§4.3　协方差与相关系数

一　协方差与相关系数的定义

数学期望和方差反映的是随机变量自身的特征，很多时候我们希望了解随机变量之间的相互关系. 例如：身高X与体重Y的关系.

显然X与Y的关系可通过$X-EX$与$Y-EY$的符号差异来反映.

$X-EX$与$Y-EY$同号：X与Y是正向关系，等价于

$$(X-EX)(Y-EY)>0,$$

$X-EX$与$Y-EY$异号：X与Y是反向关系，等价于

$$(X-EX)(Y-EY)<0.$$

因此$(X-EX)(Y-EY)$的符号反映了X与Y的相互关系．显然用$(X-EX)(Y-EY)$的均值

$$E(X-EX)(Y-EY)$$

来刻画X与Y的相互关系是可行的.

又如：(X,Y)——语文成绩与数学成绩，则总成绩$X+Y$的方差

$$D(X+Y)=DX+DY+2E(X-EX)(Y-EY),$$

$DX+DY$：

是两个变量X与Y自身变化对$X+Y$的影响；

$E(X-EX)(Y-EY)$

是两个变量共同的影响.

$E(X-EX)(Y-EY)>0$时，总成绩的离散程度变大，$E(X-EX)(Y-EY)<0$时，离散程度变小.

可见, 通过$E(X-EX)(Y-EY)$的符号, 我们可了解两个变量之间变化的关系（变化趋势在平均意义上而言）. 我们引入如下定义.

定义1　称

$$\mathrm{Cov}(X,Y)=E(X-EX)(Y-EY)$$

为随机变量X与Y的协方差（covariance）. 易得

$$\mathrm{Cov}(X,Y)=EXY-EX\cdot EY$$

计算协方差常用此公式.

其中

$$EXY=\begin{cases}\sum\limits_i\sum\limits_j x_i y_j P(X=x_i,Y=y_j), & \text{离散型},\\ \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} xyf(x,y)\mathrm{d}x\mathrm{d}y, & \text{连续型}.\end{cases}$$

显然，方差是协方差的特例，对于随机向量 $X=(X_1,X_2,\cdots,X_n)$，定义它的方差为

$$(DX_1,DX_2,\cdots,DX_n).$$

协方差虽然可以刻画两个随机变量之间的相互关系, 其弊端是它具有量纲, 数值受到量纲的影响, 下面我们寻找更好的量去刻画这种相互关系.

对于随机变量X，若它的数学期望和方差都存在，而且$DX>0$. 称

$$X^{*}=\frac{X-EX}{\sqrt{DX}}$$

为X的标准化随机变量. 显然

$$EX^{*}=0, \quad DX^{*}=1.$$

例如：正态分布的标准化就是标准正态分布，均匀分布$U[a,b]$的标准化是均匀分布$U\left[-\sqrt{3},\sqrt{3}\right]$，但指数分布的标准化不再是指数分布.

显然X^*没有量纲, 其数值不受测量单位的影响, 此外，X^*与$X-EX$同号，Y^*与$Y-EY$同号，因而可以利用EX^*Y^*来表示随机变量X与Y的相互关系.

定义2 称

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

为X与Y的相关系数（correlation coefficient）.

规定：常数与任意的随机变量的相关系数为0.

当$\rho_{XY}>0$时，称两个随机变量正相关, 当$\rho_{XY}<0$时，称为负相关. 当$\rho_{XY}=0$时，称为不相关.

可以证明：（练习题）

若$ac>0$，则 $\rho_{aX+b,cY+d}=\rho_{XY}$，

若$ac<0$，则 $\rho_{aX+b,cY+d}=-\rho_{XY}$.

例1　设(X,Y)的联合分布律为

$X \backslash Y$	0	1	
-1	0.2	0.1	0.3
2	0.4	0.3	0.7
	0.6	0.4	

求协方差Cov(X,Y)及相关系数ρ_{XY}.

解： 先求出边缘分布(见表)，得

$$E(X)=\sum_i x_i p_{i\cdot}=1.1,\quad E(Y)=\sum_j y_j p_{\cdot j}=0.4,$$

$$E(X^2)=\sum_i x_i^2 p_{i\cdot}=3.1,$$

所以

$$D(X)=E(X^2)-[E(X)]^2=3.1-1.1^2=1.89.$$

同理

$$E(Y^2)=\sum_j y_j^2 p_{\cdot j}=0.4,\qquad D(Y)=0.24.$$

$$E(XY)=\sum_i\sum_j x_i y_j p_{ij}$$

$$=0\times 0.2+(-1)\times 0.1+0\times 0.4+2\times 0.3=0.5.$$

所以

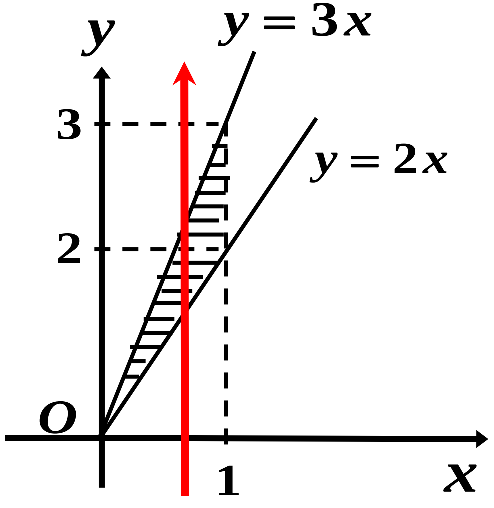
$$\text{Cov}(X,Y)=E(XY)-E(X)\cdot E(Y)=0.06.$$

$$\rho_{XY}=\frac{\text{Cov}(X,Y)}{\sqrt{D(X)D(Y)}}=\frac{0.06}{\sqrt{1.89}\sqrt{0.24}}=0.089.$$

例2　设(X,Y)的联合密度函数为

$$f(x,y)=\begin{cases}2, & 0<x<1,\quad 2x<y<3x\\ 0, & \text{其它}.\end{cases}$$

求协方差$\text{Cov}(X,Y)$及相关系数ρ_{XY}.



解：可以用以下公式直接计算

$E(X)$、$E(Y)$等.

$$E(X)=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} xf(x,y)\mathrm{d}x\mathrm{d}y=\int_0^1 \mathrm{d}x\int_{2x}^{3x} 2x\mathrm{d}y=2/3,$$

$$E(Y)=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} yf(x,y)\mathrm{d}x\mathrm{d}y=\int_0^1 \mathrm{d}x\int_{2x}^{3x} 2y\mathrm{d}y=5/3,$$

$$E(Y^2)=\int_0^1 \mathrm{d}x\int_{2x}^{3x} 2y^2\mathrm{d}y=19/6.$$

$$E(XY)=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} xyf(x,y)\mathrm{d}x\mathrm{d}y=\int_0^1 \mathrm{d}x\int_{2x}^{3x} 2xy\mathrm{d}y=5/4.$$

所以

$$D(X)=E(X^2)-[E(X)]^2=1/18,\quad D(Y)=7/18.$$

所以

$$\mathrm{Cov}(X,Y)=E(XY)-E(X)\cdot E(Y)=5/36.$$

$$\rho_{XY}=\frac{\mathrm{Cov}(X,Y)}{\sqrt{D(X)D(Y)}}=\frac{5}{2\sqrt{7}}\approx 0.9449.$$

二　协方差与相关系数的性质

协方差的性质

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

（2）　$\mathrm{Cov}(aX+b,cY+d)=ac\mathrm{Cov}(X,Y);$

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

（4）　$\mathrm{Cov}(X,X)=D(X).$

为了进一步认识相关系数的性质，先给出一个常用的定理.

定理1 柯西－施瓦兹不等式 (cauchy－schwarz)

对任意随机变量X与Y都有

$$\left|EXY\right|^2 \le EX^2 \cdot EY^2$$

等号成立当且仅当

$$P\left\{Y = t_0 X\right\} = 1.$$

这里t_0是某一个常数.

证明 对任意实数t，定义

$$u(t) = E(tX - Y)^2 = t^2 EX^2 - 2tEXY + EY^2,$$

显然对一切t，$u(t)\ge 0$，因此二次方程$u(t)=0$或者没有实根或者有一个重根. 所以判别式

$$\left[EXY\right]^2 - EX^2 \cdot EY^2 \le 0$$

即得证. 此外，方程$u(t)=0$有一个重根t_0存在的充要条件是

$$\left[EXY\right]^2 - EX^2 \cdot EY^2 = 0.$$

此时$E(t_0X-Y)^2=0$，因此

$$D(t_0X-Y)=0, \quad E(t_0X-Y)=0$$

从而

$$P\left\{t_0X-Y=0\right\}=1.$$

即为所证.

由定理可知：若两随机变量的方差存在, 则它们的协方差也存在.

相关系数的性质：

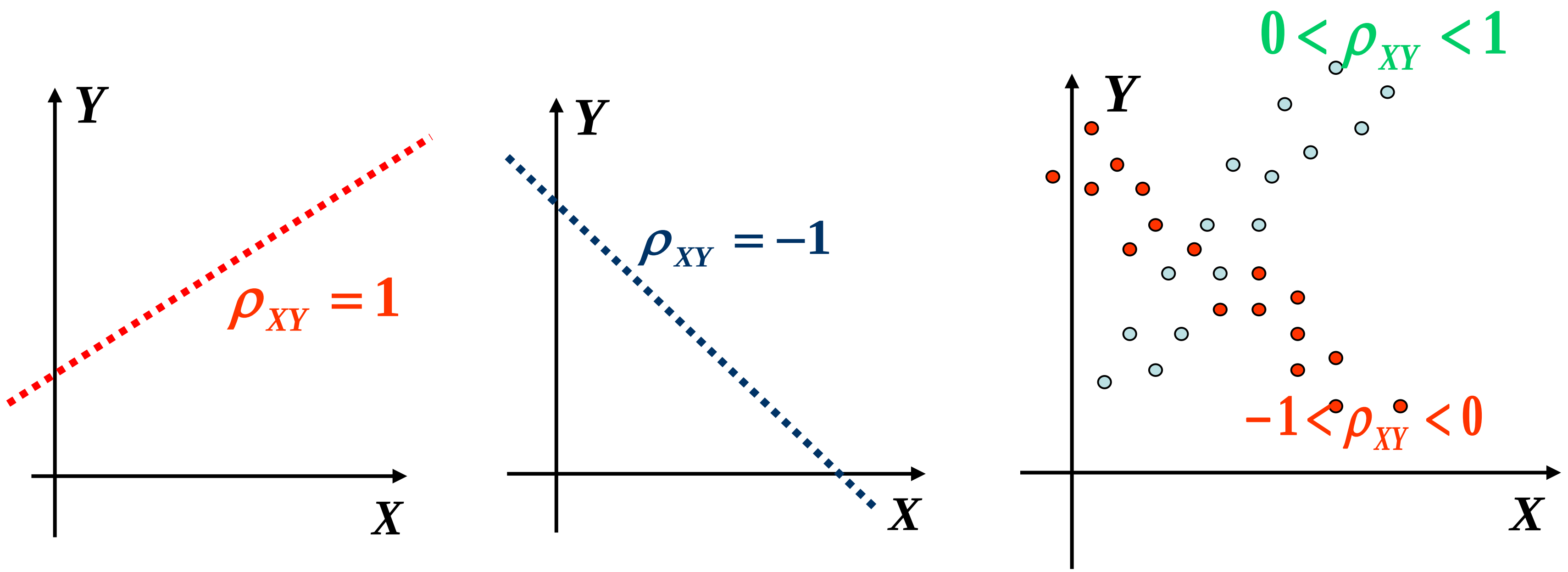
性质1　对相关系数ρ，成立$|\rho| \le 1$，并且

$$\rho = 1 \longleftrightarrow P\left\{\frac{X - EX}{\sqrt{DX}} = \frac{Y - EY}{\sqrt{DY}}\right\} = 1,$$

$$\rho = -1 \longleftrightarrow P\left\{\frac{X - EX}{\sqrt{DX}} = -\frac{Y - EY}{\sqrt{DY}}\right\} = 1.$$

ρ_{XY} 刻划了X,Y之间的线性相关程度, $\rho = \pm 1$时，X与Y存在完全线性关系. $\rho = 1$时, 称为完全正相关. $\rho = -1$时, 称为完全负相关. 极端情况是$\rho = 0$.

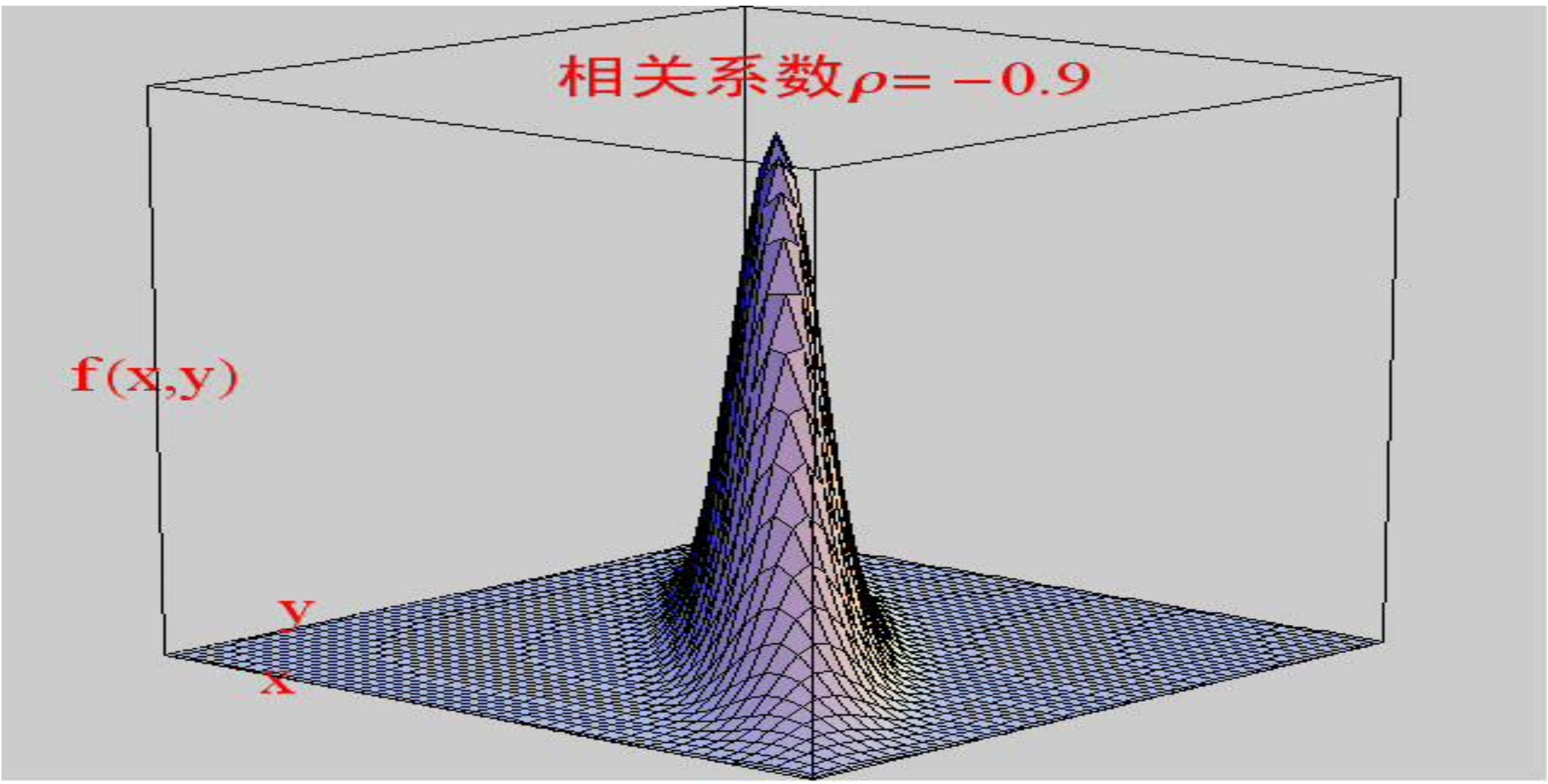
相关系数是随机变量之间线性关系强弱的一个度量(参见如下的示意图).



$|\rho|$值越接近于1，Y与X的线性相关程度越高;

$|\rho|$值越接近于0，Y与X的线性相关程度越弱;

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.



定义3 若随机变量X与Y的相关系数$\rho_{XY} = 0$,则称随机变量X与Y不相关.

性质2 下面是与不相关的几个等价命题：

（1） $\mathrm{Cov}(X, Y) = 0$;

（2） X与Y不相关;

（3） $EXY = EX \cdot EY$;

（4） $D(X + Y) = DX + DY$.

独立与不相关的关系如下：

性质3 若X与Y独立，则X与Y不相关.

证明 （略）

从该例可以看出：相关系数只是与线性关系程度的一种量度. 不相关的两个随机变量可能存在函数关系(因而关系很密切), 不过在正态分布的场合, 独立性与不相关是等价的.

性质4 对于二元正态分布

$$X\text{与}Y\text{不相关} \longleftrightarrow X\text{与}Y\text{独立}.$$

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

下面求 $\mathrm{Cov}(X,Y)$.

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

令 $\frac{x-\mu_1}{\sigma_1} = s$, $\frac{y-\mu_2}{\sigma_2} = t$, 则

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

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

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

$$=\rho\sigma_1\sigma_2\sqrt{\frac{2}{\pi}}\int_0^{+\infty}t^2e^{-\frac{1}{2}t^2}\mathrm{d}t\overset{t=\sqrt{2v}}{=}\frac{2\rho\sigma_1\sigma_2}{\sqrt{\pi}}\int_{-\infty}^{+\infty}v^{1/2}e^{-v}\mathrm{d}v,$$

$$=\frac{2\rho\sigma_1\sigma_2}{\sqrt{\pi}}\Gamma(\frac{3}{2})=\rho\sigma_1\sigma_2.\quad \rho_{XY}=\frac{\rho\sigma_1\sigma_2}{\sigma_1\sigma_2}=\rho.$$

若X与Y不相关，得$\rho=0$，因而X与Y独立.

例3 设$X \sim U[-\pi,\pi]$，$Y=\cos X$， $Z=\sin X$.

求$\mathrm{Cov}(Y,Z)$.

解：X的概率密度为 $f(x)=\begin{cases} 1/2\pi, & -\pi \le x \le \pi, \\ 0, & \text{其它}. \end{cases}$

$$E(Y)=\int_{-\infty}^{+\infty}\cos x f(x)\mathrm{d}x=\frac{1}{2\pi}\int_{-\pi}^{\pi}\cos x\mathrm{d}x=0;$$

$$E(Z)=\int_{-\infty}^{+\infty}\sin x f(x)\mathrm{d}x=\frac{1}{2\pi}\int_{-\pi}^{\pi}\sin x\mathrm{d}x=0;$$

$$E(YZ)=\int_{-\infty}^{+\infty}\sin x\cos x f(x)\mathrm{d}x=\frac{1}{4\pi}\int_{-\pi}^{\pi}\sin 2x\mathrm{d}x=0;$$

故 $\mathrm{Cov}(Y,Z)=E(YZ)-E(Y)E(Z)=0.$

这里Y,Z不相关，但它们不独立.

三　矩、协方差矩阵

下面给出矩的概念，数学期望、方差、协方差都可看作是某种矩.

定义4　对正整数k，称 $m_k = EX^k$ 为k阶原点矩.

由于$\left|X^{k-1}\right| \le 1 + \left|X^k\right|$，因此若$k$阶矩存在，则所有低阶矩都存在. 数学期望是一阶原点矩.

定义5　对正整数k，称$c_k = E(X - EX)^k$为k阶中心矩，方差是二阶中心矩.

对于多维随机向量，可定义各种混合矩，例如：

$E(X^k Y^l)$称它为X和Y的$k+l$阶混合原点矩；

$E(X - EX)^k (Y - EY)^l$称为$k+l$阶混合中心矩.

定义6 将随机向量(X_1,X_2)的四个二阶中心矩

$$c_{11}=\mathrm{Cov}(X_1,X_1),\quad c_{12}=\mathrm{Cov}(X_1,X_2),$$
$$c_{21}=\mathrm{Cov}(X_2,X_1),\quad c_{22}=\mathrm{Cov}(X_2,X_2),$$

排成矩阵的形式:

$$\begin{pmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{pmatrix}.$$

称此矩阵为(X_1,X_2)的**协方差矩阵.** 它是半正定矩阵.

类似定义n维随机变量$(X_1,X_2,\cdots,X_n)$的协方差矩阵.

$$C=(c_{ij})_{n\times n}$$

其中$c_{ij}=\mathrm{Cov}(X_i,X_j)=E\{[X_i-E(X_i)][X_j-E(X_j)]\}.$