第三节 协方差及相关系数

- 一、协方差与相关系数的概念及性质
- 二、相关系数的意义
- 三、协方差矩阵
- 四、内容小结

一、协方差与相关系数的概念及性质

1. 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

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

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

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

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

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

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

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

$$+2E[(X-E(X))(Y-E(Y))]$$

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

协方差

2. 定义3.7

设(X,Y)是二维随机变量，则称

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

为随机变量X与Y的协方差，记作Cov(X,Y)

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

而

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

称为随机变量 X 与 Y 的相关系数.

当 $\rho_{XY}=0$ 时,称 X 和 Y 不相关.

3.说明 (1) 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.

(2) 若随机变量 X 和 Y 相互独立

$$\Rightarrow \mathrm{Cov}(X,Y) = E\{[X-E(X)][Y-E(Y)]\}$$
$$= E[X-E(X)]E[Y-E(Y)] = 0.$$

(3) 若随机变量 X 和 Y 相互独立

$$\Rightarrow D(X+Y) = D(X)+D(Y)$$
$$+2E\{[X-E(X)][Y-E(Y)]\}$$
$$= D(X)+D(Y)+2\mathrm{Cov}(X,Y) = D(X)+D(Y).$$

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

4. 协方差的计算公式

(1) $\operatorname{Cov}(X,Y)=E(XY)-E(X)E(Y)$;

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

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

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

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

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

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

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

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

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

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

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

5. 性质

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

(2) $\mathrm{Cov}(aX,bY)=ab\,\mathrm{Cov}(X,Y)$ $\quad a,b$为常数;

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

例1 设$(X,Y)\sim N(\mu_1,\mu_2,\sigma_1^2,\sigma_2^2,\rho)$,试求 X 与 Y 的相关系数.

解 由 $p(x,y)=\dfrac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\exp$

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

$$\Rightarrow p_X(x)=\frac{1}{\sqrt{2\pi}\sigma_1}e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}},-\infty<x<+\infty,$$

$$p_Y(y)=\frac{1}{\sqrt{2\pi}\sigma_2}e^{-\frac{(y-\mu_2)^2}{2\sigma_2^2}},-\infty<y<+\infty.$$

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

而

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

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

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

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

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

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

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

$$+\frac{\sigma_1\sigma_2\sqrt{1-\rho^2}}{2\pi}\left(\int_{-\infty}^{+\infty}ue^{-\frac{u^2}{2}}\,\mathrm{d}u\right)\left(\int_{-\infty}^{+\infty}te^{-\frac{t^2}{2}}\,\mathrm{d}t\right)$$

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

故有 $\mathrm{Cov}(X,Y)=\rho\sigma_1\sigma_2.$

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

结论:

(1)二维正态分布密度函数中,参数 ρ 代表了 X 与 Y 的相关系数;

(2) 对于二维正态随机变量 (X,Y),

X 与 Y 不相关 $\Leftrightarrow$ X 与 Y 相互独立.

例2 已知随机变量X,Y分别服从$N(1,3^2),N(0,4^2)$, $\rho_{XY}=-1/2$,设 $Z=X/3+Y/2$.

(1) 求 Z 的数学期望和方差 .

(2) 求 X 与 Z 的相关系数 .

解 (1)由$E(X)=1,\ D(X)=9,\ E(Y)=0,\ D(Y)=16.$

得
$$E(Z)=E(\frac{X}{3}+\frac{Y}{2})$$
$$=\frac{1}{3}E(X)+\frac{1}{2}E(Y)=\frac{1}{3}.$$

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

$$=\frac{1}{9}D(X)+\frac{1}{4}D(Y)+\frac{1}{3}\mathrm{Cov}(X,Y)$$

$$=\frac{1}{9}D(X)+\frac{1}{4}D(Y)+\frac{1}{3}\rho_{XY}\sqrt{D(X)}\sqrt{D(Y)}$$

$$=1+4-2=3.$$

$$(2)\ \operatorname{Cov}(X,Z)=\operatorname{Cov}\left(X,\frac{X}{3}+\frac{Y}{2}\right)$$

$$=\frac{1}{3}\operatorname{Cov}(X,X)+\frac{1}{2}\operatorname{Cov}(X,Y)$$

$$=\frac{1}{3}D(X)+\frac{1}{2}\rho_{XY}\sqrt{D(X)}\sqrt{D(Y)}=3-3=0.$$

故 $\rho_{XY}=\operatorname{Cov}(X,Z)\big/\left(\sqrt{D(X)}\sqrt{D(Z)}\right)=0.$

例3 设 $E(X)=-2$, $E(Y)=2$, $D(X)=1$, $D(Y)=4$, $\rho_{xy}=-0.5$, 试根据切比谢夫不等式估计：$P\{|X+Y|\geq 6\}$.

解 $E(X+Y)=E(X)+E(Y)=0$

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

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

$$=1+4+2\cdot 1\cdot 2\cdot(-0.5)=3$$

$$P\{|X+Y|\geq 6\}=P\{|(X+Y)-E(X+Y)|\geq 6\}$$

$$\therefore \quad P\{|X+Y| \geq 6\}$$

$$= P\{|(X+Y) - E(X+Y)| \geq 6\}$$

$$\leq \frac{D(X+Y)}{6^2} = \frac{1}{12}.$$

$$\boxed{P\{|X - E(X)| \geq \varepsilon\} \leq \frac{D(X)}{\varepsilon^2}}$$

2. 相关系数的意义

当$|\rho_{XY}|$较大时，表明 X,Y 的线性关系联系较紧密.

当$|\rho_{XY}|$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.

3. 注意

(1) 不相关与相互独立的关系

$$\text{相互独立} \;\substack{\longrightarrow \\ \not\longleftarrow}\; \text{不相关}$$

(2) 不相关的充要条件

1° X, Y 不相关 $\Leftrightarrow \rho_{XY} = 0;$

2° X, Y 不相关 $\Leftrightarrow \mathrm{Cov}(X, Y) = 0;$

3° X, Y 不相关 $\Leftrightarrow E(XY) = E(X)E(Y).$

4. 相关系数的性质

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

(2) $|\rho_{XY}| = 1$的充要条件是,存在常数a, b使

$$P\{Y = a + bX\} = 1.$$

证(略)。

三、协方差矩阵(了解)

设 n 维随机变量 $(X_1, X_2, \cdots, X_n)$ 的二阶混合中心矩

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

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

都存在,则称矩阵

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

为 n 维随机变量的协方差矩阵.

例如　二维随机变量 (X_1, X_2) 的协方差矩阵为

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

其中　$c_{11} = E\{[X_1 - E(X_1)]^2\},$

$$c_{12} = E\{[X_1 - E(X_1)][X_2 - E(X_2)]\},$$

$$c_{21} = E\{[X_2 - E(X_2)][X_1 - E(X_1)]\},$$

$$c_{22} = E\{[X_2 - E(X_2)]^2\}.$$

由于 $c_{ij} = c_{ji}\ (i, j = 1,2,\cdots,n)$,所以协方差矩阵为对称的非负定矩阵.

协方差矩阵的应用

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

以二维正态随机变量(X_1, X_2)为例.

由于

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

引入矩阵 $X = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \quad \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}.$

及(X_1, X_2)的协方差矩阵$C = \begin{pmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{pmatrix},$

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

由此可得

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

$$=\frac{1}{\sigma_1^2\sigma_2^2(1-\rho^2)}\begin{pmatrix} \sigma_2^2 & -\rho\sigma_1\sigma_2 \\ -\rho\sigma_1\sigma_2 & \sigma_1^2 \end{pmatrix}.$$

由于

$$(X-\mu)^T C^{-1}(X-\mu)=$$

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

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

于是(X_1, X_2)的概率密度可写成

$$p(x_1, x_2)$$

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

推广

n 维随机变量$(X_1, X_2, \cdots, X_n)$的概率密度可表示为 $p(x_1, x_2, \cdots, x_n)$

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

其中$X = (x_1, x_2, \cdots, x_n)^T$,

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

四、内容小结

协方差与相关系数的定义

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

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

称 $\rho_{XY}=\dfrac{\mathrm{Cov}(X,Y)}{\sqrt{D(X)}\cdot\sqrt{D(Y)}}$ 为随机变量 X 与 Y 的相关系数.

协方差的性质

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

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

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

相关系数的意义

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

当 $|\rho_{XY}|$ 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.