Acta Cryst. (1956). 9, 652

Estimates of the Standard Deviations of the Observed Structure Factors and of the Electron Density from Intensity Data

By James A. Ibers

Shell Development Company, Emeryville, California, U.S.A.

(*Received* 9 *January* 1956 *and in revised form* 7 *February* 1956)

It is shown that an initial estimate of the standard deviation of the electron density may be obtained by a straightforward and relatively simple calculation based on estimates of the standard deviations of the observed structure factors. These latter estimates are readily obtained from the errors, observed and expected, in the intensity data.

Introduction

In the initial stages of a structure investigation an estimate of $\sigma(\varrho_o)$, the standard deviation of the electron density at the general position x, y, z, would be useful for many reasons. Such an estimate, for example, would provide the lower limit which the investigator might seek in his refinement of the structure. Thus, what might be termed 'over-refinement' of the structure, i.e. refinement to a point where the resultant $\sigma(\varrho_c)$ based perhaps on (F_o-F_c) was significantly below that allowed on the basis of errors in the intensity data, would not occur. Such over-refinement may be possible, particularly if experimental form factors and multiple-term anisotropic temperature factors are used for each atom in the asymmetric unit. Moreover, an initial estimate $\sigma(\varrho_o)$ might prevent the investigator from attempting the impossible with his data: for example, the principal objective of an investigation is the location of hydrogen atoms, and $\sigma(\varrho_o)$ is calculated to be 2 e.Å^{-3}; in this case, the objective very probably cannot be realized without more accurate intensity estimates.

It is shown below that $\sigma(\varrho_o)$ may be derived readily from estimates of the standard deviations of the observed structure factors, $\sigma(F_o)$.

Grouping of the data

In a typical three-dimensional X-ray study, intensity data are gathered around two crystallographic axes, say a and b, and estimated either visually or photometrically, or else counted electronically. The reflections within the limiting sphere may be placed in four groups: (I) reflections observed more than once, i.e. around both a and b; (II) reflections observed only once, i.e. around a but not b, or around b but not a; (III) reflections too weak to be observed around either axis; (IV) reflections absent because of the requirements of the space group. In general, the majority of the reflections will belong to Group I.

Estimation of $\sigma(F_o)$

Previous estimates of $\sigma(F_o)$ generally have been quite rough: approximations such as $\sigma(F_o) = k|F_o|^a$ (with a usually 0 or 1) have sufficed, however, for the order-of-magnitude calculations to which they have been applied (see, for example, Lipson & Cochran, 1953). That such approximations are indeed rough is indicated by the familiar observation that for photographic data the F_o's are relatively much less reliable at low and at very high F_o values than in the intermediate range.

In the estimation of $\bar{F}_o$ and of $\sigma(F_o)$ we face the problem of very small sample sizes. We shall assume that our samples are taken from a population which is normally distributed (i.e. errors are assumed to be random); we shall discuss this assumption below. The sample mean is the most efficient, unbiased estimate of the population mean regardless of the sample size. By the most efficient estimate is meant the one which gives values more closely concentrated around the true value than values derived from any other statistic. For very small sample sizes other statistics such as the mid-range and the median are nearly as efficient as the sample mean, but they offer no particular calculational advantages. Therefore, we use the sample mean as the estimate of the population mean. The situation with regard to unbiased estimates of the population standard deviation is somewhat different. The sample standard deviation may be corrected for bias, and it is then the most efficient, unbiased estimate of the population standard deviation. For very small sample sizes, however, another statistic, the range estimate, is very nearly as efficient as the sample standard deviation, and it is much more readily computed. Accordingly, we adopt the range estimate as the unbiased estimate of the population standard deviation $\sigma(y)$; we obtain

$$\sigma(y) = c|y_{\max.}-y_{\min.}| , \qquad (1)$$

where $y_{\max.}$ is the maximum and $y_{\min.}$ is the minimum of the observed sample values. Tippett (1925) has discussed the range estimate at length, and has derived

the coefficient c as a function of n, the sample size. Of importance to us are the following values: $n = 2$, $c = 0{\cdot}89$; $n = 3$, $c = 0{\cdot}59$; $n = 4$, $c = 0{\cdot}49$.* For $n = 2$, the range estimate is equivalent to the sample standard deviation corrected for bias. The application of equation (1) to reflections of Group I is obvious: For the usual case of two observations F_a and F_b we obtain

$$\sigma(F_o) = 0{\cdot}89|F_a - F_b| \,, \tag{2}$$

and

$$\bar{F}_o = \tfrac{1}{2}(F_a + F_b) \,. \tag{3}$$

Such estimates assume that both F_a and F_b are equally reliable. The application of an elaborate weighting system to calculations based on such small sample sizes does not seem justified. If, for a particular reflection, F_a is believed to be much more reliable than F_b, then the best policy is probably to reject F_b and consider that the particular reflection belongs to Group II.

For an estimate of $\sigma(F_o)$ for a particular F_o in Group II there seems little choice but to use a plot of average $\sigma(F_o)$ versus F_o, based on the data from Group I.

For Group III reflections one may use Wilson's (1949) distributions of structure factors and a method similar to that employed by Hamilton (1955) to derive:

$$\left.\begin{array}{ll} \text{Centric:} & \bar{F}_{\text{unobs.}} \sim \tfrac{1}{2}F_{\text{min.}}(\theta), \\ & \sigma(F_{\text{unobs.}}) \sim F_{\text{min.}}(\theta)/12^{\frac{1}{2}}; \\ \text{Acentric:} & \bar{F}_{\text{unobs.}} \sim \tfrac{2}{3}F_{\text{min.}}(\theta), \\ & \sigma(F_{\text{unobs.}}) \sim F_{\text{min.}}(\theta)/18^{\frac{1}{2}}; \end{array}\right\} F_{\text{min.}} \ll \bar{F}^2{}^{\frac{1}{2}}\dagger \,. \tag{4}$$

Group IV reflections have $\sigma(F_o) \equiv 0$.

The above estimates of $\sigma(F_o)$ are essentially the best that can be made with the information available to us; yet, they are unreliable. For example, for Group I reflections and $n = 2$ it can be shown that 50% of the time our estimate will lie between 0·40 and 1·44 times the true population standard deviation (Pearson, 1942). Fortunately, we are usually concerned with the combination of these estimates, for example to give an estimate of $\sigma(\varrho_o)$. Under normal circumstances we can expect that the relative error in $\sigma(\varrho_o)$ will be much less than the relative error in any individual estimate of $\sigma(F_o)$.

Estimation and use of $\sigma(\varrho_o)$

It has been shown (e.g. Cruickshank, 1949, 1950) that if the errors in F are assumed to be random, and if the data are numerous, we may write the approximate expression

$$\sigma(\varrho) = \frac{m}{V}\left[\sum_{hkl}\{\sigma(F)\}^2\right]^{\frac{1}{2}}, \tag{5}$$

where m is 1 for the centric space groups and 2 for the acentric ones. Here, $\sigma(\varrho)$ refers to *any* general position x, y, z; i.e. $\sigma(\varrho)$ is essentially independent of the general position x, y, z. Equation (5), when used, is applied at the conclusion of a structure investigation, use being made of $|F_o - F_c|$ as an estimate of $\sigma(F)$ (Cruickshank, 1949). There is no reason, however, why (5) cannot be applied at the beginning of the investigation to yield the initial estimate $\sigma(\varrho_o)$, an estimate independent of the refinement, dependent only upon the intensity data. For the calculation of such an estimate we use in (5) $\sigma(F_o)$ as derived above.

Naturally, $\sigma(\varrho_o)$ cannot be calculated in e.Å^{-3} until the intensity data are on an absolute scale. A scale can generally be found from the intensity data (e.g. Wilson's (1942) method), but in those cases where this is not possible, the calculation of $\sigma(\varrho_o)$ can still be made with profit after a reliable trial structure has been found.

If a decision is to be reached about the possibility of locating a particular atom i (assumed to be in the general position x, y, z), then $\sigma(\varrho_o)$ should be compared with the expected height $\varrho(0)$ at the center of the peak corresponding to atom i on a Fourier map. That height may be calculated from the usual expression

$$\varrho(0) = \frac{1}{2\pi^2}\int_0^{s_0} f^*(s)s^2 ds \,, \tag{6}$$

where $f^*(s)$ is the product of temperature and form factors, s is $(4\pi/\lambda)\sin\theta$, and s_0 is the limiting value of s. An approximate temperature factor, as guessed from those in similar, known structures, or as obtained from Wilson's (1942) method, should suffice for the calculation of $\varrho(0)$.

As was pointed out in the introduction, comparison of succeeding $\sigma(\varrho_c)$'s with $\sigma(\varrho_o)$ should provide information about the state of the refinement of the structure. It is important that such a comparison be made between $\sigma(\varrho_c)$ and $\sigma(\varrho_o)$ only when these are calculated with the same weighting scheme. For example, if $\sigma(\varrho_c)$ is based on a difference Fourier in which $(F_o - F_c)$ terms are omitted when F_o is unobserved (belongs to Group III), then the Group III reflections should be omitted from the calculation of $\sigma(\varrho_o)$. However, inclusion of such terms in the calculation of the difference Fourier, and hence in the calculation of $\sigma(\varrho_o)$, in accordance with equations (4), is to be recommended over their omission.

Discussion of the assumption of random errors

It may well be argued that, contrary to what has been assumed above, the errors in F_o are not random. This

* An abbreviated table of c versus n is given by Dixon & Massey (1951).

† Equations (4) will hold in general, i.e. when Wilson's distributions apply; in other specific cases the distributions which differ from those of Wilson and which arise because of the failure of the conditions assumed by him are known and could be used to derive equations analogous to (4). See Rogers, Stanley & Wilson (1955) for a summary of references pertinent to such distributions.

is true. The two or more independent estimates of F_o for reflections in Group I are not truly independent in the case of photographic data, since they depend upon the scale factor of each layer, and hence upon all of the intensity estimates. With numerous data, however, the estimates remain essentially independent.

Moreover, extinction will not produce random errors. The effects of extinction may be removed, at least in part, after the structure is known (Vand, 1955), and at that time the initial estimate $\sigma(\varrho_o)$ may easily be revised, if necessary.

Likewise, absorption will not produce random errors: $\sigma(F_o)$ will be sensitive to absorption effects. If many of the data are affected appreciably by absorption, the resultant estimated value of $\sigma(\varrho_o)$ will be significantly high*; this would indicate that no reasonable refinement of the structure is possible, and would suggest that a better method, experimental or theoretical, for the correction or elimination of absorption errors is needed.

Note that in the usual application of (5) the errors in $|F_o - F_c| = \sigma(F)$ depart from randomness not only because of the effects discussed above, but also because of the dependence of F_c on the proposed structure.

Other initial estimates

In some cases it may be useful to calculate $\sigma\{\varrho_o(x,y,z)\}$ for special x, y, z or $\sigma\{\varrho_o(x,y)\}$. The same estimates of $\sigma\{F_o(hkl)\}$ given above (or analogous estimates of $\sigma\{F_o(hk)\}$) may be used in equations analogous to (5). (For such equations see Cruickshank & Rollett, 1953.) Estimates of $\sigma\{F_o(hk)\}$, since they are usually based on two readings of the same intensity, are generally not as reliable as those of $\sigma\{F_o(hkl)\}$.

These same estimates of $\sigma(F_o)$ may be used in the calculation of such quantities as $\sigma(\partial\varrho/\partial x)$ which are needed for the evaluation of $\sigma(x_n) = \sigma(\partial\varrho/\partial x)/(\partial^2\varrho/\partial x_n^2)$, the standard deviation in the x coordinate of the nth atom due to errors in the intensity data (Cruickshank, 1949). However, at the beginning of the investigation $\sigma(x_n)$ is not likely to provide as much useful information as is $\sigma(\varrho_o)$. The reliability of such a calculation, moreover, is severely limited by lack of knowledge of $\partial^2\varrho/\partial x_n^2$, the central curvature at the center of the nth atom.

The use of $\sigma(F_o)$ in the least-squares procedure

The subject of the least-squares procedure is not completely unrelated to the above discussions, for in such a procedure where it is desired to minimize the function $\sum_{hkl} w(hkl)[|F_o(hkl)| - |F_c(hkl)|]^2$ the weights $w(hkl)$ are properly assigned proportional to $1/\sigma^2(F_o)$. The methods of estimating $\sigma(F_o)$ given above should provide weights for the least-squares procedure which are less arbitrary and more reliable than those used in the past. The contention that more reliable weights are desirable is supported by the recent results of Abrahams (1955) which point to the rather large effects that various weighting schemes have on the bond lengths derived by the least-squares procedure.

I wish to thank Prof. V. Schomaker of the California Institute of Technology and Mr D. P. Stevenson and Mr A. E. Smith of this Laboratory for helpful discussions.

References

ABRAHAMS, S. C. (1955). *Acta Cryst.* **8**, 661.
CRUICKSHANK, D. W. J. (1949). *Acta Cryst.* **2**, 65.
CRUICKSHANK, D. W. J. (1950). *Acta Cryst.* **3**, 72.
CRUICKSHANK, D. W. J. & ROLLETT, J. S. (1953). *Acta Cryst.* **6**, 705.
DIXON, W. J. & MASSEY, F. J., JR. (1951). *Introduction to Statistical Analysis*, chap. 16. New York: McGraw-Hill.
HAMILTON, W. C. (1955). *Acta Cryst.* **8**, 185.
LIPSON, H. & COCHRAN, W. (1953). *The Determination of Crystal Structures*, p. 288. London: Bell.
PEARSON, E. S. (1942). *Biometrika*, **32**, 301.
ROGERS, D., STANLEY, E. & WILSON, A. J. C. (1955). *Acta Cryst.* **8**, 383.
TIPPETT, L. H. C. (1925). *Biometrika*, **17**, 364.
VAND, V. (1955). *J. Appl. Phys.* **26**, 1191.
WILSON, A. J. C. (1942). *Nature, Lond.* **150**, 152.
WILSON, A. J. C. (1949). *Acta Cryst.* **2**, 318.

* This is not true in the rare case of absorption errors which are mainly functions of s; such errors occur if the data are obtained from a spherical or cylindrical crystal but are not corrected for absorption effects.