概率论与数理统计

第二十六讲

协方差与相关系数

- 讨论随机变量 X,Y 之间的关系

设 $(X,Y)\sim f(x,y)$, $X\sim f_X(x)$, $Y\sim f_Y(y)$, 则

X,Y 相互独立 $\Longleftrightarrow$ $f(x,y)=f_X(x)\cdot f_Y(y)$

- 问题 若 X, Y 不独立, 如何刻画它们之间的关系？

分析 若 X, Y 独立,则

$$E[(X-E(X))(Y$$

刻画随机变量之间关系的数字特征

反之, 若

$$\boxed{E[(X-E(X))(Y-E(Y))]\neq 0}$$

则 X, Y 必不相互独立.

定义 设随机变量 X, Y 的方差都存在, 记

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

则称 $\mathrm{Cov}(X,Y)$ 为 X, Y 的协方差.

易知

(1)若 X, Y 相互独立,则 $\mathrm{Cov}(X,Y)=0$

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

(3) $D(X)=E[(X-E(X))^2]=\mathrm{Cov}(X,X)$

- 协方差的基本性质：(其中 a,b 为常数)

(1) 若 X,Y 相互独立,则 $\mathrm{Cov}(X,Y)=0$

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

(3) $D(X)=E[(X-E(X))^2]=\mathrm{Cov}(X,X)$

(4) $\mathrm{Cov}(aX,bY)=ab\mathrm{Cov}(X,Y)$

(5) $\mathrm{Cov}(X_1+X_2,Y_1+Y_2)=\mathrm{Cov}(X_1,Y_1)+\mathrm{Cov}(X_1,Y_2)$
$+\mathrm{Cov}(X_2,Y_1)+\mathrm{Cov}(X_2,Y_2)$

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

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

协方差的意义

$\because X,Y$ 相互独立 $\Longrightarrow$ $\mathrm{Cov}(X,Y)=0$

$\therefore \mathrm{Cov}(X,Y)\neq 0$ $\Longrightarrow$ X,Y 必不独立

$\Longrightarrow$ X,Y 之间必存在某种关系

问题

(1) 这种关系是什么关系？

(2) 这种关系的密切程度能否用 $\mathrm{Cov}(X,Y)$ 的值的大小来表示？

分析 (2) 这种关系的密切程度能否用 $\mathrm{Cov}(X,Y)$ 的值的大小来表示？

$\forall\, k\in(-\infty,\infty)$, 由协方差的性质,有

$$\mathrm{Cov}(kX,kY)=k^2\mathrm{Cov}(X,Y)$$

故问题(2)的答案是否定的！

考虑“单位化”的随机变量,令

$$X^*=\frac{X-E(X)}{\sqrt{D(X)}},\quad Y^*=\frac{Y-E(Y)}{\sqrt{D(Y)}}$$

易知 $E(X^*)=0,\ D(X^*)=1$

$E(Y^*)=0,\ D(Y^*)=1$

分析 (2) 这种关系的密切程度能否用 $\mathrm{Cov}(X,Y)$ 的值的大小来表示？

$\forall\, k\in(-\infty,\infty)$, 由协方差的性质,有

$$\mathrm{Cov}(kX,kY)=k^2\mathrm{Cov}(X,Y)$$

故问题(2)的答案是否定的！

考虑“单位化”的随机变量,令

$$X^*=\frac{X-E(X)}{\sqrt{D(X)}},\quad Y^*=\frac{Y-E(Y)}{\sqrt{D(Y)}}$$

定义 称

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

为 X,Y 的**相关系数**.

分析 (1) 这种关系是什么关系？

考虑 X,Y 之间的线性关系. 即用随机变量

$$\hat{Y}=a+bX,\quad (a,b \text{ 为常数})$$

近似表示 Y. 考虑均方误差

$$e=E[(Y-\hat{Y})^2]=E[(Y-(a+bX))^2]$$

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

解得 $b_0=\dfrac{\mathrm{Cov}(X,Y)}{D(X)}, a_0=E(Y)-b_0E(X)$, 即有

$$\min_{a,b} e=\min_{a,b} E[(Y-(a+bX))^2]=E[(Y-(a_0+b_0X))^2]$$

$$\min_{a,b} e = \min_{a,b} E[(Y-(a+bX))^2] = \boxed{E[(Y-(a_0+b_0X))^2]}$$

$$= E[(Y-(E(Y)-b_0E(X)+b_0X))^2]$$

$$= E\{\cdots E(X))]^2\}$$

$$= E[\cdots - E(X))^2]$$

$$-2b_0E[(Y-E(Y))(X-E(X))]$$

$$= D(Y)+b_0^2D(X)-2b_0\mathrm{Cov}(X,Y)$$

$$= D(Y)+b_0\mathrm{Cov}(X,Y)-2b_0\mathrm{Cov}(X,Y)$$

$$= D(Y)-b_0\mathrm{Cov}(X,Y)$$

$$= D(Y)[1-\frac{\mathrm{Cov}^2(X,Y)}{D(X)D(Y)}]$$

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

$$b_0 = \frac{\mathrm{Cov}(X,Y)}{D(X)}$$

$$b_0D(X) = \mathrm{Cov}(X,Y)$$

从而得到关系式

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

其中 $b_0=\dfrac{\mathrm{Cov}(X,Y)}{D(X)}, a_0=E(Y)-b_0E(X).$

定理 X,Y 的相关系数 ρ_{XY} 具有下列性质

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

(2) $|\rho_{XY}|=1 \iff Y\overset{a.e}{=}a+bX$（$a,b$ 为常数)

$$\boxed{\min_{a,b} e = E[(Y-(a_0+b_0X))^2] = D(Y)(1-\rho_{XY}^2)}$$

定理 X,Y 的相关系数 ρ_{XY} 具有下列性质

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

(2) $|\rho_{XY}|=1 \Longleftrightarrow Y \overset{a.e}{=} a+bX$ （a,b 为常数)

- **相关系数的实际意义**

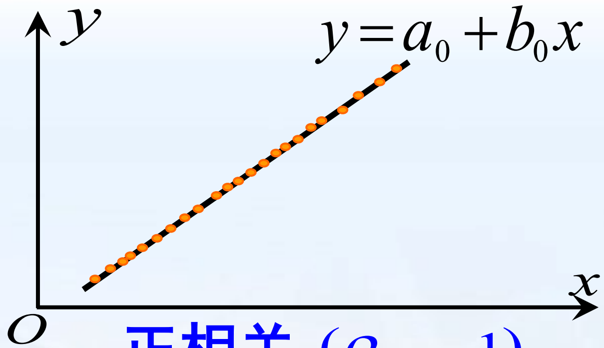
$|\rho_{XY}|$较大 $\Longrightarrow$ e 较小 $\Longrightarrow$ X,Y 之间线性关系较密切

$|\rho_{XY}|$较小 $\Longrightarrow$ e 较大 $\Longrightarrow$ X,Y 之间线性关系较弱

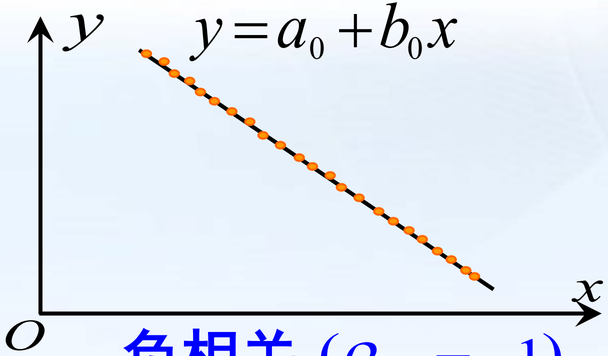
定义 设 X,Y 的相关系数为 ρ_{XY}

(1) 当 $|\rho_{XY}|=1$ 时,称 X 与 Y **相关**;

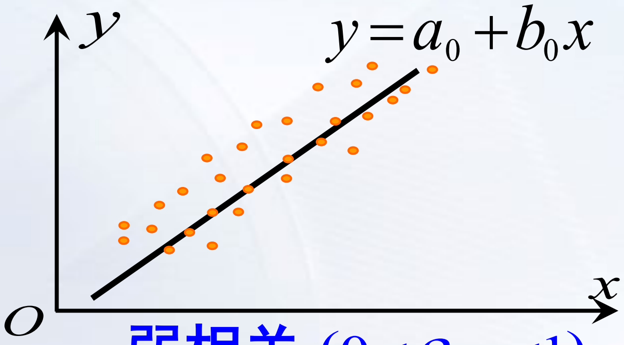
(2) 当 $\rho_{XY}=0$ 时,称 X 与 Y **不相关**.



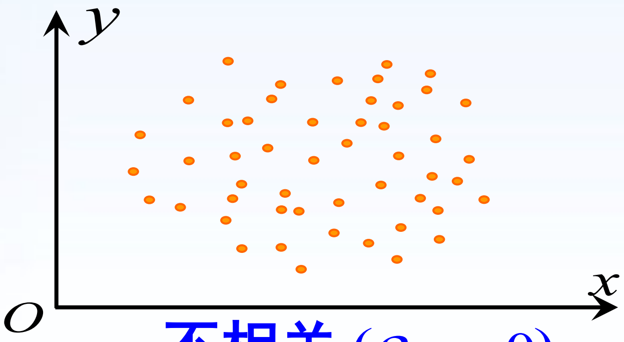
正相关 ($\rho_{XY}=1$)



负相关 ($\rho_{XY}=-1$)



弱相关 ($0<\rho_{XY}<1$)



不相关 ($\rho_{XY}=0$)

例 设 X,Y 服从单位圆域 $G: x^2+y^2\le 1$ 上的均匀分布. 讨论 X,Y 的独立性与相关性.

解 先前求得 X,Y 的边缘密度函数分别为

$$f_X(x)=\begin{cases}\frac{1}{\pi}\sqrt{1-x^2}, & |x|<1,\\ 0, & |x|\ge 1,\end{cases}\quad f_Y(y)=\begin{cases}\frac{1}{\pi}\sqrt{1-y^2}, & |y|<1,\\ 0, & |y|\ge 1.\end{cases}$$

$\because\ f(x,y)=\frac{1}{\pi}\ne f_X(x)\cdot f_Y(y)\ \ (\forall (x,y)\in G)$

$\therefore\ X,Y$ 不独立.

对单位圆域上的均匀分布, 由前讲曾求得

$$E(XY)=0, E(X)=0, E(Y)=0$$

$$\therefore\ \mathrm{Cov}(X,Y)=E(XY)-E(X)E(Y)=0$$

故 X,Y 不相关.

相关系数的实际意义

相关系数是刻画随机变量之间线性关系的数字特征

$|\rho_{XY}|$较大 ⇨ e 较小 ⇨ X,Y 之间线性关系较密切

$|\rho_{XY}|$较小 ⇨ e 较大 ⇨ X,Y 之间线性关系较弱

其中均方误差 $e=E[(Y-(a+bX))^2]$.

X,Y 相互独立

⇓ ⇑ **不一定！**

X,Y 不相关

本讲结束　谢谢大家