第九周 协方差与相关系数

9.2 协方差

多元随机变量更本质的方面是各分量之间的相互关系、相互作用，这方面最重要的数字特征是协方差与相关系数。

定义：设(X,Y)是二元随机变量，$E[(X-E(X))(Y-E(Y))]$ 称为 X,Y 的协方差，

记为$Cov(X,Y)$。

$Cov(X,a)=0$，　　　　$Cov(X,Y)=Cov(Y,X)$，

$Cov(c_1X+a,c_2Y+b)=c_1c_2\cdot Cov(X,Y)$，　　$Cov(X+Y,Z)=Cov(X,Z)+Cov(Y,Z)$，

$Cov(X,Y)==E(XY)-E(X)E(Y)$，　　若 X,Y 相互独立，$Cov(X,Y)=0$

****

例 9.2.1 从$1,2,3,4$中等可能地取 1 个数记为 X，再从$1,2,\cdots,X$中等可能地取 1 个数记为 Y。求$Cov(X,Y)$。

解：(X,Y)的联合与边缘分布列为

$X\backslash Y$	$Y=1$	$Y=2$	$Y=3$	$Y=4$	$P(X=i)$
$X=1$	$1/4$	0	0	0	$1/4$
$X=2$	$1/8$	$1/8$	0	0	$1/4$
$X=3$	$1/12$	$1/12$	$1/12$	0	$1/4$
$X=4$	$1/16$	$1/16$	$1/16$	$1/16$	$1/4$
$P(Y=k)$	$25/48$	$13/48$	$7/48$	$1/16$	1

$$E(XY)==\frac{1}{4}\cdot 1+\frac{1}{8}(2\cdot 1+2\cdot 2)+\frac{1}{12}(3\cdot 1+3\cdot 2+3\cdot 3)+\frac{1}{16}(4\cdot 1+4\cdot 2+4\cdot 3+4\cdot 4)=5$$

$$Cov(X,Y)=E(XY)-E(X)E(Y)=5-\frac{5}{2}\cdot\frac{7}{4}=\frac{5}{8}。$$

****

例 9.2.2 设随机变量 $X\sim Ge(p)$（$0<p<1$），$Y=\begin{cases}1, & X=1\\ 0, & X>1\end{cases}$，计算$Cov(X,Y)$。

解： $E(X)=\frac{1}{p}$， $E(Y)=1\cdot P(Y=1)+0\cdot P(Y=0)=1\cdot P(X=1)=p$

$$E(XY)=E(E(XY|Y))=P(Y=1)\cdot E(XY|Y=1)+P(Y=0)\cdot E(XY|Y=0)$$

$$=P(Y=1)\cdot E(XY|Y=1)=P(X=1)\cdot E(X|X=1)=p\text{ ,}$$

$$Cov(X,Y)=E(XY)-E(X)E(Y)=p-\frac{1}{p}\cdot p=p-1\text{。}$$

补充：其中 X,Y 乘积的期望也可以直接观察得到，只有 X=1，Y=1 时，X,Y 的联合概率非零， $E(XY)=1\cdot 1\cdot P(X=1,Y=1)=P(X=1)=p$ 。

****

随机变量和的方差公式

$$Var(X+Y)=E\left[(X+Y)^2\right]-E(X+Y)^2$$

$$=E(X^2)+2\cdot E(XY)+E(Y^2)-\left[E(X)^2+2\cdot E(X)E(Y)+E(Y)^2\right]$$

$$=E(X^2)-E(X)^2+E(Y^2)-E(Y)^2+2\cdot\left[E(XY)-E(X)E(Y)\right]$$

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

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

****

随机变量 (X,Y) 的协方差 $Cov(X,Y)=E\left[(X-E(X))(Y-E(Y))\right]$

若 (X,Y) 的取值，当 $X>E(X)$ 时， $Y>E(Y)$ 的可能性较大；当 $X<E(X)$ 时， $Y<E(Y)$ 的可能性较大，则 $Cov(X,Y)>0$ ；

若 (X,Y) 的取值，当 $X>E(X)$ 时， $Y<E(Y)$ 的可能性较大；当 $X<E(X)$ 时， $Y>E(Y)$ 的可能性较大，则 $Cov(X,Y)<0$ ；

若 (X,Y) 的取值，当 $X>E(X)$ 时，$Y>E(Y)$ 和 $Y<E(Y)$ 的可能性差不多；当 $X<E(X)$ 时，$Y>E(Y)$ 和 $Y<E(Y)$ 的可能性差不多，则 $Cov(X,Y)$ 会比较接近于 0。

例如本节的例 1，X 越大则 Y 取到比较大的值的可能性也越大，它们是正相关的关系，计算得协方差也为正数，等于 8 分之 5；例 2，当 X 等于 1 时 Y 等于 0,当 X 大于 1 时，Y 的取值为 0,X,Y 的变化趋势相反，它们是负相关的关系，协方差等于 p-1，是负数。但是，随机变量 X，Y 的协方差的大小还不足以充分地反映 X，Y 之间的相关程度，因为若将 X，Y 同时放大 10 倍，变为 10X 和 10Y，它们的协方差增大了 100 倍，但是它们实际的相关程度并没有发生变化，所以我们还需要引入更细致、更合理的刻画随机变量之间相关性的指标。就是相关系数。

****