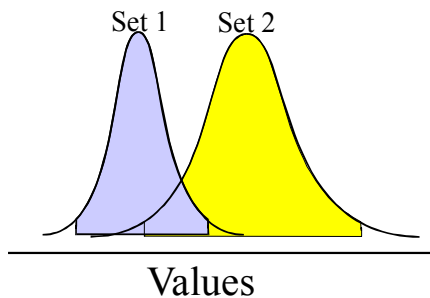


Statistics and Data Analysis

In this guide I will make use of Microsoft Excel in the examples and explanations. This should not be taken as an endorsement of Microsoft or its products. In fact, there are several other spreadsheet packages available, some are better suited to scientific calculations than Excel. Most of you will have access to Excel at home and in the labs. If you need an inexpensive spreadsheet program (actually a whole office suite) you can try StarOffice produced by Sun Microsystems (http://www.sun.com/software/product_family/staroffice.html). I haven't tried StarOffice but the price is about \$70. There is probably an academic discount that would make it even less expensive. Many of the commands won't be identical to Excel but they should be quite close.

In this course we will always use what are known as the double-sided probability tables, or two-tailed tables (T-table, F-table etc.). Probabilities are calculated by integrating the area under the Normal (Gaussian) curve where the total area under the curve has been normalized to 1. For example if the integration is from -1σ through to $+1\sigma$ the integrated area is 0.683 and we have a 68.3% probability that any single measurement will fall into the region $\mu \pm 1\sigma$. This also means that there is $1-0.683 = 0.317$ of the area outside of the integration limits. Half of this area will be above the upper limit and half will be below the lower limit of integration. These two little zones are known as the tails of the Gaussian curve.



Most of the t-tests that we will perform are asking the basic question “is there a significant overlap between one set of data and the other?” This is illustrated in the figure to the left where the overlap is obvious. You can also see that where the two curves meet the upper tail (right-hand tail) of Set 1 and the lower tail of Set 2 have not been included in the diagram but clearly they are involved in the overlapping region! By using the two-tailed probability tables for this test

we will discard these two small pieces that are in the overlapping region. Most often this won't matter (because the total area in these little tails is so small) but you should be aware that it is more appropriate to use the single tail probability tables in this case. In other cases it is appropriate to use the two-tailed probability tables. The tests are the same in both cases and the results will be largely the same except the level of confidence will change. So... **WHILE YOU ARE LEARNING TO USE STATISTICAL TESTS IN THIS CLASS IT IS ACCEPTABLE TO ALWAYS USE THE TWO-TAILED TABLES.** In the future, you should be careful and consult a good statistics text to verify that you are using the correct probability tables.

Basic Statistical Concepts

Mean Value

The mean value is the value that we EXPECT all of our data to equal. For large data sets we can calculate the mean value. For smaller data sets we can only estimate the mean value. The mean value is calculated by:

$$\mu = \sum_{i=0}^{N \rightarrow \infty} \frac{x_i}{N} \quad \text{Equation 1}$$

For a smaller data set we can calculate the average value and we use the same formula except that we use the symbol \bar{x} rather than μ and that N is a finite number. In practice \bar{x} approaches μ as N reaches 20-30.

$$\bar{x} = \sum_{i=1}^N \frac{x_i}{N} \quad \text{Equation 2}$$

Standard Deviation

The standard deviation, or the standard error, is a convenient number that measures how far away from the expected value any individual measurement is likely to be. The larger the value of the standard deviation is the larger the “scatter” in the data set. For large sets of data it is possible to calculate the standard deviation of the population. Another term that is often used in statistical calculations is the variance that is simply the standard deviation squared.

$$\sigma = \sqrt{\lim_{N \rightarrow \infty} \frac{\sum_{i=0}^N (x_i - \mu)^2}{N}} \quad \text{Equation 3}$$

The explanation of this equation is quite simple. It is the sum of all of the differences (between the individual measurements and the expected value) divided by the degrees of freedom (N). Effectively, it is *similar* to the average difference (though, if you examine the equation carefully you will see it is NOT the average difference). When dealing with a real data set some of the differences will be positive, some will be negative, they will sum to zero. To correct for this problem the differences, or errors, are squared, summed and then the root is taken.

In most practical situations we do not collect enough data to calculate the population standard deviation. Instead, we calculate our best estimate of the standard deviation and we call it the sample standard deviation.

$$s = \sqrt{\frac{\sum_{i=1}^N (x_i - \bar{x})^2}{N - 1}} \quad \text{Equation 4}$$

Again, we see that it is the root of the sum of the errors squared divided by the number of degrees of freedom. The degrees of freedom are explained below.

Statistical Tests and Calculations

Confidence intervals

In nearly all of our measurements we do not take sufficient data to actually calculate the population mean or standard deviation. This is simply a matter of expediency because it would require many measurements and would take an enormous amount of time and effort. Luckily, we can predict a range of values that will include the population mean if we have several sample measurements. We can also predict the size of the population standard deviation even if we only have a sample standard deviation.

However, as a consequence of this expediency our predictions are not always correct. The level of risk is associated with the level of confidence. When a 95% confidence level is chosen, your prediction of the range of values that contains the mean will be correct 95% of the time. If you aren't lucky your prediction will be wrong!

CASE 1

We want to know the range of values that the population mean will be found in given that we have a **single** sample measurement and the population standard deviation. We must make the assumption that σ also applies to our measurement (this is a bit risky).

We have σ from many other analyses.
We have x from our experiment.

From the math of the normal (Gaussian) distribution we can predict that:

68.3% of the time x will be within one σ of μ
95.5% of the time x will be within two σ of μ
99.7% of the time x will be within three σ of μ

$$\mu = x \pm z\sigma \quad \text{where } z \text{ is chosen by you.}$$

Equation 5

CASE 2

We repeat our experiment N times.

Recall that in the limit $\bar{x} \xrightarrow{N \rightarrow \infty} \mu$. If the variation is due to random sources the average (\bar{x}) will approach the mean (μ) as $\sqrt{\frac{1}{N}}$. This means that our average is even more likely to be closer to μ than was our single measurement in CASE 1. In fact

68.3% of the time \bar{x} will be within one σ/\sqrt{N} of μ

95.5% of the time \bar{x} will be within two σ/\sqrt{N} of μ

99.7% of the time \bar{x} will be within three σ/\sqrt{N} of μ

i.e. $\mu = \bar{x} \pm z\sigma/\sqrt{N}$ where z is chosen by you.

Equation 6

Later you will see that z is in fact just a special circumstance of the more general t statistic. For now, z is the number of standard deviations.

Now you can see the advantage of taking a few measurements. They significantly tighten the range of values that we can expect to find the mean value in.

CASE 3 (MOST REALISTIC)

Most often we don't know σ and only have s . In the limit $s \xrightarrow{N \rightarrow \infty} \sigma$, for smaller numbers of N we have to be cautious and realize that s may underestimate σ . To compensate we use the t statistic so once again we have

68.3% of the time \bar{x} will be within one ts/\sqrt{N} of μ

95.5% of the time \bar{x} will be within two ts/\sqrt{N} of μ

99.7% of the time \bar{x} will be within three ts/\sqrt{N} of μ

i.e. $\mu = \bar{x} \pm ts/\sqrt{N}$ where t is chosen by you.

Equation 7

The value of t can be found in the Students t -table. They are sorted according to the level of confidence and the number of degrees of freedom.

Statistical Tests

F Test

All of the following t-tests look for differences in the mean values between two sets of data. To perform these tests requires that the precisions (σ or s) are similar for the two test populations. To test if this is true use the F test.

$$F_{\text{calc}} = S_a^2/S_b^2 \quad \text{Where } S_a > S_b$$

Equation 8

If $F_{\text{calc}} > F_{\text{table}}$ the variances are too different and the two populations cannot be compared with the student t tests. Otherwise they are similar enough to continue testing.

T Statistic

The Students t statistic is very useful in data analysis and is used quite extensively in analytical chemistry. Intuitively, we know that if we have a population and we take a small number of samples from the population and calculate the average value it should be close to the mean value. The real question is, *how close?* The values of t in the students t-table can be used to answer that question. We can calculate a value for t using the data and then compare it to the table value. The table values can be thought of as the maximum allowable difference between the average of a small set of data and the population mean.

We have the familiar

$$\mu = \bar{x} \pm z\sigma/\sqrt{N}$$

Equation 9

Rearranged/modified to

$$z_{\text{calc}} = t_{\text{calc}} = (\bar{x} - \mu)\sqrt{N}/\sigma$$

Equation 10

Here you see that t_{calc} is related to $\bar{x} - \mu$, or the difference between the measured and expected means.

Always use the positive value of t since the tables are for the positive value!

T Tests

CASE 1: Where μ and σ are both well known (certified reference material or quality control) and we have an experiment with N measurements.

We rearrange $\mu = \bar{x} \pm ts/\sqrt{N}$ to

$$\pm t_{\text{calc}} = (\bar{x} - \mu)\sqrt{N}/\sigma$$

and calculate t_{calc} from the experimental data. If t_{calc} is greater than t_{table} then our data is too far away (i.e. $(x-\mu)$ is too large) to be due to random error (σ) alone. If t_{calc} is less than or equal to the tabulated value, the difference $(x-\mu)$ could be due to the amount of noise (error, variation) in the experiment and we predict that there is no difference between x and μ .

CASE 2: Where σ is well known and we want to compare the results of two experiments with \bar{x}_a , N_a , \bar{x}_b and N_b .

$$\pm t_{\text{calc}} = [(\bar{x}_a - \bar{x}_b)/\sigma] \sqrt{\frac{N_a \cdot N_b}{N_a + N_b}}$$

Calculate t from the experimental data. If t_{calc} is greater than t_{table} then the difference in the two sets of data is too great ($\bar{x}_a - \bar{x}_b$ is too large) to be due to random error (σ) alone. If t_{calc} is less than or equal to t_{table} the difference ($\bar{x}_a - \bar{x}_b$) could be due to the amount of noise (error, variation) in the experiments.

CASE 3: Where σ is **NOT** well known and we want to compare the results of two experiments with \bar{x}_a , N_a , \bar{x}_b and N_b . THIS IS THE MOST COMMON CASE.

$$\pm t_{\text{calc}} = [(\bar{x}_a - \bar{x}_b)/S_{\text{pool}}] \sqrt{\frac{N_a \cdot N_b}{N_a + N_b}}$$

$$\text{Where } S_{\text{pool}} = \sqrt{\frac{\sum_{i=1}^{N_a} (x_{ai} - \bar{x}_a)^2 + \sum_{j=1}^{N_b} (x_{bj} - \bar{x}_b)^2}{N_a + N_b - 2}}$$

Equation 11

A more easily calculated version is

$$S_{\text{pool}} = \sqrt{\frac{(N_a - 1)s_a^2 + (N_b - 1)s_b^2}{N_a + N_b - 2}}$$

Equation 12

Q-TEST

Q Test: Can we throw out a piece of bad data?

$$Q_{\text{calc}} = \text{gap}/\text{range} = (\text{Questionable data} - \text{closest neighbor})/(\text{biggest} - \text{smallest})$$

Equation 13

If $Q_{\text{calc}} > Q_{\text{table}}$ then the “questionable” data point is too far away from the other data and can be removed.

Most spreadsheets do not calculate the rejection quotient (Q) so I have prepared the following table for you.

Number of Measurements	Level of confidence		
	90%	95%	99%
3	0.941	0.970	0.994
4	0.765	0.829	0.926
5	0.642	0.710	0.821
6	0.560	0.625	0.740
7	0.507	0.568	0.680
8	0.468	0.526	0.634
9	0.437	0.493	0.598
10	0.412	0.466	0.568
15	0.338	0.384	0.475
20	0.300	0.342	0.425
30	0.260	0.298	0.372

Detection Limit

The detection limit is the smallest signal, or concentration, that is statistically larger than the blank signal, or concentration. It is useful to know this value because it sets the lower limit of the signals, or concentrations, that can be reported. For example if you have determined that the detection limit is 20 PPM and one of your samples calculates out to be 15 PPM you must report that no analyte was detected for that sample since the calculated value falls below the detection limit!

The detection limit equation can be derived from the following simple example: A samples' signal comprises two parts, the first part is due to the presence of analyte and the second is from non-analyte species (blank). All of our measurements are subject to noise including the blank measurement. Thus:

$$\text{Blank measurement} \quad \text{Sig} = \text{Sig}_{\text{blk}} \pm \sigma_{\text{blk}}$$

From our knowledge of the Normal distribution we know that 68.3% of the time (our level of confidence) any blank measurement will be within $1 \sigma_{\text{blk}}$ of Sig_{blk} . The converse is also true, the blank signal will be larger than $\text{Sig}_{\text{blk}} + 1\sigma_{\text{blk}}$ only 15.85 % of the time ($[100-68.3]/2$ divide by two since only larger signals apply). We can extend this reasoning to any level of confidence that we desire by consulting the Student's t table and looking up the appropriate level of confidence. Now, if we want to measure a signal that we can say is statistically larger than the blank we know that the signal must be larger than:

$$\text{Signal at detection limit} \quad \text{Sig}_{\text{@DL}} = \text{Sig}_{\text{blk}} + k\sigma_{\text{blk}}$$

Notice the \pm has been replaced by $+$, this means that the signal is larger than the blank! k is an appropriate number of "sigmas" that will give you the desired degree of confidence. Most often k is chosen as 3 to give XX% (you look it up!) confidence that the signal that has just been measured is not the same as the blank.

Converting the signal at the detection limit to a concentration detection limit.

Using the calibration general formula:

$$\text{Concentration} = (\text{Signal} - \text{signal}_{\text{blk}})/\text{slope}$$

We substitute in the equation for the signal at the detection limit.

$$\begin{aligned} \text{Concentration}_{\text{@DL}} &= (\text{Sig}_{\text{blk}} + k\sigma_{\text{blk}} - \text{signal}_{\text{blk}})/\text{slope} \\ \text{Conc}_{\text{@DL}} &= k\sigma_{\text{blk}}/\text{slope} \end{aligned}$$

There are quite a few assumptions in defining the detection limit this way and some do not agree that this is a "good" definition. However it is a useful guide and can help the analyst "get a feel" for the practical lower range of concentrations.

Linear Regression

Most of the measurements that we take in analytical chemistry do not yield the final result directly. Most often a signal is measured that is proportional to the amount of analyte present. We calibrate the instrument using standards of known concentration or amount of analyte in order to establish the relationship between the signal and concentration.

Thus we have Signal = sensitivity x concentration + blank signal
 Sig = sens. x [analyte] + blank

Which is in the same form as: $y = mx + b$

The slope and the intercept of the line can be determined graphically by using your best judgment and drawing a straight line through the data and calculating the slope and intercept from this line. There is, not too surprisingly, a better way than this and it is called the least squares method. Most hand calculators and spreadsheets are capable of doing this calculation.

The Appendix details the method of calculating these parameters using Excel, here is the mathematical model used for those calculations.

We have a data set that is composed of signals (y values) and concentrations (x values).

First we calculate a few intermediate values for convenience sake.

$$S_{xx} = \sum (x_i - \bar{x})^2$$

Equation 14

$$S_{yy} = \sum (y_i - \bar{y})^2$$

Equation 15

$$S_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y})$$

Equation 16

Where \bar{x} and \bar{y} are the averages of the entire set of x and y values respectively.

From this data set we can calculate the following calibration curve numbers.

$$\text{Slope} = m = S_{xy}/S_{xx}$$

Equation 17

$$\text{Intercept} = b = \bar{y} - m\bar{x}$$

Equation 18

In an ideal case all of the measured signals will all satisfy the equation and when they are plotted all of the data will fall onto the line. In the real world there is random noise (uncertainty, error) in our measurements and all of the data will not sit on the line. This uncertainty in the measurement of the standards leads to an uncertainty in any value calculated using the linear least squares line. The equations necessary to calculate the uncertainty are presented next.

The first thing we need to get a handle on is the magnitude of the scatter of the data “away” from the line. We will use the same general idea of a standard deviation that we used for single data sets except it will be known as the standard deviation about the regression. The standard deviation about the regression follows the general form for standard deviations (i.e. the root of: the sum of [the actual value (y_i) minus the expected value ($b+mx_i$)]² divided by the degrees of freedom).

$$S_r = \sqrt{\frac{\sum_{i=1}^N [y_i - (b + mx_i)]^2}{N - 2}} = \sqrt{\frac{S_{yy} - m^2 S_{xx}}{N - 2}}$$

Equation 19

This value allows us to calculate an uncertainty in our estimate of the slope and the intercept.

The standard deviation of the slope

$$S_m = \sqrt{\frac{S_r^2}{S_{xx}}}$$

Equation 20

The standard deviation of the intercept

$$S_b = S_r \sqrt{\frac{\sum x_i^2}{N \sum x_i^2 - (\sum x_i)^2}} = S_r \sqrt{\frac{1}{N - (\sum x_i)^2 / \sum x_i^2}}$$

Equation 21

With these values of the standard deviations it is possible to calculate the standard deviation in a calculated result.

$$S_c = \frac{S_r}{m} \sqrt{\frac{1}{M} + \frac{1}{N} + \frac{(\bar{y}_c - \bar{y})^2}{m^2 S_{xx}}}$$

Equation 22

Where

- m is the calculated slope
- M is the number of number of measurements of the unknown
- N is the number of measurements that are used to calibrate
- \bar{y}_c is the average signal of the M unknown measurements
- \bar{y} is the average of all of the signals used in the calibration

Now you must calculate the confidence interval on the calculated result(s).

We have our general equation (Equation 7) that estimates the range of values that the true value can be found in:

$$\mu = \bar{x} \pm ts/\sqrt{N}$$

in our case we will use the same notation as equation 22:

$$\mu = \bar{x} \pm ts_c/\sqrt{M} \quad \text{Where M is the number of unknown measurements.}$$

The real sticky question is: how many degrees of freedom to use when looking up t ?

Remember that when we use our calibration data (N measurements) set to determine the concentration of an unknown (M measurements) we are saying that the unknown belongs in the same data set as the calibration data. This implies that we start with M + N degrees of freedom. One degree is transferred (lost) every time we must use our calibration (or unknown) signal data.

In our calculation of S_c we have calculated m, S_r , \bar{y}_c and \bar{y} . On the first glance it would appear that we lose 4 degrees from the M+N total degrees. However, this is not the case if you examine how S_r is calculated you will see (second half of Equation 19) that we already have m and S_{yy} uses \bar{y} and other values that don't require the calibration data set. So, we can avoid losing a degree of freedom due to the calculation of S_r . In short we have N-2 (m and \bar{y}) plus M-1 (\bar{y}_c) or M+N-3 degrees of freedom!! Now go look it up in the table!!

Linear regression with Excel

Many of the Excel functions that we are going to use are not installed by default when Excel is first installed on a computer. You will have to install the Analysis Tool-Pak to have access to functions that are useful for working with and analyzing data. To install the Tool-Pak start Excel, choose TOOLS\Add-Ins then select Analysis Tool-Pak. Follow the installation instructions. In principle the Analysis Tool-Pak needs to be installed only once, occasionally though it “disappears” and will need to be reinstalled. Don’t blame me, blame Bill Gates!

If you don’t know how to enter data, equations, copy data etc in Excel check out the Appendix to this Guide.

Data Work-up

Place your data in the spreadsheet, one column for your independent variable (usually concentration) another column for the dependant variable (signal). It should look something like:

Concentration of Standard	Signal data from stds
0.00	3.07
0.00	2.79
0.00	2.90
1.00	8.52
1.00	9.88
1.00	8.44
2.00	16.47
2.00	15.19

More data...

Choose TOOLS\DATA ANALYSIS select regression. A dialog box will open. Select the dependant data as the Y-Input Range, the independent data as the X-Input-Range. Select Output Range and choose an empty place in your spreadsheet to put the Regression output. The cell that you choose will be the top left corner of the output. For the demonstration below I checked the Residuals box to calculate the residuals, normally you won’t need the residuals.

For the data set that I used the Regression output looked like:

SUMMARY OUTPUT

<i>Regression Statistics</i>	
Multiple R	0.9964
R Square	0.99282
Adjusted R Square	0.99256
Standard Error	0.92615 ← This is useful
Observations	30

ANOVA

	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression	1	3320.37	3320.37	3870.99	1.5E-31
Residual	28	24.0173	0.85776		
Total	29	3344.39			

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>	<i>Lower 95.0%</i>	<i>Upper 95.0%</i>
Intercept	2.9176	0.29977	9.73287	1.7E-10	2.30355	3.53164	2.30355	3.53164
X Variable 1	6.16012	0.09901	62.2172	1.5E-31	5.95731	6.36294	5.95731	6.36294

RESIDUAL OUTPUT

<i>Observation</i>	<i>Predicted Y</i>	<i>Residuals</i>
1	2.9176	0.15178
2	2.9176	-0.1308
3	2.9176	-0.0158
4	2.9176	-0.1228
5	2.9176	0.08558

I HAVE CHOPPED OUT A BUNCH OF ROWS HERE

27	33.7182	-0.978
28	33.7182	1.50713
29	33.7182	-1.3837
30	33.7182	0.6377

Useful numbers from the Regression output:

Intercept coefficient is, well... you guessed it, the intercept!

X Variable 1 coefficient is the slope of the curve.

Standard error of the intercept is the estimate of the standard deviation on the intercept.

Standard error of the x-variable is the estimate of the standard deviation of the slope.

Standard error is the standard error of the regression (Sr)

The following example shows that the standard error is actually the standard deviation about the regression.

To obtain an estimate of the standard deviation of the regression or the standard deviation that can be expected in all signal measurements (equation 1-34 in Skoog or equation 5-7 in Harris 5th) you must determine the difference between the individual data points and the calibration curve (expected value). Do this by summing the squares of the residuals, dividing by the remaining degrees of freedom N-2 (in the data set I have N-2 = 28), take the square root and then sum the column. Notice that this is equal to the standard error reported in the summary output.

My data looks like:

RESIDUAL OUTPUT

<i>Observation</i>	<i>Predicted Y</i>	<i>Residuals</i>	(Resid)^2
1	2.917597	0.151783	0.023038
2	2.917597	-0.13077	0.0171
3	2.917597	-0.01576	0.000248
I CHOPPED OUT SOME DATA HERE TO SAVE SPACE			
27	33.71822	-0.97798	0.956448
28	33.71822	1.507129	2.271438
29	33.71822	-1.38366	1.914527
30	33.71822	0.637703	0.406665

Estimated

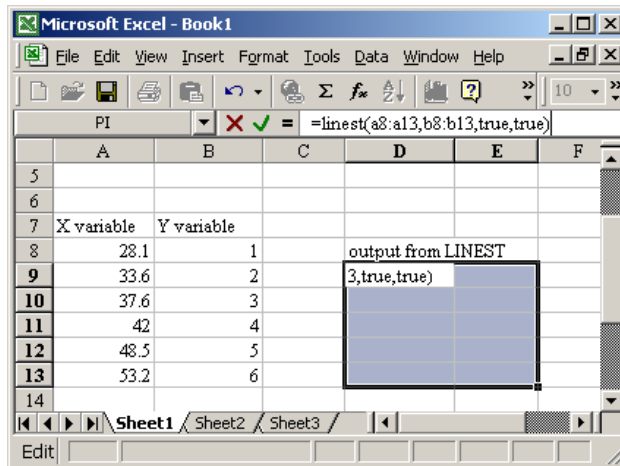
Standard deviation

Sr 0.926153

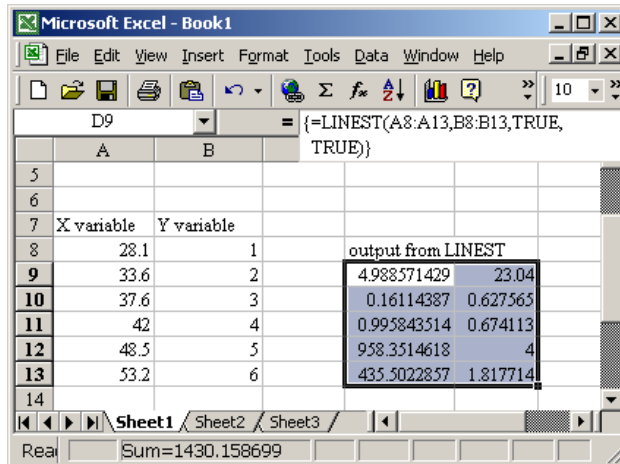
=SQRT(SUM(Data)/28))

If you don't have the tool pack for linear regression you can use the built in linear regression function (LINEST) in Excel. Use the Excel help function to get instructions on how to enter **array functions**. There is a trick for entering the function.

1. Highlight an empty block of cells (where it will give you the results) 5 rows by 2 columns.
2. Type `=linest(startY:endY,startX:endX,true,true)` in the formula bar. It should look like:



3. Hit CTRL + SHIFT + ENTER all at the same time and Excel will calculate the linear regression. Each of the cells in the 5X2 block contains a regression parameter. Your results should look like these:



the output block is arranged as:

slope	intercept
s_dev slope	s_dev inter
R_squared	Std_err_of_Y
F_stat	degrees'o_free
	(sum of resid)^2

Calculating the unknown concentration and its uncertainty.

Our calibration curve is:

$$\text{Signal} = \text{Slope} * \text{Concentration} + \text{Intercept}$$

Which rearranges to:

$$\text{Concentration} = (\text{Signal} - \text{Intercept})/\text{Slope}$$

Calculate the average unknown concentration using the average unknown signal and the values for slope and intercept.

To calculate the uncertainty we apply Equation 22. We don't have all of the terms directly from the output of the spreadsheet but they are all easily calculated.

\bar{y}_c is the average signal of the M unknown measurements

\bar{y} is the average of all of the signals used in the calibration

m , M and N are known

S_r is reported in the Regression Summary

S_{xx} is readily available by rearranging Equation 20, the standard deviation of the slope and the standard error are known.

Substitute the terms into $S_c = \frac{S_r}{m} \sqrt{\frac{1}{M} + \frac{1}{N} + \frac{(\bar{y}_c - \bar{y})^2}{m^2 S_{xx}}} = \sigma_{\text{Unknown}}$

3) Apply a suitable confidence interval now that you have the uncertainty.

$$\text{The true (mean) unknown concentration} = [\text{Unknown}] \pm t \sigma_{\text{Unknown}}/\sqrt{N}$$

Where N is the number of replicate measurements of the unknown.

This method of calculating the concentration and error makes a couple of assumptions. The largest assumption, that you should be aware of, is that the sample is a member of the same population as the standards (i.e. a sample and a standard of the same concentration would give the same signal AND would have the same standard deviation σ). Making this assumption allows us to use the error data from the calibration data in our error propagation rather than the measured sample standard deviation. If this is not the case then we have to abandon our Equation 22 for a more complex method.

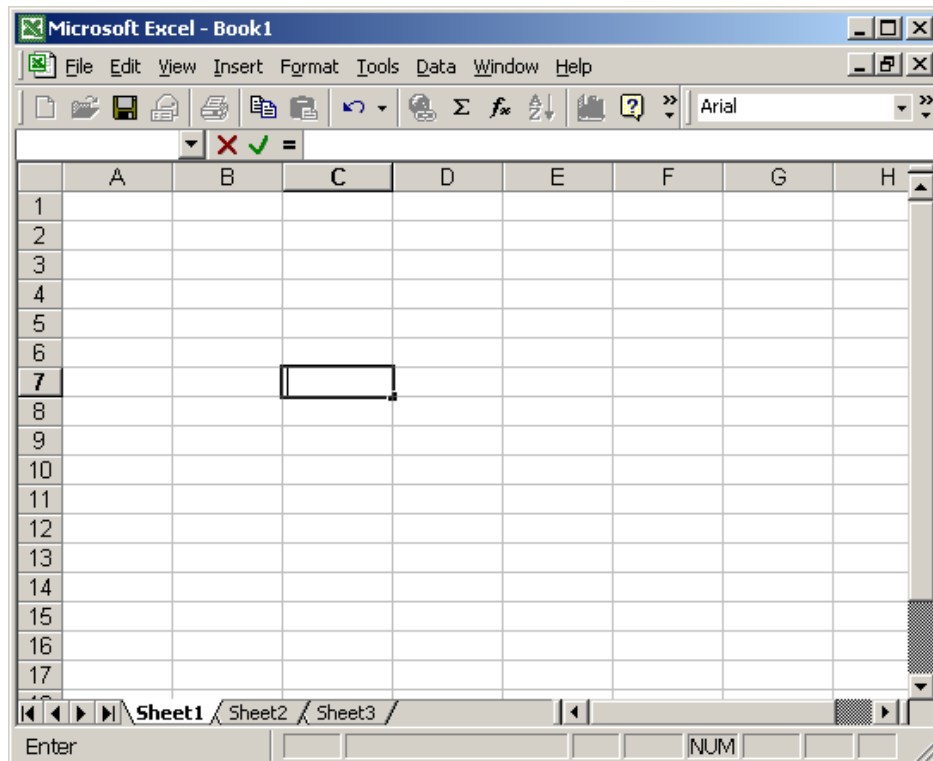
Appendix (Basics of how a spreadsheet works)

This Appendix is designed for those of you who are new to spreadsheets and are unfamiliar with how they work. If you've used one before this is probably too simple an introduction.

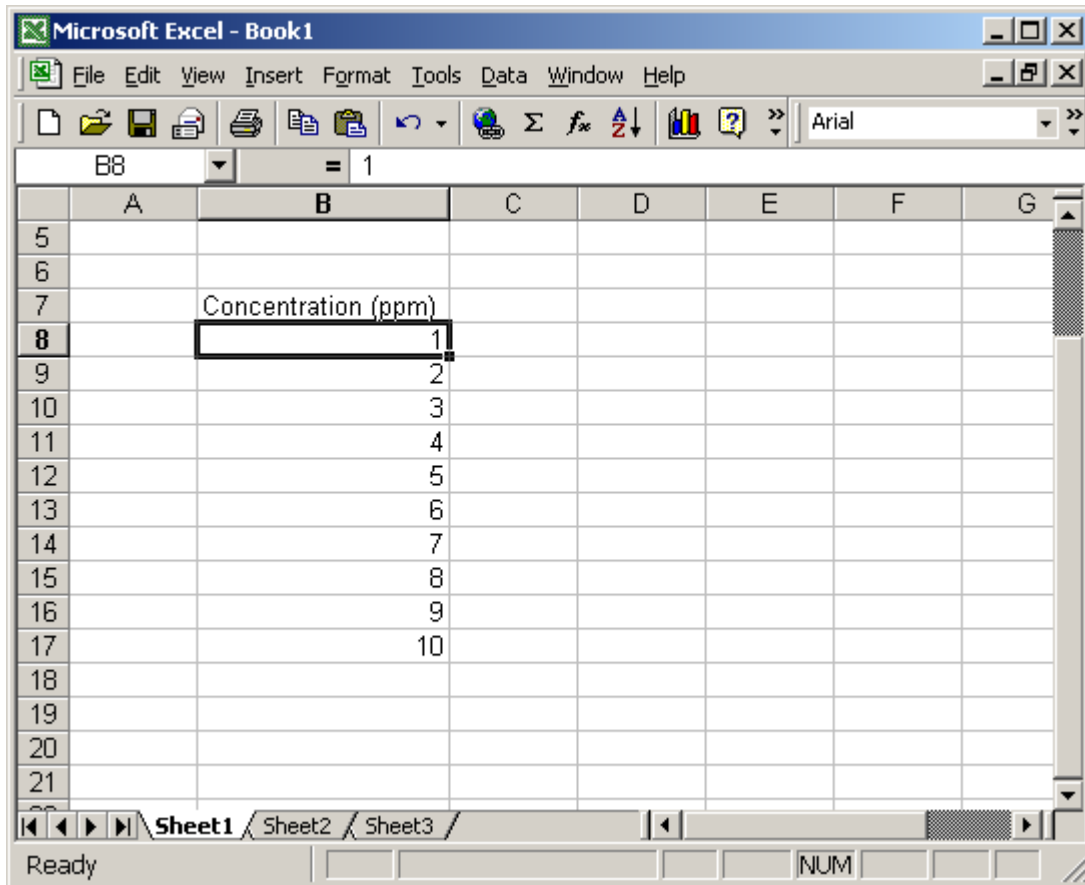
You can think of a spreadsheet as a two dimensional matrix, each cell in the matrix may contain (store) a number, text or a formula. The individual cells have unique addresses that are determined by the combination of the row and column numbers. The most powerful feature of the spreadsheet is that formulas can take data in other cells (or results from other calculations) and use them in their calculations. To “get” the value from another cell and use it in a calculation you use the address of the cell in the formula rather than the numerical value displayed (stored) in the cell. This allows you to perform long complex calculations as a series of smaller intermediate calculations that are easily checked. It also allows you to fix data entry errors by changing the values in the cells that contain the data, all of the calculations will automatically update. Sweet huh?

Entering data into a spreadsheet

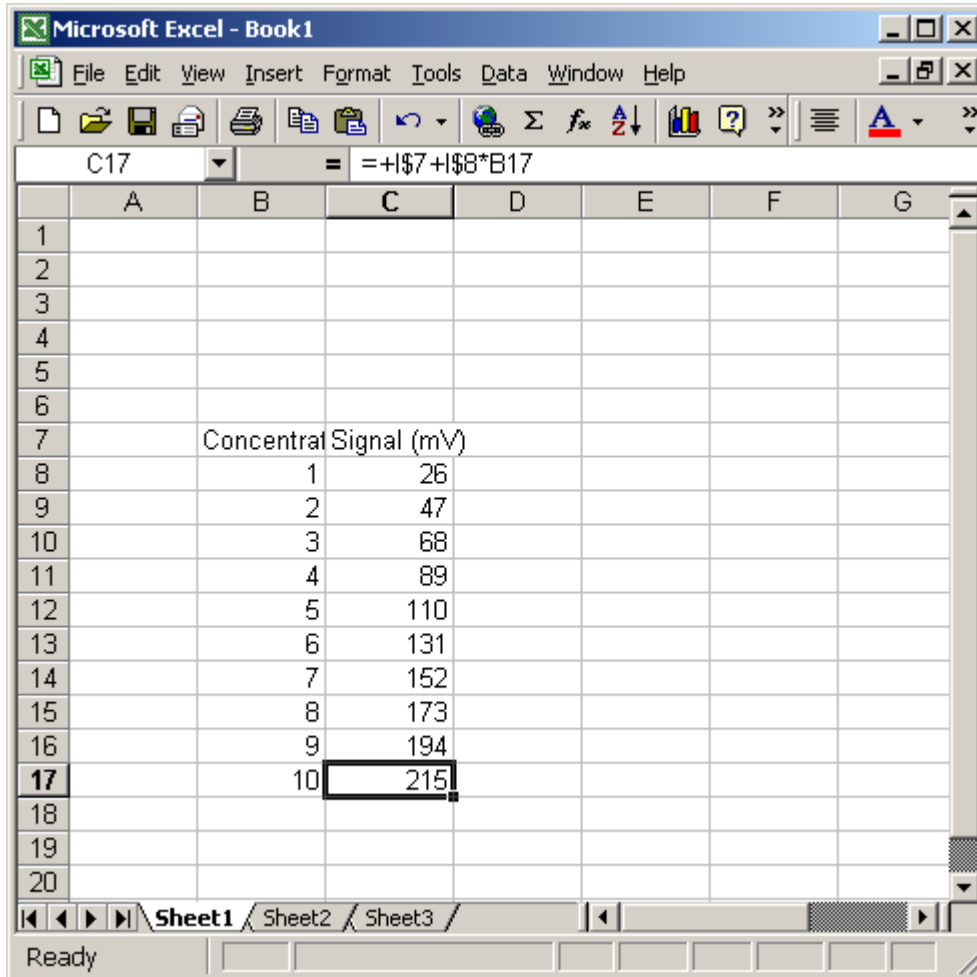
You can enter data directly into the individual cells of the spreadsheet by using the mouse to select a cell (note that it becomes highlighted along with its row and column index). In this case, cell C7 is selected. You can use the number pad on the right of the keyboard to enter the values you want. Note you must have Num Lock on (top left key on the number pad, the NUM at the bottom right of the spreadsheet tells you Num Lock in on) Most often it works out best if you enter your data in a column rather than a row. You can choose any column you like.



For this example, let's put in 1 to 10. Give the column of data a name (Concentration (ppm)) so that it will be easier to remember in the future. Notice that the words are wider than the column, so long as nothing is put into cell C7 the full text will be displayed. If text is being overwritten you can adjust the width of the column by moving the mouse pointer to the top of the column, in-between the B and C. The cursor should change to a line with two arrowheads. Use the left mouse button to "grab" the column width and drag it to the right so that it contains all of the text, this won't affect any calculations.



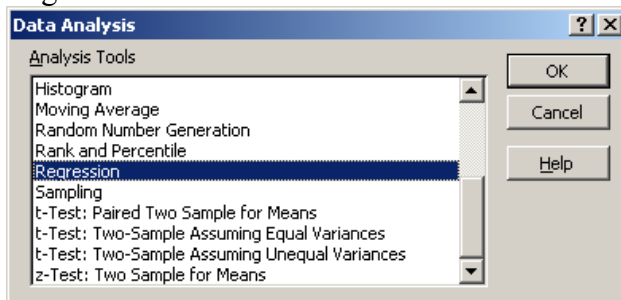
Next, enter some data. In this case I have made up some experimental linear data



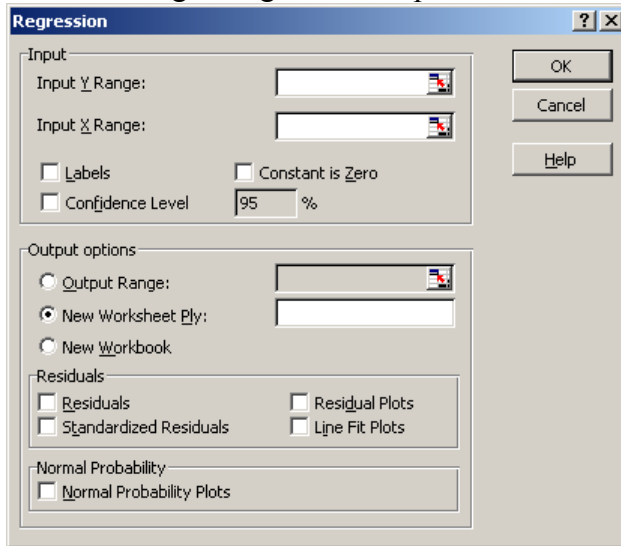
Notice how the Signal label has “overwritten” the Concentration label.

Notice the formula bar (little window at the top) contains the formula that I used to generate the data.

To perform a linear regression on the data, choose TOOLS\DATA ANALYSIS select regression. The following box should open. You will have to scroll down to select Regression

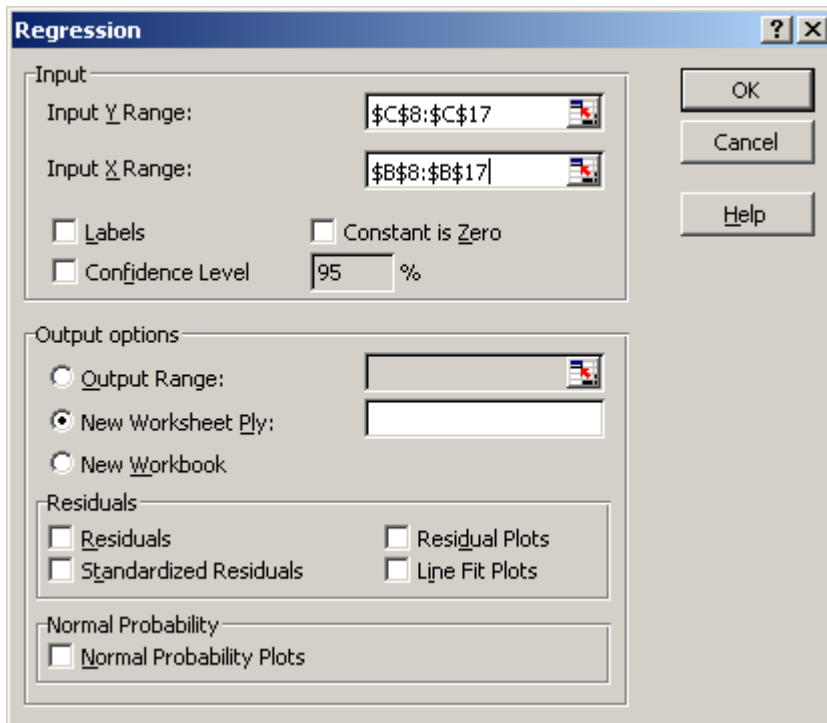


The following dialog box will open.

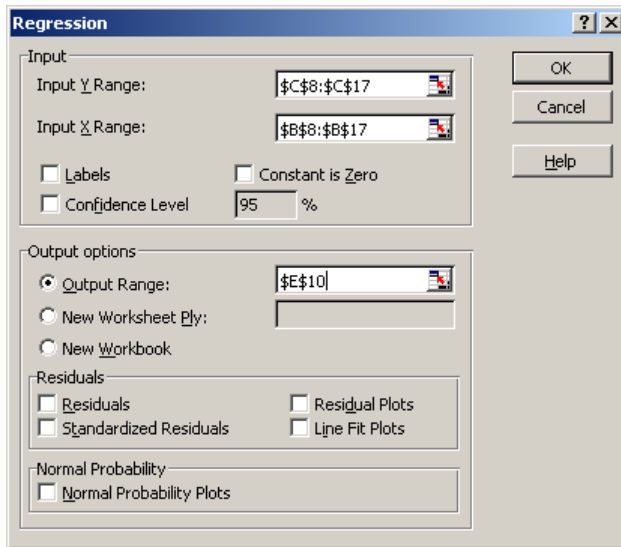


Select the dependant data as the Y-Input Range, the independent data as the X-Input-Range. Do this by clicking on the red arrow beside the data window. This will minimize the Regression window and allow you to select a block of data from the spreadsheet. Select the start of the block of data (holding down left mouse button) to the end of the data. Once you have selected the block of data hit ENTER.

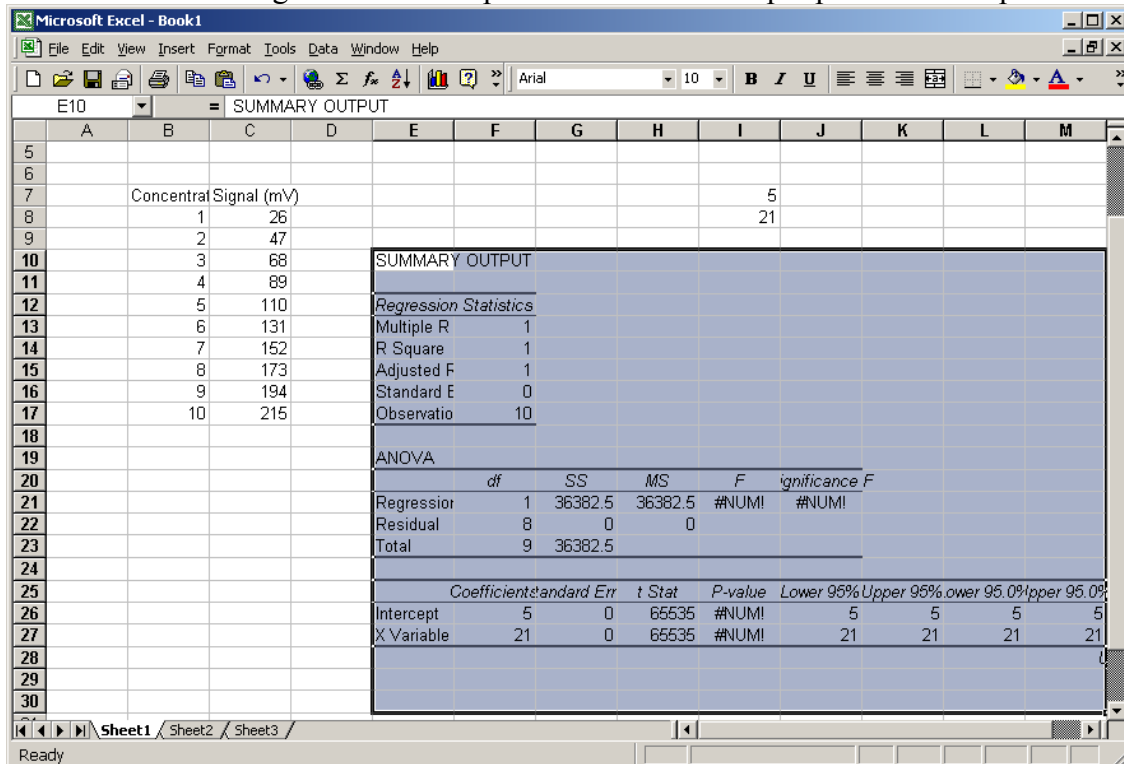
Do this for both the dependant and independent data blocks



Select Output Range and choose an empty place in your spreadsheet to put the Regression output. The cell that you choose will be the top left corner of the output. In this case I chose cell E10.



Select OK and the regression will be performed and the output placed in the spreadsheet.



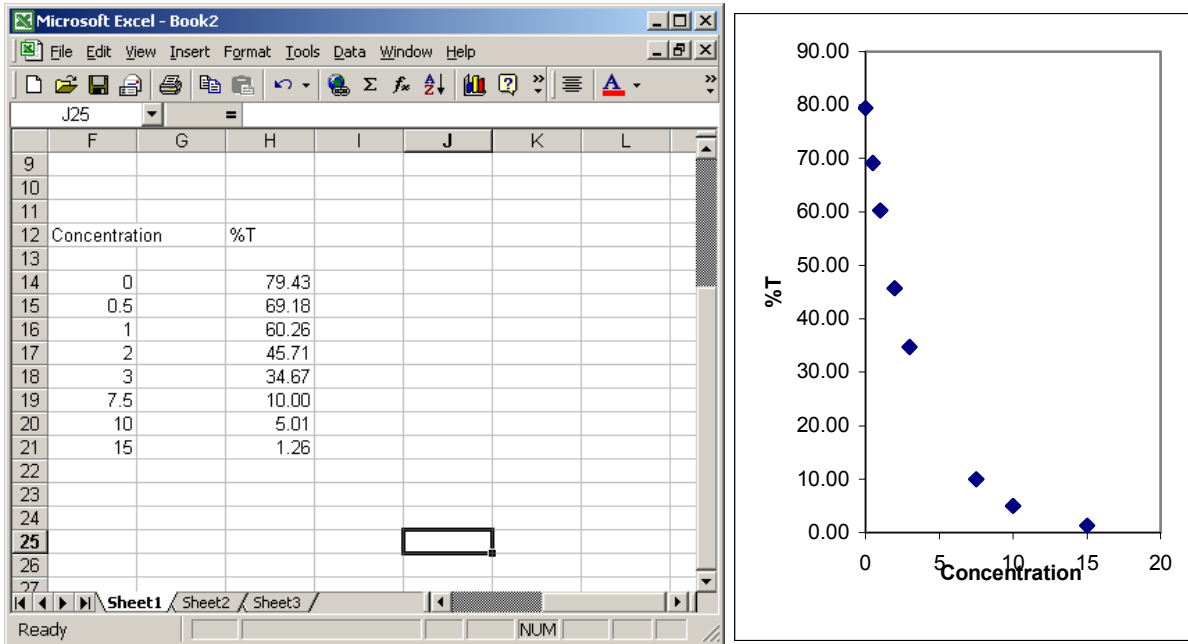
Notice that the intercept is the same the value that I used in the formula (cell I7) and the slope is identical to cell I8.

Entering formula

Formulas can be entered directly into the cells of the spreadsheet. This is particularly useful if the data that you collect must be transformed (linearized) before you can plot or calculate the slope and intercept. Alternatively, you can choose the type of formula that you would like to use and a pop-up window will open and “help” you through the process of filling in all of the important values. Two examples are shown below.

Manual formula entry

Below is an example that uses the %T signal collected from an absorbance spectrophotometer.



As you can see from the figure on the right, %T is NOT proportional to concentration

To linearize the data we transform the %T into absorbance values using the following equation.

$$\text{Abs} = -\log(\%T/100)$$

To have Excel calculate this for you move the cursor to an empty cell (use the same row as the data) and type in the following:

=-log(address/100)

- The = sign lets Excel know that the following text is a formula. You can also use a + sign and Excel will insert a = in front for you
- - is the negative of the value calculated
- **log** is a reserved word for the log function. There is a full range of functions to choose from, they can be found by going to Insert/functions. A dialog box will open that will allow you to select the type of function that you want.
- **address** is the cell address of the data that you want to use in the calculation. In the example that I am using it is cell H14. We divide by 100 to remove the percentage.

The cell should now contain:

=-log(H14/100)

Once you hit enter it will calculate and display the value (the formula is still there though).

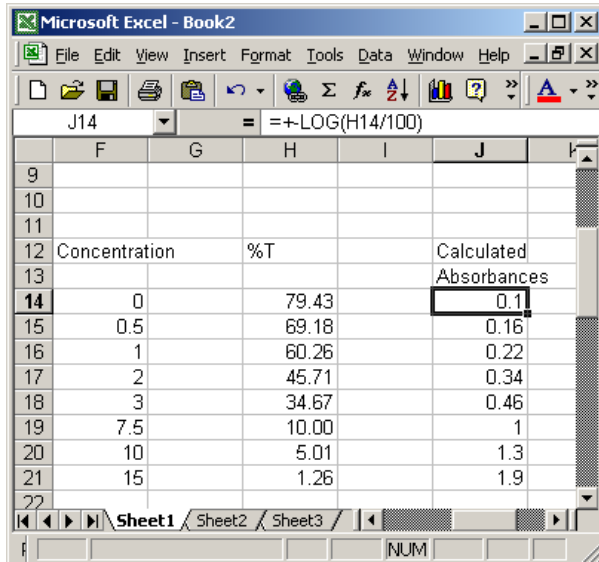
Common functions and their Excel syntax

Function	Excel syntax	Example
Sum	sum(<i>start add.:end add.</i>)	=SUM(B7:B15)
Average	average(<i>start add.:end add.</i>)	=AVERAGE(B7:B17)
Sample standard deviation	stdev(<i>start add.:end add.</i>)	=STDEV(B7:B15)
Student's T value ^{Note 1}	tin(probability,deg. of free)	=TINV(0.01,12)

- 1) The probability is actually = (1-the level of confidence). So for a 95% confidence level the probability is 0.05. You can also check to see if you are calculating the t-value correctly by comparing your results to those in the printed table!

Copying and Pasting & Relative Addressing

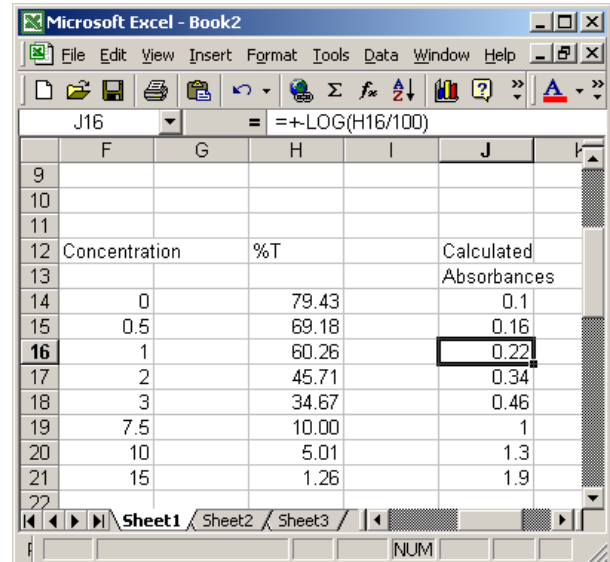
Rather than typing the formula into each cell you can copy the formula to adjacent cells. As Excel pastes the formula into the new cells it UPDATES the addresses.



The screenshot shows the Microsoft Excel interface with the following data in the spreadsheet:

	F	G	H	I	J
9					
10					
11					
12	Concentration		%T		Calculated Absorbances
13					
14	0		79.43		0.1
15	0.5		69.18		0.16
16	1		60.26		0.22
17	2		45.71		0.34
18	3		34.67		0.46
19	7.5		10.00		1
20	10		5.01		1.3
21	15		1.26		1.9
22					

The formula bar shows the formula in cell J14: `=+LOG(H14/100)`.



The screenshot shows the same Microsoft Excel interface, but the formula in cell J16 is now `=+LOG(H16/100)`. The spreadsheet data is identical to the previous screenshot, but the value in cell J16 has updated to 0.22.

	F	G	H	I	J
9					
10					
11					
12	Concentration		%T		Calculated Absorbances
13					
14	0		79.43		0.1
15	0.5		69.18		0.16
16	1		60.26		0.22
17	2		45.71		0.34
18	3		34.67		0.46
19	7.5		10.00		1
20	10		5.01		1.3
21	15		1.26		1.9
22					

Notice how the formula in cell J16 uses the contents of H16 in its calculation! This will be true for all of the formulas, they all use the corresponding cell in the H column. This automatic updating of the formula is known as Relative Addressing since the formula always uses the data that is in the cell relative to its own position. In this case two cells to the left.

Absolute addressing

Let's say that we want to subtract a constant from all of the data (for example a blank subtraction). You can setup a column that contains the constant or you can use absolute addressing in your formula rather than relative addressing. To force the formula to always use the contents of a given cell use the \$ before (and n) the cell address.

In our example we will subtract the blank signal from all of the measurements.

	F	G	H	I	J	K	L
9							
10							
11							
12	Concentration		%T		Calculated		Blank
13					Absorbances		Subtracted
14	0		79.43		0.1		0
15	0.5		69.18		0.16		0.06
16	1		60.26		0.22		0.12
17	2		45.71		0.34		0.24
18	3		34.67		0.46		0.36
19	7.5		10.00		1		0.9
20	10		5.01		1.3		1.2
21	15		1.26		1.9		1.8
22							

Notice that the formula is a mixture of relative and absolute addressing. For the cell highlighted

$$J17 \text{ (relative address)} - \$J\$14 \text{ (absolute address)} = 0.24 \text{ (calculated value)}$$

The \$ before the J forces Excel to always use column J and the \$ before 14 always forces row 14 to be used.