

COVARIANCE AND CORRELATION

Let (X, Y) be a pair of r.v.s, discrete or continuous, with means and variances

$$EX = \mu_X, \quad \text{Var } X = \sigma_X^2$$

$$EY = \mu_Y, \quad \text{Var } Y = \sigma_Y^2.$$

The following two quantities describe the strength of linear relationship between X and Y .

Defn: The covariance between X and Y is

$$\text{Cov}(X, Y) = E(X - \mu_X)(Y - \mu_Y) =: \sigma_{XY}$$

We sometimes use σ_{XY} to denote $\text{Cov}(X, Y)$.

Defn: The correlation between X and Y is

$$\text{corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var } X} \sqrt{\text{Var } Y}} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} =: \rho_{XY}$$

We sometimes use ρ_{XY} to denote $\text{corr}(X, Y)$.

Remark: The correlation ρ_{XY} is a unitless measure of the strength of linear relationship between X and Y ; the division of $\text{Cov}(X, Y)$ by $\sigma_X \sigma_Y$ results in a quantity bounded between -1 and 1 . That is, we always have

$$-1 \leq \rho_{XY} \leq 1.$$

A proof of this will be given in the addendum to these notes.

Useful expression: $\text{Cov}(X, Y) = EXY - EXEY$.

Result: For $a, b, c, d \in \mathbb{R}$, we have

$$\text{Cov}(aX + b, cY + d) = ac \text{Cov}(X, Y)$$

$$\text{corr}(aX + b, cY + d) = \text{corr}(X, Y).$$

Ex. let (X, Y) have joint pdf given by

$$f(x, y) = \frac{1}{8}(x+y) \mathbb{1}(0 < x < 2, 0 < y < 2).$$

(a) Find $\text{Cov}(X, Y)$.

We have

$$\begin{aligned} \mathbb{E}XY &= \int_0^2 \int_0^2 xy \cdot \frac{1}{8}(x+y) dx dy \\ &= \frac{1}{8} \int_0^2 \int_0^2 (x^2y + xy^2) dx dy \\ &= \frac{1}{8} \int_0^2 \left(\frac{x^3}{3}y + \frac{x^2}{2}y^2 \right) \Big|_0^2 dy \\ &= \frac{1}{8} \int_0^2 \left(\frac{8}{3}y + \frac{4}{2}y^2 \right) dy \\ &= \frac{1}{8} \left(\frac{8}{3} \frac{y^2}{2} + \frac{4}{2} \frac{y^3}{3} \right) \Big|_0^2 \\ &= \frac{1}{6} \cdot 4 + \frac{4}{6} \\ &= \frac{4}{3}. \end{aligned}$$

To find $\mathbb{E}X$ and $\mathbb{E}Y$, we need the marginal pdfs f_X and f_Y .

For $x \in (0, 2)$, we have

$$f_X(x) = \int_0^2 \frac{1}{8}(x+y) dy = \frac{1}{8} \left(xy + \frac{y^2}{2} \right) \Big|_0^2 = \frac{1}{8}(2x+2) = \frac{1}{4}(x+1).$$

So

$$\mathbb{E}X = \int_0^2 x \cdot \frac{1}{4}(x+1) dx = \frac{1}{4} \left(\frac{x^3}{3} + \frac{x^2}{2} \right) \Big|_0^2 = \frac{1}{4} \left(\frac{8}{3} + \frac{4}{2} \right) = \frac{7}{6}.$$

We also get $\mathbb{E}Y = 7/6$.

So

$$\text{Cov}(X, Y) = \frac{4}{3} - \frac{7}{6} \cdot \frac{7}{6} = \frac{48 - 49}{36} = -\frac{1}{36}.$$

(b) Find $\text{corr}(X, Y)$.

We have $\text{Var} X = \mathbb{E}X^2 - (\mathbb{E}X)^2$, with

$$\mathbb{E}X^2 = \int_0^2 x^2 \cdot \frac{1}{4}(x+1) dx = \frac{1}{4} \left(\frac{x^4}{4} + \frac{x^3}{3} \right) \Big|_0^2 = \frac{1}{4} \left(\frac{16}{4} + \frac{8}{3} \right) = \frac{5}{3},$$

so that

$$\text{Var} X = \frac{5}{3} - \left(\frac{7}{6} \right)^2 = \frac{60 - 49}{36} = \frac{11}{36}.$$

We also get $\text{Var} Y = \frac{11}{36}$.

So

$$\text{corr}(X, Y) = \frac{-1/36}{\sqrt{\frac{11}{36} \cdot \frac{11}{36}}} = -\frac{1}{11}.$$

E.g. Roll two dice and let X be the maximum and Y be the minimum of the rolls.

(i) Find $\text{Cov}(X, Y)$.

The joint pmf p of X and Y takes the values:

		$X = \max$						$P_Y(Y=y)$
		1	2	3	4	5	6	
$Y = \min$	1	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{2}{36}$	$\frac{2}{36}$	$\frac{2}{36}$	$\frac{2}{36}$	$\frac{11}{36}$
	2		$\frac{1}{36}$	$\frac{2}{36}$	$\frac{2}{36}$	$\frac{2}{36}$	$\frac{2}{36}$	$\frac{7}{36}$
	3			$\frac{1}{36}$	$\frac{2}{36}$	$\frac{2}{36}$	$\frac{2}{36}$	$\frac{7}{36}$
	4				$\frac{1}{36}$	$\frac{2}{36}$	$\frac{2}{36}$	$\frac{5}{36}$
	5					$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$
	6						$\frac{1}{36}$	$\frac{1}{36}$
$P_X(X=x)$		$\frac{1}{36}$	$\frac{2}{36}$	$\frac{5}{36}$	$\frac{7}{36}$	$\frac{7}{36}$	$\frac{11}{36}$	

$$E X = \frac{1(2) + 2(3) + 3(5) + 4(7) + 5(9) + 6(11)}{36} = \frac{161}{36}$$

$$E Y = \frac{1(11) + 2(9) + 3(7) + 4(5) + 5(3) + 6(1)}{36} = \frac{91}{36}$$

$$\begin{aligned}
 E X Y &= \frac{1}{36} \left[1^2 + 2^2 + 3^2 + 4^2 + 5^2 + 6^2 \right] \\
 &\quad + \frac{2}{36} \left[1(2+3+4+5+6) + 2(3+4+5+6) + 3(4+5+6) \right. \\
 &\quad \quad \quad \left. + 4(5+6) + 5(6) \right] \\
 &= \frac{1}{36} [91] + \frac{2}{36} [20 + 36 + 45 + 44 + 30] \\
 &= \frac{441}{36}
 \end{aligned}$$

$$\begin{aligned} \text{So } \text{Cov}(X, Y) &= \frac{441}{36} - \left(\frac{91}{36}\right)\left(\frac{161}{36}\right) \\ &= \frac{1225}{1296} \approx .945 \end{aligned}$$

(ii) Find $\text{corr}(X, Y)$.

$$\mathbb{E}X^2 = \frac{1^2(2) + 2^2(3) + 3^2(5) + 4^2(4) + 5^2(4) + 6^2(11)}{36} = \frac{791}{36}$$

$$\text{So } \text{Var } X = \frac{791}{36} - \left(\frac{161}{36}\right)^2 = \frac{2555}{1296}$$

$$\mathbb{E}Y^2 = \frac{1^2(11) + 2^2(9) + 3^2(7) + 4^2(5) + 5^2(2) + 6^2(1)}{36} = \frac{301}{36}$$

$$\text{So } \text{Var } Y = \frac{301}{36} - \left(\frac{91}{36}\right)^2 = \frac{2555}{1296}$$

$$\text{Then } \text{corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var } X} \sqrt{\text{Var } Y}} = \frac{\left(\frac{1225}{1296}\right)}{\sqrt{\frac{2555}{1296}} \sqrt{\frac{2555}{1296}}} = \frac{1225}{2555} = \frac{245}{511} \approx .479$$

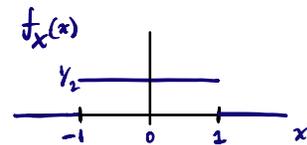
Thm: If X and Y are independent, then $\text{Cov}(X, Y) = 0$.

Proof: $\text{Cov}(X, Y) = \mathbb{E}XY - \mathbb{E}X\mathbb{E}Y = \mathbb{E}X\mathbb{E}Y - \mathbb{E}X\mathbb{E}Y = 0$.
= $\mathbb{E}X\mathbb{E}Y$, because X and Y are independent.

Remark: If $\text{Cov}(X, Y) = 0$, X and Y are not necessarily independent!

E.g.: Let $X \sim \text{Uniform}(-1, 1)$ and let $Y = X^2$ (X and Y are not independent).

$$\begin{aligned} \text{Cov}(X, Y) &= \mathbb{E}XY - \mathbb{E}X\mathbb{E}Y \\ &= \mathbb{E}X^3 - \mathbb{E}X\mathbb{E}X^2 \\ &= 0, \end{aligned}$$



since $\mathbb{E}X^3 = \int_{-1}^1 x^3 \cdot \frac{1}{2} dx = \frac{x^4}{8} \Big|_{-1}^1 = 0$ and $\mathbb{E}X = \int_{-1}^1 x \cdot \frac{1}{2} dx = 0$.

So $\text{Cov}(X, Y) = 0$, but X and Y are not independent.

Ex: let (X, Y) have joint pdf

$$f(x, y) = \frac{1}{2} \frac{1}{|x|} e^{-y/|x|} \mathbb{1}(x \in (-1, 1) \setminus \{0\}, 0 < y).$$

(a) Check whether X and Y are independent.

We cannot find a factorization of $f(x, y)$ such that

$$f(x, y) = g(x)h(y),$$

so X and Y are not independent.

(b) Compute $Cov(X, Y)$.

$$\begin{aligned} \mathbb{E}XY &= \int_{-1}^1 \int_0^{\infty} x \cdot y \frac{1}{2} \frac{1}{|x|} e^{-y/|x|} dy dx \\ &= \int_{-1}^1 \frac{x}{2} \int_0^{\infty} \frac{y}{|x|} e^{-y/|x|} dy dx \\ &\quad \underbrace{\hspace{10em}}_{\text{Expected value of Exponential}(|x|)} \\ &= \int_{-1}^1 \frac{x}{2} |x| dx \\ &= \int_{-1}^0 -\frac{x^2}{2} dx + \int_0^1 \frac{x^2}{2} dx \\ &= -\frac{x^3}{6} \Big|_{-1}^0 + \frac{x^3}{6} \Big|_0^1 \\ &= -\frac{1}{6} + \frac{1}{6} \end{aligned}$$

$$= 0.$$

To find $\mathbb{E}X$, we need the marginal pdf f_X of X .

$$\begin{aligned} f_X(x) &= \int_0^\infty \frac{1}{2} \frac{1}{|x|} e^{-y/|x|} dy \mathbb{1}(x \in (-1, 1) \setminus \{0\}) \\ &= \frac{1}{2} \mathbb{1}(x \in (-1, 1) \setminus \{0\}). \end{aligned}$$

so

$$\begin{aligned} \mathbb{E}X &= \int_{-\infty}^{\infty} \frac{1}{2} \mathbb{1}(x \in (-1, 1) \setminus \{0\}) dx \\ &= \int_{-1}^1 \frac{1}{2} dx \\ &= \left. \frac{x}{2} \right|_{-1}^1 \\ &= 0. \end{aligned}$$

single point makes no contribution to the integral, so we can just integrate from -1 to 1.

$$\text{Therefore, } \text{Cov}(X, Y) = \underbrace{\mathbb{E}XY}_{=0} - \underbrace{\mathbb{E}X}_{=0} \underbrace{\mathbb{E}Y}_{\text{doesn't matter}} = 0.$$

The covariance comes into play when we wish to compute the variance of a sum of r.v.s.

Thm: For any $a, b \in \mathbb{R}$,

$$\text{Var}(aX + bY) = a^2 \text{Var} X + b^2 \text{Var} Y + 2ab \text{Cov}(X, Y).$$

Proof:

$$\begin{aligned} \text{Var}(aX + bY) &= \mathbb{E} \left[aX + bY - \mathbb{E}(aX + bY) \right]^2 \\ &= \mathbb{E} \left[a(X - \mathbb{E}X) + b(Y - \mathbb{E}Y) \right]^2 \\ &= \mathbb{E} \left[a^2(X - \mathbb{E}X)^2 + b^2(Y - \mathbb{E}Y)^2 + 2ab(X - \mathbb{E}X)(Y - \mathbb{E}Y) \right] \\ &= a^2 \underbrace{\mathbb{E}(X - \mathbb{E}X)^2}_{\text{Var } X} + b^2 \underbrace{\mathbb{E}(Y - \mathbb{E}Y)^2}_{\text{Var } Y} + 2ab \underbrace{\mathbb{E}(X - \mathbb{E}X)(Y - \mathbb{E}Y)}_{\text{Cov}(X, Y)} \end{aligned}$$

E.g. Suppose $\text{Var} X = 2$, $\text{Var} Y = 3$, and $\text{Cov}(X, Y) = -3/2$.

Find $\text{Var}(3X - Y)$.
 $\uparrow a=3 \quad \uparrow b=-1$

$$\begin{aligned} \text{Var}(3X - Y) &= (3)^2 \text{Var} X + (-1)^2 \text{Var} Y + 2(3)(-1) \text{Cov}(X, Y) \\ &= 9(2) + 1(3) + 2(3)(-1)(-3/2) \\ &= 30 \end{aligned}$$

We can extend the previous theorem to a sum of any number n of r.v.s.

Thm: Let X_1, \dots, X_n be r.v.s and let $a_1, \dots, a_n \in \mathbb{R}$. Then

$$\text{Var} \left(\sum_{i=1}^n a_i X_i \right) = \sum_{i=1}^n a_i^2 \text{Var}(X_i) + 2 \sum_{i > j} a_i a_j \text{Cov}(X_i, X_j).$$

Proof:

$$\begin{aligned} \text{Var} \left(\sum a_i X_i \right) &= \mathbb{E} \left[\sum_{i=1}^n a_i X_i - \mathbb{E} \left(\sum_{i=1}^n a_i X_i \right) \right]^2 && (c_1 + c_2 + \dots + c_n)^2 = c_1^2 + c_1 c_2 + \dots + c_1 c_n \\ & && c_2 c_1 + c_2^2 + \dots + c_2 c_n \\ & && \vdots \quad \vdots \quad \quad \quad \vdots \\ & && c_n c_1 + c_n c_2 + \dots + c_n^2 \\ & && = \sum_{i=1}^n c_i^2 + 2 \sum_{i > j} c_i c_j \\ &= \mathbb{E} \left[\sum_{i=1}^n a_i^2 (X_i - \mathbb{E}X_i)^2 + 2 \sum_{i > j} a_i a_j (X_i - \mathbb{E}X_i)(X_j - \mathbb{E}X_j) \right] \end{aligned}$$

$$\begin{aligned}
&= \sum_{i=1}^n a_i^2 \mathbb{E}(X_i - \mathbb{E}X_i)^2 + 2 \sum_{i>j} a_i a_j \mathbb{E}(X_i - \mathbb{E}X_i)(X_j - \mathbb{E}X_j) \\
&= \sum_{i=1}^n a_i^2 \text{Var}(X_i) + 2 \sum_{i>j} a_i a_j \text{Cov}(X_i, X_j).
\end{aligned}$$

Corollary: let X_1, \dots, X_n be independent r.v.s with common variance σ^2
and let $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$. Then

$$\text{Var}(\bar{X}) = \frac{\sigma^2}{n}.$$

Proof: $\text{Var}(\bar{X}) = \text{Var}\left(\frac{1}{n} \sum_{i=1}^n X_i\right)$

$$\begin{aligned}
&= \sum_{i=1}^n \left(\frac{1}{n}\right)^2 \text{Var}(X_i) + 2 \sum_{i>j} \underbrace{\left(\frac{1}{n}\right)\left(\frac{1}{n}\right)}_{=0, \text{ since } X_i, X_j \text{ independent for all } i, j} \text{Cov}(X_i, X_j) \\
&= \left(\frac{1}{n}\right)^2 \sum_{i=1}^n \sigma^2 \\
&= \frac{\sigma^2}{n}.
\end{aligned}$$

We now introduce the bivariate Normal distribution, in which the correlation between the variables appears as a parameter.

BIVARIATE NORMAL DISTRIBUTION

The random variable pair (X, Y) has the bivariate Normal distribution if the joint pdf of (X, Y) is

$$f(x, y; \mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho) = \frac{1}{2\pi} \frac{1}{\sigma_x \sigma_y \sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)} \left[\left(\frac{x-\mu_x}{\sigma_x}\right)^2 - 2\rho \left(\frac{x-\mu_x}{\sigma_x}\right) \left(\frac{y-\mu_y}{\sigma_y}\right) + \left(\frac{y-\mu_y}{\sigma_y}\right)^2 \right]\right),$$

where

- $\mu_x \in \mathbb{R}$ and $\sigma_x^2 \in (0, \infty)$ are the marginal mean and variance of X
- $\mu_y \in \mathbb{R}$ and $\sigma_y^2 \in (0, \infty)$ are the marginal mean and variance of Y
- $\rho \in [0, 1)$ is the correlation between X and Y

We write $(X, Y) \sim \text{biv Normal}(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho)$.

The marginal distributions of X and Y are

$$X \sim \text{Normal}(\mu_X, \sigma_X^2) \quad \text{and} \quad Y \sim \text{Normal}(\mu_Y, \sigma_Y^2).$$

Note that if $\rho = 0$, the joint pdf of (X, Y) may be written as the product of the marginal pdfs of X and Y :

$$f(x, y; \mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho = 0) = \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma_X} \exp\left[-\frac{1}{2} \frac{(x - \mu_X)^2}{\sigma_X^2}\right] \cdot \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma_Y} \exp\left[-\frac{1}{2} \frac{(y - \mu_Y)^2}{\sigma_Y^2}\right].$$

Thus, if (X, Y) are bivariate Normal with correlation equal to 0, then X and Y are independent.

* This is a special case in which $\text{corr}(X, Y) = 0$ implies independence of X and Y . Keep in mind that this does not hold in general.

The bivariate Normal distribution with $\mu_X = \mu_Y = 0$ and $\sigma_X^2 = \sigma_Y^2 = 1$ is called the standard bivariate Normal distribution.

If (Z_1, Z_2) has the standard bivariate Normal distribution then the joint pdf of (Z_1, Z_2) is given by

$$\phi(z_1, z_2; \rho) = \frac{1}{2\pi} \frac{1}{\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2(1-\rho^2)} (z_1^2 - 2\rho z_1 z_2 + z_2^2)\right].$$

If $(X, Y) \sim \text{biv Normal}(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho)$ then Standard bivariate Normal distribution

$$(X, Y) \mapsto \left(\frac{X - \mu_X}{\sigma_X}, \frac{Y - \mu_Y}{\sigma_Y}\right) \sim \text{biv Normal}(0, 0, 1, 1, \rho),$$

and, likewise, if $(Z_1, Z_2) \sim \text{biv Normal}(0, 0, 1, 1, \rho)$, then

$$(Z_1, Z_2) \mapsto (\mu_X + \sigma_X Z_1, \mu_Y + \sigma_Y Z_2) \sim \text{biv Normal}(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho).$$

ADDENDUM: To show that $\rho_{XY} \in [-1, 1]$, we will make use of the following:

Thm: (Cauchy-Schwarz inequality): For any r.v.s X and Y ,

$$|\mathbb{E}XY| \leq \mathbb{E}|XY| \leq \sqrt{\mathbb{E}X^2} \sqrt{\mathbb{E}Y^2}.$$

Proof: • We first show that $|\mathbb{E}XY| \leq \mathbb{E}|XY|$.

Since $-|xy| \leq xy \leq |xy|$ for any $x, y \in \mathbb{R}$, we have if X and Y are continuous with joint pdf $f(x, y)$,

(If X, Y discrete, replace f with the joint pmf of X and Y and replace integrals with sums.)

$$\iint_{\mathbb{R}^2} -|xy| f(x, y) dx dy \leq \iint_{\mathbb{R}^2} xy f(x, y) dx dy \leq \iint_{\mathbb{R}^2} |xy| f(x, y) dx dy,$$

so that $-\mathbb{E}|XY| \leq \mathbb{E}XY \leq \mathbb{E}|XY|$, giving $|\mathbb{E}XY| \leq \mathbb{E}|XY|$.

(for any $x, a \in \mathbb{R}, a > 0, -a \leq x \leq a \Leftrightarrow |x| \leq a$)

• Now we show the second inequality $\mathbb{E}|XY| \leq \sqrt{\mathbb{E}X^2} \sqrt{\mathbb{E}Y^2}$.

We will use the fact that $ab \leq \frac{1}{2}a^2 + \frac{1}{2}b^2$ for all $a, b \in \mathbb{R}$, which we can get from

$$0 \leq (a-b)^2 = a^2 + b^2 - 2ab \Leftrightarrow 2ab \leq a^2 + b^2$$

$$\uparrow \begin{array}{l} \text{a squared quantity} \\ \text{is always non-negative} \end{array} \Leftrightarrow ab \leq \frac{1}{2}a^2 + \frac{1}{2}b^2.$$

Now, setting $a = \frac{|X|}{\sqrt{\mathbb{E}X^2}}$ and $b = \frac{|Y|}{\sqrt{\mathbb{E}Y^2}}$ we get

for any $a, b \in \mathbb{R}$,
 $|a||b| = |ab|$

$$\frac{|XY|}{\sqrt{\mathbb{E}X^2} \sqrt{\mathbb{E}Y^2}} \leq \frac{1}{2} \left(\frac{|X|}{\sqrt{\mathbb{E}X^2}} \right)^2 + \frac{1}{2} \left(\frac{|Y|}{\sqrt{\mathbb{E}Y^2}} \right)^2.$$

Taking expectations of both sides, we get

$$\frac{\mathbb{E}|XY|}{\sqrt{\mathbb{E}X^2} \sqrt{\mathbb{E}Y^2}} \leq 1$$

since $\mathbb{E} \left(\frac{|X|}{\sqrt{\mathbb{E}X^2}} \right)^2 = \frac{\mathbb{E}X^2}{\mathbb{E}X^2} = 1$ and $\mathbb{E} \left(\frac{|Y|}{\sqrt{\mathbb{E}Y^2}} \right)^2 = \frac{\mathbb{E}Y^2}{\mathbb{E}Y^2} = 1$.

This gives $\mathbb{E}|XY| \leq \sqrt{\mathbb{E}X^2} \sqrt{\mathbb{E}Y^2}$.

Corollary: The correlation ρ_{XY} between any two r.v.s X and Y is in the interval $[-1, 1]$.

Proof: Let $U = X - EX$ and $V = Y - EY$. Then

$$\rho_{XY} = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}X} \sqrt{\text{Var}Y}} = \frac{E(X - EX)(Y - EY)}{\sqrt{E(X - EX)^2} \sqrt{E(Y - EY)^2}} = \frac{EU V}{\sqrt{EU^2} \sqrt{EV^2}}.$$

So

$$|\rho_{XY}| = \frac{|EU V|}{\sqrt{EU^2} \sqrt{EV^2}} \leq \frac{E|UV|}{\sqrt{EU^2} \sqrt{EV^2}} \leq 1,$$

since by Cauchy-Schwarz $|EU V| \leq E|UV| \leq \sqrt{EU^2} \sqrt{EV^2}$.