References

FLACK, H. D. (1970). *Philos. Trans. R. Soc. London Ser. A*, **266**, 575–582.
GLAZER, A. M. (1970). *Philos. Trans. R. Soc. London Ser. A*, **266**, 635–639.
GLAZER, A. M., STADNICKA, K. & SINGH, S. (1981). In preparation.
HOSEMANN, R. & BAGCHI, S. N. (1962). *Direct Analysis of Diffraction by Matter*. Amsterdam: North Holland.
STADNICKA, K. & GLAZER, A. M. (1980). *Acta Cryst.* B**36**, 2977–2985.
STOKES, A. R. (1948). *Proc. Phys. Soc.* **61**, 382–391.
WILSON, A. J. C. (1962). *X-ray Optics*. London: Methuen.

Acta Cryst. (1981). A**37**, 808–810

Can Intensity Statistics Accommodate Stereochemistry?

BY A. J. C. WILSON

*Department of Physics, University of Birmingham, Birmingham B*15 2*TT, England*

(*Received* 6 *March* 1981; *accepted* 14 *April* 1981)

Abstract

As pointed out by French & Wilson [*Acta Cryst.* (1978), A**34**, 517–525], central-limit theorems exist for the sums of non-independent as well as of independent variables [Bernstein (1922). *Math. Ann.* **85**, 237–241; (1927). *Math. Ann.* **97**, 1–59]. The finite size and stereochemical properties of atoms make the terms summed in the calculation of structure factors non-independent, but, if a central-limit theorem is applicable, then French & Wilson's postulate that the distribution parameter is $\langle I \rangle$ and not Σ has a theoretical base as well as empirical justification. The curve of $\langle I \rangle$ *versus* $(2 \sin \theta)/\lambda$ is correlated with the Patterson function, and the question of the existence of series expansions of the Gram–Charlier or Edgeworth type for sums of non-independent variables is raised.

Central-limit theorems

The expression for the structure factor in terms of the atomic positions and the indices of reflexion,

$$F_{hkl} = \sum_{j=1}^{n} f_j \exp\{2\pi i(hx_j + ky_j + lz_j)\}, \tag{1}$$

is of the form

$$S_n = u_1 + u_2 + \ldots + u_n \tag{2}$$

considered in statistics in connexion with central-limit theorems. The usual theorem (see, for example, Cramér, 1945, pp. 213–220) depends on the assumption that the u's are independent variables, and Wilson (1949) used it to deduce the probability distribution of the structure factors and of the intensities of reflexion for crystals having the space groups $P1$ and $P\bar{1}$, though he realized that the finite size of atoms would prevent complete independence of the successive terms of (1). The assumption of complete independence implies that the mean intensity of reflexion is

$$\Sigma = \sum_{i=1}^{n} |f_i|^2 \tag{3}$$

(Wilson, 1942). The expressions derived by Wilson have been found to apply with useful accuracy to many structures, but for large-molecule structures the average intensity does not decrease monotonely with $(2 \sin \theta)/\lambda$, as predicted by (3), but shows more or less marked oscillations. French & Wilson (1978), drawing attention to generalized central-limit theorems applicable when the u's of (2) are not independent, postulated that the functional forms of the Wilson (1949) distributions would remain valid, but that the distribution parameter [S in the notation of Wilson (1950)] would be $\langle I \rangle$, the actual local value of the mean intensity, averaged over values of hkl giving approximately the same value of $(2 \sin \theta)/\lambda$, instead of the sum given in (3). [In reading their paper it must be noted that they use the symbol Σ both for this sum and for the mean intensity $\langle I \rangle$.] Rogers (1965, 1980), Ladd (1978) and others have tacitly made the same assumption, without explicit reference to central-limit theory.

There are two main generalizations of the central-limit theorem for non-independent variables. The first applies when the u's are 'almost independent' (*presque indépendantes*; Bernstein, 1922), where 'almost independent' is given a precise mathematical definition whose physical meaning is not easy to grasp. The

0567-7394/81/060808-03$01.00

second applies when each u_j is related only to a finite number, $f(n)$, of its neighbours (Bernstein, 1927), when the u's are said to be $f(n)$ dependent. The second is perhaps the case that has been considered most frequently in later work, and particularly for $f(n)$ equal to a constant, m, when the u's are said to be m dependent. This second case seems plausible for crystallographic applications, since the positions of atoms close together in the structure are closely correlated by interatomic forces, whereas those far apart will show little correlation if there is any flexibility in the asymmetric unit when unconstrained. Long-range stereochemical effects, as in pseudo-graphitic aromatic hydrocarbons, would presumably produce long-range correlation. Harker's (1953) idea of 'globs' seems equivalent to $f(n)$ dependence.

French & Wilson (1978) justified their postulate for the value of the distribution parameter on empirical grounds (their Fig. 3), but it is readily seen to be theoretically correct if two conditions are satisfied:

(i) that one of the generalized central-limit theorems is applicable, giving asymptotic normality of the distribution of F in $P\bar{1}$ under certain conditions, or of its real and imaginary parts separately in $P1$; and

(ii) that the expected value of u_j,

$$\langle u_j \rangle \equiv f_j \langle \exp\{2\pi i(hx_j + ky_j + lz_j)\}\rangle, \tag{4}$$

is equal to zero.

The distribution parameter in a normal distribution is the variance of the variable, and in the present case, for $P\bar{1}$,

$$\operatorname{var}(F) = \langle F^2 \rangle - \langle F \rangle^2. \tag{5}$$

The first term is the mean value of the intensity of reflexion, and the second term is zero if the second condition is fulfilled. Wilson (1949) considered this condition at some length, and concluded that it would be fulfilled for all but the lowest-order reflexions; there is no obvious reason why it should fail when the u's are non-independent. The French & Wilson postulate is thus justified theoretically as well as experimentally for $P\bar{1}$. A somewhat longer calculation verifies it for $P1$ also. It may be remarked that practically all experimental investigations based on the intensity distribution functions, from Howells, Phillips & Rogers (1950) onwards, have used the empirically determined value of $\langle I \rangle$, rather than the theoretical value of Σ given by (3), as the distribution parameter, and are thus consistent with the generalized central-limit theorem and the French & Wilson postulate. Obvious allowances have to be made if a significant part of the scattering is due to atoms in parameter-free positions, such as a heavy atom at the origin, or in positions that are effectively parameter-free, such as atoms in the Wyckoff positions (a), (b), (c), (d) of the space group $P121$ for the $h0l$ reflexions. Main (1975) and others have devised methods for making use of known molecular fragments in improving the determination of absolute scale (Wilson, 1942).

Patterson interpretation of $\langle I \rangle$

It is instructive to consider the difference between $\langle I \rangle$ and Σ from the point of view of the Patterson (1935) function. The intensity of reflexion is

$$I = \sum_{j,k} f_j f_k^* \exp\{2\pi i[h(x_j - x_k) + k(y_j - y_k) + l(z_j - z_k)]\} \tag{6}$$

$$= \Sigma + \sum_{j \neq k} f_j f_k^* \exp\{2\pi i\, \mathbf{s}.(\mathbf{r}_j - \mathbf{r}_k)\}, \tag{7}$$

where

$$\mathbf{r} = x\mathbf{a} + y\mathbf{b} + z\mathbf{c} \tag{8}$$

and

$$\mathbf{s} = h\mathbf{a}^* + k\mathbf{b}^* + l\mathbf{c}^*. \tag{9}$$

The intensity can thus be regarded as the structure factor of the Patterson (1935) representation of the actual structure, in which representation there are pseudo-atoms of atomic scattering factor $f_j f_k^*$ at positions given by the interatomic vectors

$$\mathbf{r}_{jk} \equiv \mathbf{r}_j - \mathbf{r}_k \tag{10}$$

(*cf.* Wilson, 1970, pp. 177–179; 1978, § 3.1). Alternatively, (7) may be regarded as expressing the intensity as the sum of the ideal average intensity Σ and terms which have the expected value zero for sufficiently large $|\mathbf{s}|$, but which may be appreciable for $\mathbf{s}$ in the observable range (*cf.* Rogers, 1965). Dispersion, implicit in the use of the complex conjugate of f_k in the expressions above, complicates the intensity distributions even when the contributions of separate atoms are assumed to be statistically independent (Wilson, 1980), and will be neglected. Its effect is probably small in most X-ray experiments, but might be more important for electron and neutron diffraction.

The fundamental correlation between atomic positions results from the finite radii of atoms, so that there is a minimum value of $|\mathbf{r}_{jk}|$. Interatomic distances are not only finite, but reasonably constant within a class of substances. For example, in aliphatic organic compounds the inter-carbon distance does not vary greatly from 1·54 Å, and in aromatic compounds it does not vary greatly from 1·39 Å. The Patterson representation will thus have a pseudo-atom of atomic scattering factor Σ at the origin, surrounded by an approximately spherical inaccessible volume of radius about $|\mathbf{r}_{jk}|_{\min}$. For crystals of the classes just mentioned there will be a pile-up of about $2n$ pseudo-atoms at about this radius. Beyond that, in flexible structures, there may be less-definite pile-ups corresponding to next-nearest neighbours, and in non-flexible structures definite pile-ups may exist for a long sequence of values of $|\mathbf{r}|$. In proteins there will be pile-ups corresponding to the repeat distances along the backbone of the helix [or 'warped zipper' in DNA? (Stokes, 1980)]. One sees readily that each favoured value of $r_{jk} \equiv |\mathbf{r}_{jk}|$ will lead to a term proportional to the spherically averaged value

$$\langle u_{jk}\rangle = \frac{f_i f_k}{4\pi}\int\limits_0^{2\pi}\int\limits_0^{\pi} \exp\{2\pi i\, s r_{jk}\cos\theta\}\sin\theta\, d\theta d\varphi \qquad (11)$$

$$= f_j f_k \frac{\sin 2\pi s r_{jk}}{2\pi s r_{jk}} \qquad (12)$$

summed over all values of j and k giving (about) the same value of r_{jk}. This is, of course, the familiar Debye expression, and produces a hump in the $\langle I\rangle$ *versus* s curve at about $s = 5/4r$. Nearest-neighbour distances with r about 1·5 Å thus correspond to values of s far off scale to the right of French & Wilson's (1978) Fig. 2, and the hump observed by them for a phosphorylase corresponds to a much larger frequent distance.

The excluded volume of radius of about one atomic diameter around the origin of the Patterson representation is closely analogous to the inaccessible volumes caused by certain symmetry elements (Wilson, 1964; Nigam, 1972; Nigam & Wilson, 1980). It corresponds, in fact, to the inaccessible volume round a centre of symmetry, and would modulate the average intensity, as a function of s, by the addition of a spherical Bessel function if the other interatomic vectors were randomly distributed over the accessible regions of Patterson space. Non-random distribution, as is observed in practice for proteins and other large molecules, will produce a further modulation, and it does not seem practicable to make any general prediction at this stage.

Is there a series expansion for a finite sum?

When the sum of a number of independent random variables tends to an ideal distribution as the number of variables increases, then under fairly general conditions the distribution function for a finite number can be expressed as the sum of the ideal distribution and some correction terms, each correction term being the product of three factors:

(i) a function of the moments of the distribution;
(ii) one of a set of orthogonal polynomials; and
(iii) the ideal distribution.

The derivation is given in many statistical texts, for example by Cramér (1945, pp. 221–231). Whether the series is genuinely convergent, or useful as an asymptotic approximation, or of no practical use, depends on the properties of the distribution. Applications of such series to intensity statistics up to about 1975 have been reviewed by Srinivasan & Parthasarathy (1976), and more recent applications have been made by Shmueli (1979) and Shmueli & Wilson (1981). Presumably analogous expansions should exist for the sum of non-independent random variables, but so far I have not found any useful references. One might guess that factors (ii) and (iii) above would not be altered, but that in (i) the functions of the moments would undergo changes analogous to the substitution of $\langle I\rangle$ for Σ as the distribution parameter.

References

BERNSTEIN, S. (1922). *Math. Ann.* **85**, 237–241.
BERNSTEIN, S. (1927). *Math. Ann.* **97**, 1–59.
CRAMÉR, H. (1945). *Mathematical Methods of Statistics.* Uppsala: Almqvist and Wiksells.
FRENCH, S. & WILSON, K. (1978). *Acta Cryst.* A**34**, 517–525.
HARKER, D. (1953). *Acta Cryst.* **6**, 731–736.
HOWELLS, E. R., PHILLIPS, D. C. & ROGERS, D. (1950). *Acta Cryst.* **3**, 210–214.
LADD, M. F. C. (1978). *Z. Kristallogr.* **147**, 279–296.
MAIN, P. (1975). *International Summer School on Crystallographic Computing*, pp. 48/1–48/17. Prague: Czechoslovak Academy of Sciences.
NIGAM, G. D. (1972). *Indian J. Pure Appl. Phys.* **10**, 655–656.
NIGAM, G. D. & WILSON, A. J. C. (1980). *Acta Cryst.* A**36**, 832–833.
PATTERSON, A. L. (1935). *Z. Kristallogr.* **90**, 517–542.
ROGERS, D. (1965). In *Computing Methods in Crystallography*, edited by J. S. ROLLETT, pp. 140–148. Oxford: Pergamon Press.
ROGERS, D. (1980). In *Theory and Practice of Direct Methods in Crystallography*, edited by M. F. C. LADD & R. A. PALMER, pp. 82–90. New York: Plenum Press.
SHMUELI, U. (1979). *Acta Cryst.* A**35**, 282–286.
SHMUELI, U. & WILSON, A. J. C. (1981). *Acta Cryst.* A**37**, 342–353.
SRINIVASAN, R. & PARTHASARATHY, S. (1976). *Some Statistical Applications in X-ray Crystallography.* Oxford: Pergamon Press.
STOKES, T. D. (1980). Preprint of paper provisionally accepted for publication in *Social Studies of Science.*
WILSON, A. J. C. (1942). *Nature (London)*, **150**, 151, 152.
WILSON, A. J. C. (1949). *Acta Cryst.* **2**, 318–321.
WILSON, A. J. C. (1950). *Acta Cryst.* **3**, 258–261.
WILSON, A. J. C. (1964). *Acta Cryst.* **17**, 1591–1592.
WILSON, A. J. C. (1970). *Elements of X-ray Crystallography.* Reading, Mass.: Addison-Wesley.
WILSON, A. J. C. (1978). *Acta Cryst.* A**34**, 986–994.
WILSON, A. J. C. (1980). *Acta Cryst.* A**36**, 945–946.