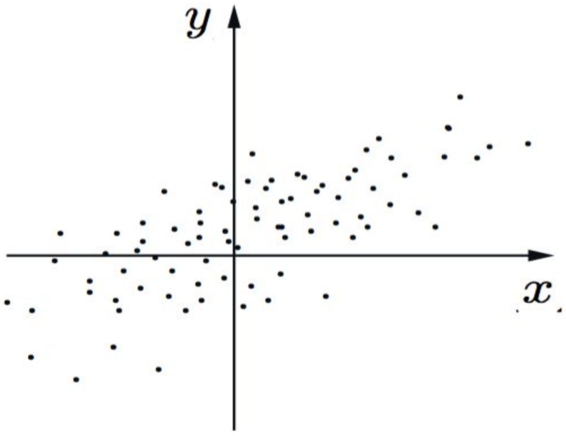
Covariance

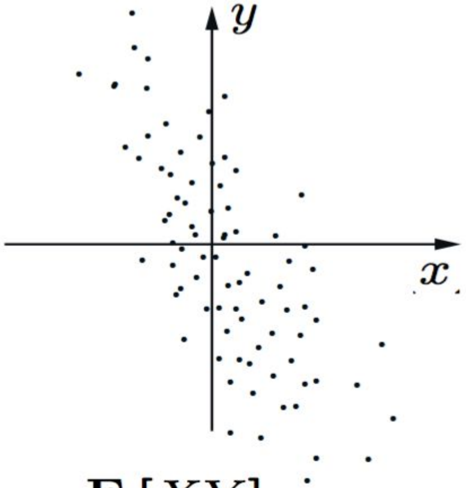
- Zero-mean, discrete X and Y
  - if independent: $\mathbf{E}[XY] =$

$$= E[x]\,E[y] = 0$$

But suppose that joint PMF is one of the following kinds



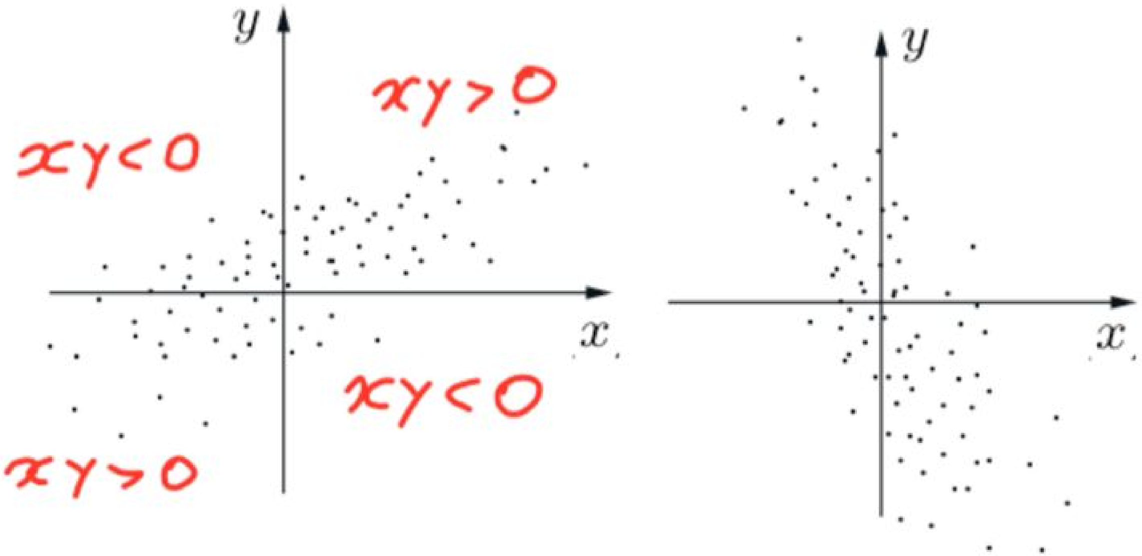
$\mathbf{E}[XY]$



$\mathbf{E}[XY]$

In both diagrams, each point is assumed to be equally likely. So we have a discrete uniform distribution on the discrete set shown. In the left diagram, most of the time, the positive values of X tend to go with the positive values of Y, and negative values of X tend to go with the negative values of Y. In the right diagram negative X-es tend to occur with the positive Ys, and positive Xes with the negative Ys, most of the time.

Thus, under our assumptions, E[XY] must take values as shown



$\mathbf{E}[XY] > 0$

$\mathbf{E}[XY] < 0$

- When X and Y have zero mean, E [X Y] is called the “covariance of X and Y”.

- Covariance of X and Y, denoted **Cov(X, Y)**, can be thought of as the degree to which X and Y covary.

- Let’s consider a general case next.

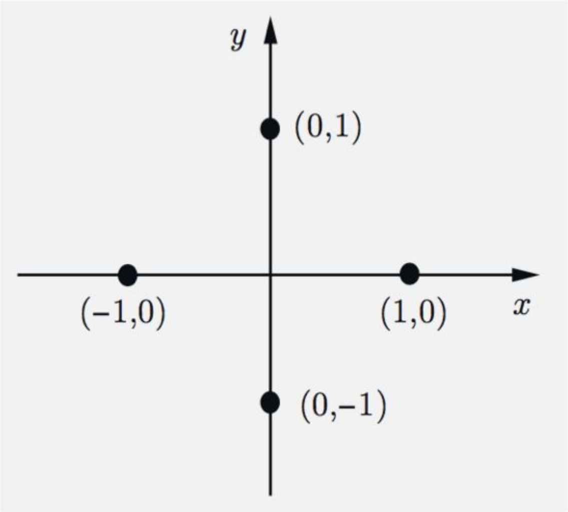
Definition for general case:

$$\boxed{\mathrm{cov}(X, Y) = \mathbf{E}\Big[(X - \mathbf{E}[X]) \cdot (Y - \mathbf{E}[Y])\Big]}$$

Definition for general case:

$$\boxed{\text{cov}(X,Y) = \mathbf{E}\Big[(X - \mathbf{E}[X]) \cdot (Y - \mathbf{E}[Y])\Big]}$$

- independent $\Rightarrow$ $\text{cov}(X,Y) = 0$ (converse is not true)

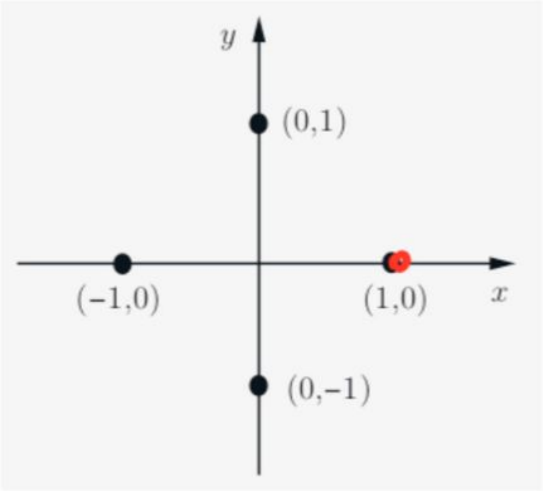


Definition for general case:

$$\boxed{\text{cov}(X,Y) = \mathbf{E}\Big[\underbrace{(X-\mathbf{E}[X])}_{}\cdot\underbrace{(Y-\mathbf{E}[Y])}_{}\Big]}$$

ind $0 = E[(X-E[X])]\,E[Y-E[Y]]$

- independent $\Rightarrow$ $\text{cov}(X,Y)=0$
  (converse is not true)



$XY = 0$

$\text{Cov} = 0$

$X = 1 \Rightarrow Y = 0$

Covariance properties

$$\text{cov}(X,X) = E\left[(X - E[X])^2\right]$$

$$= \text{var}(X) = E[X^2] - (E[X])^2$$

$$\text{cov}(aX+b, Y) =$$

(assume 0 means)

$$= E[(aX+b)Y] = aE[XY] + b\cancel{E[Y]}$$

$$= a \cdot \text{cov}(X,Y)$$

$$\text{cov}(X, Y+Z) = E[X(Y+Z)]$$

$$= E[XY] + E[XZ] = \text{cov}(X,Y) + \text{cov}(X,Z)$$

$$\boxed{\text{cov}(X,Y) = \mathbf{E}\Big[(\underbrace{X}_{} - \underbrace{\mathbf{E}[X]}_{}) \cdot (\underbrace{Y}_{} - \underbrace{\mathbf{E}[Y]}_{})\Big]}$$

$$= E[XY] - E[X E[Y]]$$

$$- E[E[X]Y] + E[E[X]E[Y]]$$

$$= E[XY] - E[X]E[Y]$$

$$\cancel{- E[X]E[Y]} \cancel{+ E[X]E[Y]}$$

$$\boxed{\text{cov}(X,Y) = \mathbf{E}[XY] - \mathbf{E}[X]\mathbf{E}[Y]}$$

The variance of a sum of random variables

$$\mathrm{var}(X_1 + X_2) = E\left[\left(X_1 + X_2 - E[X_1 + X_2]\right)^2\right]$$

$$= E\left[\left(\left(X_1 \underline{-} E[X_1]\right) + \underline{\left(X_2 - E[\right.}X_2]\right)\right)^2\right]$$

$$= E\Big[\left(X_1 - E[X_1]\right)^2 + \left(X_2 - E[X_2]\right)^2$$

$$+ 2\left(X_1 - E[X_1]\right)\left(X_2 - E[X_2]\right)\Big]$$

$$= \mathrm{var}(X_1) + \mathrm{var}(X_2) + 2\,\mathrm{cov}(X_1, X_2).$$

The variance of a sum of random variables

$$\mathrm{var}(X_1 + X_2) = \mathrm{var}(X_1) + \mathrm{var}(X_2) + 2\,\mathrm{cov}(X_1, X_2)$$

$$\mathrm{var}(X_1 + \cdots + X_n) =$$

$$\boxed{\mathrm{var}(X_1 + \cdots + X_n) = \sum_{i=1}^{n} \mathrm{var}(X_i) + \sum_{\{(i,j):\, i \neq j\}} \mathrm{cov}(X_i, X_j)}$$

The variance of a sum of random variables

$$\mathrm{var}(X_1 + X_2) = \mathrm{var}(X_1) + \mathrm{var}(X_2) + 2\,\mathrm{cov}(X_1, X_2)$$

$$\mathrm{var}(X_1 + \cdots + X_n) = E\left[(x_1 + \cdots + x_n)^2\right]$$

(assume 0 means)

$$= E\left[\sum_{i=1}^{n} X_i^2 + \sum_{\substack{i=1,\dots,n \\ j=1,\dots,n \\ i \neq j}} X_i X_j\right]$$

$\Big\}$ $n^2 - n$ terms

$$= \sum_i \mathrm{Var}(X_i) + \sum_{i \neq j} \mathrm{Cov}(X_i, X_j)$$

$$\boxed{\mathrm{var}(X_1 + \cdots + X_n) = \sum_{i=1}^{n} \mathrm{var}(X_i) + \sum_{\{(i,j):\, i \neq j\}} \mathrm{cov}(X_i, X_j)}$$

The Correlation coefficient

- The covariance between two random variables tells us something about the strength of the dependence between them. But it is not so easy to interpret qualitatively. For example, if I tell you that the covariance of X and Y is equal to 5, this does not tell you very much about whether X and Y are closely related or not.

- Another difficulty is that if X and Y are in units, let's say, of meters, then the covariance will have units of meters squared. And this is hard to interpret. A much more informative quantity is the so-called *correlation coefficient, which is a dimensionless version of the covariance.*

The Correlation coefficient

- Dimensionless version of covariance:

$$\boxed{-1 \leq \rho \leq 1}$$

$$\boxed{\begin{aligned} \rho(X,Y) &= \mathbf{E}\left[\frac{(X-\mathbf{E}[X])}{\sigma_X}\cdot\frac{(Y-\mathbf{E}[Y])}{\sigma_Y}\right] \\ &= \frac{\mathrm{cov}(X,Y)}{\sigma_X\sigma_Y} \end{aligned}}$$

- Measure of the degree of "association" between X and Y

A very important property of the correlation coefficient is this:

- It turns out that the correlation coefficient is always between -1 and 1. And this allows us to judge whether a certain correlation coefficient is big or not, because we now have an absolute scale. And so it does provide a measure of the degree to which two random variables are associated.

The Correlation coefficient

- Dimensionless version of covariance:

$$\boxed{-1 \le \rho \le 1}$$

$$\boxed{\begin{aligned} \rho(X,Y) &= \mathbf{E}\left[\frac{(X-\mathbf{E}[X])}{\sigma_X}\cdot\frac{(Y-\mathbf{E}[Y])}{\sigma_Y}\right] \\ &= \frac{\text{cov}(X,Y)}{\sigma_X\sigma_Y} \end{aligned}}$$

- Measure of the degree of "association" between X and Y

- Independent $\Rightarrow$ $\rho = 0$, "uncorrelated" (converse is not true)

- $\rho(X,X) = \dfrac{\text{var}(X)}{\sigma_X^2} = 1$

- $|\rho| = 1 \;\Leftrightarrow\; (X-\mathbf{E}[X]) = c(Y-\mathbf{E}[Y])$ (linearly related)

- $\text{cov}(aX+b,Y) = a\cdot\text{cov}(X,Y) \quad\Rightarrow\quad \rho(aX+b,Y) = \dfrac{a\,\text{cov}(X,Y)}{|a|\,\sigma_X\,\sigma_Y} = \text{sign}(a)\cdot\rho(X,Y)$

Proof of key properties of the correlation coefficient

$$\rho(X,Y) \;=\; \mathbf{E}\left[\frac{(X-\mathbf{E}[X])}{\sigma_X}\cdot\frac{(Y-\mathbf{E}[Y])}{\sigma_Y}\right]$$

$$\boxed{-1 \le \rho \le 1}$$

- Assume, for simplicity, zero means and unit variances, so that $\rho(X,Y)=\mathbf{E}[XY]$

$$0 \le \mathbf{E}\Big[\underline{(X-\rho Y)^2}\Big] = E[X^2] - 2\rho E[XY] + \rho^2 E[Y^2]$$

$$= 1 - 2\rho^2 + \rho^2 = \underline{\underline{1-\rho^2}} \qquad 1-\rho^2 \ge 0 \Rightarrow \rho^2 \le 1$$

If $|\rho| = 1$, then $X = \rho Y \Rightarrow X = Y$ or $X = -Y$

Interpreting the correlation coefficient

$$\rho(X,Y) = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y}$$

- Association does not imply causation or influence

It is important to be able to interpret the correlation coefficient correctly.

A correlation coefficient of let's say 0.5, tells us that something interesting is going on as far as the relation of X and Y is concerned. But what exactly?

It tells us that the two random variables are associated in some sense. But this is often misinterpreted to mean that there is a causal relation between the two. But this is wrong!

A large correlation coefficient in general does not indicate that there is a causal relation between the random variables.

Example: *Let r.v.* **X** *be the amount of ice cream sold at some beach*
Let r.v. **Y** *be the number of shark attacks on that beach*

These variables may indeed be correlated, but if I ate a lot of ice cream, would that cause sharks to attack me? I don't know… but correlation often reflects that there is an underlying common but perhaps hidden factor that affects both of the random variables X and Y. Here the common factor may be that it is summer and that a lot more people are swimming and eating ice cream than in the winter.

Interpreting the correlation coefficient

$$\rho(X,Y) = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y}$$

- Association does not imply causation or influence

  X: math aptitude

  Y: musical ability

- Correlation often reflects underlying, common, hidden factor

  - Assume, Z, V, W are independent

$$\rho(x,Y) = \frac{1}{\sqrt{2}\cdot\sqrt{2}} = \frac{1}{2}$$

  $X = \underline{\underline{Z}} + V$ $Y = \underline{\underline{Z}} + W$

  Assume, for simplicity, that Z, V, W have zero means, unit variances

$$\mathrm{var}(x) = \mathrm{var}(z) + \mathrm{var}(v) = 2 \Rightarrow \sigma_x = \sqrt{2} \qquad \sigma_y = \sqrt{2}$$

$$\mathrm{cov}(x,Y) = E[(Z+V)(Z+W)] = E[Z^2] + E[VZ] + E[ZW] + E[VW]$$

$$= 1 + 0 + 0 + 0$$

Correlations matter...

- A real-estate investment company invests \$10M in each of 10 states. At each state i, the return on its investment is a random variable X_i, with mean 1 and standard deviation 1.3 (in millions).

Correlations matter...

- A real-estate investment company invests $10M in each of 10 states. At each state i, the return on its investment is a random variable X_i, with mean 1 and standard deviation 1.3 (in millions).

$$\mathrm{var}(X_1 + \cdots + X_{10}) = \sum_{i=1}^{10} \mathrm{var}(X_i) + \sum_{\{(i,j):\, i \neq j\}} \mathrm{cov}(X_i, X_j)$$

- If the X_i are uncorrelated, then:

$$\mathrm{var}(X_1 + \cdots + X_{10}) = 10 \cdot (1.3)^2 = 16.9 \qquad \sigma(X_1 + \cdots + X_{10}) = 4.1$$

So, if you look at one state in isolation, it would be a pretty risky investment because the standard deviation is comparable to the mean. It's not an unlikely event to have a return that's one standard deviation below the mean. And if that happens, your return is going to be negative, and you're losing money.

Now, when you diversify, your expected return is equal to 10 and you will only lose money if the outcome is 2 and 1/2 standard deviations below the mean. And that's a fairly unlikely outcome, and so in this situation you feel very confident that you will have a positive profit.

- Suppose, however, that your assumption is wrong, and that actually the different Xi's are correlated with each other (perhaps, because the markets in different states are affected by some global phenomenon that operates on a national level).

- Also suppose that the real estate market in one state is strongly related to the behavior of the market in another state, i.e. that the correlation is pretty high, say, 0.9.

- Let's see what happens to our profits then.

Correlations matter...

- A real-estate investment company invests $10M in each of 10 states. At each state i, the return on its investment is a random variable X_i, with mean 1 and standard deviation 1.3 (in millions).

$$\text{var}(X_1 + \cdots + X_{10}) = \sum_{i=1}^{10} \text{var}(X_i) + \sum_{\{(i,j):\, i \neq j\}} \text{cov}(X_i, X_j)$$

$E[x_1 + \dots + x_{10}] = 10$

- If the X_i are uncorrelated, then:

$\text{var}(X_1 + \cdots + X_{10}) = 10 \cdot (1.3)^2 = 16.9$ $\quad \sigma(X_1 + \cdots + X_{10}) = 4.1$

- If for $i \neq j$, $\rho(X_i, X_j) = 0.9$: $\quad \text{cov}(x_i, x_j) = \rho\, \sigma_{x_i} \sigma_{x_j} = 0.9 \times 1.3 \times 1.3 = 1.52$

$\text{var}(x_1 + \quad + x_{10}) = 10 \cdot (1.3)^2 + 90 \cdot 1.52 = 154$

$\sigma(X_1 + \cdots + X_{10}) = 12.4$

- Now your expected profit is 10, but the standard deviation is 12.4. And if you happen to be one standard deviation below the expectation, which is something that has a sizable probability of occurring, then your profit is going to be negative.

- So, in the uncorrelated case, you're pretty certain that you will have a positive profit, but if the correlations actually turn out to be significant, then you're facing a very risky situation.

- To some extent, this is similar to what happened *during the great financial crisis*. That is, many investment companies thought that they were secure by diversifying and by investing in different housing markets in different states, but then when the economy moved as a whole, it turned out that there were high correlations between the different states, and so the *unthinkable*, that is large losses, actually *did occur*.