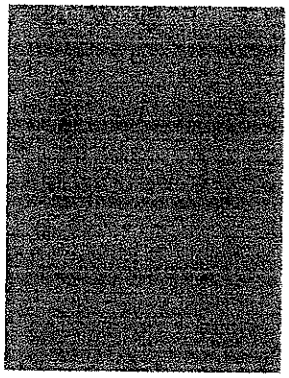
Biology statistics made simple using *Excel*

Neil Millar



Spreadsheet programs such as Microsoft *Excel* can transform the use of statistics in A-level science

Statistics is an area that most A-level biology students (and their teachers!) find difficult. The formulae are often complicated, the calculations tedious, degrees of freedom mysterious, and probability tables confusing. But in fact students need no longer grapple with any of these. In real life, biologists and statisticians rarely use calculation and tables these days, but instead use statistical packages such as *Minitab* or *SPSS*. But it isn't even necessary to buy an expensive statistics package, since spreadsheet software such as *Excel* has most of the common statistical tests built-in.

When using statistics, the first hurdle is to decide *which* statistical test to use. Figure 1 (overleaf) is a flow chart showing when to use the various tests described in this article. There are many other possible statistical tests, but this flow chart should be more than sufficient for A-level biology students. It briefly summarises the *Excel* formulae and how to interpret the results, so it can be used as a handy guide on its own once the student is familiar with the tests. This flow chart should be used when designing an experiment, not after the experiment is complete. This will ensure that the correct kind of data are collected so that the statistical test will be valid. The rest of the article describes in detail how to carry out these tests using *Excel* and how to interpret the results. It is divided into five sections:

ABSTRACT

Modern spreadsheet software, such as Microsoft *Excel*, can transform the use of statistics in biology. Instead of being difficult to do and to interpret, statistical tests become simple to do and much easier to interpret. This article describes when and how to carry out many of the most common tests (including mean, standard deviation, confidence limits, correlation, regression, *t*-test, χ^2-test and ANOVA) using *Excel*.

1	Descriptive statistics	mean, median, mode standard deviation, standard error, confidence interval
2	Graphing data	scatter graphs, bar graphs error bars, lines
3	Association statistics	Pearson coefficient, Spearman coefficient linear regression
4	Comparative statistics	paired and unpaired *t*-test Mann-Whitney *U*-test ANOVA
5	Frequency statistics	χ^2-test χ^2-test of association

1 Descriptive statistics

Most school biology experiments will involve some kind of measurement, such as time, length, mass, temperature, absorbance, etc., and in a well-designed experiment there should be a number of repeats (or replicates) of each measurement. Once some measurements have been collected the first job is usually to summarise them using descriptive statistics. *Excel* has formulae for the three measures of the centre of a distribution of replicates.

The arithmetic mean is given by the formula:

`=AVERAGE (range)`

The median is given by the formula:

`=MEDIAN (range)`

And the mode is given by the formula:

`=MODE (range)`

These formulae are illustrated in Figure 2. In many cases the quantities measured in biology will show a normal distribution, and so the mean is the most appropriate statistic to use. It is also the one students are most likely to know already, and to be able to do by hand. The median and mode are less likely to be needed for experimental data, but some A-level specifications require a knowledge of them. It is unfortunate that *Excel* uses the word 'average' for 'mean', as some textbooks use average as a general term to refer to any measure of the centre of a distribution.

A statistician will tell you that there is no point in calculating a mean without also calculating some measure of the variation or spread of the measurements, but students often don't bother because of the difficulty of the calculations. Figure 2 shows five different measures of the spread, and shows how easy they are to calculate using *Excel*.

- The range is given by the *Excel* formula:
  `=MAX (range) - MIN (range)`
  This is the simplest, but least useful.
- The variance is given by the *Excel* formula:
  `=VAR (range)`
  This is used in calculations, but has little use as a descriptive statistic since it is not in the same units as the measurements.
- The standard deviation (SD) is given by the *Excel* formula:
  `=STDEV (range)`
  This is common (since it is fairly easy to calculate by hand) and it gives a good indication of the variability of a set of data. However it is not the best statistic to use when comparing different sets of data, especially if the data sets are different sizes.
- The standard error of the mean (SE) is given by the formula:
  `=STDEV (range) / SQRT (COUNT (range))`
  This gives an indication of the confidence of the mean, and is often used as an error measurement simply because it is small rather than for any good statistical reason.
- The 95% confidence interval (CI) is given by the formula:
  `=CONFIDENCE (0.05, STDEV (range), COUNT (range))`

	A	B	C	D	E	F	G	H	I
1		group A	group B						
2		13.2	9.3						
3		14.7	17.2						
4		10.5	25.4						
5		12.0	5.1						
6		14.6	21.2						
7		14.3	1.1						
8	mean	13.22	13.22	=AVERAGE(C2:C7)					
9	median	13.75	13.25	=MEDIAN(C2:C7)					
10	mode	#N/A	#N/A	=MODE(C2:C7)					
11	range	4.20	24.30	=MAX(C2:C7)-MIN(C2:C7)					
12	variance	2.83	91.21	=VAR(C2:C7)					
13	SD	1.68	9.55	=STDEV(C2:C7)					
14	SE	0.69	3.90	=STDEV(C2:C7) / SQRT(COUNT(C2:C7))					
15	95% CI	1.35	7.64	=CONFIDENCE(0.05, STDEV(C2:C7), COUNT(C2:C7))					

Figure 2 Eight descriptive statistics. The MODE formula returns #N/A because no values are duplicated, so there is no modal value in these data. Note that *Excel* will always return the results of a calculation to about 8 decimal places. This is usually meaningless, and cells with calculated results should always be formatted to a more sensible precision (Format menu > Cells > Number tab > Number).

The value of 0.05 is used to give the 95% (0.95) confidence interval, and different values can be used for different levels of confidence, such as 0.01 for a 99% confidence interval. There is a 95% probability that the true mean lies within ± CI from the measured mean, and the upper and lower values of this range are called the confidence limits.

Of these five, the 95% confidence interval is the most useful measure of the dispersion of data around the mean, and also the easiest to understand. It is not as well known as the others because it is so difficult to calculate, but using *Excel* it is no more difficult to calculate than the others. It is the preferred statistic to use when comparing different sets of data, and when drawing error bars on a graph. Students should always be encouraged to calculate a CI whenever they calculate a mean, and to refer to it whenever they evaluate their data. If the CI is small compared to the mean then the mean is reliable, but if the CI is large compared to the mean then the mean is unreliable. In Figure 2 the two sets have the same means but different spreads, and the statistics all show that the data in group A have a smaller spread and are therefore more reliable than those in group B.

2 Graphing data

Graphs are an important part of data analysis and are closely connected to statistics, since the choice of graph is connected to the choice of statistical test, as the flow chart in Figure 1 shows. If you are investigating an association between two variables, then you should plot a scatter graph; if you are comparing different sets of data, you should plot a bar graph; and if you are collecting frequency data, then you may plot a bar or pie chart, or a graph may not be appropriate. In *Excel* it is quite easy to plot these graphs, as well as many other types. First enter the data into columns or rows, and select them. Then click on the chart wizard (or Insert menu > Chart). This wizard has four steps:

1 In Graph Type, select the type you want and press Next. Choose 'Column' for bar charts or 'XY (Scatter)' for line and scatter graphs. Do **not** choose 'Line', which plots the data against row number. This is a very common mistake.

2 In Source Data, if the sample graph looks about right, then just press Next. If it looks wrong, you can correct it by clicking on the series tab, and then the red arrow at the end of the X Values box. Then highlight the cells containing the X data in the spreadsheet and press the red arrow again. Repeat for the Y Values box.

3 In Chart Options, the most important tasks are to type in suitable titles for the graph and the two axes. You can also turn off gridlines and legend, which makes the chart look better.

4 In Graph Location, just press Finish. This puts the chart beside the data so you can see both.

Excel graphs are quite flexible and almost everything about them can be changed. Just double-click (or sometimes right-click) on the part you want to change. For example, you can move and re-shape the graph; change the background colour (white is usually best); change the shape and size of the markers (points); join the points; change the axes scales and tick marks; or add a trend line or error bars. Students should be discouraged from using 3D or shadow effects, which only serve to obscure the graph trend. It is worth taking some time to get the graph right, because you can use an existing graph as a template. Simply type the new data in place of the existing data, and the graph automatically changes. The sheet can then be saved as a new file.

Error bars

If you are plotting averages on a scatter or bar graph, then error bars are a very good way to illustrate the confidence of the data on the graph. Again, they are awkward to do by hand, but quite easy with *Excel*, and students should be encouraged to use error bars as a matter of course. Error bars usually show ± CI, although you could also plot them from SD or SE. Double-click on any data point or bar to get the Format Data Series dialogue box, and choose the Y Error Bars tab. Click in the Custom + box, and highlight the range of cells containing your confidence intervals. Repeat for the Custom – box, and then press OK. Error bars are useful for the evaluation part of student investigations. Small error bars suggest reliable data; large error bars suggest dubious data. A line of best fit should pass through the error bars, and a good question to address in an evaluation is 'Could I draw a different line through my error bars?' (in other words, do the data support a different conclusion?). Figure 3 shows a graph where a curve has been drawn, but in fact a straight line would also pass through the error bars, so a linear relation is also supported by the data.

	A	B	C
1	X	Y	CI
2	0.0	0.0	0.0
3	0.8	1.6	0.9
4	3.7	5.2	1.8
5	6.1	6.9	1.5
6	8.0	8.1	1.1
7	9.9	8.8	1.6

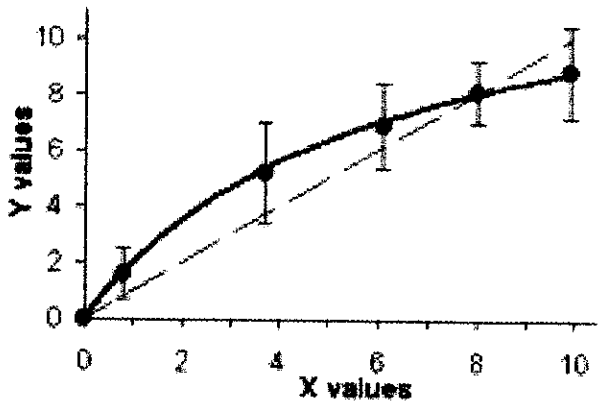


Figure 3 A scatter graph showing error bars. A curved line of best fit has been drawn through the data points, but in fact a straight line can also be drawn within the error bars, so a linear relationship is not ruled out.

Lines

Scatter graphs often have lines, which either join the data points or form a smooth 'line of best fit' (or trend line) through the middle of the points. The choice depends on the circumstances, but generally, if there should be a continuous smooth relation between X and Y, then a trend line is appropriate; otherwise the points should be joined by straight lines. Trend lines are best drawn on the graph by hand, unless you want a linear regression line (see below). To join the points with lines: double-click on any data point, select the Patterns tab and click on Line-Automatic. It is not usually a good idea to have *Excel* draw a curved or smoothed line, as these curves can be highly misleading and can create spurious peaks and troughs for which there is no evidence.

3 Association statistics

A common task in data analysis is to investigate an association between two variables. This can be a correlation to see if two variables vary together, or a regression to see how one variable affects another. We'll see how to do each of these in *Excel*. In both cases a scatter graph should be plotted first.

Correlation

A correlation tells us whether the two variables vary together, i.e. as one goes up the other goes up (or goes down). The most common tests for correlation are the Pearson product-moment correlation coefficient (r) for normally-distributed (parametric) data, and the Spearman rank-order correlation coefficient (r_s) for data that are not normally distributed (non-parametric data). Both vary from +1 (perfect correlation) through 0 (no correlation) to −1 (perfect negative correlation). In *Excel* the Pearson coefficient can be found by two alternative formulae:

`=CORREL (range 1, range 2)`

or

`=PEARSON (range 1, range 2)`

There is no direct formula for the Spearman coefficient, but it can be calculated by first making two new columns for the ranks of the original data. For each of the two variables the largest value is given a rank of 1, the next largest a rank of 2, and so on. This can most simply be done by hand, or for large data sets, by using *Excel*'s =RANK command.

The Spearman coefficient is then simply the Pearson coefficient calculated on the rank data, ignoring the original data. Both coefficients are demonstrated in Figure 4. This shows measurements on the size of breeding pairs of penguins to see if there is a correlation between the sizes of the two sexes. The Spearman coefficient r_s (0.77) is more conservative than the Pearson coefficient r (0.88), but both show a strong positive correlation.

Linear regression

Regression is used when we have reason to believe that changes in one variable cause the changes in the other. A correlation is not evidence for a causal relationship, but very often we are aware of a causal relationship and we design an experiment to