1B40 Practical Skills

Weighted mean

The normal (Gaussian) distribution with a true mean μ and standard deviation σ is

$$p(x) = \frac{1}{s\sqrt{2p}} \exp\left[-\frac{(x-m)^2}{2s^2} \right].$$

The probability of occurrence of a value x_1 is $p(x_1)$. Hence the probability P of obtaining the values $x_1, x_2, x_3, \ldots, x_n$ is

$$P(x_1, x_2, \ldots x_n) = p(x_1)\,p(x_2)\,p(x_3)\cdots p(x_n).$$

{Strictly there should be a factor 1/n! on the R.H.S as the order is irrelevant, but it can be omitted as we are only interested in the variation of *P* with its parameters.}

Thus explicitly

$$P = \left(\frac{1}{s\sqrt{2p}} \right)^n \exp\left[-\frac{\sum_{i=1}^{n} (x-m)^2}{2s^2} \right].$$

It is reasonable to assume that this should be a maximum. (**Principle of Maximum Likelihood**). This probability is a maximum when $\sum_{i=1}^{n} (x_i - m)^2$ is a minimum. This idea leads to the **Principle of Least Squares** which may be expressed as follows:

- the most probable value of any observed quantity is such that the sum of the squares of the deviations of the observations from this value is the least.

The quantity $\sum_{i=1}^{n} (x_i - l)^2$ has a minimum value when $l = \frac{1}{n}\sum_{i=1}^{n} x_i$, i.e. the mean. This follows from

$$\sum_{i=1}^{n} (x_i - l)^2 = \sum_{i=1}^{n} x_i^2 - 2l\sum_{i=1}^{n} x_i + nl^2,$$

$$\frac{d}{dl}\left(\sum_{i=1}^{n} (x_i - l)^2 \right) = -2\sum_{i=1}^{n} x_i + 2nl = 0.$$

Hence these principles lead to the often quoted result that the best estimate for μ is the arithmetic mean.

It may be, of course, that the $x_1, x_2, x_3, \ldots, x_n$ belong to different Gaussian distributions with different standard deviations. The total probability would then be

$$P = \frac{1}{\sqrt{2p}} \left(\frac{1}{s_1} \right) \left(\frac{1}{s_2} \right) \cdots \left(\frac{1}{s_n} \right) \exp\left[-\sum_{i=1}^{n} \frac{(x-m)^2}{2s_i^2} \right].$$

This will be greatest when $\sum_{i=1}^{n}\frac{\left(x_i - m\right)^2}{2s_i^2}$ is a minimum. This occurs when μ is given by the **weighted mean**

$$\overline{x} = \frac{\sum_{i=1}^{n}\frac{x_i}{s_i^2}}{\sum_{i=1}^{n}\frac{1}{s_i^2}}.$$

In general if a measurement x_i has weight w_i then the weighted mean is

$$\overline{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}.$$

The standard deviation of the weighted mean is

$$s^2 = \frac{\sum_{i=1}^{n} w_i\left(x_i - \overline{x}\right)^2}{\sum_{i=1}^{n} w_i}$$

These expressions reduce to those given earlier for the unweighted quantities if we put $w_i = 1$ for all measurements.

For the case of only two quantities,

$$\overline{x} = \frac{\frac{x_1}{s_1^2} + \frac{x_2}{s_2^2}}{\frac{1}{s_1^2} + \frac{1}{s_2^2}},$$

and using the formula for the propagation of errors on a sum of two quantities gives

$$\frac{1}{s^2} = \frac{1}{s_1^2} + \frac{1}{s_2^2}.$$

Curve fitting

We can apply the principle of least squares to the problem of fitting a theoretical formula to a set of experimental points. The simplest case is that of a straight line.

The Straight Line

In many experiments it is convenient to express the relationship between the variables in the form of the equation of a straight line i.e.

$$y = mx + c,$$

where m, the gradient of the line, and c, the intercept at x = 0, are treated as unknown parameters.

As an example, consider the compound pendulum experiment where the relationship between the period T and the adjustable parameter h is given by

$$T = 2p\sqrt{\frac{h}{g}+\frac{k^2}{gh}},$$

and the quantities h and k are defined in the script for the experiment.
Plotting T against h would yield a complicated curve which is difficult to analyse. However the relationship can be expressed as

$$T^2h = \frac{4p^2}{g}h^2 + \frac{4p^2k^2}{g}.$$

If T^2h is plotted against h^2 a straight line is expected. The benefits of this are

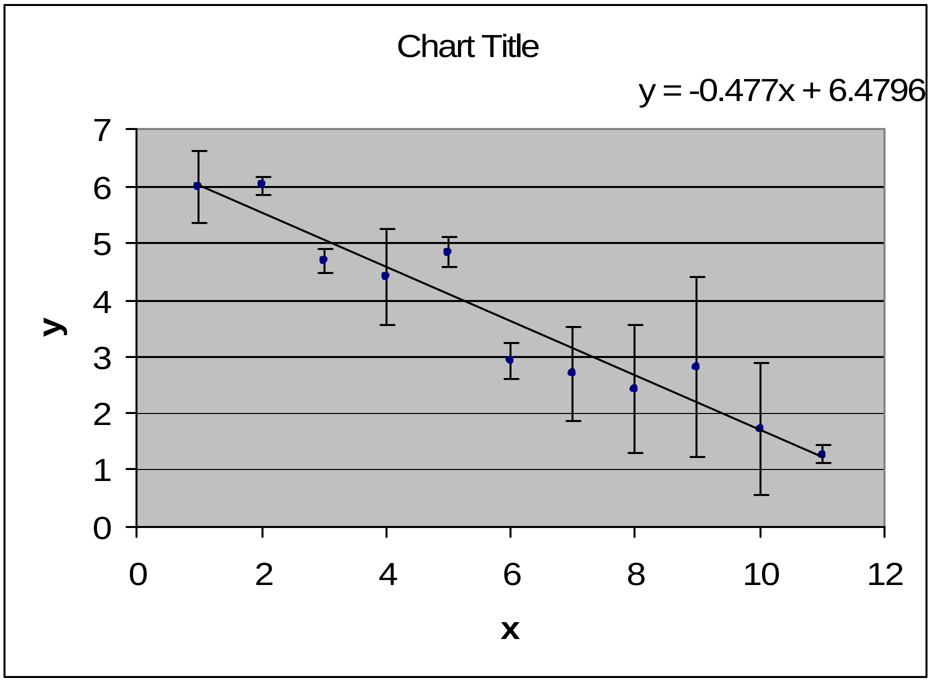
1. Results expressed as a linear graph have a satisfying immediacy of impact.
2. It is very easy to see if a set of results is progressively deviating from linearity - much easier than say detecting deviation from, for example, a parabola.
3. The line of best fit to a set of points with error bars is easy to estimate approximately by eye, and to insert with the aid of a ruler as a quick check. .
4. A mathematical method exists to calculate the line of best fit, which relates simply to the statistical consideration of errors.

Best straight line fit to a linear curve using the method of least squares (Legendre 1806)

We consider fitting a function of the form

$$y = mx + c$$

to a set of data points. The example is simple because it is *linear* in m and c *not* because y is linear in x. The data consists of the points $(x_i,\ y_i \pm s_i)$. That is we assume that the x-coordinates are known exactly as they usually correspond to the independent variable (the one under the experimenter's control), but there is an uncertainty s_i in the y-coordinate corresponding to the dependent variable that is "measured". The deviation, d_i, of each point from the straight line is taken only in the y-coordinate, $d_i = y_i - (mx_i + c)$.



Method I – Points with associated error bars

According to the principle of least squares we have to minimise

$$S=\sum_{i=1}^{n}\frac{d_i^2}{s_i^2}=\sum_{i=1}^{n}\left(\frac{y_i-mx_i-c}{s_i}\right)^2. \qquad (1.1)$$

On differentiating with respect to *m* and *c* in turn we get

$$\begin{aligned}-2\sum_{i=1}^{n}\frac{\left(y_i-mx_i-c\right)x_i}{s_i^2}&=0,\\ -2\sum_{i=1}^{n}\frac{\left(y_i-mx_i-c\right)}{s_i^2}&=0.\end{aligned} \qquad (1.2)$$

Expanding these we have

$$\begin{aligned}\sum_{i=1}^{n}\frac{x_iy_i}{s_i^2}&=m\sum_{i=1}^{n}\frac{x_i^2}{s_i^2}+c\sum_{i=1}^{n}\frac{x_i}{s_i^2},\\ \sum_{i=1}^{n}\frac{y_i}{s_i^2}&=m\sum_{i=1}^{n}\frac{x_i}{s_i^2}+c\sum_{i=1}^{n}\frac{1}{s_i^2}.\end{aligned} \qquad (1.3)$$

The last one, on dividing through by $\sum_{i=1}^{n}\frac{1}{s_i^2}$ becomes

$$\frac{\sum_{i=1}^{n}\frac{y_i}{s_i^2}}{\sum_{i=1}^{n}\frac{1}{s_i^2}} = m\frac{\sum_{i=1}^{n}\frac{x_i}{s_i^2}}{\sum_{i=1}^{n}\frac{1}{s_i^2}} + c$$

$$\overline{y} = m\overline{x} + c.$$

This shows that the best fit line passes through the weighted mean point $\left(\overline{x},\overline{y}\right)$ even if this does not correspond to an actual measured point.

The eqns (1.3) are two simultaneous equations for the two unknown, *m* and *c*. Their solution is

$$\begin{aligned} m &= \frac{[1][xy]-[x][y]}{[1][xx]-[x][x]} = \frac{[1][xy]-[x][y]}{D}, \\ c &= \frac{[y][xx]-[x][xy]}{[1][xx]-[x][x]} = \frac{[y][xx]-[x][xy]}{D}, \end{aligned} \qquad (1.4)$$

where

$$D = [1][xx]-[x][x], \qquad (1.5)$$

and the quantities in the square brackets [ ] are defined by

$$[1] = \sum_{i=1}^{n}\frac{1}{s_i^2}; \quad [x] = \sum_{i=1}^{n}\frac{x_i}{s_i^2}; \quad [y] = \sum_{i=1}^{n}\frac{y}{s_i^2}; \quad [xy] = \sum_{i=1}^{n}\frac{x_i y_i}{s_i^2}; \quad [xx] = \sum_{i=1}^{n}\frac{x_i^2}{s_i^2}. \qquad (1.6)$$

The calculation of the errors on the fitted parameters, *m* and *c*, is intricate and is done best by techniques that are beyond this introductory course (matrix methods). We simply quote the results,

$$\begin{aligned} s_m^2 &= (dm)^2 = \frac{[1]}{[1][xx]-[x][x]} = \frac{[1]}{D}, \\ s_c^2 &= (dc)^2 = \frac{[xx]}{[1][xx]-[x][x]} = \frac{[xx]}{D}. \end{aligned} \qquad (1.7)$$

These expressions may look complicated but they are easily evaluated in a computer programme or a spreadsheet. They only involve sums of terms.

Method II – no estimate of error on y

If the errors on the data points are not known we can only minimise

$$S = \sum_{i=1}^{n}\left(y_i - mx_i - c\right)^2. \qquad (1.8)$$

This is equivalent to the previous case if we set all the errors $s_i = 1$. Thus we can get the results immediately for

$$\begin{aligned} m &= \frac{n[xy]-[x][y]}{n[xx]-[x][x]}, \\ c &= \frac{[y][xx]-[x][xy]}{n[xx]-[x][x]}. \end{aligned} \tag{1.9}$$

In this case the only way to estimate the uncertainties in m and c is to use the scatter of the points about the fitted line. The mean square error s^2 in the residuals $d_i = y_i - (mx_i + c)$ is given by

$$s^2 = \frac{\sum_{i=1}^{n} d_i^2}{n-2} = \frac{S_{\min}}{n-2}. \tag{1.10}$$

{The n -2 occurs because we have only n -2 independent points, two being needed to find the slope and intercept of the line}. We then estimate the errors

$$\begin{aligned} s_m^2 &= (dm)^2 = \frac{n}{n[xx]-[x][x]} s^2 = \frac{n}{D} s^2, \\ s_c^2 &= (dc)^2 = \frac{[xx]}{n[xx]-[x][x]} s^2 = \frac{[xx]}{D} s^2, \end{aligned} \tag{1.11}$$

where

$$D = n[xx]-[x][x]. \tag{1.12}$$

Correlation in least squares fits

The original measurements x_i and y_i may be uncorrelated but the values of m and c found by both methods are correlated since they depend on the same data - the (x_i, y_i) values. The best fit line passes through the fixed point $(\bar{x}, \bar{y})$. If $\bar{x} > 0$ and the gradient is increased by its error the intercept decreases as the line pivots about the point $(\bar{x}, \bar{y})$, and vice versa. A formula for the covariance can be derived. For the weighted fits,

$$\text{cov}(m, c) = s_{mc}^2 = -\frac{[x]}{D},$$

and for the unweighted ones

$$\text{cov}(m, c) = s_{mc}^2 = -\frac{[x]}{D} s^2.$$

As an illustration, a fit to some data gave the following weighted fit parameters: $m = -0.433$, $c = 6.189$, $s_m = 0.057$, $s_c = 0.297$, $\text{cov}(m, c) = -0.014$. The table shows the values of y_i predicted and the estimated uncertainty for chosen x_i.

x	predicted y	y error with correlation	y error without correlation
-10	10.5	0.8	0.6
40	-11.1	2.0	2.3

The errors with correlation included may be smaller or larger than those calculated without it!

The table below compares the advantages and disadvantages of Method I and Method II.

	Method I	Method II
Data points with big errors	are essentially ignored in the fit	are treated like those points with small errors
Errors on m and c	are realistic in terms of the statistics of the data, s_i and n	can be (unfortunately) small if points happen to lie well on a straight line
If the data don't really lie on a straight line	the errors on m and c may be ridiculously small if statistics are large	errors will be larger
Number of data points needed to estimate m, c and errors	2	3
Can goodness of fit be tested?	Yes	No
Can method be used if s_i are unknown?	No	Yes

Excel implementation of unweighted fit

Method II is implemented in the Excel spreadsheet function LINEST. If the array formula (see Excel for ways to enter an array formula)

`=LINEST(range of known y's, range of known x's, , true)`

is entered into an array of 5 rows and 2 columns, then it returns the following results, (The words have been added to explain the quantities computed).

	LINEST		
m	-0.3961	5.8665	c
error on m	0.0465	0.3151	error on c
r^2	0.8898	0.4873	standard error on y
F	72.6618	9	number degrees freedom
regression sum of squares	17.2568	2.1374	sum of squares of residuals

Thus this example describes the line

$$y = -(0.40 \pm 0.05) + (5.9 \pm 0.3).$$