Fundamentals of Data Analysis in Physical Sciences

The intent of this discussion is to give the student a minimal basis for analyzing data generated in this laboratory. The student will find it helpful to refer to this discussion while writing each lab report. Further information on this subject can be found in many books available in the UTC library.

Outline

I. Calculations and Results

A. Obtaining the Best Result

Generally, in order to obtain the best result, several measurements of the same quantity need to be taken and averaged together. Of course, the more measurements that can be made of a quantity, the better the result. Reproducibility is an axiom of science. All measurements, to be valid determinations, must be reproducible. At least three measurements should be made on a quantity before averaging the result or performing any other statistical calculation. With at least three measurements, you can often spot personal errors in your data. If one measurement is completely different from the others, you can check the results by repeating the measurement. You should never exclude a determination from an average unless you are absolutely sure that the measurement is erroneous. The final result from a set of data can be found in different ways. A simple average (summing and dividing by the number of determinations) can be taken, or the result can by found by plotting your data on a graph. In some cases, the latter techniques gives you a better understanding of how good your results are.

B. Significant Digits

When reporting a result, it is important to consider the number of significant digits in the result. The number of significant digits in a result is indicative of the certainty of the result. The number of significant digits you should report directly depends on the measuring equipment you use and the precision of the measuring process. As an example, if you are measuring a distance, and your ruler is only marked to the nearest millimeter, you would never report a measurement as 1.2635 mm. The number of significant digits in a value infers the precision of that value.

In calculations, it is important to keep enough digits to avoid round off error. In general, keep at least one more digit than is significant in calculations to avoid round off error. It is not necessary to round every intermediate result in a series of calculations, **but it is very important to round your final result to the correct number of significant digits.**

C. Reporting your results

Results are usually reported as result ± uncertainty (or error). **The uncertainty is given to one significant digit, and the result is rounded to that place.** For example, a result might be reported as 9.8 ± 0.3 m/s^2. A more precise result would be reported as 9.795 ± 0.004 m/s^2. A result should not be reported as 9.70361 ± 0.2 m/s^2.

Units are very important to any result. It would be very ambiguous to say: The distance from UTC to downtown is 8. This is unclear whether the author means 8 miles, 8 meters, 8 minutes, or 8 apples. Similarly, the units of a value of current must be specified as mA, mA, A, etc.…

II. Dealing with Uncertainties and Errors

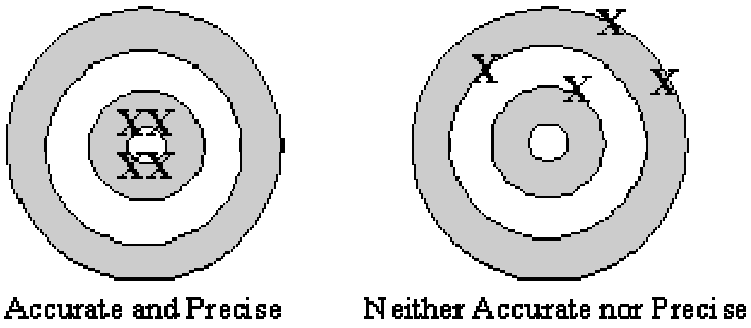
A. Types of Errors

Errors and their sources are a very important part of every scientific experiment. In discussing results, you should always analyze your errors and its possible sources. There are three types of errors that can occur within the experiment: personal error, systematic error, and random error. Personal errors are mistakes on the part of the experimenter. **It is your responsibility to make sure that there are no errors in recording data or performing calculations.**

Systematic errors tend to decrease or increase all measurements of a quantity, (for instance all of the measurements are too large). One example of a systematic error could be if an instrument is not calibrated correctly. A systematic error could also occur if for instance, in an experiment to determine the focal length of a lens, an image was out of focus during every measurement because of the imperfect eyesight of the experimenter. Systematic errors are often the hardest errors to identify, but reasonable guesses can be made.

Random errors are also known as statistical uncertainties, and are a series of small, unknown, and uncontrollable events. Statistical uncertainties are much easier to assign, because there are rules for estimating the size. If you are reading a ruler, the statistical uncertainty is half of the smallest division on the ruler. Thus, if a ruler is marked to the nearest millimeter, the statistical uncertainty associated with any measurement made with that ruler is ± 0.5 millimeters. Even if you are recording a digital readout, the uncertainty is half of the smallest place given. This type of error should *always* be recorded for any measurement.

All types of error imply precision and accuracy. Precision and accuracy in a result are not the same. Precision is a measure of the closeness of separate determinations of a value. Accuracy refers to how close the average of the determinations is to the accepted value. A result can be both accurate and precise, neither accurate nor precise, accurate but not precise, or precise but not accurate (refer to figure below).



Lack of accuracy but good precision implies systematic error, and lack of precision but good accuracy implies random error.

An example:

Accepted distance from Chattanooga to Knoxville = 125 miles

Experimental values of 151, 152, 148, and 149 miles were determined experimentally. The average value is 150 miles. This is a precise value (there is not much spread in determinations) but not accurate (there is a large deviation from accepted value). The small precision implies small random error, but the inaccuracy implies some sort of systematic error.

B. Standard Measures of Error

There are several ways to quantitatively measure the accuracy and the precision of a result. Absolute deviation and relative percentage deviation both measure the accuracy of a result but can only be used when the accepted value is known. The standard deviation measures the precision of a result. It is a widely used statistical value.

---

1. Absolute Deviation

The absolute deviation is simply the difference between an experimentally determined value and the accepted value. Using the above example, the absolute deviation of the experimentally determined distance from Chattanooga to Knoxville is:

$$|125 - 150| = 25 \text{ miles}.$$

The absolute deviation is a measure of the accuracy, *and not* the error itself. Therefore, it would be incorrect to report the value as 125 ± 25 miles.

---

2. Relative % Deviation

The relative percentage deviation is a more meaningful value than the absolute deviation because it accounts for the *relative* size of the error. The relative percentage deviation is given by the absolute deviation divided by the accepted value and multiplied by 100%. Thus, the relative % deviation in the above example is:

$$\frac{|125 - 150|}{125} \times 100\% = 20\%.$$

---

3. Standard Deviation

The standard deviation is a valid result for error, and tells about the precision of your experiment. The standard deviation is found in the following manner. First, the average value is found by summing and dividing by the number of determinations. Then the residuals are found by finding the absolute value of the difference between each determination and the average value. Third, square the residuals and sum them. Last, divide the result by the number of determinations - 1 and take the square root. From the example above:

Values	Residuals	Residuals^2	
148	=| 148-150| = 2	= 2^2 = 4	
149	=| 149-150| = 1	= 1^2 = 1	
151	=| 151-150| = 1	= 1^2 = 1	
152	=| 152-150| = 2	= 2^2 = 4	
Average = 150		sum = 4+1+1+4 = 10	Standard deviation =
			$\sqrt{\frac{10}{(4-1)}} = 1.82$

Thus, the result would be reported as 150 ± 2 miles.

The mathematical equation for the average is:

$$\text{average} = \langle x \rangle = \frac{\sum_{i=1}^{N} x_i}{N},$$

and the mathematical expression for standard deviation is:

$$\sigma = \sqrt{\frac{1}{(N-1)} \sum_{i=1}^{N} \left| x_i - \bar{x} \right|^2}.$$

C. Propagation of Errors

We have previously discussed reporting an error value for a direct measurement, but we have not discussed how errors in each measurement propagate through a calculation to yield an error in the result. In the case of a simple addition, (C = A + B) then the error in A, ΔA, and the error in B, ΔB, would simply add to give an error in C, ΔC.

$$\Delta C = \Delta A + \Delta B.$$

For the subtractive case of C = A - B, the above rule would still be true. The errors *always* add.

The cases for multiplication and division follow a different rule than addition and subtraction. Correct error propagation techniques for multiplication and division always keep in mind that relative errors add. An absolute error in the result, C, is the uncertainty, (i.e., $\pm \Delta C$), where we use Δ to denote an absolute error. The relative error, which we denote by δ, is the absolute error divided by the value (i.e., $\Delta C / C = \delta C$). The rule is:

$$\delta C = \delta A + \delta B .$$

Take the following case as an example. The volume of a cylinder is given by $V = \pi R^2 L$. If you experimentally determined the radius of the cylinder to be 5.48 ± 0.05 cm and the length of the cylinder to be 14.75 ± 0.05 cm, the volume of the cylinder will be

$$V = (3.14)\ (5.48\ \text{cm})^2\ (14.75\ \text{cm}) = 1.39 \times 10^3\ \text{cm}^3.$$

However, it would be incorrect to say that the error in the volume is ± 0.05 cm^3, simply because the errors in the measurements of R and L are ± 0.05 cm. It would also be incorrect to use the volume formula for the error propagation, (i.e., ΔV ***<u>is not</u>*** = (3.14) $(0.05cm)^2$ (0.05 cm)) The measurement of the length of the cylinder had an absolute error of 0.05 cm, as did the radius. Their relative errors are then given by:

$$\delta R = \frac{\Delta R}{R} = \frac{0.05\ \text{cm}}{5.48\ \text{cm}} = 0.00912 \ \text{ and } \delta L = \frac{\Delta L}{L} = \frac{0.05\ \text{cm}}{14.75\ \text{cm}} = 0.00339$$

.

The relative error in the result is then the sum of the relative errors of the variables. Thus,

$$\delta V = \delta R + \delta L = 0.00912 + 0.00339 = 0.0125 .$$

The absolute error in the volume, ΔV, is the relative error, δV, times the volume.

$$\Delta V = (\delta V)\ (V) = (0.0125)\left(139 \times 10^3\ \text{cm}^3\right) = 17.41\ \text{cm}^3$$

.

Therefore, the result should be reported as V = 1390 ± 20 cm^3. Remember, you should round your results to the correct number of significant digits.

A more correct error propagation for the above rules would be to add the uncertainties "in quadrature," which means to square and sum the errors. Thus for addition the most correct error would be,

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

and for multiplication,

$$\delta C = \sqrt{\delta A^2 + \delta B^2}.$$

When the function is not merely addition, subtraction, multiplication, or division, the error propagation must be defined by the total derivative of the function. Consider the case of the function: f = 1/T. The derivative of this function (remembering for error propagation to ignore negative signs) is:

$$\left|\frac{df}{dT}\right| = \frac{1}{T^2}, \quad \text{or} \quad df = \frac{1}{T^2}dT, \quad \text{or} \quad \Delta f = \frac{1}{T^2}\Delta T.$$

Similarly, for the case of a natural logarithm, or B = ln (A),

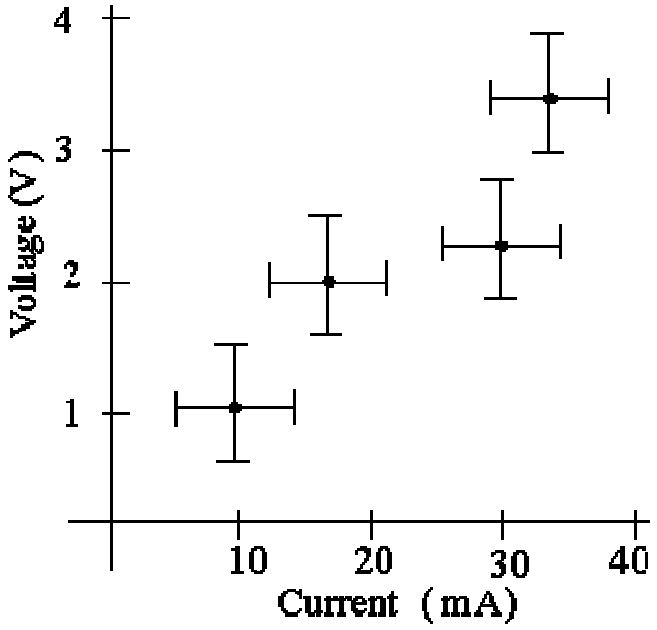
$$\left|\frac{dB}{dA}\right| = \frac{1}{A}, \quad \text{or} \quad dB = \frac{1}{A}dA, \quad \text{or} \quad \Delta B = \frac{1}{A}\Delta A.$$

III. Graphing

A. Important Aspects of a Graph

There are several things to keep in mind when making a graph of your data. **The variables must be labeled on the axes. Also, do not forget the units!**

When plotting a data point on a graph, it should be kept in mind that the measurements used to create the data point had uncertainties associated with them. The way of keeping track of these uncertainties is by using error bars. The error bar is a line drawn through the data point that covers the range of uncertainty and are as shown below.



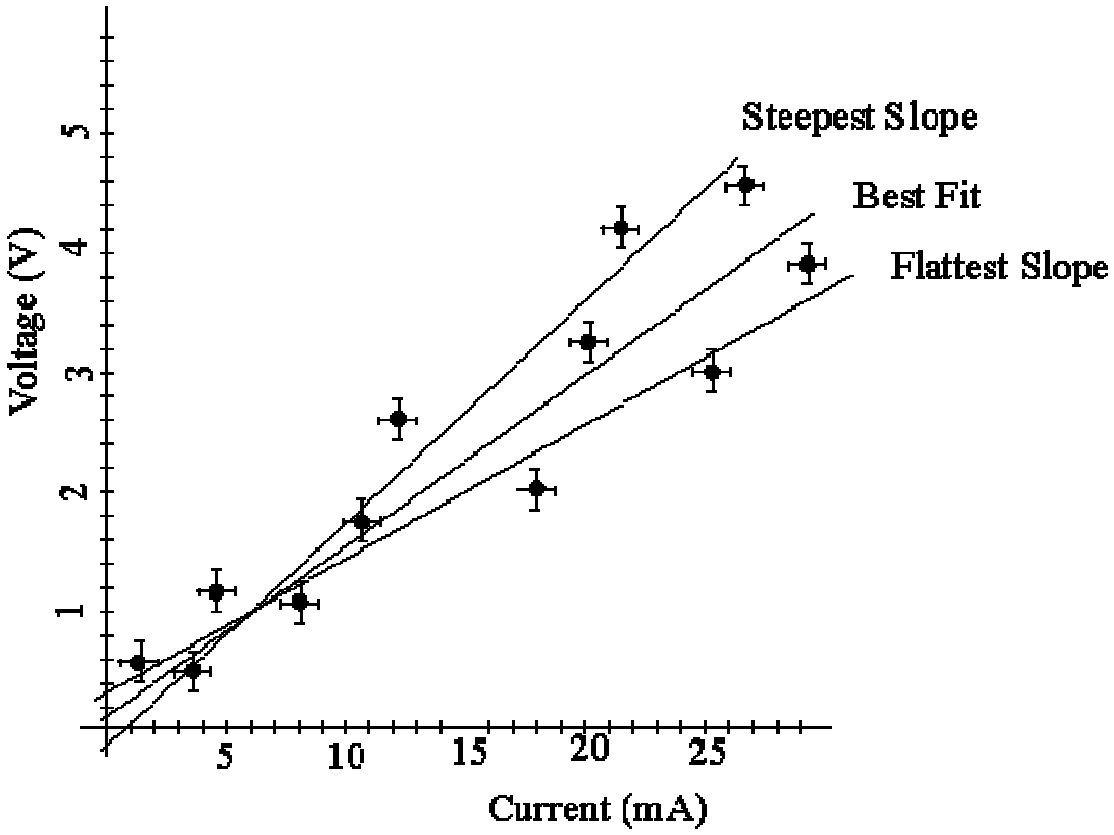
B. Fitting the data

In physical sciences, we try to find relationships between variables. In order to derive a simple relationship, we would try to draw the simplest possible curve that is compatible with the data. This is called the best fit, and is drawn through as many error bars as possible, but does not necessarily go through *all* the data points. We never simply "connect the dots". **The equation of any curve fit to the data should be clearly labeled on the graph.**

The most common fit is a linear fit. The best fit line through the data points will have the equation y = mx + b, where m is the slope of the line and b is the y-intercept. A linear fit might be used for the above graph.

As with all measurements, there is uncertainty associated with the slope and intercept of a fit. The uncertainty in the slope, Δm, is depicted below. Δm can be found by drawing the steepest slope and the flattest slope,

$$\Delta m \approx \frac{m_{max} - m_{min}}{2}.$$



A better data fit can be performed by a computer or calculator (linear regression analysis) - you can use MacCurve Fit, Excel, Data Logger or another program. Normally, the program should report m, Δm, b, Δb and R^2. It should be noted that correlation coefficient, R^2, is a measure of the linearity of the data, but is not the directly the error in the slope and should not be reported as such. A linear fit might be performed on data suspected to follow Ohm's Law (V = I R). If voltage, V, is plotted on the y-axis, and current, I, on the x-axis, then the slope of the line, m, will equal the resistance, R. The y-intercept of the graph would theoretically equal zero.

Another type of fit to data could be a exponential fit. This would be described by the equation $y = b_0 e^{b_1 x}$. This type of fit might be used to describe the relationship between voltage and current in a forward-bias diode. This relationship is $I = I_0 e^{\frac{eV}{k_B T}}$. Thus, if current, I, is plotted on the y-axis, and voltage, V, on the x-axis, then the coefficient $b_0 = I_0$ and the coefficient $b_1 = e/k_B T$.

Other fits that are frequently used include a polynomial of degree 2 ($y=A+Bx+Cx^2$), the power function ($y=Ax^B$), the logarithmic function ($y=\ln(x)$), and sinusoidal function ($y=A\sin(Bx)$).