第四章 随机变量的数字特征

§4.3 协方差 相关系数

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

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

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

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

协方差

定义1 设(X,Y)为二维随机变量，若

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

存在，则称其为随机变量X与Y的 **协方差**，记为$Cov(X,Y)$. 即

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

(1)若(X,Y)为**离散型**，$P\{X=x_i,Y=y_j\}=p_{ij}$，则

$$Cov(X,Y)=\sum_{i,j}[x_i-E(X)][y_j-E(Y)]p_{ij}$$

(2)若(X,Y)为**连续型**,其概率密度为$f(x,y)$, 则

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

➢计算公式：

1。$Cov(X,Y)=E(XY)-E(X)E(Y)$

2。$D(X\pm Y)=D(X)+D(Y)\pm 2Cov(X,Y)$

➢协方差的性质：

1。$Cov(X,X)=D(X)$；

2。$Cov(X, Y)=Cov(Y, X)$；

3。$Cov(aX, bY)=abCov(X, Y)$；　(*a*,*b*为常数)

4。$Cov(X_1+X_2, Y)=Cov(X_1, Y)+Cov(X_2, Y)$；

5。若X与Y独立，则 $Cov(X, Y)=0$.

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

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

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

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

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

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

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

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

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

例1 设(X,Y)具有概率密度

$$f(x,y)=\begin{cases}8xy, & 0\le x\le y\le 1,\\ 0, & \text{其它。}\end{cases}$$

求 $Cov(X,Y)$.

答： $\dfrac{4}{225}$

解：

$$E(X)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}xf(x,y)dxdy=\int_0^1 dx\int_x^1 x\cdot 8xydy=\frac{8}{15}$$

$$E(Y)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}yf(x,y)dxdy=\int_0^1 dx\int_x^1 y\cdot 8xydy=\frac{4}{5}$$

$$E(XY)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}xyf(x,y)dxdy=\int_0^1 dx\int_x^1 xy\cdot 8xydy=\frac{4}{9}$$

$$Cov(X,Y)=\frac{4}{9}-\frac{4}{5}\cdot\frac{8}{15}=\frac{4}{225}$$

定义2 设(X,Y)为二维随机变量，若$D(X)>0$, $D(Y)>0$

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

为随机变量X与Y的 **相关系数**，记为 ρ_{XY} .

特别，当$\rho_{XY}=0$时，称X与Y**不相关**.

[注] ★ ρ_{XY}是一个无量纲的量.

★ X, Y相互独立 $\Rightarrow$ X与Y不相关；反之不一定.

★ X与Y不相关 $\Leftrightarrow Cov(X,Y)=0$

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

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

相关系数的性质：

$1^\circ\ |\rho_{xy}| \le 1.$

$2^\circ\ |\rho_{xy}| = 1 \Leftrightarrow \exists$ 常数 a,b 使 $P\{Y = a + bX\} = 1.$

证明： 考虑用X的某个线性函数$a+bX$来近似表达Y，我们以均方误差 $e = E[(Y-(a+bX))^2]$ 来衡量以 $a+bX$ 来近似表达Y近似的好坏程度. 选取a，b 使 e 值最小.

$$e = E[(Y-(a+bX))^2]$$
$$= E(Y^2) + b^2E(X^2) + a^2 - 2bE(XY) - 2aE(Y) + 2abE(X)$$

$$\begin{cases} \dfrac{\partial e}{\partial a} = 2a + 2b\mathrm{E}(X) - 2\mathrm{E}(Y) = 0, \\ \dfrac{\partial e}{\partial b} = 2b\mathrm{E}(X^2) - 2\mathrm{E}(XY) + 2a\mathrm{E}(X) = 0, \end{cases} \Rightarrow \begin{cases} b_0 = \dfrac{\mathrm{Cov}(X,Y)}{\mathrm{D}(X)} \\ a_0 = \mathrm{E}(Y) - b_0\mathrm{E}(X) \end{cases}$$

$$\Rightarrow \boxed{\min_{a,b} E\{[Y-(a+bX)]^2\} = E\{[Y-(a_0+b_0X)]^2\} = (1-\rho^2_{XY})D(Y).}$$

$$\Rightarrow (1-\rho^2_{XY}) \ge 0 \Rightarrow |\rho^2_{XY}| \le 1$$

证2° $|\rho_{XY}|=1 \longrightarrow E\{[Y-(a_0+b_0X)]^2\}=0$

$$\| $$

$$D[Y-(a_0+b_0X)]+\{E[Y-(a_0+b_0X)]\}^2$$

$$\Longrightarrow \begin{cases} D[Y-(a_0+b_0X)]=0 \\ E[Y-(a_0+b_0X)]=0 \end{cases}$$

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

反之，若 $\exists a^*, b^*$, $P\{Y=a^*+b^*X\}=1$, $\Longrightarrow P\{Y-(a^*+b^*X)=0\}=1$

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

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

$$=E\{[Y-(a_0+b_0X)]^2\}=(1-\rho_{XY}^2)D(Y).$$

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

[注]

(1) 相关系数ρ_{XY}刻画了随机变量 Y 与X之间的"线性相关"程度: $|\rho_{XY}|$ 的值越接近于1， Y 与X 的线性相关程度越高； $|\rho_{XY}|$ 的值越接近于0， Y与X 的线性相关程度较弱.

(2) 当$\rho_{XY}=0$时,只说明Y 与X之间没有线性关系,并不能说明Y 与X之间没有其他函数关系,从而不能推出Y 与X独立.

例2 设(X, Y)均匀分布在以坐标原点为中心，R为半径的圆的内部，则随机变量X与Y不相关, 但X与Y也不相互独立.

解：由已知得

$$f(x,y)=\begin{cases}\dfrac{1}{\pi R^2}, & x^2+y^2\le R^2\\ 0, & \text{其它}\end{cases}$$

$$E(X)=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}xf(x,y)dxdy=\int_{-R}^{+R}dy\int_{-\sqrt{R^2-y^2}}^{\sqrt{R^2-y^2}}x\frac{1}{\pi R^2}dx$$

$$=0$$

$$E(Y)=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} yf(x,y)dxdy=\int_{-R}^{+R}dx\int_{-\sqrt{R^2-x^2}}^{\sqrt{R^2-x^2}} y\frac{1}{\pi R^2}dy$$

$$=0$$

$$E(XY)=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} xyf(x,y)dxdy=\int_{-R}^{+R}dx\int_{-\sqrt{R^2-x^2}}^{\sqrt{R^2-x^2}} xy\frac{1}{\pi R^2}dy$$

$$=0$$

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

故 $\rho_{XY}=0$

所以X与Y不相关

又因为

$$f_X(x)=\int_{-\infty}^{+\infty} f(x,y)dy=\begin{cases}\dfrac{2\sqrt{R^2-x^2}}{\pi R^2}, & -R\le x\le R\\ 0, & 其它\end{cases}$$

$$f_Y(y)=\int_{-\infty}^{+\infty} f(x,y)dx=\begin{cases}\dfrac{2\sqrt{R^2-y^2}}{\pi R^2}, & -R\le y\le R\\ 0, & 其它\end{cases}$$

所以 $f(x,y)\neq f_X(x)f_Y(y)$

X与Y也不相互独立

思考 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$\xi$和$\eta$的相关系数？

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

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

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

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

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

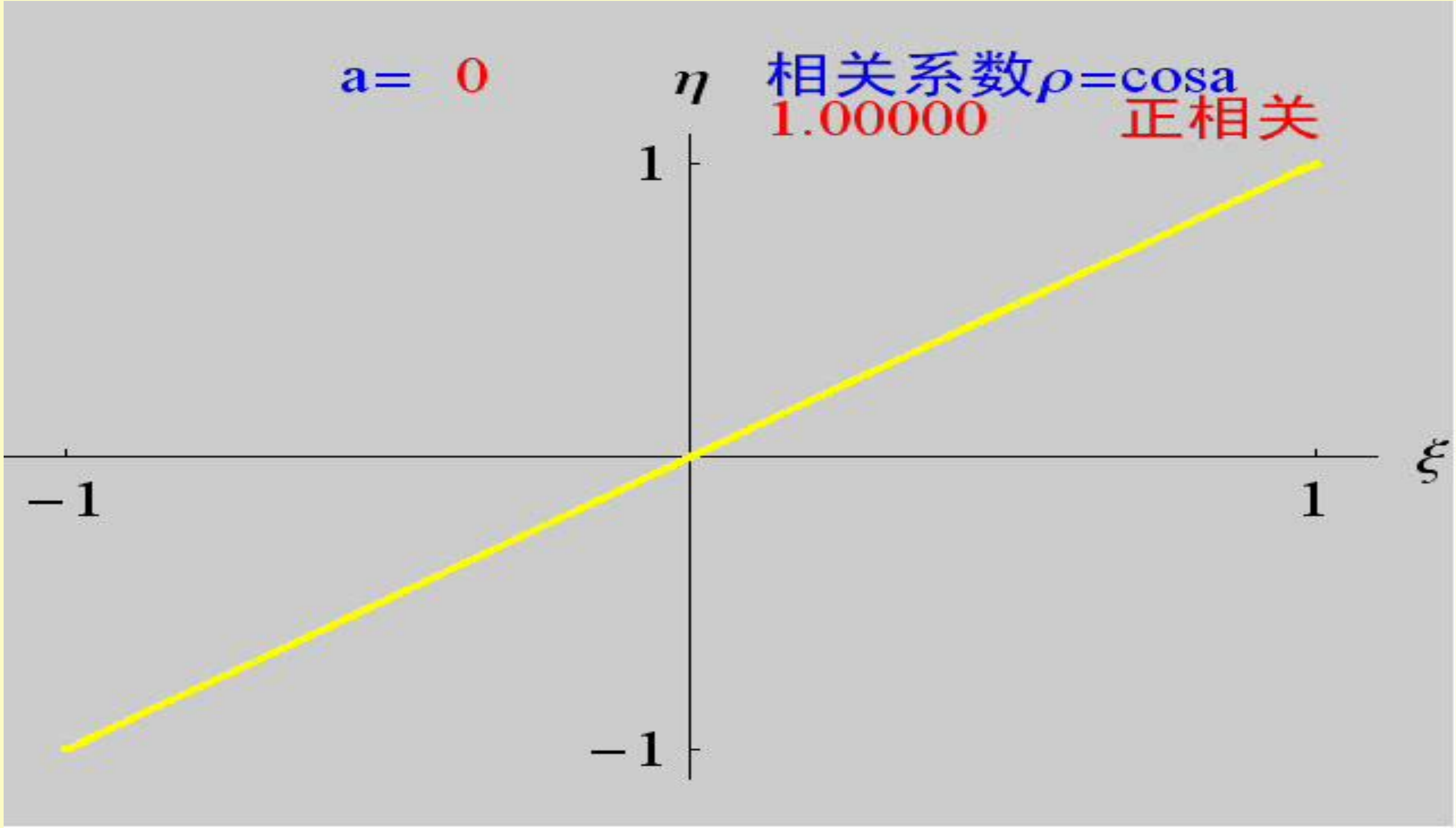
由以上数据可得相关系数为 $\rho=\cos a.$

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

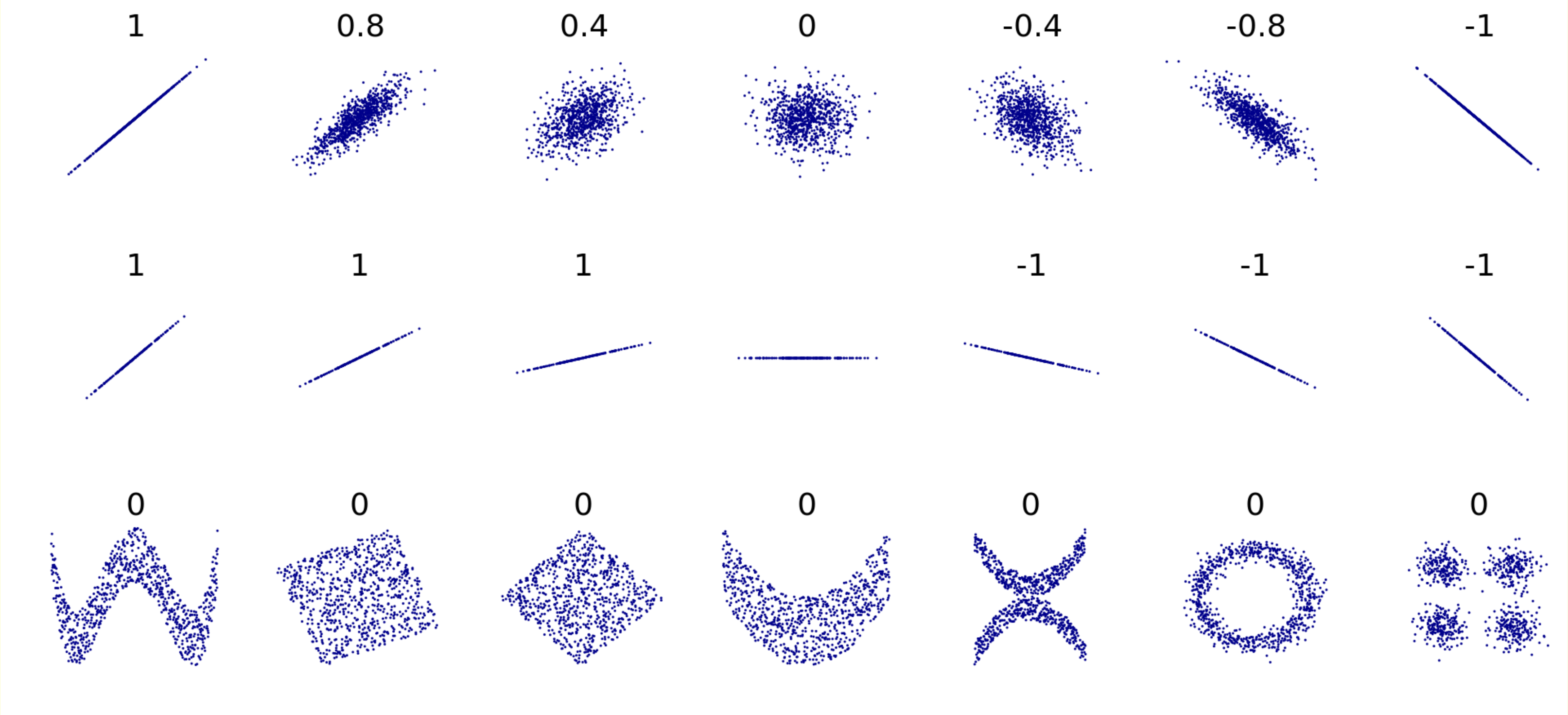
当$a=\dfrac{\pi}{2}$或$a=\dfrac{3\pi}{2}$时,$\rho=0$, ξ与η不相关.

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.



From https://en.wikipedia.org/wiki/Pearson_correlation_coefficient



几组(x, y)的点集，以及各个点集中x和y之间的相关系数。我们可以发现相关系数反映的是变量之间的线性关系和相关性的正负方向（第一行），而不是相关性的斜率（第二行），也不是各种非线性关系（第三排）。请注意：第二行的图中斜率为0，但相关系数是没有意义的，因为此时变量Y的方差是0

例3 设 $(X,Y)\sim N(\mu_1,\mu_2,\sigma_1^2,\sigma_2^2,\rho)$，则 X, Y相互独立 $\Leftrightarrow$ X, Y不相关.

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

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

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

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

令$t=\frac{1}{\sqrt{1-\rho^2}}[\frac{y-\mu_2}{\sigma_2}-\rho\frac{x-\mu_1}{\sigma_1}]$, $u=\frac{x-\mu_1}{\sigma_1}$

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

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

$$+\frac{\sigma_1\sigma_2\sqrt{1-\rho^2}}{2\pi}\int_{-\infty}^{+\infty}ue^{-\frac{u^2}{2}}du\cdot\int_{-\infty}^{+\infty}te^{-\frac{t^2}{2}}dt$$

$$=\frac{\sigma_1\sigma_2\rho}{2\pi}\int_{-\infty}^{+\infty}u^2e^{-\frac{u^2}{2}}du\cdot\int_{-\infty}^{+\infty}e^{-\frac{t^2}{2}}dt=\rho\sigma_1\sigma_2.$$

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

由上章§3.4例1知，X, Y相互独立$\Leftrightarrow\rho=0$，即$\rho_{XY}=0$, 亦即X,Y不相关．由此知二维正态随机变量的参数ρ是X与Y的相关系数，且X与Y相互独立与不相关是等价的．

例4 已知 $X \sim N(1,3^2), Y \sim N(0,4^2), \rho_{XY} = -\frac{1}{2}$,

设 $Z = \frac{X}{3} - \frac{Y}{2}$ ，求 $D(Z), \rho_{XZ}$.

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

$$= \frac{1}{9}D(X) - 2\cdot\frac{1}{2}\cdot\frac{1}{3}Cov(X,Y) + \frac{1}{2}D(Y)$$

$$= \frac{1}{9}D(X) - \frac{1}{3}\rho_{XY}\sqrt{D(X)}\sqrt{D(Y)} + \frac{1}{2}D(Y)$$

$$= \frac{1}{9}\cdot 9 - \frac{1}{3}(-\frac{1}{2})\sqrt{9}\sqrt{16} + \frac{1}{2}\cdot 16 = 11$$

$$\rho_{XZ} = \frac{Cov(X,Z)}{\sqrt{D(X)}\sqrt{D(Z)}} = \frac{1}{3\sqrt{11}}Cov(X,\frac{X}{3}-\frac{Y}{2})$$

$$= \frac{1}{3\sqrt{11}}[Cov(X,\frac{X}{3}) - Cov(X,\frac{Y}{2})] = \frac{1}{3\sqrt{11}}[\frac{1}{3}Cov(X,X) - \frac{1}{2}Cov(X,Y)]$$

$$= \frac{1}{3\sqrt{11}}[\frac{1}{3}\cdot 9 - \frac{1}{2}(-\frac{1}{2})\cdot 3\cdot 4] = \frac{2}{\sqrt{11}}$$

小结

1. 协方差与相关系数的概念及性质

(1) 定义

(2) 协方差的计算公式

(3) 性质

2. 相关系数的意义和性质

(1) 意义

(2) 性质

§4.5 矩 协方差矩阵

定义1 设X, Y为随机变量，k, l为正整数，称

$E(X^k)$ 为X的**k阶原点矩**（**k阶矩**）

$E\{[X-E(X)]^k\}$ 为X的 **k阶中心矩;**

$E(X^kY^l)$ 为X和Y的**$k+l$ 阶混合矩;**

$E\{[X-E(X)]^k[Y-E(Y)]^l\}$ 为X和Y的**$k+l$ 阶混合中心矩**

[注] (1)$E(X)$是X的一阶原点矩；

(2)$D(X)$是X的二阶中心矩；

(3)$Cov(X,Y)$是X和Y的二阶混合中心矩;

(4)以上数字特征都是随机变量函数的数学期望；

(5) 在实际应用中，高于4阶的矩很少使用.

三阶中心矩$\mu_3 = E\{[X - E(X)]^3\}$主要用来衡量随机变量的分布是否有偏 .

偏度：μ_3 / σ^3

四阶中心矩 $\mu_4 = E\{[X - E(X)]^4\}$ 主要用来衡量随机变量的分布在均值附 近的陡峭程度如何 .

峰度：μ_4 / σ^4 或 $\mu_4 / \sigma^4 - 3$

定义2 设n维随机变量$(X_1, X_2, \cdots, X_n)$的二阶混合中心矩都存在，称矩阵

$$C = \begin{pmatrix} c_{11} & c_{12} & \cdots & c_{1n} \\ c_{21} & c_{22} & \cdots & c_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ c_{n1} & c_{n2} & \cdots & c_{nn} \end{pmatrix}$$

为n维随机变量$(X_1, X_2, \cdots, X_n)$的**协方差矩阵**。

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

$$i, j = 1, 2, \cdots, n.$$

显然： $\boxed{C' = C.}$

且协方差矩阵为半正定矩阵. 为什么？

协方差矩阵的应用

协方差矩阵可用来表示多维随机变量的概率密度，从而可通过协方差矩阵达到对多维随机变量的研究

以二维正态随机变量 (X_1, X_2) 为例.概率密度

$$f(x_1,x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\exp\left\{\frac{-1}{2(1-\rho^2)}\left[\frac{(x_1-\mu_1)^2}{\sigma_1^2}\right.\right.$$

$$\left.\left.-2\rho\frac{(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2}+\frac{(x_2-\mu_2)^2}{\sigma_2^2}\right]\right\},$$

现在将上式中花括号内的式子写成矩阵形式, 为此引入下面的列矩阵

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \quad \boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}.$$

(X_1, X_2) 的协方差矩阵为

$$C = \begin{pmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix},$$

它的行列式 $\det C = \sigma_1^2\sigma_2^2(1-\rho^2)$, C 的逆矩阵为

$$C^{-1} = \frac{1}{\det C}\begin{pmatrix} \sigma_2^2 & -\rho\sigma_1\sigma_2 \\ -\rho\sigma_1\sigma_2 & \sigma_1^2 \end{pmatrix}$$

$$(\mathbf{x}-\mu)^{\mathrm{T}} C^{-1}(\mathbf{x}-\mu)$$

$$=\frac{1}{\det C}(x_1-\mu_1, x_2-\mu_2)\begin{pmatrix} \sigma_2^2 & -\rho\sigma_1\sigma_2 \\ -\rho\sigma_1\sigma_2 & \sigma_1^2 \end{pmatrix}\cdot\begin{pmatrix} x_1-\mu_1 \\ x_2-\mu_2 \end{pmatrix}$$

$$=\frac{1}{1-\rho^2}\left[\frac{(x_1-\mu_1)^2}{\sigma_1^2}-2\rho\frac{(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2}+\frac{(x_2-\mu_2)^2}{\sigma_2^2}\right].$$

于是 (X_1, X_2) 的概率密度可写成

$$f(x_1, x_2)$$

$$=\frac{1}{(2\pi)^{2/2}(\det C)^{1/2}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\mu)^{\mathrm{T}} C^{-1}(\mathbf{x}-\mu)\right\}.$$

推广 n维正态随机变量$(X_1, X_2, \cdots, X_n)$的概率密度表示为

$$f(x_1, x_2, \cdots, x_n)$$

$$= \frac{1}{(2\pi)^{n/2}(\det C)^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\mu)^{\mathrm{T}} C^{-1}(\mathbf{x}-\mu)\right\}.$$

其中$\mathbf{x} = (x_1, x_2, \cdots, x_n)^{\mathrm{T}}$,

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{pmatrix} = \begin{pmatrix} E(X_1) \\ E(X_2) \\ \vdots \\ E(X_n) \end{pmatrix}, \quad C = \begin{pmatrix} c_{11} & c_{12} & \cdots & c_{1n} \\ c_{21} & c_{22} & \cdots & c_{2n} \\ \vdots & \vdots & & \vdots \\ c_{n1} & c_{n2} & \cdots & c_{nn} \end{pmatrix},$$

n 维正态变量的性质

1° n 维正态随机变量 $(X_1, X_2, \cdots, X_n)$ 的每一个分量 $X_i, i = 1, 2, \cdots, n$ 都是正态随机变量；

反之，若 $X_1, X_2, \cdots, X_n$ 都是正态随机变量，且相互独立，则 $(X_1, X_2, \cdots, X_n)$ 是 n 维正态随机变量.

2° n 维随机变量 $(X_1, X_2, \cdots, X_n)$ 服从 n 维正态分布的充要条件：$X_1, X_2, \cdots, X_n$ 的任意的线性组合 $l_1X_1 + l_2X_2 + \cdots + l_nX_n$ 服从一维正态分布．

（必要性利用1°易证，充分性利用特征函数证明）

3° 若$(X_1, X_2, \cdots, X_n)$服从 n维正态分布，设 $Y_1, \cdots, Y_k$是 $X_j (j = 1,2,\cdots,n)$ 的线性函数，则$(Y_1, Y_2, \cdots, Y_k)$也服从多维正态分布 .

这一性质称为正态变量的线性变换不变性

4°设$(X_1, \cdots, X_n)$服从 n维正态分布，则“$X_1, X_2, \cdots, X_n$ 相互独立”与“$X_1, X_2, \cdots, X_n$ 两两不相关”是等价的 .

注：n维正态分布中一般要求协方差矩阵C 可逆。若C 不可逆，则退化为R^n的仿射子空间（维数等于C 的秩）中的正态分布。

小结

1. 矩是随机变量的数字特征.

$$\begin{cases} \text{数学期望 } E(X) \text{ 是 } X \text{ 的一阶原点矩；} \\ \text{方差 } D(X) \text{ 是 } X \text{ 的二阶中心矩；} \\ \text{协方差 } \mathrm{Cov}(X,Y) \text{ 是 } X \text{ 与 } Y \text{ 的二阶混合中心矩.} \end{cases}$$

2.正态变量是最重要的随机变量, 其性质一定要熟练掌握.

作业

- **第四章习题27(2)(4), 32, 33, 34(2), 36, 37**