三. 协方差 相关系数

1.协方差 相关系数的定义

2.协方差 相关系数的性质

3.相关与独立

4.二维正态分布的相关与独立

由期望的性质，若X与Y独立 ➡ $E(XY)=E(X)E(Y)$

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

$$\boxed{E\{(X-EX)(Y-EY)\}=0}$$

言外之意，若 X 与 Y 不独立，或许有$E(XY)-E(X)E(Y)\neq 0,$

或者说数值$E(XY)-E(X)E(Y)$大小可能反映X与Y间某种关系。

我们定义数值 $E(XY)-E(X)E(Y)=E\{(X-EY)(Y-EY)\}$

为X与Y的协方差。

1. 定义：

（1）协方差定义：设 (X,Y) 为二维随机变量，若 $E(X-EX)(Y-EY)$ 存在，则称其为随机变量 X 与 Y 的协方差，记为 $\mathrm{Cov}(X,Y)$，即：

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

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

$$=E(X-EX)^2=EX^2-(EX)^2$$

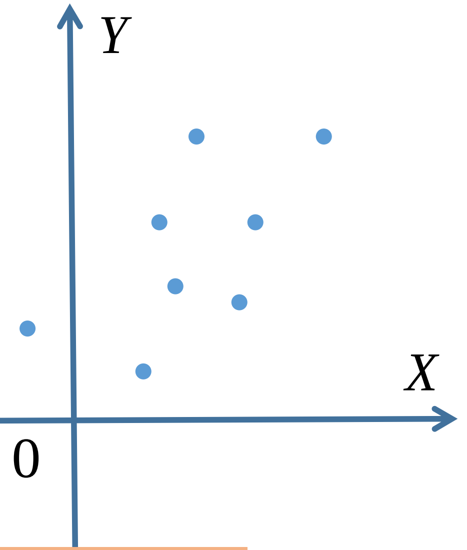
$$=D(X)$$

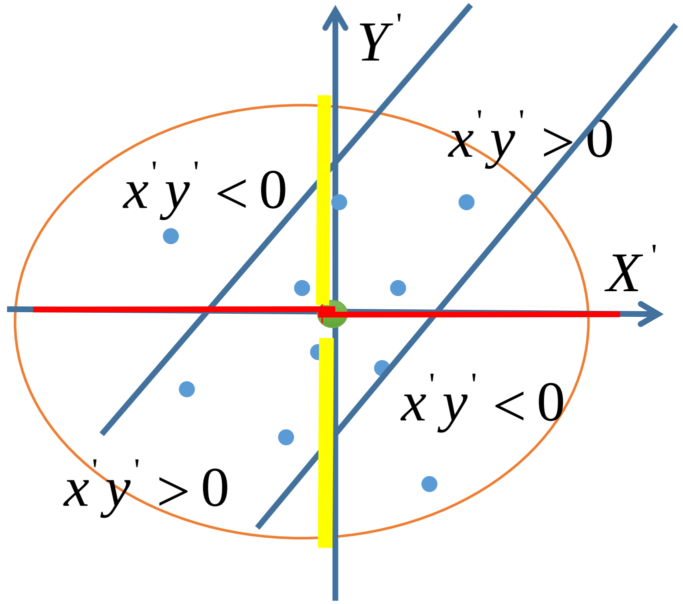
二维离散型随机变量 (X,Y) 等可能取 n 个点，$(x_1,y_1)(x_2,y_2)\cdots(x_n,y_n)$

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

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

$$=\frac{1}{n}\sum_{i,j=1}^{n}\left(\boxed{x_i}-EX\right)\left(\boxed{y_j}-EY\right)$$





(x_i',y_i')所在位置：主要1,3象限,2,4象限，或1234象限。

$$=\frac{1}{n}\sum_{i,j=1}^{n}x_i'y_j'$$ 的值的大小 $\begin{cases}\text{与}D(X)D(Y)\text{的大小有关}\\ \text{与}X\text{与}Y\text{间的线性关系趋势有关}\end{cases}$

去除$D(X),D(Y)$的因素：$\dfrac{\mathrm{Cov}(X,Y)}{\sqrt{D(X)}\sqrt{D(Y)}}$ 称为X与Y的相关系数。

(2) 相关系数：设 (X,Y) 为二维随机变量，若 $E(X-EX)(Y-EY)$ 存在，则称 $\rho_{XY}=\dfrac{\mathrm{Cov}(X,Y)}{\sqrt{D(X)}\sqrt{D(Y)}}$ 其为随机变量 X 与 Y 的相关系数。

$$\rho_{XY}=\frac{\mathrm{Cov}(X,Y)}{\sqrt{D(X)}\sqrt{D(Y)}}=\frac{E(X-EX)(Y-EY)}{\sqrt{D(X)}\sqrt{D(Y)}}$$

$$=E\left\{\frac{X-EX}{\sqrt{D(X)}}\right\}\left\{\frac{Y-EY}{\sqrt{D(Y)}}\right\}=EX^{*}Y^{*}$$

(协方差 $\mathrm{Cov}(X,Y)$ 的值含 DX，DY，及 X 与 Y 线性关系两因素)

(相关系数只含有 X 与 Y 线性关系一个因素)

2. 性质：(1) 协方差性质：

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

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

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

iii) 若 X 与 Y 独立则 $\mathrm{Cov}(X,Y)=0.$

iv) $\mathrm{Cov}(X,b)=0$

v) $\mathrm{Cov}(aX,bY)=ab\mathrm{Cov}(X,Y)$, a,b 为常数。

证：$\mathrm{Cov}(aX,bY)=E\{aX-E(aX)\}\{bY-E(bY)\}$

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

vi) $Cov(X_1 \pm X_2, Y) = Cov(X_1, Y) \pm Cov(X_2, Y)$

证：$Cov(X_1 + X_2, Y) = E\{(X_1 + X_2) - E(X_1 + X_2)\}\{Y - E(Y)\}$

$$= E\{(X_1 - EX_1) + (X_2 - EX_2)\}\{Y - EY\}$$

$$= E\{(X_1 - EX_1)(Y - EY) + (X_2 - EX_2)(Y - EY)\}$$

$$= E\{(X_1 - EX_1)(Y - EY)\} + E\{(X_2 - EX_2)(Y - EY)\}$$

$$= Cov(X_1, Y) + Cov(X_2, Y)$$

一般的，设$Y, X_1, X_2, \ldots X_n$都是随机变量，则

$$Cov\left(\sum_{i=1}^{n} X_i, Y\right) = \sum_{i=1}^{n} Cov(X_i, Y)$$

vii) $\boxed{\mathrm{Cov}(aX+b,\ Y)=a\mathrm{Cov}(X,\ Y)}+\underbrace{\mathrm{Cov}(b,\ Y)}_{=0}=a\mathrm{Cov}(X,\ Y)$

viii) 随机变量方差和协方差的关系：

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

例1.设随机变量$X_1,X_2,\ldots X_{10}$相互独立,且都服从参数为λ的的泊松分布,并设$X=\sum_{i=1}^{7}X_i$，$Y=\sum_{j=4}^{10}X_j$，求$\mathrm{Cov}(X,Y)$.

$$\mathrm{Cov}(X_1+X_2,Y)$$
$$=\mathrm{Cov}(X_1,Y)+\mathrm{Cov}(X_2,Y)$$

解：$\mathrm{Cov}(X,Y)=\mathrm{Cov}\left(\sum_{i=1}^{7}X_i,\sum_{j=4}^{10}X_j\right)$

$$=\mathrm{Cov}\left(\sum_{i=1}^{3}X_i+\sum_{i=4}^{7}X_i,\sum_{j=4}^{7}X_j+\sum_{i=8}^{10}X_i\right)$$

$$
\begin{aligned}
&= \mathrm{Cov}\left(\sum_{i=1}^{3} X_i + \sum_{i=4}^{7} X_i, \sum_{j=4}^{7} X_j + \sum_{i=8}^{10} X_i\right) = \underbrace{\mathrm{Cov}\left(\sum_{i=1}^{3} X_i, \sum_{j=4}^{7} X_j\right)}_{=0} \\
&+ \underbrace{\mathrm{Cov}\left(\sum_{i=1}^{3} X_i, \sum_{i=8}^{10} X_j\right)}_{=0} + \mathrm{Cov}\left(\sum_{i=4}^{7} X_i, \sum_{j=4}^{7} X_j\right) + \underbrace{\mathrm{Cov}\left(\sum_{i=4}^{7} X_i, \sum_{i=8}^{10} X_j\right)}_{=0} \\
&= \mathrm{Cov}\left(\sum_{i=4}^{7} X_i, \sum_{j=4}^{7} X_j\right) \\
&= \sum_{j=4}^{7} \mathrm{Cov}\left(X_j, X_j\right) \\
&= \sum_{j=4}^{7} D\left(X_j\right) = 4\lambda
\end{aligned}
$$

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

若X与Y独立，则$\mathrm{Cov}(X, Y) = 0$

(2) 相关系数的性质

i) $|\rho_{XY}| \leq 1.$ 随机变量X,Y的取值在一条直线上

ii) $|\rho_{XY}| = 1 \Leftrightarrow Y = aX + b(a,b$为常数，且$a \neq 0)$

证：i) $0 \leq D(\lambda X + Y) = \lambda^2 D(X) + 2\lambda Cov(X,Y) + D(Y)$，则

$$\{2Cov(X,Y)\}^2 - 4D(X)D(Y) \leq 0; \{2Cov(X,Y)\}^2 \leq 4D(X)D(Y)$$

$$\rho_{XY}^2 = \left\{\frac{\mathrm{Cov}(X,Y)}{\sqrt{D(X)}\sqrt{D(Y)}}\right\}^2 = \frac{\{\mathrm{Cov}(X,Y)\}^2}{D(X)D(Y)} \leq 1; \Longrightarrow |\rho_{XY}| \leq 1.$$

ⅱ) 若$Y=aX+b\left(a,b\text{为常数，且}a\neq 0\right)$

$$\rho_{XY}=\frac{\mathrm{Cov}\left(X,Y\right)}{\sqrt{D\left(X\right)}\sqrt{D\left(Y\right)}}=\frac{\mathrm{Cov}\left(X,aX+b\right)}{\sqrt{D\left(X\right)}\sqrt{D\left(aX+b\right)}}=\frac{a}{|a|}\ ;\quad |\rho_{XY}|=1.$$

反之，若$|\rho_{XY}|=1$， 考察 $D\boxed{\left(X\sqrt{D(Y)}-Y\sqrt{D(X)}\right)}$

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

$$=2D(Y)D(X)-2\sqrt{D(X)}\sqrt{D(X)}\rho_{XY}\sqrt{D(X)}\sqrt{D(Y)}=0$$

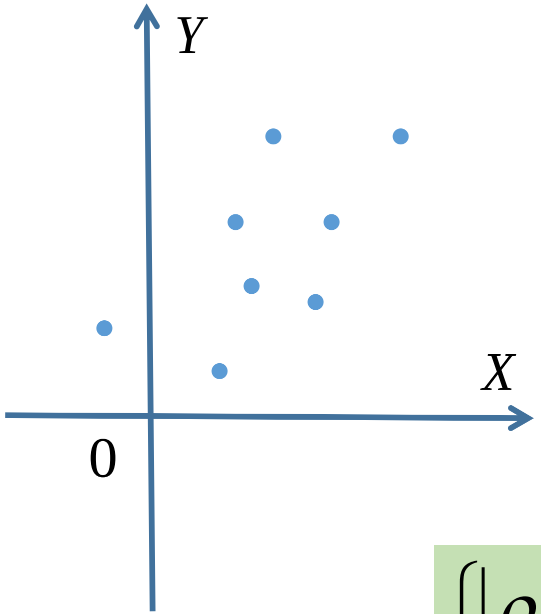
即 $D(\quad)=0$； ⟶ 则$(\quad)=$常数； ⟶ 于是$E(\quad)=(\quad)$；

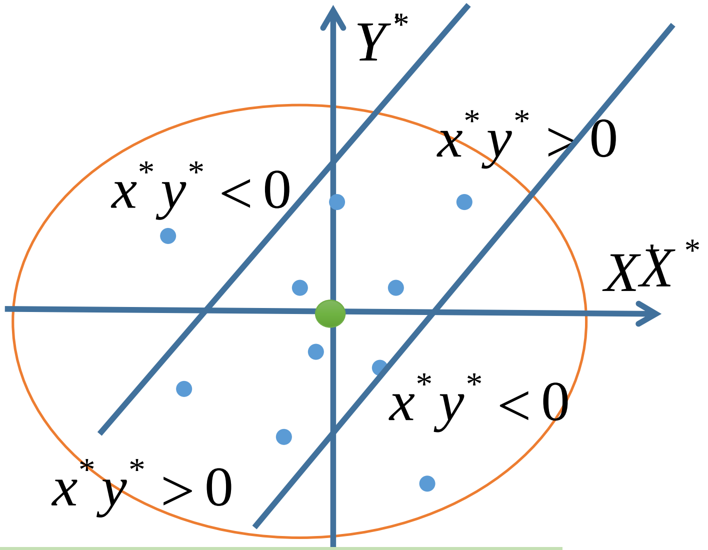
$$E\left(X\sqrt{D(Y)}-Y\sqrt{D(X)}\right)=\left(X\sqrt{D(Y)}-Y\sqrt{D(X)}\right);$$

$$Y=X\underbrace{\sqrt{D(Y)}/\sqrt{D(X)}}_{a}\underbrace{-E\left(X\sqrt{D(Y)}-Y\sqrt{D(X)}\right)/\sqrt{D(X)}}_{b}$$

二维随机变量 (X,Y)是平面上等可能 n 个点，$(x_1,y_1)(x_2,y_2)\cdots(x_n,y_n)$

$$\mathrm{Cov}(X,Y)$$
$$=E(X-EX)(Y-EY)$$
$$=\frac{1}{n}\sum_{i,j=1}^{n}(x_i-EX)(y_j-EY)$$





(x_i^*,y_i^*)所在位置:1,3象限,2,4象限，或1234象限。

$$\begin{cases} |\rho_{XY}|=1 & X与Y完全相关 \\ \rho>0 & X与Y正相关 \\ \rho<0 & X与Y负相关 \\ \rho=0 & X与Y不相关 \end{cases}$$

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

例2.已知ρ_{XY}, $W = aX + b, V = cY + d$，求ρ_{WV}

解：$\rho_{\mathrm{WV}} = \dfrac{\operatorname{cov}(W,V)}{\sqrt{DW}\sqrt{DV}}$

$$Cov(W,V) = Cov(aX + b, cY + d) = acCov(X,Y)$$

$$DW = D(aX + b) = a^2 DX, \quad DV = D(cY + d) = c^2 DY$$

$$\rho_{\mathrm{WV}} = \frac{ac\operatorname{cov}(X,Y)}{\sqrt{a^2 DX}\sqrt{c^2 DY}} = \frac{ac\operatorname{cov}(X,Y)}{\sqrt{a^2 c^2}\sqrt{DX}\sqrt{DY}} = \begin{cases} -\rho_{XY} & ac < 0 \\ \rho_{XY} & ac > 0 \end{cases}$$

3.相关与独立：

$$X\text{与}Y\text{独立}\ \underset{\text{不能决定}}{\overset{\text{必然决定}}{\rightleftarrows}}\ X\text{与}Y\text{不相关}$$

没有任何关系 没有线性关系

X与Y独立唯一判断方法：

$$F(x,y)=F_X(x)F_Y(y)$$

$$\Leftrightarrow\begin{cases} p_{ij}=p_{i\bullet}p_{\bullet j}\text{对所有}i,j\text{成立} \\ f(x,y)=f_X(x)f_Y(y) \end{cases}$$

X与Y不相关判断方法：

$$\begin{cases} \rho_{XY}=0 \\ \mathrm{Cov}(X,Y)=0 \\ E(XY)=E(X)E(Y) \\ D(X\pm Y)=D(X)+D(Y) \end{cases}$$

特别的有，X与Y独立$\Rightarrow E(XY)=E(X)E(Y)$

但 X与Y独立$\not\Leftarrow E(XY)=E(X)E(Y)$

4.二维正态分布的相关与独立：

$$\boxed{\mathrm{Cov}(X,Y)=\rho\sigma_1\sigma_2;\ \rho_{XY}=\rho}$$

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

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

$$=\rho\sigma_1\sigma_2$$

$$\rho_{XY}=\frac{Cov(X,Y)}{\sqrt{D(X)}\sqrt{D(Y)}}=\frac{\rho\sigma_1\sigma_2}{\sigma_1\sigma_2}=\rho$$

二维正态分布的重要性质：

$$\boxed{X\text{与}Y\text{独立}\Leftrightarrow\rho=0\qquad X\text{与}Y\text{独立}\Leftrightarrow X\text{与}Y\text{不相关}}$$

在二维正态分布中，随机变量X与Y要么独立，要么只存在线性关系。

正态分布的重要性质：

1. $X \sim N(\mu, \sigma^2)$，则 $aX + b \sim N(a\mu + b, a^2\sigma^2)$；特别地，$\dfrac{X-\mu}{\sigma} \sim N(0,1)$

2. 若 $(X, Y) \sim N(\mu_1, \mu_{2,}, \sigma_1^{\ 2}, \sigma_{,2}^2, \rho)$，$W = aX + bY, V = cX + dY$

则 $(W, V) \sim N(\mu_W, \mu_V, \sigma_W^2, \sigma_V^2, \rho_{WV})$

3. $(X, Y) \sim N(\mu_1, \mu_{2,}, \sigma_1^{\ 2}, \sigma_{,2}^2, \rho)$，则 $X \sim N(\mu_1, \sigma_{,1}^2)$，$Y \sim N(\mu_2, \sigma_{,2}^2)$

且均与 ρ 无关。

4. 二维正态分布两个随机变量 X 与 Y 独立 $\Leftrightarrow \rho = 0$

5.二维正态分布具有可加性：$X_i \sim N\left(\mu_i,\sigma_i^2\right), i=1,2\cdots n,$ 相互独立。

$$\sum_{i=1}^{n} a_i X_i \sim N\left(\sum_{i=1}^{n} a_i\mu_i, \sum_{i=1}^{n} a_i^2\sigma_i^2\right).$$

一般的，X_1，$X_2\cdots X_n$独立同分布，$X_i \sim N\left(\mu,\sigma^2\right), i=1,2\cdots n,$

$$\sum_{i=1}^{n} X_i \sim N\left(n\mu, n\sigma^2\right), \overline{X}_n=\frac{1}{n}\sum_{i=1}^{n} X_i \sim N\left(\mu,\frac{\sigma^2}{n}\right).$$

6. 若 $(X,Y)\sim N\left(\mu_1,\mu_{2,},\sigma_1^{\;2},\sigma_{,2}^2,\rho\right)$, 则 $X\pm Y\sim N\left(E(X\pm Y), D(X\pm Y)\right)$

$$X\pm Y\sim N\left(\mu_1\pm\mu_2,\sigma_1^2+\sigma_2^2\pm 2\rho\sigma_1\sigma_2\right)$$