DATA FITTING TECHNIQUES

CONTENTS OF THIS SECTION.

DATA FITTING TECHNIQUES

I. GRAPHICAL ANALYSIS

"*A picture is worth a thousand words*" is particularly true in the analysis of laboratory data. In many experiments it is very useful to check on the experiment by making a rough graph of the data points as they are being taken. Carefully drawn and analysed graphs are often the best way to extract the values of the unknowns and to confirm or deny the theoretical hypothesis being tested. Finally graphs are important in presenting results clearly and pleasingly in a report.

Straight line graphs have a special significance. They are the easiest to draw by hand. In addition, their interpretation is often simplest, since all straight lines plotted on an x-y graph have the form

$$y = mx + b \qquad (1)$$

where m is the slope of the line (the tangent of the angle between the line and the x-axis) and b is the intercept on the y-axis. Usually x and y depend only on the known or measured quantities while m and b contain the unknowns which you are trying to find.

Though a plot of your data points in their raw form will not generally lie on a straight line, you will find that data can usually be *made* to fit a straight line; if there are two and only two unknowns in your experiment it is often worth trying to put the theoretical relationship into the form of equation (1). Don't be discouraged if x, y, m, b turn out to be quite complicated expressions: this may be correct (see the *Example* below).

In graphs which are intended to do more than simply indicate a trend in the data, you need to indicate the uncertainty in each point. This is done by drawing "error bars" - lines corresponding to the size of the error on either side of the data point - vertically for errors in y values and horizontally for errors in x values (see Figures 3 and 4 below).

II. HAND-DRAWING OF GRAPHS.

II.1 Fitting a Straight Line.

When you are attempting to fit a straight line to some data points by hand drawing, choose a line which encompasses a majority of points and passes as close to all of them as possible, taking into account the errors (indicated by "error bars") which indicate the uncertainty in the position of the points. This "best" line is a graphical estimate of the "average" of the data; by seeing how far on either side of this "best" line you can draw other acceptable lines, you will be able to form a common-sense estimate of the error in the slope and the intercept of this line (see Figures 3 and 4). Note that **all** points need not lie on the line; remember that a scatter of points about a theoretical curve is to be expected in the real world. However, deliberately ignoring a few points (without a good reason) because they spoil a "perfect" fit to the other points is very bad science, and actually almost fraud (see section III.3 - **Rejection of Measurements**, in **EXPERIMENTAL ERRORS**)

Tips:

- use a sharp pencil or fine pen to draw your graphs.
- plot the points in pen and the smooth curve in pencil so that you may easily redraw your fitted curve.
- spread the data nicely across the available area -*i.e.*, by choosing the scales appropriately on the two axes, which may or may not include the origin at (0,0).
- avoid the use of an unnatural subdivision of intervals; use multiples of 2,5, or 10 whenever possible (not 3!!).

Example

A "compound" pendulum swings with period **T**, which is related to its unknown radius of gyration **k**, and a measured length **L** by the equation

$$T = 2\pi \sqrt{\frac{L^2 + k^2}{Lg}} \tag{2}$$

where **g** is the gravitational acceleration. The parameters **g** and **k** are to be determined. Note that (2) is not in the form of (1); a plot of **T** *versus* **L** is not a straight line (Figure 3). However, if (2) is squared and then multiplied through by **L**, it becomes

$$T^2 L = \frac{4\pi^2 L^2}{g} + \frac{4\pi^2 k^2}{g} \tag{3}$$

which *is* in the form of (1); a plot of $\mathbf{T^2L}$ *versus* $\mathbf{L^2}$ is a straight line (Figure 4). There are several other ways of putting (2) into the general form of (1). Can you see what these are?

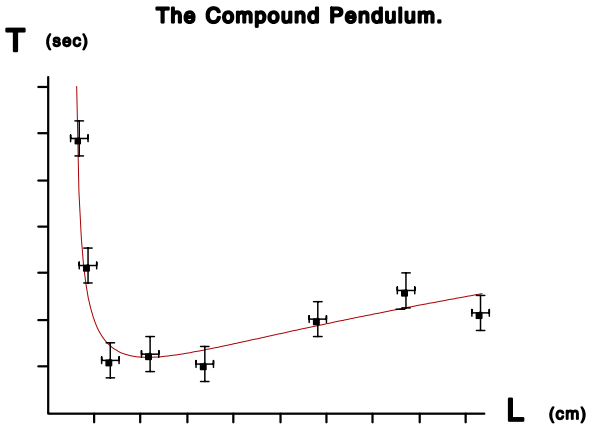

Figure 3. Note that the errors in the measured quantities **T** and **L** are represented graphically by vertical and horizontal lines respectively. These are called "error bars"

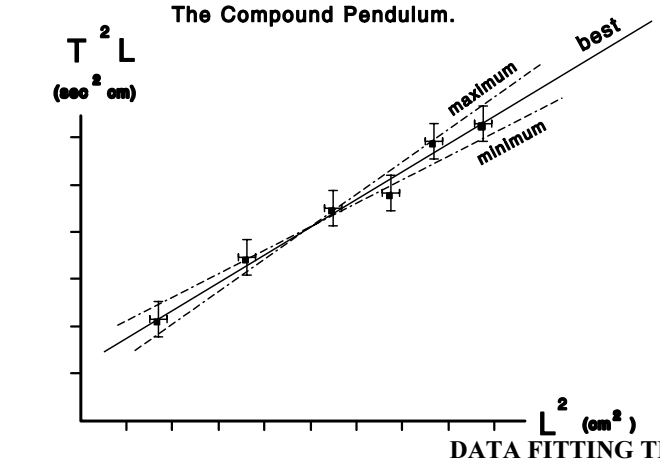

Figure 4. The solid line represents the "best" fit to the data points, in the experimenter's judgement; the dotted lines are the best guess for the maximum and minimum slopes consistent with the data.

III.2 The High Accuracy trick.

It may be that you have some data which are so accurate that you cannot do them justice on any available graph sheets - although some quite large sheets are available at the wicket in room 126. In this case you may have to resort to numerical computation. Often, however, the following trick will still enable you to get a graphical result.

This trick, useful in plotting linearly varying data of high precision, permits you to subtract off the major linearly varying part of the data you wish to investigate graphically, so that you can see small fluctuations in the data by plotting it on an expanded scale. In this procedure, in which the data follows the law

$$y = mx + b$$

you make an estimate for $m = m_e$, and then plot $(y - m_e x)$ versus x. As the values of $(y - m_e x)$ cover a very restricted range, this plot can be made to be highly expanded, and you can then determine, to high accuracy, the slope of this graph which we shall call m_{ha}.
Thus

$$(y - m_e x) = m_{ha} x + b \qquad (4)$$

So

$$m = m_e + m_{ha} \qquad (5)$$

To summarize in steps:

Step 1. - We assume that your tabulated values of x and y (x_i, y_i) fit a relation of the form of Equation (1) and that the x_i and y_i values have better-than-graph paper accuracy. Your first plot presumably gave a perfectly straight looking line even with a razor sharp pencil.

Step 2. - Estimate the slope m from your plot and call this estimate m_e.

Step 3. - Subtract $m_e x_i$ from each value of y_i so as to make a new column of data δy_i
i.e. $\delta y_i = y_i - m_e x_i$.

Step 4. - Plot δy_i against x_i on a suitably expanded scale. The slope of your new plot is the correction to m_e which gives you a more accurate value.

There is no experiment in the first year lab which is so accurate that the graphical analysis cannot be made in this way to reveal the deviations (random or otherwise) from the straight line relationship.

Some of your graphical fitting of data in the lab will be done using numerical fitting methods on a computer (see two sections ahead). Such techniques find slopes and intercepts calculated to many figures using numerical calculations. When such numerical methods are used the high accuracy trick is of less importance.

III. USING THE COMPUTER TO FIT GRAPHS.

CAUTION! *Although the computer provides a very powerful method of fitting and analysing your data, it will not evaluate your experiment, tell you about the appropriateness of your means of fitting data nor generally provide judgement on how well the performance of your experiment is going. It is up to you to check carefully how the line visually fits the displayed data and to ensure that it makes sense. If your choice of errors is poor or non-existent, the computer's calculated error in the slope and intercept will be equally flawed.* ***The thinking part of the experiment is still your responsibility****! (see the notes on the* χ^2 *of the fit in this section).*

III.1 Plotting Graphs

Although the computer has a data-plotting utility called *graph*, if you wish to plot your data points and nothing more, our advice to you is don't do it on the computer. You will learn much more about what your data looks like and how your experiment is behaving if you plot by hand. However, if you want to fit your data to a theoretical expression and extract physical values from the plot, as is almost always the case, use the utility called *fit* described below.

III.2 Fitting a Curve (Line) To Your Data

The lab computer provides an "objective" procedure for giving a best fit of a straight line or a specified polynomial to your data, along with correctly calculated error estimates on that fit. It uses the programme called *Mathematica*. Basically the computer calculates the square of the distance between your data points and the curve it is trying to find; it then adjusts the constants of the polynomial representing the curve until the sum of these squares is a minimum. It gives you the values of the constants so found, along with their errors (one standard error). In addition it draws a graph which shows your data points and the fitted curve. If the errors in your data points are large enough to be visible on the graph, the computer also draws two lines which correspond to the "maximum" and "minimum" lines which would appear in a hand-drawn graph (see **HAND-DRAWING OF GRAPHS** above); however the computer's lines are correctly calculated to be ± one standard deviation from the fitted line. The graph itself is an important visual aid to determine the reasonableness of your fit.

The programme is menu-driven, and mostly self-explanatory. To enter your data, choose the *data* menu (type *d*), then *create* (type *c*). The computer will ask you to enter the names of the variables you will enter. This refers to the mathematical symbols you want to use to specify your data.

READ THE FOLLOWING TIPS BEFORE YOU USE THE COMPUTER FOR THE FIRST TIME.

Tips:

- your variable names should be short mnemonics, in lower case; *e.g.* **s,t,vc** etc., rather than **distance, time, voltage** (the less typing, the less chance of error; *Mathematica* functions start with an upper case letter).
- always enter your **raw** data, just as you have taken it (the computer will do any calculations on it that you need - see below).
- as long as the errors on your data are *either* constants (*e.g.* 0.005, 0.3, etc.) *or* functions of the variables (*e.g.* $0.005/t^2$, 2s/t, etc.) you do NOT need to enter them (you can enter them at the time of setting up the graph). Otherwise, however, you need to define a variable name for your errors and enter the numerical value for each point.
- once you are satisfied that your data is properly entered, **back it up!** This is done by accessing the menu item *backup* and following the instructions. Your data will now remain in your own directory until you overwrite it; it can be accessed at any later time by invoking *restore.*

Once your data has been entered, choose *analyse* then *fit*. The *fit* window has a table which contains the variable names you have defined. To proceed, you must specify a *dependent variable* (which means, for the computer, that it will appear on the y-axis) and an *independent variable* (x-axis); make your own choice depending on which variable you want to appear on the y-axis (these won't necessarily have anything to do with which variable was dependent or independent in your actual experiment). You also need to define the errors in these quantities. If you want to plot the variables and errors just as they were defined, choose the appropriate "buttons"; if you want to manipulate the data somewhat (or, *e.g.* insert a constant value for an error) choose the button which reads *An expression* - you can then enter a constant value or a formula to calculate the variable you want to use.

Notes on calculations using Mathematica.

Some of the grammar you may need when you calculate mathematical expressions involving your variables is listed in the table opposite; (you can probably guess at any others you may need - see also "**Mathematica"**, listed in **References**).

N.B. Use "(" and ")" as parenthesis in your expressions; "[" and "]" are reserved for the arguments of *Mathematica* functions. Angles are measured in radians.

Algebraic Expression	*Mathematica* Expression
xy	x*y OR x y
$1/x$	1/x
x^n	x^n
$\sqrt{x}$	Sqrt[x]
$\ln(x)$	Log[x]
e^x	Exp[x]
$\log_{10}(x)$	Log[10,x]
$\cos(x)$	Cos[x]
$\sin^{-1}(x)$	ArcSin[x]
$\lvert x \rvert$	Abs[x]

To all data, the computer fits a general polynomial which has the form:

$$y = A(0) + A(1)x + A(2)x^2 + A(3)x^3 + \ldots$$

You need to specify which of the terms in the polynomial are relevant to the fit of your data; this is done by selecting the appropriate "buttons" in the window from the set:

0 **1** **2** **3** **4** **5** **6** **7** **8** **9**

to choose one or more of the coefficients

A(0) *A(1)* *A(2)* *A(3)* *A(4)* *A(5)* *A(6)* *A(7)* *A(8)* *A(9)*

(Often you will be specifying 0 and 1, which is a fit to the straight line $y = A(0) + A(1)x$ with $A(1)$ = slope and $A(0)$ = intercept on the y axis). Follow the on-screen menu to try another fit, or re-plot the graph with a variety of options, or *print* the graph on the lab printer (near the **R**esource **C**entre).

III.3 Other Useful Features of the Faraday Analysis Programmes.

Although it may not be obvious when you access *fit* or *graph* from the *analyse* menu, both programs are actually running in the Netscape web browser. You may access the program from any web browser, and have access to all data sets that you created from the *data* menu. Choose the **Data** choice from the URL http://www.upscale.utoronto.ca. If you are using a PC to run your browser, you may also upload files created on the PC and fit or graph them using the same program. Consult the on-line help for the programs for further information.

Some other tips about our analysis programs include:

1. When you use a menu item, you need usually type in only enough letters to uniquely identify the word; most of the time only the first letter is sufficient.

2. Once you have used *create* to enter your data, you can always look at it by choosing the menu item *show* which is available on the *data* and the *analyse* menus (the latter also gives you the option to print the data file).

3. There are several ways in which you can change your data after you have created it. You should almost never have to type in the full set of data again, since these programmes will make almost all the corrections you want.

A) *edit*. This is the simplest of these programmes, and should be used only for simple fixes. You can move around your data file, deleting, changing and adding data. To exit "*edit*" , click on *File* (top left corner of the screen) to pull down the menu there. Choose the *End* option; you will then be prompted to *Save* before exiting. **CAUTION**! Occasionally you will get a message which reads *Wrong columns or non-numbers found* ; this often means that you have left the cursor sitting on a new line instead of exactly at the end of your data. Move the cursor to the end of the file and use the Backspace button to delete any extra spaces or line returns till the cursor is sitting immediately after the last digit of your last data point. Then try exiting again.

B) *massage*. You will find this programme in the *analyse* menu. Clicking on *massage* will lead you to another menu with several options. The two which you will find most useful are now discussed.

C) recalc. This option presents you with the opportunity to add variables to your data set which are either constants or functions of the ones you have already defined. For example, in the Boyle's Law experiment you may mistakenly have added mm of mercury (from the barometer reading) to cm of mercury (manometer reading), and all your pressure data has to be changed. Or you may have a very complicated expression to calculate from the raw data which you have entered in the file (the Flywheel experiment is a good example). You are asked how many new variables you want to define, their names, and how the computer is to calculate them from the data already in your file.

NOTE! You are first asked if you want to keep the original data - it is usually a good idea to answer *yes* to this question.

D) addvar. You may decide that you want to add the values of a variable which you had forgotten to enter the first time. This option will ask you how many variables you want to add, and their names. It will present you, a line at a time, with the data that is already in your file. You just have to type in the new values. The programme will exit automatically when you reach the last line.

III.4 The Chi-Squared (χ^2) test

Along with the fitted parameters, *A(0), A(1), A(2) etc.* (let us call the number of these parameters **m**) a value called *Chi-Squared* is printed at the top of your graph (the Greek symbol is χ, pronounced like "*cry*", without the "*r*"). This is a statistical quantity which tells you something about the quality of your fit. (You will come across the quantity χ^2 in procedures in which averages are derived from data or in which curves are fitted to data, curve fitting being just another form of averaging.) For the case of the curve fitting described above, the value is calculated by taking the sum of the squares of the deviation of each of the **n** data points from the fitted curve and dividing by the square of the error (= standard deviation) in each point;

i.e

$$\chi^2 \equiv \sum_{1}^{n} \left[\frac{\text{observed value - BOLDexpected value}}{\text{standard deviation}} \right]^2 \qquad (6)$$

PERCENTILE VALUES FOR THE CHI-SQUARED DISTRIBUTION WITH v DEGREES OF FREEDOM

v	99.5 %	99%	97.5 %	95%	90%	75%	50%	25%	10%	5%	2.5%	1%	0.5%
1	7.88	6.63	5.02	3.84	2.71	1.32	0.46	0.10	0.02	0.00	0.00	0.00	0.00
2	10.6	9.21	7.38	5.99	4.61	2.77	1.39	0.58	0.21	0.10	0.05	0.02	0.01
3	12.8	11.3	9.35	7.81	6.25	4.11	2.37	1.21	0.58	0.35	0.22	0.16	0.07
4	14.9	13.3	11.1	9.49	7.78	5.39	3.36	1.92	1.06	0.71	0.48	0.30	0.21
5	16.7	15.1	12.8	11.1	9.24	6.63	4.35	2.67	1.61	1.15	0.83	0.56	0.41
6	18.5	16.8	14.4	12.6	10.6	7.84	5.35	3.45	2.20	1.64	1.24	0,87	0.68
7	20.3	18.5	16.0	14.1	12.0	9.04	6.35	4.25	2.83	2.17	1.69	1.24	0.99
8	22.0	20.1	17.5	15.5	13.4	10.2	7.34	5.07	3.49	2.73	2.18	1.65	1.34
9	23.6	21.7	19.0	16.9	14.7	11.4	8.34	5.90	4.17	3.33	2.70	2.09	1.73
10	25.2	23.2	20.5	18.3	16.0	12.5	9.34	6.74	4.87	3.94	3.25	2.56	2.16
11	26.8	24.7	21.9	19.7	17.3	13.7	10.3	7.58	5.58	4.57	3.82	3.05	2.60
12	28.3	26.2	23.3	21.0	18.5	14.8	11.3	8.44	6.30	5.23	4.40	3.57	3.07
13	29.8	27.7	24.7	22.4	19.8	16.0	12.3	9.30	7.04	5.89	5.01	4.11	3.57
14	31.3	29.1	26.1	23.7	21.1	17.1	13.3	10.2	7.79	6.57	5.63	4.66	4.07
15	32.8	30.6	27.5	25.0	22.3	18.2	14.3	11.0	8.55	7.26	6.26	5.23	4.60

The size of this quantity depends on how well the curve fits the data, taking into account the fact that if the errors are large, a greater degree of deviation is expected. Its value will obviously depend also on the number of terms in the sum, equal to the number of data points. For a good fit, we might expect that, on average, each point might deviate from the fitted curve by approximately the value of the stated error - *i.e.* the sum in equation (14) would, on average, have a value approximately equal to the number of data points, or approximately the number of degrees of freedom, if the number of extracted parameters is small.

If we have overestimated the errors, χ^2 would be too low; if even some of the points deviate from the fitted curve by more than we expect from the size of the errors, χ^2 will be too high. In fact, if we define the *degrees of freedom*, **v**, as $\mathbf{v = n - m}$, it can be shown that, on average, the "expected value" of χ^2 is equal to **v**; if your value of χ^2 deviates greatly from **v**, you will already begin to suspect the goodness of your fit to the data.

This qualitative understanding can be made quantitative; χ^2 has a statistical distribution which is well-known. The calculated values in the Chi-Square Distribution table give the probability that, in a random process, a given value of χ^2 would be smaller than the value that you actually obtained. If this probability is very large (*i.e.*, hardly anyone would get a larger value than you) we conclude that the hypothesis is probably incorrect. As you might guess a too-small value of χ^2 should also be questioned. Why?

III.5 Confidence Levels

Suppose that the result of a measurement of some quantity x yields a result of $\bar{x}$. We have already seen that this means that there is a 68% chance that the mean of x lies between the values $\bar{x} - \boldsymbol{\sigma_m}$ and $\bar{x} + \boldsymbol{\sigma_m}$, a 95% chance that it lies between the values $\bar{x} - \mathbf{2}\boldsymbol{\sigma_m}$ and $\bar{x} + \mathbf{2}\boldsymbol{\sigma_m}$; we could also say *e.g.*

"*with 99% confidence, the mean of* $\bar{x}$ *has a value lying between* $\bar{x} - \mathbf{3}\boldsymbol{\sigma_m}$ *and* $\bar{x} + \mathbf{3}\boldsymbol{\sigma_m}$ ".

In the same way the χ^2 test can give an idea of our confidence in a given result. If the value of the probability obtained from the χ^2 table is greater than 0.95 (95% confidence level) - *i.e.* that in only 5% of the random samples from the population in question would the value of χ^2 be greater than the one actually found - we say that our hypothesis (that the fit is a good one) is rejected at the 5% significance level and we should then regard our particular hypothesis regarding the data as being suspect. Values at the 1% level of probability are said to be highly significant. Note that the χ^2 test is a negative test and only tells us if our hypothesis regarding the data distribution is incorrect.

Examples:

Suppose a fit gives a χ^2 of 4.2 for 6 degrees of freedom; from the table we see that this value corresponds to a probability somewhere between 0.25 and 0.5, which would be a quite acceptable fit.

If, on the other hand, the χ^2 was 17.3, we should reject the validity of the fit *at the 1% confidence level*; in this case we might start by checking that we had not made a mistake in measuring one or more of our data points and by ensuring that our estimate of error was sufficiently large. If, on the other hand, a value for χ^2 for this fit was 0.71, we would be concerned that the fit was "too good", and suspect that we had over-estimated the errors. Once the data has been satisfactorily completed, a value of χ^2 giving a probability outside the 5% to 95% limits may indicate that the theoretical hypothesis concerning the data is suspect.