第三节　协方差及相关系数

一、协方差与相关系数的概念及性质

二、相关系数的意义

三、小结

一、协方差与相关系数的概念及性质

1. 问题的提出

若随机变量 X 和 Y 相互独立,那么

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

若随机变量 X 和 Y 不相互独立

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

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

$$=D(X)+D(Y)+2E\{[X-E(X)][Y-E(Y)]\}.$$

协方差

2. 定义

量 $E\{[X-E(X)][Y-E(Y)]\}$ 称为随机变量 X 与 Y 的协方差. 记为 $\mathrm{Cov}(X,Y)$, 即

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

而

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

称为随机变量 X 与 Y 的相关系数.

3. 说明

(1) X 和 Y 的相关系数又称为标准协方差,它是一个无量纲的量.

(2) 若随机变量 X 和 Y 相互独立

$$\Rightarrow \mathrm{Cov}(X,Y) = E\{[X-E(X)][Y-E(Y)]\}$$
$$= E[X-E(X)]E[Y-E(Y)]$$
$$= 0.$$

(3) 若随机变量 X 和 Y 相互独立

$$\Rightarrow D(X+Y) = D(X)+D(Y)$$
$$+2E\{[X-E(X)][Y-E(Y)]\}$$
$$= D(X)+D(Y)+2\mathrm{Cov}(X,Y) = D(X)+D(Y).$$

4. 协方差的计算公式

(1) $\mathrm{Cov}(X,Y)=E(XY)-E(X)E(Y)$;

(2) $D(X+Y)=D(X)+D(Y)+2\mathrm{Cov}(X,Y)$.

证明 (1) $\mathrm{Cov}(X,Y)=E\{[X-E(X)][Y-E(Y)]\}$

$=E[XY-YE(X)-XE(Y)+E(X)E(Y)]$

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

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

$$(2)D(X+Y)=E\{[(X+Y)-E(X+Y)]^2\}$$

$$=E\{[(X-E(X))+(Y-E(Y)]^2\}$$

$$=E\{[X-E(X)]^2\}+E\{[Y-E(Y)]^2\}$$

$$+2E\{[X-E(X)][Y-E(Y)]\}$$

$$=D(X)+D(Y)+2\mathrm{Cov}(X,Y).$$

5. 性质

(1) $\mathrm{Cov}(X,Y)=\mathrm{Cov}(Y,X)$;

(2) $\mathrm{Cov}(aX,bY)=ab\,\mathrm{Cov}(X,Y)$, a,b 为常数;

(3) $\mathrm{Cov}(X_1+X_2,Y)=\mathrm{Cov}(X_1,Y)+\mathrm{Cov}(X_2,Y)$.

例1 设$(X,Y)\sim N(\mu_1,\mu_2,\sigma_1^2,\sigma_2^2,\rho)$,试求 X 与 Y 的相关系数.

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

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

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

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

$$\Rightarrow E(X)=\mu_1,\ E(Y)=\mu_2,\ D(X)=\sigma_1^2,\ D(Y)=\sigma_2^2.$$

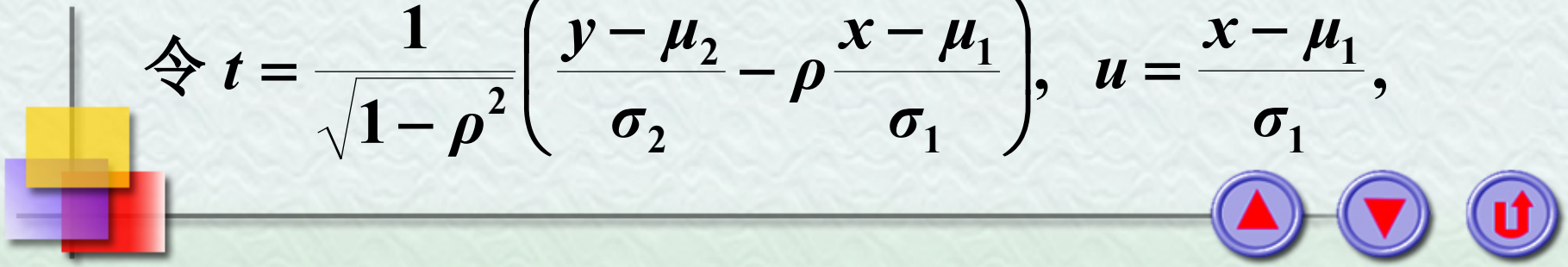
而

$$\mathrm{Cov}(X,Y)=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}(x-\mu_1)(y-\mu_2)f(x,y)\,\mathrm{d}x\,\mathrm{d}y$$

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

$$\cdot\,\mathrm{e}^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}}\,\mathrm{e}^{-\frac{1}{2(1-\rho^2)}\left[\frac{y-\mu_2}{\sigma_2}-\rho\frac{x-\mu_1}{\sigma_1}\right]^2}\mathrm{d}y\,\mathrm{d}x.$$

令 $t=\dfrac{1}{\sqrt{1-\rho^2}}\left(\dfrac{y-\mu_2}{\sigma_2}-\rho\dfrac{x-\mu_1}{\sigma_1}\right),\quad u=\dfrac{x-\mu_1}{\sigma_1},$



$$\mathrm{Cov}(X,Y)$$

$$=\frac{1}{2\pi}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}(\sigma_1\sigma_2\sqrt{1-\rho^2}tu+\rho\sigma_1\sigma_2u^2)\mathrm{e}^{-\frac{u^2}{2}-\frac{t^2}{2}}\,\mathrm{d}t\,\mathrm{d}u$$

$$=\frac{\rho\sigma_1\sigma_2}{2\pi}\left(\int_{-\infty}^{+\infty}u^2\mathrm{e}^{-\frac{u^2}{2}}\,\mathrm{d}u\right)\left(\int_{-\infty}^{+\infty}\mathrm{e}^{-\frac{t^2}{2}}\,\mathrm{d}t\right)$$

$$+\frac{\sigma_1\sigma_2\sqrt{1-\rho^2}}{2\pi}\left(\int_{-\infty}^{+\infty}u\mathrm{e}^{-\frac{u^2}{2}}\,\mathrm{d}u\right)\left(\int_{-\infty}^{+\infty}t\mathrm{e}^{-\frac{t^2}{2}}\,\mathrm{d}t\right)$$

$$=\frac{\rho\sigma_1\sigma_2}{2\pi}\sqrt{2\pi}\cdot\sqrt{2\pi},$$

故有 $\mathrm{Cov}(X,Y)=\rho\sigma_1\sigma_2.$

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

结论

(1) 二维正态分布密度函数中,参数 ρ 代表了 X 与 Y 的相关系数;

(2) 二维正态随机变量 X 与 Y 相关系数为零等价于 X 与 Y 相互独立.

例2 已知随机变量X,Y分别服从$N(1,3^2),N(0,4^2)$，$\rho_{XY}=-1/2$，设$Z=X/3+Y/2$.

(1) 求Z的数学期望和方差.

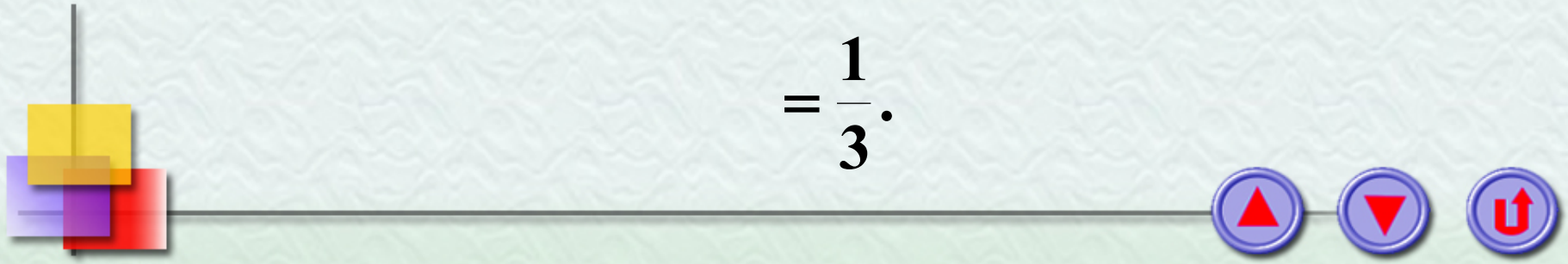
(2) 求X与Z的相关系数.

(3) 问X与Z是否相互独立？为什么？

解 (1)由$E(X)=1, D(X)=9, E(Y)=0, D(Y)=16.$

得 $$E(Z)=E(\frac{X}{3}+\frac{Y}{2})=\frac{1}{3}E(X)+\frac{1}{2}E(Y)$$

$$=\frac{1}{3}.$$



$$D(Z)=D(\frac{X}{3})+D(\frac{Y}{2})+2\mathrm{Cov}(\frac{X}{3},\frac{Y}{2})$$

$$=\frac{1}{9}D(X)+\frac{1}{4}D(Y)+\frac{1}{3}\mathrm{Cov}(X,Y)$$

$$=\frac{1}{9}D(X)+\frac{1}{4}D(Y)+\frac{1}{3}\rho_{XY}\sqrt{D(X)}\sqrt{D(Y)}$$

$$=1+4-2=3.$$

$$(2)\ \mathrm{Cov}(X,Z)=\mathrm{Cov}(X,\frac{X}{3}+\frac{Y}{2})$$

$$=\frac{1}{3}\mathrm{Cov}(X,X)+\frac{1}{2}\mathrm{Cov}(X,Y)$$

$$=\frac{1}{3}D(X)+\frac{1}{2}\rho_{XY}\sqrt{D(X)}\sqrt{D(Y)}=3-3=0.$$

故 $\rho_{XY}=\mathrm{Cov}(X,Z)/(\sqrt{D(X)}\sqrt{D(Z)})=0.$

(3)由二维正态随机变量相关系数为零和相互独立两者是等价的结论,可知:X与Z是相互独立的.

二、相关系数的意义

1. 问题的提出

问 a,b 应如何选择,可使 $aX+b$ 最接近 Y?

接近的程度又应如何来衡量?

设 $e=E[(Y-(a+bX))^2]$

则 e 可用来衡量 $a+bX$ 近似表达 Y 的好坏程度.

当 e 的值越小,表示 $a+bX$ 与 Y 的近似程度越好.

确定 a,b 的值,使 e 达到最小.

$$e = E[(Y-(a+bX))^2]$$

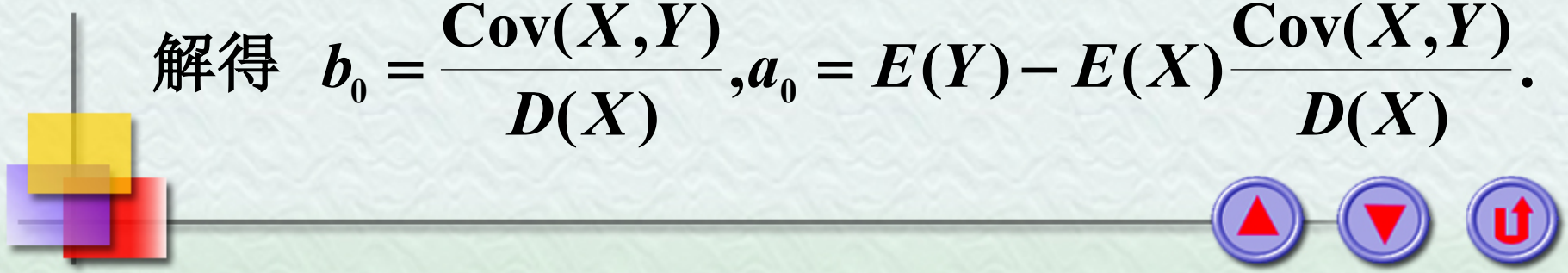
$$= E(Y^2)+b^2E(X^2)+a^2-2bE(XY)+2abE(X)$$

$$-2aE(Y).$$

将 e 分别关于 a,b 求偏导数,并令它们等于零,得

$$\begin{cases} \dfrac{\partial e}{\partial a} = 2a+2bE(X)-2E(Y)=0, \\ \dfrac{\partial e}{\partial b} = 2bE(X^2)-2E(XY)+2aE(X)=0. \end{cases}$$

解得 $b_0 = \dfrac{\mathrm{Cov}(X,Y)}{D(X)}, a_0 = E(Y)-E(X)\dfrac{\mathrm{Cov}(X,Y)}{D(X)}.$



将 a_0, b_0 代入 $e = E[(Y-(a+bX))^2]$ 中,得

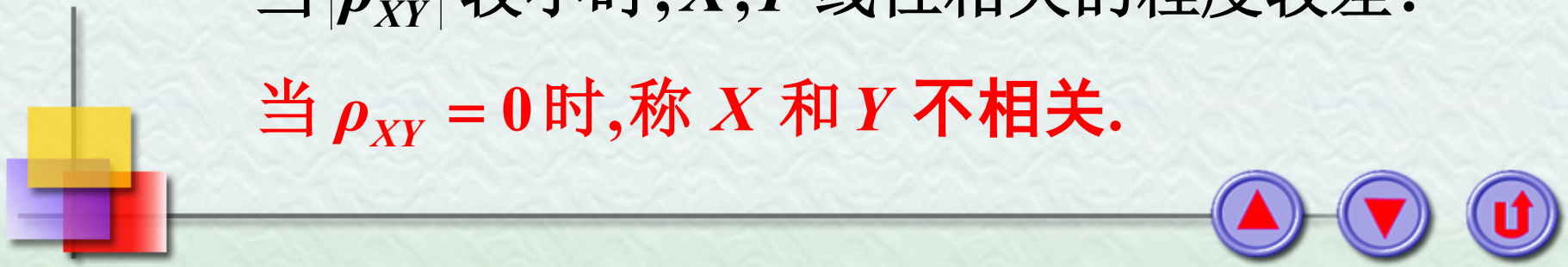
$$\begin{aligned}\min_{a,b} e &= E[(Y-(a+bX))^2] \\ &= E[(Y-(a_0+b_0X))^2] \\ &= (1-\rho_{XY}^2)D(Y).\end{aligned}$$

2. 相关系数的意义

当 $|\rho_{XY}|$ 较大时 e 较小, 表明 X, Y 的线性关系联系较紧密.

当 $|\rho_{XY}|$ 较小时, X, Y 线性相关的程度较差.

当 $\rho_{XY}=0$ 时,称 X 和 Y 不相关.



例3　设 θ 服从 $[0,2\pi]$ 的均匀分布，$\xi=\cos\theta,\eta=\cos(\theta+a)$，这里 a 是常数，求 ξ 和 η 的相关系数？

解　$$E(\xi)=\frac{1}{2\pi}\int_0^{2\pi}\cos x\,\mathrm{d}x=0,$$

$$E(\eta)=\frac{1}{2\pi}\int_0^{2\pi}\cos(x+a)\,\mathrm{d}x=0,$$

$$E(\xi^2)=\frac{1}{2\pi}\int_0^{2\pi}\cos^2 x\,\mathrm{d}x=\frac{1}{2},$$

$$E(\eta^2)=\frac{1}{2\pi}\int_0^{2\pi}\cos^2(x+a)\,\mathrm{d}x=\frac{1}{2},$$

$$E(\xi\eta)=\frac{1}{2\pi}\int_0^{2\pi}\cos x\cdot\cos(x+a)\,\mathrm{d}x=\frac{1}{2}\cos a,$$

由以上数据可得相关系数为 $\rho=\cos a.$

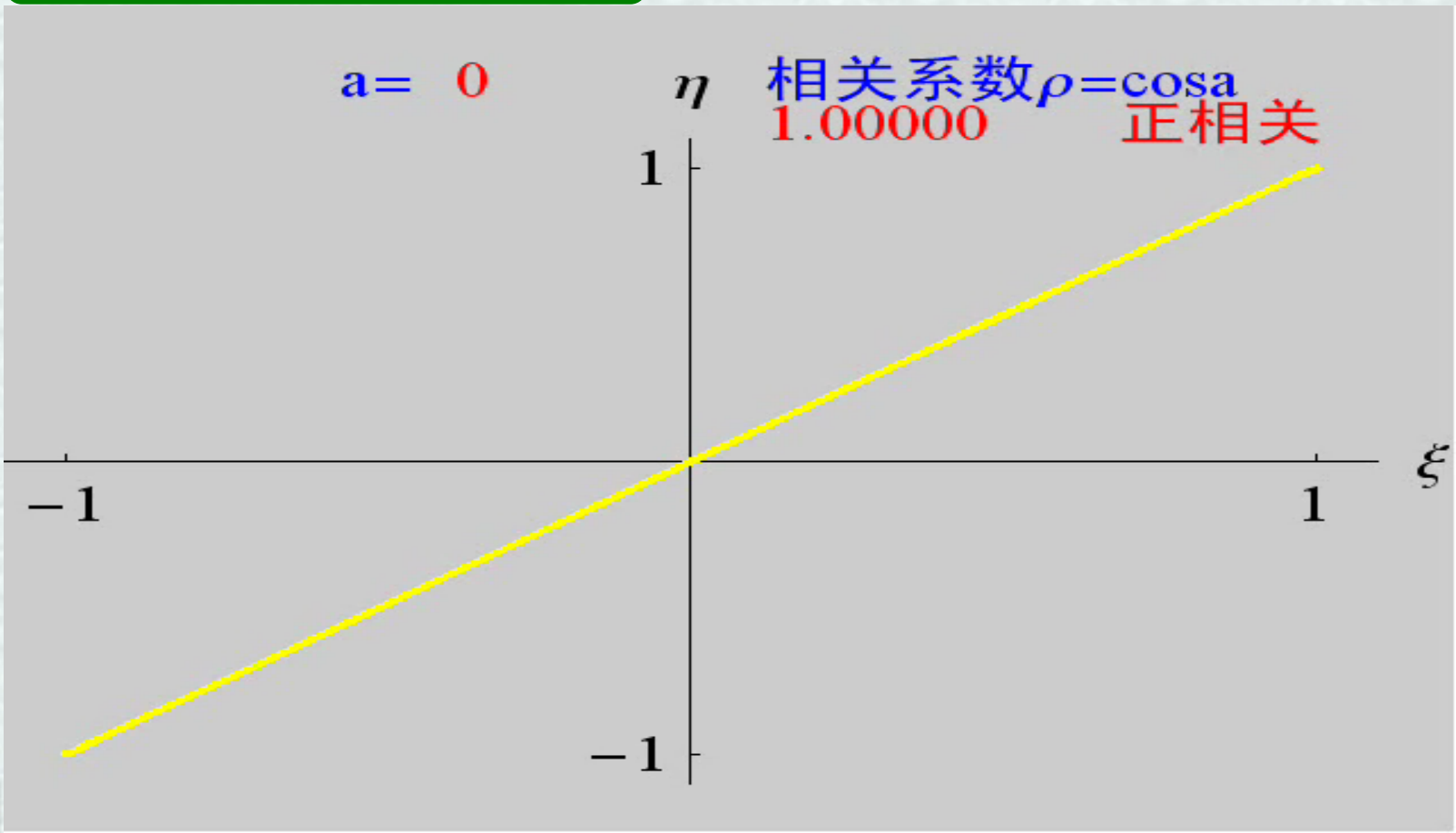
$$\left.\begin{array}{l}\text{当}a=0\text{时},\rho=1,\xi=\eta,\\ \text{当}a=\pi\text{时},\rho=-1,\xi=-\eta,\end{array}\right\}\text{存在线性关系.}$$

当 $a=\dfrac{\pi}{2}$ 或 $a=\dfrac{3\pi}{2}$ 时，$\rho=0$，ξ 与 η 不相关.

但 $\xi^2+\eta^2=1$，因此 ξ 与 η 不独立.

动画演示 ξ 与 η 的相关关系.

单击图形播放/暂停 ESC键退出



3. 注意

(1) 不相关与相互独立的关系

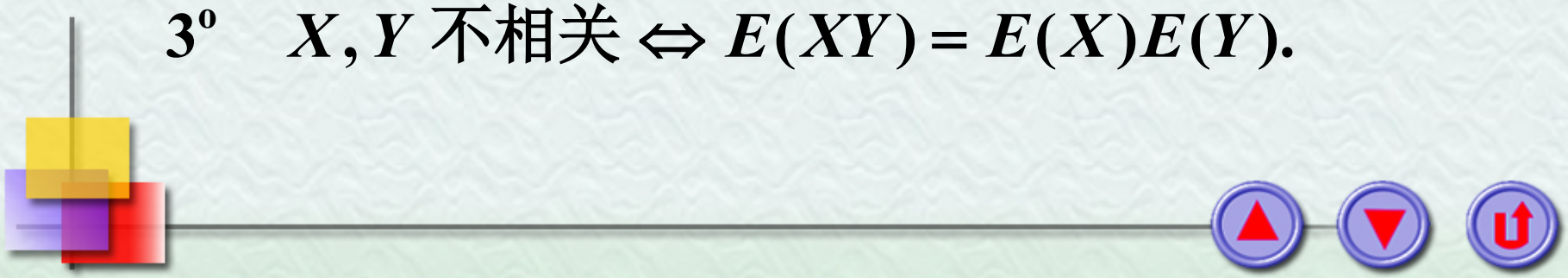
相互独立 $\rightleftharpoons$ 不相关

(2) 不相关的充要条件

1^{o} X, Y 不相关 $\Leftrightarrow \rho_{XY} = 0;$

2^{o} X, Y 不相关 $\Leftrightarrow \mathrm{Cov}(X, Y) = 0;$

3^{o} X, Y 不相关 $\Leftrightarrow E(XY) = E(X)E(Y).$



4. 相关系数的性质

(1) $|\rho_{XY}| \le 1.$

(2) $|\rho_{XY}| = 1$ 的充要条件是：存在常数 a, b 使

$$P\{Y = a + bX\} = 1.$$

证明 (1) $\min\limits_{a,b} e = E[(Y - (a + bX))^2]$

$$= (1 - \rho_{XY}^2) D(Y) \ge 0$$

$$\Rightarrow 1 - \rho_{XY}^2 \ge 0$$

$$\Rightarrow |\rho_{XY}| \le 1.$$

(2) $|\rho_{XY}|=1$ 的充要条件是,存在常数 a,b 使

$$P\{Y=a+bX\}=1.$$

事实上,$|\rho_{XY}|=1 \Rightarrow E[(Y-(a_0+b_0X))^2]=0$

$$\Rightarrow 0=E[(Y-(a_0+b_0X))^2]$$

$$=D[Y-(a_0+b_0X)]+[E(Y-(a_0+b_0X))]^2$$

$$\Rightarrow D[Y-(a_0+b_0X)]=0,$$

$$E[Y-(a_0+b_0X)]=0.$$

由方差性质知

$$P\{Y-(a_0+b_0X)=0\}=1, \text{ 或 } P\{Y=a_0+b_0X\}=1.$$

反之,若存在常数 a^*, b^* 使

$$P\{Y = a^* + b^* X\} = 1 \Leftrightarrow P\{Y - (a^* + b^* X) = 0\} = 1,$$

$$\Rightarrow P\{[Y - (a^* + b^* X)]^2 = 0\} = 1,$$

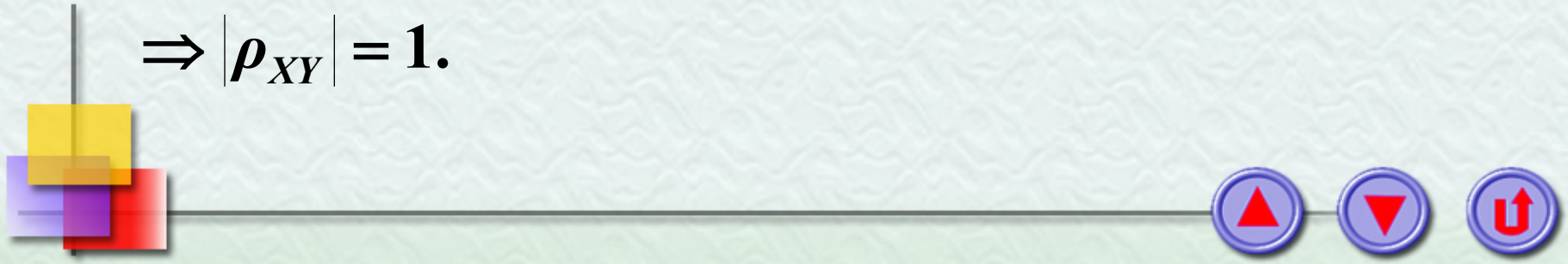
$$\Rightarrow E\{[Y - (a^* + b^* X)]^2\} = 0.$$

故有

$$0 = E\{[Y - (a^* + b^* X)]^2\} \geq \min_{a,b} E[(Y - (a + bX))^2]$$

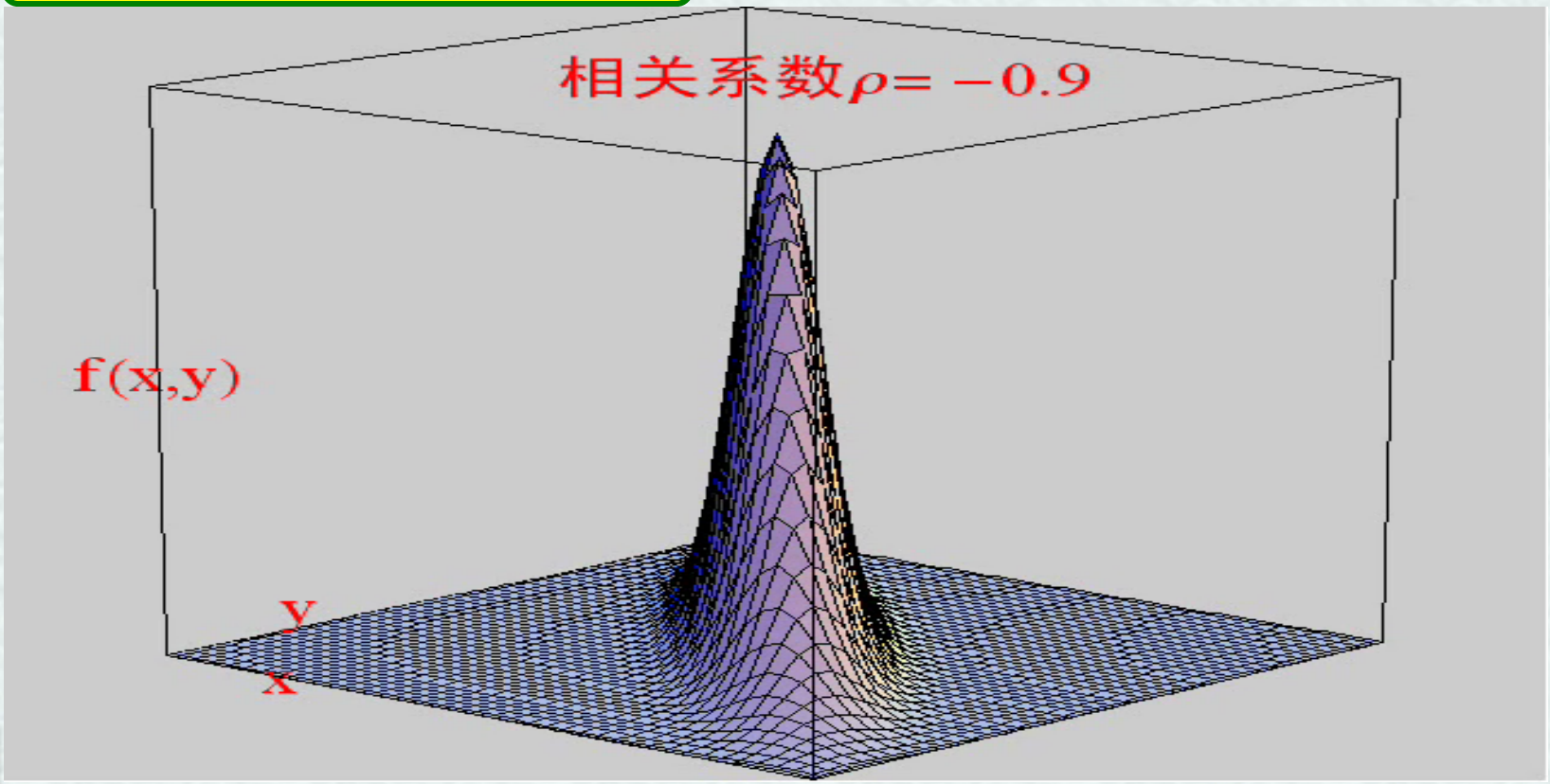
$$= E\{[Y - (a_0 + b_0 X)]^2\} = (1 - \rho_{XY}^2) D(Y)$$

$$\Rightarrow |\rho_{XY}| = 1.$$



二维正态随机变量 (X,Y) 的概率密度曲面与相关系数 $\rho_{XY}=\rho$ 的关系.

单击图形播放/暂停　ESC键退出



三、小结

相关系数的意义

当 $|\rho_{XY}|$ 较大时，X, Y 的线性相关程度较高.

当 $|\rho_{XY}|$ 较小时，X, Y 的线性相关程度较差.

当 $\rho_{XY} = 0$ 时，X 和 Y 不相关.