§ 4.3 协方差和相关系数

问题 对于二维随机变量(X,Y):

$$\boxed{\text{已知联合分布}} \underset{\not\longleftarrow}{\longrightarrow} \boxed{\text{边缘分布}}$$

对二维随机变量,除每个随机变量各自的概率特性外,相互之间可能还有某种联系问题是用一个怎样的数去反映这种联系.

我们已经知道当X与Y相互独立时，有

$$E([X-E(X)][Y-E(Y)])=0$$

也就是说，当$E([X-E(X)][Y-E(Y)])\neq 0$时，$X$与$Y$不独立。

数值 $E([X-E(X)][Y-E(Y)])$

一定程度上反映了 X, Y 之间的某种关系

协方差的定义

定义 称 $E([X-E(X)][Y-E(Y)])$ 为 X,Y 的协方差.记为

$$\operatorname{cov}(X,Y)=E([X-E(X)][Y-E(Y)])$$

若 $(\boldsymbol{X},\boldsymbol{Y})$ 为离散型,

$$\operatorname{cov}(X,Y)=\sum_{i=1}^{\infty}\sum_{j=1}^{\infty}[x_i-E(X)][y_j-E(Y)]p_{ij}$$

若 $(\boldsymbol{X},\boldsymbol{Y})$ 为连续型,

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

协方差是统计学上表示两个随机变量之间关系的变量，在确定有价证券组合收益率时，协方差的含义是:

如果其是正值，则表明证券A和证券B的收益有相互一致的变动趋向，即一种证券的收益高于预期收益，另一种证券的收益也高于预期收益；一种证券的收益低于预期收益，另一种证券的收益也低于预期收益。如果得到的是负值，则表明证券A和证券B的收益相互抵销的趋向，即一种券的收益高于预期收益，则另一种证券的收益低于预期的收益，反之亦然。

协方差的计算

- $\operatorname{cov}(X,Y)=E(XY)-E(X)E(Y)$

$$=\pm\frac{1}{2}\big(D(X\pm Y)-D(X)-D(Y)\big)$$

协方差的性质

- $\operatorname{cov}(X,Y)=\operatorname{cov}(Y,X)=E(XY)-E(X)E(Y)$
- $\operatorname{cov}(aX,bY)=ab\operatorname{cov}(X,Y)$
- $\operatorname{cov}(X+Y,Z)=\operatorname{cov}(X,Z)+\operatorname{cov}(Y,Z)$
- $\operatorname{cov}(X,X)=D(X)$

协方差的数值虽然在一定程度上反映了X与Y相互间的联系，但它还受X与Y本身数值大小的影响。比如说，令X与Y各自增大K倍，即

$$X_1 = KX, Y_1 = KY$$

这时X_1与Y_1的相互联系和X与Y间的相互间的联系应该是一样的，可是反映这种联系的协方差却增大了K^2倍，即

$$\operatorname{cov}(X_1, Y_1) = K^2 \operatorname{cov}(X, Y)$$

所以协方差是一个绝对量，不能进行大小的比较。为了克服这一缺点，在计算随机变量的协方差之前，先对X与Y标准化，因此引入相关系数的定义

相关系数的定义

若$D(X)>0, D(Y)>0$,称

$$E\left(\frac{(X-E(X))(Y-E(Y))}{\sqrt{D(X)}\sqrt{D(Y)}}\right)=\frac{\operatorname{cov}(X,Y)}{\sqrt{D(X)}\sqrt{D(Y)}}$$

为X,Y的 相关系数，记为

$$\boxed{\rho_{XY}=\frac{\operatorname{cov}(X,Y)}{\sqrt{D(X)}\sqrt{D(Y)}}} \longrightarrow \text{无量纲的量}$$

例1 已知 X,Y 的联合分布为

$Y \backslash X$ (p_{ij})	1	0
1	p	0
0	0	q

$$0 < p < 1$$

$$p + q = 1$$

解

X	1	0
P	p	q

Y	1	0
P	p	q

$$E(X) = p,\quad E(Y) = p,\qquad E(XY) = p,$$

$$\operatorname{cov}(X,Y) = E(XY) - E(X)E(Y) = pq$$

$$D(X) = pq,\quad D(Y) = pq,\qquad \rho_{XY} = \frac{\operatorname{cov}(X,Y)}{\sqrt{D(X)}\sqrt{D(Y)}} = 1$$

例、设**X,Y**的概率密度为

$$f(x,y)=\begin{cases} x+y, 0\le x\le 1, 0\le y\le 1 \\ 0, \text{其他} \end{cases}$$

求**Cov(X,Y),ρ**$_{XY}$。

解、$E(X)=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} xf(x,y)dydx=\int_0^1\int_0^1 x(x+y)dydx=\dfrac{7}{12}$

$$E(X^2)=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} x^2 f(x,y)dydx=\int_0^1\int_0^1 x^2(x+y)dydx=\frac{5}{12}$$

同理$E(Y)=\dfrac{7}{12}, E(Y^2)=\dfrac{5}{12}$。又

$$E(XY)=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} xyf(x,y)dxdy=\int_0^1 dx\int_0^1 xy(x+y)dy=\frac{1}{3}$$

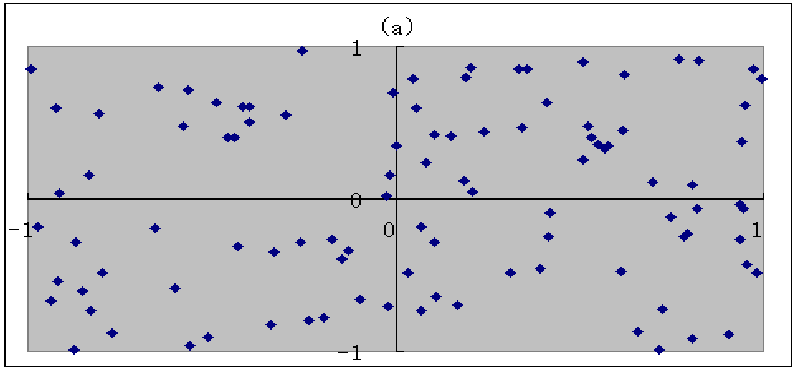
$$D(Y)=D(X)=E(X^2)-E^2(X)=\frac{5}{12}-(\frac{7}{12})^2=\frac{11}{144}$$

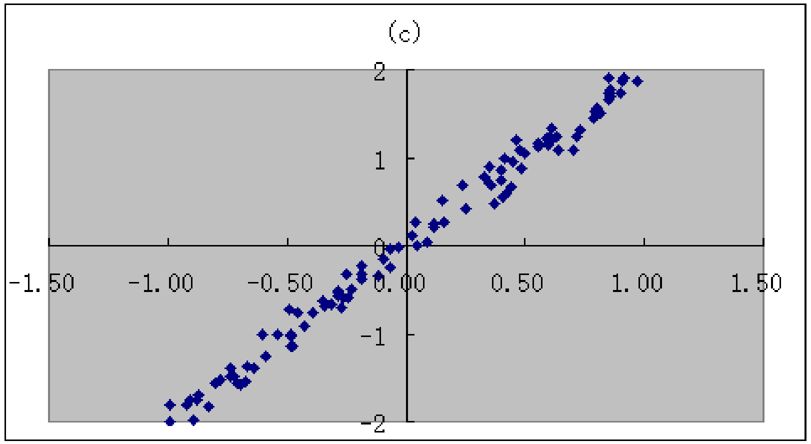
$$Cov(X,Y)=E(XY)-E(X)E(Y)=\frac{1}{3}-(\frac{7}{12})^2=-\frac{1}{144}$$

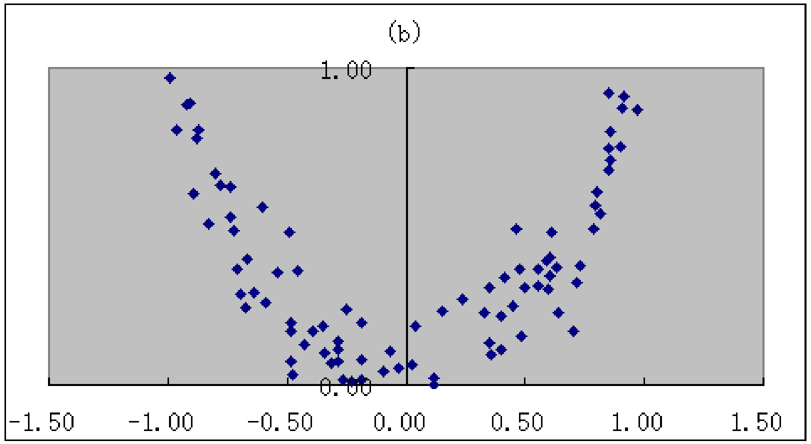
$$\rho_{XY}=\frac{\operatorname{cov}(X,Y)}{\sqrt{D(X)}\sqrt{D(Y)}}=\frac{-\frac{1}{144}}{\frac{11}{144}}=-\frac{1}{11}$$

如何描述两个随机变量之间的关系?

若(X, Y)的全部可能取值坐标如图a, b, c, X与Y的关系各是什么?







19

讨 论

考虑用直线 $$Y = aX + b$$

逼近 (X,Y)的所有可能取值点.

1.参数a,b取什么值时,(X,Y)的所有可能取值点与直线最接近?—如何定义(X,Y)与直线的距离?

2.在什么情况下, (X,Y)的所有可能取值落在一条直线上?

3.在什么情况下, X与Y没有线性关系?

考虑Y与X的线性函数aX+b的平均距离”：

$$e = E\{[Y-(aX+b)]^2\} = E(Y^2)+a^2E(X^2)+b^2$$
$$-2aE(XY)+2abE(X)-2bE(Y) \qquad (1)$$

取a,b使e最小:

令

$$\frac{\partial e}{\partial a} = 2aE(X^2)-2E(XY)+2bE(X)=0$$

$$\frac{\partial e}{\partial b} = 2b+2aE(X)-2E(Y)=0$$

解得:

$$\begin{cases} a_0 = \dfrac{E(XY)-E(X)E(Y)}{D(X)} \\ b_0 = E(Y)-E(X)\dfrac{E(XY)-E(X)E(Y)}{D(X)} \end{cases} \qquad (2)$$

19

将(2)带入(1),得

$$\min_{a,b} E\left\{[Y-(aX+b)]^2\right\}$$

$$= D(Y)\left[1-\frac{\left(E(XY)-E(X)E(Y)\right)^2}{D(X)D(Y)}\right] \qquad (3)$$

$$= D(Y)\left[1-\rho^2(X,Y)\right]$$

相关系数的性质

- $|\rho_{XY}| \le 1$
- $|\rho_{XY}| = 1$，称X与Y完全相关

$\Longleftrightarrow$ 存在常数a,b,使 $P\{Y = aX + b\} = 1$

- $|\rho_{XY}| = 0$　称X与Y不相关

相关系数只是随机变量间线性关系强弱的一个度量，因而说得更确切一些，应该把它称作线性相关系数，只是大家习惯了，所以叫相关系数。

当 $|\rho|=1$ 时，上述性质表明X与Y间存在线性关系，并且 $\rho=1$ 时正线性相关，$\rho=-1$ 时负线性相关，当 $\rho=0$ 时，称X与Y是不相关的或者零相关的。

前面已有：当X与Y独立时，必**Cov(X,Y)=0**。从而 $\rho=0$ 于是X与Y不相关，确切地讲，是不线性相关。问题：当X与Y不相关时，X与Y是否一定独立？

问题：当X与Y不相关时，X与Y是否一定独立？

回答：否！理由：若X与Y不相关，则它们之间

不存在线性关系，但它们之间可能存在别的函数关系，从而是不独立的。请看下例：

例、$X \sim N(0,1), Y = X^2$，求ρ_{XY}。

解 $E(X)=0, D(X)=1,$

$$\operatorname{cov}(X,Y) = E(XY) - E(X)E(Y) = E(XY)$$

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

$$\therefore \rho_{XY} = 0$$

但X与Y之间有关系。但没有线性关系。

例、$X \sim N(0,1), Y = X^2$，求ρ_{XY}。

解 $E(X)=0, D(X)=1,\quad E(Y)=E(X^2)=D(X)+E^2(X)=1$

$$\operatorname{cov}(X,Y)=E(XY)-E(X)E(Y)=E(XY)$$

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

$\therefore \rho_{XY} = 0$

但X与Y之间有关系。但没有线性关系。

例2 设 $(X,Y)\sim N(\mu_1,\sigma_1^2;\mu_2,\sigma_2^2;\rho)$，求 ρ_{XY}

解

$$\operatorname{cov}(X,Y)=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}(x-\mu_1)(y-\mu_2)f(x,y)dxdy$$

$$\overset{\frac{x-\mu_1}{\sigma_1}=s}{\underset{\frac{y-\mu_2}{\sigma_2}=t}{=\!=}}\frac{\sigma_1\sigma_2}{2\pi\sqrt{1-\rho^2}}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}ste^{-\frac{1}{2(1-\rho^2)}(s-\rho t)^2-\frac{1}{2}t^2}dsdt$$

$$\overset{令 s-\rho t=u}{=}\frac{\sigma_1\sigma_2}{2\pi\sqrt{1-\rho^2}}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}t(\rho t+u)e^{-\frac{u^2}{2(1-\rho^2)}-\frac{1}{2}t^2}dudt$$

$$=\frac{\sigma_1\sigma_2\rho}{2\pi\sqrt{1-\rho^2}}\int_{-\infty}^{+\infty}e^{-\frac{u^2}{2(1-\rho^2)}}du\int_{-\infty}^{+\infty}t^2e^{-\frac{1}{2}t^2}dt=\sigma_1\sigma_2\rho$$

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

若 $(X,Y)\sim N(\mu_1,\sigma_1^2,\mu_2,\sigma_2^2,\rho)$,

则 X,Y 相互独立 $\Longleftrightarrow$ X,Y 不相关

- $\rho_{XY} = 0$ $\longleftrightarrow$ X, Y 不相关

$\longleftrightarrow$ $\operatorname{cov}(X, Y) = 0$

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

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

X, Y 相互独立 $\overset{\Longrightarrow}{\not\Longleftarrow}$ X, Y 不相关

但若 (X, Y) 服从二维正态分布，

X, Y 相互独立 $\longleftrightarrow$ X, Y 不相关

例4 设 $(X, Y) \sim N(1,4;\ 1,4;\ 0.5),\ Z = X + Y,$ 求 ρ_{XZ}

解 $$E(X) = E(Y) = 1, D(X) = D(Y) = 4,$$

$$\rho_{XY} = 1/2, \qquad \operatorname{cov}(X, Y) = 2$$

$$\operatorname{cov}(X, Z) = \operatorname{cov}(X, X) + \operatorname{cov}(X, Y) = 6$$

$$D(Z) = D(X+Y) = D(X) + D(Y) + 2\operatorname{cov}(X, Y) = 12$$

$$\therefore \quad \rho_{XZ} = .3/\sqrt{12} = \sqrt{3/2}.$$

附录　矩在线性回归中的应用

若 X, Y 是两个r.v., 用X 的线性函数去逼近 Y 所产生的平均平方误差为

$$E[Y-(aX+b)]^2$$

当取 $$\hat{a}=\frac{\operatorname{cov}(X,Y)}{D(X)},$$

$$\hat{b}=E(Y)-\hat{a}E(X)=E(Y)-\rho_{XY}\sqrt{\frac{D(Y)}{D(X)}}E(X)$$

平均平方误差最小.

附例 设 X, Y 相互独立, 且都服从 $N(0, \sigma^2)$,

$U = aX + bY$，$V = aX - bY$，a,b 为常数，

且都不为零，求ρ_{UV}

解

$$\operatorname{cov}(U,V) = E(UV) - E(U)E(V)$$

$$= a^2E(X^2) - b^2E(Y^2) - \left[aE(X) + bE(Y)\right]\left[aE(X) - bE(Y)\right]$$

由

$$\left.\begin{array}{l} E(X) = E(Y) = 0, \\ D(X) = D(Y) = \sigma^2 \end{array}\right\} \longrightarrow \begin{array}{l} E(X^2) = \sigma^2 \\ E(Y^2) = \sigma^2 \end{array}$$

$$\longrightarrow \operatorname{cov}(U,V) = (a^2 - b^2)\sigma^2$$

而

$$D(U) = a^2D(X) + b^2D(Y) = (a^2 + b^2)\sigma^2$$

$$D(V) = a^2D(X) + b^2D(Y) = (a^2 + b^2)\sigma^2$$

故

$$\rho_{UV} = \frac{a^2 - b^2}{a^2 + b^2}$$

继续讨论

a,b 取何值时, U与V 不相关?

此时, U与V 是否独立?

若 $a=b$, $\rho_{UV}=0$, 则 U, V 不相关.

但 $U\sim N(0,\ 2a^2\sigma^2)$, $V\sim N(0,\ 2a^2\sigma^2)$,

$$\begin{cases} U=a(X+Y) \\ V=a(X-Y) \end{cases} \Rightarrow \begin{cases} X=\dfrac{1}{2a}(U+V) \\ Y=\dfrac{1}{2a}(U-V) \end{cases}$$

$$f_{UV}(u,v)=\left|\begin{vmatrix} \dfrac{1}{2a} & \dfrac{1}{2a} \\ \dfrac{1}{2a} & -\dfrac{1}{2a} \end{vmatrix}\right| f_{XY}\left(\frac{1}{2a}(u+v),\frac{1}{2a}(u-v)\right)$$

$$f_{UV}(u,v)=\frac{1}{2a^2}f_X\left(\frac{1}{2a}(u+v)\right)f_Y\left(\frac{1}{2a}(u-v)\right)$$

$$=\frac{1}{2a^2}\frac{1}{2\pi\sigma^2}e^{-\frac{\left(\frac{u+v}{2a}\right)^2+\left(\frac{u-v}{2a}\right)^2}{2\sigma^2}}$$

$$=\frac{1}{2\pi(\sqrt{2}a\sigma)^2}e^{-\frac{u^2+v^2}{4a^2\sigma^2}}$$

$(U,V)\sim N(0,2a^2\sigma^2;0,2a^2\sigma^2;0)$, 且$U,V$相互独立