Topic #10: Correlation

In probability theory and statistics, correlation, also called correlation coefficient, indicates the strength and direction of a linear relationship between two random variables. In general statistical usage, correlation or co-relation refers to the departure of two variables from independence, although correlation does not imply causation. In this broad sense there are several coefficients, measuring the degree of correlation, adapted to the nature of data.

A number of different coefficients are used for different situations. The best known is the Pearson product-moment correlation coefficient, which is obtained by dividing the covariance of the two variables by the product of their standard deviations. Despite its name, it was first introduced by Francis Galton.

Mathematical properties

The correlation ?X, Y between two random variables X and Y with expected values μX and μY and standard deviations sX and sY is defined as:

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E((X-\mu_X)(Y-\mu_Y))}{\sigma_X \sigma_Y}$$

where E is the expected value of the variable and cov means covariance. Since μX = E(X), sX2 = E(X2) - E2(X) and likewise for Y, we may also write

$$\rho_{X,Y} = \frac{E(XY) - E(X)E(Y)}{\sqrt{E(X^2) - E^2(X)}\,\sqrt{E(Y^2) - E^2(Y)}}$$

The correlation is defined only if both of the standard deviations are finite and both of them are nonzero. It is a corollary of the Cauchy-Schwarz inequality that the correlation cannot exceed 1 in absolute value. The correlation is 1 in the case of an increasing linear

relationship, -1 in the case of a decreasing linear relationship, and some value in between in all other cases, indicating the degree of linear dependence between the variables. The closer the coefficient is to either -1 or 1, the stronger the correlation between the variables.

If the variables are independent then the correlation is 0, but the converse is not true because the correlation coefficient detects only linear dependencies between two variables. Here is an example: Suppose the random variable X is uniformly distributed on the interval from -1 to 1, and Y = X2. Then Y is completely determined by X, so that X and Y are dependent, but their correlation is zero; they are uncorrelated. However, in the special case when X and Y are jointly normal, independence is equivalent to uncorrelatedness.

A correlation between two variables is diluted in the presence of measurement error around estimates of one or both variables, in which case disattenuation provides a more accurate coefficient.