

FACT SHEET

Using *Excel* for descriptive statistics

Introduction

Biologists no longer routinely plot graphs by hand or rely on calculators to carry out difficult and tedious statistical calculations. These can be carried out using a variety of statistical software packages, including *SPSS* and *Minitab*. These are expensive. As an advanced level science student, you are more likely to have access to Microsoft *Excel*.

Excel contains many statistical functions which can be used for most of the graphs and calculations that you need to be able to use. Using such software has many advantages, not least by making your calculations and conclusions from the data much more reliable:

- calculations and graph plots are much easier to carry out
- arithmetical errors are far less likely
- results of tests are much easier to interpret

Nonetheless, OCR Unit 14 *Ecology and managing the environment* specifications still insist that *initial* calculations are carried out by hand:

A range of methods is available to display ecological data, e.g. line graphs, bar graphs, histograms, kite diagrams, pictographs, pie graphs, rose diagrams and scatter graphs. Calculations and statistical tests, carried out by hand calculation, or using computers, are particularly useful in ecology and are necessary to determine whether differences in data are the result of chance. It needs to be emphasised, though, that the first time candidates use statistics, calculations need to be carried out manually.

Check with your teacher

When you carry out statistical calculations, **ask your teacher** what you need to record before you submit work for assessment. You may need to show the stages of your calculations.

General note on using *Excel*

In calculations, *Excel* usually defaults to 8 decimal places, so **always format cells to give suitable precision**:

- click on the cell that you want your calculated value to appear in
- click *Format* to get the menu and then click *Cells...* (or right click your cell and click *Format cells*)
- In the *Category* list, click *Number*: change the *Decimal places* to fit your original data.

Descriptive statistics

Although entering a large set of numbers into a spreadsheet can take some time, once they are there you can carry out a wide range of operations and calculations very easily and quickly. You are also much less likely to make mistakes than when you use a calculator.

Tip: If you have a large set of data to enter, work with a partner. One person can type in the numbers, while the other reads them out. This makes it much easier and faster. The person reading out the data can also cross out each item in turn, to make sure they are all entered and there are no repeats. You can also check that the number of entries matches the original list.

'Average'

If a series of repeated measurements are taken, it is usually useful to summarise them by finding an 'average' value or 'measure of central tendency'. If the values are typed into an Excel spreadsheet, the values for *arithmetic mean*, *median* and *mode* respectively can be found using the formulae:

=AVERAGE(number 1, number 2...)

calculates the total of the values divided by the number of values

=MEDIAN(number 1, number 2...)

finds the value at the centre of the range of values

=MODE(number 1, number 2...)

finds the most frequent value, gives #N/A if there are no repeated values

Confusingly, 'average' is sometimes used to mean *arithmetic mean* and sometimes any one of the central measures: *mean*, *mode* or *median*. It is better to avoid using 'average'.

If the three values coincide, you are likely to have a normal distribution with an even spread either side of the mean. If they vary, you need to consider *why*. Is it because there is a definite pattern or is it because your data are not a representative sample of the population?

Using *Excel* to find the mean, mode and median:

- 1 Enter your raw data (the set of measurements).
- 2 Click on the cell where you want the value to appear (mean, mode or median).
- 3 Type in the function up to the first bracket (it will appear in the **fx** box at the top).
- 4 Click and drag over the cells with the numbers that you wish to include in the calculation.
- 5 Press ENTER key ↵ (*Excel* will add the final bracket to your formula).
- 6 The value will appear in your cell chosen in 2.

- 7 *Excel* usually defaults to 8 decimal places, so always format cells to give suitable precision:
- click *Format* to get the menu and then click *Cells...* (or right click your cell and click *Format cells*)
 - in the *Category* list, click *Number*
 - change the *Decimal places* to fit your original data.

Generally speaking, you should not exceed the number of significant figures used in your original data and should normally drop down to one less after calculations. You may need to use different numbers of decimal places in different cells.

Using *Excel* to determine spread

- *Range* (difference between the largest and the smallest value) is given by:
 $=\text{MAX}(\text{number 1, number 2...})-\text{MIN}(\text{number 1, number 2...})$
- *Variance* measures spread by comparing squares of values with the mean. It is often used in other calculations. Since it does not have the same units as the measurements, it is not often used as a descriptive statistic. It is found using:
 $=\text{VAR}(\text{number 1, number 2...})$
- *Standard deviation* is the square of the variance and so restores the original units. It is a useful measurement of the spread of a set of data. It is found using:
 $=\text{STDEV}(\text{number 1, number 2...})$
- *The standard error of the mean* gives a measurement of confidence of the mean. Use:
 $=\text{STDEV}(\text{number 1, number 2...}) / \text{SQRT}(\text{COUNT}(\text{number 1, number 2...}))$
- *The 95% confidence interval (CI)* gives a value for which the true mean has a 95% probability of lying $\pm\text{CI}$ from the measured mean. It is the best method to use to compare sets of data or to draw error bars on a graph. It is found using:
 $=\text{CONFIDENCE}(0.05,\text{STDEV}(\text{number 1, number 2...}),\text{COUNT}(\text{number 1, number 2...}))$
Note: 0.01 can be used instead of 0.05 to give the 99% confidence interval, and so on.

Using *Excel* to evaluate data:

Repeating a method to get a second or third value for the mean gives you some idea of the consistency of the results that you can obtain by using it. However, this does not give you a measure of the consistency of the individual measurements and can take a long time. You can use *95% confidence interval to evaluate your data more effectively*:

- If the CI is small compared with the mean, your mean is reliable.
- If the CI is large compared with the mean, the mean is unreliable.

You should consider whether this is due to chance, because of using a small sample, or if it is due to a flaw in the method. If the method is at fault, repeating the same method to get a new set of results is likely to give very inconsistent values for the mean.

If a mean is found to be unreliable, it suggests a new approach is needed. Sometimes increasing the sample size may be enough (which you can test with a new CI calculation).

Evaluating investigations

Using the 95% confidence interval is one of the most useful ways to support a discussion of the reliability and validity of your results. Without *Excel* it is a difficult calculation to carry out. Using *Excel* makes it very easy indeed! It is very helpful to know that most of your results are within a given distance of the value you are measuring.

Reliability refers to repeatability. If your results are consistent you are more likely to have a representative sample that is a true reflection of whatever it is that you are investigating. The 95% confidence interval will be small compared to your mean and the mean is likely to be close to its true value, you can have confidence in it.

Reliability does not necessarily mean that your results and the conclusions are valid, it's just more likely. Valid results are true and fit for their purpose. You could be making a consistent error. Measuring length instead of width might give you very reliable, but invalid values!

A large 95% confidence interval compared with your mean indicates unreliable results. This could be caused by using a poor method or by using a method poorly, causing errors. It may be due to chance effects when using too small a sample. If this is the case, obtaining more results should reduce the confidence interval. Whatever the cause, you need to try to find evidence for it and suggest ways to overcome it.

Graphs

Graphs are important for data analysis. They allow trends and patterns to be observed and identified and show which data are reliable (consistent and fit with the rest).

Excel can be used to plot a variety of graphs, including:

- *scatter graphs* - for investigating an association between two variables
- *bar charts* - for frequency data (counts) or comparing different sets of data
- *pie charts* – for visual comparisons of frequency data (not good for numerical analysis)

Using *Excel* to plot graphs:

- 1 Enter the data into neighbouring columns and click and drag to select them. For *scatter graphs* put the data for the *independent variable* in the *first* column to go on the *x*-axis at the bottom. This is the variable that can be altered to cause a change in the dependent variable, plotted on the *y*-axis up the side. For *bar charts* put the names of the categories in the first column.
- 2 Click on the *chart wizard* icon (or click *Insert* and then click *Chart...* in the menu)
- 3 The wizard has four steps:
 - a Use *Chart Type* to select the type of graph you want:

Usually choose from:

Column for **bar charts** or *XY (Scatter)* for **line** and **scatter graphs** (note: **not** 'Line', which confusingly plots the data against row number). You should normally choose the default option – already highlighted - which does not add lines.

Click to highlight the *Chart sub-type* diagram that you want.

Click *Next*.

- b Look at the sample graph in *Data Range*, if it appears to have used the data correctly, click *Next*.

If it looks wrong, correct it by clicking the *Series* tab, then the red arrow at the end of the X Values box. In the spreadsheet, highlight the cells containing the data you want on the bottom axis of your graph and click the red arrow again.

Repeat for the Y Values box and click *Next*.

- c In the *Chart Options* window, you can type in *Titles* for the graph, label the axes (type the label for your axes in the *Value* boxes – usually *name / units*). You can alter or turn off the gridlines (*Gridlines* tab) and remove the legend (*Legend* tab, then click to remove the tick from the *Show legend* box).
- d In the *Chart Location* window, click *Finish* to put the chart in the spreadsheet with the data. You can click on the chart and then use *File > Print preview* to check the size that the data markers will appear in the printed chart. *File > Print* can be used to print a full page chart, or you can *Copy* the chart and place it into a document.

Notes

- Most things in *Excel* graphs can be changed by right clicking or double clicking on the part that you want to alter.

For example, you can:

- double click on any empty graph area to *Format plot area*
- right click on any empty graph area to get a menu with *Format plot area*, *Chart type*, *Source data*, *Chart options* or *Location*
- click on an axis to *Format Axis*

You can use:

- Format the plot area* to alter borders or background fill. The border and background shading can be removed by clicking *None* in the border and area sections.
 - Format Axis* to alter the *Font* of the label (including the use of superscripts)
 - Scale* to change the range used (e.g. if a large section of the graph is empty)
 - Format Chart Title* or *Format Axis* to change font (size or type and add effects like superscripts)
 - Click or click and drag on text to edit it. A single click will give you a box with handles to move titles. A second click allows you to type and edit. A right click gives you formatting options.
- New graphs using the same layout
- If you alter the data in the spreadsheet, the graph plots (data points) will automatically be altered to fit the new data.