A.6 Combining Data: Fits and Averages

We have already seen how repeated measurements of a quantity can be averaged to obtain an improved estimate of its true value. This is a simple case of combining data. We shall often meet more complex cases: for example, we want to combine two measurements 1.5 ± 0.2 and 1.70 ± 0.05. It is intuitively obvious that the true value is likely to be between 1.5 and 1.70, but probably closer to the latter (the more precise value).

The Weighted Mean

Suppose you have measured the same quantity using several different techniques, which naturally give different errors. How do you combine the results to get the best possible estimate of the true value?

Clearly, you want to somehow weight your data so that the more precise values have more influence on the final answer. For Gaussian variables with independent errors, the appropriate weight is $1/\sigma_i^2$. In other words, we define the weighted mean by

$$\boxed{\bar{x}_w = \frac{\sum_i x_i/\sigma_i^2}{\sum_i 1/\sigma_i^2}}$$

and the standard error of the weighted mean by

$$\boxed{\sigma_{\bar{x}_w} = \left(\frac{1}{\sum_i \frac{1}{\sigma_i^2}}\right)^{\frac{1}{2}}}$$

Both of these expressions reduce to the usual values when the $\sigma_i = \sigma$ (the errors are all the same):

$$\bar{x}_w = \frac{\sum_i x_i/\sigma^2}{\sum_i 1/\sigma^2} = \frac{1/\sigma^2 \sum_i x_i}{1/\sigma^2 \sum_i (1)} = \frac{\sum_i x_i}{N} = \bar{x}$$

and

$$\sigma_{\bar{x}_w} = \left(\frac{1}{\sum_i \frac{1}{\sigma^2}}\right)^{1/2} = \left(\frac{1}{\frac{1}{\sigma^2}\sum_i (1)}\right)^{1/2} = \left(\frac{\sigma^2}{\sum_i (1)}\right)^{1/2} = \left(\frac{\sigma^2}{N}\right)^{1/2} = \frac{\sigma}{\sqrt{N}} = \sigma_{\bar{x}}$$

In the weighted average, data points with large uncertainties are guaranteed to contribute almost nothing to the overall mean. Provided that N is reasonably large, the weighted and unweighted means should be roughly the same. If they give drastically different values, it is likely that your error estimates are off.

The Unweighted Linear Least Squares Fit

A common situation is the case where two variables or suitable functions of two variables are linearly related:

$$y = a + bx,$$

where a and b are unknown constants. Typically, the data consist of N pairs of observations (x_i, y_i) and the desired physical quantity is a or b or both. The problem is to derive the values of a and b which best fit the observations, and the corresponding uncertainties σ_a and σ_b.

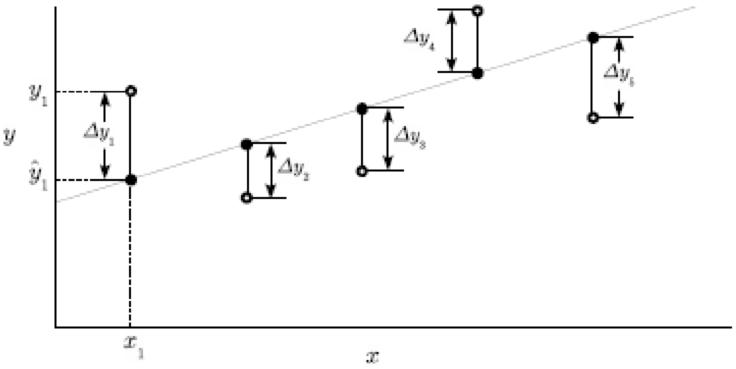

Figure A.6: x-y plot showing residuals. From Kirkup (2002).

To find the best line through x-y data, we need to decide on a measure of the goodness of fit of the line to the data. Fig. A.6 graphically depicts what we want to do. At each point (x_i, y_i), there is a fitted value of y $\hat{y}_i = a + bx_i$. The best fit line should somehow minimize the differences $\Delta y_i = y_i - \hat{y}_i$ between the fitted values and the data points, known as the *residuals* of the fit.

The standard best-fit algorithm used by most statistics packages is the Method of Least Squares, which involves minimizing the quantity

$$\chi^2 = \sum_{i=1}^{N} \frac{(y_i - \hat{y}_i)^2}{\sigma_i^2},$$

where σ_i is the uncertainty in the data point y_i. The expression χ^2 (chi-square) is the weighted sum of squares of the residuals. For simplicity, let's assume for now that $\sigma_i = \sigma$, i.e., all the errors are the same. To find the best fit line, we write χ^2 in terms of the fit parameters a and b and differentiate with respect to a and b:

$$\begin{aligned}
\chi^2 &= \sum_i \left(\frac{y_i - a - bx_i}{\sigma}\right)^2 \\
\frac{\partial \chi^2}{\partial a} &= 0 = -\frac{2}{\sigma^2}\sum (y_i - a - bx_i) \\
\frac{\partial \chi^2}{\partial b} &= 0 = -\frac{2}{\sigma^2}\sum x_i(y_i - a - bx_i)
\end{aligned}$$

Manipulating these equations to solve for a and b, we find

$$\begin{aligned}
a &= \frac{\sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i}{D} \\
b &= \frac{N \sum x_i y_i - \sum x_i \sum y_i}{D}
\end{aligned}$$

where

$$D = N \sum x_i^2 - \left(\sum x_i\right)^2$$

The uncertainties in the fit parameters are tedious to derive; if all of the uncertainties are equal ($\sigma_i = \sigma$), you can show that

$$\begin{aligned} \sigma_a &= \sigma \left(\frac{\sum x_i^2}{D} \right)^{1/2} \\ \sigma_b &= \sigma \left(\frac{n}{D} \right)^{1/2} \end{aligned}$$

These expressions look intimidating, but take comfort in two things: they are actually rather easy to calculate using a spreadsheet; and most statistics packages contain functions that automatically estimate a, b, σ_a, and σ_b for you[7].

The Weighted Linear Least Squares Fit

On a few rare occasions (to be discussed momentarily), you will want to perform a weighted least squares fit to your data that accounts for different σ_i. In this case, the expressions for the fit parameters and their standard errors change slightly:

$$\begin{aligned} a_w &= \frac{\sum \frac{x_i^2}{\sigma_i^2} \sum \frac{y_i}{\sigma_i^2} - \sum \frac{x_i}{\sigma_i^2} \sum \frac{x_i y_i}{\sigma_i^2}}{E} \\ b_w &= \frac{\sum \frac{1}{\sigma_i^2} \sum \frac{x_i y_i}{\sigma_i^2} - \sum \frac{x_i}{\sigma_i^2} \sum \frac{y_i}{\sigma_i^2}}{E} \end{aligned}$$

and

$$\begin{aligned} \sigma_{a,w} &= \left(\frac{\sum x_i^2/\sigma_i^2}{E} \right)^{1/2} \\ \sigma_{b,w} &= \left(\frac{\sum 1/\sigma_i^2}{E} \right)^{1/2} \end{aligned}$$

where the demoninator E is given by

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

The bad news is that few statistics packages automate these functions. The good news is that you will need to use them rarely, if ever.