Method Sheet 109

Basic correlation analysis using Microsoft Excel

Overview

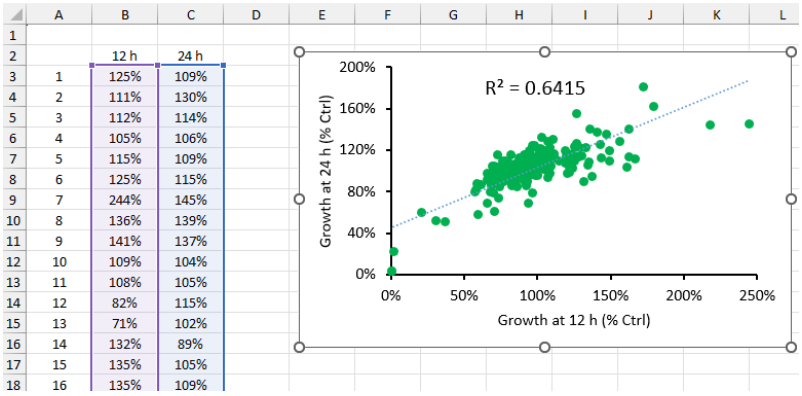
This method sheet explains how to perform basic correlation analysis to seek evidence of potential associations between two distinct variables. This approach is helpful when exploring whether any relationships may exist between the physicochemical properties of hit compounds (such as molecular weight, lipophilicity, or surface area) and their potential potency as drug leads. Calculating a correlation coefficient helps to identify whether any such relationship exists for a specific physical characteristic. This type of analysis also helps move your dissertation beyond simple observation and toward an understanding of the underlying "structure-activity relationship" (SAR), which is a key component of successful drug discovery.

A key point to remember in conducting these analyses is that correlation is not the same as causation. Even if you find a strong correlation with a highly significant p-value for a particular relationship, it does not mean that the physical characteristic you have found *causes* the change in bioactivity. This is a limitation that you should acknowledge in your report or dissertation when reporting the results of correlation analyses.

The following steps will show how to calculate the **Pearson "r" value** for a given pair-wise comparison. An "r" value of +1 or -1 represents a completely linear relationship, and 0 represents no relationship whatsoever.

Method

1) Create a data table in a fresh Excel worksheet in which your bioassay data values (e.g. bacterial growth, tumour cell viability etc.) are in a single column in the same order as the Puretitre compound IDs (i.e. from compound 1 in the first row, through to compound 200 in the last row - the analysis will not work if this is not in place).
2) Just to the right of this column, insert a second set of values which are also ordered strictly in exactly the same order of compound 1 at the start to compound 200 at the end of the column.
3) Highlight all of the data in both of the columns, then insert a 'Scatter plot'.
4) Click on the large 'Chart Design' button at the top ribbon, then 'Add Chart Element' then 'Trendline' then 'More trendline options'.
5) In the 'Format Trendline' menu that appears at the right of the screen, click on 'Display R-squared value on chart', you should now see something like this:



	A	B	C
1			
2		12 h	24 h
3	1	125%	109%
4	2	111%	130%
5	3	112%	114%
6	4	105%	106%
7	5	115%	109%
8	6	125%	115%
9	7	244%	145%
10	8	136%	139%
11	9	141%	137%
12	10	109%	104%
13	11	108%	105%
14	12	82%	115%
15	13	71%	102%
16	14	132%	89%
17	15	135%	105%
18	16	135%	109%

6) In this example, we are comparing the growth of *E. coli* at 12 hours with the growth at 24 hours in response to each of the 200 *Puretitre* compounds.

7) If you find a significant correlation in your later analysis, it may be helpful to show this scatter plot in your dissertation, since the closer the dots align with the trendline, the stronger the relationship is likely to be.
8) Take a note of the R^2 value shown on the chart - this is a measure of how strongly correlated the two variables are.
9) Now we must find out if this degree of correlation is statistically significant for the number of samples we have - this requires first calculating the 't-statistic'.
10) Prepare a table below your chart with the following headings: R2, r, n, t, P-value.
11) Below the R2 heading, type the R^2 value shown on the scatter plot - we will assume for this example that this value will be in cell `E21`.
12) In the cell just to the right of that, calculate the square root of this value with:

```
=SQRT(E21)
```

13) In this example, it gives a result of `0.801`.
14) In the cell to the right of that value (below the 'n' heading), type the number of different observations there are for each column, in this case it is 200 because 200 different compounds were rested.
15) In the cell below the 'T' heading, type the following formula:

```
=F21*SQRT((G21-2)/(1-F21^2))
```

16) This calculates the T-statistic for the specific 'r' and 'n' values you have just calculated in the table.
17) Remember to adjust the cell references in this formula to ensure they match with the cell locations of your own data.
18) Finally, we can convert the T-statistic into a p-value, using the following equation:

```
=T.DIST.2T(H21,G21-2)
```

19) Your data table should now look like this:

	E	F	G	H	I
19					
20	R2	r	n	T	P-value
21	0.6415	0.801	200	18.82	0.000

20) In this example, the p-value is extremely small (below 0.001) this means a very significant correlation exists between the two measurements.
21) Tip: After you have created this table once, you can use it again and again to calculate new p-values simply by typing the relevant R^2 value into the first cell of the second row.

Calculating the Pearson r coefficient without charting a scatter plot

1) Excel offers a shortcut to finding the Pearson ‘r’ value without having to prepare a scatter plot and insert a trendline.
2) Arrange your data in two columns, ensuring they are aligned by correct compound ID, as shown above.
3) In an empty cell, type the following formula:

```
=CORREL(B3:B202,C3:C202)
```

4) This formula assumes your first column data (without the heading) are in cells B3 to B202, and the second column data are in cells C3 to C202 - remember to change the cell references to match where the cells are in the first and second data columns of your own worksheet.
5) The value given is the r value, so you don’t have to perform a square root operation.
6) Instead, you can place this formula in the second position of the table above, and the rest of the table will calculate the p-value automatically.

Interpreting the p-value

1) We typically assume that if a p-value is lower than 0.05, the results are “statistically significant”.
2) In this example, it would mean that there is a significant correlation between the two variables.
3) Bear in mind, however, that if multiple correlation tests have been performed, you will have to adjust the threshold of your p-values accordingly to account for this (see Method Sheet 111 for how to do this).

Notes

- The Pearson correlation test requires that the data are normally distributed and have a linear relationship between one variable and the other.
- If it seems likely that your data are not normally distributed, you should perform the Spearman’s Rank Correlation test, which does not require the data to be normally distributed.
- If a compound has a value in column A but is missing a value in column B, you must delete that entire row, as Excel's CORREL function cannot handle empty cells.
- In addition to a significant p-value, in drug discovery, we often look also for 'r' values above 0.7, since these suggest a strong, meaningful relationship between a physical property and biological activity.