The Correlation Coefficient

The strength of the relationship between two variables that might be observed on a (bivariate) scatter plot can be measured by the *covariance* or *correlation* between the two variables. The covariance is defined as

$$\mathrm{cov}_{xy} = s_{xy}^2 = (1/n)\sum_{i=1}^{n}(x_i - \overline{X})(y_i - \overline{Y}).$$

As is true for the variance, the magnitude of the covariance depends on the units in which the individual variables are measured. Moreover, if the individual variables have unequal variances, then the covariance may reflect that situation more than it does the strength of the relationship between variables.

An alternative measure, the correlation (or correlation coefficient), rescales the variance by dividing it by the product of the standard deviations of each variable, which consequently removes the influence of the unit of measurement or scale and variability of the individual variables. The correlation coefficient is defined as

$$\begin{aligned} r_{xy} &= \frac{\mathrm{cov}_{xy}}{s_x s_y} \\ &= \frac{\sum_{i=1}^{n}(x_i - \overline{X})(y_i - \overline{Y})}{(\sum_{i=1}^{n}(x_i - \overline{X})^2)^{1/2}(\sum_{i=1}^{n}(y_i - \overline{Y})^2)^{1/2}} \end{aligned}$$

The correlation coefficient, r_{xy}, ranges between +1 and -1.

The significance of the correlation coefficient (i.e. the result of a test of the hypothesis that $r_{xy} = 0$ can be judged using the following statistic

$$t = \frac{r(n-2)^{1/2}}{(1-r^2)^{1/2}}$$

which may be compared with a t-distribution with $(n-2)$ degrees of freedom.