Weighted averages are frequently used in scientific calculations. Denoting the measurements to be averaged as x_i and their weights w_i (i=1...n), we can straightforwardly calculate the weighted mean as:

$$\bar{x}_{wtd} = \frac{\sum_{i=1}^{n} w_i \, x_i}{\sum_{i=1}^{n} w_i} \tag{1}$$

where the subscript *wtd* indicates a weighted mean. In the trivial case that all the w_i are equal, this formula is equivalent to the familiar unweighed mean.

Weighted averages are used in at least two different cases. In the first case (which we'll call Case I), one wants to give more weight to some points than to others, because they are considered to be more important. For example, one might want to calculate a flow-weighted mean concentration because the total load is the total flow times the flow-weighted mean (not the arithmetic mean). In this case, the weights differ but the uncertainties associated with the individual x_i are assumed to be the same.

Alternatively (Case II), one might have a set of measurements in which all of the x_i are equally important, but some of the x_i have larger uncertainties than others. In this case, it makes sense to give more weight to the x_i that are more certain. The optimal allocation of weights is to make each point's weight *inversely proportional to the square of its uncertainty*. One can show that this so-called "inverse variance weighting" scheme is optimal in the sense that it minimizes the uncertainty in the weighted mean.

Now, what's the right way to calculate the standard error of a weighted mean? It turns out that the answer to this question depends on whether one is dealing with the first case outlined above, or the second. In Case II, we calculate the standard error by the usual approach, but we include the weights in the formulas. We start by calculating the variance of the measurements around the mean. The variance is the mean of the squared deviations around the mean, and when both of those kinds of means are weighted, we get:

$$Var(x)_{wtd} = (s_x^2)_{wtd} = \frac{\sum_{i=1}^{n} w_i \left(x_i - \bar{x}_{wtd}\right)^2}{\sum_{i=1}^{n} w_i} \frac{n}{n-1} = \left(\frac{\sum_{i=1}^{n} w_i x_i^2}{\sum_{i=1}^{n} w_i} - \left(\bar{x}_{wtd}\right)^2 \right) \frac{n}{n-1} \tag{2}$$

where the factor of n/n-1 is intended to account for the number of degrees of freedom, thus giving us (we hope!) an unbiased estimate of the variance of the population from which the x_i's were sampled. We can straightforwardly calculate the weighted standard deviation as the square root of the weighted variance. We can then apply the Central Limit Theorem, and estimate the standard error of the mean as:

$$(s_{\bar{x}})_{wtd} = \frac{(s_x)_{wtd}}{\sqrt{n}} = \sqrt{\frac{Var(x)_{wtd}}{n}} \tag{3}$$

If all the weights are equal, this approach yields the familiar unweighted standard error. If the weights are unequal, but are inversely proportional to the uncertainties in the individual x_i (inverse variance weighting, Case II), this approach yields an unbiased estimate of the uncertainty in the weighted mean.

So far, so good. However, big problems can arise if these formulas are used in Case I, when the uncertainties in the individual x_i are similar but some points are simply more important than others. Statistical software packages generally do not distinguish between Cases I and II as outlined above, and use equations (2) and (3) for all sorts of weighted averages. It turns out that these equations are exactly right in Case II, but are biased -- potentially *very* biased -- in Case I. To see what the problem is, let's imagine that we have 100 measurements x_1...x_{100}, but the weight for one measurement (say, the 27th point, x_{27}) is 10,000 times greater than the weights for any of the others.

Because this one point has roughly 100 times as much weight as all the other points combined, the weighted mean will typically be very close to x_{27}. There's nothing wrong with that; it's simply a reflection of the weighting.

But now, look what will happen when we calculate the weighted variance using equation (2). The weighted variance will be utterly dominated by just one term, the one in which we measure the deviation of x_{27} from the weighted mean. But that deviation will inevitably be very small, since the weighted mean is almost entirely composed of x_{27} itself! The more uneven the weightings are, the smaller the weighted variance and the weighted standard error will be (assuming the x's are the same). But as the weighting becomes more uneven, the weighted mean becomes more and more *unstable*, since it is increasingly dominated by one or a few points with high weights. To capture the statistical uncertainty in the mean, the weighted standard error should become larger as the weightings become more uneven; instead, it becomes smaller. What is going wrong?

What's going awry is that as the weightings become more uneven, and thus the weighted mean and weighted variance become dominated by just a few points, the effective degrees of freedom become smaller (potentially *much* smaller) than n. As an extreme example, consider a situation where all of the points except one have a weight of zero; the effective number of data points is 1, not n, and the effective number of degrees of freedom is *zero*, not n-1!

We can (and should!) correct for this loss of degrees of freedom. The right way to do it is to estimate the effective number of measurements as

$$n_{eff} = \frac{\left(\sum_{i=1}^{n} w_i\right)^2}{\sum_{i=1}^{n}\left(w_i^2\right)} \tag{4}$$

and then use n_{eff} instead of n in the formulas on the previous page. Note that n_{eff} behaves as it should. As the weightings become more and more even, n_{eff} converges to n. If the weightings are dominated by a single point, n_{eff} will converge to 1, and if the weightings are dominated by (say) k points with equal weight, with the rest of the points having trivial weights, n_{eff} will converge to k.

Using n_{eff}, the (now unbiased!) estimate of the weighted variance in Case I is:

$$Var(x)_{wtd} = (s_x^2)_{wtd} = \frac{\sum_{i=1}^{n} w_i\left(x_i - \bar{x}_{wtd}\right)^2}{\sum_{i=1}^{n} w_i}\,\frac{n_{eff}}{n_{eff}-1} = \left(\frac{\sum_{i=1}^{n} w_i x_i^2}{\sum_{i=1}^{n} w_i} - \left(\bar{x}_{wtd}\right)^2\right)\frac{\left(\sum_{i=1}^{n} w_i\right)^2}{\left(\sum_{i=1}^{n} w_i\right)^2 - \sum_{i=1}^{n}\left(w_i^2\right)} \tag{5}$$

The corresponding (now unbiased!) standard error of the weighted mean in Case I is:

$$(s_{\bar{x}})_{wtd} = \frac{(s_x)_{wtd}}{\sqrt{n_{eff}}} = \sqrt{\frac{Var(x)_{wtd}}{n_{eff}}} = \sqrt{\left(\frac{\sum_{i=1}^{n} w_i x_i^2}{\sum_{i=1}^{n} w_i} - \left(\bar{x}_{wtd}\right)^2\right)\frac{\sum_{i=1}^{n}\left(w_i^2\right)}{\left(\sum_{i=1}^{n} w_i\right)^2 - \sum_{i=1}^{n}\left(w_i^2\right)}} \tag{6}$$

Using Monte Carlo methods, one can show that this standard error gives reliable estimates of how much the weighted mean can be expected to deviate from the true mean of the population from which the x_i's were sampled, assuming that they all have similar uncertainties (Case I).

Reference: Bevington, P. R., Data Reduction and Error Analysis for the Physical Sciences, 336 pp., McGraw-Hill, 1969.