What is the optimal estimator of *F*?

Ian J. Tickle, Global Phasing Ltd.

Here we define *F* as the real structure amplitude (**not** the complex structure factor) which corresponds to the real part of the electron density. In the presence of an anomalous signal and in the absence of experimental errors a good approximation for $\hat{F}$ is the simple arithmetic mean $\hat{F}_{mean}$ of $\hat{F}[+]$ and $\hat{F}[-]$:

$$\hat{F}_{mean}=(\hat{F}[+]+\hat{F}[-])/2$$

where $\hat{F}$ signifies 'true value of *F*' throughout.

For example for a 10000 MW protein with a single Sm^{3+} ion at full occupancy the maximum $\hat{F}$ at the L edge and zero θ is ~ 230, whereas assuming *f"* = 17, the maximum difference between $\hat{F}_{mean}$ and $\hat{F}$ (when $\varphi_H = \varphi_P$) is $\sqrt{(230^2 + 17^2)} - 230 = 0.63$, so in this case the maximum fractional error in $\hat{F}$ as a consequence of assuming that $\hat{F}=\hat{F}_{mean}$ is ~ 0.3%, which is of course negligible in the grand scheme of things.

The question we wish to answer here is: in the presence of experimental errors, is the observed F_{mean} still the optimal estimator of $\hat{F}$ and if not then what is? An alternative estimator which has been proposed is the variance-weighted mean F_{wmean} :

$$F_{wmean}=\frac{F[+]/\sigma^2(F[+])+F[-]/\sigma^2(F[-])}{1/\sigma^2(F[+])+1/\sigma^2(F[-])}$$

This has the superficial advantage that F_{wmean} has a lower variance than F_{mean} (in fact the variance-weighted mean is always a minimum variance estimator). The argument supporting this is that if one of $\sigma(F[+])$ or $\sigma(F[-])$ is significantly higher than the other one, using the weighted mean has the effect that the higher one doesn't 'pollute' the estimate of $\hat{F}$ with random errors. However, the price to be paid is that if $\sigma(F[+])$ and $\sigma(F[-])$ do differ, then F_{wmean} is also a biased estimator of $\hat{F}_{mean}$ and therefore also of $\hat{F}$, where 'bias' is defined in the usual way as the deviation of the expectation of the estimator from its true value. In contrast, as shown above, F_{mean} is effectively always an unbiased (though possibly high variance) estimator of $\hat{F}$.

A better (if not optimal) measure of the true error is one that strikes the right balance between a pure measure of precision (*i.e.* the variance, σ^2) and a pure measure of accuracy (*i.e.* the bias, δ), where:

$$\sigma^2(F)=\langle(F-\langle F\rangle)^2\rangle$$

$$\delta(F)=\langle F\rangle-\hat{F}$$

One such measure is the expected mean-squared error (*MSE*), defined as the expectation of the squared error $\langle\epsilon^2(F)\rangle$, where the error $\epsilon(F)$ is the deviation of the estimator from its true value:

$$\epsilon(F)=F-\hat{F}$$

$$\langle\epsilon^2(F)\rangle=\langle(F-\hat{F})^2\rangle=\sigma^2(F)+\delta^2(F)$$

The point here is that it would seem much more sensible to measure errors relative to the true value, rather than to the possibly bogus population or (even worse) sample mean. This of course would seem to require that the true value is known *a priori* – an apparent impossibility (more on this further down)! The optimal estimator is then the one that minimises the expected *MSE*, thereby achieving the optimal balance of variance and bias. In general neither F_{mean} nor F_{wmean} is a minimum *MSE* estimator of $\hat{F}$, so neither is optimal in the sense of minimal expected *MSE*.

To simplify things, we assume that the optimal estimator of $\hat{F}$, whatever it is, is some weighted linear combination of $F[+]$ and $F[-]$:

$$F_{opt} = x\,F[+] + (1-x)\,F[-] \quad \text{with } 0 \le x \le 1.$$

i.e. F_{opt} must certainly lie somewhere between $F[+]$ and $F[-]$. This general definition of F_{opt} obviously also covers both F_{mean} and F_{wmean}, by appropriate selection of the value of x.

F_{opt} has variance:

$$\sigma^2(F_{opt}) = x^2\sigma^2(F[+]) + (1-x)^2\sigma^2(F[-])$$

and bias:

$$\delta(F_{opt}) = \langle F_{opt}\rangle - \hat{F} = \langle x\,F[+] + (1-x)\,F[-]\rangle - (\hat{F}[+] + \hat{F}[-])/2$$

Therefore the expected *MSE* is:

$$\langle \epsilon^2(F_{opt})\rangle = x^2\sigma^2(F[+]) + (1-x)^2\sigma^2(F[-]) + (\langle x\,F[+] + (1-x)\,F[-]\rangle - (\hat{F}[+] + \hat{F}[-])/2)^2$$

Straightforward differentiation of the expected *MSE* with respect to x and setting the derivative to zero gives the solution for the optimal value of x that we are seeking:

$$x = \frac{(\hat{F}[+] - \hat{F}[-])^2/2 + \sigma^2(F[-])}{(\hat{F}[+] - \hat{F}[-])^2 + \sigma^2(F[+]) + \sigma^2(F[-])}$$

from which an expression for F_{opt} is obtained using its definition above.

Unfortunately, as it stands this expression for F_{opt} is unusable because it contains the true values $\hat{F}[+]$ and $\hat{F}[-]$ which obviously are unknown. We therefore substitute these by their observed values $F[+]$ and $F[-]$. I think this is the only questionable step in this derivation (since by the same argument we could substitute $F[+]$ and $F[-]$ into the expression for $\hat{F}_{mean}$ to get F_{mean} but then that contradicts the conclusion that $\mathrm{F_{opt}}$ is a better estimate than F_{mean}). However this step seems to be justified by the results (see Table below).

Results

Taking the Sm^{3+} example again, since the differences between the alternative estimates of $\hat{F}$ will be greatest when the anomalous difference is a maximum, I used the values of $\hat{F}[+]$ and $\hat{F}[-]$ at the maximum of the anomalous difference as a function of the phase difference ($\varphi_P - \varphi_H$): $\hat{F}[+]=578$ and $\hat{F}[-]=1369$, so that $\hat{F}=\hat{F}_{mean}=973$ (arbitrary scale). I then compared the calculated values of the sample standard deviation $s(F)$, the sample bias $d(F)$ and the sample $\sqrt{MSE}(F)$ for the three alternative estimates: F_{mean}, F_{wmean} and F_{opt} by using a random-number generator to give a normal distribution of errors for $F[+]$ and $F[-]$ with various specified standard deviations (using a sample size of 10^6 in each case).

$\sigma(F[+])$	$\sigma(F[-])$	$s(F_{mean})$	$s(F_{wmean})$	$s(F_{opt})$	$d(F_{mean})$	$d(F_{wmean})$	$d(F_{opt})$	$\sqrt{}MSE(F_{mean})$	$\sqrt{}MSE(F_{wmean})$	$\sqrt{}MSE(F_{opt})$
1	1	1	1	1	0	0	0	1	1	1
10	10	7	7	7	0	0	0	7	7	7
100	100	71	71	71	0	0	0	71	71	71
1	10	5	1	5	0	-388	0	5	388	5
10	1	5	1	5	0	388	0	5	388	5
10	100	50	10	50	0	-388	-6	50	388	50
100	10	50	10	49	0	388	6	50	388	50
100	1000	466	99	200	20	-387	-232	467	400	306
1000	100	386	99	171	88	389	274	396	402	323

where:

$s(F)=\sqrt{\overline{(F-\overline{F})^2}}$ (sample standard deviation)

$d(F)=\overline{F}-\hat{F}$ (sample bias)

$\sqrt{MSE(F)}=\sqrt{\overline{(F-\hat{F})^2}}=\sqrt{(s^2(F)+d^2(F))}$ (sample root-*MSE*)

Conclusions

1. As expected, when $\sigma(F[+])=\sigma(F[-])$ all three estimates of *F* are equal (*x* = 0.5), with equal sample standard deviation and root-*MSE*, and zero sample bias.

2. Otherwise, when $\sigma(F[+])$ and $\sigma(F[-])$ differ significantly, although the sample standard deviation of the weighted mean is always the lowest, this is more than offset by a much higher sample bias in all cases. This results in a rather poor estimate (as measured by large *MSE*) by F_{wmean} for large values of $F[+]/\sigma(F[+])$ or $F[-]/\sigma(F[-])$. The use of F_{wmean} in this case would result in significant error; either F_{mean} or F_{opt} provides a much more accurate estimate of $\hat{F}_{mean}$ than does F_{wmean}.

3. For small values of $F[+]/\sigma(F[+])$ or $F[-]/\sigma(F[-])$ F_{wmean} is indeed a better estimate than F_{mean}, but only marginally, and in any case F_{opt} is a better estimate than either. Even without the use of $\mathrm{F_{opt}}$, F_{wmean} does not appear to offer any significant advantage over F_{mean}, and indeed as shown above comes with the possibility of a severe disadvantage in the form of a large bias in the case of accurately measured $F[+]$ and $F[-]$.

4. The sample *MSE* of F_{opt} is never higher than the lower of the sample *MSE* values for F_{mean} and F_{wmean}, and therefore appears to be a more optimal estimator over the whole range of values of $\sigma(F[+])$ and $\sigma(F[-])$.