软件工程统计方法

随机变量及其分布（三）

陈振宇

南京大学软件学院

Email:zychen@software.nju.edu.cn

Homepage:software.nju.edu.cn/zychen

内 容 提 纲

- 统计学导论
- 描述统计
- 概率计算基础
- 随机变量及其分布
- 统计量及其抽样分布
- 参数估计
- 参数假设检验
- 非参数假设检验
- 方差分析
- 回归分析

1 本节内容

- 随机变量
- 分布函数
- 数学期望
- 方差
- 常用分布
- 二维随机向量
- 随机变量函数的分布

2 二维随机向量

在实际问题中，对于某些随机试验的结果需要同时用两个或两个以上的随机变量来描述. 例如，为了研究某一地区学龄前儿童的发育情况，对这一地区的儿童进行抽查，对于每个儿童都能观察到他的身高H和体重W. 在这里，样本空间$S = \{e\} = \{$某地区的全部学龄前儿童$\}$，而$H(e)$和$W(e)$是定义在S上的两个随机变量.

Definition 1 (二维随机向量) 设实验E的样本空间为$S = \{w\}, X = X(w)$和$Y = Y(w)$是定义在S上的随机变量，由它们构成的一个向量(X, Y)叫做二维随机变量或二维随机向量.

二维离散随机向量

Definition 2 (联合分布) 设(X,Y)是二维随机向量, x,y是任意实数, 称二元函数

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

为二维随机向量(X,Y)的联合分布函数.

请注意:

$$P\{x_1<X\le x_2,y_1<Y\le y_2\}=F(x_2,y_2)-F(x_2,y_1)-F(x_1,y_2)+F(x_1,y_1)$$

二维随机变量

Definition 3 (二维离散型随机变量) 若二维随机变量(X,Y)的可能值(x_i,y_i)只有有限对或可列无限对, 则称(X,Y)是离散型二维随机变量.

(X,Y)是离散型二维随机变量⇔X和Y都是离散型随机变量.

Definition 4 (联合分布) 称$P\{X=x_i,Y=y_j\}p_{ij},i,j=1,2,\cdots$为$(X,Y)$的联合分布律(概率函数). 且满足

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

离散二维随机变量的联合分布律通常列成一个二维表.

Definition 5 (分布函数) 二维随机变量(X,Y)的分布函数定义为

$$F(x,y)=\sum_{x_i\le x}\sum_{y_j\le y}p_{ij}$$

二维连续随机变量

Definition 6 (二维连续型随机变量) 对任意(X,Y),如果存在非负函数$f(x,y)$,使对任意实数对(x,y)有$F(x,y) = \int_{-\infty}^{y}\int_{-\infty}^{x} f(u,v)dudv$. 则称$(X,Y)$为二维连续型随机变量, 其中$f(x,y)$称为$(X,Y)$的联合概率密度函数.

二维随机变量具有以下性质:

- $f(x,y) \geq 0$
- $F(-\infty,-\infty) = 0, F(+\infty,+\infty) = 1$
- F在(x,y)点连续, 则
$$\frac{\partial^2 F(x,y)}{\partial x \partial y} = f(x,y)$$
- 在任意平面G上的概率为

$$P\{(X,Y) \in G\} = \int\int_G f(x,y)dxdy$$

二维随机变量

$$F(x,y) = 1 - e^{-0.01x} - e^{-0.01y} + e^{-0.01(x+y)}, x \geq 0, y \geq 0$$

其它$F(x,y) = 0$。

求

(1) $P(X < 120, Y < 120)$

(2) $P(X > 120, Y > 120)$

(3) $P(Y \leq X)$

$$(1)P(X < 120, Y < 120) = F(120,120) = 1 - e^{-1.2} - e^{-1.2} + e^{-2.4}$$

$$= (1 - e^{-1.2})^2$$

二维随机变量

$$F(x,y)=1-e^{-0.01x}-e^{-0.01y}+e^{-0.01(x+y)},x\geq 0,y\geq 0$$

$$\begin{aligned}(2)P(X>120,Y>120)\\ =F(+\infty,+\infty)-F(120,+\infty)-F(+\infty,120)+F(120,120)\\ =1-(1-e^{-1.2})-(1-e^{-1.2})+(1-e^{-1.2})^2=e^{-2.4}\end{aligned}$$

$$(3)f(x,y)=\frac{\partial^2 F}{\partial x\partial y}=(0.01)^2e^{-0.01(x+y)}$$

$$P(Y\leq X)=\int\int_{y\leq x}f(x,y)dxdy=\int_{-\infty}^{+\infty}dx\int_{-\infty}^{x}f(x,y)dy$$

$$=\int_0^{+\infty}dx\int_0^{x}(0.01)^2e^{-0.01(x+y)}dy=\int_0^{+\infty}(-0.01e^{-0.02x}+0.01e^{-0.01x})dx$$

$$=(0.5e^{-0.02x}-e^{-0.01x})|_0^{+\infty}=(0-0.5)-(0-1)=0.5$$

3 边缘分布

Definition 7 设(X,Y)为二维随机变量，则称随机变量X的概率分布为(X,Y)关于X的边缘分布；随机变量Y的概率分布为(X,Y)关于Y的边缘分布，其分布函数，密度函数和分布律分别记为：$F_X(x)$，$F_Y(y)$；$f_X(x)$，$f_Y(y)$；$p_{i\cdot}$，$p_{\cdot j}$.

对于离散二维随机变量(X,Y)，有

$$F_X(x)=F(x,\infty)=\sum_{x_i\le x}\sum_{j=1}^{\infty}p_{ij},F_Y(y)=F(\infty,y)=\sum_{i=1}^{\infty}\sum_{y_j\le y}p_{ij}$$

$$p_{i\cdot}=P\{X=x_i\}=\sum_{j=1}^{\infty}p_{ij},p_{\cdot,j}=P\{Y=y_j\}=\sum_{i=1}^{\infty}p_{ij}$$

对于连续二维随机变量(X,Y)，有

$$F_X(x)=\int_{-\infty}^{x}\int_{-\infty}^{\infty}f(x,y)dxdy,F_Y(y)=\int_{-\infty}^{\infty}\int_{-\infty}^{y}f(x,y)dxdy$$

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

4　二维随机变量条件概率

对于多个随机事件可以讨论它们的条件概率, 同样地, 对于多个随机变量也可以讨论它们的条件分布.
设(X,Y)是二维离散型随机变量, 其分布率为$P\{X=x_i, Y=y_j\}=p_{ij}$, 其边缘概率分别为$p_{i\cdot}, p_{\cdot j}$. 则条件概率定义为

$$P\{X=x_i \mid Y=y_j\}=\frac{P\{X=x_i, Y=y_j\}}{P\{Y=y_j\}}=\frac{p_{ij}}{p_{\cdot j}}$$

$$P\{Y=y_j \mid X=x_i\}=\frac{P\{X=x_i, Y=y_j\}}{P\{X=x_i\}}=\frac{p_{ij}}{p_{i\cdot}}$$

设(X,Y)是二维连续型随机变量, 其概率密度为$f(x,y)$, 其边缘概率密度分别为$f_X(x)$, $f_Y(y)$. 则条件概率密度定义为

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

其条件概率分布定义为

$$F_{X|Y}(x|y)=P\{X\le x|Y=y\}=\int_{-\infty}^{x}\frac{f(x,y)}{f_Y(y)}dx$$

$$F_{Y|X}(y|x)=P\{Y\le y|X=x\}=\int_{-\infty}^{y}\frac{f(x,y)}{f_X(x)}dy$$

例 子

Example 1 对一群人进行吸烟X和身体健康Y调查. $X=1$健康, $X=0$一般, $X=-1$不健康; $Y=0$不吸烟, $Y=1$每天不多于*15*支, $Y=2$每天多于*15*支. *(X,Y)*的联合分布律如下*:*

$X\|Y$	*0*	*1*	*2*
1	*0.35*	*0.04*	*0.025*
0	*0.025*	*0.15*	*0.04*
-1	*0.02*	*0.1*	*0.25*

*(1)*试求X, Y的边缘分布律*; (2)*试求$P(X=-1|Y=2)$的值.

解: (1)X,Y的边缘分布律分别为:

X	1	0	-1	Y	0	1	2
$p_{i\cdot}$	0.415	0.215	0.37	$p_{\cdot j}$	0.395	0.290	0.315

(2)

$$P(X=-1|Y=2)=\frac{0.25}{0.315}=0.794$$

例 子

Example 2 设X在(0,1)上随机取值. 当观察到$X = x(0 < x < 1)$时, Y在区间$(x, 1)$上均匀分布, 求Y的概率密度$f_Y(y)$.

解: 对于任意$x(0 < x < 1)$, 在$X = x$的条件下, Y的条件概率密度为

$$f_{Y|X}(y|x) = \frac{1}{1-x}, x < y < 1; f_{Y|X}(y|x) = 0, \text{其它}$$

故(X, Y)的概率密度为

$$f(x, y) = f_X(x) f_{Y|X}(y|x) = 1 \cdot \frac{1}{1-x}$$

所以Y的边缘概率度为

$$f_Y(y) = \int_{-\infty}^{\infty} f(x, y) dx = -\ln(1-y), 0 < y < 1$$

5 二维随机变量独立性

Definition 8 设$F(x,y)$及$F_X(x)$, $F_Y(y)$分别是二维随机变量(X,Y)的分布函数及边缘分布函数. 若对所有x,y有

$$P\{X \leq x, Y \leq y\} = P\{X \leq x\}P\{Y \leq y\}$$

则称随机变量X和Y是相互独立的. 等价命题有

$$F(x,y) = F_X(x)F_Y(y)$$

$$f(x,y) = f_X(x)f_Y(y)$$

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

对于连续, 只要条件几乎出出成立即可.

例 子

Example 3 X, Y服从同一分布, 其分布律为

X, Y	-1	0	1
p	1/4	1/2	1/4

已知$P(X = Y) = 0$, 判断X, Y是否相关, 是否独立.

解: 求X, Y的联合概率和边缘概率

$X \mid Y$	-1	0	1	$p_{\cdot j}$
-1	0	1/4	0	1/4
0	1/4	0	1/4	1/2
1	0	1/4	0	1/4
$p_{i\cdot}$	1/4	1/2	1/4	

$$E(X) = E(Y) = -1 * 1/4 + 0 * 1/2 + 1 * 1/4 = 0$$

$$E(XY) = (-1) * (0) * 1/4 + 0 * (-1) * 1/4 + 1 * 0 * 1/4 + 1 * 0 * 1/4 = 0$$

所以$COV(X, Y) = 0$, X, Y不相关.

$$p_{-1,-1} = 0 \neq p_{-1\cdot} p_{\cdot -1} = 1/4 * 1/4$$

所以X, Y不独立.

例 子

Example 4 (X,Y)的概率密度如下, 问X,Y是否独立?

$$f(x,y)=6e^{-(2x+3y)},x>0,y>0$$

解: X,Y的边缘概率密度分别为:

$$f_X(x)=\int_{-\infty}^{\infty}6e^{-(2x+3y)}dy=2e^{-2x},x>0$$

$$f_Y(y)=\int_{-\infty}^{\infty}6e^{-(2x+3y)}dx=2e^{-3y},y>0$$

$$f(x,y)=f_X(x)f_Y(y)$$

所以X,Y相互独立.

例 子

Example 5 设X, Y相互独立, 已知(X, Y)的联合分布律如下

p_{ij}	0	1	2	$p_{i\cdot}$
1	0.01	0.2		
2	0.03			
$p_{\cdot j}$				

求未知概率值.

例 子

Example 6 设X,Y是两个相互独立的随机变量,X在*(0,1)*上服从均匀分布,Y的概率密度为*:*

$$f_Y(y) = \frac{1}{2}e^{-\frac{y}{2}}, y > 0$$

1. 求$f(x,y)$

2. 设有a的二次方程$a^2 + 2aX + Y = 0$,求此方程有实根的概率.

$$(1) f(x,y) = f_X(x) f_Y(y) = \frac{1}{2}e^{-\frac{y}{2}}, 0 < x < 1, y > 0$$

$$(2) P(X^2 \geq Y) = \int\int_{X^2 \geq Y} \frac{1}{2}e^{-\frac{y}{2}} dxdy$$

$$= \int_0^1 dx \int_0^{x^2} \frac{1}{2}e^{-\frac{y}{2}} dy = \int_0^1 (1 - e^{-\frac{x^2}{2}}) dx$$

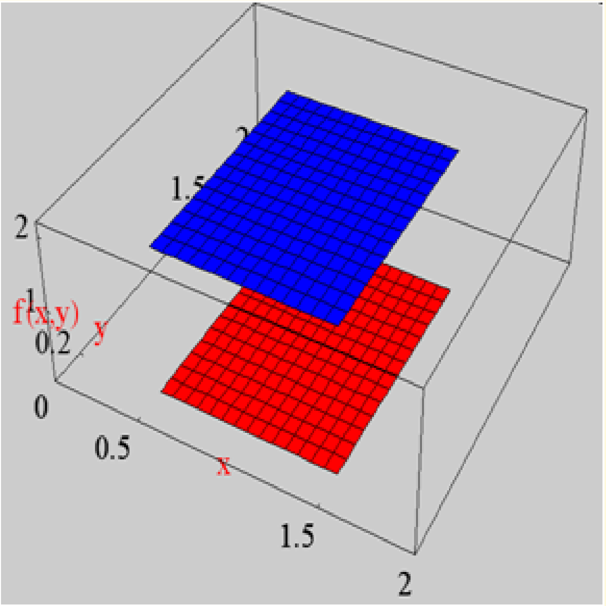
$$= 1 - \int_0^1 (e^{-\frac{x^2}{2}}) dx = 1 - \sqrt{2\pi}(\Phi(1) - \Phi(0)) = 0.1448$$

6 二维均匀分布

Definition 9 设G是平面上的有界区域,其面积为A.若二维随机变量(X,Y)具有概率密度

$$f(x,y)=\frac{1}{A},(x,y)\in G$$

其它$f(x,y)=0$, 称(X,Y)在G上的二维均匀分布。



二维均匀分布

Example 7 二维随机变量(X,Y)是在$x^2+y^2\le 1$上的均匀分布,即

$$f(x,y)=\frac{1}{\pi},x^2+y^2\le 1$$

其它$f(x,y)=0$, 求$f_{X|Y}(x|y)$。

先求边缘密度函数。因$x^2+y^2\le 1$时，$f(x,y)=\frac{1}{\pi}$，所以

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

$$f_{X|Y}(x|y)=\frac{f(x,y)}{f_Y(y)}=\frac{\frac{1}{\pi}}{\frac{2}{\pi}\sqrt{1-y^2}}$$

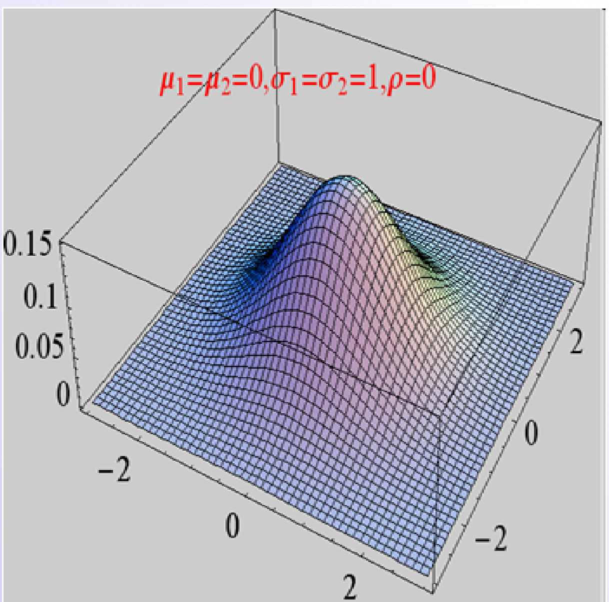
$$f_{X|Y}(x|y)=\frac{1}{2\sqrt{1-y^2}},-\sqrt{1-y^2}\le x\le\sqrt{1-y^2}$$

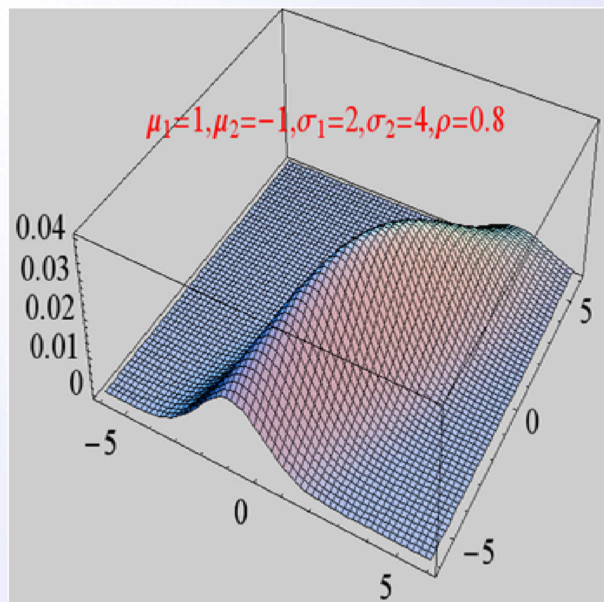
7 二维正态分布

Definition 10 如果随机变量(X,Y)的概率密度为:

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

$-\infty<x,y<\infty$. 则称(X,Y)服从参数为$\mu_1,\mu_2,\sigma_1,\sigma_2,\rho$的二维正态分布, 记为$(X,Y)\sim N(\mu_1,\mu_2,\sigma_1^2,\sigma_2^2,\rho)$.





二维正态分布

求X, Y的边缘密度函数.

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)dy = \int_{-\infty}^{\infty} \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}[\frac{(x-\mu_1)^2}{\sigma_1^2} - 2\rho\frac{(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2} + \frac{(y-\mu_2)^2}{\sigma_2^2}]} dy$$

$$= \int_{-\infty}^{\infty} \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}} e^{-\frac{1}{2(1-\rho^2)}[\frac{y-\mu_2}{\sigma_2} - \rho\frac{x-\mu_1}{\sigma_1}]^2} dy$$

$$= \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma_2\sqrt{1-\rho^2}} e^{-frac[y-(\mu_2+\rho\frac{\sigma_2}{\sigma_1}(x-\mu_1))]^2 2\sigma_2^2(1-\rho^2)} dy$$

$$= \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}}$$

同理可得

$$f_Y(y) = \frac{1}{\sqrt{2\pi}\sigma_2} e^{-\frac{(y-\mu_2)^2}{2\sigma_1^2}}$$

8 数字特征

$$E(X+Y)=E(X)+E(Y)$$

证明:

$$E(X+Y)=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}(x+y)f(x,y)dxdy$$

$$=\int_{-\infty}^{+\infty}xf(x,y)dxdy+\int_{-\infty}^{+\infty}yf(x,y)dxdy=E(X)+E(Y)$$

X和Y独立

$$E(XY)=E(X)E(Y)$$

证明:

$$E(XY)=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}(xy)f(x,y)dxdy=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}(xy)f_X(x)f_Y(y)dxdy$$

$$=\int_{-\infty}^{+\infty}xf_X(x)dx\int_{-\infty}^{+\infty}yf_Y(y)dy=E(X)E(Y)$$

数 字 特 征

X和Y独立

$$D(X+Y)=D(X)+D(Y)$$

证明：

$$D(X+Y)=E(((X+Y)-E(X+Y))^2)=E(((X-E(X))+(Y-E(Y)))^2)$$

$$=E((X-E(X))^2)+E((Y-E(Y))^2)+2E((X-E(X))(Y-E(Y)))$$

X和Y独立，则$X-E(X)$和$Y-E(Y)$独立，所以$E((X-E(X))(Y-E(Y)))=E(X-EX)E(Y-EY)=0$,

$$D(X+Y)=D(X)+D(Y)$$

9 协方差

Definition 11 (协方差) $E\{[(X-EX)][(Y-EY)]\}$称为随机变量X和Y的协方差$Cov(X,Y)$, 即$Cov(X,Y)=E\{[(X-EX)][(Y-EY)]\}$.

将定义式展开, 易得: $Cov(X,Y)=E(XY)-E(X)E(Y)$.

协方差的性质:

1. $Cov(X,Y)=Cov(Y,X)$;
2. $Cov(aX,bY)=abCov(X,Y), a,b$为常数;
3. $Cov(X_1+X_2,Y)=Cov(X_1,Y)+Cov(X_2,Y)$.

相关系数

Definition 12 (相关系数) 设随机变量X, Y的数学期望、方差都存在, 称

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

为随机变量X, Y的相关系数.

相关系数的两条重要性质:

1. $|\rho_{XY}| \leq 1$;
2. $|\rho_{XY}| = 1$的充要条件为两个随机变量X和Y有线性关系, 即关系式$Y = aX + b$成立的概率为1(a, b为常数).

10　随机变量的函数分布

在一些应用中，我们直接得到的随机变量X，通常进行函数变换$Y = g(X)$才能得到我们需要的随机变量。例如我们需要测量圆的面积Y，通常是测量半径X，然后经过$Y = \pi X^2$计算圆的面积。X的概率密度函数已知，求问Y的概率密度函数。

随机变量函数分布的基本解法。

- 若Y为离散随机变量，先写出Y的可能取值：$y_1, \cdots, y_n$，再找出$Y = y_i$的等价事件$X \in D$，求得$P(Y = y_i) = P(X \in D)$。
- 若Y为连续随机变量，先写出Y的分布函数：$F_Y(y) = P_Y(Y \le y)$，找出$Y \le y$的等价事件$(X \in D)$，得$F_Y(y) = P(X \in D)$，再求出Y的概率密度函数$f_Y(y)$。

随机变量的函数分布

Example 8 假设随机变量X的分布律如下：

x	*-1*	*0*	*1*
$f_X(x)$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

令$Y = X^2$, 求Y的概率分布律。
解：

$$P(Y = 0) = P(X = 0) = \frac{1}{2}$$

,

$$P(Y = 1) = P(X = 1) + P(X = -1) = \frac{1}{2}$$

即Y的概率分布律为

y	*0*	*1*
$f_Y(y)$	$\frac{1}{2}$	$\frac{1}{2}$

随机变量的函数分布

Example 9 设随机变量X的概率密度为

$$f_X(x) = \frac{x}{8}, 0 < x < 4$$

求$Y = X^2$的概率密度函数。
解：求Y的分布函数

$$F_Y(y) = P\{Y \leq y\} = p\{X^2 \leq Y\} = P\{-\sqrt{y} < X < \sqrt{y}\}$$

当$y \leq 0$时，$F_Y(y) = 0$; 当$y \geq 16$时，$F_Y(y) = 1$; 当$0 < y < 16$时，

$$F_Y(y) = P\{0 < X < \sqrt{y}\} = F_X(\sqrt{y}) = \int_{-\infty}^{\sqrt{y}} f_X(t)dt$$

当$f(x)$连续时，有$\frac{d}{dx}\int_a xf(t)dt = f(x)$，$\frac{d}{dx}\int_a g(x)f(t)dt = f(g(x))g'(x)$，可得

$$f_Y(y) = \frac{1}{2\sqrt{y}} f_X(\sqrt{y}) = \frac{1}{16}, 0 < y < 16$$

内 容 回 顾

- ❑ 随机变量、概率分布、概率密度
- ❑ 二项分布、泊松分布、正态分布
- ❑ 二维随机变量
- ❑ 数学期望、方差