2 Variable Statistics
(according to IB)

X and Y are two variables

$x_1, x_2, x_3 \dots x_i \dots x_n$
$y_1, y_2, y_3 \dots y_i \dots y_n$

Σ is summation

$\sum_{i=1}^{n} = x_1 + x_2 + x_3 \dots x_i \dots x_n$

or more simply Σx

$\bar{x}$ x bar is the sample mean of x.

$\bar{x} = \Sigma x / n$

$\bar{y} = \Sigma y / n$

S_{xx} is the variance of x using $\bar{x}$ as the best estimate of the mean μ

$$S_{xx} = \frac{\Sigma (x - \bar{x})^2}{n}$$

$$= \frac{\Sigma x^2}{n} - \frac{2\bar{x}\Sigma x}{n} + \frac{n\bar{x}^2}{n}$$

$$= \frac{\Sigma x^2}{n} - \bar{x}^2$$

We say the mean of the squares minus the square of the means

This form is simpler to calculate and more stable.

The standard deviation

$$S_x = \sqrt{S_{xx}}$$

S_{xy} is the covariance of X and Y. It is a measure of how the two variables change together.
The sign shows the tendency in the linear relationship between the two variables.

$$S_{xy} = \frac{\Sigma (x - \bar{x})(y - \bar{y})}{n}$$

$$= \frac{\Sigma xy}{n} - \frac{\bar{x}\Sigma y}{n} - \frac{\bar{y}\Sigma x}{n} + \bar{x}\bar{y}$$

$$= \frac{\Sigma xy}{n} - \bar{x}\bar{y} - \bar{y}\bar{x} + \bar{x}\bar{y}$$

$$= \frac{\Sigma xy}{n} - \bar{x}\bar{y}$$

Compare this with S_{xx}

The magnitude of the covariance is difficult to interpret directly so it is normalised to produce the correlation coefficient r a dimensionless measure

$$r = \frac{S_{xy}}{S_x S_y}$$

Using the two simplifications already shown and cancelling the ns we get

$$r = \frac{\Sigma xy - n\bar{x}\bar{y}}{\sqrt{\Sigma x^2 - n\bar{x}^2}\sqrt{\Sigma y^2 - n\bar{y}^2}}$$

The correlation coefficient is a measure of the degree of linearity between the two variables X and Y.

Now if we draw a line of best fit between the plots the equation is given by

$$y = mx + c$$

Now the gradient is given by

$$m = \frac{S_{xy}}{S_{xx}}$$

strictly this is y on x

and the line passes through the point $(\bar{x}, \bar{y})$

So we have

$$(y - \bar{y}) = m(x - \bar{x})$$

rearranging we determine

$$c = \bar{y} - m\bar{x}$$

$$\text{or } \bar{y} - \frac{S_{xy}}{S_{xx}}\bar{x}$$

and is the y-intercept of the line of best fit.

The coefficient of determination r^2 is a measure of the strength of association between the two variables X and Y. For linear regression it is the proportion explained by the model.

$0 \leq r^2 < 0.25$ very weak
$0.25 \leq r^2 < 0.5$ weak
$0.5 \leq r^2 < 0.75$ moderate
$0.75 \leq r^2 < 0.9$ strong
$0.9 \leq r^2 < 1.0$ very strong
$r^2 = 1$ is perfect.

For linear correlation we have

$$r^2 = r^2$$

but be careful

↓ this is the coefficient of determination and is a thing all in its own right

this is the correlation coefficient squared.

Not a lot of people realise this!

rg ∞

Linear Regression

$$y - \bar{y} = \frac{S_{xy}}{S_{x^2}}(x - \bar{x})$$