RESEARCH PAPERS

Acta Cryst. (1998). A**54**, 957–970

The Probability Distribution of Structure Factors with Non-Integral Indices. I. The *P*1 Case

Carmelo Giacovazzo[a]* and Dritan Siliqi[a,b]

[a]*IRMEC c/o Dipartimento Geomineralogico, Università di Bari, Campus Universitario, Via Orabona 4, 70125 Bari, Italy, and* [b]*Laboratory of X-ray Diffraction, Department of Inorganic Chemistry, Faculty of Natural Sciences, Tirana, Albania. E-mail: c.giacovazzo@area.ba.cnr.it*

(*Received 27 October 1997; accepted 4 March 1998*)

Abstract

The probability distribution of structure factors with non-integral indices is derived. The distributions are first studied in the one-dimensional case, to understand their main features, then the three-dimensional case is treated. Only the *P*1 group is taken into consideration. For integral values of the indices, the distributions coincide with those provided by Wilson statistics but may strongly differ from them when the indices are (or are close to) half-integrals and are sufficiently small. In these cases, the moduli and phases of the reflections may be accurately estimated in the absence of any structural information. Conditional distributions are also derived which are able to estimate moduli or phases by exploiting the prior information on the specific crystal structure.

1. Symbols and notation

N: number of atoms in the unit cell
f_j: scattering factor of the jth atom (thermal factor included)
$\mathbf{h}$: three-dimensional index with integral components (h, k, l)
$\mathbf{p}$: three-dimensional index with rational components (p_1, p_2, p_3)
F: structure factor
φ: phase of the structure factor

$$\sum\nolimits_1 = \sum_{j=1}^{N} f_j$$

$$\sum\nolimits_2 = \sum_{j=1}^{N} f_j^2$$

Z_j: atomic number of the jth atom

2. Introduction

The statistical properties of the structure factors with integral indices (referred to as 'Wilson statistics') have been carefully investigated since the first contribution by Wilson (1942). The reader will find a updated description of the subject in the splendid monograph by Shmueli & Weiss (1995).

Properties of structure factors with non-integral indices have been used in different contexts inside the crystallographic phase problem. We quote:

(*a*) Boyes-Watson *et al.* (1947) determined the signs of the centrosymmetric structure factors *via* the use of the intensities at non-integral Miller indices;

(*b*) Sayre (1952) underlined that the sign problem in centrosymmetric crystal structures is solvable if the intensities of the reflections with half-integral indices are known;

(*c*) Mishnev (1996) applied the discrete Hilbert transform, previously introduced by Ramachandran (1969), to express structure factors with non-integral indices in terms of standard ones;

(*d*) Zanotti *et al.* (1996) applied the Mishnev results (involving half-integral-index reflections) to extend and improve phase information;

(*e*) In molecular replacement methods, the rotation function (Rossmann & Blow, 1962) may be calculated in reciprocal space as

$$\sum_{\mathbf{p}} |F_{\mathbf{p}}|^2 \left(\sum_{\mathbf{h}} |F_{\mathbf{h}}|^2 G_{\mathbf{hp}} \right),$$

which involves summations over integral indices $\mathbf{h}$ and non-integral indices $\mathbf{p}$.

In spite of the above applications (and others which for shortness are not quoted), no attempt has been made so far to define the statistical properties of the structure factors with non-integral indices. This is the main job of this paper, which may be considered a propaedeutic of the use of such reflections in the phase problem.

3. About the basic assumptions

Wilson statistics hold if one of the following assumptions are made:

(*a*) the atomic positions are assumed to be random variables;

(*b*) the structure is fixed while $\mathbf{h}$ is allowed to vary over reciprocal space.

© 1998 International Union of Crystallography
Printed in Great Britain – all rights reserved

Assumption (*a*) answers different questions like: what is the expected average value of a given reflection **h**? How is the $|F_{\mathbf{h}}|^2$ value of a given reflection **h** distributed around the expected value? Assumption (*b*) answers questions like: for a given set of structure factors, what is the expected average value of the $|F_{\mathbf{h}}|^2$'s? How are the $|F_{\mathbf{h}}|^2$ values of the set distributed around the average?

Evidently, the two statistical approaches are distinct; however, Weyl's (1916) theorem proves that both can be described by the same formulas. The Weyl theorem may be expressed as: when an $\mathbf{r}_j$ vector has rationally independent x_j, y_j, z_j components then the fractional part of $\mathbf{h}\cdot\mathbf{r}_j$ is uniformly distributed within the interval $(0, 1)$ when **h** varies in the domain of the integer numbers. As a consequence, $2\pi\mathbf{h}\cdot\mathbf{r}_j$ is uniformly distributed in the interval $(0, 2\pi)$, no matter if $\mathbf{r}_j$ varies over rational numbers in the interval $(0, 1)$ or **h** over the integer components.

Unfortunately, Weyl's theorem cannot be applied to structure factors with non-integral indices, therefore assumptions (*a*) and (*b*) will lead to different results. Since reflections with different indices may have quite different statistical properties, assumption (*b*), even if practicable, would obscure important properties of the distributions. In particular, it is not relevant to the phase problem. Thus, in all our calculations we will adopt assumption (*a*), under the explicit condition that the atomic coordinates are uniformly distributed in the interval $(0, 1)$. It is worthwhile noting that different choices for this interval [*e.g.* the range $(-1/2, 1/2)$] do not affect the Wilson statistics but they do affect the statistics of the structure factors with non-integral indices. Obviously, our mathematical approach may be applied to any interval but the mathematical results we obtain in this paper are strictly dependent on the stated assumption that the atomic coordinates lie in the interval $(0, 1)$. The modifications to be expected for different intervals are discussed in §13.

Our final formulas are rather more complicated than Wilson's distributions. To explain their features, we will first treat the one-dimensional problem and then we will extend our calculations to three dimensions.

4. The one-dimensional acentric distributions

Let us consider a one-dimensional crystal, with period *a*: no element of symmetry is present. We have

$$F_p = A_p + iB_p = \sum_{j=1}^{N} f_j \exp(2\pi i p x_j),$$
$$0 \le x_j < 1 \quad \text{for} \quad j = 1, \ldots, N.$$

The characteristic function of the distribution $P(A_p, B_p)$, say

$$C(u, v) = \langle \exp i(uA_p + vB_p)\rangle,$$

may be written in terms of the cumulants K_{rs} of the distribution. If only terms up to second order are considered, we have

$$C(u, v) \approx \exp[i(uK_{10} + vK_{01}) - \tfrac{1}{2}(u^2K_{20} + v^2K_{02} + 2uvK_{11})].$$

In their turn, the K_{rs}'s may be expressed in terms of the moments m_{rs} of $P(A_p, B_p)$:

$$K_{10} = m_{10} = \langle A_p\rangle$$
$$K_{01} = m_{01} = \langle B_p\rangle$$
$$K_{20} = m_{20} - m_{10}^2 = \langle A_p^2\rangle - \langle A_p\rangle^2$$
$$K_{02} = m_{02} - m_{01}^2 = \langle B_p^2\rangle - \langle B_p\rangle^2$$
$$K_{11} = m_{11} - m_{10}m_{01} = \langle A_pB_p\rangle - \langle A_p\rangle\langle B_p\rangle.$$

The expressions of the cumulants have been explicitly given to emphasize the fact that, unlike for Wilson's statistics, the moments m_{10}, m_{01}, m_{11} are nonvanishing when p is a non-integral value. Then

$$\begin{aligned} P(A_p, B_p) &= (2\pi)^{-2} \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} C(u, v)\exp[-i(uA_p + vB_p)]\,du\,dv \\ &= (2\pi)^{-2} \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \exp\{-i[u(A_p - K_{10}) + v(B_p - K_{01})]\} \\ &\quad\times \exp[-(u^2K_{20} + v^2K_{02} + 2uvK_{11})/2]\,du\,dv. \end{aligned}$$

The integral may be calculated by repeated application of the standard formula

$$\int_{-\infty}^{+\infty} \exp(itu - \tfrac{1}{2}qu^2)\,du = (2\pi/q)^{1/2}\exp[-t^2/(2q)].$$

The conclusive result is

$$P(A, B) = (2\pi)^{-1}\Delta^{-1/2}\exp\{-(2\Delta)^{-1}[K_{02}(A - K_{10})^2 + K_{20}(B - K_{01})^2 - 2(A - K_{10})(B - K_{01})K_{11}]\} \tag{1}$$

where

$$\Delta = (K_{02}K_{20} - K_{11}^2).$$

The distribution (1) may be expressed in terms of $|F_p|$ and φ_p by standard techniques. We obtained

$$\begin{aligned} P(|F_p|, \varphi_p) = \exp[-q_1/(2\Delta)](2\pi)^{-1}\Delta^{-1/2}|F_p| \\ \times \exp\big(-(1/2\Delta)|F_p|^2\{(K_{02} + K_{20})/2 \\ + [(K_{02} - K_{20})/2]\cos 2\varphi - K_{11}\sin 2\varphi\} \\ - (|F_p|/\Delta)[(K_{01}K_{11} - K_{02}K_{10})\cos\varphi \\ + (K_{10}K_{11} - K_{20}K_{01})\sin\varphi]\big), \end{aligned} \tag{2}$$

where

$$q_1 = (K_{02}K_{10}^2 + K_{20}K_{01}^2 - 2K_{11}K_{01}K_{10}). \tag{3}$$

From (2), the following marginal distributions may be found:

(*a*)

$$P(|F_p|) = \exp[-q_1/(2\Delta)](2\pi)^{-1}(|F_p|/\Delta^{1/2}) \times \exp\{(-|F_p|^2/2\Delta)[(K_{02}+K_{20})/2]\}q_2, \quad (4)$$

where

$$q_2 = \int_0^{2\pi} \exp\left\{(-|F_p|^2/2\Delta)\{[(K_{02}-K_{20})/2]\cos 2\varphi - K_{11}\sin 2\varphi\} - (|F_p|/\Delta)[(K_{01}K_{11}-K_{02}K_{10})\cos\varphi + (K_{10}K_{11}-K_{20}K_{01})\sin\varphi]\right\} d\varphi \quad (5)$$

is a factor which does not depend on φ.

(*b*)

$$P(\varphi) = \exp[-q_1/(2\Delta)](2\pi)^{-1}\Delta^{-1/2} \times \int_0^{\infty} \big(|F_p| \exp(-|F_p|^2/2\Delta)\{[(K_{02}+K_{20})/2] + [(K_{02}-K_{20})]/2\cos 2\varphi - K_{11}\sin 2\varphi\} - (|F_p|/\Delta)[(K_{01}K_{11}-K_{02}K_{10})\cos\varphi + (K_{10}K_{11}-K_{20}K_{01})\sin\varphi]\big)\, d|F_p|. \quad (6)$$

The distributions (2), (4) and (6) are the first results of this paper and will be analysed in the following sections. In particular, we note that, unlike for Wilson's statistics, phase values can be assigned to non-integral-index reflections. However, their role can be better understood if the origin problem is considered (see §13).

5. The cumulants for the one-dimensional acentric case

Let us denote

$$c_p = \sin(2\pi p)/(2\pi p), \quad s_p = [1-\cos(2\pi p)]/(2\pi p).$$

By analogy, c_{2p} and s_{2p} are the values of c and s calculated for the reflection with index $2p$, *i.e.*

$$c_{2p} = \sin(4\pi p)/(4\pi p), \quad s_{2p} = [1-\cos(4\pi p)]/(4\pi p).$$

Accordingly,

$$c_{p/2} = \sin(\pi p)/(\pi p), \quad s_{p/2} = [1-\cos(\pi p)]/(\pi p).$$

From Appendix *A*, the following expressions for the cumulants arise:

$$K_{10} = \textstyle\sum_1 c_p$$
$$K_{01} = \textstyle\sum_1 s_p$$
$$K_{20} = 0.5\textstyle\sum_2(1 + c_{2p} - 2c_p^2)$$
$$K_{02} = 0.5\textstyle\sum_2(1 - c_{2p} - 2s_p^2)$$
$$K_{11} = 0.5\textstyle\sum_2(s_{2p} - 2c_p s_p).$$

If p is an integer different from zero (say $p = h$), then

$$K_{10} = K_{01} = K_{11} = 0, \quad K_{02} = K_{20} = \textstyle\sum_2/2$$

and (1), (4) and (6) reduce to the classical Wilson distributions

$$P(A_h, B_h) \approx \left(\pi\textstyle\sum_2\right)^{-1} \exp\left[-(A_h^2+B_h^2)/\textstyle\sum_2\right]$$
$$P(|F_h|) \simeq 2|F_h|\textstyle\sum_2^{-1} \exp\left(-|F_h|^2/\textstyle\sum_2\right),$$
$$P(\varphi_h) = 1/2\pi.$$

The study of the distributions (4) and (6) for non-integral values of p requires supplementary observations.

6. About the expected value of $|F_p|^2$

According to Appendix *A*,

$$\langle|F_p|^2\rangle = m_{20} + m_{02} = \textstyle\sum_2[1-(c_p^2+s_p^2)] + \sum_1^2(c_p^2+s_p^2),$$

which may be arranged in the simpler formula

$$\langle|F_p|^2\rangle = \textstyle\sum_2(1-c_{p/2}^2) + \sum_1^2 c_{p/2}^2. \quad (7)$$

It is easily verified that, for $p = 0$, $c_p = c_{p/2} = 1$; furthermore, $c_p = 0$ when p is an integer or a half-integer, $c_{p/2} = 0$ for integral values of p, $c_{p/2} = \pm 1/(\pi p)$ for half-integral values of p. As a consequence:

(*a*) For integral values of p (provided $p \neq 0$), Wilson's relation

$$\langle|F_p|^2\rangle_W = \textstyle\sum_2 \quad (8)$$

is obtained.

(*b*) For equal-atom structures, $\sum_1^2 = N\sum_2$ and

$$\langle|F_p|^2\rangle = \textstyle\sum_2[1+(N-1)c_{p/2}^2].$$

Accordingly, $\langle|F_p|^2\rangle$ may attain large values at half-integral (and small) values of p: *i.e.*

$$\langle|F_p|^2\rangle \approx \textstyle\sum_2[1+(N-1)/(\pi^2p^2)].$$

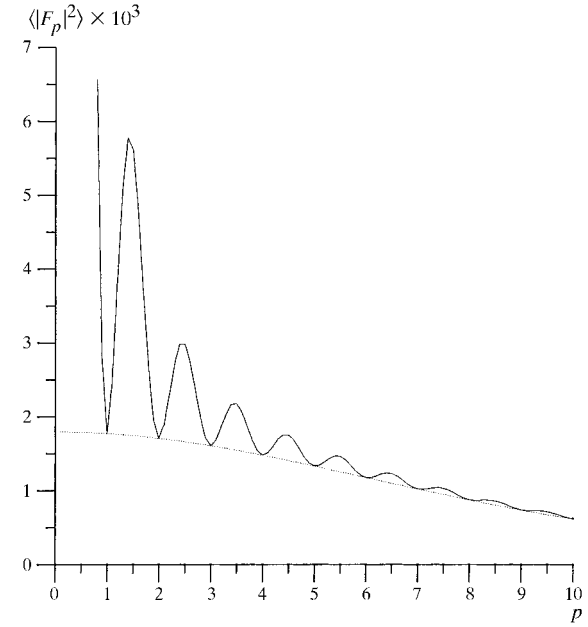
(*c*) For $p = 0$, the relation $\langle|F_p|^2\rangle = \sum_1^2$ arises, which agrees with the well known relation $F_0 = \sum_{j=1}^N Z_j$. It is thus suggested that our statistical approach should hold also in the vicinity of (and at) $p = 0$.

The main features of (7) may be appreciated by plotting (see Fig. 1*a*) $\langle|F_p|^2\rangle$ *versus* p for a 50 equal-atom random structure (RAND50 in code; all the atoms are assumed to be carbon, with the same isotropic temperature factor $B_T = 5$ Å^2). In the same figure, we also draw $\langle|F_p|^2\rangle_W$ (broken line). We note:

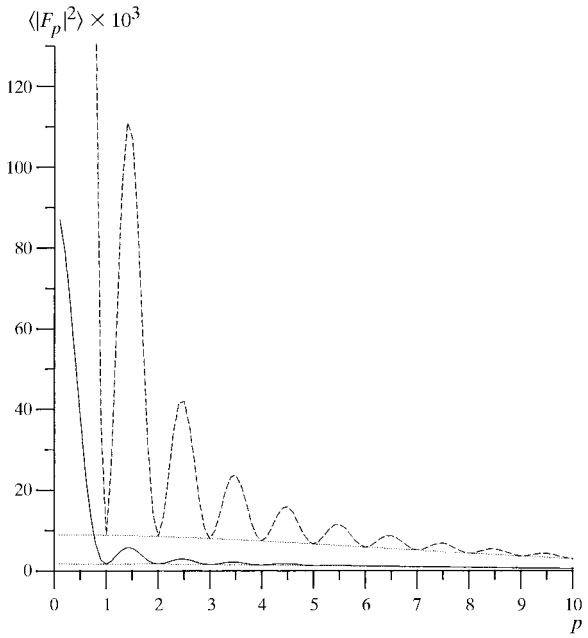
(*a*) Both $\langle|F_p|^2\rangle$ and $\langle|F_p|^2\rangle_W$ decay with $\sin\theta/\lambda$ but the first is an oscillating function with maxima at half-integral values of p (but for $p = 1/2$) and minima at integral values. The amplitudes of the oscillations decay with p but are still non-negligible up to $p = 9.5$. The practical consequence of (7) is that the moduli of the structure factors with indices close to some half-integer are expected to be larger (on average) than the moduli of the reflections with integral indices. This trend is

clearly confirmed by Table 1, where the $|F|$ values for integral and half-integral indices are given up to $p = 10$.

(*b*) $\langle|F_p|^2\rangle$ regularly decays in the interval (0, 1) (thus $p = 0.5$ is the only half-integer index that does not correspond to a relative maximum). The absolute maximum is attained at $p = 0$, where $\langle|F_p|^2\rangle$ assumes the value $(\sum_{j=1}^{N} Z_j)^2 = 90\,000$, corresponding to $F_0 = 300$.



(*a*)



(*b*)

Fig. 1. (*a*) RAND50: $\langle|F_p|^2\rangle$ as given by (7) (full line) is compared with the $\sum_2$ function (broken line); (*b*) the $\langle|F_p|^2\rangle$ function is plotted against p for RAND50 (full line) and RAND250 (broken line). The corresponding $\langle|F_p|^2\rangle_W$ curves are plotted by dotted lines.

Table 1. *RAND50 : list of |F|'s and φ's for integral and half-integral indices up to p = 10*

p	$\|F\|$	φ (°)
0.5	945.72	93
1.0	73.00	280
1.5	317.11	79
2.0	2.93	150
2.5	143.19	73
3.0	62.87	9
3.5	110.99	58
4.0	195.41	25
4.5	300.21	96
5.0	72.90	162
5.5	135.88	103
6.0	36.06	194
6.5	126.65	93
7.0	96.32	198
7.5	67.79	3
8.0	101.93	58
8.5	186.21	104
9.0	180.87	171
9.5	97.75	228
10.0	46.83	287

(*c*) The amplitudes of the oscillations increase with N. In Fig. 1(*b*), we plot the $\langle|F_p|^2\rangle$ curves for RAND50 and for RAND250 [this last being an equal-atom (C) random structure, with $N = 250$ and $B_T = 5$ Å^2]. For this last structure, the amplitudes of the oscillations are much larger than those calculated for RAND50; furthermore, at least at small values of the half-integral indices, they are even larger than $\langle|F_p|^2\rangle_W$. The size of this effect and its trend against the structural complexity may be appreciated from Fig. 2, where the ratio $\langle|F_p|^2\rangle/\sum_2$ is shown for RAND50, RAND250 and RAND500, this last being an equal-atom (C) random structure with $N = 500$ and $B_T = 5$ Å^2. The ratio is expected to be 1 at large values of p: the reader may observe that the convergence rate is low for large values of N.

The above observations suggest that the distributions of the structure factors with non-integral indices are expected to be different from Wilson's distributions; the differences are expected to be larger for structure factors with half-integral (or close to half-integral) indices.

7. The normalized one-dimensional acentric distribution $P(|E_p|)$

The concept of a normalized structure factor preserves its full meaning even for non-integral indices. We define

$$E_p = F_p/\langle|F_p|^2\rangle^{1/2}. \qquad (9)$$

It may be useful to note that, according to the above definition, $E_0 = 1$ (while, in Wilson statistics, $E_0 = N^{1/2}$).

The study of distributions (4) and (6) will be made after having: (*a*) expressed them in terms of E rather than of F; and (*b*) replaced in (4) and (6) the cumulant expressions derived in §5. We obtain for $P(|E_p|)$ the following formula:

$$P(|E_p|) \approx (2\pi)^{-1} \exp(-v_0) v_1 |E_p| \exp(-|E_p|^2 v_2) q_2, \tag{10}$$

where

$$q_2 = \int_0^{2\pi} \exp[-|E_p|^2(v_3 \cos 2\varphi - v_4 \sin 2\varphi) - |E_p|(v_5 \cos\varphi + v_6 \sin\varphi)]\, d\varphi.$$

The expressions for v_i, for $i = 0, \ldots, 6$, are given below:

$$\begin{aligned}
v_0 &= q_1/(2\Delta) = \left(\textstyle\sum_1^2 / \sum_2\right) c_{p/2}^2 (1 - c_p)/\delta, \\
\delta &= (1 - c_p)(1 + c_p - 2c_{p/2}^2) \\
v_1 &= 2\langle |F_p|^2 \rangle \Big/ \left(\textstyle\sum_2 \delta^{1/2}\right) \\
v_2 &= \langle |F_p|^2 \rangle (1 - c_{p/2}^2) \Big/ \left(\textstyle\sum_2 \delta\right) \\
v_3 &= \langle |F_p|^2 \rangle [-c_{2p} + (c_p^2 - s_p^2)] \Big/ \left(\textstyle\sum_2 \delta\right) \\
v_4 &= \langle |F_p|^2 \rangle (s_{2p} - 2c_p s_p)] \Big/ \left(\textstyle\sum_2 \delta\right) \\
v_5 &= 2\left(\textstyle\sum_1 / \sum_2\right) c_p (c_p - 1) \langle |F_p|^2 \rangle^{1/2} / \delta \\
v_6 &= 2\left(\textstyle\sum_1 / \sum_2\right) s_p (c_p - 1) \langle |F_p|^2 \rangle^{1/2} / \delta.
\end{aligned}$$

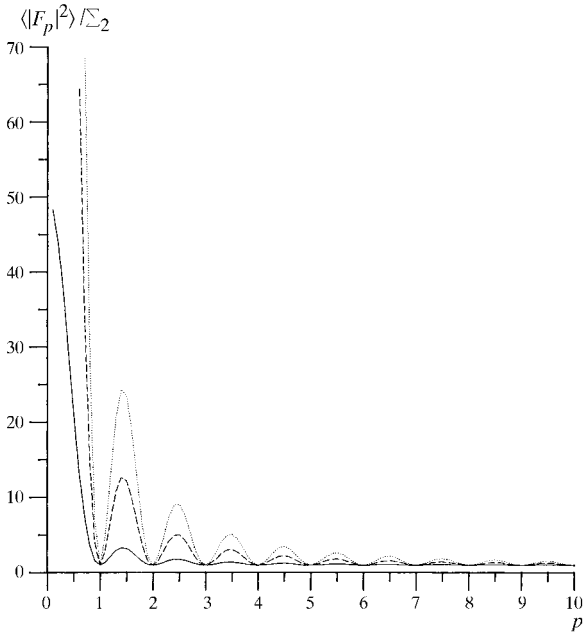


Fig. 2. The ratio $\langle |F_p|^2 \rangle / \sum_2$ is plotted against p for the three random structures RAND50 (full line), RAND250 (broken line) and RAND500 (dotted line).

For integral values of p, $\delta = 1$, $v_1 = 2$, $v_2 = 1$, $v_0 = v_3 = v_4 = v_5 = v_6 = 0$; then (10) reduces to the acentric Wilson distribution

$$P(|E|) = 2|E| \exp(-|E|^2).$$

Let us now evaluate q_2 for non-integral values of p. Denoting

$$\begin{aligned}
v_3 &= X_2 \cos 2\theta_2, \quad v_4 = X_2 \sin 2\theta_2, \\
v_5 &= X_1 \cos\theta_1, \quad v_6 = X_1 \sin\theta_1
\end{aligned}$$

yields

$$q_2 = \int_0^{2\pi} \exp[-|E_p|^2 X_2 \cos 2(\varphi + \theta_2) - |E_p| X_1 \cos(\varphi - \theta_1)]\, d\varphi,$$

where

$$\begin{aligned}
X_2 &= (v_3^2 + v_4^2)^{1/2}, \quad \theta_2 = 0.5 \tan^{-1}(v_4/v_3), \\
X_1 &= (v_5^2 + v_6^2)^{1/2}, \quad \theta_1 = \tan^{-1}(v_6/v_5).
\end{aligned}$$

We then expand $\exp[-|E_p|^2 X_2 \cos 2(\varphi + \theta_2)]$ in a series of Bessel functions according to

$$\exp[-|E_p|^2 X_2 \cos 2(\varphi + \theta_2)] = I_0(|E_p|^2 X_2) + 2 \sum_{n=1}^{\infty} I_n(-|E_p|^2 X_2) \cos 2n(\varphi + \theta_2).$$

The application of the relations

$$\int_0^{2\pi} \cos(n\varphi) \exp(-X \cos\varphi)\, d\varphi = 2\pi I_n(X)$$

$$\int_0^{2\pi} \sin(n\varphi) \exp(-X \cos\varphi)\, d\varphi = 0$$

gives

$$\begin{aligned}
q_2 &= 2\pi I_0(|E_p|^2 X_2) I_0(|E_p| X_1) + 2 \sum_{n=1}^{\infty} I_n(-|E_p|^2 X_2) \\
&\quad \times \int_0^{2\pi} \cos 2n(\varphi + \theta_2) \exp[-|E_p| X_1 \cos(\varphi - \theta_1)]\, d\varphi \\
&= 2\pi \Big[I_0(|E_p|^2 X_2) I_0(|E_p| X_1) \\
&\quad + 2 \sum_{n=1}^{\infty} \cos 2n(\theta_1 + \theta_2) I_n(-|E_p|^2 X_2) I_{2n}(|E_p| X_1) \Big].
\end{aligned}$$

The final result is

$$\begin{aligned}
P(|E_p|) = \exp(-v_0) v_1 |E_p| \exp(-|E_p|^2 v_2) \\
\times \Big[I_0(|E_p|^2 X_2) I_0(|E_p| X_1) \\
+ 2 \sum_{n=1}^{\infty} \cos 2n(\theta_1 + \theta_2) \\
\times I_n(-|E_p|^2 X_2) I_{2n}(|E_p| X_1) \Big].
\end{aligned} \tag{11}$$

An unexpected feature is immediately observed for (11): while Wilson's distribution $P(|E|)$ is universal (*i.e.* it

holds for any structure, protein or small molecule, provided $2\pi\mathbf{h}\cdot\mathbf{r}$ is uniformly distributed over the trigonometric circle), (11) does depend on the structural complexity through the coefficients v_i, $i = 0, \ldots, 6$, δ and $\sum_i$, $i = 1, 2$. Therefore, we have always to specify for which structure a given $P(|E|)$ is calculated.

The Bessel series in (11) is rapidly convergent: about 20 terms are sufficient for a good approximation of $P(|E_p|)$. However, (11) is not easily computable for p too close to zero. For clearness, we will discuss here the main features of $P(|E_p|)$ in the interval $(1, \infty)$, while the properties of the distribution in the interval (0, 1) will be analysed in §9.

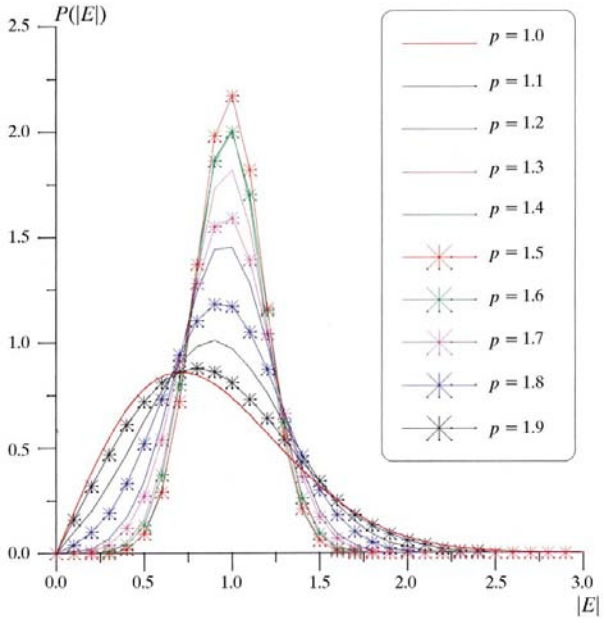


Fig. 3. RAND250: the $P(|E|)$ distribution is plotted for selected values of p between 1 and 2.

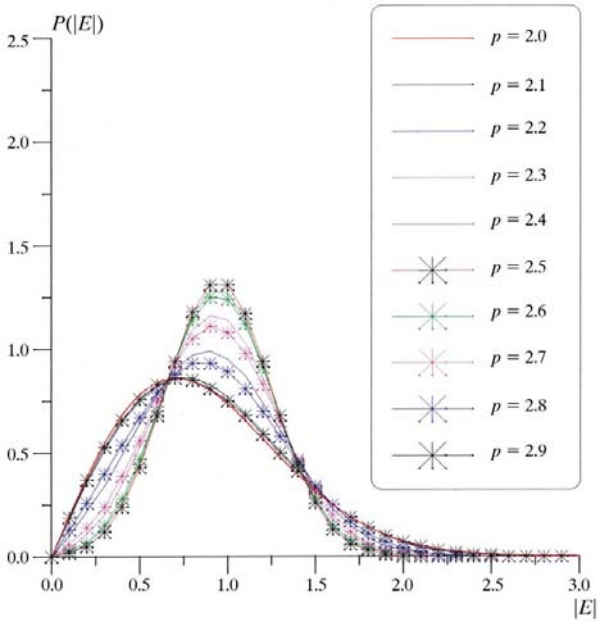


Fig. 4. RAND250: the $P(|E|)$ distribution is plotted for selected values of p between 2 and 3.

The distribution (11), as obtained for RAND250, is plotted in Fig. 3 for selected values of p. We note:

(*a*) Among the curves drawn in Fig. 3, that corresponding to $p = 1.9$ is the closest to the Wilson distribution, that corresponding to $p = 1.5$ is the furthest away. The criterion to rank the curves is the following: if p is very close to some integral number and/or is very large then $P(|E_p|)$ will be very close to the Wilson distribution. Accordingly, 1.9 is the p value that (among the selected values) is the closest to some integer (*i.e.* 2) and has the largest modulus. Then the curve corresponding to $p = 1.1$ will follow since it has the same minimal 'distance' from an integer number but has smaller modulus: The order of the curves may then easily be established: *i.e.* 0.9, 0.1, 0.8, 0.2, 0.7, 0.3, 0.6, 0.4, 0.5.

(*b*) The Wilson distribution ($p = 1$) is the flattest curve of the set, the sharpest one is that corresponding to $p = 1.5$. The figure suggests that good estimates of $|E_p|$ could be made for small values of p provided they are close to half-integral values.

(*c*) As soon as p approaches some half-integral value, the curve becomes more symmetric around the mode. The mode of the distribution occurs at an $|E|$ value close to unity for $p = 1.5$ and decreases up to about 0.71 (the value of the Wilson's distribution) when p approaches some integral number.

(*d*) As soon as p approaches 1.5, the mode of the distribution moves towards values very close to the expected value of $|E_p|$.

In order to show how the distributions depend on the integral part of p, we show in Fig. 4 curves corresponding to selected values of p between 2.0 and 2.9. Its comparison with Fig. 3 suggests that the general features described for Fig. 3 hold for Fig. 4 too (the 'order' of curves, the relative variance, the mode variation, ...); however, the deviations from Wilson's distribution are smaller in Fig. 4 than in Fig. 3. As a consequence, the predictability of the $|E_p|$ values will rapidly decrease when p increases.

An effect of the structural complexity is that the deviations of the $P(|E|)$ curves from Wilson's distributions will increase with N. In Fig. 5, we show the curves calculated for selected values of p in the range (2.0, 2.9) for RAND500 (compare with the corresponding curves in Fig. 4).

8. The distribution $P(\varphi)$ in the one-dimensional acentric case

The distribution (6) may be written in the following shorter form:

$$P(\varphi_p) = (2\pi)^{-1} \exp(-v_0) 2\langle |F_p|^2 \rangle \Big/ \Big(\sum_2 \delta^{1/2} \Big) \times \int_0^\infty |E_p| \exp(-|E_p|^2 \mu - 2\nu |E_p|)\, \mathrm{d}|E_p|, \quad (12)$$

where

$$\mu = v_2 + v_3 \cos 2\varphi - v_4 \sin 2\varphi,$$
$$\nu = (v_5 \cos\varphi + v_6 \sin\varphi)/2.$$

The integral in (12) may be estimated *via* the formula (Gradshteyn & Ryzhik, 1965)

$$\int_0^\infty x \exp(-\mu x^2 - 2\nu x)\, \mathrm{d}x = (2\mu)^{-1} - \nu(2\mu)^{-1}(\pi/\mu)^{1/2} \times \exp(\nu^2/\mu)[1 - \Phi(\nu/\mu^{1/2})],$$

where $\Phi(x)$ is the probability integral defined by

$$\Phi(x) = (2/\pi^{1/2}) \int_0^x \exp(-t^2)\, \mathrm{d}t.$$

Finally, we obtain

$$P(\varphi_p) = (2\pi)^{-1} \exp(-v_0)(2\langle |F_p|^2 \rangle) \Big/ \Big(\sum_2 \delta^{1/2} \Big)(1/2\mu) \times \{1 - \nu(\pi/\mu)^{1/2} \exp(\nu^2/\mu)[1 - \Phi(\nu/\mu^{1/2})]\}. \quad (13)$$

For integral values of p, $v_0 = 0$, $\mu = 1$, $\delta = 1$ and $\nu = 0$, then $P(\varphi_p) = 1/2\pi$, in agreement with Wilson's results. $P(\varphi_p)$ is however non-uniform when $p \neq h$. In Fig. 6, we show, for RAND250, the $P(\varphi_p)$ curves for selected values of p in the interval (1.1, 1.9). We note:

(*a*) the distributions are unimodal;

(*b*) the curves become sharper when p gets nearer the half-integral value;

(*c*) the mode regularly moves from 17° (corresponding to $p = 1.1$) to 160° (corresponding to $p = 1.9$).

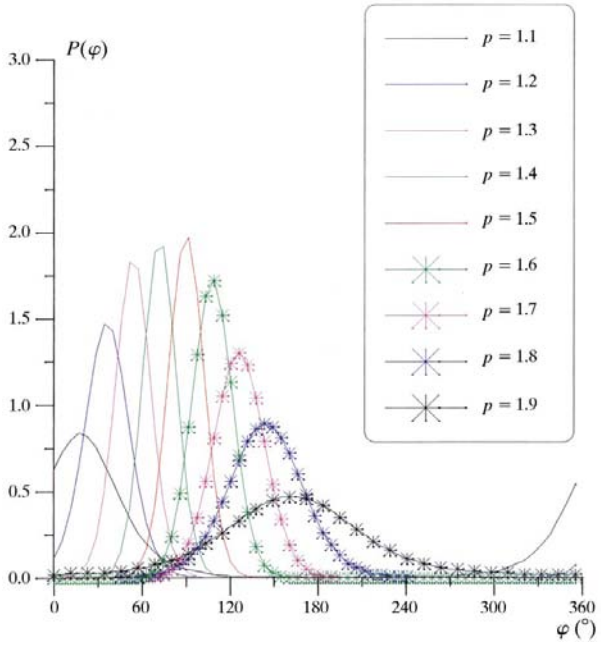


Fig. 6. RAND250: the $P(\varphi)$ curves for selected values of p in the interval (1.1, 1.9).

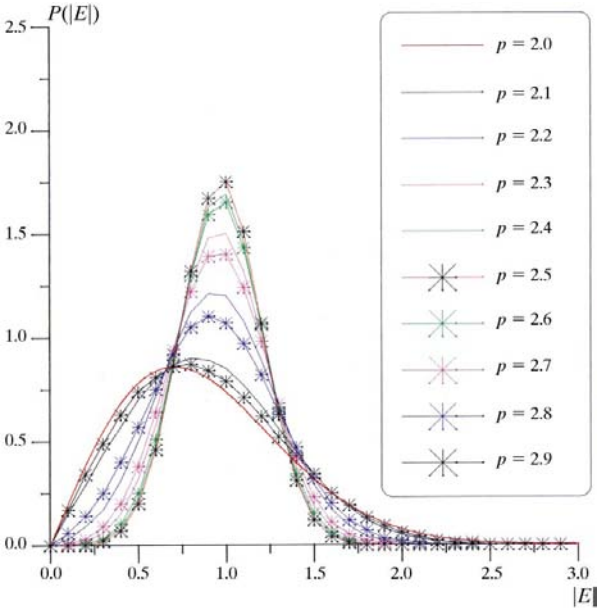


Fig. 5. RAND500: the $P(|E|)$ distribution for selected values of p between 2 and 3.

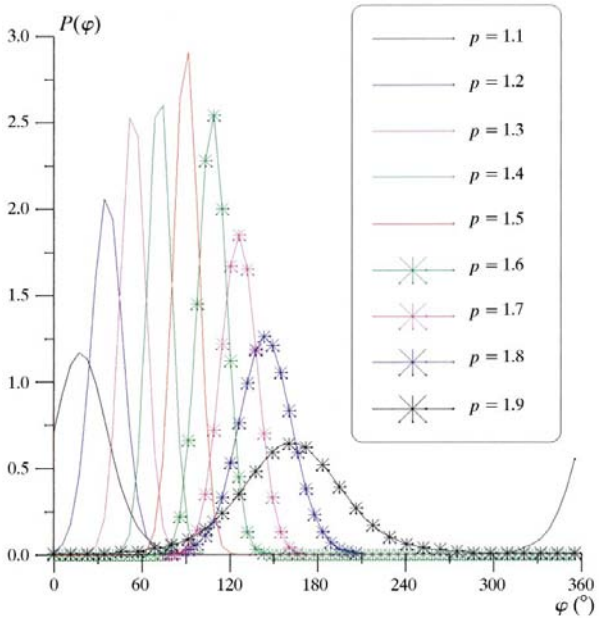


Fig. 7. RAND500: the $P(\varphi)$ curves for selected values of p in the interval (1.1, 1.9).

As an effect of the structural complexity, the curves are sharper with increasing values of N (compare Fig. 7 obtained for RAND500 with Fig. 6). Furthermore, the distributions become flatter when the integral part of p increases (compare Fig. 8 with Fig. 6). From the above considerations, it may be argued that, if p is close to or coincident with a half-integral value, then φ_p is expected to be close to $\pi/2$, with good accuracy provided p is not too large.

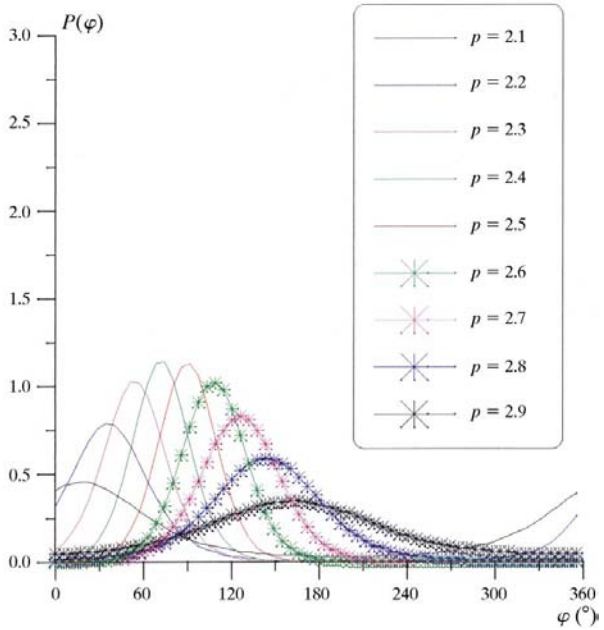


Fig. 8. RAND250: the $P(\varphi)$ curves for selected values p in the interval (2.1, 2.9).

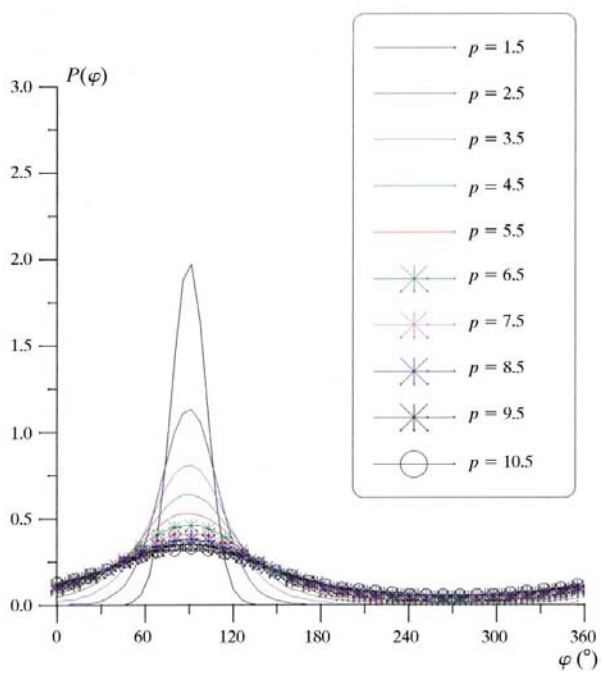


Fig. 9. RAND250: the $P(\varphi)$ curves for some selected half-integral values of p.

In order to provide the reader with some insight about the long-range behaviour of $P(\varphi)$, we show in Fig. 9 for RAND250 the distributions at half-integral values from 1.5 to 10.5. Even at $p = 10.5$, the phase distribution is far from being flat. The reader can verify in Table 1 that the 'true' phases of the reflections with half-integral indices will agree with our expectations.

Some considerations about the phase predictability are useful. When p is not an integer, the phases can be predicted just because $2\pi\mathbf{h} \cdot \mathbf{r}_j$ is not uniformly distributed on the trigonometric circle. This feature may be qualitatively perceived by estimating φ_p as $\tan^{-1}(\langle B\rangle/\langle A\rangle)$. We have

$$\begin{aligned}(\varphi_p)_{\text{est}} &= \tan^{-1}(m_{01}/m_{10}) \\ &= \tan^{-1}[(1-\cos 2\pi p)/\sin 2\pi p] \\ &= \tan^{-1}[2(\sin^2 \pi p)/\sin 2\pi p].\end{aligned}$$

Since the numerator is always positive, φ_{est} is restricted to the interval $(0, \pi)$ and will be closer to 0° if p is smaller than the closest half-integer, closer to π if p is larger than the closest half-integer.

9. About the features of $P(|E_p|)$ and $P(\varphi_p)$ for $0 \le p < 1$

In accordance with Appendix B, distribution (11) is computable for $p \ge 0.8$; for $p < 0.8$, the arguments of the exponential and of the Bessel functions are too large even for a modern computer. Such behaviour is not unexpected: indeed, $P(|E_p|)$ at $p = 0$ should coincide with the δ function,

$$P(|E_p|) = \delta(|E_p| - 1),$$

centred at $E_0 = 1$. This is because at $p = 0$ E_p is perfectly estimated *via* the algebraic (and therefore certain) relationship $E_p = 1$.

Similarly, the distribution $P(\varphi_p)$ is not computable for $p < 0.6$: however, there is no doubt that it is expected to coincide with the δ function

$$P(|\varphi_p|) = \delta(\varphi_p)$$

centred at $\varphi = 0$ (indeed, E_0 is real and positive). We can then look at the distribution $P(|E_p|)$ and $P(\varphi_p)$ for p in the interval $(0, 1)$ as a family of curves approaching δ functions when $p \to 0$. This behaviour may be perceived by observing Figs. 10 and 11: for RAND250 in Fig. 10, $P(|E_p|)$ is drawn for $p = 1, 0.9, 0.8$, and in Fig. 11 $P(\varphi_p)$ is drawn for $p = 0.9, 0.8, 0.7, 0.6$.

10. The conditional distribution $P(\varphi_p \| E_p|)$ in the one-dimensional acentric case

It may occur that $|F_p|$, and therefore $|E_p|$, are known from other sources. This information may be used as prior for a more accurate estimate of φ_p. We have

$$\begin{aligned} P(\varphi_p||E_p|) &= P(\varphi_p, E_p) \Big/ \int_0^{2\pi} P(\varphi_p, |E_p|)\,\mathrm{d}\varphi_p \\ &= \exp[-|E_p^2|(v_3 \cos 2\varphi - v_4 \sin 2\varphi) \\ &\quad - |E_p|(v_5 \cos\varphi + v_6 \sin\varphi)]/q_2. \end{aligned} \tag{14}$$

To see how (14) varies with $|E_p|$ for a fixed p, we draw it in Fig. 12 for three values of $|E_p|$ (RAND250 and $p = 1.5$ have been used). We see that (14) strongly depends on the known $|E_p|$ value: it is sharper if $|E_p|$ is known to be large; it is flat if $|E_p|$ is known to be small. In conclusion, the confidence one should have in φ_p will depend on the prior information on $|E_p|$: if this last is unknown, the distribution (11) may be used.

The conclusions of this section answer a rather intriguing question: 'let us suppose that $\varphi_h \approx \pi$. How can we trust in the distribution (11) if this always estimates $\varphi_p = 0$ for p very close to and larger than h? Such an estimate indeed should be in conflict with the expected continuity of the function φ_p. The question may be answered *via* the use of (14). If $\varphi_h \approx \pi$ in the p interval immediately following the h position [say $(h, h + 0.1)$], we can expect that the true phases vary very rapidly: that will be in better agreement with (14) only if small $|E_p|$ are associated with the p values in the interval. As an example, we plot in Figs. 13(*a*), (*b*), for RAND250, the true $|E_p|$ and φ_p values against p for p between 5.5 and 7. Since $\varphi_6 = 194^\circ$ and $|E_6| = 0.47$, the $|E_p|$ values, for p immediately following $h = 6$, collapse to very small values of $|E_p|$ to fit better the relationship (14).

11. The conditional distribution $P(|E_p| \| \varphi_p)$ in the one-dimensional acentric case

It may occur that φ_p is known from other sources. This information may be used as prior for a more accurate estimate of $|E_p|$. We have

$$\begin{aligned} p(|E_p||\varphi_p) &= \frac{|E_p| \exp(-|E_p|^2\mu - 2\nu|E_p|)}{\int_0^\infty |E_p| \exp(-|E_p|^2\mu - 2\nu|E_p|)\,\mathrm{d}|E_p|} \\ &= |E_p| \exp(-|E_p|^2\mu - 2\nu|E_p|) \\ &\quad \times \left((2\mu)^{-1}\{1 - \nu(\pi/\mu)^{1/2} \exp(\nu^2/\mu) \right. \\ &\quad \left. \times [1 - \Phi(\nu/\mu^{1/2})]\}\right)^{-1}. \end{aligned} \tag{15}$$

To check how the distribution (15) varies against φ_p, we draw it in Fig. 14 for RAND250 and $p = 1.5$. We see that the expected value of $|E_p|$ when φ_p is known to be equal to $\pi/2$ is close to unity, as in the case in which φ_p is unknown (see Fig. 5). If we introduce in (15) the prior knowledge that $\varphi_p = 2\pi$, then the expected value of $|E_p|$ remarkably diminishes. The rational is the following: the prior information that the phase φ_p is far from its expected value will generate, through (15), estimates for $|E_p|$ concentrated about small values of $|E|$ but the relative distribution may be sharp.

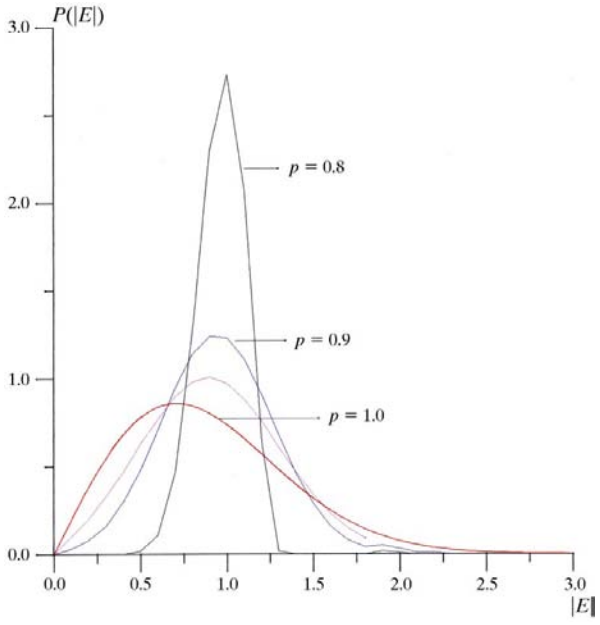


Fig. 10. RAND250: the $P(|E|)$ curves for p = 1, 0.9, 0.8.

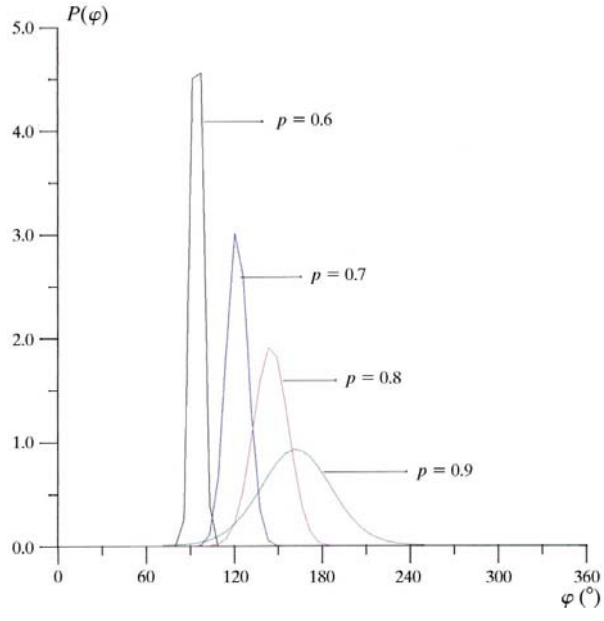


Fig. 11. RAND250: the $P(\varphi)$ curves for p = 0.9, 0.8, 0.7, 0.6.

12. The distribution of structure factors with non-integral indices in *P*1

In $P1$ (three-dimensional case),

$$\begin{aligned} F_{\mathbf{p}} &= \sum_{j=1}^{N} f_j \cos 2\pi(p_1x_j + p_2y_j + p_3z_j) \\ &\quad + i\sum_{j=1}^{N} f_j \sin 2\pi(p_1x_j + p_2y_j + p_3z_j) \\ &= A_{\mathbf{p}} + iB_{\mathbf{p}}. \end{aligned}$$

Suppose now that x_j, y_j, z_j are randomly and independently distributed in the interval $(0, 1)$. Then the distributions (2), (4) and (6) will still hold provided the corresponding cumulants are available. We obtain

$$\begin{aligned} k_{10} &= \langle A_{\mathbf{p}}\rangle = \textstyle\sum_1 c_{\mathbf{p}} \\ k_{01} &= \langle B_{\mathbf{p}}\rangle = \textstyle\sum_1 s_{\mathbf{p}}, \end{aligned}$$

where

$$\begin{aligned} c_{\mathbf{p}} &= c_{p_1}c_{p_2}c_{p_3} - c_{p_1}s_{p_2}s_{p_3} - s_{p_1}s_{p_2}c_{p_3} - s_{p_1}c_{p_2}s_{p_3} \\ &= \cos\pi(p_1 + p_2 + p_3)c_{p_1/2}c_{p_2/2}c_{p_3/2} \\ s_{\mathbf{p}} &= s_{p_1}c_{p_2}c_{p_3} - s_{p_1}s_{p_2}s_{p_3} + c_{p_1}s_{p_2}c_{p_3} + c_{p_1}c_{p_2}s_{p_3} \\ &= \sin\pi(p_1 + p_2 + p_3)c_{p_1/2}c_{p_2/2}c_{p_3/2} \\ c_{p_i} &= \sin(2\pi p_i)/(2\pi p_i) \\ s_{p_i} &= [1 - \cos(2\pi p_i)]/(2\pi p_i). \end{aligned}$$

The cumulants of order two may be expressed in terms of cumulants of order one as follows:

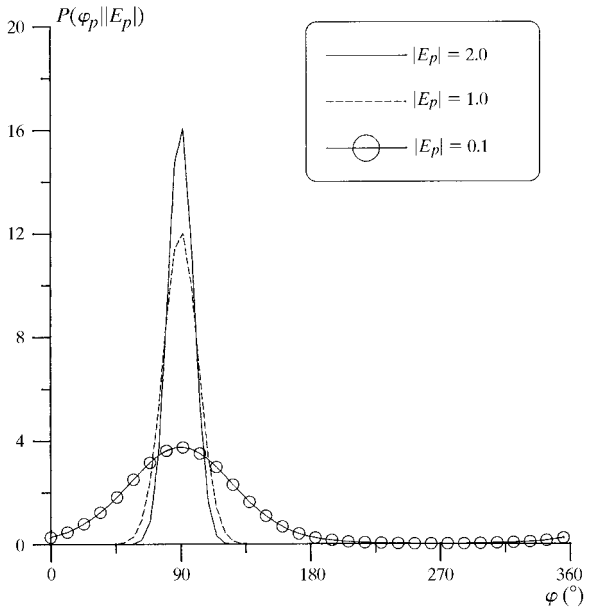


Fig. 12. RAND250: the $P(\varphi||F_{\mathbf{p}}|)$ curves for $p = 1.5$ and selected values of $|E_{\mathbf{p}}|$.

$$\begin{aligned} K_{20} &= \langle A_{\mathbf{p}}^2\rangle - \langle A_{\mathbf{p}}\rangle^2 = \left(\textstyle\sum_2/2\right)(1 + c_{2\mathbf{p}} - 2c_{\mathbf{p}}^2) \\ K_{02} &= \left(\textstyle\sum_2/2\right)(1 - c_{2\mathbf{p}} - 2s_{\mathbf{p}}^2) \\ K_{11} &= \left(\textstyle\sum_2/2\right)(s_{2\mathbf{p}} - 2c_{\mathbf{p}}s_{\mathbf{p}}), \end{aligned}$$

where $c_{2\mathbf{p}}$ and $s_{2\mathbf{p}}$ are the c and s values calculated for the reflections with indices $2p_1, 2p_2, 2p_3$. The probability distribution $P(|F_{\mathbf{p}}|)$ will now be

$$P(|F_{\mathbf{p}}|) \approx \exp(-t_1)(2\pi)^{-1}|F_{\mathbf{p}}|\Delta^{-1/2}\exp(-t_2|F_{\mathbf{p}}|^2)q_2, \tag{16}$$

where

$$\begin{aligned} t_1 &= (K_{02}K_{10}^2 + K_{20}K_{01}^2 - 2K_{11}K_{01}K_{10})/(2\Delta) \\ \Delta &= (K_{02}K_{20} - K_{11}^2) \\ q_2 &= 2\pi\Big[I_0(|F_{\mathbf{p}}|^2X_2)I_0(|F_{\mathbf{p}}|X_1) \\ &\qquad + 2\sum_{n=1}^{\infty}\cos 2n(\theta_1 + \theta_2)I_n(-|F_{\mathbf{p}}|^2X_2)I_{2n}(-|F_{\mathbf{p}}|X_1)\Big] \\ X_2 &= (t_3^2 + t_4^2)^{1/2} \\ X_1 &= (t_5^2 + t_6^2)^{1/2} \\ t_2 &= (K_{02} + K_{20})/(4\Delta) \\ t_3 &= (K_{02} - K_{20})/(4\Delta) \\ t_4 &= K_{11}/(2\Delta) \\ t_5 &= (K_{01}K_{11} - K_{02}K_{10})/\Delta \\ t_6 &= (K_{10}K_{11} - K_{20}K_{01})/\Delta \\ \theta_2 &= 0.5\tan^{-1}(t_4/t_3) \\ \theta_1 &= \tan^{-1}(t_6/t_5). \end{aligned}$$

The distribution in terms of the normalized modulus $P(|E_{\mathbf{p}}|)$ is trivially obtained from (16) *via* the transformation

$$E_{\mathbf{p}} = F_{\mathbf{p}}/\langle|F_{\mathbf{p}}|^2\rangle^{1/2} = F_{\mathbf{p}}/(m_{20} + m_{02})^{1/2},$$

where

$$\begin{aligned} m_{20} &= 0.5\textstyle\sum_2(1 + c_{2\mathbf{p}} - 2c_{\mathbf{p}}^2) + \sum_1^2 c_{\mathbf{p}}^2 \\ m_{02} &= 0.5\textstyle\sum_2(1 - c_{2\mathbf{p}} - 2s_{\mathbf{p}}^2) + \sum_1^2 s_{\mathbf{p}}^2. \end{aligned}$$

The distribution (6) for $P1$ can be written as

$$\begin{aligned} P(\varphi_{\mathbf{p}}) \approx \exp(-t_1)(2\pi)^{-1}\Delta^{-1/2}(1/2\mu)\{1 - \nu(\pi/\mu)^{1/2} \\ \times \exp(\nu^2/\mu)[1 - \Phi(\nu/\mu^{1/2})]\}, \end{aligned} \tag{17}$$

where

$$\begin{aligned} \mu &= t_2 + t_3\cos 2\varphi - t_4\sin 2\varphi \\ \nu &= (t_5\cos\varphi + t_6\sin\varphi)/2. \end{aligned}$$

In their turn, the conditional probabilities $P(\varphi_{\mathbf{p}}||F_{\mathbf{p}}|)$ and $P(|F_{\mathbf{p}}||\varphi_{\mathbf{p}})$ can be expressed as follows:

$$P(|F_{\mathbf{p}}||\varphi_{\mathbf{p}}) = |F_{\mathbf{p}}| \exp(-|F_{\mathbf{p}}|^2\mu - 2\nu|F_{\mathbf{p}}|) \times (2\mu)^{-1}\{1 - \nu(\pi/\mu)^{1/2} \exp(\nu^2/\mu) \times [1 - \Phi(\nu/\mu^{1/2})]\}^{-1} \quad (18)$$

$$P(\varphi_{\mathbf{p}}||F_{\mathbf{p}}|) = \exp[-|F_{\mathbf{p}}^2|(t_3 \cos 2\varphi - t_4 \sin 2\varphi) - |F_{\mathbf{p}}|(t_5 \cos \varphi + t_6 \sin \varphi)]/q_2. \quad (19)$$

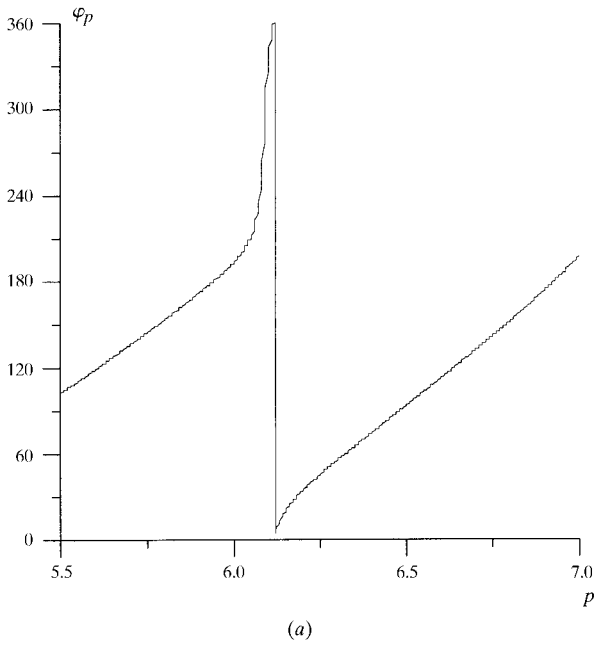


Fig. 13. (*a*) RAND250: true $\varphi_{\mathbf{p}}$ values against p for p lying in the interval (5.5, 7). (*b*) RAND250: true $|E_{\mathbf{p}}|$ values against p for p lying in the interval (5.5, 7).

The behaviour of the above distributions [*i.e.* from (16) to (19)] are quite similar to those described for the corresponding one-dimensional case, thus we will not expand further. However, the three-dimensionality introduces a larger variety of reflection types: *e.g.* reflections $h + 0.5, 0, 0$, $h + 0.5, k + 0.5, 0$, $h + 0.5, k + 0.5, l + 0.5$ may have quite different distributions. To provide a fast if qualitative insight into the problem, we analyse the ratio $\langle|F_{\mathbf{p}}|^2\rangle / \sum_2$. The expected value

$$\langle|F_{\mathbf{p}}|^2\rangle = \sum_2 [1 - (c_{\mathbf{p}}^2 + s_{\mathbf{p}}^2)] + \sum_1^2 (c_{\mathbf{p}}^2 + s_{\mathbf{p}}^2)$$

reduces, for an equal-atom structure, to

$$\langle|F_{\mathbf{p}}|^2\rangle = \sum_2 [1 + (N - 1)(c_{\mathbf{p}}^2 + s_{\mathbf{p}}^2)].$$

Accordingly,

$$\langle|F_{\mathbf{p}}|^2\rangle \Big/ \sum_2 = 1 + (N - 1)(c_{\mathbf{p}}^2 + s_{\mathbf{p}}^2) \quad (20)$$

is an oscillating function: the oscillations are stronger for large structures. The algebraic form of (20) does not allow immediate specification of where the maxima and minima of $\langle|F_{\mathbf{p}}|^2\rangle / \sum_2$ are. A further algebraic analysis led us to the simple expression

$$\langle|F_{\mathbf{p}}|^2\rangle \Big/ \sum_2 = 1 + (N - 1)c_{p_1/2}^2 c_{p_2/2}^2 c_{p_3/2}^2,$$

from which the following rule arises: the largest values of $\langle|F_{\mathbf{p}}|^2\rangle / \sum_2$ occur when all of p_1, p_2 and p_3 are half-integer or zero. When one of the indices is an integer (different from zero) then $\langle|F_{\mathbf{p}}|^2\rangle / \sum_2 = 1$ as for the Wilson distributions. In accordance with the above rule, the **p** vectors with components $(1/2, 0, 0)$, $(0, 1/2, 1/2)$ or $(1/2, 1/2, 1/2)$ will correspond to maxima of

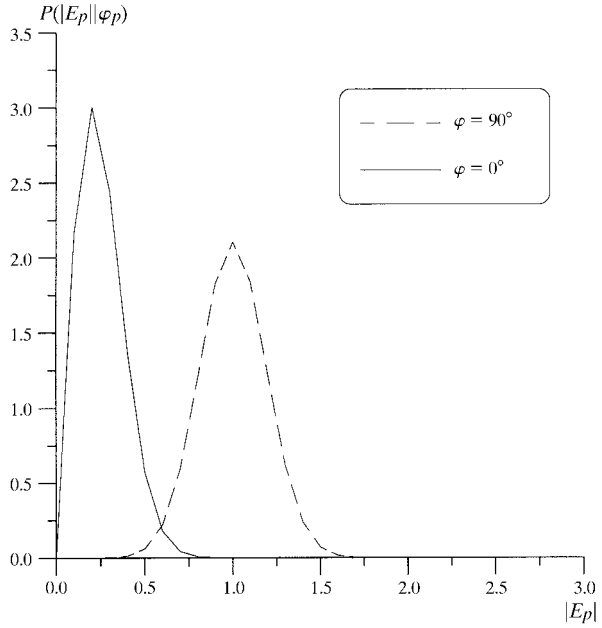


Fig. 14. RAND250: the $P(|E_{\mathbf{p}}||\varphi_{\mathbf{p}})$ curves for $p = 1.5$ and $\varphi_{\mathbf{p}} = 0, \pi/2$.

Table 2. *Jamilas structure-factor moduli and phases calculated from the published atomic parameters for a selected set of indices*

p_1	p_2	p_3	$\|F_\mathbf{p}\|$	φ (°)
0.00	0.00	0.00	739.89	360.00
0.00	0.00	0.50	521.57	80.00
0.00	0.00	1.50	103.19	131.00
0.00	0.50	0.00	464.16	91.00
0.00	0.50	0.50	284.91	170.00
0.00	0.50	1.50	64.63	192.00
0.00	1.50	0.00	121.05	97.00
0.00	1.50	0.50	51.72	107.00
0.00	1.50	1.50	53.59	151.00
0.50	0.00	0.00	463.36	82.00
0.50	0.00	0.50	370.31	162.00
0.50	0.00	1.50	101.71	230.00
0.50	0.50	0.00	275.02	171.00
0.50	0.50	0.50	218.82	250.00
0.50	0.50	1.50	74.70	289.00
0.50	1.50	0.00	90.98	201.00
0.50	1.50	0.50	33.09	229.00
0.50	1.50	1.50	79.90	283.00
1.50	0.00	0.00	118.50	55.00
1.50	0.00	0.50	67.81	142.00
1.50	0.00	1.50	157.40	342.00
1.50	0.50	0.00	70.78	122.00
1.50	0.50	0.50	17.84	212.00
1.50	0.50	1.50	126.30	91.00
1.50	1.50	0.00	41.57	354.00
1.50	1.50	0.50	69.54	69.00
1.50	1.50	1.50	84.90	95.00

$\langle|F_\mathbf{p}|^2\rangle/\sum_2$: the one-dimensional case suggests that the phases of such reflections are predictable. On the contrary, the structure factors with indices $(1/2, 1, 0)$, $(2, 1/2, 1/2)$ or $(1, 3, 1/2)$ will satisfy the Wilson's distributions: no phase value may be predicted for them. In order to have more insight into the phase predictability, let us apply the relation

$$\langle\varphi_\mathbf{p}\rangle = \tan^{-1}(m_{01}/m_{10}) = \tan^{-1}(s_\mathbf{p}/c_\mathbf{p}) \qquad (21)$$

to reflections for which all the p_1, p_2, p_3 are half-integers or zero. From the definitions of $c_\mathbf{p}$ and $s_\mathbf{p}$ given in §12, it is easily derived that

$$\langle\varphi_{1/2,0,0}\rangle = \langle\varphi_{0,1/2,0}\rangle = \langle\varphi_{0,0,1/2}\rangle = \pi/2$$
$$\langle\varphi_{1/2,1/2,0}\rangle = \langle\varphi_{0,1/2,1/2}\rangle = \langle\varphi_{1/2,0,1/2}\rangle = \pi$$
$$\langle\varphi_{1/2,1/2,1/2}\rangle = 3\pi/2.$$

It is worthwhile noting that, as an effect of the three-dimensionality, the intensity oscillations die down quickly with increasing values of p_1, p_2, p_3. Accordingly, the phase distributions soon also become flat.

The correctness of our predictions may be checked *via* Table 2, where we show, for Jamilas (Dobson *et al.*, 1990), a $P1$ crystal structure with chemical formula $K_4C_{64}H_{28}N_8O_{28}S_4$, the structure-factor moduli and phases calculated from the published atomic parameters. They agree with the values $\langle\varphi_\mathbf{p}\rangle$ provided by (21).

13. On the assumption about the atomic positions

We have assumed throughout this paper that the atomic coordinates x_j, y_j, z_j lie in the interval $(0, 1)$. What is the effect of a different assumption on the probability distribution of the structure factors with non-integral indices? Let us define

$$F'_\mathbf{p} = \sum_{j=1}^{N} f_j \exp[2\pi i(p_1x'_j + p_2y'_j + p_3z'_j)] = |F'_\mathbf{p}| \exp(i\varphi'_\mathbf{p})$$

as the structure factor when the shift of the origin

$$\mathbf{T}_0 = X_0\mathbf{a} + Y_0\mathbf{b} + Z_0\mathbf{c}$$

has been applied. Then the new coordinates will satisfy the conditions

$$-X_0 < x'_j < 1 - X_0$$
$$-Y_0 < y'_j < 1 - Y_0$$
$$-Z_0 < z'_j < 1 - Z_0.$$

In accordance with the known relationship

$$F'_\mathbf{p} = \exp(-2\pi i\mathbf{p}\cdot\mathbf{T}_0)F_\mathbf{p},$$

we will have

$$\varphi'_\mathbf{p} = \varphi_\mathbf{p} - 2\pi\mathbf{p}\cdot\mathbf{T}_0, \qquad |F'_\mathbf{p}| = |F_\mathbf{p}|.$$

The application of our statistical approach to the above case will lead to a shifted (along the φ axis) phase distribution $P(\varphi_\mathbf{p})$ and to the same distribution $P(|F_\mathbf{p}|)$.

14. Conclusions

We started a new theme in the area of structure-factor statistics: the derivation of the distribution of the structure factors with non-integral indices. Our main results may be summarized as:

(*a*) The distribution $P(|F_\mathbf{p}|)$ may be quite different from Wilson's distribution $P(|F_\mathbf{h}|)$. The differences increase when $\mathbf{p}$ approaches vectors with half-integral indices, decrease when the integral part of the index components increase and/or when $\mathbf{p}$ approaches some reciprocal vector with integer components.

(*b*) The phase $\varphi_\mathbf{p}$ may be predicted with good reliability when $|\mathbf{p}|$ is not large and its components are close to half-integers.

(*c*) $P(|E_\mathbf{p}|)$ and $P(\varphi_\mathbf{p})$ are not universal, as in Wilson statistics, but depend on the structural complexity.

The first question to answer is the following: can $P(\varphi)$ be directly applied to the solution of the phase problem? Our answer is negative: indeed, each phase estimate available through (17) is independent of the structural features and only depends on the structural complexity. However, no phase relationship involving the structure factors with non-integral indices can prescind the

Table 3. *RAND250: values of the parameters v_0, v_1, v_2, X_1 and X_2 at selected values of p in the interval (0.01, 1)*

p	v_0	v_1	v_2	X_1	X_2
0.01	5773512954	93667128	2886950144	11547033600	181249264
0.03	71170419	3464737	35606716	142341680	6664132
0.05	9196025	746492	4605739	18392354	1418570
0.08	1392895	181119	699451	2785909	334126
0.1	566933	91113	285380	1133943	165536
0.3	5967	2983	3174	11942	2664
0.5	535	468	320	1073	42
0.7	66	88	88	134	19
1.0	0	2	1	0	0

information provided by (17). A first example is constituted by the conditional probability distributions $P(\varphi_{\mathbf{p}}||F_{\mathbf{p}}|)$ and $P(|F_{\mathbf{p}}||\varphi_{\mathbf{p}})$, which may lead to estimates that may confirm or modify the estimates provided by $P(\varphi_{\mathbf{p}})$ and $P(|F_{\mathbf{p}}|)$ according to the available prior information. In particular, we should look at the conditional distributions as simple ways of using the structural information: *vice versa*, they should be potentially useful for solving the phase problem.

In the following papers of this series, we will first describe the structure-factor statistic for the $P\bar{1}$ case and then new and more useful phase relationships that can be directly applied for the solution of the phase problem.

APPENDIX *A*
Calculation of the cumulants for the acentric one-dimensional case

We will suppose that the variables x_j, $j = 1, \ldots, N$, are independently and uniformly distributed in the interval (0, 1). Then

$$\begin{aligned}
\langle A_p\rangle &= \left\langle \sum_{j=1}^{N} f_j \cos(2\pi p x_j)\right\rangle \\
&= \textstyle\sum_1[\sin(2\pi p x)/2\pi p]_0^1 = \sum_1 c_p \\
\langle B_p\rangle &= \textstyle\sum_1[\cos(2\pi p x)/2\pi p]_1^0 \\
&= \textstyle\sum_1\{[1-\cos(2\pi p)]/2\pi p\} = \sum_1 s_p \\
\langle A_p^2\rangle &= \left\langle \left[\sum_{j=1}^{N} f_j \cos(2\pi p x_j)\right]^2\right\rangle \\
&= \sum_{j_1,j_2=1}^{N} f_{j_1} f_{j_2} \langle \cos(2\pi p x_{j_1}) \cos(2\pi p x_{j_2})\rangle \\
&= \textstyle\sum_2\langle \cos^2 2\pi p x\rangle \\
&\quad + \sum_{j_1\neq j_2=1}^{N} f_{j_1} f_{j_2} \langle \cos(2\pi p x_{j_1}) \cos(2\pi p x_{j_2})\rangle.
\end{aligned} \tag{22}$$

Since x_{j_1} and x_{j_2} have been assumed to be statistically independent of each other, the average of the product of the two cosines at the right-hand side of (22) will be equal to the product of the average. Therefore,

$$\langle A_p^2\rangle = 0.5\textstyle\sum_2(1+c_{2p}) + \left(\sum_{j_1\neq j_2=1}^{N} f_{j_1} f_{j_2}\right) c_p^2. \tag{23}$$

Since

$$\textstyle\sum_1^2 = \sum_2 + \sum_{j_1\neq j_2}^{N} f_{j_1} f_{j_2},$$

(23) reduces to

$$\begin{aligned}
\langle A_p^2\rangle &= 0.5\textstyle\sum_2(1+c_{2p}) + \left(\sum_1^2 - \sum_2\right)c_p^2 \\
&= 0.5\textstyle\sum_2(1+c_{2p}-2c_p^2) + \sum_1^2 c_p^2.
\end{aligned}$$

Accordingly,

$$K_{20} = \langle A_p^2\rangle - \langle A_p\rangle^2 = 0.5\textstyle\sum_2(1+c_{2p}-2c_p^2). \tag{24}$$

In an equivalent way, the relations

$$\begin{aligned}
\langle B_p^2\rangle &= \tfrac{1}{2}\textstyle\sum_2(1-c_{2p}-2s_p^2) + \sum_1^2 s_p^2, \\
K_{02} &= \langle B_p^2\rangle - \langle B_p\rangle^2 = 0.5\textstyle\sum_2(1-c_{2p}-2s_p^2)
\end{aligned}$$

may be derived. The last cumulant to calculate is

$$\begin{aligned}
K_{11} &= \langle A_p B_p\rangle - \langle A_p\rangle\langle B_p\rangle \\
&= \sum_{j_1,j_2=1}^{N} f_{j_1} f_{j_2} \langle \cos(2\pi p x_{j_1}) \sin(2\pi p x_{j_2})\rangle \\
&\quad - \textstyle\sum_1^2[\sin(2\pi p)/2\pi p]\{[1-\cos(2\pi p)]/2\pi p\} \\
&= \tfrac{1}{2}\left(\sum_{j=1}^{N} f_j^2\right)\langle \sin(4\pi p x)\rangle \\
&\quad + \sum_{j_1\neq j_2=1}^{N} f_{j_1} f_{j_2} \langle \cos(2\pi p x_{j_1}) \sin(2\pi p x_{j_2})\rangle \\
&\quad - \textstyle\sum_1^2[\sin(2\pi p)/2\pi p]\{[1-\cos(2\pi p)]/2\pi p\} \\
&= \tfrac{1}{2}\textstyle\sum_2(s_{2p} - 2c_p s_p).
\end{aligned}$$

It may be worthwhile noting that the second-order moments all depend on both $\sum_1$ and $\sum_2$ but, as in the Wilson statistics, the second-order cumulants depend only on $\sum_2$.

APPENDIX *B*
About the compatibility of $P(|E_p|)$ and $P(|\varphi_p|)$

We give in Table 3, for RAND250, the values of the parameters v_0, v_1, v_2, X_1, X_2 at selected values of p in the interval $(0.01, 1)$. It is easily seen that (11) is not immediately computable for p close to zero (say $p < 0.9$) since those values are too large arguments for exponential and Bessel functions.

The computability of $P(|E_p|)$ may be improved by expanding I_n in an asymptotic series (to be used for very large arguments):

$$I_n(x) = (2\pi x)^{-1/2}(\exp x)S_n(x), \qquad x > 0,$$

where

$$S_n(x) = [1 - (\mu - 1)/8x + (\mu - 1)(\mu - 9)/2!(8x)^2 - (\mu - 1)(\mu - 9)(\mu - 25)/3!(8x)^3 + \ldots]$$

and $\mu = 4n^2$.

Since the behaviour of $P(|E_p|)$ in the interval $(0, 1)$ may be inferred from other considerations (see §9), we decided not to explicitly calculate it in such an interval. Analogously, $P(\varphi_p)$ depends [see (13)] on the parameters μ and ν, the values of which are also fixed by the v_i's. The consequence is that $P(\varphi_p)$ is hardly computable for $p < 0.7$. Since its behaviour may be predicted (see §9), we decided not to calculate it in such an interval.

One of the authors (DS) undertook this work with the support of the 'ICTP Programme for Training and Research in Italian Laboratories, Trieste, Italy'.

References

Boyes-Watson, J., Davidson, E. & Perutz, M. F. (1947). *Proc. R. Soc. London Ser. A*, **191**, 83–132.
Dobson, C., Fattah, J., Prout, C. K., Twyman, J. M. & Watkin, D. J. (1990). Personal communication.
Gradshteyn, I. S. & Ryzhik, I. M. (1965). *Table of Integrals, Series, and Products.* New York: Academic Press.
Mishnev, A. F. (1996). *Acta Cryst.* A**52**, 629–633.
Ramachandran, G. N. (1969). *Mater. Res. Bull.* **4**, 525–534.
Rossmann, M. G. & Blow, D. M. (1962). *Acta Cryst.* **15**, 39–45.
Sayre, D. (1952). *Acta Cryst.* **5**, 843.
Shmueli, U. & Weiss, G. H. (1995). *Introduction to Crystallographic Statistics.* Oxford University Press.
Weyl, H. (1916). *Math. Ann.* **77**, 313.
Wilson, A. J. C. (1942). *Nature (London)*, **150**, 151–152.
Zanotti, G., Fogale, F. & Capitani, G. (1996). *Acta Cryst.* A**52**, 757–765.