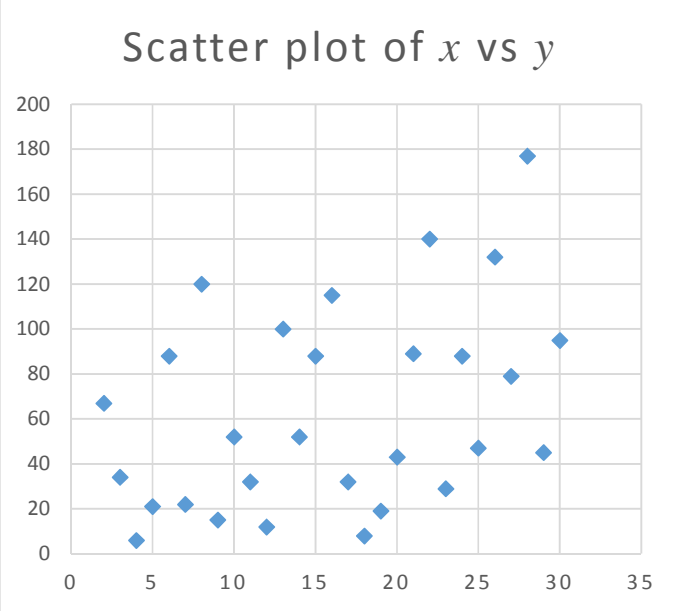
The Pearson Correlation Coefficient

Scatterplots are an important visual tool for examining the relationship between pairs of variables, such as x's and y's. We can do a statistical estimate of their relationships and a test of significance for them.

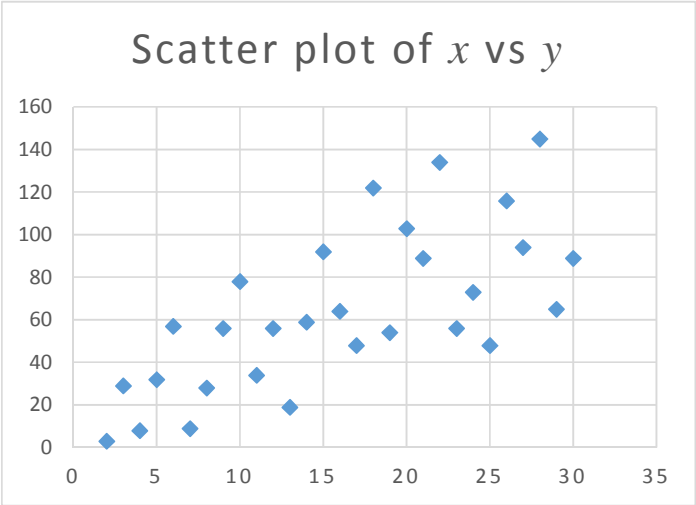
For two variables measured on interval or ratio level, such as a series of concentrations of reference standard solution (x's) and their respective instrument responses (y's), the most common measure of association is the Pearson correlation coefficient, also called the product-moment correlation coefficient, written as ρ (the Greek letter rho) for a population and r for a sample.

Pearson's r has a range of (−1, 1), with 0 indicating no relationship between these two variables and the larger absolute values indicating a stronger relationship between the variables (assuming neither variable is a constant).

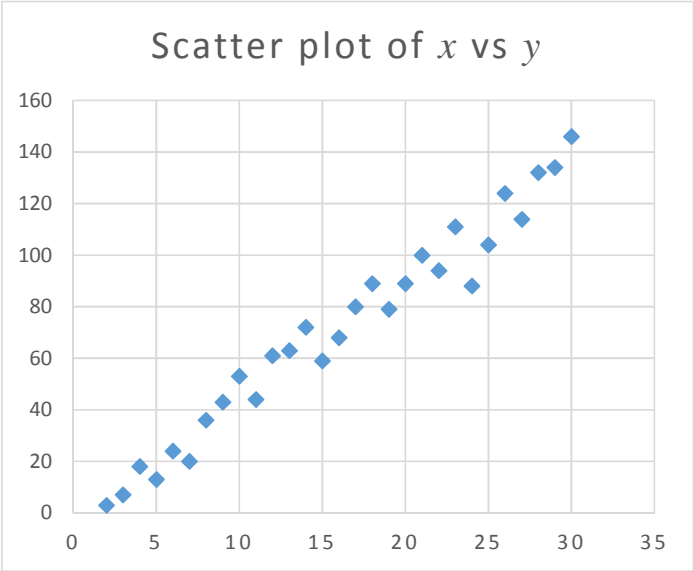
The following figures shows examples of various data displays with different correlation coefficients, r :



$r = 0.400$



$r = 0.702$



$r = 0.983$

In plotting an instrumental calibration curve between concentrations of standard solutions versus instrument responses, a high Pearson correlation coefficient of >0.99 is normally desired.

Although correlation coefficients are often calculated using computer software easily, we should learn how to calculate them based on the first principles. The formula for the Pearson correlation coefficient involving sums of squares (SS) is given below:

$$r = \frac{SS_{xy}}{\sqrt{SS_x SS_y}} \qquad [1]$$

where

$$SS_x = \sum_{i=1}^{n} (x_i - \bar{x})^2 \quad [2]$$

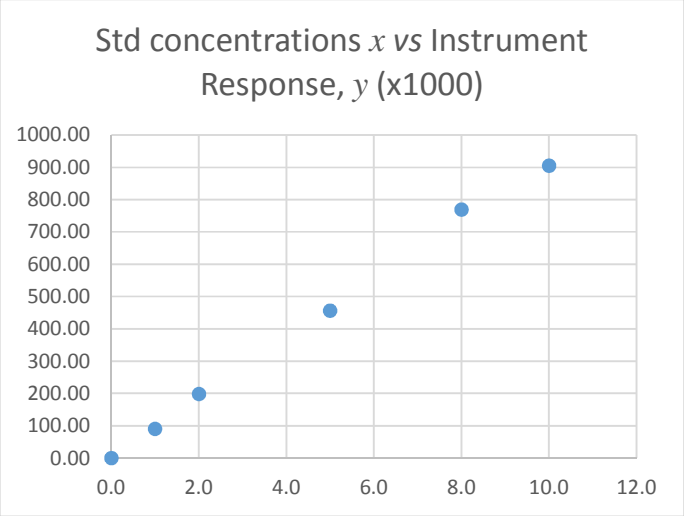
$$SS_y = \sum_{i=1}^{n} (y_i - \bar{y})^2 \quad [3]$$

$$SS_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \quad [4]$$

The use of these formulae become clearer after working through an example. Suppose we had a series of chromium standard solutions prepared for a calibration curve using an ICPOES instrument. The instrument intensity responses against the respective concentrations were tabulated as below:

Std Conc, x mg/L	Intensity Response, y (x1000)
0.0	0.08
1.0	90.10
2.0	198.55
5.0	456.91
8.0	768.90
10.0	905.66

The scatter plot of the above data is shown in the following figure:



We can then use a spreadsheet to calculate the Pearson correlation coefficient of this calibration curve easily as shown below:

	x	y(x10³)	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
	0.0	0.08	-4.333	-403.29	18.778	162642.25	1747.59
	1.0	90.10	-3.333	-313.27	11.111	98136.14	1044.22
	2.0	198.55	-2.333	-204.82	5.444	41950.16	477.91
	5.0	456.91	0.667	53.54	0.444	2866.68	35.69
	8.0	768.90	3.667	365.54	13.444	133617.16	1340.30
	10.0	905.66	5.667	502.30	32.111	252300.59	2846.34
Mean =	4.3	403.37					
				SS =	81.333	691512.980	7492.053

By using equation [1], we have r = 0.999
The above correlation coefficient result shows a strong positive relationship between the concentrations of chromium solution and the ICP intensity responses.

It may now be noted that although the correlation coefficient indicates the strength and direction of the linear relationship between two variables, we might also want to know how much of the variation in one variable can be accounted for by the other variable. To find this, we can calculate the *coefficient of determination*, which is simply r^2. In the above example,
$r^2 = 0.999^2 = 0.998$