GRAPHICAL ANALYSIS OF DATA

Often in physics, we graph data and determine physical parameters from a curve fit. In these cases, we can use the errors in the fit parameters calculated by the graphing software as the absolute errors. While the particulars of these calculations are beyond the scope of this class, we should take a moment to consider how the computer determines the best fit to data.

Many physical relationships link two quantities through proportionality. For example, Hooke's law, $F = -k\,x$, states that the force on a spring is proportional to its extension. A number of the labs during the semester will have similar linear relationships.

Let's assume that you have taken the following data (a negative distance means the spring was compressed rather than stretched):

x (cm) (±0.2 cm)	F (N)	u(F) (N)
-3.0	20	5
-1.0	7.5	1.5
1.5	-7.5	2.2
3.0	-10	2
4.0	-15	3

- Plot these points on graph paper, including error bars representing the uncertainty, and draw the best straight line representing the data. (Use a full sheet of graph paper and make the graph as large as possible.)
- Find the equation describing this line (i.e. find the slope and y-intercept).

Comment in your notebook on your strategy for determining the "best" possible fit.

Residuals: One standard way of measuring how well a straight line fits the data is to see how much each point misses the straight line in a vertical direction. These are called **residuals**. For a series of data points *x* and *y* and a fit equation *y(x)=mx + b*, the residual for a single point y_i is

$$R_i = y_i - (mx_i + b) = y_i - y(x_i)$$

Since some of these distances will be positive and others negative, we square each one. If we add up the sum of the squares of the residuals, we tell a measure of how good a straight line fits the data.

$$\sum_{i=1}^{N} {R_i}^2 = \sum_{i=1}^{N} (y_i - y(x_i))^2$$

The lower the sum of the squares of the residuals, the better the fit. Some software will calculate a "Root Mean Square Error", or "RMSE", which is

$$RMSE = \sqrt{{\sum {R_i}^2}/{N}}$$

Calculate the RMSE for your best fit line. (You can estimate your residuals by reading them off the graph if you like.)

Least-Squares Fit: Graphing software packages like Kaleidagraph and LoggerPro (the two we will use in Physics 3) use a least squares analysis to fit the line. The formulas involved in finding the best fit and uncertainties are quite complex, but they key idea is pretty simple. To get the best fit to a data set, one must minimize the sum of the squares of the residuals for the data points. This gives the value of the constants for the slope and y-intercept. The software computes these constants and then estimates the uncertainties in the slope and y-intercept based on the size of $\sum {R_i}^2$.

Now graph and fit your data in Kaleidagraph. Use the reference at the front of the lab manual for help in formatting your data table and making the plot.

- First, enter the data into the table. You should have three columns: one for position, one for force, and one for the uncertainty in the force. Make sure your columns are labelled.
- Next, make a graph of force vs. position. (This means position is on the horizontal axis.) Include the error bars for both position and force uncertainty.
- Perform a "General" curve fit to find a best fit line. Follow the instructions in the Kaleidagraph reference to perform a weighted fit. This takes the size of the uncertainities into account. A data point with large error bars is given less weight or importance in finding the best fit.

Notice that you get lots of information in the Equation Box. In the equation

m1 + m2*M0

m1 represents the y-intercept, *m2* the slope, and *M0* the independent variable (the data plotted on the horizontal axis). Notice there is a column labeled "Error", which gives the uncertainty estimates in the parameters *m1* and *m2*.

In our example, the absolute value of the slope represents the spring constant, *k*. Based on your graph, what is *k ± u(k)*? What are the units of this constant?

To see how the size of the uncertainties affects the weighting of the fit, try changing the value of the uncertainty of one of the force values. (Make a large u(F) really small or a small u(F) really big.) Update the plot at see what happens to the slope and its uncertainty. (At the top of the Data window, the second button from the left is the shortcut to update the plot.)

There are two other statistics displayed in the box. The value *R*, is a correlation coefficient that quantizes "goodness of fit". The formula for this is not very pretty, but essentially *R*=1 is a perfect fit to the data. (We really do not use this number in Physics 3.)

The value identified rather cryptically as *Chisq* is the statistic **chi-squared** (χ^2), defined as

$$\chi^2 = \sum_{i=1}^{N} \frac{(y_i - y(x_i))^2}{u(y_i)^2}$$

Notice this is based both on the residuals and on the size of the uncertainty in the dependent variable. (We do not use this statistic much, either.)

THE MATHEMATICAL DETAILS

If you want to know more about how slope, y-intercept, and their uncertainties are calculated by the computer, this is for you. Otherwise, you can stop reading now!

The least squares fitting method solves for values of slope, *m*, and y-intercept, *b*, that minimize the sum of the squares of the residuals.

$$\sum_{i=1}^{N} {R_i}^2 = \sum_{i=1}^{N} (y_i - y(x_i))^2 = \sum_{i=1}^{N} \big(y_i - (mx_i + b)\big)^2$$

Let's call this quantity *S*. If you expand this expression, you get:

$$\mathrm{S} = \sum {y_i}^2 + 2\mathrm{m} \sum x_i y_i - 2b \sum y_i + m^2 \sum {x_i}^2 + 2mb \sum x_i + Nb^2$$

To minimize *S*, the best choices for *m* and *b* are values that make the derivatives of *S* with respect to those values equal to zero.

$$\frac{\partial S}{\partial m} = 0 \text{ and } \frac{\partial S}{\partial b} = 0$$

(The "curly d" indicates a partial derivative. This simply means to take the derivative while holding everything constant except for the variable of interest.)

Solving these two equations gives the solutions:

$$m = \frac{N\sum x_i y_i - \sum x_i \sum y_i}{N\sum {x_i}^2 - (\sum x_i)^2}$$

$$b = \frac{\sum {x_i}^2 \sum y_i - \sum x_i y_i \sum x_i}{N\sum {x_i}^2 - (\sum x_i)^2}$$

And by propagating the uncertainties in *x* and *y* (something we will learn about in the next lab), the uncertainties are found to be:

$$u(m) = \sqrt{\frac{\sum {R_i}^2}{N\sum {x_i}^2 - (\sum x_i)^2}\left(\frac{N}{N-2}\right)}$$

$$u(b) = \sqrt{\frac{\sum {R_i}^2 \sum {x_i}^2}{N(N\sum {x_i}^2 - (\sum x_i)^2)}\left(\frac{N}{N-2}\right)}$$

where

$$R_i = y_i - (mx_i + b).$$

(Whew!)