第三节　协方差与相关系数

一、　协方差

二、　相关系数

一、协方差 (covariance)

1. 概念的引入

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

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

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

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

协方差

2. 定义

若$E(X)$, $E(Y)$和$E[(X\text{-}E(X))(Y\text{-}E(Y))]$都存在.

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

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

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

3. 协方差的计算公式

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

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

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

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

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

注：若随机变量 X 和 Y 相互独立,则$Cov(X,Y)=0$.

4. 简单性质

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

(2) $\mathrm{Cov}(X, a)=0$, a为常数

(3) $\mathrm{Cov}(aX, bY)=ab\cdot\mathrm{Cov}(X,Y)$, a, b是常数

(4) $\mathrm{Cov}(X+Y,Z)=\mathrm{Cov}(X,Z)+\mathrm{Cov}(Y,Z)$

证：$\mathrm{Cov}(X+Y,Z)=\mathrm{E}[(X+Y)Z]-E(X+Y)E(Z)$

$=E(XZ)+E(YZ)-E(X)E(Z)-E(Y)E(Z)$

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

由(1), (3), (4)可得协方差具有双线性性。

(5) $\mathrm{D}(aX\pm bY)=a^2\mathrm{D}(X)+b^2\mathrm{D}(Y)\pm 2abE[(X-E(X))(Y-E(Y))]$

$=a^2\mathrm{D}(X)+b^2\mathrm{D}(Y)\pm 2ab\cdot\mathrm{Cov}(X, Y)$

例4.3.1 设 $X \sim N(0,1), Y = X^2$, 求 $\mathrm{Cov}(X,Y)$.

解： 因为X~N(0,1)，所以$E(X)$=0. 又

$$E(Y) = \int_{-\infty}^{+\infty} x^2 \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx = -\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} x d e^{-\frac{x^2}{2}}$$

$$= -\frac{1}{\sqrt{2\pi}} \left[x e^{-\frac{x^2}{2}} \Big|_{-\infty}^{+\infty} - \int_{-\infty}^{+\infty} e^{-\frac{x^2}{2}} dx \right] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-\frac{x^2}{2}} dx$$

$$= 1$$

所以 $Cov(X,Y) = E(XY) - E(X)E(Y)$

$$= E(X^3) = \int_{-\infty}^{+\infty} x^3 \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx = 0.$$

协方差的意义

X,Y相互独立 $\Longrightarrow$ $Cov(X,Y)=0.$

$Cov(X,Y)\neq 0$ $\Longrightarrow$ X,Y不相互独立

$\Longrightarrow$ 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Cov(X,Y)的值的大小来表示?

对任意实数k，由协方差的性质，

$$Cov(kX,kY)=k^2Cov(X,Y),$$

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，$(kX)^*=X^*$，$(kY)^*=Y^*$，

$$Cov((kX)^*,(kY)^*)=Cov(X^*,Y^*)=\boxed{\frac{Cov(X,Y)}{\sqrt{D(X)}\cdot\sqrt{D(Y)}}},$$

相关系数

二、相关系数

1、定义： 设D(X)>0, D(Y)>0, 协方差Cov(X,Y)存在，则称

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

为随机变量X和Y的相关系数 .

注：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.

$$\rho_{XY}=Cov(X^*,Y^*).$$

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$a+bX$ (a, b为常数)近似表示Y. 接近的程度可以用均方误差来衡量。

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.

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

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

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

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

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

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

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

定理 X,Y的相关系数ρ_{XY}具有下列性质：

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

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

$$|\rho_{XY}|=1 \Leftrightarrow$$ 存在常数a,b,使得$P\{Y=aX+b\}=1$

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

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

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

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

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

由方差性质知

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

反之,若存在常数 a^*, b^* 使

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

$$\Rightarrow D[Y - (a^* + b^* X)] = 0$$

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

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

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

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

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

2. 相关系数的意义

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

当$\left|\rho_{XY}\right|$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**.

当 $|\rho_{XY}|=1$ 时, 称 X 和 Y 线性**相关**.

当 $0<|\rho_{XY}|<1$ 时, 称 X 和 Y 弱**相关**.

当 $\rho_{XY}>0$ 时, 称 X 和 Y为正（弱）相关.

当 $\rho_{XY}<0$ 时, 称 X 和 Y为负（弱）相关。

3. 相关系数的性质

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

证2:由Cauchy-Schwarz不等式可得

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

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

因此, $|Cov(X,Y)| \le \sqrt{D(X)D(Y)}$

从而, $|\rho_{XY}| = \left|\dfrac{Cov(X,Y)}{\sqrt{D(X)D(Y)}}\right| \le 1.$

(2) X与Y相互独立 $\Rightarrow \rho_{XY}=0$, 即X与Y不相关.

但反之不成立.

注: 若Cov(X, Y)≠0，则X与Y一定不独立.

(即逆否命题成立)

(3) X与Y不相关 $\Leftrightarrow \rho_{XY}=0$

$$\Leftrightarrow Cov\left(X,Y\right)=0$$

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

$$\Leftrightarrow D(X\pm Y)=D(X)+D(Y)$$

例4.3.2 设连续型随机变量(X, Y)的密度函数为

$$f(x,y)=\begin{cases}\dfrac{1}{\pi}, x^2+y^2<1\\ 0,\ 其他\end{cases}$$,验证X,Y不相关，但是X,Y不相互独立.

解： 易求边缘密度函数分别为

$$f_X(x)=\int_{-\infty}^{+\infty}f(x,y)dy=\begin{cases}\int_{-\sqrt{1-x^2}}^{+\sqrt{1-x^2}}\dfrac{1}{\pi}dy, -1<x<1\\ 0, \qquad 其他\end{cases}=\begin{cases}\dfrac{2\sqrt{1-x^2}}{\pi}, -1<x<1\\ 0, \qquad 其他\end{cases},$$

$$f_Y(y)=\int_{-\infty}^{+\infty}f(x,y)dx=\begin{cases}\int_{-\sqrt{1-y^2}}^{+\sqrt{1-y^2}}\dfrac{1}{\pi}dx, -1<y<1\\ 0, \qquad 其他\end{cases}=\begin{cases}\dfrac{2\sqrt{1-y^2}}{\pi}, -1<y<1\\ 0, \qquad 其他\end{cases}.$$

$f(x,y)\neq f_X(x)f_Y(y)$,所以X,Y不相互独立.

$$E(XY)=\iint\limits_{x^2+y^2<1} xy\cdot\frac{1}{\pi}dxdy=\int_0^{2\pi}d\theta\int_0^1\frac{1}{\pi}r^3\sin\theta\cos\theta dr=0,$$

$$E(X)=\iint\limits_{x^2+y^2<1} x\cdot\frac{1}{\pi}dxdy=\int_0^{2\pi}d\theta\int_0^1\frac{1}{\pi}r^2\cos\theta dr=0,$$

$$E(Y)=\iint\limits_{x^2+y^2<1} y\cdot\frac{1}{\pi}dxdy=\int_0^{2\pi}d\theta\int_0^1\frac{1}{\pi}r^2\sin\theta dr=0.$$

于是$\rho=\dfrac{cov(X,Y)}{\sqrt{D(X)}\sqrt{D(Y)}}=\dfrac{E(XY)-E(X)E(Y)}{\sqrt{D(X)}\sqrt{D(Y)}}=0,$

故X与Y不相关.

X与Y既不相关也不相互独立的例子！

独立 VS 不相关，不独立 VS 相关

X与Y相互独立 $\Rightarrow$ X与Y不具有任何关系

$\Rightarrow$ X与Y不具有线性关系

$\Rightarrow$ X与Y不相关

X与Y不相关 $\Rightarrow$ X与Y不具有线性关系

$\Rightarrow$ X与Y可能具有其它关系

$\Rightarrow$ X与Y可能不独立

不相关是就线性关系而言，相互独立是就一般关系而言的.

例4.3.3 设$(X,Y)\sim N(\mu_1,\mu_2,\sigma_1{}^2,\sigma_2{}^2,\rho)$，求Cov$(X,Y)$,并证明X与Y相互独立当且仅当X与Y不相关.

解： 已知 $X\sim N(\mu_1,\sigma_1{}^2), Y\sim N(\mu_2,\sigma_2{}^2)$,则有

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

$$
\begin{aligned}
Cov(X,Y) &= E[(X-EX)(Y-EY)] \\
&= E[(X-\mu_1)(Y-\mu_2)] \\
&= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}(x-\mu_1)(y-\mu_2)\underline{f(x,y)}dxdy \\
&= \rho\sigma_1\sigma_2 \\
&= \rho\sqrt{D(X)}\sqrt{D(Y)}.
\end{aligned}
$$

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

$f(x,y)$

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

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

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

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

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

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

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

0

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

$$\rho_{XY} = \rho.$$

已知结论： 设$(X,Y) \sim N(\mu_1,\mu_2,\sigma_1^2,\sigma_2^2,\rho)$，

则X与Y相互独立的充分必要条件是$\rho = 0$.

二维正态分布，有

X与Y相互独立 $\Leftrightarrow \rho = 0 \Leftrightarrow \rho_{XY} = 0$

对其它分布，X与Y相互独立 $\underset{不一定}{\overset{必定}{\rightleftarrows}}$ $\rho_{XY} = 0$