4.3 协方差、相关系数

数学科学学院

电子科技大学
UNIVERSITY OF ELECTRONIC SCIENCE
AND TECHNOLOGY OF CHINA

协方差与相关系数

下面介绍的协方差、相关系数是描述随机变量之间相互关系的数字特征.

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

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

定义 若$E\{[X-E(X)][Y-E(Y)]\}$存在，称

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

为随机变量(X,Y)的协方差(Covariance).

有 $D(X)=Cov(X,X)$;

$$D(X\pm Y)=D(X)+D(Y)\pm 2Cov(X,Y)$$

性质 协方差性质

1) 对称性 $Cov(X,Y)= Cov(Y,X)$；

2) 齐次性 $Cov(aX+c, bY+d) = ab\, cov(X,Y)$；

3) 双线性性 $Cov(X_1+X_2, Y) = Cov(X_1, Y)+Cov(X_2, Y)$，

$$Cov(X, Y_1+Y_2) = Cov(X, Y_1)+Cov(X, Y_2)$$

证明：

$$Cov(aX,bY)=E\{[aX+c-E(aX+c)][bY+d-E(bY+d)]\}$$
$$=ab\,E\{[X-E(X)][Y-E(Y)]\}=ab\,cov(X,Y)$$

常用计算公式： $Cov(X,Y)=E(XY)-E(X)E(Y)$

定义 设二维随机变量X,Y的方差存在，且$D(X)>0, D(Y)>0$称

$$\rho_{XY}=\frac{E\{[X-E(X)][Y-E(Y)]\}}{\sqrt{D(X)D(Y)}}$$

为随机变量X与Y的**相关系数(Correlation Coefficient)**.

注： 1) ρ_{XY}无量纲

相关系数是标准化随机变量的协方差

$$2)\ \rho_{XY}=E\left\{\frac{[X-E(X)]}{\sqrt{D(X)}}\frac{[Y-E(Y)]}{\sqrt{D(Y)}}\right\}$$

$$=E\{X^*Y^*\}=Cov(X^*,Y^*)$$

性质 相关系数性质 设随机变量X,Y的相关系数$\rho_{X,Y}$存在，则

1） $-1 \le \rho_{X,Y} \le 1$;

2） $\rho_{X,Y}=1$或-1 **等价于** X与Y以概率1线性相关. 即存在a，$b(a \neq 0)$使得 $P\{Y=aX+b\}=1$

思路： 使用标准化随机变量简化计算

证： 1） $0 \le D(X^* \pm Y^*) = D(X^*) + D(Y^*) \pm 2\rho_{X,Y} = 2(1 \pm \rho_{X,Y})$,

因此$(1 \pm \rho_{X,Y}) \ge 0$，进而 $-1 \le \rho_{X,Y} \le 1$

2） $\rho_{X,Y}=-1$时 $D(X^*+Y^*)=0$，$E(X^*+Y^*)=0$

可以推出 $P(X^*+Y^*=0)=1$

还原可得

$$P\left(Y=-\frac{\sqrt{D(Y)}}{\sqrt{D(X)}}X+\frac{\sqrt{D(Y)}}{\sqrt{D(X)}}E(X)+E(Y)\right)=1$$

系数小于0

$\rho_{X,Y}=1$时同理

期望与概率为0的区域无关

反过来，若 $P\{Y=aX+b\}=1$，则

$$E(Y)=aE(X)+b, D(Y)=a^2D(X)$$

计算协方差可得

$$\begin{aligned}Cov(X,Y)&=E\{[X-E(X)][Y-E(Y)]\}\\&=E\{[X-E(X)][aX+b-E(aX+b)]\}\\&=aD(X)\end{aligned}$$

因此 $$\rho_{X,Y}=\frac{Cov(X,Y)}{\sqrt{D(X)}\sqrt{D(Y)}}=\frac{a}{|a|}=\pm 1$$

设随机变量X,Y的相关系数存在，

1）$\rho_{X,Y}=1$，称X, Y正相关；　$P\{Y=aX+b\}=1.$　$(a>0)$

2）$\rho_{X,Y}=-1$，称X, Y负相关；$(a<0)$

3）$\rho_{X,Y}=0$，称X, Y不相关.

例：设$ac\neq 0$，讨论$\rho_{aX+b,cY+d}$和$\rho_{X,Y}$的关系。

方法：标准化后算协方差。

$$\rho_{aX+b,cY+d}=\begin{cases}\rho_{X,Y}, & ac>0.\\ -\rho_{X,Y}, & ac<0.\end{cases}$$

注1 $\rho_{X,Y}=0$仅说明X,Y之间没有线性关系，但可能有其他非线性关系.

注2 若X,Y相互独立，则$\rho_{X,Y}=0$。 因为独立时$E(XY)=E(X)E(Y)$

逆不真： 由$\rho_{X,Y}=0$不能得到X与Y相互独立.

反例：若X服从对称分布，则$Y=X^2$与X不相关但不独立

例4：举出一个例子：两个随机变量不相关但不独立

X	-1	0	1
P	0.4	0.2	0.4

取值和概率
都要对称

注意到：$P(X^2=1,\ X=1)=P(X=1)$ 因此$Y=X^2$与X不独立

例：若X服从标准正态分布，证明$Y = X^2$与X不相关但不独立

注意到：$P(X^2 \le a,\ X \le b) = P(-\sqrt{a} \le X \le \sqrt{a},\ X \le b)$
$= P(-\sqrt{a} \le X \le min\{\sqrt{a},\ b\})$

只要 $min\{\sqrt{a},\ b\} = \sqrt{a}$,

则$P(X^2 \le a,\ X \le b) = P(X^2 \le a)$，就不独立。

另一方面： $E(X) = 0, E(X^2) = D(X) = 1, E(X^3) = 0$

因此 $cov(X, X^2) = E(X^3) - E(X)E(X^2) = 0$

最后，因为 $D(X) > 0$，$D(X^2) > 0$ 所以相关系数为0

例 设随机变量$X \sim U(0, 2\pi)$，$Y=\cos X$, $Z=\cos(X+\pi/2)$,讨论Y与Z的相关性.

解：可以证明$cov(Y,Z)=0$(书上有计算)，并且显然Y,Z的方差非0，因此X与Y不相关。

但因

$$Y^2+Z^2=1,$$

所以Y与Z有确定的函数关系.

是否会感觉：相关系数没什么用？

相关系数可以明确计算，在科研、现实中都大有用处。

例 设(X,Y)服从二维正态分布$N(0,1;0,1;\rho)$，计算(X,Y)的相关系数ρ_{XY}。

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

相关系数是标准化随机变量的协方差

二维正态分布$N(0,1;0,1;\rho)$的密度函数

$$f(x,y)=\frac{1}{2\pi\sqrt{1-\rho^2}}e^{-\frac{x^2-2\rho xy+y^2}{2(1-\rho^2)}}$$

计算：$E(XY)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{xy}{2\pi\sqrt{1-\rho^2}}e^{-\frac{x^2-2\rho xy+y^2}{2(1-\rho^2)}}dxdy$

思路：选择一个变量，将其配凑成完全平方项

$$E(XY)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{xy}{2\pi\sqrt{1-\rho^2}}e^{-\frac{x^2-2\rho xy+\rho^2y^2-\rho^2y^2+y^2}{2(1-\rho^2)}}dxdy$$

$$=\int_{-\infty}^{\infty}ye^{-\frac{y^2}{2}}\left[\int_{-\infty}^{\infty}\frac{x}{2\pi\sqrt{1-\rho^2}}e^{-\frac{(x-\rho y)^2}{2(1-\rho^2)}}dx\right]dy$$

变成$\frac{t^2}{2}$

$$=\int_{-\infty}^{\infty}ye^{-\frac{y^2}{2}}\left[\int_{-\infty}^{\infty}\frac{t\sqrt{1-\rho^2}+\rho y}{2\pi}e^{-\frac{t^2}{2}}dt\right]dy=\rho$$

因此二维正态随机变量不相关等价于$\rho=0$

n维随机变量的协方差矩阵

定义 设n 维随机变量$(X_1, X_2,\ldots, X_n)$ 的协方差

$$C_{ij} = Cov(X_i, X_j)$$

均存在, 称$C=(C_{ij})$为$(X_1, X_2,\ldots, X_n)$的协方差矩阵.

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

性质 协方差矩阵的元素满足

1) $C_{ii} = D(X_i)$;

2) $C_{ij} = C_{ji}$

对称矩阵

性质 协方差矩阵是半正定矩阵，即对任意向量$\alpha=(\alpha_1,\ldots,\alpha_n)$有

$$\alpha C\alpha^T=\sum_{i=1}^{n}\sum_{j=1}^{n}C_{ij}\alpha_i\alpha_j\geq 0.$$

证： $\alpha C\alpha^T$

$$=\sum_{i=1}^{n}\sum_{j=1}^{n}\int_{-\infty}^{\infty}\cdots\int_{\infty}^{\infty}[x_i-E(X_i)][x_j-E(X_j)]\alpha_i\alpha_j f(x_1,\ldots,x_n)dx_1\ldots dx_n$$

$$=\int_{-\infty}^{\infty}\cdots\int_{\infty}^{\infty}\sum_{i=1}^{n}\sum_{j=1}^{n}[x_i-E(X_i)][x_j-E(X_j)]\alpha_i\alpha_j f(x_1,\ldots,x_n)dx_1\ldots dx_n$$

$$=\int_{-\infty}^{\infty}\cdots\int_{\infty}^{\infty}\left(\sum_{i=1}^{n}\alpha_i[x_i-E(X_i)]\right)^2 f(x_1,\ldots,x_n)dx_1\ldots dx_n\geq 0$$

例 (X,Y)在以原点为圆心的单位圆内服从均匀分布

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

讨论X,Y的相关性和独立性

相关性：计算相关系数是否为0。

独立性：讨论联合分布和边缘分布的关系。

独立一定不相关，不相关不一定独立

计算可得 $Cov(X, Y)=0$.

可以验证 $D(X)>0$，$D(Y)>0$，因此$\rho_{X,Y}=0$，X,Y不相关。

另一方面，边缘概率密度函数为：

$$f_X(x)=\begin{cases}\int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}}\frac{1}{\pi}dy=\frac{2\sqrt{1-x^2}}{\pi}, & 0\le x\le 1,\\ 0, & \text{其它}\end{cases}$$

$$f_Y(y)=\begin{cases}\frac{2\sqrt{1-y^2}}{\pi}, & 0\le y\le 1,\\ 0, & \text{其它}\end{cases}$$

可以看出当$x^2+y^2\le 1$，$f(x,y)\neq f_X(x)f_Y(y)$，X,Y不独立。

因此X,Y不相关，但不独立

例 设二维随机变量 (X,Y) 的联合概率密度为

$$f(x,y)=\begin{cases}6xy, & 0<x<1, 0<y<2(1-x);\\ 0, & \text{其它}.\end{cases}$$

求：(X,Y)的协方差矩阵.

分析 计算(X,Y)的协方差矩阵, 就是计算X、Y 的方差和协方差.

$$E(X)=\frac{2}{5},\quad E(Y)=\frac{4}{5},$$

$$E(X^2)=\frac{1}{5},\quad E(Y^2)=\frac{4}{5}$$

$$E(XY)=-\frac{4}{75},\quad D(X)=\frac{1}{25},\quad D(Y)=\frac{4}{25}$$

因此(X,Y) 的协方差矩阵为

$$\begin{bmatrix}\frac{1}{25} & -\frac{4}{75}\\ -\frac{4}{75} & \frac{4}{25}\end{bmatrix}$$

5、设随机变量 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 均服从标准正态分布，二者的相关系数为 0，则下列说法中正确的个数有（　　）.

（1）$\boldsymbol{X}$ 与 $\boldsymbol{Y}$ 一定独立；（2）$\boldsymbol{D}(\boldsymbol{X}+3\boldsymbol{Y})=10$；（3）$3\boldsymbol{X}\sim \boldsymbol{N}(0,3)$；（4）$\mathrm{cov}(2\boldsymbol{X},3\boldsymbol{Y})=0$.

（A）1 个；　　（B）2 个；　　（C）3 个；　　（D）4 个.

4. 某人途经一个十字路口，所经方向有 50%时间亮红灯，遇红灯需等待直至绿灯，等待时间在区间$[0,20]$(单位：秒)上服从均匀分布. 用X表示此人的等待时间，求X的分布函数，并分析X是否为离散型或连续型随机变量，说明理由.

5. 设随机变量X,Y相互独立，且$X\sim N(0,1)$，Y的概率分布为$P\{Y=0\}=P\{Y=1\}=0.5$，求 $P\{XY<z\},z\in R$.（用Φ表示结果即可）

22. 设 P{X=0}=P{X=1}=1/2，Y~U(0,1)且 X,Y 相互独立，求 X+Y 的概率分布.

全概率公式本质上是一个思想！

(1999) 设随机变量 X_i 服从分布

X_i	-1	0	1
P	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

, $i=1,2$, 且 $P\{X_1X_2=0\}=1$, 则 $P\{X_1=X_2\}=$ ______.

X_1 \ X_2	-1	0	1
-1	**0**	**1/4**	**0**
0	**1/4**	**0**	**1/4**
1	**0**	**1/4**	**0**

由题意知

$$P(X_1=1,\ X_2=-1)=0$$
$$P(X_1=-1,\ X_2=-1)=0$$
$$P(X_1=1,\ X_2=1)=0$$
$$P(X_1=-1,\ X_2=1)=0$$

利用边缘分布律

$$P(X_1=-1,\ X_2=1)=1/4$$