Correlation and Trends

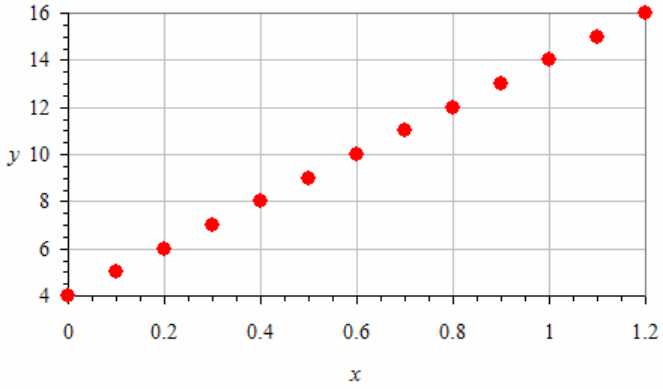
Author: John M. Cimbala, Penn State University
Latest revision: 28 January 2010

Introduction

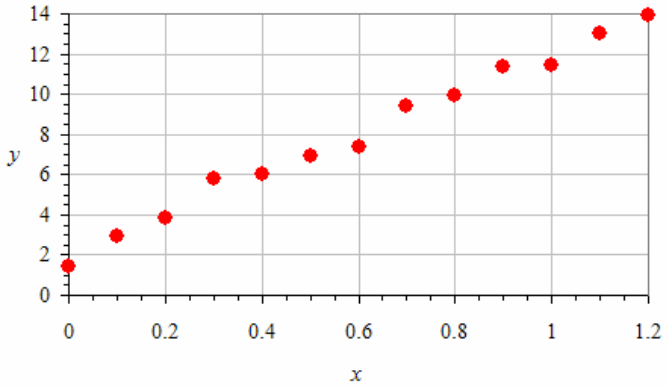
- In engineering analysis, we often want to fit a trend line or curve to a set of *x*-*y* data.
- Consider a set of *n* measurements of some variable *y* as a function of another variable *x*.
- Typically, *y* is some measured ***output*** as a function of some known ***input***, *x*.
- In general, in 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.

Correlation coefficient

- The ***linear correlation coefficient*** r_{xy} is defined as $$r_{xy} = \frac{\sum_{i=1}^{i=n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{i=n}(x_i - \overline{x})^2 \sum_{i=1}^{i=n}(y_i - \overline{y})^2}}.$$
- In the equation above, the ***mean value of x*** and the ***mean value of y*** are defined in the usual manner as $\overline{x} = \frac{1}{n}\sum_{i=1}^{i=n} x_i$ and $\overline{y} = \frac{1}{n}\sum_{i=1}^{i=n} y_i$.
- Some observations about the linear correlation coefficient are worth noting:
  - By definition, r_{xy} must always lie between –1 and 1, i.e., $-1 \le r_{xy} \le 1$.
  - The linear correlation coefficient is always ***nondimensional***, regardless of the dimensions of *x* and *y*.
  - If $\mathbf{r_{xy} = 1}$, it means that *y* ***increases*** with *x* in a ***perfectly linear fashion***, with ***no scatter***:

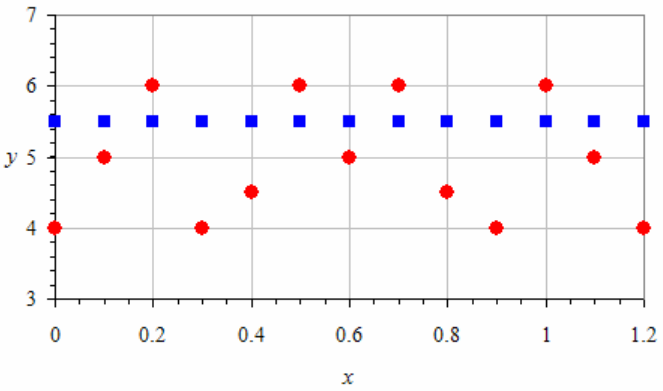


  - If $\mathbf{0 < r_{xy} < 1}$, it means that in general, *y* ***increases*** with *x*, but with ***some scatter***. Here, r_{xy} = 0.995, as calculated from the set of data points plotted below. The closer r_{xy} is to one, the less scatter in the data.

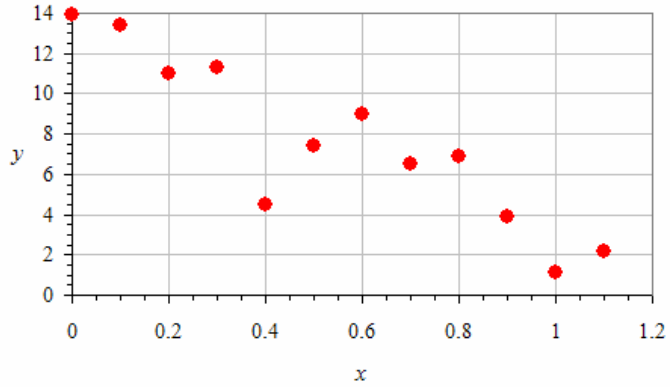


  - If $\mathbf{r_{xy} = 0}$, it means that *y* is ***uncorrelated*** with *x*, and there is ***no trend***. There may or may not be scatter, as illustrated below. Both sets of data have zero correlation, even though the circles have lots of scatter.

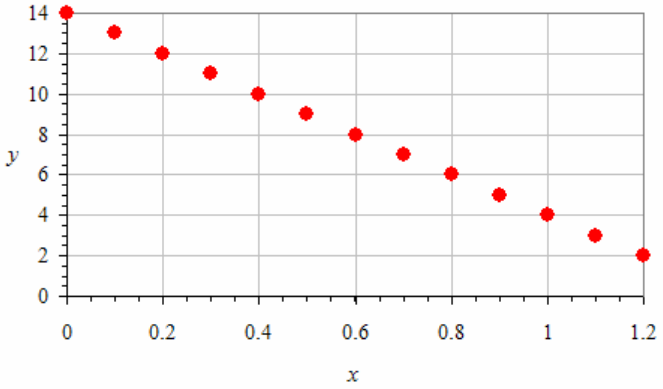
If there is scatter, the scatter must be ***symmetric*** (circles) in order for r_{xy} to equal *exactly* zero. If there is no scatter (squares), r_{xy} cannot be calculated with the above equation, since there will be a zero in the denominator. However, r_{xy} is zero for perfectly flat data (the squares below). It is extremely unlikely for r_{xy} to ever be exactly zero in an actual measurement.



- If **$-1 < r_{xy} < 0$**, it means that in general, *y* ***decreases*** with *x*, but with ***some scatter***. Here, r_{xy} = -0.924, as calculated from the set of data points plotted below.



- Finally, if **$r_{xy} = -1$**, it means that *y* ***decreases*** with *x* in a ***perfectly linear fashion***, with ***no scatter***:



- In Excel, the linear correlation coefficient can be found easily:
  - Excel 2003: Tools-Data Analysis-Correlation.
  - Excel 2007: Click the Data tab; in the area called *Analysis*, Data Analysis-Correlation.
  - Select both columns of data (*x* and *y*) as the *Input Range*.
  - Select a cell for the top-left corner of the output as the *Output Range*, and OK.
  - The value shown in the lower left corner of the resulting table is the linear correlation coefficient, r_{xy}.

The critical value of r_{xy}

- How large does r_{xy} need to be so that the observed trend is *real*, and not just pure chance?
- The answer is that it depends on two parameters:
  - The desired confidence level.
  - The number of data pairs.

- Let r_t be defined as the ***critical value of*** r_{xy}. It is evaluated as $r_t = \sqrt{\dfrac{k(t_{\alpha/2})^2}{k(t_{\alpha/2})^2 + \text{df}}}$, where
  - k is the number of independent variables used in the correlation. For linear correlation with only one independent variable, $k = 1$. [This is the case we are discussing here, namely, y as a function of x.]
  - $t_{\alpha/2}$ is the critical t-test value – a function of level of significance and df as defined previously. $t_{\alpha/2}$ is obtained from a table as previously. Or, if using Excel, $t_{\alpha/2}$ = TINV(α,df) as previously.
  - df is the degrees of freedom. In this case, $\text{df} = n - k - 1$, where n is the number of data pairs.
  - **Example**: For n = 24 (24 data pairs), α = 0.05 (95% confidence), and k = 1 (one independent variable x), we calculate df = 24 – 1 – 1 = 22, $t_{\alpha/2}$ = TINV(0.05,22) = 2.073873, and
    $r_t = \sqrt{\dfrac{1(2.073873)^2}{1(2.073873)^2 + 22}} = 0.404386$ which we round to 5 significant digits as r_t = 0.40439.
- We find a table listing r_t as a function of n (number of data pairs) and c (confidence level) or α (significance level) in many statistics books. Recall, α = significance level = 1 – c, as previously defined. A brief table of r_t is provided below for the case with k = 1. A longer table is provided on the Exams tab on the website.

Values of r_t (critical values) for linear correlation coefficient

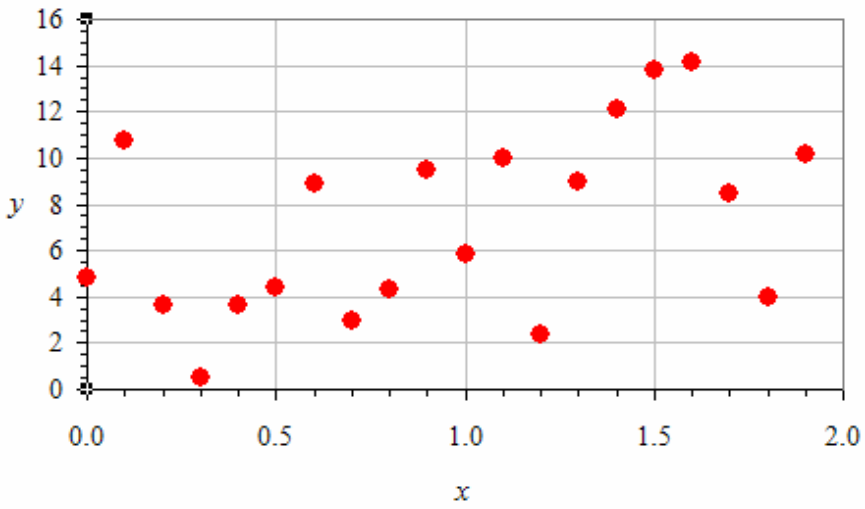
c = →	80%	90%	92.5%	95%	97%	98%	99%	99.5%
α = →	0.2	0.1	0.075	0.05	0.03	0.02	0.01	0.005
n = ↓								
3	0.95106	0.98769	0.99307	0.99692	0.99889	0.99951	0.99988	0.99997
4	0.80000	0.90000	0.92500	0.95000	0.97000	0.98000	0.99000	0.99500
5	0.68705	0.80538	0.83994	0.87834	0.91377	0.93433	0.95874	0.97404
6	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

- In standard engineering practice, 95% confidence level is used. Thus, c = 0.95 and α = 1 – 0.95 = 0.05. The 95% confidence column of the r_t table is therefore used for standard engineering confidence, and is highlighted for convenience.

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

- **Example**:
  Given: 20 data pairs (y vs. x) are available, as listed below, along with a scatter plot of the data.

x	y
0.0	4.800
0.1	10.729
0.2	3.600
0.3	0.500
0.4	3.600
0.5	4.400
0.6	8.900
0.7	3.000
0.8	4.300
0.9	9.500
1.0	5.800
1.1	10.000
1.2	2.400
1.3	9.000
1.4	12.100
1.5	13.800
1.6	14.100
1.7	8.500
1.8	4.000
1.9	10.200



To do: Determine if there is a significant trend between y and x.
Solution:

- *By eye*, it seems that there *may* be a trend that y increases with x, but there is too much scatter to tell for sure. The linear correlation coefficient is probably small. It is not obvious from the plot whether the trend is real or simply a result of chance. ***By eye, we cannot be confident that a trend exists***.
- The linear correlation coefficient is calculated from the above equation, or using Excel's built-in Correlation macro: $\boldsymbol{r_{xy} = 0.480}$.
- The critical correlation table is used at $n = 20$ and $\alpha = 1 - 0.95 = 0.05$ for the standard engineering confidence level of 95%. From the table, the critical value is $r_t = 0.44376$.
- r_t is compared to the actual value of r_{xy}. Since $r_{xy} > r_t$ (0.480 > 0.44376), the conclusion is:
  Yes, there is a trend between y and x, to the standard engineering confidence level of 95%.
- Suppose, however, that 98% confidence level is required. Then, at $n = 20$ and 98% confidence, the table yields $r_t = 0.51550$.
- r_t is compared to the actual value of r_{xy}. Since $r_{xy} < r_t$ (0.480 < 0.51550) the conclusion is:
  No, there is not a trend between y and x, to the more stringent confidence level of 98%.
- If we want to give a more accurate confidence level, we can use the equation given above, or interpolate in the table. For 20 points, we interpolate for $r_{xy} = 0.480$ between 0.44376 at 95% confidence and 0.48551 at 97% confidence. The result is a confidence level of 96.7%. Our final conclusion:
  ***There is a trend between y and x, to a confidence level of 96.7%* (to three digits)**.

Discussion: In cases like this, we cannot judge *by eye* whether there is a trend or not, so the mathematical procedure discussed here is useful, and eliminates the possibility of bias on the part of a researcher who may be trying to show that a trend exists. In the present case, he or she is 96.7% confident that a trend exists. ***This researcher should report that there is a trend in the data to a confidence level of* 96.7%**.