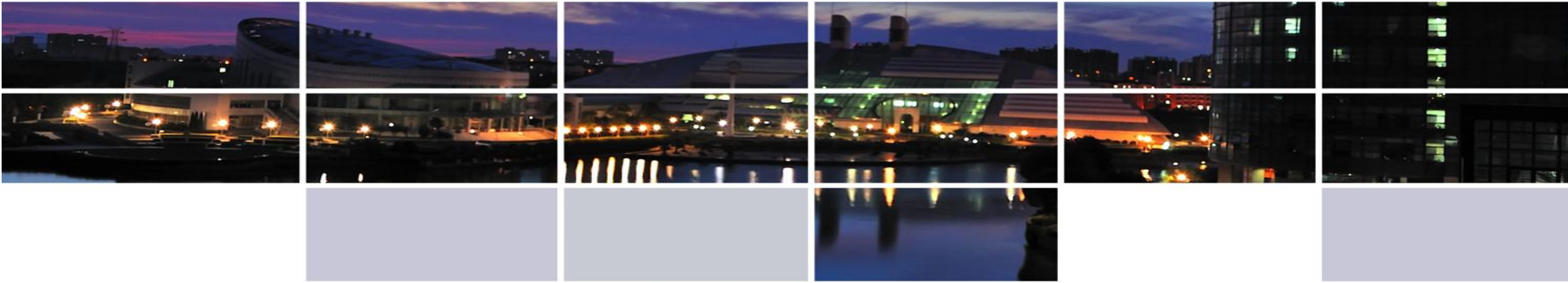
浙江大学
ZHEJIANG UNIVERSITY



第32讲 协方差与相关系数

上一讲中方差性质3:

设X，Y是两个随机变量，则有

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

特别地，若X与Y相互独立时，则有 $D(X+Y)=D(X)+D(Y)$.

即，若X与Y相互独立时，则有 $E\{[X-E(X)][Y-E(Y)]\}=0$

那么 $E\{[X-E(X)][Y-E(Y)]\}\neq 0$

X与Y不相互独立

协方差

定义： 数值 $E\{[X-E(X)][Y-E(Y)]\}$ 为随机变量X与Y的协方差，记为$Cov(X,Y)$，即

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

此时 $D(X+Y)=D(X)+D(Y)+2Cov(X,Y)$

协方差$Cov(X,Y)$反映了随机变量X与Y的线性相关性：

- 当 $Cov(X,Y)>0$ 时，称X与Y**正相关**；
- 当 $Cov(X,Y)<0$ 时，称X与Y**负相关**；
- 当 $Cov(X,Y)=0$ 时，称X与Y**不相关**.

由于 $E\{[X-E(X)][Y-E(Y)]\}$

$$=E\{XY-XE(Y)-YE(X)+E(X)E(Y)\}$$

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

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

- **协方差的计算公式:**

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

例1： 设X, Y相互独立，且均服从参数为$1/2$的$0-1$分布，记 $U=X-Y$，

$V=X+Y$，求U与V的协方差.

解： 由于 $X\sim B(1,\ 1/2)$，$Y\sim B(1,\ 1/2)$，且两者独立，则

$U \backslash V$	0	1	2
-1	0	$1/4$	0
0	$1/4$	0	$1/4$
1	0	$1/4$	0

$\{U=-1\}=\{X-Y=-1\}=\{X=0,Y=1\}\Rightarrow\{X+Y=1\}=\{V=1\}$

故而$P(U=-1,V=1)=P(X=0,Y=1)=P(X=0)P(Y=1)=1/4$;

故而$P(U=-1,V=0)=P(U=-1,V=2)=0$;

从而

$U \backslash V$	0	1	2	$P(U=i)$
-1	0	$1/4$	0	$1/4$
0	$1/4$	0	$1/4$	$1/2$
1	0	$1/4$	0	$1/4$
$P(V=j)$	$1/4$	$1/2$	$1/4$	

故 $E(UV)=(-1)\times1\times1/4+0\times0\times1/4+0\times2\times1/4+1\times1\times1/4=0;$

$$E(U)=(-1)\times1/4+0\times1/2+1\times1/4=0;$$

因此 $Cov(U,V)=E(UV)-E(U)E(V)=0.$

注意到 $Cov(X,Y)=E\{[X-E(X)][Y-E(Y)]\}$

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

■ 协方差的性质：

1. $Cov(X,Y)=Cov(Y,X)$;
2. $Cov(X,X)=D(X)$;
3. $Cov(aX,bY)=ab\cdot Cov(X,Y)$，其中$a,b$为两个实数;
4. $Cov(X_1+X_2,Y)=Cov(X_1,Y)+Cov(X_2,Y)$.

思考:

(1) $Cov(3X+2Y,X)=?$

$$=Cov(3X,X)+Cov(2Y,X)$$

$$=3Cov(X,X)+2Cov(Y,X)=3D(X)+2Cov(X,Y).$$

(2) $D(3X-2Y)=D(3X+(-2Y))$

$$=D(3X)+D(-2Y)+2Cov(3X,-2Y)$$

$$=9D(X)+4D(Y)-12Cov(X,Y).$$

或者利用$Cov(3X-2Y,3X-2Y)$来做.

例2： 对于例1：设X, Y相互独立，且均服从参数为$1/2$的$0-1$分布，记 $U=X-Y$, $V=X+Y$, 利用协方差性质求U与V的协方差.

解： 由于

$$
\begin{aligned}
Cov(U,V) &= Cov(X-Y, X+Y) \\
&= Cov(X,X)+Cov(X,Y)+Cov(-Y,X)+Cov(-Y,Y) \\
&= D(X)+Cov(X,Y)-Cov(Y,X)-D(Y) \\
&= D(X)-D(Y),
\end{aligned}
$$

而X与Y同分布，因此它们的方差相等，故

$$Cov(U,V)=0.$$

协方差是有量纲的数字特征，为了消除其量纲的影响，引入一个新概念：

定义： 数值 $$\boxed{\rho_{XY}=\frac{Cov(X,Y)}{\sqrt{D(X)D(Y)}}}$$

称为随机变量X与Y的**相关系数**，是没有量纲的.

若记标准化变量 $$X^*=\frac{X-E(X)}{\sqrt{D(X)}},Y^*=\frac{Y-E(Y)}{\sqrt{D(Y)}},$$

则 $\rho_{XY}=Cov(X^*,Y^*)$.

相关系数的性质：

1. $|\rho_{XY}| \le 1$;

2. $|\rho_{XY}| = 1 \Leftrightarrow$ 存在常数a,b，使$P(Y = a + bX) = 1$

特别的，$\rho_{XY} = 1$时，$b > 0$; $\rho_{XY} = -1$时，$b < 0$.

说明：上面的性质也可写为：

当$D(X)D(Y) \ne 0$时,有 $(Cov(X,Y))^2 \le D(X)D(Y)$,

其中等号当且仅当X与Y之间有严格的线性关系, 即

存在常数a,b, 使$P(Y = a + bX) = 1$.

相关系数是一个用来表征两个随机变量之间**线性关系密切程度**的特征数,有时也称为“线性相关系数”.

当$|\rho_{XY}|$较大时，表明X与Y的线性关系程度较好;

当$|\rho_{XY}|$较小时，表明X与Y的线性关系程度较差.

特别地,

当$|\rho_{XY}|=1$ 时，表明X与Y之间以概率1存在线性关系;

当$\rho_{XY}=0$ 时，表明X与Y之间没有线性关系，称两个变量**不相关**.

例3： 抛一枚均匀的硬币十次，若令X，Y分别表示出现正面和反面的次数，求X和Y的相关系数ρ_{XY}.

解： 由题意知，$Y=10-X$,

那么X和Y与相关系数ρ_{XY}为

$$\rho_{XY}=\frac{Cov(X,Y)}{\sqrt{D(X)D(Y)}}=\frac{Cov(X,10-X)}{\sqrt{D(X)D(10-X)}}=\frac{-D(X)}{\sqrt{D(X)D(X)}}=-1.$$

$$\because Cov(X,10-X)=Cov(X,10)+Cov(X,-X)=E(X\cdot 10)-E(X)\cdot 10-D(X)$$

$$D(10-X)=D(X)$$