Data Analysis

From PHYS 161, you should have learned:

1. Quantitative scientific statements include appropriately precise numerical values, uncertainties of matching precision, and units to identify what in fact the numbers refer to.

2. The number of decimal places of the uncertainty number determines the number of decimal places of the reported number; the two must agree.

3. Agreement of results or of a result with prediction is a matter of differences relative to, or in units of, the uncertainty of the difference.

4. When a numerical result depends on the arithmetical combination of measurements, the uncertainty of the result must be determined by error propagation.

In PHYS 161, furthermore, you had reference to formulas for a few simple, specific arithmetical combinations. When certain assumptions are met, these can be applied in series, to get a final uncertainty for a more complicated expression. A general approach is generally more efficient and comprehensive.

Propagation of Errors

Let us say that the final result, \( z \), depends on two, independent sets of measurements, \( x \) and \( y \), according to some functional relationship \( f \):

\[
z = f(x, y).
\]

(1)

Knowing the functional relationship \( f \) as well as the uncertainties of \( x \) and \( y \), \( \sigma_x \) and \( \sigma_y \), respectively, we determine the uncertainty of \( z \), \( \sigma_z \), by

\[
\sigma_z = \sqrt{\left(\frac{\partial f}{\partial x}\right)^2 \sigma_x^2 + \left(\frac{\partial f}{\partial y}\right)^2 \sigma_y^2},
\]

(2)

where \( \frac{\partial f}{\partial x} \) is the partial derivative of \( f \) with respect to \( x \) (that is, the derivative of \( f \) is taken with respect to \( x \) as if \( x \) is the only variable in \( f \), while all other terms in the function are treated as constants). And similarly for \( y \). This expression can be extended to any number of independent measurements included in the relationship.

Suppose that \( z = xy^2 \). Then, \( \frac{\partial z}{\partial x} = y^2 \) and \( \frac{\partial z}{\partial y} = 2xy \), so

\[
\sigma_z = \sqrt{y^4\sigma_x^2 + 4x^2y^2\sigma_y^2} = xy^2 \sqrt{\left(\frac{\sigma_x}{x}\right)^2 + \left(\frac{2\sigma_y}{y}\right)^2}.
\]

1The assumption here is that statistical uncertainties are uncorrelated and far larger than systematic uncertainties.

2If you don’t know the elements of this list, now is a very good time to learn them.
Linearization

An experiment is often designed to discover the relationship between variables, rather than the values of the variables. This means that a functional form and the parameters of the function are determined. If the relationship happens to be linear (proportional), a scatter plot of the values of the dependent variable against the corresponding values of the independent variable exhibits a constant slope. Non-linear relations, on the other hand, can result in complicated or hard to distinguish curves, and the determination of both form and function can be tricky.

Some spreadsheet and data-analysis software packages, such as Excel, provide a catalog of possible fits. Many physical relationships can be described with the standard functions available in, say, Excel, such as a line, a polynomial, a power law, an exponential, or a trigonometric function. Yet, particularly over a small domain of the independent variable, these can prove to be difficult to distinguish by eye. The fitting algorithms of these packages often provide a measure of the quality of a certain fit, the so-called coefficient of determination, \( R^2 \). Its value lies between 0 and 1: the closer to 1, the better the fit describes the distribution. Thus, one can run through a set of possible functions to choose a best representation.

Excel also provides the equation of a fit function. That is, it gives the function’s parameters, but not the uncertainties of those parameters, and so the information is incomplete. Some functions—particularly, the power law and the exponential—can, however, be made to appear graphically linear if the data are transformed appropriately, a procedure we might call linearization. Such transforming of data so that their graph appears linear thereby could help reveal the nature of the relationship should \( R^2 \) values be too similar to make a reasonable choice between functions. Better yet, once linearized, the transformed data can be run through Excel’s regression utility, which will provide sufficient information to determine the parameters as well as their uncertainties.

Consider the power law \( y = Cx^j \), where \( C \) and \( j \) are parameters to be determined. Taking the natural logarithm of both sides of this equation

\[
\ln y = \ln C + j \ln x
\]

so that \( \ln y = \ln C + j \ln x \), which has the form of a straight line, \( y = b + mx \), where \( b \) is the \( y \)-intercept of the line and \( m \) is its slope. So, if the relationship between \( y \) and \( x \) is a power law, then the graph of \( \ln y \) versus \( \ln x \) will appear linear. The regression of \( \ln y \) versus \( \ln x \) yields a value for the slope, which gives the power \( j \) and the intercept \( \ln C \). The parameter \( C \) can then be found by taking the exponential of the intercept. The regression utility also provides the uncertainty of the slope and intercept values. The value for the power \( j \) that will be reported can therefore simply be read off. If we label the intercept \( b = \ln C \), then, again, \( C = e^b \), and, as you should be able to show, the uncertainty of \( C \) is \( \sigma_C = C \sigma_b \), where \( \sigma_b \) is the uncertainty of the intercept.

Now, consider the exponential \( y = De^{kx} \), where \( D \) and \( k \) are parameters to
be determined. Taking the natural log of both sides of this equation
\[ \ln y = \ln D e^{kx} = \ln D + \ln e^{kx} = \ln D + kx \]
so that \( \ln y = \ln D + kx \), which has the form of a straight line, \( y = b + mx \), where \( b \) is the \( y \)-intercept of the line and \( m \) is its slope. So, if the relationship between \( y \) and \( x \) is an exponential, then the graph of \( \ln y \) versus \( x \) will appear linear. The regression of \( \ln y \) versus \( x \) yields a value for the slope which gives the power \( k \) and the intercept \( \ln D \). The parameter \( D \) can then be found by taking the exponential of the intercept. The regression utility also provides the uncertainty of the slope and intercept values. The value for the power \( k \) that will be reported can therefore simply be read off. If we label the intercept \( b = \ln D \), then, again, \( D = e^b \), and, as you should be able to show, the uncertainty of \( D \) is \( \sigma_D = D \sigma_b \), where \( \sigma_b \) is the uncertainty of the intercept.

This approach to linearization works when the relationship is a power law or exponential, but not otherwise. Some equally simple and commonplace relationships, such as those represented by polynomials or trigonometric functions, will not become linear when the data are transformed with logarithms.

Methods for determining the parameters of such functions as sinusoids and multi-term polynomials (recall the function of position with respect to time of an object under the influence of a constant force, \( x = x_0 + v_0(t - t_0) + \frac{1}{2}a(t - t_0)^2 \), where \( x \) is the dependent variable, \( t \) the independent variable, and \( x_0, v_0, a, \) and \( t_0 \) are all parameters to be determined from the data) are beyond the scope of this course. If the relationship is a two-term polynomial, of the form \( y = Ax^n + C \), where \( A, n, \) and \( C \) are all parameters, you might get a hint of \( n \) from the Excel fitting function using the polynomial option and seeing that all coefficients of terms with powers other than \( n \) are much smaller than \( A \). If such is the case, then one can transform the independent variable by raising it to the power \( n \) and checking that the graph of the dependent variable against the transformed independent variable is a straight line. Regression will then give you the slope, which is the value of the parameter \( A \), and the intercept, the parameter \( C \), with their respective uncertainties.