Esssentials of Applied Data Analysis IPSA-USP Summer School 2018

Covariance and Correlation

Leonardo Sangali Barone
leonardo.barone@usp.br

jan/18

Functions of random variables

We can sum random variables. For example: $Z = X + Y$

What is the expected value of Z?

$$E[Z] = E[X + Y] = E[X] + E[Y]$$

We can multiply random variables. For example: $Z = X * Y$

What is the expected value of Z? It dependes. If X and Y are independent of each other, then

$$E[Z] = E[X * Y] = E[XY] = E[X] * E[Y]$$

What if they are not independent?

Covariance

The covariance of a variable is defined as

$$Cov(X, Y) = E[(X - E[X]) * (Y - E[Y])]$$

But wait, this looks familiar!

$$\begin{gathered} Cov(X, X) = E[(X - E[X]) * (X - E[X])] = \\ = E[(X - E[X])^2 = E[X^2] - (E[X])^2 = Var[X] \end{gathered}$$

Another way to look at the covariance:

$$Cov(X, Y) = E[XY] - E[X] * E[Y]$$

Wait again! Isn't $E[XY]$ the expected value of the multiplication of X and Y? Yes it is, so we can rearrange the formula to obtain:

$$E[XY] = E[X] * E[Y] + Cov(X, Y)$$

If X and Y are independent of each other, then

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

Or we can simply say that

$$Cov(X, Y) = 0$$

So the covariance is a measure of how two variables are related to each other. More elegantly, it is a measure of how two variables vary together (covary). One problem with the covariance is that is can be any number, hence, we can't compare two covariances. But there is a way to "standardize" covariances.

Correlation

The correlation of two variables is defined as:

$$\rho_x y = Corr(X, Y) = \frac{Cov(X, Y)}{\sigma_x * \sigma_y}$$

Where σ_x and σ_y are the standard deviations of X and Y.

One important property of the correlation is that, differently from the covariance, it is bounded:

$$-1 \leq Corr(X, Y) \leq 1$$

As we will see in the future, correlation is a measure of linear relation among two random variables.

Variance of the sum

Finally, when we sum two variables, for example: $Z = X + Y$, the variance of the sum is:

$$Var[Z] = Var[X + Y] = Var[X] + Var[Y] + 2 * Cov(X, Y)$$

Since $Cov(X,Y) = 0$ under indepence, if X and Y are indepent we can simplify the variance to:

$$Var[Z] = Var[X+Y] = Var[X] + Var[Y]$$