

DATA and ERROR ANALYSIS

Performing the experiment and collecting data is only the beginning of the process of completing an experiment in science. Understanding the results of any given experiment is always the central goal of the experiment. Presenting those results in a clear concise manner completes the experiment. This overview of the complete process is as valid in an instructional laboratory course as in a research environment. You will not have learned any physics if you did not understand the experiment. Presenting the results of your experimental work is fundamentally saying, **"This is what I did and this is what I learned."**

Putting together your presentation of the results should help you clarify the results to yourself. (If your instructor can clearly see what you did and what you learned, you might get a better grade.)

Data analysis should NOT be delayed until all of the data is recorded. Take a low point, a high point and maybe a middle point, and do a quick analysis and plot. This will help one avoid the problem of spending an entire class collecting bad data because of a mistake in experimental procedure or an equipment failure.

First and foremost, data analysis means understanding what your results mean. When analyzing your data, try to think through the physical processes which are occurring. Write your train of thought down. Ultimately, the goal is for you to understand physics and the world a bit better. Your understanding of your results probably occurs in stages, with each stage being a refinement and possibly more mathematical than the previous stage.

For example, one might first note that as time increases so does distance. Next a quick graph of distance vs time might verify this understanding but the relationship is NOT linear, i.e. the data does not form a straight line. By further work, one might discover that distance increase linearly with the square of the time. Or sometimes the mathematical relationship may remain hidden.

Relate each successive stage of your understanding and interpretation of your results to the physical principles that are involved. In the above example, one might note that the change in position with time is caused by velocity that is in turn caused by an acceleration from the gravitational force. Finally, develop the related mathematics. Equations are nearly meaningfully unless they are related to the

physical laws. (Remember to identify all the variables and constants in you equations.)

Sometimes, your results will not support and may even contradict the physical explanation suggested by the manual or your instructor. Say so! But of course then a few suggestions as to the reason for this apparent failure of the physical laws, would be in order. Do NOT just say "The equipment was a piece of sh_t!" Try to explain what went wrong or what competing effects have come into play.

One of the reasons that you are encouraged to record everything that is going on as it is going on, is that this information may help explain bad results. For example, partly for fun, you note each time your lab partner sneezes. Later while looking at the data, you discover that each data point that was being collected during a sneeze deviates from the pattern of the rest of the data. This may give you good reason for dropping "bad" data.

The quality of the data, determines to a great extent, what conclusions can be reached from them. If you are looking for a small effect, say a total change of 1 mm, and the uncertainties in your data is 2 mm then you really can not make any solid conclusions. (See the section on error analysis below.)

When one considers the **quality of a measurement** there are two aspects to consider. The first is if one were to repeat the measurement, how close would new results be to the old, i.e., how reproducible is the measurement? Scientists refer to this as the **precision** of the measurement.

Secondly, a measurement is considered "good" if it agrees with the true value. This is known as the **accuracy** of the measurement. But there is a potential problem in that one needs to know the "true value" to determine the accuracy.

A good measurement must be close to the "true value" and be reproducible. In this experiment, if someone made one measurement of g and got 9.79 m/s^2 , it would be an accurate measurement. But if next time they tried they got 4.1 m/s^2 , no one would believe that they were anything but lucky in the first measurement. Similarly, if one group got values of $7.31, 7.30, 7.33$, and 7.29 m/s^2 their results are reproducible but not really very good.

Accuracy vs. Precision: These two words do not mean the same thing.

"**Accuracy**" deals with how close is a measured value to an accepted or "true" value.

"**Precision**" deals with how reproducible is a given measurement.

Because, *precision* is a measure of how reproducible a measurement is, one can gain some knowledge of the precision simply by taking a number of measurements and comparing them.

If the true value is not known, the *accuracy* of a measurement is more difficult to know. Here are a few things to consider when trying to estimate (or explain) the quality of a measurement:

1. How well does your equipment make the needed measurements? Two examples of problems: a meter stick with 2 mm worn off one end and trying to measure 0.01 mm with a meter stick.
2. How does the lack of precision, or *uncertainty* in any one measured variable effect the final calculated value? For example, if the data is expected to fall on a straight line, some uncertainties may only shift the intercept and leave the slope unchanged.
3. Making a plot that shows each of your measurements, can help you "see" your uncertainty. Also certain uncertainties are sidestepped by extracting a slope from a plot and calculating the final value from this slope.
4. One can estimate the uncertainty after making multiple measurements. First, note that when plotted, about $\frac{1}{3}$ of the data points will be outside of the *error bars* as they are normally drawn at one *standard deviation*. For example if there are six measurements, one expects that 2 of the points (probably one too high and one too low) will be outside the normal error bars. One can draw a reasonable set of error bars based on this assumption, by drawing the upper part of the error bar between the highest two points and the lower part of the error bar between the lowest two points.

Data analysis is seldom a straight forward process because of the presence of uncertainties.

Data can not be fully understood until the associated uncertainties are understood.

◆ ERROR ANALYSIS

The words "error" and "uncertainty" are used to describe the same concept in measurement. It is unfortunate that the term, "error" is the standard scientific word because usually there is no mistake or error in making a measurement. Frequently, the uncertainties are dominated by natural irregularities or differences in what is being measured.

Types of Error: All measurements have errors. Errors may arise from three sources:

- a) **Careless errors:** These are due to mistakes in reading scales or careless setting of markers, etc. They can be eliminated by repetition of readings by one or two observers.
- b) **Systematic errors:** These are due to built-in errors in the instrument either in design or calibration. Repetition of observation with the same instrument will not show a spread in the measurements. They are the hardest source of errors to detect.
- c) **Random errors:** These always lead to a spread or distribution of results on repetition of the particular measurement. They may arise from fluctuations in either the physical parameters due to the statistical nature of the particular phenomenon or the judgement of the experimenter, such as variation in response time or estimation in scale reading.

Taking multiple measurements helps reduce uncertainties.

◆ DETERMINING ERRORS:

Although it is interesting and reassuring to compare your results against "accepted values," it is suggested that error analysis be done, before this comparison. One reason is that the accuracy of a set of measurements depends on how well the experiment was done, not how close the measurement was to the accepted value. One could get close to the accepted value, by sloppiness and luck.

Do not base your experimental uncertainties on the accepted values.

Determining the source of uncertainty and the magnitude of this uncertainty is often difficult. Some errors or uncertainties are caused by natural fluctuations or irregularities. These can not be eliminated. To estimate these uncertainties one frequently uses mathematical methods similar to those discussed in the section below titled, Averages and Deviation.

Another method of estimating uncertainties is to assign an uncertainty to the measurement equal to the finest scale reading on the measuring instrument. For example, if a ruler may be marked in millimeters, then the uncertainty in any measurement with this ruler can be given as 1 mm. But with practice, one might be able to interpolate the scale and reduce the error to 0.25 mm.

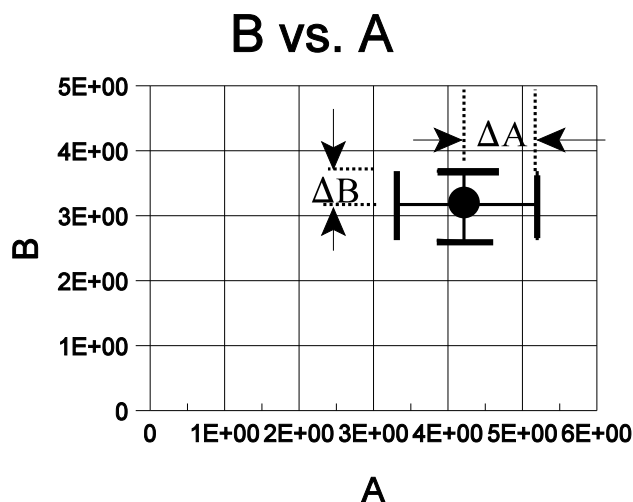
◆ EXPRESSING ERRORS:

For each measured value, A , there is an **estimated error**, ΔA . The complete result is given by $A \pm \Delta A$. This means that the “true value” probably lies between a maximum value of $A + \Delta A$ and a minimum value of $A - \Delta A$. Sometimes the terms **relative error** and **percent error** are used, where:

$$\text{Relative Error} = \frac{\text{Estimated Error}}{\text{Estimated Value}} = \frac{\Delta A}{A}$$

$$\text{Percent Error} = \text{Relative Error} \times 100\% = \frac{\Delta A}{A} \times 100\%$$

Errors are expressed in graphs by using error bars. Consider a data point (A, B) and the associated uncertainties are ΔA and ΔB respectively. The vertical error bar is drawn from $B - \Delta B$ to $B + \Delta B$. Similarly, the horizontal error bar is drawn from $A - \Delta A$ to $A + \Delta A$. For example if $A = 4.2 \pm 0.8$ and $B = 3.2 \pm 0.5$ then this point would be plotted like:



◆ PROPAGATION OF ERRORS

When measured values are used to calculate other values, the uncertainties in these measured values causes uncertainties in the calculated values.

Calculating the uncertainties in the calculated values is called **error propagation**. For the few simple cases that are discussed below, let C be a function of A and B and the associated uncertainties are ΔC , ΔA , and ΔB , respectively.

1) Product with a Constant

Here $C = k A$ where k is a constant. Then

$$\Delta C = k \Delta A.$$

This rule can be applied if k is a measured quantity with a relatively negligible uncertainty, for example, if k were the gravitational constant, g .

2) Addition and Subtraction

Here $C = A + B$ or $C = A - B$. In either case:

$$\Delta C = \sqrt{(\Delta A)^2 + (\Delta B)^2}$$

Note

that ΔC is less than $\Delta A + \Delta B$. This is an expression of the fact that the uncertainties in A and B are independent of each other. (In math lingo, one could say that ΔA , and ΔB are orthogonal or perpendicular to each other. Note how the calculation of ΔC is identical to the Pythagorean theorem for the sides of a right triangle.)

Why should the same formula work for addition and subtraction? Notice that the original uncertainties are squared.

3) Multiplication or Division

If $C = A B$ or $C = A/B$ then:

$$\frac{\Delta C}{C} = \sqrt{\left(\frac{\Delta A}{A}\right)^2 + \left(\frac{\Delta B}{B}\right)^2}$$

$$\Delta C = C \times \sqrt{\left(\frac{\Delta A}{A}\right)^2 + \left(\frac{\Delta B}{B}\right)^2}$$

Again the assumption is that the uncertainties in B and A are independent. An example of when this is not true is $C = A^2$. This brings up another rule.

4) Raised to a Power

This is the case where $C = A^n$, where n is a constant. In this case:

$$\frac{\Delta C}{C} = |n| \frac{\Delta A}{A}$$

5) Graphical Analysis of Uncertainties in Slopes and Intercepts.

If the slope or intercept of a line on a plot is the required calculated value (or the required value is calculated from these values) then the uncertainty of the slope and intercept will also be required.

Graphically one can estimate these uncertainties. First draw the best line possible, and then draw the two lines that just barely pass through the data. The differences of these slopes and intercepts from those of the best fit line provide an estimate of the uncertainties in these quantities.

In the graph below the **solid line** is a good fit to the data and the while the **dotted line** and the **dashed line** represent the extremes for lines that just fit the data.. The dashed line has the steepest slope and will be referred to as the max. line, while the dotted line will be referred to as the min. line. The slopes and y- intercepts for these three lines are:

	Slope	(dif.)	y-inter.	(dif.)
Best fit	1.0		2.0	
Max. line	1.16	(0.16)	-1.5	(-3.5)
Min. Line	0.81	(-0.19)	5.2	(3.2)

Thus if the data was expected to fit the equation:

$$y = a + bx,$$

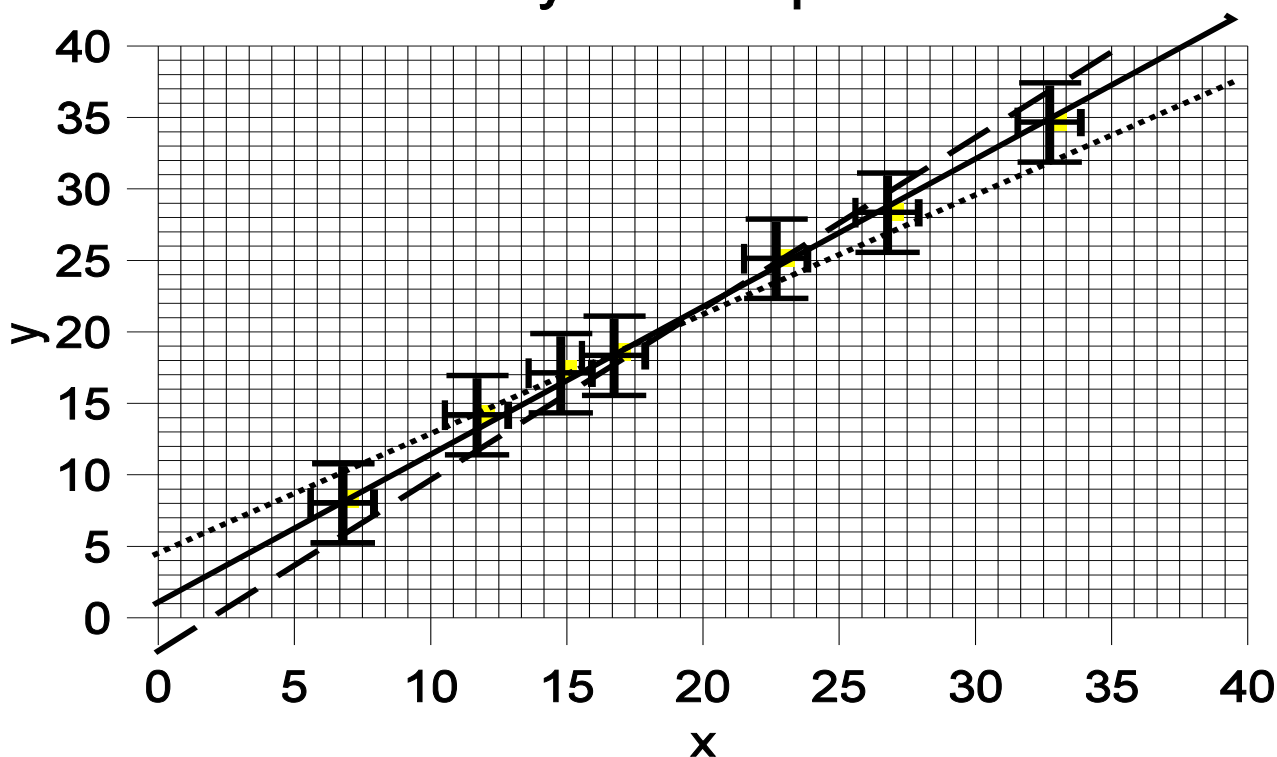
then one would estimate the constants as:

$$a = 2.0 \pm 3.4$$

$$b = 1.0 \pm 0.2.$$

The uncertainties are based on averaging the absolute value of the differences (labeled (dif.)).

Graphical Error Analysis of Slope and y-intercept



6) General Formula (Advanced)

All of the above examples of propagation of errors are special cases of the a general formula. Consider a calculated variable z that is a function of two measured variables x and y , then one writes:

$$z = z(x, y). \quad (7)$$

If the uncertainties associated with x and y are Δx and Δy , respectively. The uncertainty, Δz in z is:

$$\Delta z = \sqrt{\left(\frac{\partial z}{\partial x}\right)^2 (\Delta x)^2 + \left(\frac{\partial z}{\partial y}\right)^2 (\Delta y)^2}. \quad (8)$$

Here a symbol $\frac{\partial z}{\partial x}$ that may be somewhat unfamiliar.

It is the partial derivative of z with respect to x . Partial derivatives are used when a variable is a function of more than one variable. To evaluate a partial

derivative $\frac{\partial z}{\partial x}$, one just takes the normal derivative

$\frac{dz}{dx}$ but treats the variable y as a constant. That is one

does not do any chain rule, $\frac{dz}{dy} \frac{dy}{dx}$, kind of stuff.

Note that the uncertainties are assumed to be independent of each other and add as if they are vectors at right angles to each other.

AVERAGES and DEVIATIONS

The *average* of a series of measurements is one of the most common methods of analyzing data. The average, \bar{x} , or *arithmetic mean* for a series of n numbers: $x_1, x_2, x_3, \dots, x_n$ is defined as:

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \dots + x_n}{n} = \sum_{i=1}^n \frac{x_i}{n}$$

Example: The length of a table was measured three times with the following results in meters:

$$x_1 = 1.42, \quad x_2 = 1.45, \quad \text{and} \quad x_3 = 1.41.$$

Then the average is:

$$\bar{x} = \frac{1.42 + 1.45 + 1.41}{3} = 1.43 \text{ meters}$$

Often one wants to compare individual measurements to the average. The *deviation* is a simple quantity that is frequently used for this kind of comparison. The deviation δx_i of the measurement labeled i , from the average is:

$$\delta x_i = x_i - \bar{x}.$$

Continued Example: The deviations for the above example are:

$$\delta x_1 = 1.42 - 1.43 = -0.01 \text{ meters},$$

$$\delta x_2 = 1.45 - 1.43 = 0.02 \text{ meters},$$

$$\delta x_3 = 1.41 - 1.43 = -0.02 \text{ meters}.$$

STANDARD DEVIATION: If the deviations of a measurement were averaged, the result would be zero because of high and low values would cancel each other. Generally one expresses the fluctuation about the average measurement with by calculating and quoting the **standard deviation**, σ of the n measurements.

$$\sigma = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}}$$

Continued Example: $\sigma = 0.0212$ meter

$$\sigma = \sqrt{\frac{(-0.01)^2 + (+0.02)^2 + (-0.02)^2}{3 - 1}} \quad (20)$$

Finally, the term, standard deviation of the mean defined as:

$$\sigma_m = \frac{\sigma}{\sqrt{n}}. \quad (21)$$

The standard deviation of the mean, σ_m , is the standard measure for describing of the precision of a measurement, i.e. how well a number of measurements agree with themselves. Thus:

$$\Delta x = \sigma_m$$

$$\Delta x = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}} \quad (23)$$

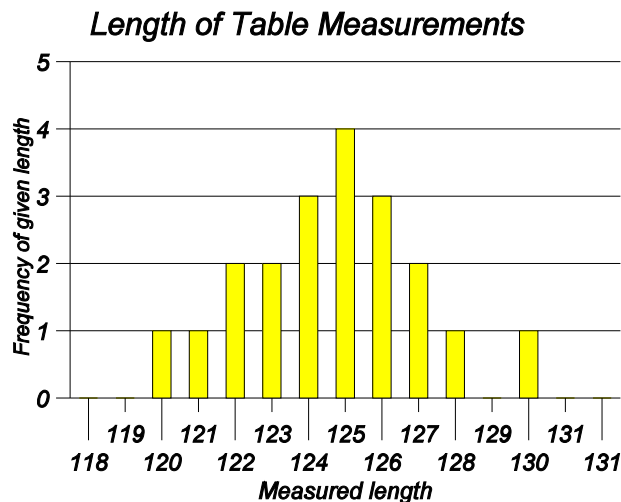
Continued Example: $\Delta x = \sigma_m = 0.012$ meter

◆ GAUSSIAN or NORMAL DISTRIBUTION (Advanced)

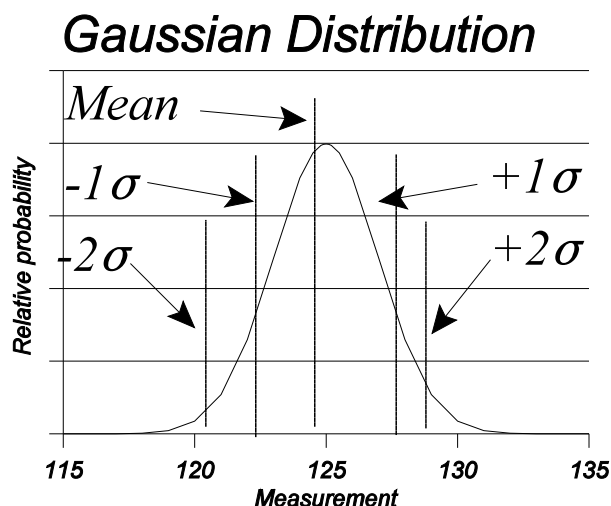
When analyzing data a hidden or implicit assumption is usually made about how results of a number of measurements of the same the quantity will fluctuate about the “true” value of this quantity. One assumes that the distribution of a large number of measurements is a “Nor-mal” or “Gaussian distribution.” Here the term distributed refers to a plot of the number of times the measurement results in a given value (or range of values) vs. the measured result. For example, the result (sort in increasing order) of a table twenty times is (in meters)

1.20,
1.21,
1.22, 1.22,
1.23, 1.23,
1.24, 1.24, 1.24,
1.25, 1.25, 1.25, 1.25,
1.26, 1.26, 1.26,
1.27, 1.27,
1.28,
1.30.

then the results could be displayed in a plot



A plot of a Gaussian distribution is a bell curve:



If the measurements are distributed as a Gaussian distribution then for a large number of measurements:

1. The most frequently occurring value is also the average or arithmetic mean of the all of the measurements.
2. The curve is symmetric about the mean.
3. If an additional measurement is made, there is a 68% chance that the measured value would be within 1 **standard deviation**, 1σ , of the mean and a 95% chance that the measured value would be within 2 standard deviations, 2σ of the mean.

The assumption that measurements are distributed in a Gaussian is sometimes based on solid theoretical grounds, e.g cosmic ray counting. Other times it is based on empirical evidence that a large number of measurements have been made and the distribution matches a Gaussian quite well. But frequently it is just assumed because most random measurements are distributed this way. The reason that scientist make this assumption is that it allows a relatively straightforward theoretical analysis of the uncertainties.

◆ LEAST SQUARES ANALYSIS

(Advanced)

Although extracting a slope or a y-intercept from a graph is relatively straightforward, the method has several limitations. The result is effected by the users skill and bias. Also precision may be lost in the process of graphing and extracting the slope and intercept. Evaluating the uncertainties using graphical methods is usually even more difficult. If data is only evaluated graphically, two different people evaluating the same data would get different slopes.

Part of the scientific method is that the techniques must be reproducible. Usually when a curve (e.g. a straight line) is fitted to some data, the method used is **least squares analysis**. Although this method is quite general, only the limited case of fitting a straight line to data is presented here. For this discussion the following assumptions are made:

- 1) The uncertainties are the same for all of the data points.
- 2) The uncertainties are only in the dependant variable, y .
- 3) The uncertainties are all random and the multiple measurements for a given value of the independent variable, x , would be distributed according to a Gaussian.

The data in this case would be pairs of measurements (x_i and y_i) and the goal is to determine the straight line ($y = a + bx$) that best models this data. One wants to make the minimize the difference between the measured value y_i , and $a + bx_i$ which the associated value calculated from the straight line. Explicitly one is looking for values of **a** and **b** the minimize

$$\chi^2 = \sum_{i=1}^n (y_i - (a + bx_i))^2 \quad (22)$$

χ^2 is pronounced chi-squared and is an indicator of the goodness of fit.

With out any prove the following results of the linear least squares analysis of data yields a straight line, $y = a + bx$, are given:

$$a = \frac{1}{\Delta'} \left[\left(\sum (x_i^2) \right) \left(\sum y_i \right) - \left(\sum x_i \right) \left(\sum (x_i y_i) \right) \right]$$

$$b = \frac{1}{\Delta'} \left[n \times \sum (x_i y_i) - \left(\sum x_i \right) \left(\sum y_i \right) \right]$$

$$\Delta' = n \times \sum (x_i^2) - \left(\sum x_i \right)^2$$

Here all of the sums are from $i = 1$ to n .

Note that all of the individual sums can be built up a point at a time, which make calculation relatively simple. This is especially easy if, **one uses a spreadsheet**. With x and y in the two columns labeled **A** and **B**, one can have column **C** be x^2 , column **D** be y^2 , and column **E** be the product xy . Then sum each column and calculate the values

Many calculators and spreadsheet programs provide access to least squares analysis routines. You are encouraged to learn to use these programs

If one needs to include uncertainties σ_i associated with the y_i in this calculation the above formula becomes:

$$a = \frac{1}{\Delta'} \left[\sum \left(\frac{x_i^2}{\sigma_i^2} \right) \sum \left(\frac{y}{\sigma_i^2} \right) - \sum \left(\frac{x_i}{\sigma_i^2} \right) \sum \left(\frac{x_i y_i}{\sigma_i^2} \right) \right]$$

$$b = \frac{1}{\Delta'} \left[\sum \left(\frac{1}{\sigma_i^2} \right) \sum \left(\frac{x_i y_i}{\sigma_i^2} \right) - \sum \left(\frac{x_i}{\sigma_i^2} \right) \sum \left(\frac{y_i}{\sigma_i^2} \right) \right]$$

$$\Delta' = \sum \left(\frac{1}{\sigma_i^2} \right) \sum \left(\frac{x_i^2}{\sigma_i^2} \right) - \left(\sum \left(\frac{x_i}{\sigma_i^2} \right) \right)^2$$

REFERENCE: Although there are many good books on data and error analysis, the following two books are the authors standard references. The first book is considered to be a standard reference in physics.

Data Reduction and Error Analysis for the Physical Sciences, Second Edition, by Philip R. Bevington and D.Keith Robinson, McGraw-Hill Inc., 1992

Statistical Treatment of Data, by Hugh D. Young, McGraw-Hill Book Company Inc., New York, 1962.