Topic 3.Correlation

4. The relationship bettween two data

In engineering analysis, it is often desired to fit a trend line or curve to a set of x-y data. Consider a set of n measurements of some variable y as a function of another variable x.

Typically, y is some measured output as a function of some known input, x.
In general, in such a set of measurements, there may be:

- Some scatter (precision error or random error).
- A trend – in spite of the scatter, y may show an overall increase with x, or perhaps an overall decrease with x.

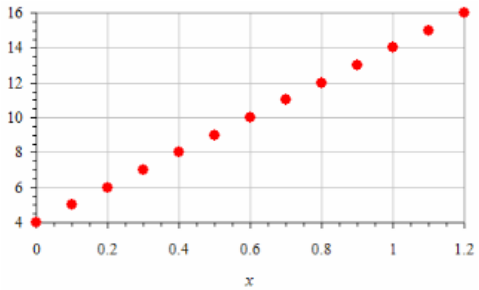
The linear correlation coefficient is used to determine if there is a trend. If there is a trend, regression analysis is used to find an equation for y as a function of x that provides the best fit to the data.

A. Correlation

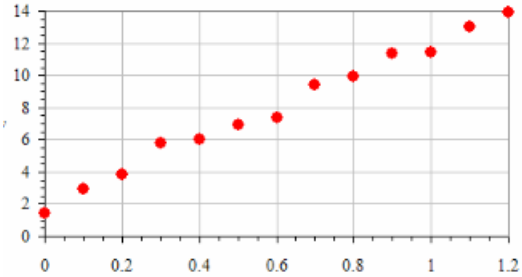
Correlation quantifies the degree to which two variables are related. Correlation does not fit a line through the data points. You simply are computing a correlation coefficient (r) that tells you how much one variable tends to change when the other one does. When r is 0.0, there is no relationship. When r is positive, there is a trend that one variable goes up as the other one goes up. When r is negative, there is a trend that one variable goes up as the other one goes down.
Linear correlation coefficient is defined as:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

In the equation above, the mean value of x and the mean value of y are defined in the usual manner.
If *rxy* = 1, it means that *y increases* with *x* in a *perfectly linear fashion*, with *no scatter*:

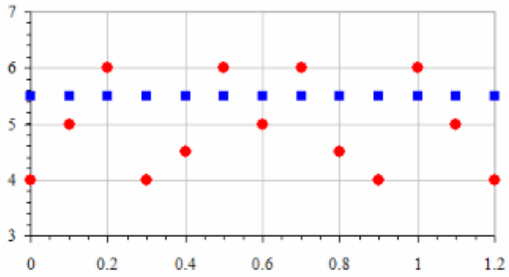


If **0 < r_{xy} < 1**, it means that in general, *y* ***increases*** with *x*, but with ***some scatter***. Here, *r* = 0.995, as calculated from the set of data points plotted below. The closer r_{xy} is to one, the less scatter in the data.

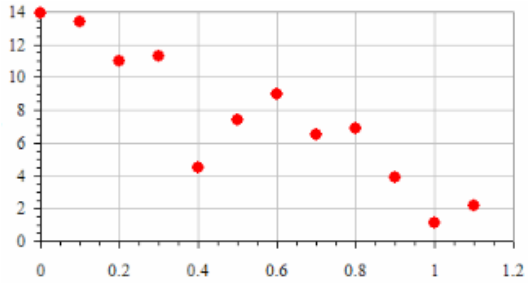


If ***rxy* = 0**, it means that *y* is ***uncorrelated*** with *x*, and there is ***no trend***. There may or may not be scatter, as illustrated below. Both sets of data have zero correlation, even though the circles have lots of scatter.

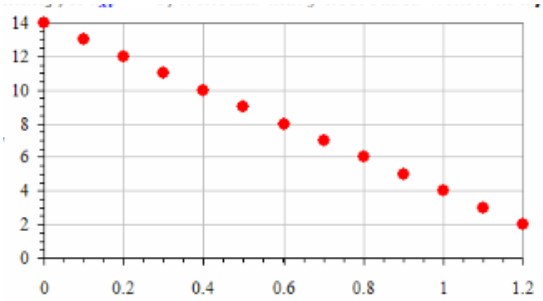
Topic 3.Correlation



If **-1 < *rxy* < 0**, it means that in general, *y* ***decreases*** with *x*, but with ***some scatter***. Here, *r* = -0.924, ascalculated from the set of data points plotted below.



Finally, if ***rxy* = -1**, it means that *y* ***decreases*** with *x* in a ***perfectly linear fashion***, with ***no scatter***:



The critical value of rxy

How large does r_{xy} need to be so that the observed trend is real, and not just pure chance? it depends on two parameters:

- The desired **confidence level**.
- The **number of data pairs**.

r_t can be defined as the critical value of r_{xy} . r_t is generally given in tables as a function of n (number of data pairs) and c (confidence level) or α (significance level)

By definition, if the actual r_{xy} (the linear correlation coefficient) is larger than r_t (the critical value), we areconfident (to some confidence level) that the trend is real.On the other hand, if the actual rxy is smaller than rt , we cannot be confident (again to some confidence level) that the trend is real.

Topic 3.Correlation

Values of r_t (critical values) for linear correlation coefficient								
$c = \rightarrow$	80%	90%	92.5%	95%	97%	98%	99%	99.5%
$\alpha = \rightarrow$	0.2	0.1	0.075	0.05	0.03	0.02	0.01	0.005
$n = \downarrow$								
3	0.95106	0.98769	0.99307	0.99692	0.99889	0.99951	0.99988	0.99997
4	0.80000	0.90000	0.92500	0.95000	0.97000	0.98000	0.99000	0.99500
5	0.68705	0.80538	0.83994	0.87834	0.91377	0.93433	0.95874	0.97404
6	0.60840	0.72930	0.76718	0.81140	0.85503	0.88219	0.91720	0.94170
7	0.55086	0.66944	0.70809	0.75449	0.80206	0.83287	0.87453	0.90556
8	0.50673	0.62149	0.65985	0.70673	0.75599	0.78872	0.83434	0.86974
9	0.47159	0.58221	0.61982	0.66638	0.71613	0.74978	0.79768	0.83591
10	0.44280	0.54936	0.58606	0.63190	0.68148	0.71546	0.76459	0.80461
11	0.41866	0.52140	0.55713	0.60207	0.65114	0.68510	0.73479	0.77589
12	0.39806	0.49726	0.53202	0.57598	0.62434	0.65807	0.70789	0.74961
13	0.38022	0.47616	0.50998	0.55294	0.60049	0.63386	0.68353	0.72553
14	0.36456	0.45750	0.49043	0.53241	0.57911	0.61205	0.66138	0.70344
15	0.35069	0.44086	0.47295	0.51398	0.55980	0.59227	0.64114	0.68311
16	0.33828	0.42590	0.45719	0.49731	0.54227	0.57425	0.62259	0.66434
17	0.32710	0.41236	0.44290	0.48215	0.52627	0.55774	0.60551	0.64696
18	0.31696	0.40003	0.42986	0.46828	0.51158	0.54255	0.58971	0.63083
19	0.30770	0.38873	0.41791	0.45553	0.49804	0.52852	0.57507	0.61580
20	0.29921	0.37834	0.40689	0.44376	0.48551	0.51550	0.56144	0.60176
22	0.28414	0.35983	0.38723	0.42271	0.46303	0.49209	0.53680	0.57627
24	0.27114	0.34378	0.37016	0.40439	0.44338	0.47158	0.51510	0.55370
26	0.25977	0.32970	0.35516	0.38824	0.42603	0.45341	0.49581	0.53355
28	0.24972	0.31722	0.34184	0.37389	0.41055	0.43718	0.47851	0.51542
30	0.24075	0.30606	0.32991	0.36101	0.39664	0.42257	0.46289	0.49900

B. Fitting a treand line

In engineering analysis, it is often desired to fit a trend line or curve to a set of x-y data. Consider a set of n measurements of some variable y as a function of another variable x. If the correlation analysis gives satisfactory results, this means that there is a trend and a regression analysis is useful. Regression analysis is used to find an equation for y as a function of x that provides the best fit to the data.

Linear regression analysis is also called linear least-squares fit analysis. The goal of linear regression analysis is to find the "best fit" straight line through a set of y vs. x data.
An equation for a straight line that attempts to fit the data pairs is chosen as

$$Y = ax + b$$

In the above equation, *a* is the ***slope*** (*a* = *dy*/*dx* – most of us are more familiar with the symbol *m* rather than *a* for the slope of a line), and *b* is the ***y-intercept*** – the *y* location where the line crosses the *y* axis (in other words, the value of *Y* at *x* = 0). **An upper case *Y*** is used for the fitted line to distinguish the fitted data from the *actual* data values, *y*.
coefficients a and b are optimized for the best possible fit to the data.
For each data pair, error can be defined as follows:

$$e_i = Y_i - y_i = ax_i + b - y_i$$

ei is also called residual.
Define ***E*** as the ***sum of the squared errors*** of the fit – a global measure of the error associated with

$$E = \sum_{i=1}^{i=n} e_i^2 = \sum_{i=1}^{i=n} (ax_i + b - y_i)^2$$

It is now assumed that the best fit is the one for which E is the smallest. In other words, coefficients a and b that minimize E need to be found. These coefficients are the ones that create the best-fit straight line Y = ax + b.
How can a and b be found such that E is minimized? Well, as any good engineer knows, to find a minimum (or maximum) of a quantity, that quantity is differentiated, and the derivative is set to zero. Here, two partial derivatives are required, since E is a function of two variables, a and b. Therefore;

Topic 3.Correlation

$$\frac{\partial E}{\partial a}=0 \text{ and } \frac{\partial E}{\partial b}=0$$

By solving these two equations simultaneously, we obtain a and b with given formulas:

$$a=\frac{n\sum_{i=1}^{i=n}x_i y_i-\left(\sum_{i=1}^{i=n}x_i\right)\left(\sum_{i=1}^{i=n}y_i\right)}{n\sum_{i=1}^{i=n}x_i^2-\left(\sum_{i=1}^{i=n}x_i\right)^2} \text{ and } b=\frac{\left(\sum_{i=1}^{i=n}x_i^2\right)\left(\sum_{i=1}^{i=n}y_i\right)-\left(\sum_{i=1}^{i=n}x_i\right)\left(\sum_{i=1}^{i=n}x_i y_i\right)}{n\sum_{i=1}^{i=n}x_i^2-\left(\sum_{i=1}^{i=n}x_i\right)^2}$$