概率论与数理统计

协方差及相关系数

主讲人：曾华琳

信息科学与技术学院

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) $Cov(aX,bY)=ab\,Cov(X,Y)$　a,b 是常数

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

一、协方差

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，$Cov(X,Y)=0$.

一、协方差

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

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

一、协方差

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，对协方差进行标准化，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 。

二、相关系数

相关系数的性质： $1.|\rho|\le 1$

由于方差$D(Y)$是正的，故必有$1-p^2\ge 0$,所以$|p|\le 1$。

证：由方差的性质和协方差的定义知，对任意实数 b，有

$$0\le D(Y-bX)= b^2D(X)+D(Y)-2b\ Cov(X,Y)$$

令 $b=\dfrac{Cov(X,Y)}{D(X)}$ **，则上式为**

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

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

二、相关系数

相关系数的性质：

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。

二、相关系数

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

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

$$E(XY)=E(X\cos X)=\int_{-\frac{1}{2}}^{\frac{1}{2}} x\cos x f(x)dx=0$$

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

二、相关系数

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。

相关系数的性质： 3 . $|\rho| = 1$

$\Longleftrightarrow$ 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。

二、相关系数

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，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

二、相关系数

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

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

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

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

最佳逼近

$L(X)=a_0+b_0X$

二、相关系数

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。

二、相关系数

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 $\Leftrightarrow$ X 与 Y 不相关

谢 谢 大 家