Averaging multiple measurements
Some slightly random thoughts based on experience in writing data processing software
Phil Evans September 2012

There are problems in getting the “best” estimate of I and $\sigma(I)$ from a set of measured I_i, $\sigma(I_i)$, and I don't know how best to do this. It is clear that any averaging (merging) of measurements is losing information and ideally we should work with unmerged data. However that is not easy!

(1) estimation of the best mean intensity <I>

We can agree that $\langle I \rangle = \sum_i w_i I_i \big/ \sum_i w_i$ but what are the weights?

As Luc says, probably the most common method (and that used in my data processing programs SCALA and AIMLESS) is to use $w_i = 1/\sigma_i^2$, ie 1/variance weights

This has several problems:
1. $\sigma(I)$ estimates from integration programs are based (at least partly) on Poisson (counting) statistics, and this underestimates the total error. Additional errors come from (a) inaccurate "gain" of the detector, leading to a scale error on the counting statistic (b) fluctuation errors of various sorts which lead to an error proportional to the (true) intensity (c) other things (unknown unknowns). Data processing programs generally adjust the raw $\sigma(I)$ estimate upwards with a scale and a fraction of I^2: the value of I used for this correction is the averaged scaled <I>, to reduce bias, but this begs the question of how to do the averaging!
2. the weighting is still biased, with stronger measurements tending to get a lower weight (larger $\sigma(I)$)
3. if the different observations are on very different scales, then the variation of weights is very severe. For example, if we collect a fast pass through the data to extend the dynamic range of eg a CCD detector, with say a factor of 10 in the dose rate, then the ratio of weights will be around 100 which seems much too large a difference.

There is an argument (and I've lost the reference) for using a weight = 1/sqrt(scale factor), ie an observation which is scaled up a lot gets a lower weight. This is arises from a very simplified assumption of measurement statistics, pure Poisson counting statistics and no background. This is a much less severe weighting than 1/variance. I use this weighting when testing for outliers, but not for the final merged intensity.

Probably the "best" (unbiased) weight estimate is somewhere in between

Where does the ShelX weight that Luc quotes come from?
ie $w = \frac{I}{\sigma^2}\ if\ \frac{I}{\sigma} > 3, else\ w = 3/\sigma$

(2) estimation of σ(<I>)

Consistent with the above, in SCALA & AIMLESS I use the "external variance"
$\sigma(\langle I \rangle)^2 = 1 \Big/ \sum_i w_i$
In most cases the multiplicity of measurement is not high enough to give a reliable estimate of the "internal variance", though I can see there is a case for using the internal variance if it is larger, ie if the scatter of observations for a (unique) reflection is larger than expected from its $\sigma(I)$, then maybe $\sigma(I)$ should be increased to take this into account. You certainly can't use just the internal variance, since with a small sample this could be = 0 by chance.

However, I adjust ("correct") the $\sigma(I)$ values so that on average they match the observed scatter, and this is not easy to optimise if the $\sigma(I)$ values are taken as the maximum of two possible values. I suppose I could optimise the corrections, but then use the internal estimate for the final output $\sigma(\langle I \rangle)$ value, either taking the maximum or some weighted mean (but how weighted?)