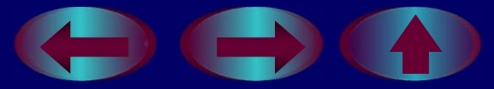
第三节　协方差及相关系数

- 协方差
- 相关系数
- 课堂练习
- 小结　布置作业

前面我们介绍了随机变量的数学期望和方差，对于二维随机变量（X，Y），我们除了讨论X与Y的数学期望和方差以外，还要讨论描述X和Y之间关系的数字特征，这就是本讲要讨论的

协方差和相关系数



一、协方差

1.定义　量$E\{[X-E(X)][Y-E(Y)]\}$称为随机变量X和Y的协方差,记为$Cov(X,Y)$，即

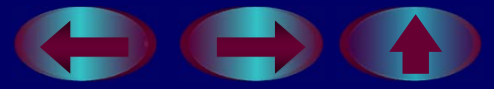
$$Cov(X,Y)=E\{[X-E(X)][Y-E(Y)]\}$$

2.简单性质

(1) $Cov(X,Y)=Cov(Y,X)$

(2) $Cov(aX,bY)=abCov(X,Y)$　a,b 是常数

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



3. 计算协方差的一个简单公式

由协方差的定义及期望的性质，可得

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

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

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

即 $$Cov(X,Y)=E(XY)-E(X)E(Y)$$

可见，若X与Y独立，$Cov(X,Y)=0$.

特别地 $Cov(X,X)=E(X^2)-E(X)^2=D(X)$

4. 随机变量和的方差与协方差的关系

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

- 当Cov(X,Y)>0时，称X与Y正相关，即X与Y同时增加或同时减少；
- 当Cov(X,Y)<0时，称X与Y负相关，即X增加Y减少，或X减少Y增加；
- 当Cov(X,Y)=0时，称X与Y不相关（/线性无关/线性不相关）

协方差的大小在一定程度上反映了X和Y相互间的关系，但它还受X与Y本身度量单位的影响. 例如：

$$Cov(kX, kY)=k^2Cov(X,Y)$$

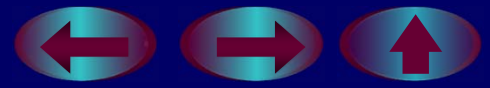
为了克服这一缺点，对随机变量X和Y进行标准化，

$$X^*=\frac{X-E(X)}{\sqrt{D(X)}}, Y^*=\frac{Y-E(Y)}{\sqrt{D(Y)}}$$

$$Cov(X^*,Y^*)=E(X^*Y^*)-E(X^*)E(Y^*)$$

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

这就引入了相关系数 .



二、相关系数

定义： 设$D(X)>0, D(Y)>0,$ 称

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

为随机变量 X 和 Y 的相关系数 .

在不致引起混淆时， 记 ρ_{XY} 为 ρ .

- 衡量线性独立的无量纲数

相关系数的性质：

1. $|\rho| \le 1$

证：　　考虑以X的线性函数$a+bX$来近似表示Y，

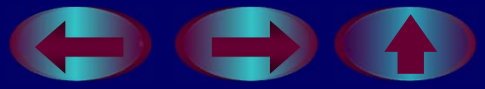
以均方误差

$$e = E\{[Y-(a+bX)]^2\}$$

来衡量以 $a+bX$ 近似表示 Y 的好坏程度：

e 值越小表示 $a+bX$ 与 Y 的近似程度越好.

用微积分中求极值的方法，
求出使e 达到最小时的 a,b



$$e = E\{[Y-(a+bX)]^2\}$$

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

$$\begin{cases} \dfrac{\partial e}{\partial a} = 2a + 2bE(X) - 2E(Y) = 0 \\ \dfrac{\partial e}{\partial b} = 2bE(X^2) - 2E(XY) + 2aE(X) = 0 \end{cases}$$

解得

$$\begin{cases} b_0 = \dfrac{Cov(X,Y)}{D(X)} \\ a_0 = E(Y) - b_0E(X) \end{cases}$$

这样求出的最佳逼近为

$L(X)=a_0+b_0X$

这样求出的最佳逼近为$L(X)=a_0+b_0X$ $\begin{cases} b_0=\dfrac{Cov(X,Y)}{D(X)} \\ a_0=E(Y)-b_0E(X) \end{cases}$

这一逼近的剩余是

$$E[(Y-L(X))^2]=D[Y-a_0-b_0X]+[E(Y-a_0-b_0X)]^2$$

$\because E(Y-a_0-b_0X)=E(Y)-a_0-b_0E(X)=0$

$\therefore E[(Y-L(X))^2]=D[Y-a_0-b_0X]$

$=D[Y-b_0X]+D[a_0]-2E[(a_0-E(a_0))(Y-b_0X-E(Y-b_0X))]$

$=D[Y-b_0X]$

$=D[Y]+b_0^2D[X]-2Cov[Y,b_0X]=D[Y]+b_0^2D[X]-2b_0Cov[X,Y]$

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

$$=D(Y)\left[1-\frac{[Cov(X,Y)]^2}{D(X)D(Y)}\right]=D(Y)[1-\rho^2]$$

由于方差$D(Y)$是正的,故必有

$1-\rho^2\geq 0$,所以$|\rho|\leq 1$。

2. X和Y独立时，$\rho=0$，但其逆不真.

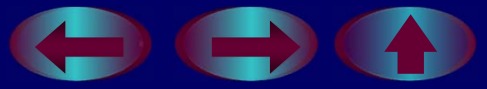
由于当X和Y独立时，$Cov(X,Y)=0$.

故 $$\rho=\frac{Cov(X,Y)}{\sqrt{D(X)D(Y)}}=0$$

但由 $\rho=0$ 并不一定能推出X和Y独立.

- 当$|\rho|=0$时，X和Y无线性关系，
  即 X和Y不相关，但不排除X和Y存在其他联系

请看下例.



例1 设X服从(-1/2, 1/2)内的均匀分布，而$Y=\cos X$,

不难求得 $Cov(X,Y)=0$，

事实上，X的密度函数

$$f(x)=\begin{cases}1 & -\frac{1}{2}<x<\frac{1}{2}\\ 0 & \text{其它}\end{cases}$$ 可得$E(X)=0$

$$E(XY)=E(X\cos X)=\int_{-\frac{1}{2}}^{\frac{1}{2}} x\cos x f(x)dx$$

$$=\left[x\sin x\right]_{-\frac{1}{2}}^{\frac{1}{2}}-\int_{-\frac{1}{2}}^{\frac{1}{2}} sixdx=0$$

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

因而 $\rho=0$， 即X和Y不相关 .

但Y与X有严格的函数关系，即X和Y不独立 .

3. $|\rho| = 1 \Leftrightarrow$ 存在常数 $a,b(b\neq 0)$，

使 $P\{Y= a + b X\}=1$,

即 X 和 Y 以概率 1 线性相关.

- $|\rho|$ 越接近于1，则X和Y的线性关系越显著

相关系数刻划了X和Y间“线性相关”的程度.

考虑以X的线性函数$a+bX$来近似表示Y,

这样求出的最佳逼近为$L(X)=a_0+b_0X$

这一逼近的剩余是

$$E[(Y-L(X))^2]= D(Y)(1- \rho^2)$$

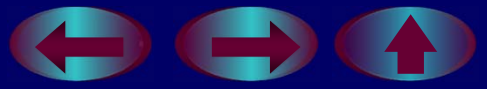
可见，若 $\rho = \pm 1$，Y与X有严格线性关系；

若 $\rho =0$， Y 与 X 无线性关系；

若$0<|\rho|<1$,

$|\rho|$ 的值越接近于1, Y 与 X 的线性相关程度越高；

$|\rho|$ 的值越接近于0, Y与X的线性相关程度越弱.



前面，我们已经看到：

若 X 与 Y 独立，则X与Y不相关，

但由X与Y不相关，不一定能推出X与Y独立.

但对下述情形，独立与不相关等价

若(X,Y)服从二维正态分布，则
X与Y独立 $\Longleftrightarrow$ X与Y不相关

根据P109 例2可得，二维正态分布R.V.(X,Y)

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

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

在第三章§4中已经讲过，若(X,Y)服从二维正态分布，那么 X 和 Y 相互独立的充要条件为$\rho=0$. 现在知道 $\rho=\rho_{XY}$，故知对于二维正态随机变量(X,Y)来说，X 和 Y 不相关与 X 和 Y 相互独立是等价的.

四、小结

这一节我们介绍了协方差、相关系数、相关系数是刻划两个变量间线性相关程度的一个重要的数字特征.

注意独立与不相关并不是等价的.

当(X,Y) 服从二维正态分布时，有

X 与 Y 独立 $\Leftrightarrow$ X 与 Y 不相关