3.2 边缘分布 marginal distribution

二维随机变量 (X,Y) ,是两个随机变量视为一个整体，来讨论其取值规律的，我们可用分布函数来描述其取值规律。

$$F(x,y)=P\{X\le x,Y\le y\}$$

问题：能否由二维随机变量的分布来确定两个一维随机变量的取值规律呢？如何确定呢？

——边缘分布问题

边缘分布 marginal distribution

设二维随机变量 (X,Y) 的分布函数为 $F(x,y)$，

$$F_X(x) = P\{X \le x\} = P\{X \le x, Y < +\infty\} = F(x,+\infty)$$

$$F_Y(y) = P\{Y \le y\} = P\{X < +\infty, Y \le y\} = F(+\infty, y)$$

依次称为二维随机变量 (X,Y) 关于 X 和关于 Y 的边缘分布函数.

$$\boxed{F_X(x) = F(x,+\infty)} \qquad \boxed{F_Y(y) = F(+\infty, y)}$$

边缘分布实际上是高维随机变量的某个(某些)低维分量的分布。

二维离散型R.v.的边缘分布

如果二维离散型随机变量（**X,Y**）的联合分布律为

$$P\{X=x_i, Y=y_j\}=p_{ij} \qquad i,j=1,2,3,\cdots$$

对于二维离散型随机变量(X, Y)，有

$$\mathrm{F}_X(x)=F(x,+\infty)=\sum_{x_i\le x}\sum_{j=1}^{+\infty}p_{ij}$$

与一维离散型随机变量X的分布函数$\mathrm{F}_X(x)=\sum_{x_i\le x}P\{X=x_i\}$比较，得X的分布律

$$P\{X=x_i\}=\sum_{j=1}^{+\infty}p_{ij}\overset{\text{记为}}{=}p_{i\cdot}$$

同样，Y的分布律$P\{Y=y_j\}=\sum_{i=1}^{+\infty}p_{ij}\overset{\text{记为}}{=}p_{\cdot j}$

二维离散型R.v.的边缘分布

如果二维离散型随机变量（**X,Y**）的联合分布律为

$$P\{X = x_i, Y = y_j\} = p_{ij} \qquad i, j = 1,2,3,\cdots$$

即

Y \ X	X_1	X_2	X_3	…
Y_1	p_{11}	p_{21}	p_{31}	…
Y_2	p_{12}	p_{22}	p_{32}	…
Y_3	p_{13}	p_{23}	p_{33}	…
…	…	…	…	…

二维离散型R.v.的边缘分布

Y \ X	X_1	X_2	X_3	…	$P_{\cdot j}$
Y_1	p_{11}	P_{21}	p_{31}	…	$P_{\cdot 1}$
Y_2	P_{12}	P_{22}	p_{32}	…	$P_{\cdot 2}$
Y_3	P_{13}	p_{23}	p_{33}	…	$P_{\cdot 3}$
…	…	…	…	…	…
$p_{i\cdot}$	$p_{1\cdot}$	$p_{2\cdot}$	$p_{3\cdot}$	…	

关于**X**的边缘分布 $$p_{i\cdot} = P\{X = x_i\} = \sum_j p_{ij}$$

关于**Y**的边缘分布 $$p_{\cdot j} = P\{Y = y_j\} = \sum_i p_{ij}$$

二维离散型R.v.的边缘分布律

关于**X**的边缘分布

X	x_1	x_2	x_3	…
概率	$P_{1\cdot}$	$P_{2\cdot}$	$P_{3\cdot}$	…

$$p_{i\cdot} = P\{X = x_i\} = \sum_j p_{ij}$$

第**i**列之和

关于**Y**的边缘分布

Y	y_1	y_2	y_3	…
概率	$P_{\cdot 1}$	$P_{\cdot 2}$	$P_{\cdot 3}$	…

$$p_{\cdot j} = P\{Y = y_j\} = \sum_i p_{ij}$$

第**j**行之和

二维离散型R.v.的边缘分布

例1　设二维离散型随机变量（**X,Y**）的联合分布律为

Y \ X	0	1	1/3
-1	0	1/3	1/12
0	1/6	0	0
2	5/12	0	0

求关于**X**、**Y**的边缘分布

解　关于**Y**的边缘分布为

Y	-1	0	2
概率	5/12	1/6	5/12

关于**X**的边缘分布

X	0	1	1/3
概率	7/12	1/3	1/12

（**X**，**Y**）的联合分布列

Y \ X	0	1	1/3
-1	0	1/3	1/12
0	1/6	0	0
2	5/12	0	0

例1　设一个整数N等可能在1,2…10这10个值中取1个值。设D=D(N)是能整除N的正整数的个数，F=F(N)是能整除N的素数的个数（注意1不是素数）。试写出D和F的联合分布律，并求边缘分布律。

解　先将试验的样本点及D,F的取值情况列出，如下表

样本点	1	2	3	4	5	6	7	8	9	10
D	1	2	2	3	2	4	2	4	3	4
F	0	1	1	1	1	2	1	1	1	2

所以D的所有可能的取值为1,2,3,4，F的所有可能的取值为0,1,2，容易得到(D,F)取(i,j)的概率(i=1,2,3,4,j=0,1,2)

例如，P{D=1,F=0}=1/10, P{D=2,F=1}=4/10,可得D和F的联合分布律及边缘分布律为

样本点	1	2	3	4	5	6	7	8	9	10
D	1	2	2	3	2	4	2	4	3	4
F	0	1	1	1	1	2	1	1	1	2

F \ D	1	2	3	4	P{F=j}
0	1/10	0	0	0	1/10
1	0	4/10	2/10	1/10	7/10
2	0	0	0	2/10	2/10
P{D=i}	1/10	2/10	3/10	4/10	1

二维连续型随机变量的边缘分布

关于 X 的边缘分布函数为

$$F_X(x) = F(x,+\infty) = \int_{-\infty}^{x}\left[\int_{-\infty}^{+\infty} f(x,y)dy\right]dx$$

与一维连续型随机变量**X**的分布函数

$$F_X(x) = F(x,+\infty) = \int_{-\infty}^{x} f_X(x)dx$$

比较，得**X**的概率密度为 $$f_X(x) = \int_{-\infty}^{+\infty} f(x,y)dy$$

关于 Y 的边缘分布函数为

$$F_Y(x) = F(+\infty, y) = \int_{-\infty}^{y}\left[\int_{-\infty}^{+\infty} f(x,y)dx\right]dy$$

与一维连续型随机变量**Y**的分布函数

$$F_X(y) = F(+\infty, y) = \int_{-\infty}^{y} f_Y(y)dy$$

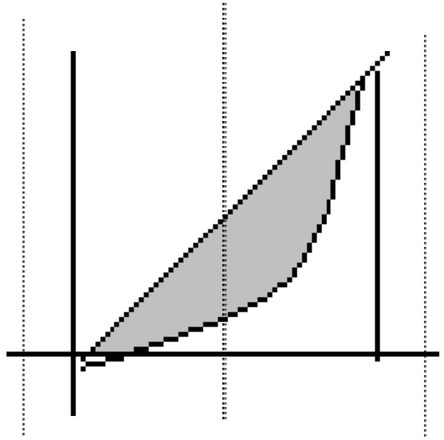
比较，得**Y**的概率密度为 $f_Y(y) = \int_{-\infty}^{+\infty} f(x,y)dx$

定义：分别称

$$f_X(x) = \int_{-\infty}^{+\infty} f(x,y)dy \qquad f_Y(y) = \int_{-\infty}^{+\infty} f(x,y)dx$$

为**(X,Y)**关于**X**和**Y**的边缘概率密度函数。

例3. 设(X, Y)的概率密度为



$$f(x,y)=\begin{cases} c & x^2 \le y < x \\ 0 & others \end{cases}$$

（1）求常数c; (2)求关于X的边缘概率密度

解:(1)由归一性

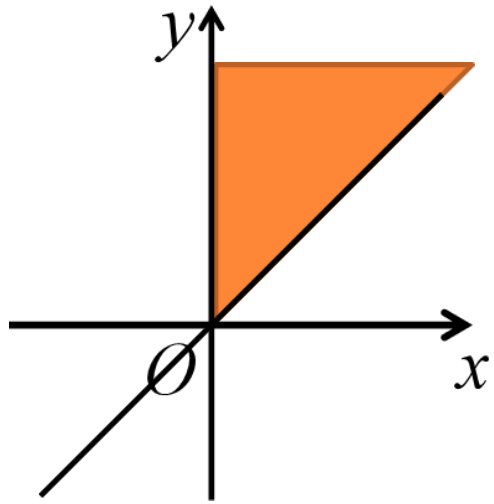
$$\int_0^1 dx\int_{x^2}^{x} c\,dy = 1 \quad \Rightarrow c = 6$$

$$(2)\quad f_X(x)=\int_{-\infty}^{\infty} f(x,y)dy=\begin{cases} 0 & x<0 \ \ or \ \ x>1 \\ \int_{x^2}^{x} 6dy = 6(x-x^2) & 0\le x\le 1 \end{cases}$$

例 设（**X, Y**）的联合概率密度为

$$f(x,y)=\begin{cases} e^{-y} & 0<x<y \\ 0 & \text{其它} \end{cases}$$

求**X**，**Y**的边缘概率密度函数



解

关于**X**的边缘分布密度为

$$f_X(x)=\int_{-\infty}^{+\infty} f(x,y)dy$$

当**x>0**时

$$f_X(x)=\int_{x}^{+\infty} e^{-y}dy=e^{-x}$$

当**x**≤**0**时

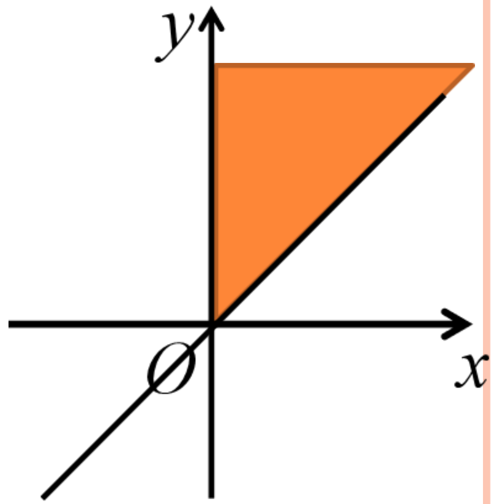
$$f_X(x)=0$$

所以，关于**X**的边缘分布密度为

$$f_X(x)=\begin{cases}e^{-x} & x>0\\ 0 & x\le 0\end{cases}$$

关于**Y**的边缘分布密度为

$$f_Y(y)=\int_{-\infty}^{+\infty}f(x,y)dx$$



当**y≤0**时

$$f_Y(y)=0$$

当**y>0**时

$$f_Y(y)=\int_0^y e^{-y}dx$$

$$=ye^{-y}$$

所以，关于**Y**的边缘分布密度为

$$f_Y(y)=\begin{cases}ye^{-y} & y>0\\ 0 & y\le 0\end{cases}$$

二维正态分布

若 $r.v.(X,Y)$ 的联合为

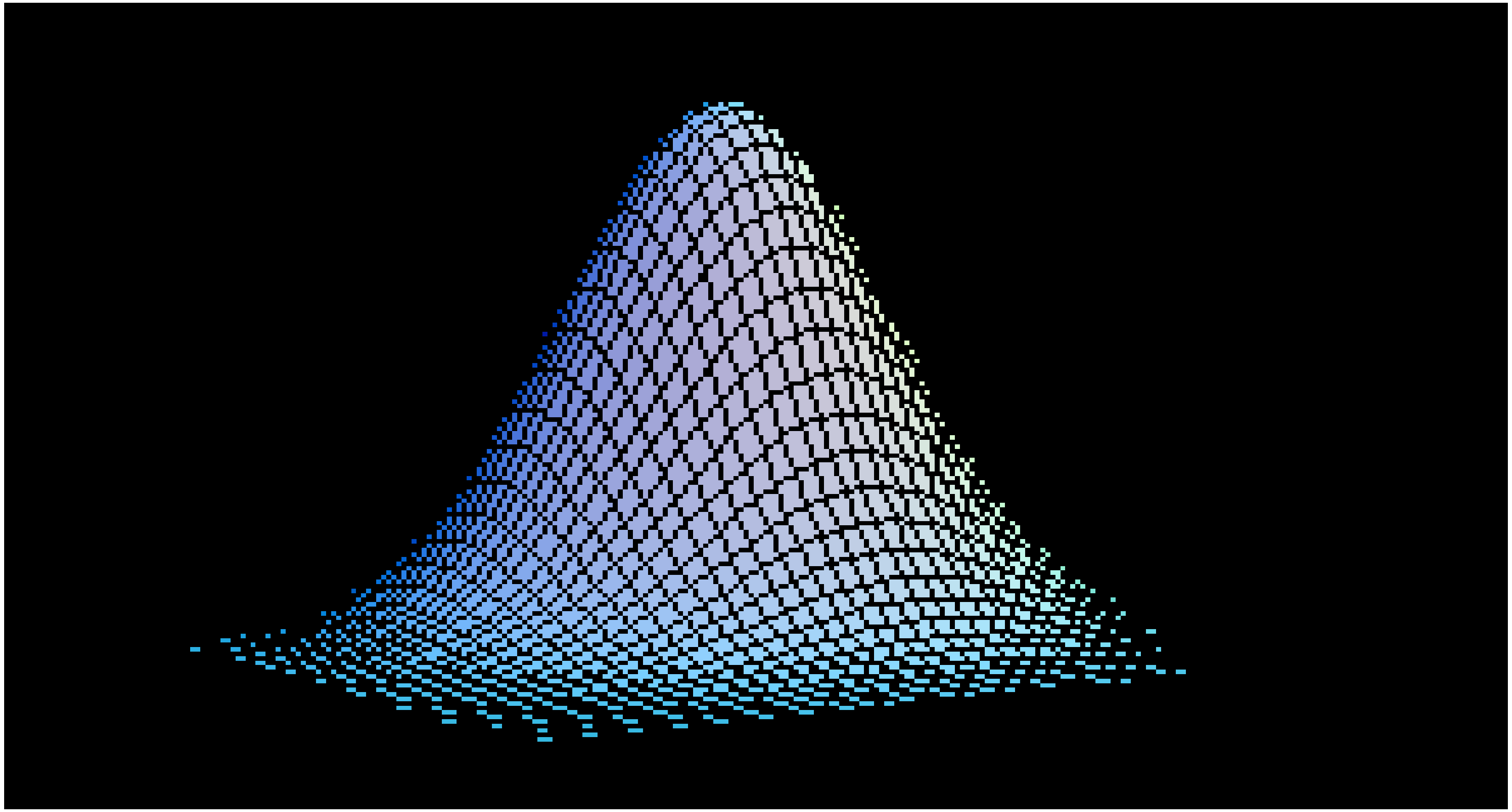
$$f(x,y)=\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\times$$

$$e^{-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_1)^2}{\sigma_1{}^2}-2\rho\frac{(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2}+\frac{(y-\mu_2)^2}{\sigma_2{}^2}\right]}$$

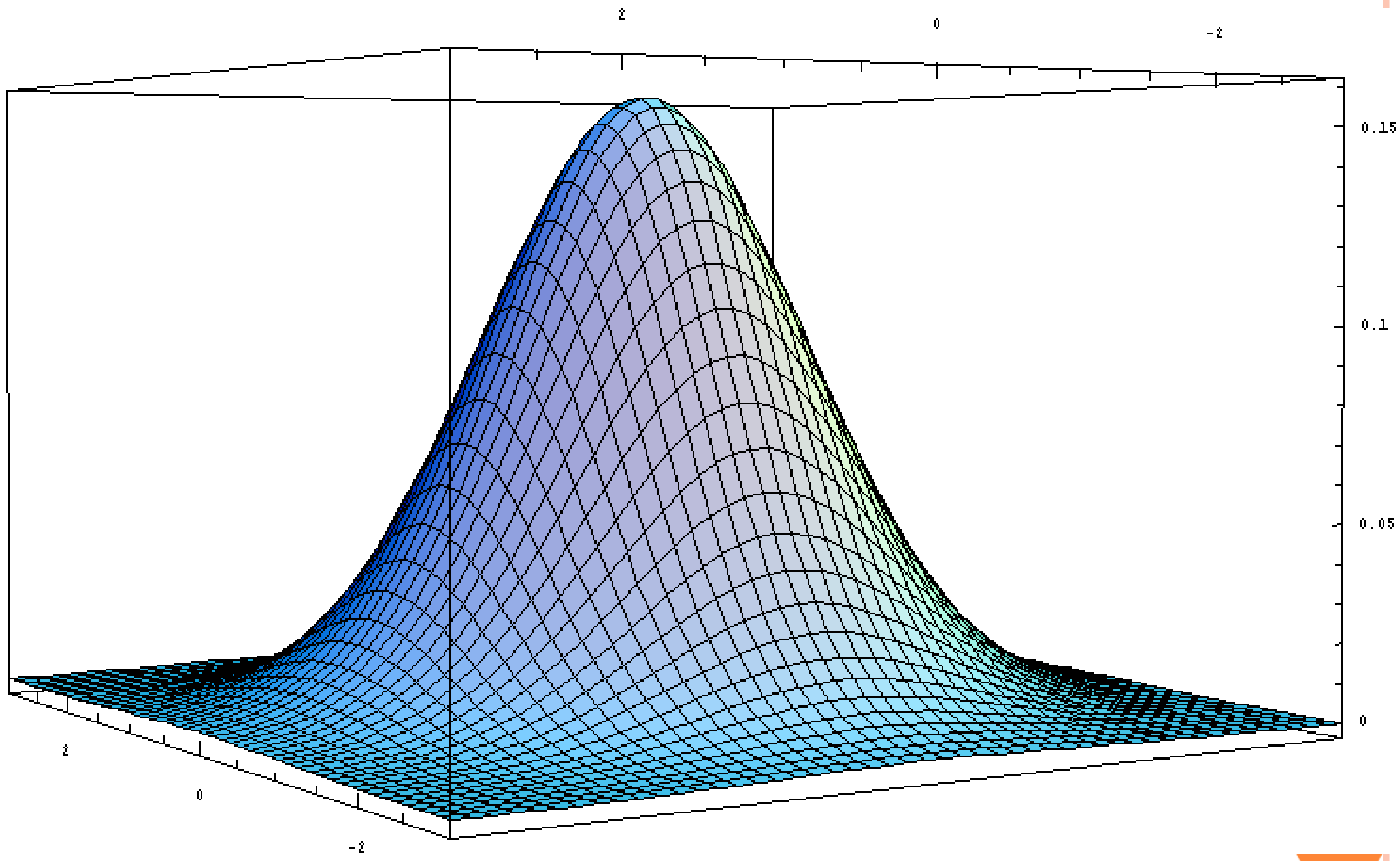
$$-\infty<x<+\infty,-\infty<y<+\infty$$

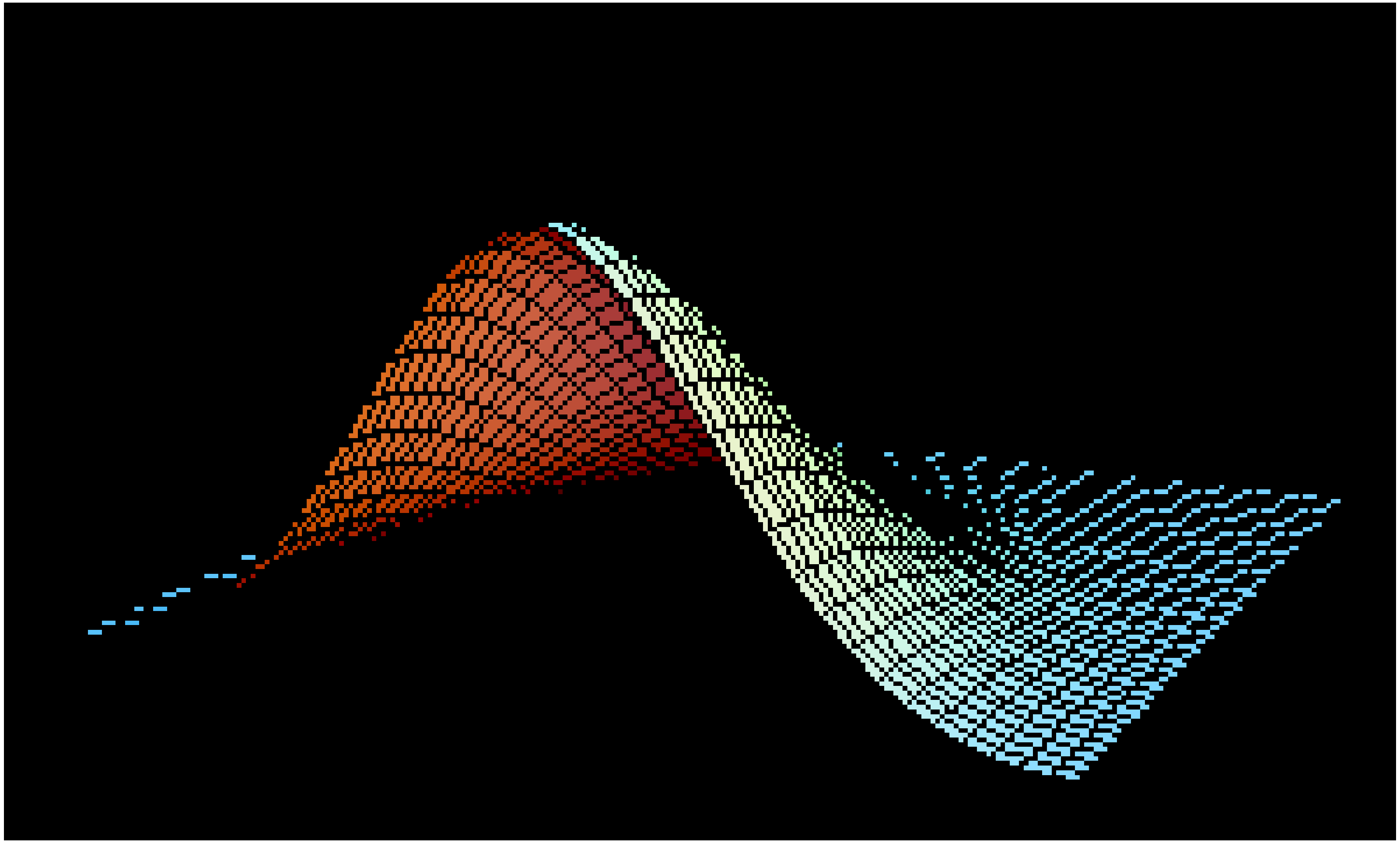
则称 (X,Y) 服从参数为 $\mu_1,\sigma_1^2,\mu_2,\sigma_2^2,\rho$ 的正态分布，记作 $(X,Y)\sim N(\mu_1,\sigma_1^2;\mu_2,\sigma_2^2;\rho)$，

其中 $\sigma_1,\sigma_2>0$，$-1<\rho<1$.



二维正态分布图


2
0
-2
0.15
0.1
0.05
0
2
0
-2



二维正态分布剖面图

正态分布的边缘分布仍为正态分布

$$f_X(x)=\frac{1}{\sqrt{2\pi}\sigma_1}e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}},-\infty<x<+\infty$$

$$f_Y(y)=\frac{1}{\sqrt{2\pi}\sigma_2}e^{-\frac{(y-\mu_2)^2}{2\sigma_2^2}},-\infty<y<+\infty$$

例4 设（X，Y）的联合分布密度函数为

$$f(x,y)=\frac{1}{2\pi}e^{-\frac{x^2+y^2}{2}}(1+\sin x\sin y),\quad -\infty<x,y<+\infty$$

求关于X，Y的边缘分布密度函数

解 关于**X**的分布密度函数为

$$f_X(x)=\int_{-\infty}^{+\infty}f(x,y)dy$$

$$=\int_{-\infty}^{+\infty}\frac{1}{2\pi}e^{-\frac{x^2+y^2}{2}}(1+\sin x\sin y)dy$$

$$=\int_{-\infty}^{+\infty}\frac{1}{2\pi}e^{-\frac{x^2+y^2}{2}}dy+\int_{-\infty}^{+\infty}\frac{1}{2\pi}e^{-\frac{x^2+y^2}{2}}\sin x\sin y\,dy$$

$$=\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}\cdot\int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{y^2}{2}}dy$$

$$+\frac{1}{2\pi}e^{-\frac{x^2}{2}}\sin x\cdot\int_{-\infty}^{+\infty}e^{-\frac{y^2}{2}}\sin y dy$$

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

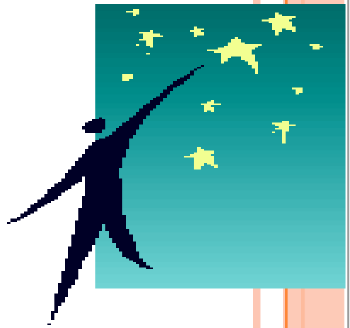
所以，$X\sim N(0,1)$

同理可得 $Y\sim N(0,1)$

不同的联合分布，可有相同的边缘分布。

可见，联合分布可以确定边缘分布，但边缘分布不能确定联合分布

由脚印估计罪犯身高



公安人员根据收集到的罪犯脚印，通过公式

$$身高=脚印长度\times 6.876$$

算出罪犯的身高.这个公式是如何推导出来的?

设一个人身高为 X，脚印长度为 Y. 显然，两者之间是有统计关系的，故应作为二维随机变量 (X,Y) 来研究.

由于影响人类身高与脚印的随机因素是大量的、相互独立的，且各因素的影响又是微小的, 可以叠加的. 故由中心极限定理知 (X,Y) 可以近似看成服从二维正态分布 $N(u_1,\sigma_1^2,u_2,\sigma_2^2;\rho)$.

其中参数 $u_1,\sigma_1^2;u_2,\sigma_2^2;\rho$ 因区域、民族、生活习惯的不同而有所变化，但它们都能通过统计方法而获得.

现已知罪犯的脚印长度为 y，要估计其身高就需计算条件期望，条件密度为

$$f_{X|Y}(x|y)=\frac{f(x,y)}{f_Y(y)}$$

$$=\frac{\sqrt{2\pi}\sigma_2}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\cdot\frac{\exp\{-\frac{1}{2(1-\rho^2)}[\frac{(x-u_1)^2}{\sigma^2}-2\rho\frac{(x-u_1)(y-u_2)}{\sigma_1\sigma_2}+\frac{(y-u_2)^2}{\sigma_2^2}]\}}{\exp\{-\frac{(y-u_2)^2}{2\sigma_2^2}\}}$$

这正是正态分布

$$N(u_1+\rho\frac{\sigma_1}{\sigma_2}(y-u_2),\sigma_1^2(1-\rho^2))$$

的密度函数，因此

$$E(X|Y=y)=u_1+\rho\frac{\sigma_1}{\sigma_2}(y-u_2)$$

如果按中国人的相应参数代入上式，即可得出以脚印长度作自变量的身高近似公式.

第三节　条件分布

问题的提出

在第一章中已经介绍了条件概率的概念，即

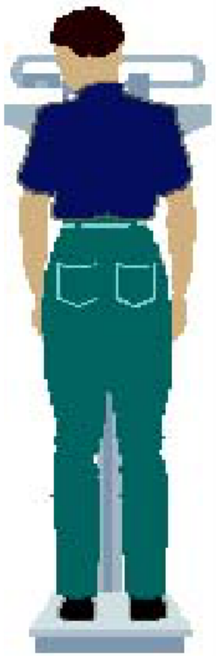
在事件*B*发生的条件下事件 *A* 发生的条件概率:

$$P(A \mid B)=\frac{P(AB)}{P(B)}$$

现设有两个随机变量 *X*, *Y*，若问:在给定*Y*取某个或某些值的条件下，求随机变量*X*的概率分布.

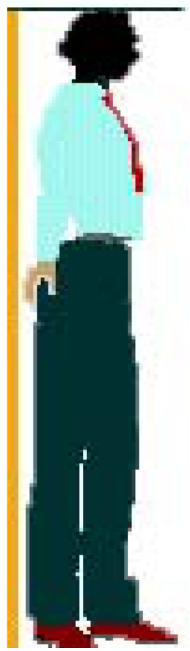
这个分布就是条件分布.

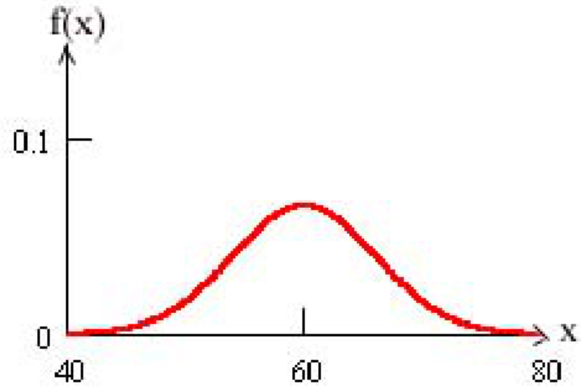
例如: 考虑某大学的全体学生，从其中随机抽取一个学生，分别以X和Y表示其体重和身高．则 X 和Y 都是随机变量，它们都有一定的概率分布.



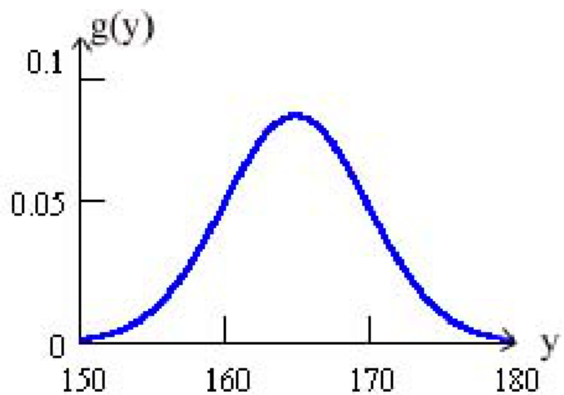
体重X

身高Y





体重X的分布



身高Y的分布

现在若限制 $1.7 < Y < 1.8$ (米)，在这个条件下

求：X 的条件分布

这就意味着要从该校的学生中把身高在1.7米 和1.8米 之间的那些人都挑出来，然后在挑出的学生中求其体重的分布.

显然：

这个分布与不加这个条件时的分布会很不一样.

例如：

在条件分布中体重取大值的概率会显著增加 .

一. 离散型随机变量的条件分布

1. 定义: 若 **(X,Y)** 是二维随机变量,其联合分布律为

$$P\{X=x_i,Y=y_j\}=P_{ij}\ ,\quad i,j=1,2,\cdots,\ \text{(X, Y)}$$

关于 **X** 和 **Y** 的边缘分布律为 $P\{X=x_i\}=P_{i\cdot}$,

$P\{Y=y_j\}=P_{\cdot j}$,且 $P_{i\cdot}>0, P_{\cdot j}>0$

则在事件$\{Y=y_j\}$ 已发生的条件下事件$\{X=x_i\}$发生的概率为:

$$P\{X=x_i \mid Y=y_j\}=\frac{P\{X=x_i,Y=y_j\}}{P\{Y=y_j\}}=\frac{P_{ij}}{P_{\cdot j}}\qquad i=1,2\cdots$$

亦称为**X** 在$\{Y=y_j\}$ 下的**条件分布律.**

同理可定义： y 在条件 $X=x_i$ 下的条件分布律为:

$$P\{Y=y_j \mid X=x_i\}=\frac{P\{X=x_i,Y=y_j\}}{P\{X=x_i\}}=\frac{P_{ij}}{P_{i\cdot}} \quad j=1,2,\cdots$$

2. 性质:

1^0. $P\{X=x_i \mid Y=y_j\}\geq 0$

2^0. $\sum_{i=1}^{\infty} P\{X=x_i \mid Y=y_j\}=\sum_{i=1}^{\infty}\frac{P_{ij}}{P_{\cdot j}}=\frac{1}{P_{\cdot j}}\sum_{i=1}^{\infty} P_{ij}=\frac{1}{P_{\cdot j}}\cdot P_{\cdot j}=1$

3^0. $\sum_{j=1}^{\infty} P\{Y=y_j \mid X=x_i\}=1$

例、在一汽车工厂中，一辆汽车有两道工序是由机器人完成的。其一是紧固3只螺栓，其二是焊接2处焊点。以X表示由机器人紧固的螺栓紧固得不良数目，以Y表示由机器人焊接的不良焊点的数目。根据累计的资料知（X,Y）具有分布律

$X \backslash Y$	0	1	2	3
0	0.840	0.030	0.020	0.010
1	0.060	0.010	0.008	0.002
2	0.010	0.005	0.004	0.001

（1）、求在X=1的条件下，Y的条件分布律；

（2）、求在Y=0的条件下，X的条件分布律；

二. 连续型随机变量的条件分布

引言

在离散型中条件分布律是由条件概率引出的，其中对任意的 $P\{X=x\}>0,\ P\{Y=y\}>0$

但注意到：在一维随机变量的讨论中指出过连续型与离散型随机变量根本区别之一就在于对于连续型随机变量而言 P{X=x}=0，P{Y=y}=0。因此，对于连续型随机变量就无法用条件概率去引出条件分布的概念了。所以必须从分布函数着手并且加以极限的方法引出条件分布的概念.

定义. 给定y，设对任意固定的正数ε>0，极限

$$\lim_{\varepsilon\to 0} P\{X \le x \mid y-\varepsilon < Y \le y+\varepsilon\}$$

$$= \lim_{\varepsilon\to 0} \frac{P\{X \le x, y-\varepsilon < Y \le y+\varepsilon\}}{P\{y-\varepsilon < Y \le y+\varepsilon\}}$$

存在，则称此极限为在条件条件下X的条件分布函数.

记作

$$F_{X|Y}(x \mid y) \equiv P\{X \le x \mid Y = y\}$$

可证当 $f_y(y) \neq 0$ 时

$$F_{X|Y}(x \mid y) = \frac{\int_{-\infty}^{x} f(u,v)du}{f_Y(y)}$$

若记 $f_{X|Y}(x|y)$ 为在Y=y条件下X的条件概率密度，

则由(3.3.3)知,当 $f_Y(y)\neq 0$ 时,

$$f_{X|Y}(x|y)=\frac{\partial F_{X|Y}(x|y)}{\partial x}=\frac{f(x,y)}{f_Y(y)}$$

类似定义，当 $f_X(x)\neq 0$ 时

$$f_{Y|X}(y|x)=\frac{\partial F_{Y|X}(y|x)}{\partial y}=\frac{f(x,y)}{f_X(x)}$$

例. 设随机变量 (X,Y) 的分布密度为：

$$f(x,y)=\begin{cases}3x & 0<x<1,\ 0<y<x\\ 0 & (x,y)\text{ 在其它域}\end{cases}$$

求: (1) X,Y 的边缘分布密度.

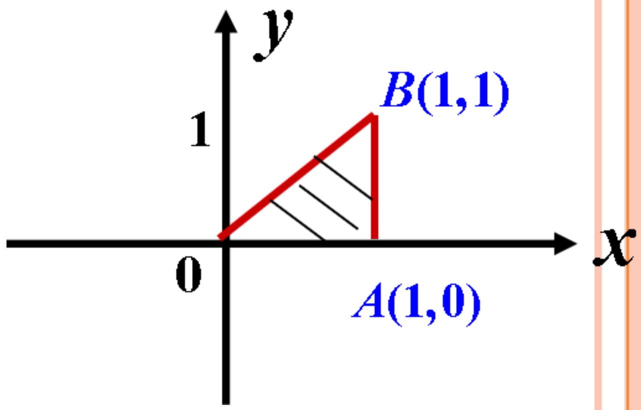
(2) X,Y 的条件分布密度.

(3)求 $P\{\frac{2}{3}>X>0\mid Y=\frac{1}{2}\}$

解： (1) 由已知的$f(x,y)$可知：

当$0<x<1$时

$$f_X(x)=\int_{-\infty}^{+\infty} f(x,y)\,dy=\int_0^x 3x\,dy=3x^2$$



当$x\le 0$或$y\ge 1$时

$$f_X(x)=0$$

$\therefore$ X的边缘概率密度为：

$$f_X(x)=\begin{cases}3x^2 & 0<x<1\\ 0 & 其它\end{cases}$$

同理，当 0 < y <1 时

$$f_Y(y)=\int_{-\infty}^{+\infty} f(x,y)dx=\int_y^1 3x dx=\frac{3}{2}(1-y^2)$$

当 $y\le 0$ 或 $y\ge 1$ 时 $\quad f_Y(y)=0$

$\therefore\quad Y$ 的边缘概率密度为:

$$f_Y(y)=\begin{cases}\frac{3}{2}(1-y^2) & 0<y<1\\ 0 & \text{其它}\end{cases}$$

(2). 依条件概率密度的定义可知:

对于使 $f(x,y), f_X(x), f_Y(y)$ 为非零值的区域有:

$\therefore$ 当$0<y<1$时，在$Y=y$的条件下X的条件概率为

$$\therefore f_{X|Y}(x|y)=\frac{f(x,y)}{f_Y(y)}=\begin{cases}\dfrac{2x}{1-y^2} & y<x<1\\ 0 & \text{其他}x\end{cases}$$

$\therefore$ 当$0<x<1$时，在$X=x$的条件下Y的条件概率为

$$\therefore f_{Y|X}(y|x)=\frac{f(x,y)}{f_X(x)}=\begin{cases}\dfrac{1}{x} & 0<y<x\\ 0 & \text{其他}y\end{cases}$$

$$f_{X|Y}(x|y=\frac{1}{2})=\begin{cases}\frac{8}{3}x & \frac{1}{2}<x<1\\ 0 & \text{其他}\end{cases}$$

$$\therefore P\{\frac{2}{3}>X>0\,|\,Y=\frac{1}{2}\}=\int_0^{\frac{2}{3}} f_{X|Y}(x|y=\frac{1}{2})dx$$

$$=\int_{\frac{1}{2}}^{\frac{2}{3}}\frac{8}{3}xdx=\frac{7}{27}$$

例　设（X, Y) 在圆域 $x^2+y^2\le 1$ 上服从均匀分布

$$f(x,y)=\begin{cases}\dfrac{1}{\pi} & x^2+y^2\le 1\\ 0 & 其它\end{cases}$$

求条件概率密度 $f_{X|Y}(x\mid y)$

解 $f_Y(y)=\int_{-\infty}^{+\infty} f(x,y)dx$

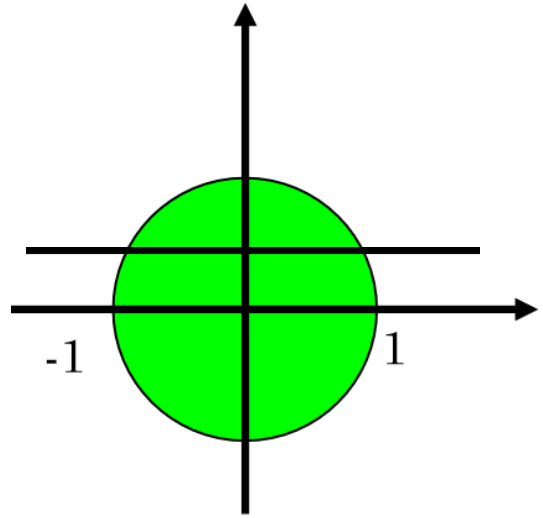
当$y\in[-1,1]$ 时

$$f_Y(y)=\int_{-\sqrt{1-y^2}}^{\sqrt{1-y^2}}\frac{1}{\pi}dx$$

$$=\frac{2}{\pi}\sqrt{1-y^2}$$

当$y\notin[-1,1]$ 时

$$f_Y(y)=0$$



所以，关于**Y**的边缘

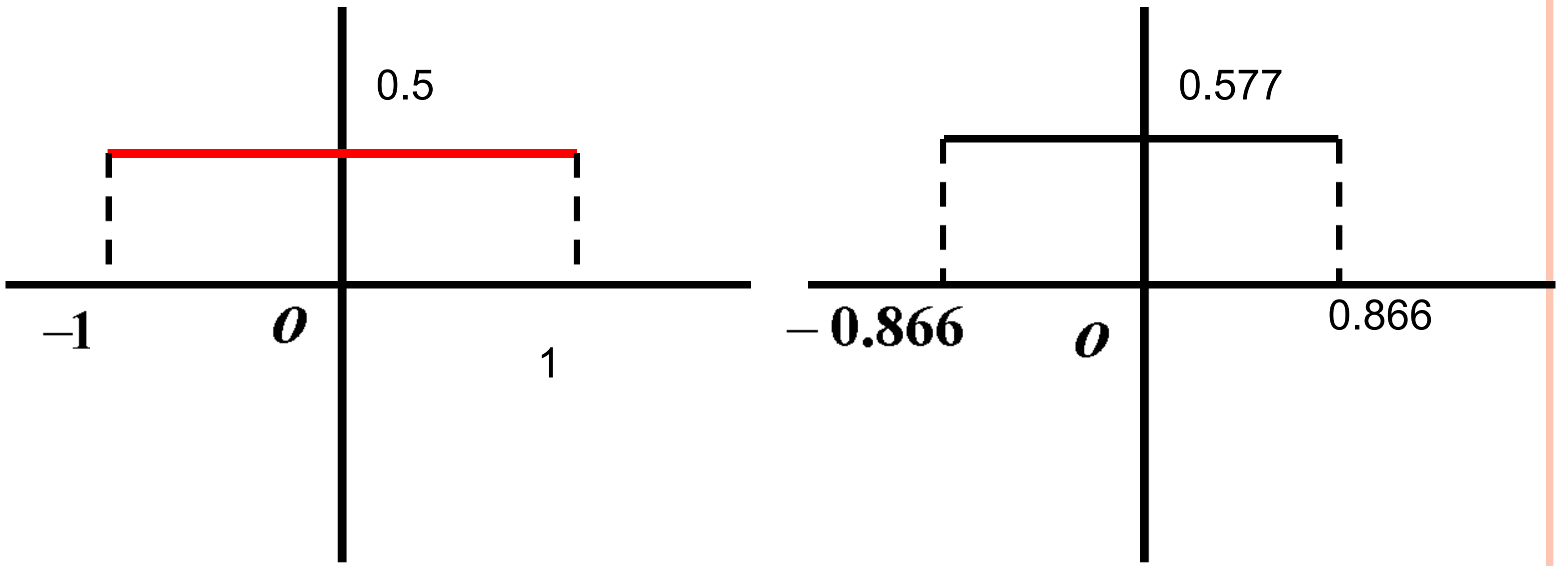
分布密度函数为

$$f_Y(y)=\begin{cases}\frac{2}{\pi}\sqrt{1-y^2} & y\in[-1,1]\\ 0 & \text{其它}\end{cases}$$

$\therefore$ 当$-1 < y < 1$时，在$Y = y$的条件下X的条件概率为

$$f_{X|Y}(x|y) = \frac{f(x,y)}{f_Y(y)} = \begin{cases} \dfrac{1}{2\sqrt{1-y^2}} & -\sqrt{1-y^2} \le x \le \sqrt{1-y^2}, \\ 0 & \text{其他}x \end{cases}$$

当$y = 0$和$y = \dfrac{1}{2}$时$f_{X|Y}(x|y)$的图形分别如下图所示.


0.5
–1
0
1
0.577
– 0.866
0
0.866