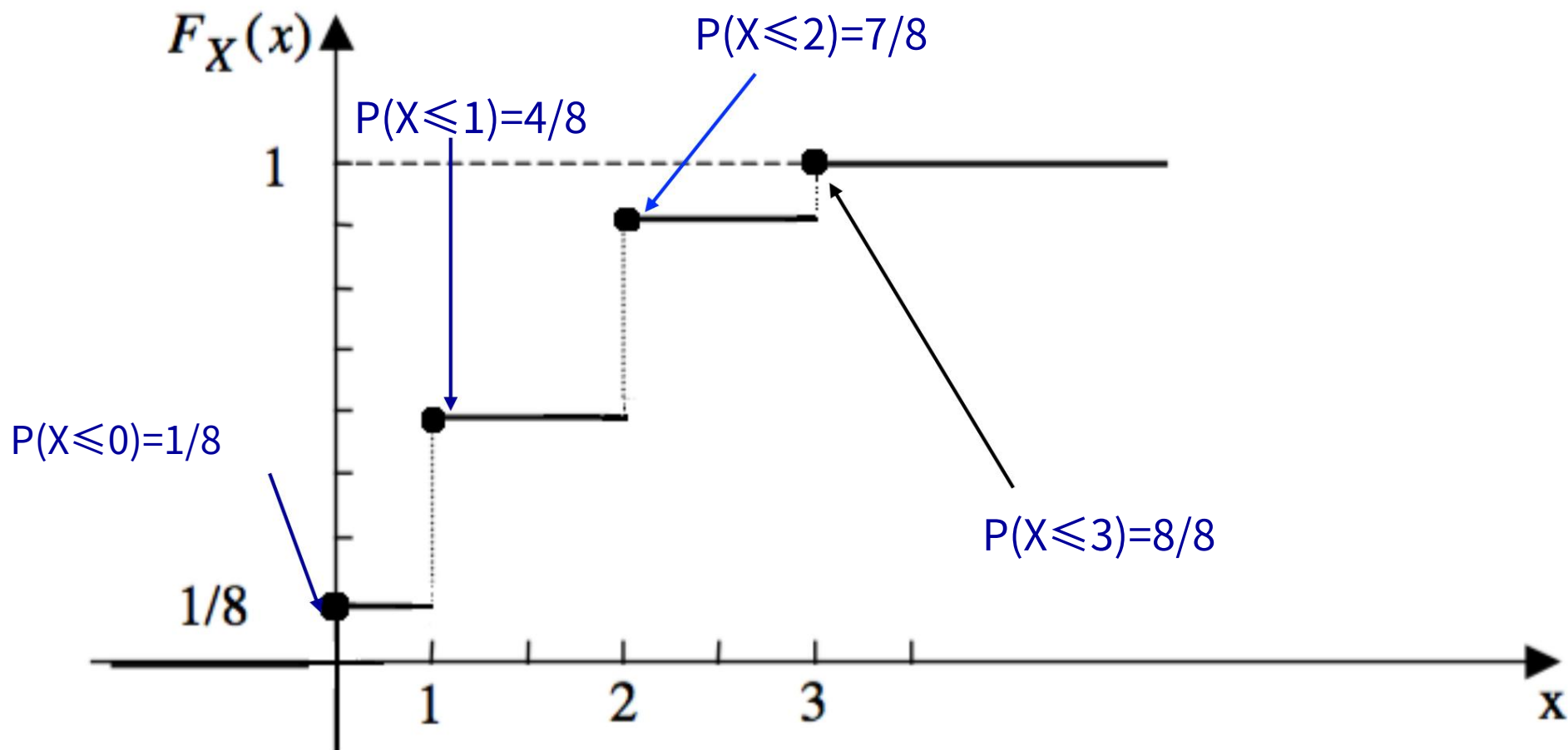
$$P(X = 0) = P(HHH) = 1/8$$

$$P(X = 1) = P(HHT \text{ 或 } HTH \text{ 或 } THH) = 3/8$$

$$P(X = 2) = P(HTT \text{ 或 } THT \text{ 或 } TTH) = 3/8$$

$$P(X = 3) = P(TTT) = 1/8$$

图:累积分布函数 (CDF)



阈值x

## 累积分布函数 (CDF) 的属性:

---

累积分布函数验证以下属性:

a) 有界且标准化

$$0 \leq F_X(x) \leq 1$$

b) 单调非递减

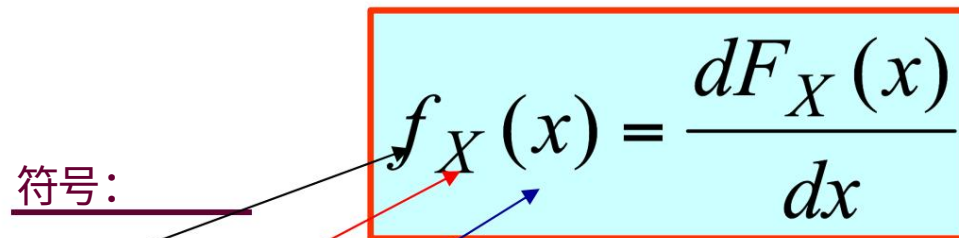
$$F_X(x+\varepsilon) \geq F_X(x), \varepsilon \geq 0$$

c) 右连续 (当  $x$  接近  $x_0$  时  $F(x)$  的极限  
右边 (大于  $x_0$  的值) 是  $F(x_0)$ 。

d) CDF 的限制是:

$$\lim_{x \rightarrow -\infty} F_X(x) = 0 \quad \lim_{x \rightarrow +\infty} F_X(x) = 1$$

## 概率密度函数:



The diagram shows the formula  $f_X(x) = \frac{dF_X(x)}{dx}$  enclosed in a light blue box with a red border. Three arrows originate from the text labels below: a black arrow points from '符号:' to the entire formula, a red arrow points from 'f : 概率密度函数' to the  $f$  in the numerator, and a blue arrow points from 'x : 实际阈值' to the  $x$  in the numerator.

$$f_X(x) = \frac{dF_X(x)}{dx}$$

符号:

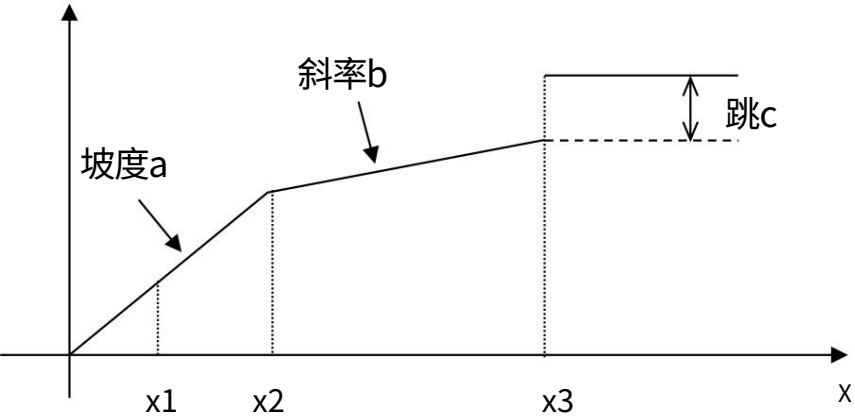
f : 概率密度函数

X : 随机变量 (f的下标)

x : 实际阈值

d./dx : 关于x (阈值)的导数

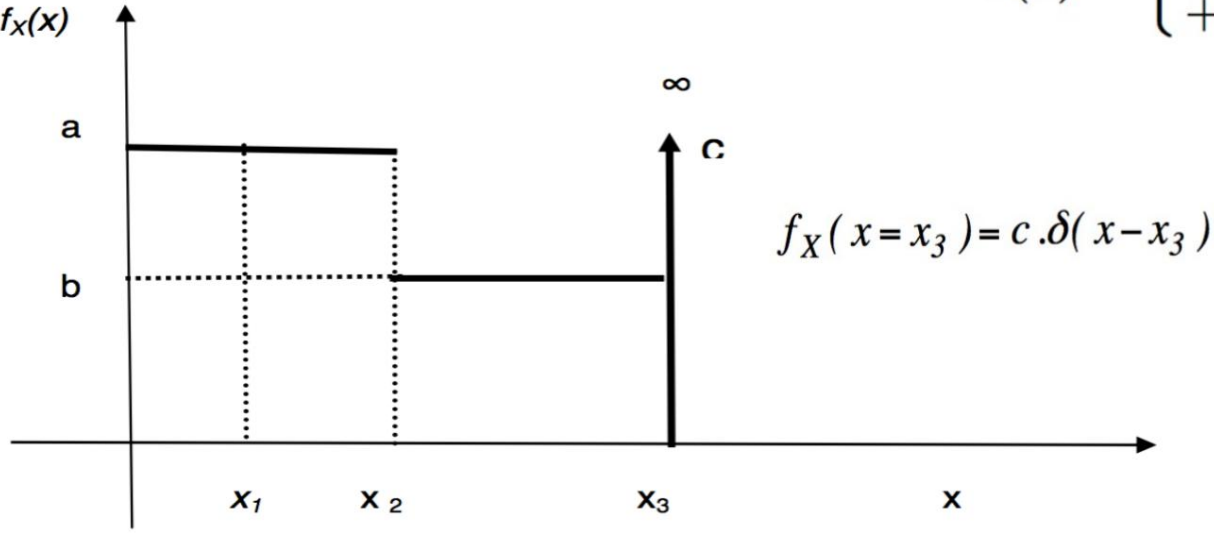
不连续累积分布函数的示例：



提醒：

$$\delta(x) = \begin{cases} 0, & x \neq 0 \\ +\infty, & x = 0 \end{cases}$$

密度：





## 概率密度函数的性质

1) 曲线下面积标准化:

$$F_X(x) = \int_{-\infty}^x f_X(u) du$$

特别是

$$F_X(+\infty) = \int_{-\infty}^{+\infty} f_X(u) du = 1$$

2) 概率密度函数处处非负:

$$f_X(x) \geq 0$$

(因为它是单调非递减函数的导数)

备注:如果函数  $f(x)$  非负且积分等于 1,则可以将其视为随机变量的概率密度函数。

## 连续随机变量

定义:如果一个随机变量的累积分布函数处处连续,则该随机变量是连续的。

### 连续随机变量密度的解释

---

$$P(x_1 \leq X \leq x_2) = \int_{x_1}^{x_2} f(x) dx$$

如果  $x_1 = x$  且  $x_2 = x + dx$  我们得到:  $P(x \leq X \leq x + dx) = \int_x^{x+dx} f(x) dx$

那么,  $f(x)dx$  可以看作变量  $X$  属于区间  $[x, x+dx]$  的概率。

密度可以解释为以下之间的比例因子:属于给定区间的概率和该区间的长度。

所以:

- $f(x)$  不是概率。
- $P(X = x) = 0$  (对于连续 rv)

## 离散随机变量

$X$  :随机变量,可以采用有限或至多可数无限的离散值集 (例如整数集) 。

对于每个值 $x_i$ 我们有:

$$P(X = x_i) = p_i \neq 0 \quad \text{和}$$

$$\sum_i p_i = 1$$

累积分布函数是不连续的并且具有阶梯状的外观

$$F_X(x) = \sum_i P(X = x_i) \cdot H(x - x_i)$$

$$f_X(x) = \sum_i P_i \cdot \delta(x - x_i)$$

在哪里:

Ø  $H(x)$ : 海维赛阶跃函数

$$\forall x \in \mathbb{R}, H(x) = \begin{cases} 0 & \text{si } x < 0 \\ 1 & \text{si } x \geq 0. \end{cases}$$

Ø  $\delta(x)$ :狄拉克 $\delta$ 函数

$$\delta(x) = \begin{cases} 0, & x \neq 0 \\ +\infty, & x = 0 \end{cases}$$

## 示例:抛硬币

两种结果1随  $(\quad) =$  与  $+ = 1$   
 $(\quad) =$

机变量 (指标) $X: \Omega \rightarrow$  !

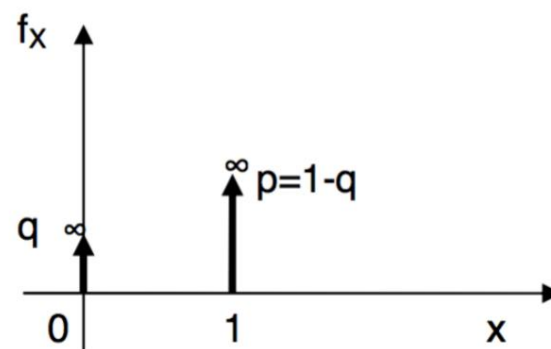
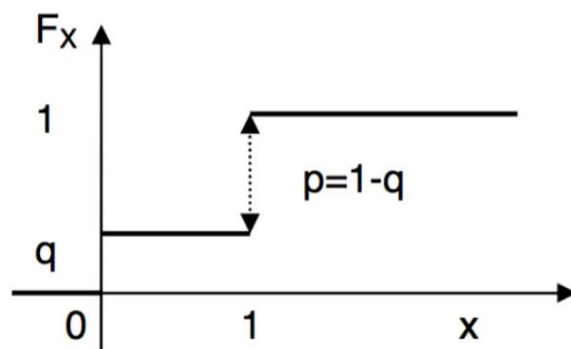
$$\begin{aligned} (\quad) &= 1 \\ (\quad) &= 0 \end{aligned}$$

累积分布函数

$$F_X(x) = q \cdot H(x) + p \cdot H(x-1)$$

概率密度函数

$$f_X(x) = q \cdot \delta(x) + p \cdot \delta(x-1)$$



概率在 0 和 1 之间变化。密度趋于无穷大。极限是狄拉克-德尔塔函数。

## 众所周知的离散概率分布的例子

Ø 伯努利分布: X有2个值: 0和1

$$P\{X = 1\} = p \quad P\{X = 0\} = q = 1 - p$$

Ø 参数为n和p的二项分布:

Ø X: 一系列n次伯努利实验的成功次数。

Ø X 取整数值 X: 0, 1, ..., n

$$0 \leq k \leq n \quad P\{X = k\} = \binom{n}{k} p^k q^{n-k} \quad \text{克肯克} \quad \text{与 } p + q = 1$$

Ø 参数  $\geq 0$  的泊松分布

X: 在固定时间间隔 (或空间) 内发生的事件数量。

假设: 事件以平均速率发生, 并且与上次事件以来的时间无关。

X 采用整数值 X: 0, 1, ..., 例如: 每小时的呼叫次数。

$$P\{X = k\} = \frac{e^{-\lambda} \lambda^k}{k!} \quad \text{我} \quad \text{立方}$$

## 著名的连续概率分布示例 (1/4)

---

Ø 参数均匀分布 和: \_\_\_\_\_

X 取区间  $[a, b]$  中的值

$$f(x) = \frac{1}{b-a} \quad \text{如果 } a \leq x \leq b$$

$f(x) = 0$  其他地方

Ø 带参数的指数分布: \_\_\_\_\_

X: 泊松过程中事件之间的时间。示例: 两次通话之间的时间。

$$f(x) = \lambda e^{-\lambda x} \quad \lambda > 0; x \geq 0$$

具有  $x_0$  位置参数和  $\lambda > 0$  尺度参数的柯西分布

---

$$f(x) = \frac{1}{\pi \left[ 1 + \left( \frac{x - x_0}{\lambda} \right)^2 \right]^{3/2}}; \lambda > 0$$

## 著名的连续概率分布示例 (2/4)

Ø 正态分布（也称为拉普拉斯高斯分布，或简称高斯分布），参数  $\mu$  和  $\sigma^2 > 0$ 。

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

我们表示  $X \sim (\mu, \sigma^2)$

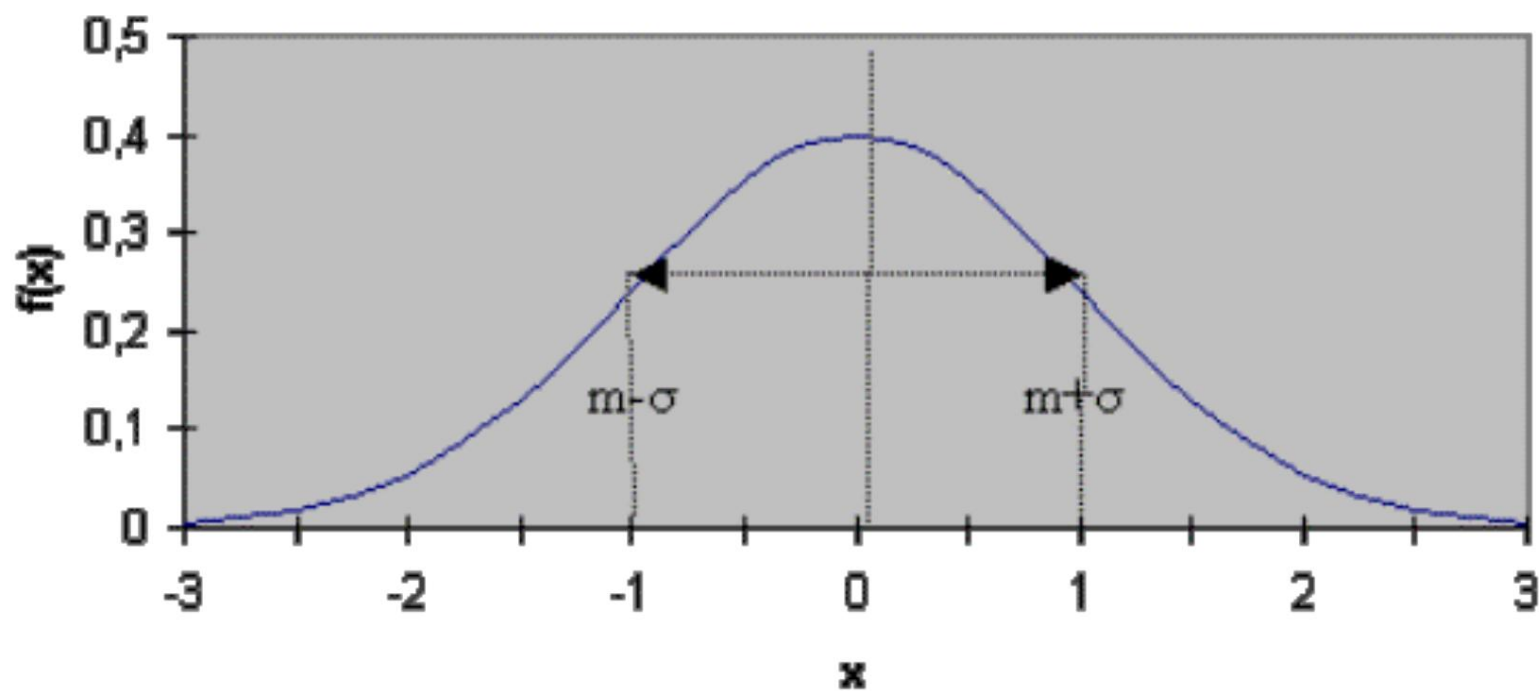
- 如果  $\mu = 0$ ,  $X$  是中心随机变量
- 如果  $\mu = 0$  且  $\sigma^2 = 1$ , 则该分布称为正态标准分布：

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$$

## 著名的连续概率分布示例 (3/4)

标准正态分布的概率密度

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





# 众所周知的连续概率分布的例子 (4/4)



累积表  
分布函数CDF  
(标准正态分布的。

~ ( , )

( ) = . ”  
%

例子 ( ≤ 1.96 ) = 0.975

z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,0	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986
3,0	0,9987	0,9987	0,9987	0,9988	0,9988	0,9989	0,9989	0,9989	0,9990	0,9990
3,1	0,9990	0,9991	0,9991	0,9991	0,9992	0,9992	0,9992	0,9992	0,9993	0,9993
3,2	0,9993	0,9993	0,9994	0,9994	0,9994	0,9994	0,9994	0,9995	0,9995	0,9995
3,3	0,9995	0,9995	0,9995	0,9996	0,9996	0,9996	0,9996	0,9996	0,9996	0,9997
3,4	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9998
3,5	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998
3,6	0,9998	0,9998	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999
3,7	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999
3,8	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999
3,9	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000

## 如何解读标准高斯分布的CDF表？

首先,标准化变量:

如果  $X \sim (\mu, \sigma^2)$ , 那么  $Z = \frac{X - \mu}{\sigma} \sim N(0, 1)$

接下来, 计算  $\Phi(z) = P(Z \leq z)$  其中  $Z \sim N(0, 1)$

$$\frac{\Phi(z)}{\sigma}$$

如果  $z = 0,12$  ; 我们有:  $\Phi(0,12) = 0,5478$

如果  $z = 0,05$  ; 我们有:  $\Phi(0,05) = 0,5199$

个位和十分位 和	z 的百分之一						
	0	1	2	3	4	5	6
0,0						(0,05)	
0,1			0,5478				
0,2					(和)		
0,3							
0,4							