3.3 协方差与相关系数

现在我们来考虑多维随机向量的数字特征，以二维的情况为例. 设(X, Y)为二维随机向量，X, Y本身都是一维随机变量，可以定义其均值方差，在本节中我们记

$$E(X) = m_1, E(Y) = m_2, \mathrm{Var}(X) = \sigma_1^2, \mathrm{Var}(Y) = \sigma_2^2$$

这些都在上两节中已讨论过了，没有什么新东西. 在多维随机向量

中，最有兴趣的数字特征是反映分量之间的关系的那种量，其中最重要的，是本节要讨论的协方差和相关系数.

定义 3.1 称 $E[(X-m_1)(Y-m_2)]$ 为 X,Y 的协方差，并记为 $\mathrm{Cov}(X,Y)$*.

"协"即"协同"的意思. X 的方差是 $(X-m_1)$ 与 $(X-m_1)$ 的乘积的期望，如今把一个 $X-m_1$ 换为 $Y-m_2$，其形式接近方差，又有 X,Y 二者的参与，由此得出协方差的名称. 由定义看出：$\mathrm{Cov}(X,Y)$ 与 X,Y 的次序无关：$\mathrm{Cov}(X,Y)=\mathrm{Cov}(Y,X)$. 直接由定义得到协方差的一些简单性质. 例如，若 c_1,c_2,c_3,c_4 都是常数，则

$$\mathrm{Cov}(c_1X+c_2,c_3Y+c_4)=c_1c_3\mathrm{Cov}(X,Y) \tag{3.1}$$

又易知

$$\mathrm{Cov}(X,Y)=E(XY)-m_1m_2 \tag{3.2}$$

这些简单性质的证明都留给读者.

下面的定理包含了协方差的重要性质.

定理 3.1 1° 若 X,Y 独立，则 $\mathrm{Cov}(X,Y)=0$.

2° $[\mathrm{Cov}(X,Y)]^2\leqslant\sigma_1^2\sigma_2^2$. 等号当且仅当 X,Y 之间有严格线性关系（即存在常数 a,b 使 $Y=a+bX$）时成立.

证 1°的证明由定理 1.2 直接得出，因据此定理，当 X,Y 独立时有 $E(XY)=m_1m_2$. 为证明 2°，需要两点预备事实：

a. 若 a,b,c 为常数，$a>0$，而二次三项式 $at^2+2bt+c$ 对 t 的任何实值都非负，则必有 $ac\geqslant b^2$.

b. 若随机变量 Z 只能够取非负值，而 $E(Z)=0$，则 $Z=0$.

为了不打断此处的讨论，我们将这两点事实的证明放到后面，现考虑

$$E[t(X-m_1)+(Y-m_2)]^2=\sigma_1^2t^2+2\mathrm{Cov}(X,Y)t+\sigma_2^2 \tag{3.3}$$

---

* Cov 是协方差 Covariance 的缩写.

由于此式左边是一个非负随机变量的均值，故它对任何 t 非负. 按预备事实 a，有

$$\sigma_1^2\sigma_2^2 \geqslant [\mathrm{Cov}(X,Y)]^2 \tag{3.4}$$

进一步，如果(3.4)成立等号，则(3.3)右边等于$(\sigma_1 t \pm \sigma_2)^2$. ±号视 $\mathrm{Cov}(X,Y)>0$ 或<0 而定，为确定计，暂设 $\mathrm{Cov}(X,Y)>0$，则(3.3)右边为$(\sigma_1 t+\sigma_2)^2$. 此式在 $t=t_0=-\sigma_2/\sigma_1$ 时为 0. 以 $t=t_0$ 代入(3.3)，有

$$E[t_0(X-m_1)+(Y-m_2)]^2=0$$

再按预备事实 b，即知 $t_0(X-m_1)+(Y-m_2)=0$，因而 X,Y 之间有严格线性关系.

反之，若 X,Y 之间有严格线性关系 $Y=aX+b$，则 $\sigma_2^2=\mathrm{Var}(Y)=\mathrm{Var}(aX+b)=\mathrm{Var}(aX)=a^2\mathrm{Var}(X)=a^2\sigma_1^2$. 又因 $m_2=E(Y)=aE(X)+b=am_1+b$，有 $Y-m_2=(aX+b)-(am_1+b)=a(X-m_1)$. 于是

$$\begin{aligned}\mathrm{Cov}(X,Y) &= E[(X-m_1)a(X-m_1)] = a[E(X-m_1)^2]\\ &= a\sigma_1^2\end{aligned}$$

因此

$$[\mathrm{Cov}(X,Y)]^2 = a^2\sigma_1^4 = \sigma_1^2(a^2\sigma_1^2) = \sigma_1^2\sigma_2^2$$

即(3.4)成立等号，这就证明了 2°全部结论.

现证明用到的两个预备事实. 对 a，注意到若 $ac<b^2$，则 $at^2+2bt+c=0$ 有两个不同的实根 $t_1<t_2$，而 $at^2+2bt+c=a(t-t_1)(t-t_2)$. 取 t_0 使 $t_1<t_0<t_2$，则将有 $at_0^2+2bt_0+c=a(t_0-t_1)(t_0-t_2)<0$，与 $at^2+2bt+c$ 对任何t 非负矛盾，这证明了 a. b 的证明很简单：若 $Z\neq 0$，则因 Z 只能取非负值，它必以一定的大于 0 的概率取大于 0 的值，这将导致 $E(Z)>0$，与 $E(Z)=0$ 的假定不合.

定理 3.1 给“相关系数”的定义打下了基础.

定义 3.2 称 $\mathrm{Cov}(X,Y)/(\sigma_1\sigma_2)$ 为 X,Y 的相关系数,并记为* $\mathrm{Corr}(X,Y)$.

形式上可以把相关系数视为"标准尺度下的协方差".协方差作为 $(X-m_1)(Y-m_2)$ 的均值,依赖于 X,Y 的度量单位,选择适当单位使 X,Y 的方差都为 1,则协方差就是相关系数.这样就能更好地反映 X,Y 之间的关系,不受所用单位的影响.

由定理 3.1 立即得到:

定理 3.2 1° 若 X,Y 独立,则 $\mathrm{Corr}(X,Y)=0$. 2° $-1\leqslant\mathrm{Corr}(X,Y)\leqslant 1$,或 $|\mathrm{Corr}(X,Y)|\leqslant 1$,等号当且仅当 X 和 Y 有严格线性关系时达到.

对这个定理我们要加以几条重要的解释.

1. 当 $\mathrm{Corr}(X,Y)=0$(或 $\mathrm{Cov}(X,Y)=0$,一样),称 X,Y"不相关".本定理的 1°说明由 X,Y 独立推出它们不相关.但反过来一般不成立:由 $\mathrm{Corr}(X,Y)=0$ 不一定有 X,Y 独立.下面是一个简单的例子.

例 3.1 设 (X,Y) 服从单位圆内的均匀分布,即其密度函数为

$$f(x,y)=\begin{cases}\pi^{-1}, & \text{当 } x^2+y^2<1\\ 0, & \text{当 } x^2+y^2\geqslant 1\end{cases}$$

用第二章公式(2.9),(2.10),容易得出 X,Y 有同样的边缘密度函数 g:

$$g(x)=\begin{cases}2\pi^{-2}\sqrt{1-x^2}, & \text{当 } |x|<1\\ 0, & \text{当 } |x|\geqslant 1\end{cases}$$

这个函数关于 0 对称,因此其均值为 0.故 $E(X)=E(Y)=0$,而

$$\mathrm{Cov}(X,Y)=E(XY)=\frac{1}{\pi}\iint_{x^2+y^2<1}xy\,\mathrm{d}x\,\mathrm{d}y=0$$

故 $\mathrm{Corr}(X,Y)=0$.但 X,Y 不独立,因为其联合密度 $f(x,y)$ 不等于其边缘密度之积 $g(x)g(y)$.

---

* Corr 是相关 Correlation 的缩写.

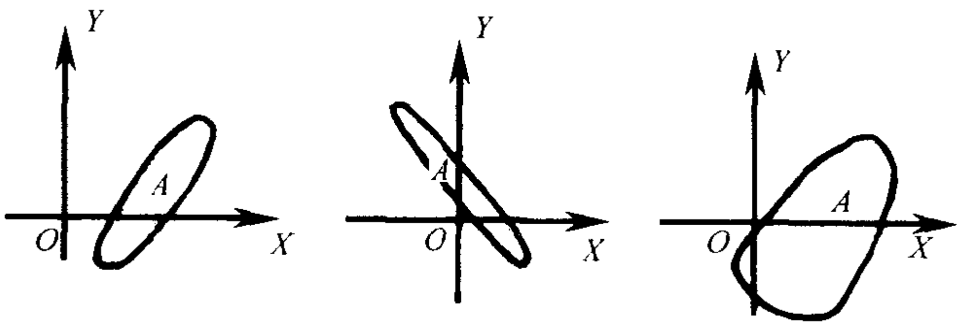
2. 相关系数也常称为“线性相关系数”. 这是因为,实际上相关系数并不是刻画了 X,Y 之间“一般”关系的程度,而只是“线性”关系的程度. 这种说法的根据之一就在于,当且仅当 X,Y 有严格的线性关系时,才有 $|\mathrm{Corr}(X,Y)|$ 达到最大值 1. 可以容易举出例子说明:即使 X 与 Y 有某种严格的函数关系但非线性关系, $|\mathrm{Corr}(X,Y)|$ 不仅不必为 1,还可以为 0.

例 3.2 设 $X \sim R(-1/2,1/2)$,即区间 $[-1/2,1/2]$ 内的均匀分布,而 $Y=\cos X$, Y 与 X 有严格函数关系. 但因 $E(X)=0$,由(3.2)有

$$\mathrm{Cov}(X,Y) = E(XY) = E(X\cos X) = \int_{-1/2}^{1/2} x\cos x \mathrm{d}x = 0$$

故 $\mathrm{Corr}(X,Y)=0$. 你看, X,Y 说是“不相关”,它们之间却有着严格的关系 $Y=\cos X$. 足见这样的相关只能指线性而言,一超出这个范围,这个概念就失去了意义.

3. 如果 $0<|\mathrm{Corr}(X,Y)|<1$,则解释为: X,Y 之间有“一定程度的”线性关系而非严格的线性关系. 何谓“一定程度”的线性关系? 我们可以用图 3.6 所示的情况来说明. 在这三个图中,我们都假定 (X,Y) 服从所画出的区域 A 内的均匀分布(即其联合密度 $f(x,y)$ 在 A 内为 $|A|^{-1}$,在 A 外为 0, $|A|$ 为区域 A 的面积). 在这三个图中, X,Y 都无严格的线性关系,因为由 X 之值并不能决定 Y 之值. 可是由这几个图我们都能“感觉”出, X,Y 之间存在



(a) (b) (c)

图 3.6

着一种线性的“趋势”.这种趋势,在(a)已较显著且是正向的(X 增加时 Y 倾向于增加),这相应于 $\mathrm{Corr}(X,Y)$ 比较显著地大于 0. 在(b),这种线性趋势比(a)更明显,程度更大,反映 $|\mathrm{Corr}(X,Y)|$ 比(a)的情况更大,但为负向的.至于(c),则多少有一点儿线性倾向,但已甚微弱:$\mathrm{Corr}(X,Y)$ 虽仍大于 0 但已接近 0.

4. 上面谈到的“线性相关”的意义,还可以从最小二乘法的角度去解释:设有两个随机变量 X,Y,现在想用 X 的某一线性函数 $a+bX$ 来逼近 Y,问要选择怎样的常数 a,b,才能使逼近的程度最高? 这逼近程度,我们就用“最小二乘”的观点来衡量,即要使 $E[(Y-a-bX)^2]$ 达到最小.

仍以 m_1,m_2 记 $E(X),E(Y)$,σ_1^2 和 σ_2^2 记 $\mathrm{Var}(X),\mathrm{Var}(Y)$. 引进常数

$$c=a-(m_2-bm_1)$$

则

$$\begin{aligned}E[(Y-a-bX)^2]&=E[(Y-m_2)-b(X-m_1)-c]^2\\&=\sigma_2^2+b^2\sigma_1^2-2b\mathrm{Cov}(X,Y)+c^2,\end{aligned}$$

为使此式达到最小,必须取 $c=0,b=\mathrm{Cov}(X,Y)/\sigma_1^2=\sigma_1\sigma_2\mathrm{Corr}(X,Y)/\sigma_1^2=\sigma_1^{-1}\sigma_2\mathrm{Corr}(X,Y)$. 这样求出最佳线性逼近为(记 $\rho=\mathrm{Corr}(X,Y)$)

$$L(X)=m_2-\sigma_1^{-1}\sigma_2\rho m_1+\sigma_1^{-1}\sigma_2\rho X \tag{3.5}$$

这一逼近的剩余是

$$\begin{aligned}E[(Y-L(X))^2]&=\sigma_2^2+b^2\sigma_1^2-2b\mathrm{Cov}(X,Y)\\&=\sigma_2^2+(\sigma_1^{-1}\sigma_2\rho)^2\sigma_1^2-2(\sigma_1^{-1}\sigma_2\rho)\sigma_1\sigma_2\rho\\&=\sigma_2^2(1-\rho^2)\end{aligned}\tag{3.6}$$

如果 $\rho=\pm1$,则 $E[(Y-L(X))]^2=0$ 而 $Y=L(X)$.这时 Y 与 X 有严格线性关系,已于前述.若 $0<|\rho|<1$,则 $|\rho|$ 愈接近 1,剩余愈小,说明 $L(X)$ 与 Y 的接近程度愈大,即 X,Y 之间线性关系的“程度”愈大.反之,$|\rho|$ 愈小,则二者的线性关系程度愈小,当 $\rho=0$ 时,剩余为 σ_2^2.这时 X 的线性作用已毫不存在.因为,仅取一

个与 X 无关的常数 m_2,已可把 Y 逼近到 σ_2^2 的剩余,因 $E(Y-m_2)^2=\sigma_2^2$. ρ 的符号的意义也由(3.5)得到解释:当 $\rho>0$ 时,$L(X)$中 X 的系数大于0,即 Y 的最佳逼近 $a+bX$ 随 X 增加而增加.这就是正向相关.反之,$\rho<0$ 表示负向相关.

由于相关系数只能刻画线性关系的程度,而不能刻画一般的函数相依关系的程度,在概率论中还引进了另一些相关性指标,以补救这个缺点.但是,这些指标都未能在应用中推开.究其原因,除了这些指标在性质上比较复杂外,还有一个重要原因:在统计学应用上,最重要的二维分布是二维正态分布.而对二维正态分布而言,相关系数是 X,Y 的相关性的一个完美的刻画,没有上面指出的缺点.其根据有两条:

1. 若(X,Y)为二维正态,则即使允许你用任何函数 $M(X)$ 去逼近 Y(仍以 $E[(Y-M(X))^2]$最小为准则,那你所得到的最佳逼近,仍是由(3.5)式决定的 $L(X)$.故在这个场合,只须考虑线性逼近已足,而这种逼近的程度完全由相关系数决定.

2. 当(X,Y)为二维正态时,由 $\mathrm{Corr}(X,Y)=0$ 能推出 X,Y 独立.即在这一特定场合,独立与不相关是一回事.我们前已指出,这在一般情况并不成立.

第一点的证明超出本书范围.第二点则不难证明.事实上我们将验证:若(X,Y)有二维正态分布 $N(a,b,\sigma_1^2,\sigma_2^2,\rho)$,则 $\mathrm{Corr}(X,Y)=\rho$.而当 $\rho=0$ 时,按第二章(2.7)式易知,(X,Y)的联合密度可表为 X,Y 各自的密度 $f_1(x)$和 $f_2(y)$之积,因而 X,Y 是独立的.

为证明此事,注意到 $E(X)=a,E(Y)=b$.利用第二章(2.7)式的 $N(a,b,\sigma_1^2,\sigma_2^2,\rho)$的密度函数的形式.有

$$\begin{aligned}\mathrm{Cov}(X,Y) &= E[(X-a)(Y-b)]\\ &= \left(2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}\right)^{-1}\iint_{-\infty}^{\infty}(x-a)(y-b)\exp\Big[-\frac{1}{2(1-\rho^2)}\\ &\quad\cdot\left(\frac{(x-a)^2}{\sigma_1^2}-\frac{2\rho(x-a)(y-b)}{\sigma_1\sigma_2}+\frac{(y-b)^2}{\sigma_2^2}\right)\Big]\mathrm{d}x\mathrm{d}y.\end{aligned}$$

注意到

$$\frac{(x-a)^2}{\sigma_1^2}-\frac{2\rho(x-a)(y-b)}{\sigma_1\sigma_2}+\frac{(y-b)^2}{\sigma_2^2}$$

$$=\left(\frac{x-a}{\sigma_1}-\frac{\rho(y-b)}{\sigma_2}\right)^2+\left(\sqrt{1-\rho^2}\,\frac{y-b}{\sigma_2}\right)^2$$

作变数代换

$$u=\frac{1}{\sqrt{1-\rho^2}}\left(\frac{x-a}{\sigma_1}-\frac{\rho(y-b)}{\sigma_2}\right),v=\frac{y-b}{\sigma^2}$$

可将上面的重积分化为

$$\mathrm{Cov}(X,Y)=(2\pi)^{-1}\iint_{-\infty}^{\infty}\left[\sqrt{1-\rho^2}\sigma_1u+\sigma_1\rho v\right]$$

$$\cdot\sigma_2v\exp\left[-\frac{u^2+v^2}{2}\right]\mathrm{d}u\,\mathrm{d}v$$

因为

$$\iint_{-\infty}^{\infty}uv\exp\left(-\frac{u^2+v^2}{2}\right)\mathrm{d}u\,\mathrm{d}v=\int_{-\infty}^{\infty}u\mathrm{e}^{-u^2/2}\mathrm{d}u\int_{-\infty}^{\infty}u\mathrm{e}^{-v^2/2}\mathrm{d}v=0$$

$$\iint_{-\infty}^{\infty}v^2\exp\left(-\frac{u^2+v^2}{2}\right)\mathrm{d}u\,\mathrm{d}v=\int_{-\infty}^{\infty}\mathrm{e}^{-u^2/2}\mathrm{d}u\int_{-\infty}^{\infty}v^2\mathrm{e}^{-v^2/2}\mathrm{d}v=2\pi$$

得到 $\mathrm{Cov}(X,Y)=\rho\sigma_1\sigma_2$. 又 $\mathrm{Var}(X)=\sigma_1^2$, $\mathrm{Var}(Y)=\sigma_2^2$. 于是

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