Some Notes on the Statistical Treatment of Data

StatOnly2c.doc 10/10/2002 6:14 PM

by Prof. Keith Ruddick

Introduction

As a general rule, a result quoted without any estimate of its uncertainty is useless. Its validity cannot be considered in any absolute way, or compared with other equivalent results. The purpose of these notes is to present an introduction to some of the techniques of error (uncertainty) analysis and of data reduction. The notes will necessarily be brief; many of the concepts should be already familiar. The topics, along with specific examples, will be covered at greater length in several lectures. Experimental usefulness, rather than mathematical rigor, will be emphasized.

There are a great number of books which cover this subject at various levels: older texts which contain much useful background and presenting numerical methods for hand computations, and *newer* ones which provide actual computer programs (Library call numbers: Dewey 519, Lib.Cong. QA278). The *book Data Reduction and Error Analysis for the Physical Sciences* by P.R. Bevington and D. Robinson is at an appropriate level but presents programs in FORTRAN. Equivalent C programs can be found *in Numerical Recipes in C, The Art of Scientific Computing, 2nd Edition* by W.H.Press, B.P.Flannery, S.A.Teukolsky, W.T.Vetterling, although that book is somewhat mathematical in its approach.

1. Significant Figures

All numerical measurements or results must be expressed with appropriate significant figures, which express the precision of the number. For example, writing the values 1, 1.0, and 1.00 imply that the values lie between 0.5 and 1.5, 0.95 and 1.05, and 0.995 and 1.005, respectively.

A measurement and its experimental error should have the same significance, e.g., 273 ± 6 (or equivalently, $((273 \pm 0.06) \times 10^3)$).

When adding or multiplying two numbers with different precisions, for example, the precision of the result is determined by the number with the least precision. For example

1.093 + 2.5 = 3.6 **not** 3.593

1.093 x 2.5 = 2.7 **not** 2.7325

Round-off errors will appear if you are not sensible. For example

0.98 x 1.04 = 1.02 **not** 1.0

In this case each number is known to a precision of about 1% and the result should express this fact. But

1.9 x 6.135 = 11.7 **not** 12 (Never throw precision away!)

It is legitimate to quote an experimental result to one more significant figure than the experimental precision would dictate, since in a computation one significant figure is sometimes lost, as shown above. As a rule, however, errors are usually given to only 1 significant figure; an error itself has uncertainty and so it is meaningless to quote a high precision for it.

2. Systematic vs. random errors

All measurements have uncertainties. It is the aim of the experimenter to reduce these to a minimum. Apart from gross errors which lead to obviously wrong results, such as misreading the scale of an instrument, there are two classes of errors which must be considered, and their effects estimated. One of these is *random errors*: Whenever you make a series of independent measurements of some quantity, these measurements will generally be spread over a range of values, with about half the fluctuations lying above the "best" (or average) value, and half below. The range of the measurements is a measure of their precision. The source of these random errors can often not be identified directly, but can arise from errors in judgment when reading a scale to the smallest division, or from small fluctuations in temperature, line voltage, etc., which can affect the measuring instruments. Random errors can usually be treated by statistical analysis and form the basis of these notes.

Systematic errors can not be treated statistically. They can arise, for example, from miscalibrated instruments or occasionally from observer bias. A very common error made by students is to assume that the zero-point on an instrument scale really is zero. There is generally no reason to assume this, since the zeroing of any instrument is part of its calibration, along with setting any scales on the instrument. Corrections for systematic errors can often be made to the data provided their presence is known and understood. This is not always the case, however. Consider the table below, which shows the results of two sets of measurements of the length of a piece of wood, with two different rulers and on two separate occasions:

Ruler, Temperature	Result (cm)
Steel, 10° C	80.35 ± .05
Plastic, 10° C	80.12 ± .08
Steel 30° C	80.17 ± .05
Plastic, 30°C	80.06 ± .08

The quoted (random) errors in the data were presumably found by a suitable analysis of the spread of the measurements, but these data also contain some obvious systematic errors, related to the material of the ruler used, and to the temperature at which the measurements were made. (The measurements made with the steel ruler are greater than those with the plastic, and higher temperature results are smaller than those made at lower temperatures).

What is the length of the piece of wood? Notice that measurements made with the same ruler have the same "precision", which is an indication of the exactness of the measurements made with that instrument, but the ultimate "accuracy" of the measurement will depend on how well the experimenter is able to account for the thermal expansion of his rulers, and in the case of the plastic ruler, expansion due to the relative humidity of the air at the time of the measurement. These types of systematic error can often be removed from the data, e.g., if the appropriate expansion coefficients are known. Such corrections will not increase the precision of the experiment, but will increase the accuracy in the final result for the length of the piece of wood, which must presumably be quoted for a specific temperature. The quoted error must be increased to accommodate the additional uncertainties arising from the corrections. These are, of course, just estimates of the exact corrections.

Very often, there are no specific rules to follow, and you must use your own good judgment in handling errors, especially systematic errors. As a guiding principle, be conservative, be honest with yourself and in the presentation of your results. The uncertainty in a result is usually expressed as a single number, but may often include an estimate of the both systematic and random errors, separately.

3. Measurements of a Single Quantity: Random Errors

If a quantity x is measured many times N, the individual measurements x_i form a distribution from which the experimenter wishes to extract a "best" value for x. Usually, but not always, the "best" value is the mean $\bar{x}$ of the distribution:

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \qquad \text{(A.1.)}$$

(For an asymmetrical distribution, the "best" value might be the *median*, which is that point above which half of the x_i lie, with half below, or it might be the *most probable value*, which is the peak of the distribution. Of course, all three quantities are equivalent for a symmetrical distribution.) The spread or width of the distribution is an indication of the precision of the measurement. As a measure of this width we consider the deviations of the individual measurements from the mean, i.e.

$$d_i = \bar{x} - x_i \qquad \text{(A.2.)}$$

From our definition of $\bar{x}$, the sum of the deviations must equal zero. The mean deviation, defined in terms of the magnitudes of the deviations:

$$\bar{d} = \frac{1}{N}\sum_{i=1}^{N} \left|\bar{x} - x_i\right| \qquad \text{(A.3.)}$$

is often useful, but a more useful measure of the dispersion of a measurement is the *standard deviation* s. The variance of the distribution s^2 is the average of the squared deviation:

$$s^2 = \frac{1}{N}\sum_{i=1}^{N} \left(\bar{x} - x_i\right)^2 \qquad \text{(A.4.)}$$

The standard deviation is then the square root of the variance, or the root mean square of the deviations.

To be more mathematically correct, we should recognize that the measured distribution is a "sample" distribution, which will be different every time we make a different series of measurements of x, rather than the "parent" distribution for which the mean is the exact, or "true" value. The best experimental estimate of the parent or "true" standard deviation is given by:

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(\bar{x} - x_i)^2} \qquad \text{(A.5.)}$$

The quantity N -1 is the *number of degrees of freedom* in the problem, which equals the number of data points minus the number of fit parameters. If you ever have to worry about the difference between $\sqrt{N}$ and $\sqrt{N-1}$ in your data, then the data are probably not very good!

Note that this s is a measure of the uncertainty in a *single* measurement; it measures the width of the distribution of all measurements and is independent of the number of measurements! But, clearly, the more times we make a measurement, the better we expect to be able to determine the mean. We shall show later that the standard deviation of the mean is given by:

$$s_{\bar{x}} = \frac{s}{\sqrt{N}} = \sqrt{\frac{\sum_{i=1}^{N}(\bar{x} - x_i)^2}{N(N-1)}} \qquad \text{(A.6.)}$$

i.e. in order to improve the accuracy by a factor 2, you must make 4 times as many measurements.

Occasionally, the term *probable error* is used to describe the precision of a measurement. This is defined such that there is a 50% probability of exceeding this error if another measurement is made. It is a little greater than the standard deviation: The exact relationship between the two depends on the form of the distribution.

We emphasize that these definitions are general, and apply to any distribution of measurements. There are specific theoretical distributions which can be used as models of actual data and which are very useful. We shall discuss them and their applicability shortly.

4. Propagation of Errors

Suppose we measure several quantities *a, b, c*,... each with its own standard deviation, s_a, s_b, s_c ... and then use these values to determine a quantity *y = f(a,b,c...)*. What is the standard deviation of the quantity *y*?

We can differentiate the function to find how changes $\Delta a, \Delta b, \Delta c$... in each of the *a, b, c*... affect the value of *y*:

$$\Delta y = \Delta a\left(\frac{\partial y}{\partial a}\right)_{bc\ldots} + \Delta b\left(\frac{\partial y}{\partial b}\right)_{ac\ldots} + \ldots \qquad \text{(A.7.)}$$

This is, of course, really the first term in a Taylor expansion, and corresponds to assuming that the partial derivatives do not change over the ranges $\Delta a, \Delta b$,... (For large errors, we must include terms in $\partial^2 y / \partial a \partial b$ and the cross terms , etc.)

If the function $y=ab$, or $y=a/b$ for example, then

$$\frac{\Delta y}{y} = \frac{\Delta a}{a} + \frac{\Delta b}{b} \qquad \frac{\Delta y}{y} = \frac{\Delta a}{a} - \frac{\Delta b}{b}$$

respectively.

In general, we do not know the absolute errors $\Delta a, \Delta b, \Delta c$, in the measurements of $a,b,c,$ but rather a quantity such as their standard deviations s_a, s_b, s_c, (or the probable errors). However, from the above equations it is intuitively likely that the variances should add in the form:

$$s_y^2 = s_a^2\left(\frac{\partial y}{\partial a}\right)^2 + s_b^2\left(\frac{\partial y}{\partial b}\right)^2 \ldots \qquad \text{(A.8.)}$$

That this is indeed true follows from:

$$s_a^2 = \frac{1}{N}\sum_{i=1}^{N}(\bar{a} - a_i)^2 = \frac{1}{N}\sum_{i=1}^{N}\Delta a^2 \qquad \text{(A.9.)}$$

and then

$$s_y^2 = \frac{1}{N}\sum_{i=1}^{N}\Delta y^2 = \frac{1}{N}\sum\left[\Delta_a^2\left(\frac{\partial y}{\partial a}\right)^2 + \Delta_b^2\left(\frac{\partial y}{\partial b}\right)^2 + \ldots + \Delta_a\Delta_b\frac{\partial^2 y}{\partial a \partial b} + \ldots\right] \qquad \text{(A.10.)}$$

which leads to equation A.8 if we neglect the cross-terms: This will be valid if the *a's* and *b's* are independent, but if they are related this correlation must be accounted for.

Thus, for example, for a function $y=abc$, or $y=ab/c$, we obtain:

$$\frac{s_y^2}{y^2} = \frac{s_a^2}{a^2} + \frac{s_b^2}{b^2} + \frac{s_c^2}{c^2} \qquad \text{(A.11.)}$$

and for $y=a+b+c$:

$$s_y^2 = s_a^2 + s_b^2 + s_c^2 \qquad \text{(A.12.)}$$

5. Some Important Distributions

a) Binomial Distribution

Consider an experiment, such as tossing a pair of dice to obtain two sixes. The experiment has only two possible outcomes, with probability p for success, and probability $q=(1-p)$ for failure. (Obviously, in this case p=1/36 and q=35/36). For a series of N such experiments, the probability of throwing sixes for the first k times, say, and then not throwing sixes for the remaining $(N-k)$ times is

$$\text{Prob} = p^k q^{(N-k)} = p^k(1-p)^{(N-k)} \qquad \text{(A.13.)}$$

But if we ask what is the chance of obtaining a pair of sixes for any k out of N tries, in any sequence, we must multiply the above expression by the number of ways this may occur. The result is known as the binomial distribution:

$$P_B(p,k,N) = \frac{N!}{k!(N-k)!}p^k(1-p)^{(N-k)} \qquad \text{(A.14.)}$$

It is left as an exercise to the student to prove this, and to show that the mean and variance of the distribution are given by:

$$\bar{x} = Np \qquad \text{(A.15.)}$$

$$s^2 = Np(1-p) \qquad \text{(A.16.)}$$

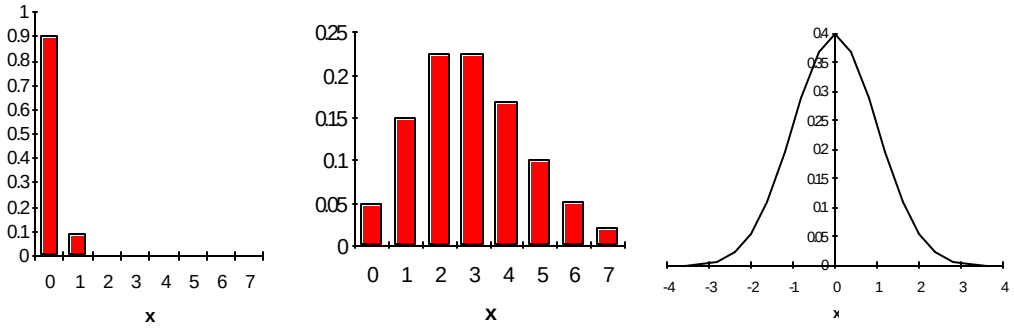
b) Poisson Distribution

The binomial distribution becomes unwieldy for large *N*. But in the very common situation where *p* is very small, (remember that the mean $\bar{x} = Np$), it can be approximated by the Poisson distribution, which has a much simpler analytic form. This distribution is:

$$P_p(x,\bar{x}) = \frac{\bar{x}^x}{x!} e^{-\bar{x}} \qquad \text{(A.17.)}$$

The variance of this distribution is $\bar{x}$, (from $s = \sum (\bar{x} - x)^2 P(x)$, so that the standard deviation of the mean $\bar{x}$ is $\sqrt{\bar{x}}$. A Poisson distribution is then completely specified by its mean $\bar{x}$ and its standard deviation $s = \sqrt{\bar{x}}$.

Both the binomial and Poisson distributions are asymmetric, and they are also discrete. The Poisson distribution is especially useful in determining the statistics of radioactive decay, where in general, *N*, the number of parent nuclei is enormous, and the probability of decay is small. It gives the statistics of the number of decays observed in a given interval, so that, for example, if 100 counts are observed in a given time interval, the standard deviation will be $\sqrt{100} = 10$. Beware, however, the thoughtless application of this rule at very low rates: it is the square root of the mean of the *parent* distribution that gives the standard deviation. For 100 counts, the rule works well because the mean probably lies in the range 90 to 110 and taking a square root then yields close to the same answer for s. But if you obtain 1 count, say, in a given time interval, the mean is not well-determined from this single measurement, and actually has a reasonable chance of being anywhere from about 0.1 to 3. So you cannot say that the s is 1 in this case. The Poisson distributions for means of 0.1 and 3.0 are shown in Figures A.5a and A.5b.



• Figure A.1. a b c

c) Gaussian Distribution

When $\bar{x}$ becomes large, the Poisson distribution becomes very symmetric and eventually can be expressed in a more useful form, the *Gaussian distribution*, often referred to as the *normal distribution:*

$$P_G(x,\bar{x},s) = \frac{1}{s\sqrt{2p}} e^{\left[-\frac{(\bar{x}-x)^2}{2s^2}\right]} \qquad \text{(A.18.)}$$

This is a continuous function in *x*, so that the probability of any measurement falling within an interval *dx* around *x* is given by:

$$dP_G = P_G(x,\bar{x},s)dx \qquad \text{(A.19.)}$$

The probability of a measurement lying $1s$, $2s$ from the mean, respectively, is 0.683, and 0.954. This is illustrated in Figure A.5c. It is also easy to show that $1s$ = 1.48 x the probable error.

The Gaussian distribution forms the basis of most error analyses, especially because of its simple analytical form. It represents many experimentally observed distributions extremely well; in practice, however, there are often longer tails in an experimental distribution than predicted by the formula above.

6. Estimating the Mean And Its Error: Method of Least Squares

We previously asserted that the "best" value from a symmetric distribution of measurements is the *mean* of the measurements. It is simple to show that this choice corresponds to minimizing the variance of the distribution. i.e.:

$$s^2 = \frac{1}{N}\sum_{i=1}^{N}(\bar{x} - x_i)^2 \qquad \text{(A.20.)}$$

from which

$$\frac{ds^2}{d\bar{x}} = \frac{1}{N}\sum_{i=1}^{N}(2\bar{x} - 2x_i) \qquad \text{(A.21.)}$$

Setting this equal to zero as the condition for a minimum in s^2, we find:

$$2N\bar{x} - 2\sum_{i=1}^{N} x_i = 0 \qquad \text{i.e.} \qquad \bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \qquad \text{(A.22.)}$$

This is an example of the *principle of least squares*: the "best" value from a distribution of measurements is that which corresponds to minimum variance, (i.e. the least sum of the squares of the deviations). The principle of least squares is itself a consequence of *the principle of maximum likelihood*, which is at the foundation of much of the field of statistics, as we shall now show:

Suppose we have a (Gaussian) distribution with mean $\bar{x}$. The probability of obtaining any given measurement x_i within the range dx_i is:

$$P_i = \frac{1}{s\sqrt{2p}} e^{\left[-\frac{(\bar{x}-x_i)^2}{2s^2}\right]} dx_i \qquad \text{(A.23.)}$$

and then the probability of obtaining a particular distribution of measurements $x_1, x_2, \ldots x_N$ is:

$$P(x_1, x_2 \ldots x_N) = P(x_1)P(x_2)\ldots P(x_N) = \prod P(x_i) = \left(\frac{dx}{s\sqrt{2p}}\right)^N e^{\left[-\sum_{i=1}^{N}\frac{(\bar{x}-x_i)^2}{2s^2}\right]} \qquad \text{(A.24.)}$$

This probability is a maximum when the exponent is a minimum, which corresponds to the least squares condition.

Weighted Mean

If each measurement in the expression for the total probability has a different s_i, then the exponent which must be minimized is:

$$\sum_{i=1}^{N} \frac{(\bar{x} - x_i)^2}{2s_i^2} \qquad \text{(A.25.)}$$

Differentiating this with respect to $\bar{x}$ then gives:

$$\bar{x} = \frac{\sum \frac{x_i}{s_i^2}}{\sum \frac{1}{s_i^2}}, \quad \text{or:} \quad \bar{x} = \frac{\sum w_i x_i}{\sum w_i} \quad ; \quad w_i = \frac{1}{s_i^2} \qquad \text{(A.26.)}$$

where the w_i, the reciprocals of the variances, are called the weights of the measurements. This is the prescription for determining a ***weighted mean***, when each measurement has a different weight.

What is the error of the mean? In equation (A.8.) we found a general expression for the variance s_y^2 of a function *y=f(a,b..).* It is equal to the sum of terms $s_a^2 \cdot (\partial y / \partial a)^2$ Applying this to our definition of the mean:

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i \qquad \text{(A.27.)}$$

we find:

$$\frac{\partial \bar{x}}{\partial x_i} = \frac{\partial}{\partial x_i} \left(\frac{\sum x_i}{N} \right) = \frac{1}{N} \qquad \text{(A.28.)}$$

Thus, if the standard deviations of the data points are all equal, i.e., $s_i = s$, the estimated error of the mean is given by:

$$s_{\bar{x}}^2 = \sum_{i=1}^{N} \frac{s^2}{N^2} = \frac{s^2}{N} \qquad \text{(A.29.)}$$

as stated earlier. And if the uncertainties are not equal, we evaluate the partial derivatives from our expression for the weighted mean:

$$\frac{\partial \bar{x}}{\partial x_i} = \frac{\partial}{\partial x_i} \left(\frac{\sum \left(x_i / s_i^2 \right)}{\sum \left(1/s_i^2 \right)} \right) = \frac{1/s_i^2}{\sum \left(1/s_i^2 \right)} \qquad \text{(A.30.)}$$

which leads to an expression for the uncertainty in the weighted mean:

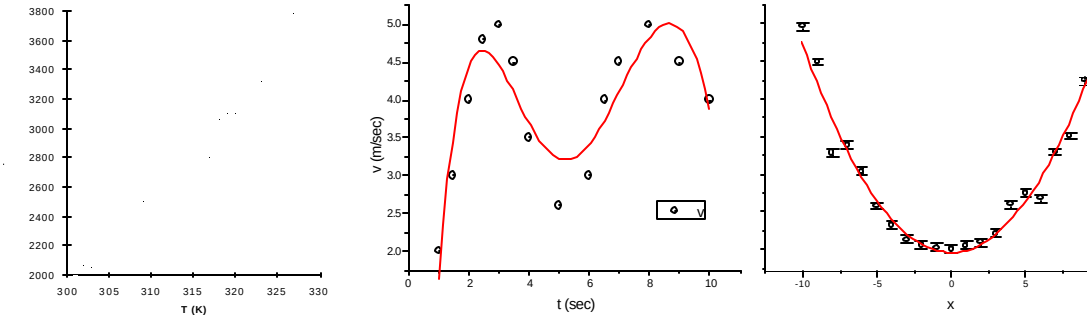
$$s_{\bar{x}}^2 = \frac{1}{\sum \left(1/s_i^2 \right)} \qquad \text{(A.31.)}$$

7. Graphical Presentation of Data

A well-drawn graph can provide an overall view of a set of data which is not possible from a table of numbers. There are good ways to represent individual data points.

Suppose you are plotting some measurements y_i as a function of a dependent variable x_i , e.g. electrical resistance vs. temperature. If there are many data points with small relative errors, and the dependence of *y* on *x* is fairly slow, then it may be reasonable to represent each data point by a single point as in Figure A.5.a. If, however, the dependence is fairly rapid and there are limited data points, it is usually better to draw a circle around the point, as in Figure A.5.b. Notice that in this figure, we drew a line through the data points to guide the eye; without this line the data would have been much more

difficult to comprehend. However, the line we drew did not go through all data points, we did not simply connect all points by straight lines. Such a *smooth* line is valid since sudden changes in slope are probably unphysical; this was justified since we presumably had some knowledge of the likely errors in the measurements.



• Figure A.2. a b c

Generally, if you know the standard deviations of each data point, these should be indicated on the graph by vertical lines through each point. Again, it may be useful to draw a smooth line through your data, in order to guide the eye, or actually to determine a slope, or to illustrate a fit to the data. Figure A.5.c shows an example.

Graphs serve several functions, in addition to summarizing data: A graph is usually necessary for observing fine features in the data, for evaluation of "backgrounds" to experiments, to show likely bad data points, etc. A good graph is often necessary to show that your data are *convincing*, and should therefore be given careful thought. It should be accurately labeled, and scales should be chosen for optimum effect, e.g. log-log to demonstrate power laws, etc. A very useful rule to use in evaluating data shown in graphical form follows from our discussion of standard deviations earlier**: A fitted line should pass through roughly two-thirds of the error bars, missing about one-third.** If the line misses significantly more than one-third of the data points, the errors have been underestimated, and if it passes through more, the errors are too conservative.

8. Fitting Data to a Straight Line y = a + bx

It is very often required to determine the "best fit" of some data to a known functional form. We first consider fits to a straight line, i.e. given a series of N measurements of a quantity y_i as we vary some x_i, find the "best" values of a and b, along with their errors, where the data are expected to have the form $y=a+bx$. We assume that the x_i have errors which are much smaller than y_i, which is usually the case. We use an extension of the least squares method which we have already used for determining the best value for the mean of a distribution of measurements of a single quantity. This process is sometimes called *linear regression*. For any arbitrary values of a and b, we can calculate the deviation d_i between each of the measured values y_i and the corresponding calculated values:

$$d_i = y_i - y(x_i) = (y_i - a)bx_i \qquad \text{(A.32.)}$$

The probability of obtaining a particular y_i is then given by:

$$P(y_i) \propto \frac{1}{s_{y_i}} e^{\left[-\frac{(y_i - y(x_i))^2}{2s_i^2}\right]} dx_i \qquad \text{(A.33.)}$$

where s_i is the standard deviation of a particular y_i. The probability of obtaining the full set of measurements $y_1, y_2, \ldots$ is given by the product of probabilities $P(y_1)P(y_2)\ldots = P$. As before, we need to maximize the probability P.

a) The Case When All s's Are Equal.

In this situation, maximizing P corresponds to minimizing the sum of the squared deviations with respect to the parameters a and b. i.e. For:

$$\sum d_i^2 = \sum (y_i - a - bx_i)^2 \qquad \text{(A.34.)}$$

the conditions for minimum deviation are:

$$\frac{\partial}{\partial a}\sum d_i^2 = \frac{\partial}{\partial b}\sum d_i^2 = 0 \qquad \text{(A.35.)}$$

which produces two simultaneous equations for a and b. The solution of these equations gives the required fit values for a and b:

$$a = \frac{S_y S_{xx} - S_x S_{xy}}{D} \;;\quad b = \frac{SS_{xy} - S_x S_y}{D} \;; \qquad \text{(A.36.)}$$

where

$$S, S_x, S_y, S_{xx}, S_{yy} = \sum_{i=1}^{N} \left(1, x_i, y_i, x_i^2, x_i y_i\right) \qquad \text{(A.37.)}$$
$$D = SS_{xx} - S_x^2$$

To determine the standard deviations of a and b, we again use the general expression for the effect of the variations $\partial z/\partial y_i$ on the standard deviation in a parameter z, so that:

$$s_a^2 = \sum s_i^2 \left(\frac{\partial a}{\partial y_i}\right)^2 \;; s_b^2 = \sum s_i^2 \left(\frac{\partial b}{\partial y_i}\right)^2 \qquad \text{(A.38.)}$$

Since the standard deviations s_i of each measurement y_i have been assumed to be the same, we can estimate these directly from the data. We find:

$$s_i^2 = s^2 = \frac{1}{N-2}\sum (y_i - a - bx_i)^2 \qquad \text{(A.39.)}$$

which is the function we have just minimized. (The factor N-2, the number of degrees of freedom, is the number of data points N minus the number of parameters (2) to be determined in the fit).

We can differentiate the solutions for a and b given above to obtain:

$$\frac{\partial a}{\partial y_1} = \frac{S_{xx} - x_1 S_x}{D} \;; \frac{\partial b}{\partial y_1} = \frac{Sx_1 - S_x}{D} \;; \qquad \text{(A.40.)}$$

Some algebra then gives:

$$s_a^2 = s^2 \frac{S_{xx}}{D} \;; s_b^2 = s^2 \frac{S}{D} \;; \qquad \text{(A.41.)}$$

where s is given by the equation quoted above, which can also be written:

$$s^2 = \frac{1}{N-1}\left(S_{yy} + Sa^2 + b^2 S_{xx} - 2aS_y - 2bS_{xy} + 2abS_x\right) \qquad \text{(A.42.)}$$

This form may be more suitable for computation.

b) Weighted Fits

It should now be obvious how you make a weighted least squares fit to a straight line with each of the y_i having *different* errors. The function to be minimized now is not simply the sum of the squares of the deviations, but the sum of squared deviations divided by the corresponding s_i. This function is called c^2 or chi-squared (hard 'ch'). i.e. we must minimize:

$$c^2 = \sum \frac{d_i^2}{s_i^2} = \sum \frac{(y_i - a - bx_i)^2}{s_i^2} \qquad \text{(A.43.)}$$

The results are the same as the equations above if we redefine:

$$S = \sum_{i=1}^{N} \frac{1}{s_i^2} \; ; S_x = \sum_{i=1}^{N} \frac{x_i}{s_i^2} \; ;$$
$$S_y = \sum_{i=1}^{N} \frac{y_i}{s_i^2} \; ; S_{xy} = \sum_{i=1}^{N} \frac{x_i y_i}{s_i^2} \qquad \text{(A.44.)}$$

The standard deviations are:

$$s_a^2 = \frac{S_{xx}}{D} \qquad s_b^2 = \frac{S}{D} \qquad \text{(A.45.)}$$

In this case, of course, the s_i are estimated separately for each measurement before the fit. In the previous case, it was the spread of measurements about the best fit that allowed us to determine the s's from the fit itself.

While equations B.36. to B.45. are correct they are susceptible to round-off errors. Such errors occur when a large number is divided by a very small one. *Numerical Recipes in C* provides the following improved linear regression formula which should be used in all your computer programs and spreadsheets. It uses the temporary variables t_i and S_{tt} to circumvent the round-off problem:

$$t_i = \frac{1}{s_i}\left(x_i - \frac{S_x}{S}\right) \text{ and } S_{tt} = \sum_{i=1}^{N} t_i^2 \qquad \text{(A.46.)}$$

Then

$$a = \frac{S_y - S_x b}{S} \; , \; b = \frac{1}{S_{tt}} \sum_{i=1}^{N} \frac{t_i y_i}{s_i} \qquad \text{(A.47.)}$$

and

$$s_a^2 = \frac{1}{S}\left(1 + \frac{S_x^2}{SS_{tt}}\right) , s_b^2 = \frac{1}{S_{tt}} \; , Cov(a,b) = -\frac{S_x}{SS_{tt}} \qquad \text{(A.48.)}$$

Occasionally, a situation arises where the errors in x are not small compared to those in y, as we have been assuming. This can be accounted for by increasing the s_y's using:

$$s_y^2 \rightarrow s_y^2 + b s_x^2 \qquad \text{(A.49.)}$$

You should realize, of course, that some very important functional forms can be reduced to a straight line: Any power law, or exponential, can be fitted with the prescription given above.

9. Fitting to Non-linear Functions

The function $c^2 = \sum (d_i / s_i)^2$ can be defined for any set of measurements, where the d_i represent the deviations from any arbitrary functional form. The prescription for minimizing the function is the same as above, but now if we introduce *m* parameters to be fitted, (in a polynomial for example), we find *m* simultaneous equations to be solved. This is usually best done by means of matrix methods. You may consult Bevington or *Numerical Recipes* for the appropriate code.

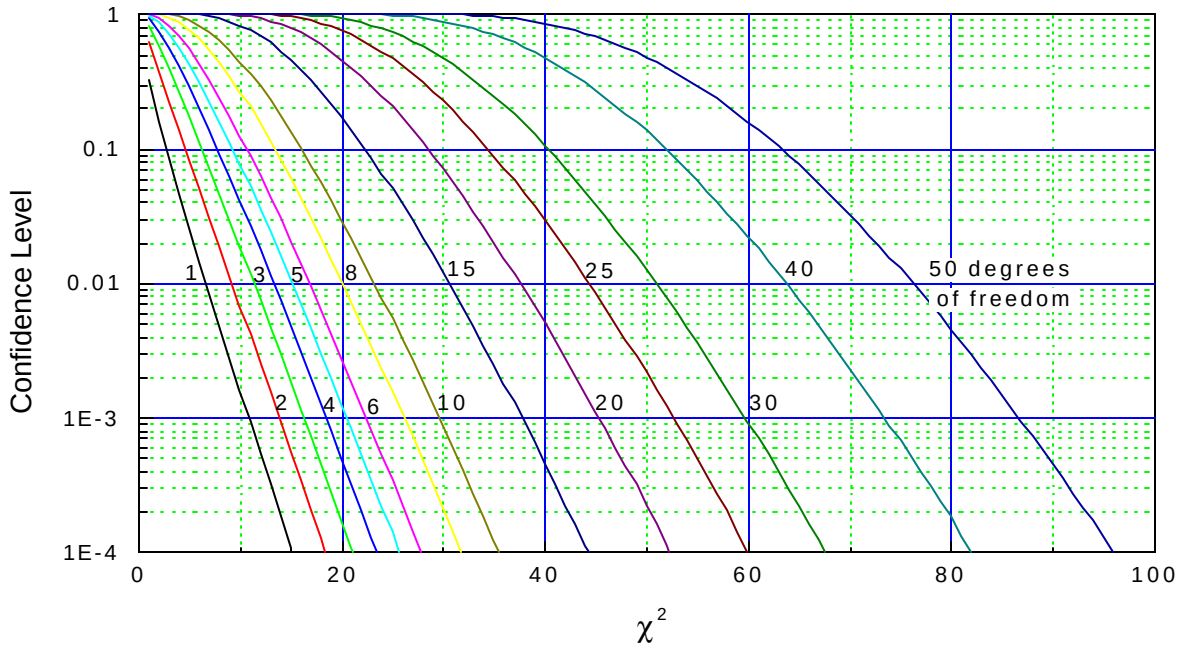
When fitting to an arbitrary function, the programs incorporate sophisticated mathematical minimization techniques. In particularly complex fits there may be problems due to finding local minima in a multi-dimensional space, and various methods must be used for checking whether a minimum is a true minimum or just a local one. Such techniques are beyond the level of these notes.

10. Goodness of Fit; more on Chi Square

One important question we have not yet asked is "How good is the fit?" This question can be answered by reference to the magnitude of the c^2 function after the minimization has been completed.

It is straightforward to obtain an intuitive feeling for c^2: Suppose we have made a fit to a theoretical curve. Then the final (minimum) value of c^2 is given by the sum of *N* individual terms $(y_{measured} - y_{theoretical})^2 / s^2$ as has been described. On average, we expect each measured data point to lie about 1*s* from the theoretical curve, if the fit is "good". On average then, each term in the sum for χ^2 contributes an amount 1 to the final sum, and the final value of c^2 should be roughly equal to *N* if the fit is good. Alternatively the reduced c^2 or c^2/DF which is the total c^2 divided by the *number of degrees of freedom* should be on order 1 for a good fit. (Remember that the number of degrees of freedom is the number of data points minus the number of parameters to be determined in a fit, e.g., *N*-2 for a straight line fit, *N*-5 for a fourth order polynomial, etc.)

Obviously a large value for the reduced c^2 implies a poor fit, with at least some data points lying far from the theoretical curve, and a very small c^2 implies either that the assumed errors are much too small, or, perhaps, that the theoretical function may be simpler than assumed. For a more quantitative measure of the *goodness of fit*, we must refer to the theoretical c^2 probability distribution, or more usually to tables or graphs of this function such as the one shown in figure A.3. This distribution function is derived assuming that the errors themselves are normally distributed. The accompanying graph shows *confidence level* as a function of c^2 and number of degrees of freedom. The confidence level is the probability that a random repeat of the experiment would result in a poorer (larger) c^2.



• Figure A.3. Confidence Level vs. Chi Square for different degrees of freedom

For example, if you do an experiment which consists of measuring 10 data points which you then fit to a straight line, and you find c^2=15, the graph shows a confidence level of 0.05 (there are 8 degrees of freedom), while if c^2=10, the confidence level would be 0.43. The first fit would be marginal: Confidence levels between 0.05 and 0.95 are usually taken as "acceptable".

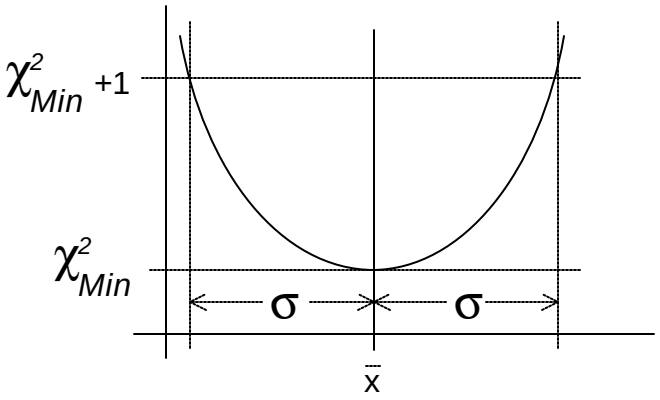
An occasionally useful non-analytical means of obtaining an error estimation for a one-parameter fit is obtained by determining the behavior of c^2 about its minimum value, $c^2(\bar{x})$. If we allow our estimation of the mean to vary by $\Delta\bar{x}$, then:

$$c^2(\bar{x} \pm \Delta\bar{x}) = \sum \frac{(x_i - \bar{x} \pm \Delta\bar{x})^2}{s_i^2} = \sum \frac{(x_i - \bar{x})^2}{s_i^2} \pm 2\Delta\bar{x} \sum \frac{(x_i - \bar{x})}{s_i^2} + (\Delta\bar{x})^2 \sum \frac{1}{s_i^2} \qquad \text{(A.50.)}$$

the second of these terms is zero, and therefore:

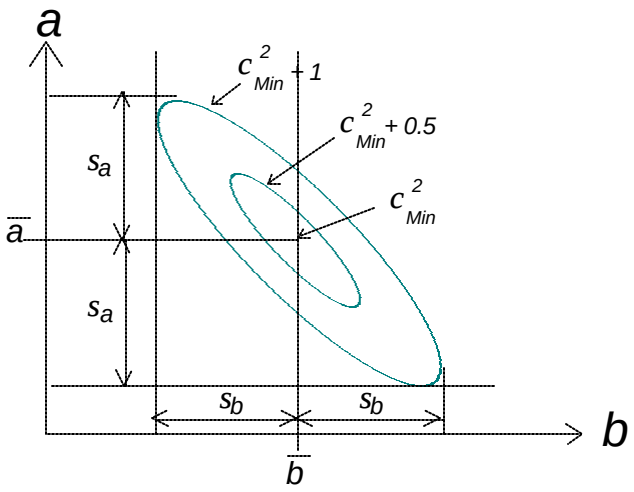
$$c^2(\bar{x} \pm s_x) = c^2(\bar{x}) + 1 \qquad \text{(A.51.)}$$

i.e. for a one parameter fit, one standard deviation "error" on the parameter is determined from where c^2 increases its value by 1. Also, $c^2_{\min} + 4$ determines the two standard deviation point; the c^2 function is parabolic about its minimum.



• Figure A.4.

For a multiple parameter fit, things are not quite so simple, e.g., for two parameters *a* and *b*, say, from a fit to a straight line, the $c^2(a,b)$ function forms a 3-dimensional surface, which can be represented by contours of constant c^2 on the (*a,b*) plane, as shown in Figure A.5. Notice that in general this form is an ellipse, indicating that *a* and *b* are *correlated*. (The correlation is obvious, since if we increase the slope of a straight line, in general we must decrease the intercept with the *y*-axis). The standard deviations are given by the rectangles drawn about the $c^2_{\min}+1$ contour. Strictly speaking, for such multiparameter fits, we must quote the *error matrix*, or *covariance matrix* from which the *correlated* errors can be determined. We shall not discuss this here.



• Figure A.5.

11. Some Comments On "Correlation Coefficient"

Many students have access to "canned" error analysis programs which they occasionally apply to fit their data. Some of these least squares fitting programs are fine, (they do the required summations described in Section 6) but essentially none of them will do a weighted fit, which is what is most often wanted. The computer prints out results, usually without errors, making them useless, and also a correlation coefficient "*r*" which the student then quotes without understanding it, but in lieu of a quoted error; this is generally a number greater then 0.99 and therefore must be good!(?). Naturally, no students in the present class are guilty of such behavior.

Most of these programs come as a package for use in the social sciences, where one of the aims is to determine whether there exists any correlation at all between variables. In the hard sciences, such a situation is very rare: There is generally a known or hypothesized functional relationship for which parameters can be determined, or which can be tested by means of c^2 or other tests.

We can develop a quantitative measure of the degree of correlation between two variables using the formulations already given. This will be the "linear correlation coefficient" *r*. Consider some data consisting of pairs of measurements, (x_i,y_i) , of which either *x* or *y* can be considered the dependent variable. We can do a linear least squares fit to the line $y=a+bx$ and determine a best value for the slope *b*, already given:

$$b = \frac{N\sum x_i y_i - \sum x_i \sum y_i}{N\sum x_i^2 - \left(\sum x_i\right)^2} \qquad \text{(A.52.)}$$

Weighting factors $1/s_i^2$ may be included. If there is no correlation between *x* and *y* the slope $b=0$. Thus the existence of a slope is an indication of a correlation; but the slope itself is not a direct measure of the correlation, since it might be very small although finite.

The data can also be fit to a straight line of the form $x=a'+b'y$ and the best fit now gives:

$$b' = \frac{N\sum x_i y_i - \sum x_i \sum y_i}{N\sum y_i^2 - \left(\sum y_i\right)^2} \qquad \text{(A.53.)}$$

Again, no correlation will result in $b'=0$. However, if there is a perfect correlation between the *x* and *y*, then obviously $b'= 1/b$, or $bb'=1$. We define the linear correlation coefficient *r* as:

$$r = \sqrt{bb'} \qquad \text{(A.54.)}$$

There exists a probability distribution for *r*, given in most statistics books which can be used to directly indicate the degree of correlation corresponding to a given value of *r*. The concept of linear correlation coefficient can be extended to multiple correlations in a general way, but we repeat that *r* is of limited use in the "hard" sciences.