数据科学基础知识

第一部分:概率论

ISEP第二年
2023-2024

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

概率论

第四次会议（2023 年 10 月 20 日）：

第 5 章（之二）:A 的变换
实值随机变量

第6章:二维随机
变量

第 5 章（之二）:实数的变换随机变量

– 随机变量的确定性函数 · 密度fX(x) 的变换 · 示例

· 特殊情况

问题描述

设 X 为密度fX(x)和 ga 确定性实函数的随机变量
变量 X 的。

X

概率密度随机变量fX(x)

转化g

和

概率密度的随机变量

: 已知

未知

变换 :确定性函数 ,Y=g(X)

变换示例： Y = X2； Y = aX + B

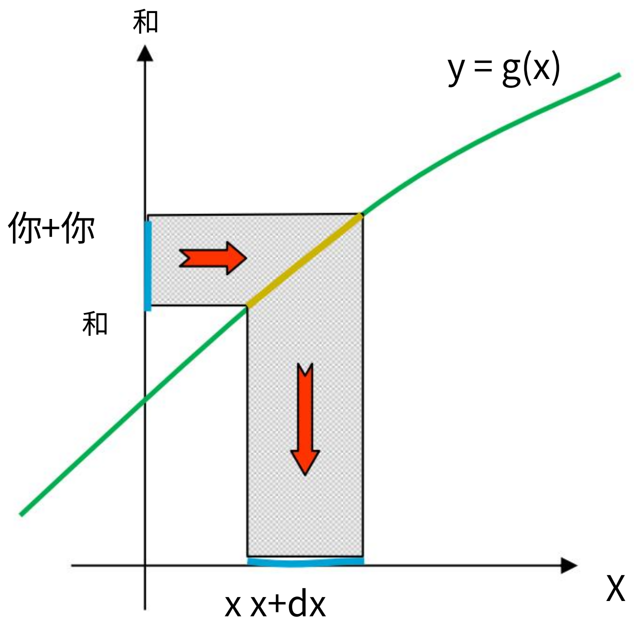
可能会发生三种情况：

- X 和 Y y之间一一对应 ⟷ X
- 非一对一对应 ⟷ (xi) 对于 i = 1, 2, ,.,, n
- 特殊情况

情况1:X和Y一一对应

子情况 1.1:

如果 g 单调递增



事件: 创建者:

$$\{y \leq Y \leq y + dy\} \Leftrightarrow \{x \leq X \leq x + dx\}$$

鉴于以下事实:

y固定只有一个先行词 x

dy>0 且 dx>0

- $P\{y \leq Y \leq y + dy\} = P\{x \leq X \leq x + dx\}$
- $$f_Y(y)\,dy = f_X(x)\,dx$$

那么,Y的密度为:

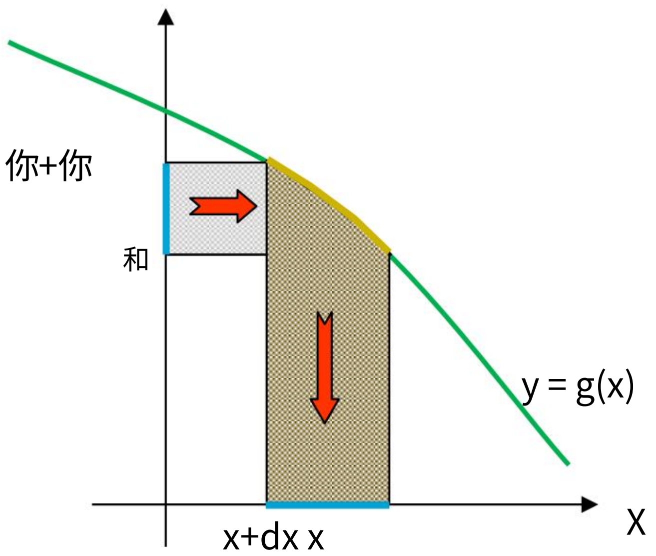
$$f_Y(y) = f_X(x)\frac{dx}{dy} = f_X(g^{-1}(y))\frac{dx}{dy}$$

其中g-1是g的反函数

情况1:X和Y一一对应

子情况 1.2:

如果g单调递减



鉴于以下事实:

y固定只有一个先行词 x

dy > 0 但dx < 0

- $P\{y \le Y \le y+dy\} = P\{x+dx \le X \le x\}$
- $f_Y(y)\,dy = -f_X(x)\,dx$

那么,Y的密度为:

$$f_Y(y) = f_X(x)\left|\frac{dx}{dy}\right| = f_X(g^{-1}(y))\left|\frac{dx}{dy}\right|$$

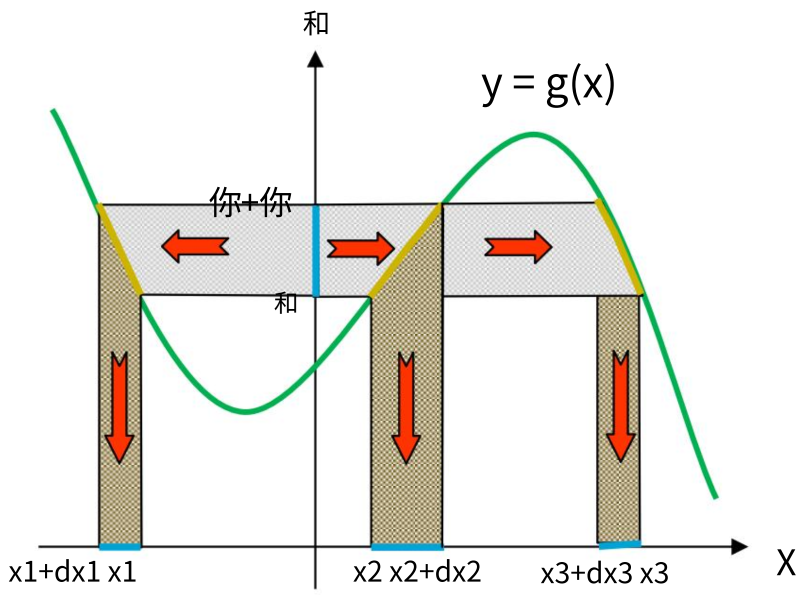
事件:

创建者:

$$\{y \le Y \le y+dy\} \Leftrightarrow \{x+dx \le X \le x\}$$

情况2:X和Y之间的对应关系不是一一对应（1/2）

X 中可能有多个值xi对应于同一值 y



y 可以有多个先行词

$$y = g(x_i) \qquad i = 1, 2, \ldots$$

（例如（x1， x2， x3）

dy>0 但dx1<0、 dx2>0、 dx3<0

事件：

创建者：

$$\{y \le Y \le y + dy\} \Leftrightarrow \{x_1 + dx_1 \le X \le x_1\} \text{或者} \{x_2 \le X \le x_2 + dx_2\} \text{或者} \{x_3 + dx_3 \le X \le x_3\}$$

情况2:X和Y之间不是一对一的对应关系（2/2）

$$\{y \le Y \le y + dy\} \Leftrightarrow \{x_1 + dx_1 \le X \le x_1\} \text{或者} \{x_2 \le X \le x_2 + dx_2\} \text{或者} \{x_3 + dx_3 \le X \le x_3\}$$

自事件发生以来

$$\{x_1 + dx_1 \le X \le x_1\},\ \{x_2 \le X \le x_2 + dx_2\},\ \{x_3 + dx_3 \le X \le x_3\}$$

是不相交的 $$f_Y(y)\left|dy\right| = \sum_i f_X(x_i)\left|dx_i\right|$$

根据公理3：

$$\boxed{f_Y(y) = \sum_i f_X(x_i)\left|\frac{dx}{dy}\right|_{x=x_i}}$$

其中xi是方程x = g的解 $^{-1}$(y) 当 y 固定时。

示例（1/2）：

设 X 为密度为fX(x) 的随机变量，

变换 Y = X2。目的是计算fY(y)？

反函数为： $x = \pm\sqrt{y}$ y>0

导数的计算： $dy = 2 \cdot x \cdot dx$

$$\left[g_i^{-1}(y)\right]' = \frac{dx}{dy} = \pm\frac{1}{2\sqrt{y}}$$

密度fY(y) 将为：

$$f_Y(y) = \frac{1}{2\sqrt{y}} \cdot \left[f_X\left(\sqrt{y}\right) + f_X\left(-\sqrt{y}\right)\right]$$

示例（2/2）：

应用： fX(x) 是瑞利密度：

$$f_X(x) = \frac{x}{\sigma^2} \cdot \exp\left(-\frac{x^2}{2\sigma^2}\right) \qquad x \geq 0$$

$$f_X(x) = 0 \quad \text{别处}$$

警告 :x ⩾ 0 只有正根属于定义域

$$f_Y(y) = \frac{1}{2\sqrt{y}} \cdot \frac{\sqrt{y}}{\sigma^2} \cdot \exp\left(-\frac{\left(\sqrt{y}\right)^2}{2\sigma^2}\right) = \frac{1}{2\sigma^2} \cdot \exp\left(-\frac{y}{2\sigma^2}\right)$$

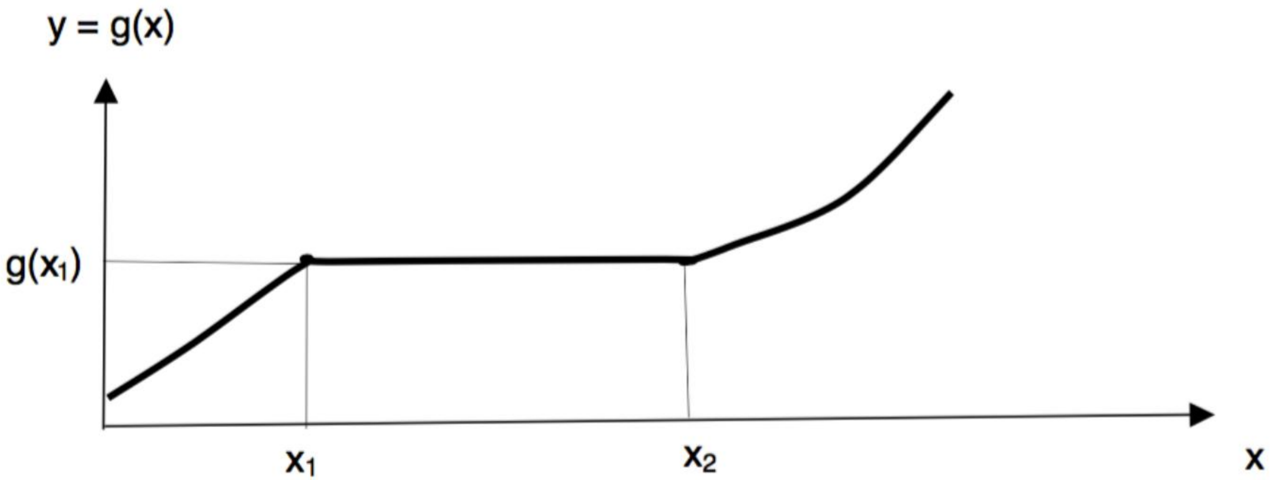
y >=0

它是一个指数分布，参数为 “#!

$\frac{!}{}$

特殊情况 1:g(x) 在区间内为常数

让我们假设 g 是连续的,对于除了区间[x1 , x2]之外的所有 x 来说,g 是不递减的



$$g(x_1) = g(x_2)$$

$$P\{Y = g(x_1)\} = P\{x_1 \le X \le x_2\} = \int_{x_1}^{x_2} f_X(x) \cdot dx$$

区间[x1,x2]中 x 的所有值都转换为 Y=g(x1) 的单个值。
并且 P(Y=g(x1)) ≥ 0（Y 属于长度为 0 的区间的概率）。所以：

Ø密度， fY(y) 是点 y = g(x1) 处的狄拉克分布

Ø CDF 的不连续性， FY(y)

$$F_Y\{y = g(x_1^+)\} = F_Y\{y = g(x_1^-)\} + P\{Y = g(x_1)\}$$

具体案例示例 (1/2)

随机变量 X 在区间 [− 1,1] 内服从均匀分布。

考虑转型 ( )

= %如果 如果 ≥ 0 如果 则为 0

对于x < 0： (= 0 ∋ (< 0 ∋ ! ” $ ∫ #! = ! ”

对于x ⩾ 0 $$f_Y(y) = f_X(x) = \frac{1}{2}$$

因此,Y的概率密度函数为：

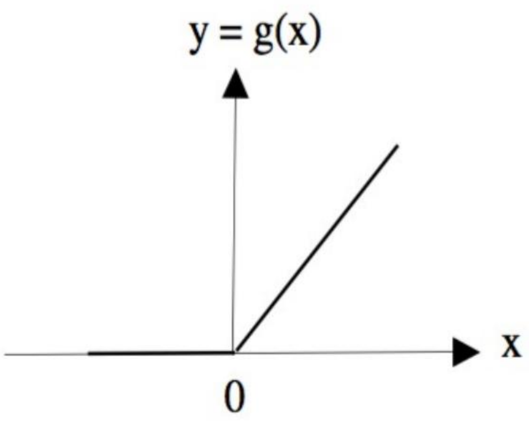
$ ( ) = ! ” ( ) ⇔ = = ( ) ! ” , =0

$ ( ) = ! ”, 0 < ⩽ 1

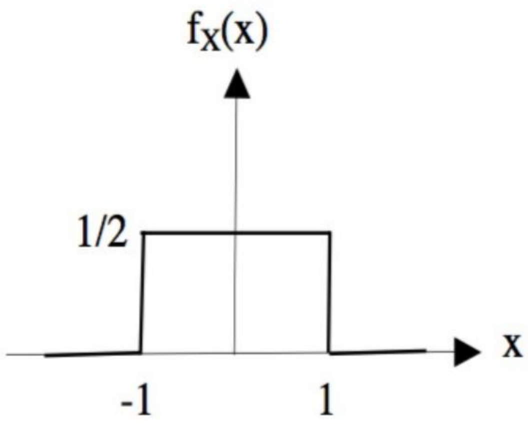
Y的 CDF为：

$ ( ) = ! − + % “”, 0⩽y⩽1

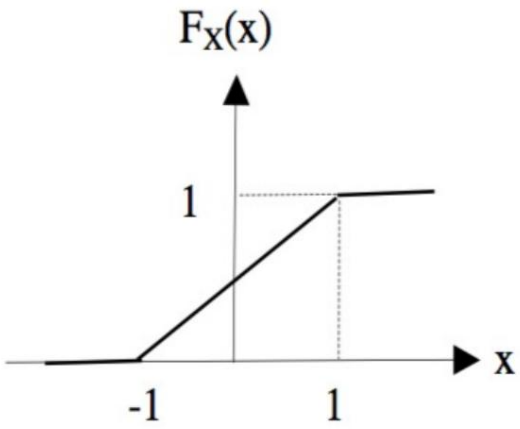
特殊情况示例 (2/2)



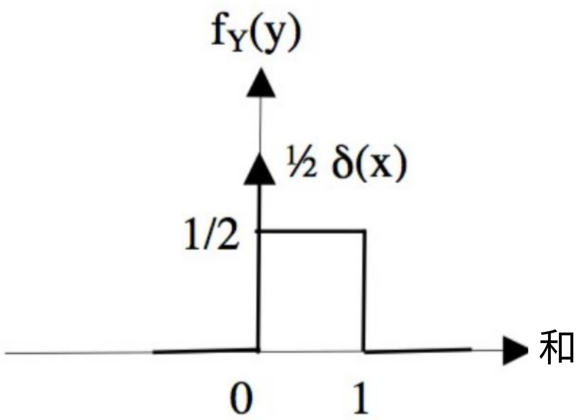
转型



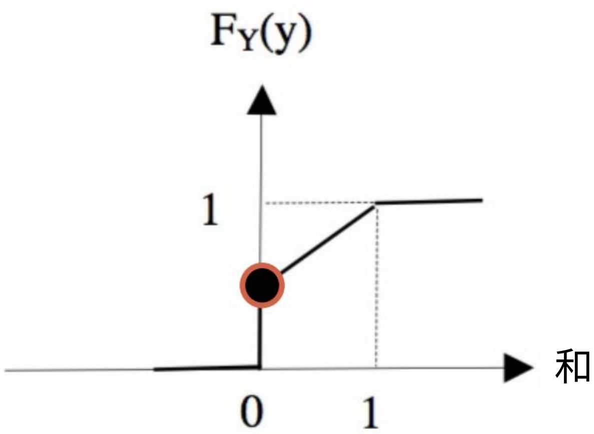
X 的密度



X 的累积分布函数



Y 的密度



Y 的 CDF

第 6 章（开头）:二维随机变量

- 二维离散随机变量
- 二维连续随机变量

第 6 章（开头）:二维随机变量

- 二维离散随机变量
- 二维连续随机变量

一维随机变量的提醒

sional random variable

一维随机变量 X 完全由以下数据之一定义：

» 其累积分布函数（CDF） FX(x)

» 其概率密度函数 FX(x)

» 其特色功能 x(x)

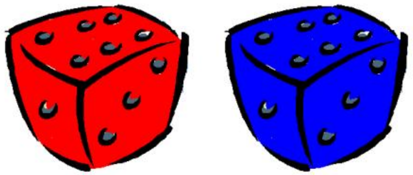
由其矩mn, n描述

通过变换函数创建新的随机变量的可能性。

所有这些概念都将扩展到两个随机的变量，（X,Y)

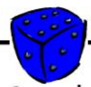
示例 :实验

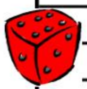
掷两个骰子,一红一蓝



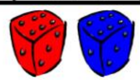
- 36 种可能的结果
:(出现的数字
在每个骰子的上表面)
  - 骰子是

对称 => 等概率
（所有结果具有相同的概率）





	1	2	3	4	5	6
1	(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
2	(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
3	(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
4	(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
5	(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
6	(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)



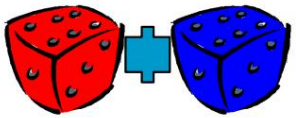
红色骰子(l=1,2,3,4,5,6),蓝色骰子(k= 1,2,3,4,5,6)

$$P\{\omega_n\} = \frac{1}{36}, \; n = 1 \text{ à } 36$$

二维随机变量的定义（1/2）

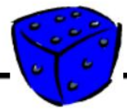
为了定义随机变量对 (X,Y),我们定义两个映射
从 Ω 到实数集：

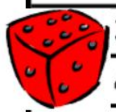
X（红色骰子） Y（2 个骰子的总和）



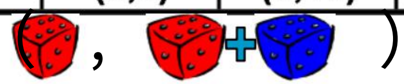
$$
\begin{aligned}
X(\omega_n) &= X(l,k) = l = x_i \\
Y(\omega_n) &= Y(l,k) = l+k = y_j
\end{aligned}
$$

该对 (X,Y) 的可能结果：





	1	2	3	4	5	6
1	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)	(1,7)
2	(2,3)	(2,4)	(2,5)	(2,6)	(2,7)	(2,8)
3	(3,4)	(3,5)	(3,6)	(3,7)	(3,8)	(3,9)
4	(4,5)	(4,6)	(4,7)	(4,8)	(4,9)	(4,10)
5	(5,6)	(5,7)	(5,8)	(5,9)	(5,10)	(5,11)
6	(6,7)	(6,8)	(6,9)	(6,10)	(6,11)	(6,12)



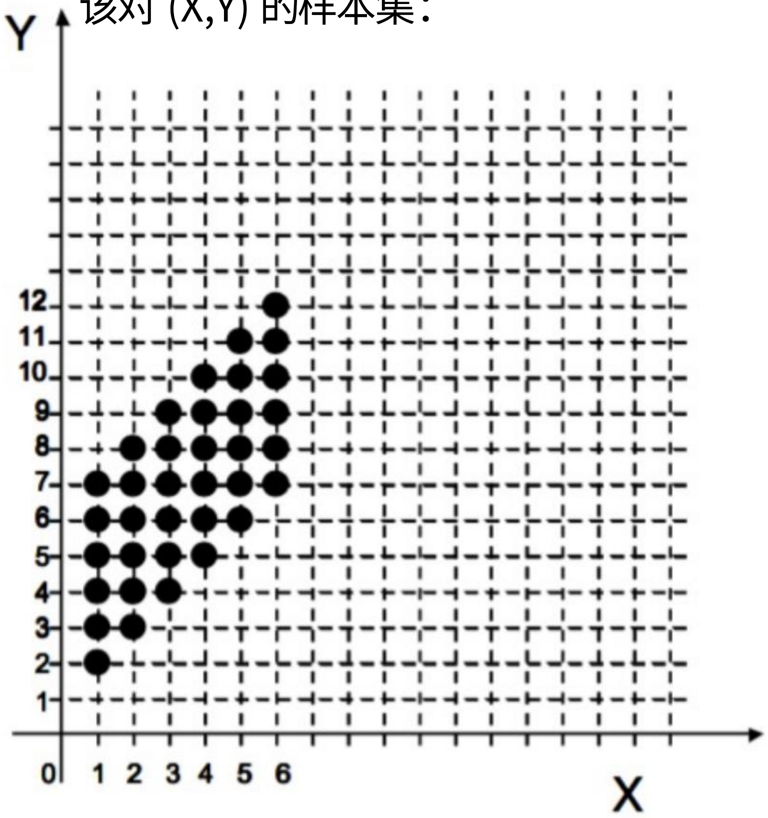
二维随机变量的定义（2/2）

X 的可能结果： $x_i = 1,2,\ldots 6$（对于 i=1 到 6）

Y的可能结果： $y_j = 2,3,\ldots 12$（对于 j=1 到 11）

这对夫妇 (X, Y) 不能
取任意值
组合(x_i , y_j)。例如，值 (4,3)
或 (2,10) 永远不会出现！

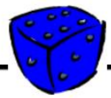
该对 (X,Y) 的样本集：



(X,Y) 对的概率密度： P(X=xi且Y=yj)=Pij

这两个映射在两者之间建立了一对一的关系（双射）

结果 (ωn) 和有序对(xi , y j)。



	1	2	3	4	5	6
1	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)	(1,7)
2	(2,3)	(2,4)	(2,5)	(2,6)	(2,7)	(2,8)
3	(3,4)	(3,5)	(3,6)	(3,7)	(3,8)	(3,9)
4	(4,5)	(4,6)	(4,7)	(4,8)	(4,9)	(4,10)
5	(5,6)	(5,7)	(5,8)	(5,9)	(5,10)	(5,11)
6	(6,7)	(6,8)	(6,9)	(6,10)	(6,11)	(6,12)

夫妻的每个值有 36 个结果： (x_i, y_j)

$$p_{i,j} = P\{\omega_n\} = \frac{1}{36},$$ 对于n=1到 36

列联表

定义 (X, Y) 对概率分布的表格

Y	x1=1	x2=2	x3=3	x4=4	x5=5	x6=6	P Yj
y1=2	1/36	0	0	0	0	0	1/36
y2=3	1/36	1/36	0	0	0	0	2/36
y3=4	1/36	1/36	1/36	0	0	0	3/36
y4=5	1/36	1/36	1/36	1/36	0	0	4/36
y5=6	1/36	1/36	1/36	1/36	1/36	0	5/36
y6=7	1/36	1/36	1/36	1/36	1/36	1/36	6/36
y7=8	0	1/36	1/36	1/36	1/36	1/36	5/36
y8=9	0	0	1/36	1/36	1/36	1/36	4/36
y9=10	0	0	0	1/36	1/36	1/36	3/36
y10=11	0	0	0	0	1/36	1/36	2/36
y11=12	0	0	0	0	0	1/36	1/36
P Xi	6/36	6/36	6/36	6/36	6/36	6/36	1

联合律 (X,Y)

$$P_{ij} \ge 0$$

$$\sum_i \sum_j P_{ij} = 1$$

Y 的边际律

$$P(Y = y_j) = \sum_i P_{ij}$$

表中所有条目的总和
必须始终为 1！

X 的边际法则

$$P(X = x_i) = \sum_j P_{ij}$$

有条件分配

给定X等于固定值,例如X = x3 = 3,计算当X取值 3时Y的条件概率。 ，我们可以

Y	x1=1	x2=2	x3=3	x4=4	x5=5	x6=6	P Yj
y1=2	1/36	0	0	0	0	0	1/36
y2=3	1/36	1/36	0	0	0	0	2/36
y3=4	1/36	1/36	1/36	0	0	0	3/36
y4=5	1/36	1/36	1/36	1/36	0	0	4/36
y5=6	1/36	1/36	1/36	1/36	1/36	0	5/36
y6=7	1/36	1/36	1/36	1/36	1/36	1/36	6/36
y7=8	0	1/36	1/36	1/36	1/36	1/36	5/36
y8=9	0	0	1/36	1/36	1/36	1/36	4/36
y9=10	0	0	0	1/36	1/36	1/36	3/36
y10=11	0	0	0	0	1/36	1/36	2/36
y11=12	0	0	0	0	0	1/36	1/36
P Xi	6/36	6/36	6/36	6/36	6/36	6/36	1

的条件概率
X 给定 Y 固定

的条件概率
给定 X Y

$$P(Y = y_j | X = x_i) = \frac{P_{ij}}{P_{X_i}} = \dots \frac{1}{6}$$

$$P(X = x_i | Y = y_j) = \frac{P_{ij}}{P_{Y_j}} = \dots \frac{1}{3}$$

总结几个离散随机变量 (X,Y)

该对 (X,Y) 具有以下结果：

= !, = 1,2, ⋯

= “, = 1,2, ⋯

联合概率函数 (= 和= ! „) = ! , ”

X的边际分布 (= %) = & %,& = % &

Y的边际分布 (= &) = & %,& = & %

给定 Y 的 X 的条件分布 (= % | = &) = %,& / Σ%%,& = %,& / &

给定 X 的 Y 的条件分布 (= & | = %) = %,& / Σ&%,& = %,& / %

独立

如果对于任何i, j ,两个变量X和Y是独立的：

(= % ∩ = &) = = %( ) (= &) 或者 %& = ∀ % & ,

如果它们是独立的,则条件分布为

与边际分布相同。

(= ! | = ”) = ! , ∀ ”

(= ” |= !) = ” , ∀ !

第 6 章（开头）:二维随机变量

- 二维离散随机变量
- 二维连续随机变量

累积分布函数 (CDF)

定义：

一维变量提醒: $F_X(x) = P(X \leqslant x)$

对于二维随机变量 (X,Y)， CDF 是
事件相交的概率： $\{X \leqslant x\}$ 和 $\{Y \leqslant y\}$

$$\boxed{F_{X,Y}(x, y) = (\{\ \leqslant \ \} \cap \{\ \leqslant \ \})}$$

符号：

F :联合累积分布函数

X,Y :随机变量

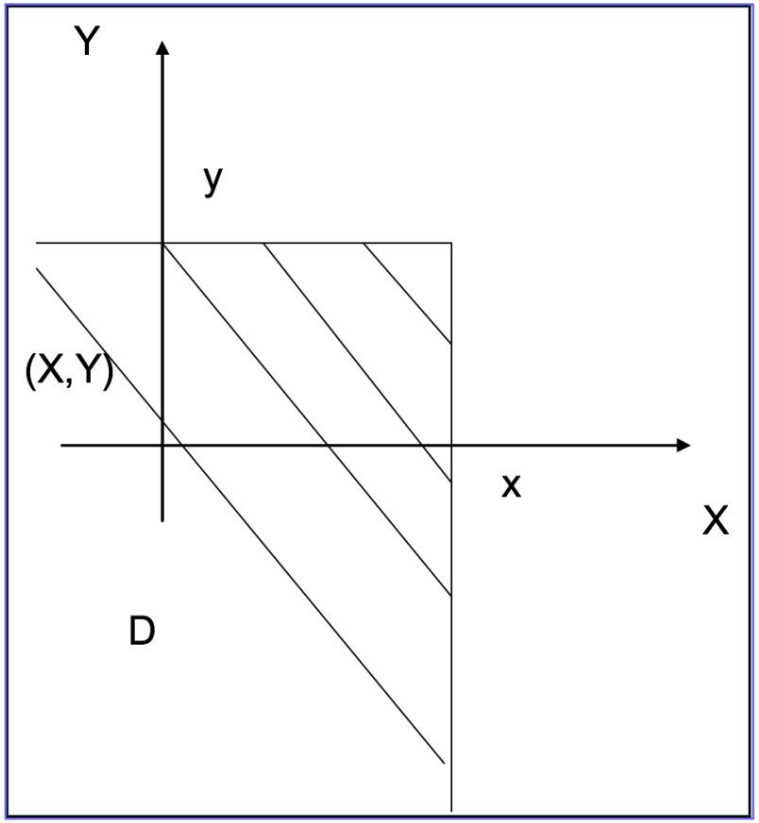
x, y :实际阈值

P :概率

图形表示（非常有用！）：

笛卡尔坐标系中的表示

FXY(x, y)： (X,Y) ∈ D 的概率



边际累积分布函数

X 的边际累积分布函数

它只是X的累积分布函数

$$F_X(x) = \quad (\{ \quad \leqslant = \}) \qquad (\{ \quad \leqslant \quad \}, \forall \quad)$$

只需考虑二维累积分布函数并设置：

$y = +\infty$ （或域$D_{x,y}$中$y = y_{max}$ ）

$$\boxed{F_X(x) = \quad (\{ \quad \leqslant \quad , \forall < + \text{无穷大}\}) = F_{X,Y}(x, +\text{无穷大})}$$

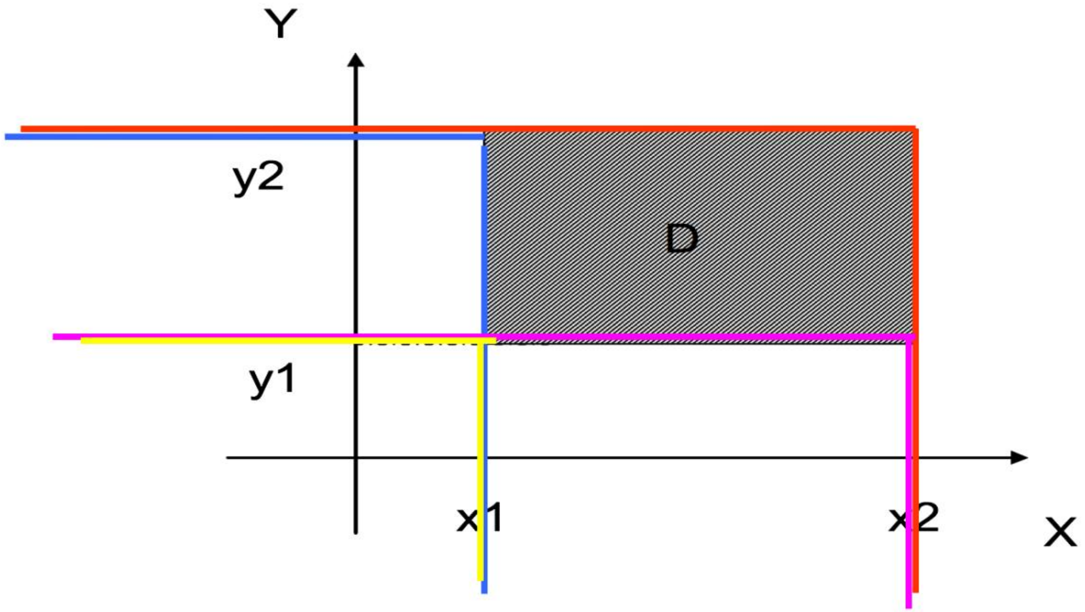
Y 的边际累积分布函数

$$\boxed{F_Y(y) = F_{X,Y}(+\infty, y)}$$

矩形的概率计算（二维区间）

我们的目标是根据FX ,Y (x,y)计算(X ,Y)对属于矩形D的概率：

:{ ε [,] 和 ε [,]}



({ % < ⩽ &, % < ⩽ & })

= FX,Y (&, &)－FX,Y (%, &)－FX,Y (&, %)+FX,Y (%, %)

联合概率密度函数

一维情况：

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

二维情况：

$$f_{X,Y}(x,y) = \frac{\partial}{\partial y}\left(\frac{\partial F_{X,Y}(x,y)}{\partial x}\right) = \frac{\partial^2 F_{X,Y}(x,y)}{\partial y \partial x}$$

符号：

f :概率密度函数

X、Y:随机变量

x, y :实际阈值

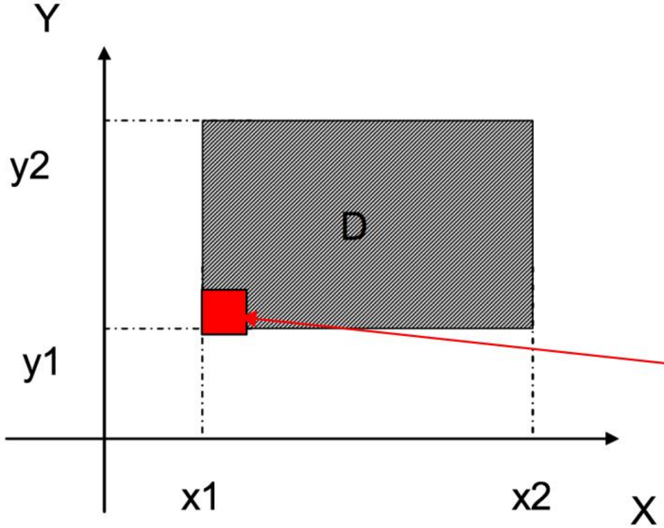
$\frac{\partial^2}{\partial y \partial x}$:相对于 x（阈值)和 y（阈值)的导数

概率密度的无穷小解释

让我们计算域 D（矩形)上的积分：

$$\int_{x1}^{x2}\int_{y1}^{y2} f_{X,Y}(x,y)dxdy = \underbrace{F_{X,Y}(x_2,y_2)-F_{X,Y}(x_1,y_2)-F_{X,Y}(x_2,y_1)+F_{X,Y}(x_1,y_1)}$$

({ % < ≤ &, % < ≤ & })



在无穷小域dS=dxdy中密度
可以认为是常数

$$x_1 = x,\ x_2 = x + dx$$

$$y_1 = y,\ y_2 = y + dy$$

({ < ≤ + , < ≤ + }) = ,), ( )

解释： (X,Y)对在点(x,y)的邻域属于dS的概率与该点的联合密度函数的值成正比。

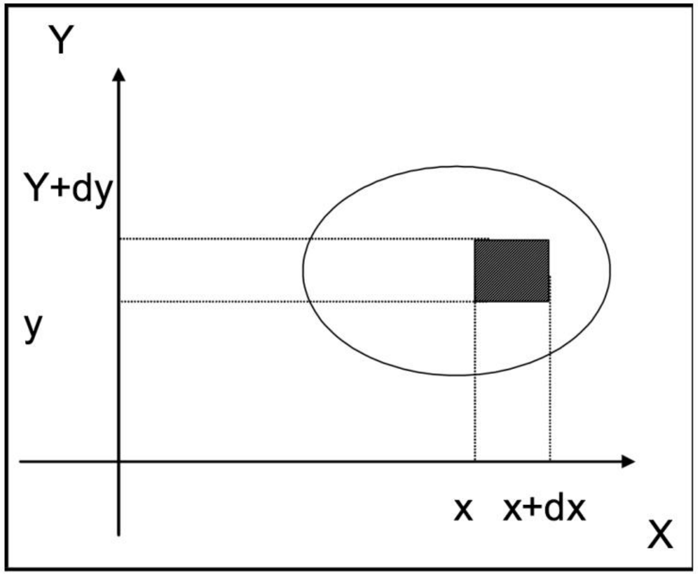
结论 :对于连续随机变量 ,{X = x, Y = y} 的概率为零
因为 dS 被视为零。

使用密度：

密度允许计算几个随机的概率
变量 (X, Y) 属于任意域 D

$$P(\{(X,Y)\in D\})=\int_D f_{X,Y}(x,y)\,dxdy$$

只需将 D 分解为不相交的元素并应用概率公理 3 即可。



积分表示由联合概率密度函数定义的表面之间的体积

和域 D。

与累积分布函数的关系：

$$\boxed{\int_{-\infty}^{x_2}\int_{-\infty}^{y_2} f_{X,Y}(x,y)dxdy = F_{X,Y}(x_2,y_2)}$$

通过取x2 = + ∞， y2 = + ∞,我们得到：

$$\boxed{\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f_{X,Y}(x,y)dxdy = F_{X,Y}(+\infty,+\infty) = 1}$$

特性：

- 表面下的体积为 1。

- 密度是非负的。

$$\boxed{f_{X,Y}(x,y) \geq 0}$$

X 和 Y 的边际密度：

X 的边际密度

$$\boxed{f_X(x)=\int_{-\infty}^{+\infty} f_{X,Y}(x,y)\,dy}$$

Y 的边际密度

$$\boxed{f_Y(y)=\int_{-\infty}^{+\infty} f_{X,Y}(x,y)\,dx}$$

提示:计算另一个变量在定义域上的积分,使其消失

条件分布

给定 X 的 Y 的条件分布

$$f_Y(y|X=x)=\frac{f_{X,Y}(x,y)}{\int_{-\infty}^{+\infty} f_{X,Y}(x,y)dy}=\frac{f_{X,Y}(x,y)}{f_X(x)}$$

给定 Y 的 X 的条件分布

$$f_X(x|Y=y)=\frac{f_{X,Y}(x,y)}{\int_{-\infty}^{+\infty} f_{X,Y}(x,y)dx}=\frac{f_{X,Y}(x,y)}{f_Y(y)}$$

如果 X 和 Y 独立会发生什么？

提醒:如果事件A和B是独立的：

$$P(A\cap B) = P(A)P(B)$$

如果 A = {X ≤ x} 且 B = {Y≤ y},则P(X≤ x, Y≤y) = P(X≤ x)P(Y≤y)：

$$\boxed{F_{X,Y}(x,y) = F_X(x)F_Y(y)}$$

耦合的联合累积分布函数等于边际累积分布函数的乘积。

密度怎么样？

$$\boxed{f_{X,Y}(x,y) = f_X(x)f_Y(y)}$$

联合密度也是边缘密度的乘积。