University of Minnesota

School of Physics and Astronomy

Physics 4052 Lab Manual

Spring 2001

Appendix A

Some Notes on the Statistical Treatment of Data

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.

A.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, $((273\pm0.06)\times10^3$).

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 10% 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.

A.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.

A.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* σ. The variance of the distribution σ^2 is the average of the squared deviation:

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

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:

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

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 σ 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:

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

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.

A.4. Propagation of Errors

Suppose we measure several quantities *a, b, c*,... each with its own standard deviation, $\sigma_a, \sigma_b, \sigma_c$ ... 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 Δa, Δb, Δc... in each of the *a, b, c*... affect the value of *y*:

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

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 Δa, Δb,... (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 Δa, Δb, Δc, in the measurements of *a,b,c,* but rather a quantity such as their standard deviations $\sigma_a, \sigma_b, \sigma_c$, (or the probable errors). However, from the above equations it is intuitively likely that the variances should add in the form:

$$\sigma_y^2 = \sigma_a^2\left(\frac{\partial y}{\partial a}\right)^2 + \sigma_b^2\left(\frac{\partial y}{\partial b}\right)^2 \ldots \tag{A.8.}$$

That this is indeed true follows from:

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

and then

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

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:

$$\frac{\sigma_y^2}{y^2} = \frac{\sigma_a^2}{a^2} + \frac{\sigma_b^2}{b^2} + \frac{\sigma_c^2}{c^2} \tag{A.11.}$$

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

$$\sigma_y^2 = \sigma_a^2 + \sigma_b^2 + \sigma_c^2 \tag{A.12.}$$

A.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

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

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:

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

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:

$$\bar{x} = Np \tag{A.15.}$$

$$\sigma^2 = Np(1-p) \tag{A.16.}$$

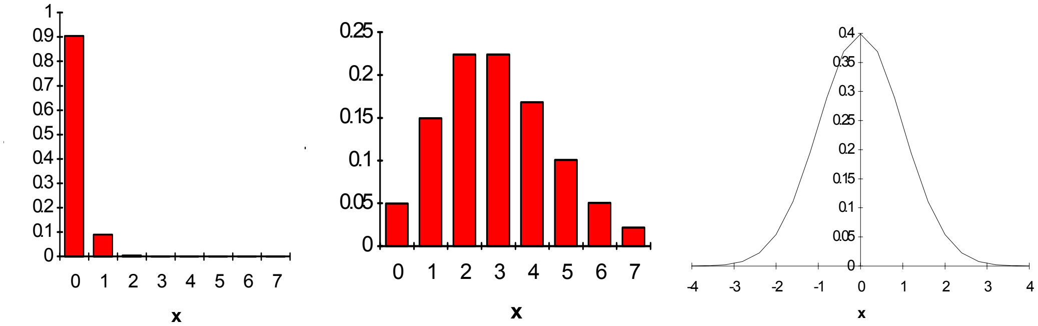
b) Poisson Distribution

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 $\sigma = \sum (\bar{x} - x)^2 P(x)$, 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 $\sigma = \sqrt{\bar{x}}$.

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 σ. 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 σ 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:*

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

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

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

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

A.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.:

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

from which

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

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

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

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 $\overline{x}$. The probability of obtaining any given measurement x_i within the range dx_i is:

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

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

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

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 σ_i , then the exponent which must be minimized is:

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

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

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

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 σ_y^2 of a function *y=f(a,b..).* It is equal to the sum of terms $\sigma_a^2 \cdot (\partial y/\partial a)^2$ Applying this to our definition of the mean:

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

we find:

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

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

$$\sigma_{\overline{x}}^2 = \sum_{i=1}^{N} \frac{\sigma^2}{N^2} = \frac{\sigma^2}{N} \qquad \text{(A.29.)}$$

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

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

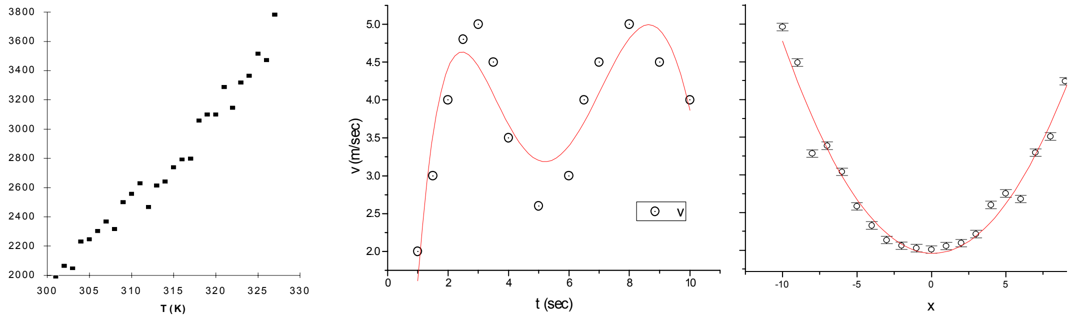
which leads to an expression for the uncertainty in the weighted mean:

$$\sigma_{\overline{x}}^2 = \frac{1}{\sum (1/\sigma_i^2)} \qquad \text{(A.31.)}$$

A.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.

A.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:

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

where σ_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 σ'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

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

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:

$$\sigma_a^2 = \sum \sigma_i^2 \left(\frac{\partial a}{\partial y_i}\right)^2 \;; \sigma_b^2 = \sum \sigma_i^2 \left(\frac{\partial b}{\partial y_i}\right)^2 \qquad \text{(A.38.)}$$

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

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

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:

$$\sigma_a^2 = \sigma^2 \frac{S_{xx}}{D} \;; \sigma_b^2 = \sigma^2 \frac{S}{D} \;; \qquad \text{(A.41.)}$$

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

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

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 σ_i . This function is called χ^2 or chi-squared (hard 'ch'). i.e. we must minimize:

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

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

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

The standard deviations are:

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

In this case, of course, the σ_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 σ'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:

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

Then

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

and

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

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 σ_y's using:

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

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.

A.9. Fitting to Non-linear Functions

The function $\chi^2 = \sum (d_i/\sigma_i)^2$ 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.

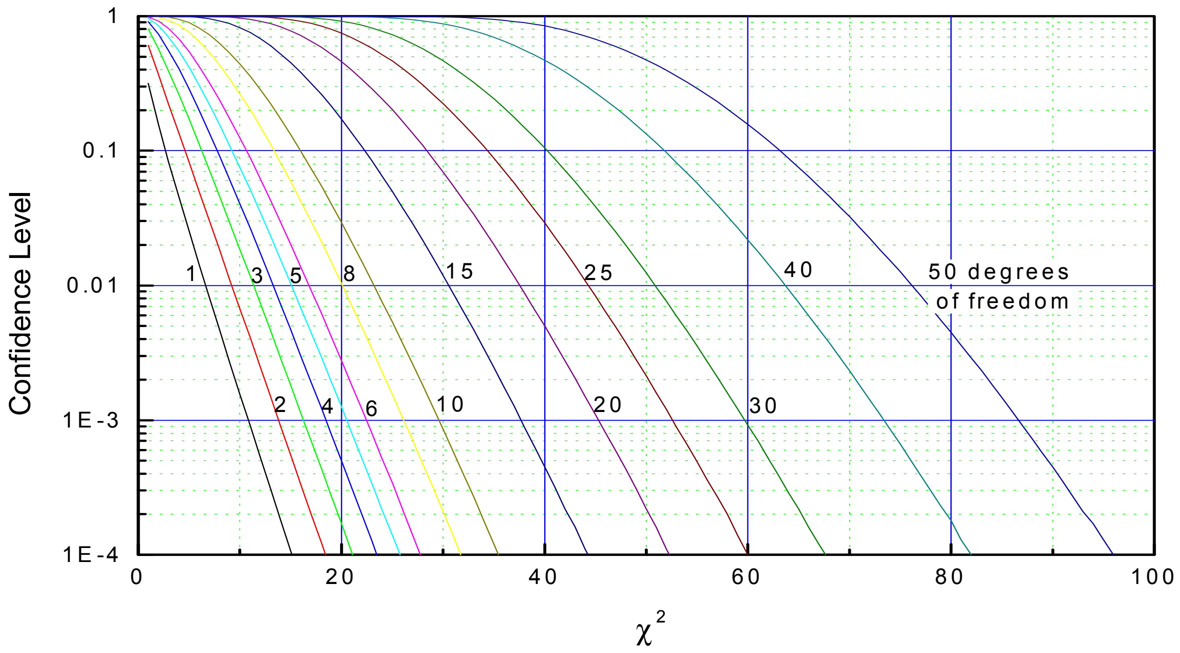
When fitting to an arbitrary function, the programs incorporate sophisticated mathematical minimization techniques. In particularly complex fits there may be problems due to finding local minima in a multi-dimensional space, and various methods must be used for checking whether a minimum is a true minimum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

A.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 χ^2 function after the minimization has been completed.

It is straightforward to obtain an intuitive feeling for χ^2: Suppose we have made a fit to a theoretical curve. Then the final (minimum) value of χ^2 is given by the sum of *N* individual terms $(y_{measured} - y_{theoretical})^2/\sigma^2$ as has been described. On average, we expect each measured data point to lie about 1σ 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 χ^2 should be roughly equal to *N* if the fit is good. Alternatively the reduced χ^2 or χ^2/DF which is the total χ^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 χ^2 implies a poor fit, with at least some data points lying far from the theoretical curve, and a very small χ^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 χ^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 χ^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) χ^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 χ^2=15, the graph shows a confidence level of 0.05 (there are 8 degrees of freedom), while if χ^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".

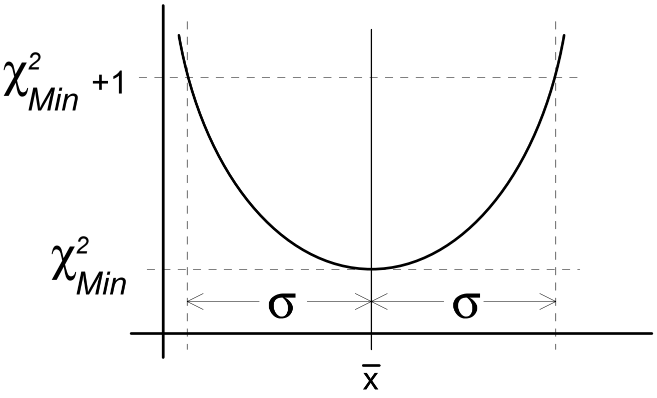
An occasionally useful non-analytical means of obtaining an error estimation for a one-parameter fit is obtained by determining the behavior of χ^2 about its minimum value, $\chi^2(\bar{x})$. If we allow our estimation of the mean to vary by $\Delta\bar{x}$, then:

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

the second of these terms is zero, and therefore:

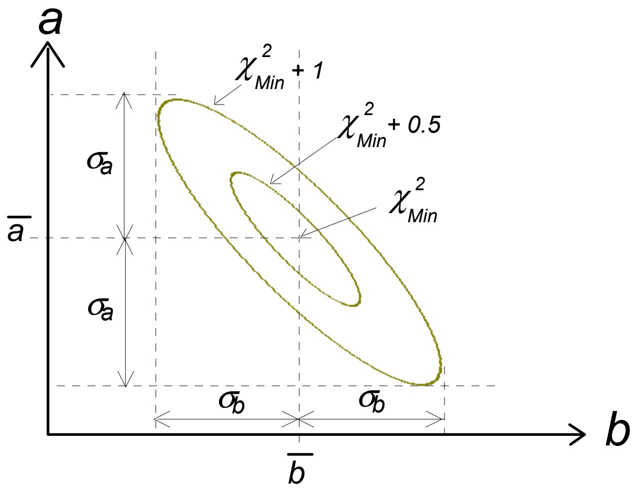
$$\chi^2(\bar{x}\pm\sigma_x) = \chi^2(\bar{x})+1 \qquad \text{(A.51.)}$$

i.e. for a one parameter fit, one standard deviation "error" on the parameter is determined from where χ^2 increases its value by 1. Also, $\chi^2_{\min}+4$ determines the two standard deviation point; the χ^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 $\chi^2(a,b)$ function forms a 3-dimensional surface, which can be represented by contours of constant χ^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 $\chi^2_{\min}+1$ 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.

A.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 χ^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:

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

Weighting factors $1/\sigma_i^2$ 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.

Appendix

B Fits Using X^2

B.1. Introduction

Good explanations of the χ^2 least squares fit (LSF) procedure can be found in Bevington and also in appendix A of this manual. The reason for including this appendix is to present the LSF in a less mathematical manner and instead emphasize the reasons why one would employ such a fitting procedure. In addition, it is sometimes said that fitting data is not just science but also an art. It is the intent of this write-up to strengthen your skills by pointing out what to look for in a fit and what to avoid and finally, how to read and interpret the results of a fit.

B.2. Least Squares Fit

A LSF calculates the best fit of a straight line through data points where each data point has an error or uncertainty associated with it; the definition of what constitutes a "best" fit determines how the fit is done.

A "good" fit is one where the fitted line is as close as possible to the data points. Therefore, the "offset", $\Delta_i = y_{measured_i} - y_{fit_i}$, is what we want to minimize. One possible method to do so is to add all the offsets together and then adjust ("wiggle") the line until the sum reaches a minimum. However, it can be shown that uncorrelated errors add in quadrature. So, instead, we add all the squares of the offsets and then "wiggle" the line till the sum is a minimum; this is the definition of a "best" fit.

The previous paragraph skipped over a crucial fact: how do we define or measure these "offsets" between a fit line and the data points when the errors are not identical? For example, if the error for a particular data point is very small, indicating that this particular data point is very well known, then it is more important for a "best" fit to pass close to such a point, as compared to a data point where the error bars are large because the location of the point is not very well known. Therefore, when the errors are not all identical, i.e. in a situation with weighted errors, the offset must be normalized. The normalization is achieved by dividing the offset by the magnitude of the data point's error i.e. its standard deviation.

$$\chi_i = \frac{y_{i\_measured} - y_{i\_fit}}{\sigma_i} \qquad \text{(B.1.)}$$

This quantity is called the "chi of i" and it represents how many error bars or standard deviations (as defined by that particular data point) it is off from the best fit. χ^2 for a fit is defined as:

$$\chi^2 = \sum_{i}^{N} \chi_i^2 \qquad \text{(B.2.)}$$

and you may think of this quantity as the square of the total errors of your fit. Since χ^2 grows larger the more data points you fit, a further normalization is useful:

$$\chi_v^2 \cong \frac{\chi^2}{N} \qquad \text{(B.3.)}$$

where *N* corresponds to the total number of data points fitted and χ_v^2 is called the "reduced Chi squared." (If you want to be picky, in two dimensions χ^2 should be divided by the degrees of freedom of the fit, i.e. *N*-2 but in a "good" analysis *N* >> 2, so you should be able to ignore the 2.) Another way of thinking of χ_v^2 is that it represents the square of the average error of your fit, i.e. how many (average) error bars squared (or sigma's) your data points (on the average) are away from the best fit.

Note that the units for χ_i, χ^2 and χ_v^2 are sigmas or sigmas squared and you should think of them in terms of Gaussian probabilities! For the sake of consistency, use these normalizations even when all the errors are identical. That way the errors are always expressed in similar quantities, i.e. sigmas or sigmas squared.

As previously mentioned, minimizing the sum of the individual offsets, i.e. minimizing χ^2, results in a best fit. Luckily, in two dimensions, there is no need to wiggle the fit curve to determine the minimum because it can be determined mathematically. See section B.8. of this manual for a derivation of all the necessary equations. Nevertheless, in more than 2 dimensions "wiggling" the curve is often the only way to determine the minimum. Various computer programs to find the best fit in more than two dimensions often resort to clever random walk algorithms but even those can not always tell with certainty whether they found the "real" minimum or just a local minimum.

Probably the most misunderstood and hardest concept to grasp about LSF fitting is how the magnitude of χ_v^2 (or χ^2) relates to the quality of the fit. The misunderstanding seems to come from the fact that we have stated that a best fit is one where the χ^2 is minimum. A logical, but not necessarily correct conclusion is to assume that the smaller χ^2 is, the better the quality of the fit. In the same manner, people sometimes adjust various parameters in their data to get a χ_v^2 that is as close as possible to 0. Also comparing different fits and assuming that the one with lowest χ_v^2 must have the best data set is simply not always correct! Read on and you should start to understand why a χ_v^2 close to 0 is not at all an indicator of a good fit.

B.3. Example: Fitting a Linear Signal With Random Noise

To illustrate what has been covered so far and to get a better understanding about the various χ's consider a LSF of a linear signal y_i=a+bx_i with some random noise y_{noise}. Under these circumstances, the signal observed or measured is:

$$y_{Measured_i} = y_i + y_{Noise_i} \qquad \text{(B.4.)}$$

Repeated measurement of $y_{Measured_i}$ allows us to determine the standard deviation of the signal, σ_{y_i}. If a LSF is performed using *N* values of x_I, $y_{Measured_i}$ and σ_{y_i}, what values for χ_i, χ^2, χ_v^2 should we expect?

Let's begin with the χ_v^2 : From the definition of the χ_v^2 we can expect that a LSF calculates that on the average, "most" (i.e. about 2/3) of the data is within **one** standard deviation from the LSF fit. Therefore, χ_v^2 should be close to unity.

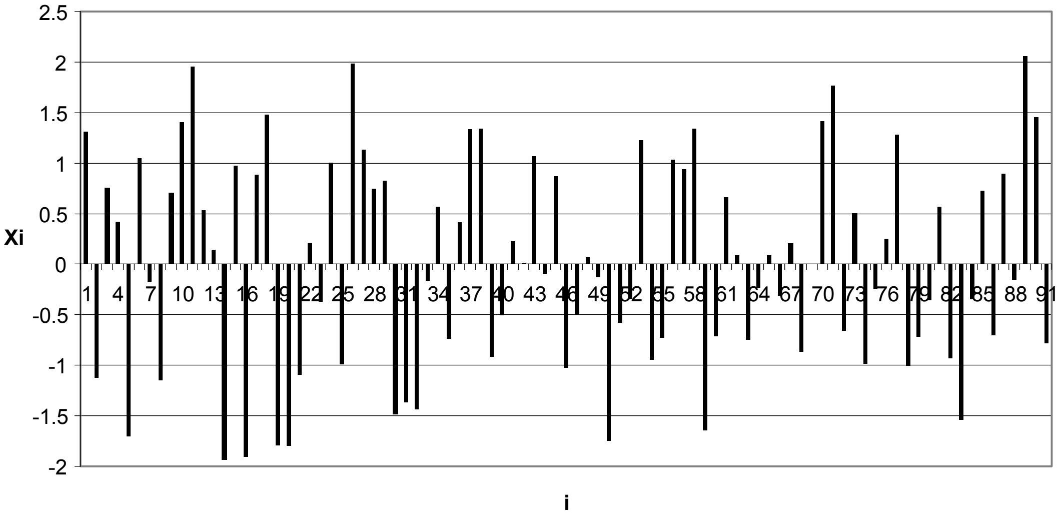
At first glance, this may seem not a very good result but, actually that is the best you can expect for random noise on a clean linear signal! The reason is that originally, if you correctly defined the size of the error bars by determining the σ_{y_i}'s, you indicated that the actual values had a pretty good chance to lie within one error bar. The fit confirmed the original guess, which is as good as it gets.

What happens to the χ_v^2 if your assumed errors, σ_{y_i} were either too small or too large? First consider the case where the experimenter is overly confident of the data and assumes that the errors, σ_{y_i}, are much smaller than what they are in reality. Under these circumstances, a LSF would have calculated a large value for χ_v^2 because on the average, the data was many (of these small) error bars away from the best fit. In other words, the average error squared, χ_v^2, would have been much larger than one error bar, indicating that most of the data does not fall within one sigma of the fitted line, as it should. For example, a reduced χ_v^2 of 4 indicates that to collect 2/3 of the data you have to go at 2 sigmas away from the best fit, a χ_v^2 of 9 indicates the majority of the data spread over three sigmas.

In the opposite case, where the experimenter was too cautious and assumed error bars or standard deviations that were far larger than was needed, the reduced χ_v^2 will be less than unity. Again, this indicates that most of the fitted data lies within less than one single standard average deviation.

From these arguments, and equation B.3 it follows that for this LSF with random noise, the χ^2 should have been approximately equal to *N*, the number of data points analyzed.

Finally, what do you expect a bar graph of χ_i vs. *i* to look? Such a graph is a representation of how many σ's a particular data point, $y_{Measured_i}$, is "off" from the best fit value that your program calculated. As can be seen from definition B.32 and B,43, a negative value of χ_i indicates that the data point fell below the best fit value, a positive χ_i indicates that the data point is above the fitted value. For "good" data, i.e. data with some random noise, it's equally likely for the points to fall below the fitted value as to be above it. In other words, χ_I should randomly vary in sign. Figure B.1 represents a picture of how χ_i for "good" data should look like:



• Figure B.1. Note three facts about the graph: the sign of the data is randomly positive and negative, most of the magnitudes are between -1 and +1 and there is no "trend" in the data, i.e. it appears random.

What about the magnitude of χ_I ? In a Gaussian distribution 63% of all error fall within one sigma. Therefore, about two thirds of your χ_I should be between -1 and +1. Furthermore, less than 5% of your χ_I should have a magnitude larger than 2.

To sum up, here is a table of what you should expect for a two dimensional LSF for *N* linear data with random noise:

χ^2	*N*
χ^2_υ	1
χ_i	2/3 of the data points should fall in the interval -1 to +1.

So far nothing has been said about the variables that we are most interested in, namely the intercept of the straight line fit, *a*, and the slope, *b* and their uncertainties σ_a and σ_b and how they are related to σ_{y_i}. The purpose of a LSF is to determine *a* and *b* with the highest degree of accuracy, i.e., the smallest σ_a and σ_b; the values of a and b in themselves are not very meaningful unless we also know their accuracies. From equations B.41. we can see that σ_a and σ_b are directly proportional to the errors, σ_{y_i}. Therefore, it may be tempting to "adjust" by decreasing it to obtain "a more accurate result". This is clearly unethical (=cheating) and meaningless. On the other hand, in "real" experiments, one is seldom certain what the "real" errors are; in these situations, all one can do is to make an educated and honest guess and then do the fit. Luckily, we can verify these assumptions by checking χ^2_υ. Only if χ^2_υ is near unity can we say with certainty that our σ_{y_i} were reasonable and, therefore, also our σ_a and σ_b ! That is why we put so much emphasis on understanding the meaning of χ^2 because if you want to understand the data analysis of an experiment, you must understand the error contributions!

B.4. Analysing a Fit

Before you use the results obtained by the least squares fit program you should always check the value obtained for the reduced χ_υ^2 and you should also inspect the individual χ_i vs. i. Knowing how to read and interpret these two factors can give you very important information about the data and the analysis. For example, it will tell you if your fit to a straight line is justified, if you have any bad data points and if your assignment of the error bars is reasonable.

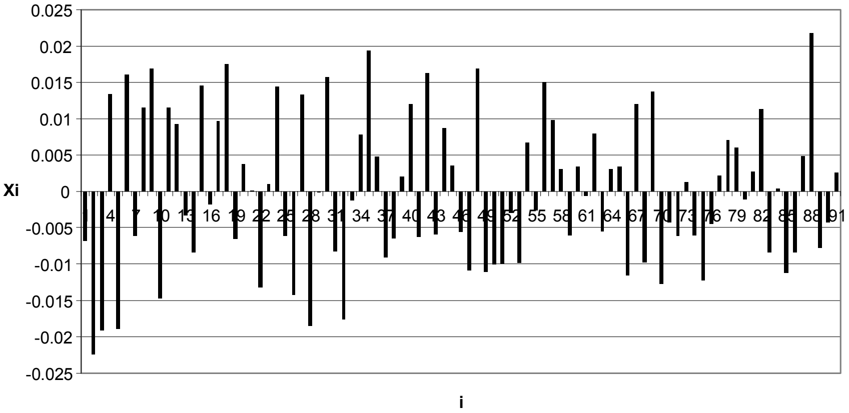
First, check the value obtained for the reduced chi square, χ_υ^2 . If it is significantly larger or smaller (by a factor of 3 or more) than unity, then you have a problem. (Don't panic yet, read the following paragraphs.) If your χ_υ^2 is close to unity, you should feel relieved because your analysis seems be going in the right direction and it confirms that your guesses concerning your error analysis seems to be correct. Nevertheless, you should not assume that everything is perfect until you have also checked the χ_i vs. *i*.

Second, look at a bar graph of χ_i vs. i. If your χ_i data does not look like the picture B.1 and your χ_υ^2 is not anywhere near what it should be, then the most common mistakes are: a bug in your spreadsheet program, faulty error bars, bad data points, wrong theory. We now will cover each of these points.

B.5. Troubleshooting χ2 analysis

If all your χ_i are positive then something is wrong with your spread sheet. A common mistake is to assume that $\chi_i = \sqrt{\chi_i^2}$. Clearly this does not hold for negative values of χ_i .

If the appearance of your χ_i data at first glance looks fine, i.e. you see no particular trend in the χ_i and the values of χ_i are randomly distributed, i.e. are equally likely to be positive or negative, but the average magnitudes of the χ_i (and of χ_υ^2) will be either very large or very small, (see figure B.2.) then you probably calculated your error bars incorrectly or made some unrealistic assumptions about your errors. (Of course, this depends on "how incorrectly" incorrect is!)



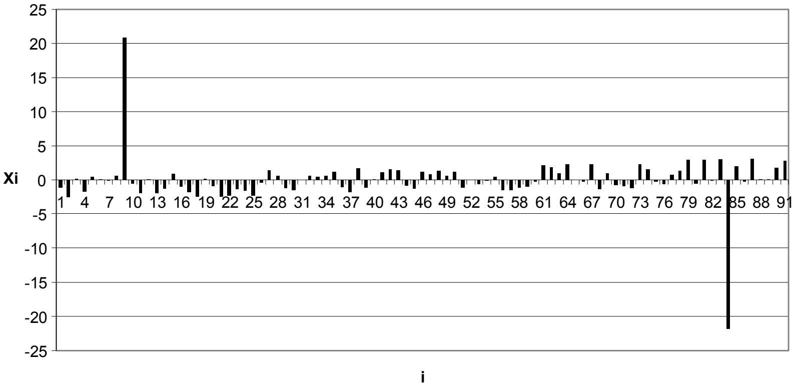
- Figure B.2.

This graph will tell you whether the errors you chose were too large or to small. Remember that χ_i is a representation of how many of <u>your</u> error bars you are away from the fitted result. If your error bars were too large, then on the average you will be much less than one unit away from the fitted curve. On the other hand, if your error bars are too small, then on the average you will be many error bars or away from it and your individual χ_i will be large! Since χ_υ^2 is a normalized average of all χ_i, then it follows that if the magnitudes of the χ_i are incorrect and it is very likely that χ_υ^2 will be incorrect too!

Assuming that there are no mathematical mistakes in your error analysis, what does it mean if you find (from the χ^2 analysis) that your errors bars are incorrect? Keep in mind that an error analysis is usually based on some "educated" guesses and the χ^2 analysis can tell you if these "guesses" were reasonable, i.e., if the assumptions made were too conservative (χ_υ^2 << 1) or too liberal (χ_υ^2 >> 1). In any case, if your χ_υ^2 is less than 2 or 3 (sigmas squared), do not bother "fudging" with the errors and stick instead with your original assumptions. If the analysis indicates that you made your errors too small, then it could imply that you have neglected other sources of errors. If it indicates that you made your errors too large then it means your instruments are better than you assumed originally.

Bad Data Points

If you have a few bad data points you may see something similar to the picture below. In addition to the two large spikes, indicating the "bad" data points, notice also the "trend" in the χ_i : χ_i on the left tend to be mostly negative while the ones on the right are mainly positive. This indicates that the fitted line has been skewed from a best fit by the bad data points.



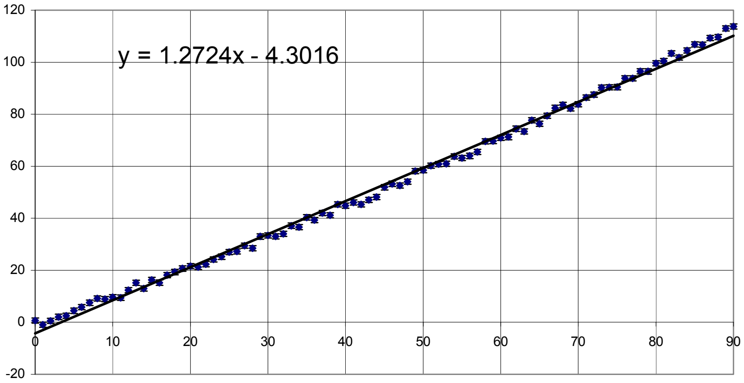
- Figure B.3.

In case you only have a **few** bad data points and if they are **many** sigmas off, you may remove them from your data. Depending on your definition of "few" and "many", you may be entering the more murky side of data analysis. Whenever you are not certain if what you are doing is appropriate or not, then always document exactly what you did with your data; in the analysis section of your report or paper describe how many data points you deleted out of how many and what criterion that you used and then let the reader make the final judgment!

Wrong Theory

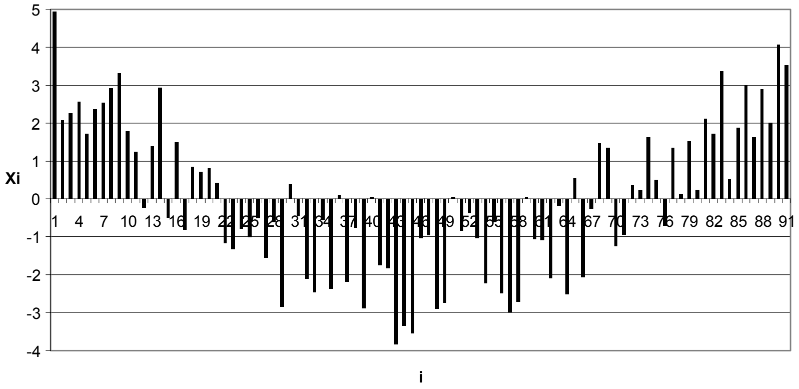
Sometimes the assumptions made when fitting the data to a straight line were wrong and the fitted data should not have been fitted to a straight line. This happens if the theory that predicates the behavior of the data is incorrect; linearizing it will not help. In reality though, it is much more common that the overall theory is correct but that a small second order effect is present in the data.

For example, the data in figure B.4. below looks fairly linear and a fit results in a χ^2_υ of 3.8 which is not great but acceptable.



- Figure B.4.

Looking at the χ_i of the data from figure B.4 clearly reveals a trend in χ_i and shows a small quadratic dependence in addition to the linear one.



• Figure B.5.

It is difficult to explain how to deal with such a result because of the different situations and magnitudes in which they arise. The two most common procedures are: If you have lots of time and/or the data is very important to you, filter out the first order effect and study both the first and the second order effect. (Sounds easier than it is.) Second, describe the second order effect in your paper, report etc. and leave it at that.

Nevertheless, keep in mind, if you should observe a "substantial" trend in your χ_i indicating that you are fitting data that is not linear, then you are doing something you should not be doing and the results may be all but meaningless!