Fundamentals of data, graphical, and error analysis

1. Data measurement and Significant Figures

Whenever we take a measurement, there are limitations to the data and how well we can measure a specific value for a quantity. The number of significant figures in a reported value is what determines how well the measurement can be made. For an analog instrument, the significant figures in the experimental measurement include all the numbers that can be read directly from the instrument scale plus one additional estimated number. For a digital instrument, the significant figures in the experimental measurement include all numbers that can be directly read from the scale. Writing the number in scientific notation explicitly shows all the significant figures in your measurement. This procedure is helpful in expressing the significant figures in very large (and very small numbers).

Digits are normally counted as significant if they are not zero. General rules for zeros are: if a *trailing* zero is reported to the right of a decimal point, the zero is significant. *Leading* zeros are not significant. If a trailing zero is to the left of the decimal point, it is not significant.

Examples:

The number 3,564,000 and 0.0001230 each have four significant figures.

These are easily expressed in scientific notation as: 3.564×10^6 and 1.230×10^{-4}.

Multiplication and Division: In the multiplication or division of two or more measurements, the number of significant figures in the final answer is equal to the least number of significant figures in the measurements.

Examples: 1. $v = \frac{d}{t} = \frac{2345.2 \text{ m}}{5.2 \text{ s}} = 451.0$ m/s. The result would be reported as 4.5×10^2 m/s

2. $d = vt = (20.5 \text{ m/s}) \times (30.52 \text{ s}) = 625.66$ m. The result is: 6.26×10^2 m.

Addition and Subtraction: In addition and subtraction, the number of significant figures in the final answer depends on the number of decimal places in the measurements. No more numbers than the fewest decimal places should be retained.

Examples: (123.56 m) + (12.351 m) = 135.911 m. The result is: 135.91 m.

(123.56 m) - (12.3 m) = 111.26 m. The result is 111.3 m.

In more complex experiments or calculations where several steps are involved, the most precise result is obtained by keeping all the digits throughout the calculation without regard to significant figures. The end result should be reported to the correct number of significant digits based on the original measurement of data. Therefore, **the most general rule for determining how many significant figures your result should have is to determine how many significant figures your initial measurement had, and report the final result to the same number of significant figures**.

2. Graphs

Graphs provide a visual representation of data and provide a faster form of communication than tables of data. Graphs may also be used as an analytical tool to determine the value of some quantity that signifies a relationship between variables.

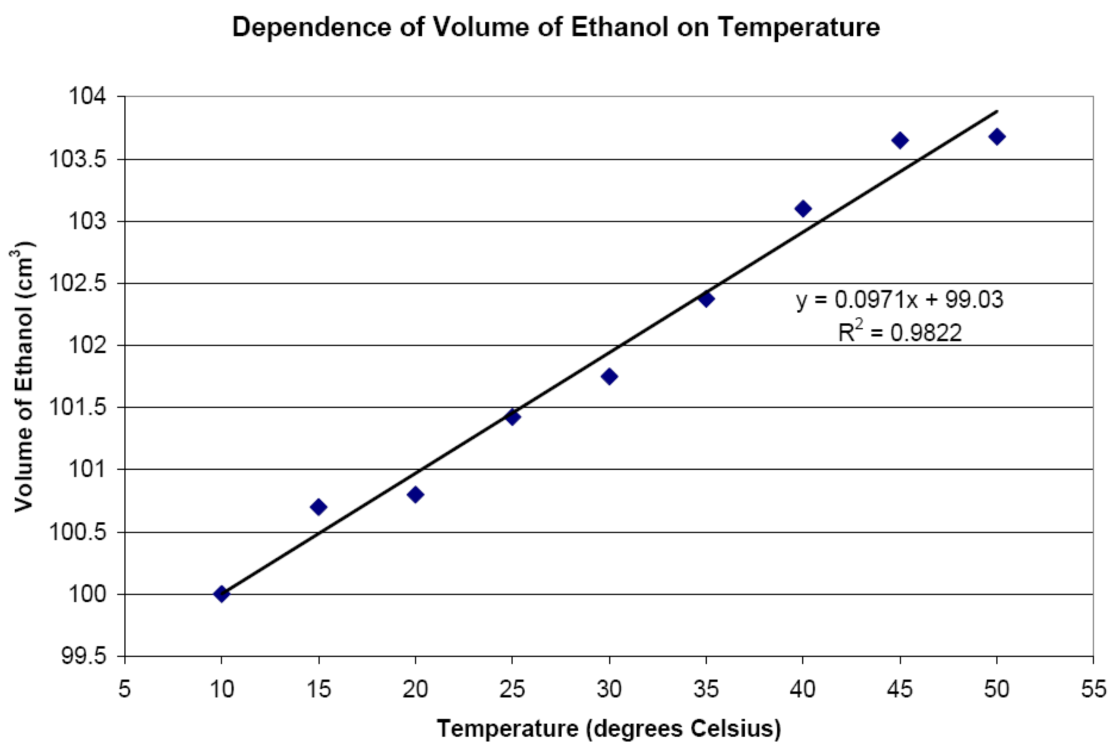
The independent variable (the variable that you are changing in the experiment) is plotted on the horizontal or x-axis. The dependent variable (the variable that you are measuring and that changes in response to manipulation of the independent variable) should be plotted on the vertical or y-axis. The scale of the graph should be chosen so that the data occupies as much of the available space as possible – this allows for accurate visual determination of quantities from the graph. The scale on each axis should be divided evenly, but the scales on both axes do not have to be the same size and in some cases do not have to start at zero.

The following should be included with every meaningful graph: (1) A self-explanatory title. This should be in the form of "(The dependent variable) as a function of (the independent variable)" or "(quantity on the y-axis) vs. (quantity on the x-axis)" or "The dependence of (the dependent variable) on (the independent variable)". (2) Axis labels with units. If either the variable *OR* the unit is missing, the graph does not make sense.

For accurate data interpretation, a best-fit line (with an appropriate equation that describes it) needs to be drawn on the graph. It is incorrect to connect the data points with straight line segments. This implies a direct mathematical relationship between the points, and does not account for measurement error. The best fit line should pass as close to as many points as possible, but need not pass through all of them. Many mathematical functions can be used as a best fit; common ones are linear fits and exponential fits. When choosing the type of equation to fit, it needs to follow the theory for the relationship between the independent and dependent variables. Fitting data to a polynomial if the theory shows an exponential relationship would be incorrect. You should choose the equation with the fewest number of variables that fits the data. It is sometimes possible to fit experimental data to a complex equation where the fit is extremely good, but additional terms in the equation must have physical meaning. When using a computer program to determine a best-fit equation for the data, always give an estimate of the goodness of fit for the equation (example, R^2 or MSE).

Example of a good graph:

Let's say we are doing an experiment to determine the volume thermal expansion of ethanol. In this experiment, we have an initial volume of ethanol (100 cm^3) at a temperature of 10 degrees Celsius. We raise the temperature to different values, measuring the change in the volume of ethanol (dependent variable) as a function of the temperature (independent variable). A good graph of our experimental data, that is *well spaced, has a title, axis labels with units, and a best-fit line with equation and goodness of fit* would look like:



3. Error Analysis

A. Types of Errors

Errors and their sources are a very important part of every scientific experiment, because they indicate how certain you are in the accuracy of your final reported result. In discussing results, you should always analyze your errors and the possible sources. There are three types of errors that can occur within the experiment: measurement uncertainty, random error, and systematic 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.

1. Measurement uncertainty: how precisely the value can be measured with a given measurement tool. For analog measuring tools the uncertainty is ± 0.5 of the smallest division; for digital devices the uncertainty is one in the smallest place given. This type of error should be recorded for all measurements, as it will be important to the determination of error in the final result.

Example: For a ruler or meterstick (analog device), the statistical uncertainty is half of the smallest division on the ruler. Thus, if a ruler is marked to the nearest millimeter, and you measure a value of 46.2 mm, the measurement would be recorded as 46.2 ± 0.5 mm.

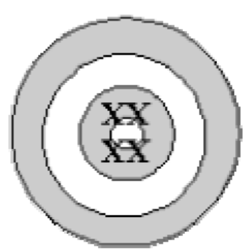
2. Random errors are also known as statistical uncertainties, and are a series of small, unknown, and uncontrollable events, such as a fluctuation in temperature or voltage from the wall outlet. This is the “noise” in an experiment. They come into play when there are multiple measurements of series of random events. To estimate the result, one needs to find average and standard deviation (see below). Taking multiple measurements of the same data point reduces the random error associated with a quantity. This is also known as increasing the signal to noise ratio.

3. 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.

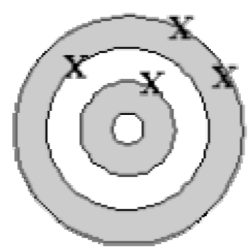
All types of error imply precision and accuracy. Precision and accuracy in a result are not the same.

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.

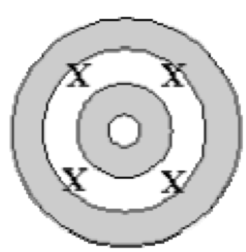
A result can be both accurate and precise, neither accurate nor precise, accurate but not precise, or precise but not accurate. High precision does not imply high accuracy. Lack of accuracy but good precision implies systematic error, and lack of precision but good accuracy implies random error.



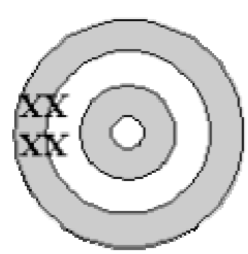
Accurate and Precise



Neither Accurate nor Precise



Accurate but not Precise



Precise but not Accurate

Example: Given that the accepted distance from Chattanooga to Knoxville = 125 miles, but 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 the 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 and can be used to describe the precision of your data when the accepted value of a measured quantity is not known. It is a widely used statistical value.

1. 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:

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

The absolute deviation is a measure of the accuracy of your determination, *and not* the error itself. It would be incorrect to report the value as 150 ± 25 miles.

2. Relative % 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, then multiplied by 100%. Thus, the relative % deviation in the above example is:

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

The percent deviation is not the error in your measurement; hence, it would be incorrect to report your result as 150 miles ± 20%. The proper way of using this statistic would be to report: The distance from Chattanooga to Knoxville was determined to be 150 miles. This represents a 20% deviation from the accepted value for that distance.

3. Standard Deviation is a valid result for error, and tells about the precision of your experiment. For a large data set with a normal distribution, 66% of the data falls within one standard deviation of the mean value; 95% of the data falls within two standard deviations. 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 minus one and take the square root. The mathematical expression for the average of a number of determinations is:

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N}$$ and the expression for the standard deviation is $$\sigma = \sqrt{\frac{1}{(N-1)} \sum_{i=1}^{N} \left| x_i - \bar{x} \right|^2}$$

Many calculators and software programs have this function; following is the computational method used.

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

So the standard deviation = $\sqrt{\frac{10}{(4-1)}} = 1.82$

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

C. Propagation of Errors

We have discussed reporting an error value for a direct measurement, but have not discussed how errors in each measurement propagate through a calculation to yield an error in the result.

Addition and Subtraction: For 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.

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, denoted by δ, is the absolute error divided by the value, $(\Delta C/C) = \delta C$. The rule for both multiplication and division is:

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

Example: Sides A and B of a rectangle are measured: A=5.4±0.05 cm, B=7.8 ±0.05 cm.

The area of the rectangle would be: $C=A*B = 42.12$ cm^2.

The relative error in the area would be: $\delta C = \delta A + \delta B =$

$$\frac{\Delta C}{C} = \frac{\Delta A}{A} + \frac{\Delta B}{B} = \frac{0.05}{5.4} + \frac{0.05}{7.8} = 0.0157$$

So the absolute error for C, $\Delta C = C * (\delta C) = 42.12 * 0.0157 = 0.6613$

The error should be rounded to 1 significant digit, so $\Delta C = 0.7$ cm^2.

Then round the result to the same decimal place: C = 42.1 cm^2

The result should be reported as: the area of the rectangle is 42.1 ± 0.7 cm^2

Error propagation for more complex functions than addition/subtraction, and multiplication/division is beyond the scope of this course.

Adapted by Dr. Kristin Whitson from material prepared by Mrs. C. I. Lane and Dr. Tatiana Allen. Modified Sept. 2009.