Correlation between two random variables

The covariance between random variables X and Y is defined as

$$\mathrm{Cov}(X,Y) = \mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y).$$

The correlation between X and Y is the covariance between standardized variables $X^* = (X-\mu_X)/\sigma_X$ and $Y^* = (Y-\mu_Y)/\sigma_Y$, where $\mu_X = \mathrm{E}(X)$, $\mu_Y = \mathrm{E}(Y)$ and $\sigma_X^2 = \mathrm{Var}(X)$, $\sigma_Y^2 = \mathrm{Var}(Y)$. The correlation between X and Y is therefore defined as

$$\rho_{X,Y} = \mathrm{Cor}(X,Y) = \mathrm{Cov}(X^*,Y^*) = \mathrm{E}(X^*Y^*) - \mathrm{E}(X^*)\mathrm{E}(Y^*) = \frac{\mathrm{Cov}(X,Y)}{\sigma_X\sigma_Y}.$$

The correlation is a measure of dependence between X and Y. Some properties:

- $-1 \le \rho_{X,Y} \le 1$ for any random variables X and Y
- If $\rho_{X,Y} = 1$, then $Y = \alpha + \beta X$ with $\beta > 0$
- If $\rho_{X,Y} = -1$, then $Y = \alpha + \beta X$ with $\beta < 0$
- If X and Y are independent, then $\rho_{X,Y} = 0$. **The opposite is not true in the general case**.

Some useful properties of the covariance that can be derived from the definition. For any random variables X, Y, Z, W and constants a, b, c, d we have

- $\mathrm{Var}(aX + bY) = a^2\mathrm{Var}(X) + b^2\mathrm{Var}(Y) + 2ab\mathrm{Cov}(X,Y)$
- $\mathrm{Cov}(X,Y) = 0$ if X and Y are independent
- $\mathrm{Cov}(X,X) = \mathrm{Var}(X)$
- $\mathrm{Cov}(X,Y+Z) = \mathrm{Cov}(X,Y) + \mathrm{Cov}(X,Z)$
- $\mathrm{Cov}(X+Y,Z) = \mathrm{Cov}(X,Z) + \mathrm{Cov}(Y,Z)$
- $\mathrm{Cov}(aX,bY) = ab\mathrm{Cov}(X,Y)$
- $\mathrm{Cov}(aX+bY,cZ+dW) = ac\mathrm{Cov}(X,Z) + ad\mathrm{Cov}(X,W) + bc\mathrm{Cov}(Y,Z) + bd\mathrm{Cov}(Y,W)$ **(linearity property)**.

You can use all these properties without proof.

Bivariate normal distribution

X and Y have a standard bivariate normal distribution with correlation ρ if we can write

$$X = Z_1, \quad Y = \rho Z_1 + \sqrt{1-\rho^2} Z_2,$$

where Z_1 and Z_2 are two independent standard normal random variables. From the definition, X and Y also have the standard normal distribution and

$$\begin{aligned} \rho_{X,Y} &= \mathrm{Cor}(X,Y) = \mathrm{Cov}(X,Y) = \mathrm{Cov}(Z_1, \rho Z_1 + \sqrt{1-\rho^2} Z_2) \\ &= \rho \mathrm{Cov}(Z_1, Z_1) + \sqrt{1-\rho^2} \mathrm{Cov}(Z_1, Z_2) = \rho. \end{aligned}$$

In the general case, if $\mu_X = \mathrm{E}(X), \mu_Y = \mathrm{E}(Y)$ and $\sigma_X^2 = \mathrm{Var}(X), \sigma_Y^2 = \mathrm{Var}(Y)$, then X and Y have a bivariate normal distribution with correlation ρ if $X^* = (X - \mu_X)/\sigma_X$ and $Y^* = (Y - \mu_Y)/\sigma_Y$ have a standard bivariate normal distribution with correlation ρ, i.e.,

$$X^* = Z_1, \quad Y^* = \rho Z_1 + \sqrt{1-\rho^2} Z_2,$$

or

$$X = \mu_X + \sigma_X Z_1, \quad Y = \mu_Y + \rho \sigma_Y Z_1 + \sqrt{1-\rho^2} \sigma_Y Z_2.$$

Some useful properties:

- The conditional distribution of Y given $X = x$ is a normal distribution with mean $\mu^* = \mu_Y + \rho \sigma_Y (x - \mu_X)/\sigma_X$ and variance $\sigma^{*2} = \sigma_Y^2 (1 - \rho^2)$
- In particular, if $\rho = 0$, the conditional distribution is the same as marginal: it is normal with mean $\mu^* = \mu_Y$ and variance $\sigma^{*2} = \sigma_Y^2$. **Zero correlation implies independence for bivariate normal random variables X and Y**
- If X and Y have a bivariate normal distribution, then $W_1 = aX + bY$ and $W_2 = cX + dY$ also have a bivariate normal distribution (with different parameters)
- X and Y have a bivariate normal distribution if and only if any linear combination $aX + bY$ has a univariate normal distribution
- Bivariate normality is a stronger assumption than univariate normality: if $X \sim N(\mu_X, \sigma_X^2)$ and $Y \sim N(\mu_Y, \sigma_Y^2)$, then it does NOT imply that X and Y have a bivariate normal distribution.

Example 1: Assume that X and Y have a bivariate distribution with parameters $\mu_X = \mu_Y = 1$, $\sigma_X^2 = \sigma_Y^2 = 2.5$ and $\rho = 0.8$.

(a) Find the distribution of $W = 2X - 3Y$,

(b) Find the joint distribution of $W_1 = X + 2Y$ and $W_2 = 2X + Y$.

Solution:

(a) The distribution of W is normal because it is a linear combination of two random variables that have a bivariate normal distribution. We need to find the parameters of this normal distribution, mean and variance:

$$\mathrm{E}(W) = 2\mathrm{E}(X) - 3\mathrm{E}(Y) = 2 - 3 = -1,$$

$$\mathrm{Var}(W) = 2^2 \cdot \mathrm{Var}(X) + 3^2 \cdot \mathrm{Var}(Y) - 2 \cdot 2 \cdot 3 \cdot \mathrm{Cov}(X, Y) = 10 + 22.5 - 12 \cdot 2 = 8.5$$

because $\mathrm{Cov}(X, Y) = \sigma_X \sigma_Y \mathrm{Cor}(X, Y) = 0.8 \cdot 2.5 = 2$. So, $\mathbf{W \sim N(-1, 8.5)}$.

(b) The joint distribution of W_1 and W_2 is bivariate normal because both W_1 and W_2 are linear combinations of two random variables that have a bivariate normal distribution. We need to find the parameters of this bivariate normal distribution, $\mu_1 = \mathrm{E}(W_1), \mu_2 = \mathrm{E}(W_2), \sigma_1^2 = \mathrm{Var}(W_1), \sigma_2^2 = \mathrm{Var}(W_2)$ and $\rho = \mathrm{Cor}(W_1, W_2)$. We find:

$$\mu_1 = \mathrm{E}(X) + 2\mathrm{E}(Y) = 1 + 2 = \mathbf{3}, \quad \mu_2 = 2\mathrm{E}(X) + \mathrm{E}(Y) = 2 + 1 = \mathbf{3},$$

$$\sigma_1^2 = \mathrm{Var}(X) + 2^2 \cdot \mathrm{Var}(Y) + 2 \cdot 2 \cdot \mathrm{Cov}(X, Y) = 2.5 + 10 + 8 = \mathbf{20.5},$$
$$\sigma_2^2 = 2^2 \cdot \mathrm{Var}(X) + \mathrm{Var}(Y) + 2 \cdot 2 \cdot \mathrm{Cov}(X, Y) = 10 + 2.5 + 8 = \mathbf{20.5},$$

$$\begin{aligned}
\mathrm{Cov}(W_1, W_2) &= 2\mathrm{Cov}(X, X) + \mathrm{Cov}(X, Y) + 4\mathrm{Cov}(X, Y) + 2\mathrm{Cov}(Y, Y) \\
&= 2\mathrm{Var}(X) + 5\mathrm{Cov}(X, Y) + 2\mathrm{Var}(Y) = 5 + 10 + 5 = 20, \\
\mathrm{Cor}(W_1, W_2) &= \mathrm{Cov}(W_1, W_2)/(\sigma_1 \sigma_2) = 20/20.5 = \mathbf{0.976}.
\end{aligned}$$

Example 2: Assume that X and Y have a standard bivariate normal distribution with the correlation $\rho = 0.6$.

(a) Find the conditional expectation $\mathrm{E}(Y^2 | X = 2Y)$,

(b) Find k such that the conditional variance $\mathrm{Var}(X + Y | X = kY) = 1.6$.

Solution:

(a) We need to find the second moment of W where W has a conditional distribution of Y given $X = 2Y$. Note that the event "$X = 2Y$" is equivalent to "$V = X - 2Y = 0$". The joint distribution of V and Y is bivariate normal because both $V = X - 2Y$ and Y are linear combinations of two random variables, X and Y, that have a bivariate normal distribution. We find the parameters of the distribution of (Y, V):

$$\mu_Y = \mathrm{E}(Y) = 0, \quad \sigma_Y^2 = \mathrm{Var}(Y) = 1, \quad \mu_V = \mathrm{E}(V) = \mathrm{E}(X) - 2\mathrm{E}(Y) = 0,$$

$$\sigma_V^2 = \mathrm{Var}(V) = \mathrm{Var}(X) + 2^2 \cdot \mathrm{Var}(Y) - 2 \cdot 2 \cdot \mathrm{Cov}(X, Y) = 1 + 4 - 4 \cdot 0.6 = 2.6,$$

because $\text{Cov}(X, Y) = \sigma_X \sigma_Y \text{Cor}(X, Y) = 0.6$. Finally,

$$\text{Cov}(V, Y) = \text{Cov}(X, Y) - 2\text{Cov}(Y, Y) = \text{Cov}(X, Y) - 2\text{Var}(Y) = 0.6 - 2 = -1.4,$$

$$\rho_{V,Y} = \text{Cor}(V, Y) = \text{Cov}(V, Y)/(\sigma_V \sigma_Y) = -1.4/\sqrt{2.6} = -0.868.$$

The conditional distribution of Y given $V = 0$ is normal with mean and variance:

$$\mu^* = \mu_Y + \rho_{V,Y}\sigma_Y(0 - \mu_V)/\sigma_V = 0 - 0.868(0 - 0)/\sqrt{2.6} = 0,$$

$$\sigma^{*2} = \sigma_Y^2(1 - \rho_{V,Y}^2) = 1 - 0.754 = 0.246,$$

So $W \sim N(\mu^*, \sigma^{*2})$ and therefore $\text{E}(W^2) = \text{Var}(W) + \{\text{E}(W)\}^2 = (\mu^*)^2 + \sigma^{*2} = \mathbf{0.246}$.

(b) We can write:

$$\text{Var}(X + Y|X = kY) = \text{Var}(W_2 = X + Y|W_1 = X - kY = 0).$$

We need to find the conditional distribution of W_2 given $W_1 = 0$. The joint distribution of W_1 and W_2 is bivariate normal because both $W_1 = X - kY$ and $W_2 = X + Y$ are linear combinations of two random variables, X and Y, that have a bivariate normal distribution. We find the parameters of the distribution of (W_1, W_2):

$$\mu_1 = \text{E}(W_1) = \text{E}(X) - k\text{E}(Y) = 0, \quad \mu_2 = \text{E}(W_2) = \text{E}(X) + \text{E}(Y) = 0,$$

$$\sigma_1^2 = \text{Var}(W_1) = \text{Var}(X) + k^2 \cdot \text{Var}(Y) - 2 \cdot k \cdot \text{Cov}(X, Y) = 1 + k^2 - 2 \cdot k \cdot 0.6 = k^2 - 1.2k + 1,$$

$$\sigma_2^2 = \text{Var}(W_2) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y) = 1 + 1 + 2 \cdot 0.6 = 3.2,$$

$$\begin{aligned}
\text{Cov}(W_1, W_2) &= \text{Cov}(X - kY, X + Y) \\
&= \text{Cov}(X, X) + \text{Cov}(X, Y) - k\text{Cov}(Y, X) - k\text{Cov}(Y, Y) \\
&= 1 + 0.6 - 0.6k - k = 1.6(1 - k), \\
\rho_{1,2} = \text{Cor}(W_1, W_2) &= \text{Cov}(W_1, W_2)/(\sigma_1\sigma_2) = 1.6(1 - k)/\sqrt{3.2(k^2 - 1.2k + 1)}.
\end{aligned}$$

It follows that the conditional distribution of W_2 given $W_1 = 0$ is normal with mean and variance:

$$\mu^* = \mu_2 + \rho_{1,2}\sigma_2(0 - \mu_1)/\sigma_1 = 0,$$

$$\sigma^{*2} = \text{Var}(W_2|W_1 = 0) = \sigma_2^2(1 - \rho_{1,2}^2) = 3.2 \cdot \left(1 - \frac{1.6^2(1 - k)^2}{3.2(k^2 - 1.2k + 1)}\right) = 1.6,$$

We solve for k:

$$1.6(1 - k)^2 = k^2 - 1.2k + 1 \quad \Rightarrow \quad 0.6k^2 - 2k + 0.6 = 0 \quad \Rightarrow \quad \mathbf{k_1 = 1/3},\ \mathbf{k_2 = 3}.$$