W7 DATA ANALYSIS 2

Drawing Simple Graphs

In some experiments, large amounts of data may be recorded and manipulation is performed using computer software. Although sophisticated, specialist software exists for analysing scientific data, spreadsheet programs suffice for many purposes. This workshop uses MS Excel, as this is the most common spreadsheet program found on home, business and university computers. Instructions are given for (a) the version of Excel in Office 2007 and (b) older versions of Excel.

The data that you will use in this workshop can be downloaded as .csv files from the First Year Chemistry website:

http://firstyear.chem.usyd.edu.au/calculators/data.shtml

Task 1
Open the file "data1.csv" from this webpage and copy and paste the data in the two columns into a fresh worksheet in MS Excel. Figure W7-1 shows a snapshot of what your spreadsheet should look like.

The results of over 500 titrations, showing the frequency with which different titre volumes were measured, is contained in the data.

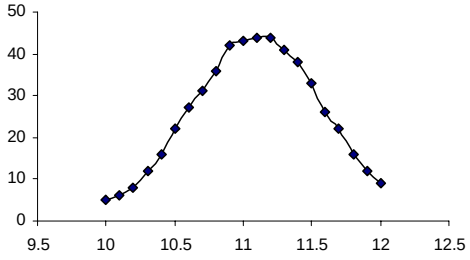
As outlined in W6, experimental data like this usually show the Gaussian distribution. To confirm that this is the case, plot the data as an XY scatter plot:
(i) Highlight the data columns.
(ii) Either:
(a) (Excel 2007) Click on the '**Insert**' tab and in the '**Chart**' group, choose to plot an XY scatter plot, or
(b) (Older versions) From the '**Insert**' menu, select '**Chart**' and then choose to plot the data as an XY scatter plot.

Figure W7-1

M7

	A	B	C
1	Titre range	Frequency	
2	10	5	
3	10.1	6	
4	10.2	8	
5	10.3	12	
6	10.4	16	
7	10.5	22	
8	10.6	27	
9	10.7	31	

Figure W7-2



Your graph should look something like that in Figure W7-2. It shows the expected 'bell shape' of the Gaussian distribution.

Gaussian distributions of data are usually a good indication of random fluctuations in repetitive measurements.

Data Modelling

Many types of measured data can be modelled using simple functions. The three most common mathematical functions found in nature are the straight line, the Gaussian distribution and exponential decay.

The straight line can vary in its *gradient* and its *intercept with the y-axis* and is described by the mathematical function:

$$f(\mathrm{x}) = m\mathrm{x} + c$$

Exponential decay can vary in its *initial value* and in its *rate of decay* and is described by the mathematical function:

$$f(\mathrm{x}) = a\mathrm{e}^{-k\mathrm{x}} + b$$

The Gaussian distribution can vary in its *background value*, its *height*, its *width* and *central value* and is described by the mathematical function:

$$f(\mathrm{x}) = a\mathrm{e}^{-k(\mathrm{x}-x_0)^2} + b$$

Tasks 2 and 3 below illustrate data modelling and fitting for exponential decay. They are applicable to any mathematical function.

Task 2
Open the file "data2.csv" and again copy and paste all of the data into a blank spreadsheet.

The data contains measurements on the concentration of a drug in a patient's blood. Follow the steps in Task 1 to plot a graph showing how the concentration changes with time. The graph should show that the drug decays exponentially with time.

As shown above, exponential decay requires three parameters: *a*, *b* and *k*. Guessed values for these parameters are given in cells F6, F7 and F8 respectively. The next task is to plot the exponential decay curve for these parameters.

In cell C2, type the formula below and then press ENTER.

=$F$6*EXP(-$F$8*A2)+$F$7

This calculates the value of *f*(x) for the exponential decay function using the value in cell A2 for x and the values in cell F6, F7 and F8 for the *a*, *b* and *k* parameters respectively. A value should appear in cell C2.

Now, copy the formula down the entire C column. (You can do this by simply clicking cell C2 and either double clicking on the black square that appears in the bottom right hand corner or dragging the black square down the column.) Values should appear in all of the cells. If you click in any one of them, you should see the formula used to calculate the value in that cell. The use of the **$** signs in the formula ensures that the cells containing the *a*, *b* and *k* parameters are still correctly referenced. Note that the referenced cell in column A has changed from A2 and corresponds to the same line of the spreadsheet.

Give the column of data a title by clicking in cell C1 and typing "Theoretical". Press ENTER.

Now, graph this new column of data on the same chart as follows:
(i) Click on the chart.
(ii) Either:

(a) (Excel 2007) From the '**Chart Tools**' group, choose '**Select Data**'.
Click on the '  ' icon and then select the data in columns A, B and C.
Press ENTER twice.

(b) (Older versions) From the '**Chart**' menu, select '**Add Data**'.
Click on the '  ' icon and then select the data in column C.
Press ENTER twice.

On the spreadsheet, adjust the values of the *a*, *b* and *k* parameters until the theoretical curve matches the experimental curve as closely as possible.

Optimal value of a =

Optimal value of b =

Optimal value of k =

Data Fitting

Modelling in this way produces a reasonable guess for the function parameters. The simplest way of assessing how good a fit it gives is to use the 'mean squared error'. This is related to the standard deviation, but is easier to program.

Task 3
In cell D2, type the formula for the square of the difference between the model (column C) and the experimental data (column B):

=(C2-B2)^2

Again, copy the formula down the entire D column.

At the bottom of the D column (in cell D75), type the following to calculate the average of data in the column:

=AVERAGE(D2:D74)

The closer this value is to zero, the better the fit between the data and the model. Depending on how well you achieved task 2, this value could be large or small.

Mean squared error value = (to 3 significant figures)

The next part of the task uses the 'Solver' facility in Excel to minimise this error by automatically varying the function parameters. To do this:
(i) Click in cell D75.
(ii) Either:
(a) (Excel 2007) Select '**Solver**' from the '**Analysis**' tab in the '**Data**' menu*, or
(b) (Older versions) Select '**Solver**' from the '**Tools**' menu*.
(iii) Choose to minimise the value in cell D75 by selecting 'Min'.
(iv) Choose to vary the parameters (the values in cells F6, F7 and F8) by clicking on the '  ' icon and selecting the relevant cells in column F.
(v) Click 'Solve'.
(vi) Step (v) may need repeating several times to achieve the best fit.

Mean squared error value after minimisation =

Optimal value of a =

Optimal value of b =

Optimal value of k =

Linear Relationships

The most commonly used function is actually the straight line. Data that do not give a straight line are often manipulated to produce one, thus making the fitting process easier.

Task 4
Open the file "data3.csv" and again copy and paste all of the data into a blank spreadsheet.

The data contains measurements of how the rate of a reaction changes with temperature. Empirically the rate, measured using the ***rate constant***, ***k***, is related to the temperature through the **Arrhenius Equation**,

$$k = A\mathrm{e}^{-E_\mathrm{a}/RT}$$

where A is called the 'frequency factor', E_a is the 'activation energy' and R is the gas constant (R = 8.314 J K^{-1} mol^{-1}. A is related to how often molecules collide and E_a is related to the energy the molecules need to break the bonds etc that are required to start the reaction. Taking natural logs on both sides gives,

$$\ln k = \ln A - \frac{E_\mathrm{a}}{RT}$$

This has the form of a straight line: if $\ln k$ is plotted as the y-axis and $^1/_T$ is plotted as the x-axis, the gradient is $^{-E_\mathrm{a}}/_R$ and the y-intercept is $\ln A$.

* If '**Solver**' does not appear in the menu, either
(a) (Excel 2007) Click on the '**Office Button**' in the top left hand corner, follow '**Excel Options**' | '**Add-Ins**', select '**Excel Add-ins**' from the '**Manage**' drop down, click on "**Go...**" and select '**Solver Add-in**', or
(b) (Older versions) Click '**Add-Ins**' in the '**Tools**' menu and then tick the '**Solver Add-in**' box.

In cell C2, type the formula:

=1/A2

Again, copy the formula down the entire C column. This puts the $^1/_T$ values in column C. These form the values on the *x*-axis of the plot.

In cell D2, type the formula:

=ln(B2)

Again, copy the formula down the entire D column. This puts the $\ln k$ values in column D. These form the values of the *y*-axis of the plot.

Draw an XY scatter graph of $\ln k$ *vs* $^1/_T$.

Perform a regression analysis on the data as follows:
(i) Either
  (a) (Excel 2007) From the '**Data**' menu† in the '**Analysis**' tab, select '**Data Analysis**' and then choose '**Regression**', or
  (b) (Older versions) From the '**Tools**' menu†, select '**Data Analysis**' and then choose '**Regression**'.

(ii) By clicking on the '  ' icons, select the data in column C for the *x*-values and the data in column D for the *y*-values.

The regression analysis output provides the slope ("X Variable") and intercept ("Intercept") with ***standard errors***. Data lying within two standard errors can be used with confidence.

Record the slope and y-intercept of the straight line and then calculate the value of E_a.

Slope = Intercept =

Activation energy = J mol^{-1}

Demonstrator's Initials	

† If '**Data Analysis**' does not appear in the menu, either
  (a) (Excel 2007) Click on the '**Office Button**' in the top left hand corner, follow '**Excel Options**' | '**Add-Ins**', select '**Excel Add-ins**' from the '**Manage**' drop down, click on "**Go...**" and select '**Analysis ToolPak**', or
  (b) (Older versions) click '**Add-Ins**' in the '**Tools**' menu and then tick the '**Analysis ToolPak**' box.