# 第九周 协方差与相关系数

## 9.1. 随机变量函数的期望

### 随机变量函数的期望

$n$ 维随机变量 $X=(X_1,X_2,\cdots,X_n)$，若 $Z=g(X_1,X_2,\cdots,X_n)$，则

离散情形：$E(Z)=\sum_{i_1}\cdots\sum_{i_n}g(x_1,x_2,\cdots,x_n)P(X_1=x_1,\cdots,X_n=x_n)$

连续情形：$E(Z)=\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}g(x_1,x_2,\cdots,x_n)f(x_1,x_2,\cdots,x_n)dx_1\cdots dx_n$

$Var(Z)=E[(Z-E(Z))^2]=E(Z^2)-E(Z)^2$

********************************************************

例 9.1.1 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 $E(X+2Y)$ 与 $Var(X+2Y)$。

解：$(X,Y)$的联合分布列为

| $X\backslash Y$ | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| 1 | 1/4 | 0 | 0 | 0 |
| 2 | 1/8 | 1/8 | 0 | 0 |
| 3 | 1/12 | 1/12 | 1/12 | 0 |
| 4 | 1/16 | 1/16 | 1/16 | 1/16 |

$$E(X+2Y)=\sum_{i=1}^{4}\sum_{k=1}^{4}(x_i+2y_k)P(X=x_i,Y=y_k)$$

$$=\frac{1}{4}\cdot(1+2\cdot1)+\frac{1}{8}(2+2\cdot1+2+2\cdot2)+\frac{1}{12}\cdot\sum_{k=1}^{3}(3+2\cdot k)+\frac{1}{16}\cdot\sum_{k=1}^{4}(4+2\cdot k)=6$$

$$E[(X+2Y)^2]=\sum_{i=1}^{4}\sum_{k=1}^{4}(x_i+2y_k)^2P(X=x_i,Y=y_k)=\sum_{i=1}^{4}\frac{1}{4i}\sum_{k=1}^{i}(x_i+2y_k)^2=\frac{259}{6}$$

$$Var(X+2Y)=E[(X+2Y)^2]-E(X+2Y)^2=\frac{43}{6}\text{。}$$

$(X,Y)$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 |

$E(X)=\frac{1}{4}(1+2+3+4)=\frac{5}{2}$， $E(Y)=\frac{25}{48}+2\cdot\frac{13}{48}+3\cdot\frac{7}{48}+4\cdot\frac{3}{48}=\frac{7}{4}$

$E(X+2Y)=E(X)+E(2Y)=\frac{5}{2}+2\cdot\frac{7}{4}=6$， $Var(X+2Y)\neq Var(X)+Var(2Y)$

**************************************************************

随机变量和的期望等于期望的求和

$$E(X_1+X_2)=\sum_i\sum_j(i+j)\cdot P(X_1=i,X_2=j)$$

$$=\sum_i\sum_j i\cdot P(X_1=i,X_2=j)+\sum_i\sum_j j\cdot P(X_1=i,X_2=j)$$

$$=\sum_i i\cdot\sum_j P(X_1=i,X_2=j)+\sum_i j\cdot\sum_j P(X_1=i,X_2=j)$$

$$=\sum_i i\cdot P(X_1=i)+\sum_i j\cdot P(X_2=j)=E(X_1)+E(X_2)$$

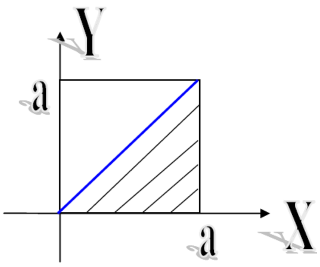
$$E(X_1+X_2+\cdots+X_n)=E(X_1)+E(X_2)\cdots+E(X_n)$$

若 $X_1,X_2,\cdots,X_n$ 相互独立， $E(X_1X_2\cdots X_n)=E(X_1)E(X_2)\cdots E(X_n)$，

若 $X_1,X_2,\cdots,X_n$ 相互独立， $Var(X_1+X_2+\cdots+X_n)=Var(X_1)+Var(X_2)\cdots+Var(X_n)$

**************************************************************

例 9.1.2 在长度为 $a$ 的线段上随机任取两点，求两点距离的期望与方差。




解： 设两个点到线段一个固定端点的距离分别为随机变量 $X$ 与 $Y$，则 $(X,Y)$ 的密度函数为

$$f_{XY}(x,y)=\begin{cases}1/a^2, & 0<x,y<a\\ 0, & 其他\end{cases}$$

$$E(|X-Y|)=\int_0^a\int_0^a|x-y|\cdot\frac{1}{a^2}dxdy=\int_0^a dx\int_0^x 2\cdot(x-y)\cdot\frac{1}{a^2}dy=\frac{a}{3}$$

$$E(|X-Y|^2)=\int_0^a\int_0^a(x-y)^2\cdot\frac{1}{a^2}dxdy=\frac{a^2}{6}，\quad Var(|X-Y|)=\frac{a^2}{18}。$$

**************************************************************

## 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 |

$$E(XY)==\frac{1}{4}\cdot1+\frac{1}{8}(2\cdot1+2\cdot2)+\frac{1}{12}(3\cdot1+3\cdot2+3\cdot3)+\frac{1}{16}(4\cdot1+4\cdot2+4\cdot3+4\cdot4)=5$$

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

************************************************************

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

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

$$=P(Y=1)\cdot E\left(XY|Y=1\right)=P(X=1)\cdot E\left(X|X=1\right)=p,$$

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

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

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

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

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

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

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

**********************************************************

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。

**********************************************************

## 9.3 相关系数

相关系数 二元随机变量 $(X,Y)$，$Var(X)\cdot Var(Y)>0$，则 $X,Y$ 的相关系数定义为

$$Corr(X,Y)=\frac{Cov(X,Y)}{\sqrt{Var(X)}\cdot\sqrt{Var(Y)}}=\frac{Cov(X,Y)}{\sigma_X\cdot\sigma_Y},$$

将随机变量做方差为 1 的标准化：$Var\left(\frac{X}{\sigma_X}\right)=1$，$Var\left(\frac{Y}{\sigma_Y}\right)=1$，

相关系数是将随机变量做方差为 1 的标准化后的协方差，$Corr(X,Y)=Cov\left(\frac{X}{\sigma_X},\frac{Y}{\sigma_Y}\right)$

性质： $-1\le Corr(X,Y)\le 1$。

**********************************************************

定理： $[Cov(X,Y)]^2\le Var(X)\cdot Var(Y)={\sigma_X}^2\cdot{\sigma_Y}^2$

证明：对任意参数 $t\in R$， $Var(tX+Y)=Var(X)\cdot t^2+2Cov(X,Y)\cdot t+Var(Y)\ge 0$

二次函数 $ax^2+bx+c\ge 0$，则其根的判别式 $b^2-4ac\le 0$，

所以，上述关于 $t$ 的二次函数的判别式小于等于 0，

$[2Cov(X,Y)]^2-4Var(X)\cdot Var(Y)\le 0$，即

$$[Cov(X,Y)]^2 \le Var(X)\cdot Var(Y) = \sigma_X{}^2 \cdot \sigma_Y{}^2$$

*****************************************************************

相关系数的绝对值不会大于 1

$$[Cov(X,Y)]^2 \le Var(X)\cdot Var(Y) = \sigma_X{}^2 \cdot \sigma_Y{}^2 \Rightarrow \left|\frac{Cov(X,Y)}{\sigma_X \cdot \sigma_Y}\right| \le 1$$

$$\Rightarrow \quad |Corr(X,Y)| = \left|\frac{Cov(X,Y)}{\sqrt{Var(X)}\cdot\sqrt{Var(Y)}}\right| = \left|\frac{Cov(X,Y)}{\sigma_X \cdot \sigma_Y}\right| \le 1$$

$$\Rightarrow \quad -1 \le Corr(X,Y) \le 1$$

$Corr(X,Y) > 0$ 正相关，$Corr(X,Y) < 0$ 负相关，$Corr(X,Y) = 0$ 不相关。

*****************************************************************

相关系数的性质

(1) $Corr(X,a) = 0$，　　(2) $Corr(X,Y) = Corr(Y,X)$，

(3) $Corr(c_1X+a, c_2Y+b) = \begin{cases} Corr(X,Y), & c_1c_2 > 0 \\ -Corr(X,Y), & c_1c_2 < 0 \\ 0, & c_1c_2 = 0 \end{cases}$，

$$Corr(c_1X+a, c_2Y+b) = \frac{Cov(c_1X+a, c_2Y+b)}{\sqrt{Var(c_1X+a)}\cdot\sqrt{Var(c_2Y+b)}}$$

$$= \frac{c_1\cdot c_2\cdot Cov(X,Y)}{|c_1\cdot c_2|\cdot\sqrt{Var(X)}\cdot\sqrt{Var(Y)}} = \frac{c_1\cdot c_2}{|c_1\cdot c_2|} Corr(X,Y)$$

*****************************************************************

例 9.3.1 随机变量 $X \sim U[0,1]$，若 $Y = X^2$，试求 $Corr(X,Y)$。

解： 由均匀分布的数学期望与方差的结论知 $E(X) = \frac{1}{2}$, $Var(X) = \frac{1}{12}$

且 $E(X^n) = \int_0^1 x^n dx = \frac{1}{n+1}$.

所以 $E(Y)=E(X^2)=\frac{1}{3}$， $E(Y^2)=E(X^4)=\frac{1}{5}$，

于是 $Var(Y)=E(Y^2)-(E(Y))^2=\frac{1}{5}-\frac{1}{9}=\frac{4}{45}$，

$$Corr(X,Y)=\frac{Cov(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}}=\frac{E(XY)-E(X)E(Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}}=\frac{E(X^3)-E(X)E(Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}}$$

$$=\frac{\frac{1}{4}-\frac{1}{2}\times\frac{1}{3}}{\sqrt{\frac{1}{12}\times\sqrt{\frac{4}{45}}}}=\frac{\sqrt{15}}{4}\approx 0.968。$$

************************************************************

例 9.3.2 计算二维随机变量 $(X,Y)\sim N(\mu_1,\mu_2,\sigma_1^2,\sigma_2^2,\rho)$ 的相关系数 $Corr(X,Y)$。

解： 考虑 $(X,Y)$ 的联合密度函数，

$$f(x,y)=\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\exp\left\{\frac{-1}{2(1-\rho^2)}\left[\frac{(x-\mu_1)^2}{\sigma_1^2}-2\rho\frac{(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2}+\frac{(y-\mu_2)^2}{\sigma_2^2}\right]\right\}$$

若令 $X_1=X-\mu_1$， $Y_1=Y-\mu_2$，则 $(X_1,Y_1)$ 的联合密度函数为

$$f(x_1,y_1)=\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\exp\left\{\frac{-1}{2(1-\rho^2)}\left[\frac{x_1{}^2}{\sigma_1^2}-2\rho\frac{x_1y_2}{\sigma_1\sigma_2}+\frac{y_2{}^2}{\sigma_2^2}\right]\right\}$$

所以 $(X_1,Y_1)\sim N(0,0,\sigma_1^2,\sigma_2^2,\rho)$，根据例 8.4.2，有 $E(X_1Y_1)=\rho\sigma_1\sigma_2$。

$Cov(X_1,Y_1)=E(X_1Y_1)-E(X_1)E(Y_1)=\rho\sigma_1\sigma_2$，

$$Corr(X,Y)=Corr(X-\mu_1,Y-\mu_2)=Corr(X_1,Y_1)=\frac{Cov(X_1,Y_1)}{\sqrt{Cov(X_1)}\cdot\sqrt{Cov(Y_1)}}=\frac{\rho\sigma_1\sigma_2}{\sigma_1\sigma_2}=\rho。$$

所以，二维正态分布的 5 个参数都有明确的概率意义，$\mu_1,\mu_2$ 为 $X,Y$ 的期望，$\sigma_1^2,\sigma_2^2$ 为 $X,Y$ 的方差，而 $\rho$ 则为 $X,Y$ 的相关系数。

************************************************************

## 9.4 相关与独立

### 相关与独立

随机变量 $X,Y$ 独立时，$Cov(X,Y)=E(XY)-E(X)E(Y)=E(X)E(Y)-E(X)E(Y)=0$

随机变量 $X,Y$ 独立 $\Rightarrow$ $X,Y$ 不相关

随机变量 $X,Y$ 不相关 $\not\Rightarrow$ $X,Y$ 相互独立

***************************************************************************************************************************************

**例 9.4.1** 把一枚均匀硬币抛掷三次，设 $X$ 为三次抛掷中正面出现的次数，而 $Y$ 为正面出现的次数与反面出现的次数之差的绝对值。试求 $X$ 与 $Y$ 的联合分布律以及 $X$ 与 $Y$ 的相关系数，并判断 $X$ 与 $Y$ 是否独立？

解： $X$ 可能取值为 0，1，2，3，而 $Y$ 的可能取值为 1，3，且

$$P(X=0,Y=3)=\frac{1}{8},\ P(X=1,Y=1)=\frac{3}{8},\ P(X=2,Y=1)=\frac{3}{8},\ P(X=3,Y=3)=\frac{1}{8},$$

其余的均为 0，因此，$X$ 与 $Y$ 联合分布律和边缘分布律为：

| $Y\backslash X$ | 0 | 1 | 2 | 3 | $P(Y=y)$ |
|---|---|---|---|---|---|
| 1 | 0 | 3/8 | 3/8 | 0 | 3/4 |
| 3 | 1/8 | 0 | 0 | 1/8 | 1/4 |
| $P(X=x)$ | 1/8 | 3/8 | 3/8 | 1/8 | 1 |

从而

$$E(X)=0\times\frac{1}{8}+1\times\frac{3}{8}+2\times\frac{3}{8}+3\times\frac{1}{8}=\frac{3}{2},\quad E(X^2)=0^2\times\frac{1}{8}+1^2\times\frac{3}{8}+2^2\times\frac{3}{8}+3^2\times\frac{1}{8}=3,$$

$$Var(X)=E(X^2)-E(X)^2=3-\left(\frac{3}{2}\right)^2=\frac{3}{4}。$$

$$E(Y)=1\times\frac{6}{8}+3\times\frac{2}{8}=\frac{3}{2},\quad E(Y^2)=1^2\times\frac{6}{8}+3^2\times\frac{2}{8}=3,$$

$$Var(Y)=E(Y^2)-E(Y)^2=3-(\frac{3}{2})^2=\frac{3}{4}$$

而 $E(XY)=1\times1\times\frac{3}{8}+2\times1\times\frac{3}{8}+0\times3\times\frac{1}{8}+3\times3\times\frac{1}{8}=\frac{9}{4}$。

故 $Corr(X,Y)=\frac{E(XY)-E(X)E(Y)}{\sqrt{Var(X)\cdot Var(Y)}}=\frac{\frac{9}{4}-\frac{3}{2}\times\frac{3}{2}}{\sqrt{\frac{3}{4}\times\frac{3}{4}}}=0$，即 $X$ 与 $Y$ 不相关。

又由于 $0=P(X=0,Y=1)\neq P(X=0)P(Y=1)=\frac{1}{8}\times\frac{6}{8}$，所以 $X$ 与 $Y$ 不独立。

************************************************************

例 9.4.2 考虑 $X\sim N(0,1)$ 与 $Y=X^2$ 的相关性和独立性。

解：显然 $X,Y$ 不独立 （思考：对不独立的关系是否有直观上的理解），

利用概率定义验证：$(X,Y)$ 的联合密度函数 $f(x,y)=\begin{cases}\varphi(x), & x\in R, y=x^2\\ 0, & 其他\end{cases}$，

抛物线 $y=x^2$ 以外的点 $(x,y)$，联合密度 $f(x,y)=0$；

而 $X,Y$ 的边缘密度 $f_X(x)$ 和 $f_Y(y)$ 均不为 0，对抛物线 $y=x^2$ 以外的点

$f(x,y)=f_X(x)\cdot f_Y(y)$ 均不成立，所以 $X,Y$ 不独立。

$Cov(X,Y)=Cov(X,X^2)=E(X^3)-E(X)E(X^2)=0$， $X,Y$ 不相关。

************************************************************

$X\sim N(0,1)$ 和 $Y=X^2$， $X,Y$ 不相关，但它们显然具有很强的关联。

实际上，相关系数反映的是随见变量之间在线性关系意义下的相关程度。

定理： $Corr(X,Y)=\pm1$ 的充要条件是 $X,Y$ 之间几乎处处有线性关系，即存在常数 $a,b$，使得 $P(Y=aX+b)=1$。

所以也称（线性）相关系数，不相关指的是不存在线性相关的关系，相关系数并不能有效地表达非线性的相关关系。

************************************************************

**定理：** $Corr(X,Y)=\pm1$ 的充要条件是 $X,Y$ 之间几乎处处有线性关系，即存在常数 $a,b$，使得 $P(Y=aX+b)=1$。

**充分性** $Y=aX+b\Rightarrow Var(Y)=a^2Var(X)\Rightarrow\sigma_Y=|a|\cdot\sigma_X$

$$Y=aX+b\Rightarrow Cov(X,Y)=a\cdot Cov(X,X)=a\cdot Var(X)=a\cdot\sigma_X^{\ 2}$$

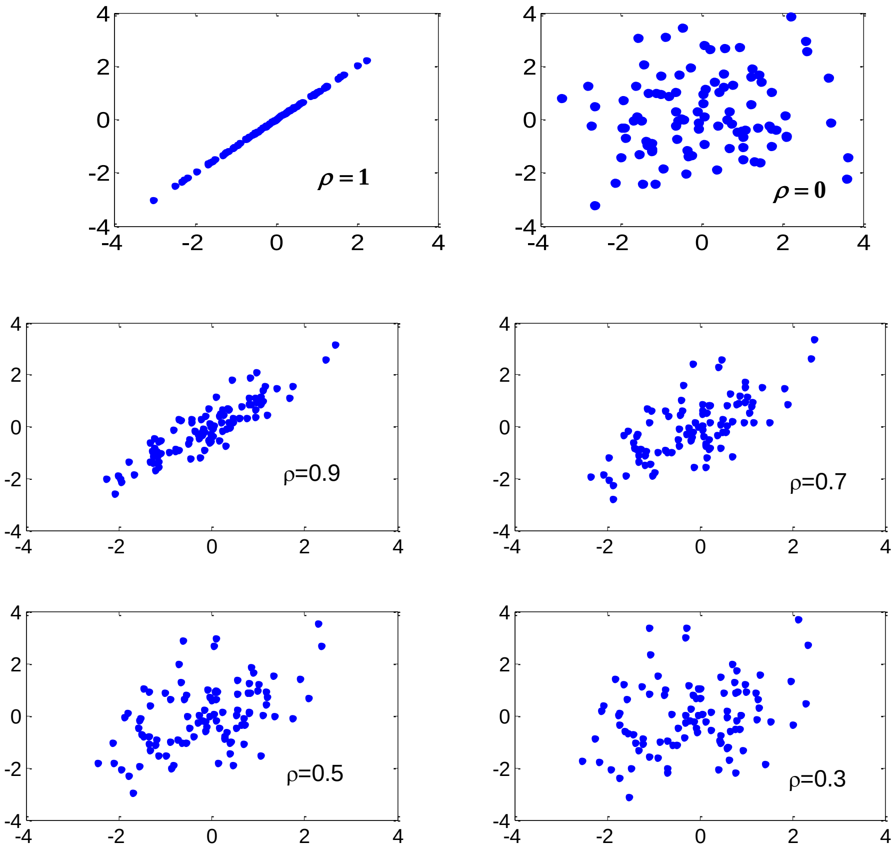
$$Corr(X,Y)=\frac{Cov(X,Y)}{\sigma_X\cdot\sigma_Y}=\frac{a\cdot\sigma_X^{\ 2}}{|a|\cdot\sigma_X^{\ 2}}=\frac{a}{|a|}=\pm1\text{。}$$

**必要性** $Var\left(\frac{X}{\sigma_X}\pm\frac{Y}{\sigma_Y}\right)=Var\left(\frac{X}{\sigma_X}\right)+Var\left(\frac{Y}{\sigma_Y}\right)\pm2Cov\left(\frac{X}{\sigma_X},\frac{Y}{\sigma_Y}\right)=2(1\pm Corr(X,Y))$

$$Corr(X,Y)=\pm1\Rightarrow Var\left(\frac{X}{\sigma_X}\mp\frac{Y}{\sigma_Y}\right)=0\Rightarrow P\left(\frac{X}{\sigma_X}\mp\frac{Y}{\sigma_Y}=c\right)=1$$

*************************************************************

**不同相关程度的示意**




*************************************************************