

The mathematics of the correlation coefficient are important.

But it is perhaps more important to be able to interpret it 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.

As an example, suppose that X somehow quantifies the mathematical aptitude of a person.

And Y somehow quantifies the musical ability of a person.

In general, it has been found that mathematical aptitude and musical ability are correlated.

People who score high on one will score high on the other as well.

Is there a causal relation?

If you study math a lot and you become very good at math, does it mean that you would become a better musician?

Not necessarily.

Or if you practice the violin day in and day out, does it mean that you will score better in the math exam?

Again, not necessarily.

Perhaps what is going on is that there's a certain feature of the human brain and when that feature is well developed, then that feature helps both in math and in musical ability.

And this is a typical situation of how a correlation coefficient may arise.

That is often a correlation coefficient that's significant, reflects that there is an underlying common but perhaps hidden factor that affects both of the random variables X and Y .

Let's us go through a simple numerical example that models a situation of this kind.

Suppose that Z , V , and w are independent random variables.

And that we have two more random variables defined by these relations.

Not that there's no direct influence from X to Y or from Y to X .

But on the other hand, there's a common underlying factor, this random variable Z that affects both X and Y . Because of this, we expect that X and Y will somehow have some kind of relation or association between them.

And we would like to measure the strength of that association.

The way to measure it will be in terms of the correlation coefficient, which we will now proceed to compute.

To have a complete example in our hands and in order to also keep things simple, let's us assume that the basic underlying random variable Z , V , and W all have 0 means and unit variances.

And now let us take the definition of the correlation coefficient and start calculating.

Let us look at the variance of X . Because X is the sum of two independent random variables, its variance is going to be the sum of those variances.

And we have assumed that each one of those variances is equal to 1.

So the variance of X is equal to 2.

And that implies that the standard deviation of X is equal to the square root of 2.

By a similar argument, the standard deviation of Y is also equal to the square root of 2.

Now, let us look at the covariance between X and Y .

Because X and Y have 0 means, the covariance is just the expected value of the product of the two random variables.

And using the definition of what these two random variables are, it's this particular product here.

We expand the product into a sum of four terms.

And take the expected value of each one of the four terms.

Which leaves us with this particular expression here.

Now, Z has 0 mean and unit variance.

Therefore, the expected value of Z squared is equal to 1.

How about the next term?

V and Z are independent.

So the expected value of the product is the product of the expected values.

But the expected values are zero, so this term is zero.

And with a similar argument, the other terms are zero as well.

So the co-variance is equal to 1.

And from this, we can conclude our calculation and write that the correlation coefficient between X and Y is equal to 1 divided by the square root of 2 times square root of 2, which is $1/2$.

This example also serves to give you a rough idea of what it may mean to have a correlation coefficient of $1/2$.

It means that the two random variables have some common elements.

And they also have some idiosyncratic elements.

And these two elements are roughly equal in weight.

If V and W were completely absent, the correlation coefficient would have been 1.

If on the other hand V and W had a huge variance, so as to completely hide the effect of Z , then the value of the correlation coefficient would have been much, much smaller perhaps closer to 0.

And in the extreme case of course where Z is completely absent, then X and Y are independent, and we get a correlation coefficient of 0.