

2.1. STATISTICAL PROPERTIES OF THE WEIGHTED RECIPROCAL LATTICE

Table 2.1.5.1. *Some properties of gamma and beta distributions*

If  $x_1, x_2, \dots, x_n$  are independent gamma-distributed variables with parameters  $p_1, p_2, \dots, p_n$ , their sum is a gamma-distributed variable with  $p = p_1 + p_2 + \dots + p_n$ .

If  $x$  and  $y$  are independent gamma-distributed variables with parameters  $p$  and  $q$ , then the ratio  $u = x/y$  has the distribution  $\beta_2(u; p, q)$ .

With the same notation, the ratio  $v = x/(x + y)$  has the distribution  $\beta_1(v; p, q)$ .

Differences and products of gamma-distributed variables do not lead to simple results. For proofs, details and references see Kendall & Stuart (1977).

Name of the distribution, its functional form, mean and variance
<p>Gamma distribution with parameter <math>p</math>:</p> $\gamma_p(x) = [\Gamma(x)]^{-1} x^{p-1} \exp(-x); \quad p \leq x < \infty, \quad p > 0$ <p>mean: <math>\langle x \rangle = p</math>; variance: <math>\langle (x - \langle x \rangle)^2 \rangle = p</math>.</p>
<p>Beta distribution of first kind with parameters <math>p</math> and <math>q</math>:</p> $\beta_1(x; p, q) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} x^{p-1} (1-x)^{q-1}; \quad 0 \leq x \leq \infty, \quad p, q > 0$ <p>mean: <math>\langle x \rangle = p/(p+q)</math>;</p> <p>variance: <math>\langle (x - \langle x \rangle)^2 \rangle = pq/[(p+q)^2(p+q+1)]</math>.</p>
<p>Beta distribution of second kind with parameters <math>p</math> and <math>q</math>:</p> $\beta_2(x; p, q) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} x^{p-1} (1+x)^{-p-q}; \quad 0 \leq x < \infty, \quad p, q > 0$ <p>mean: <math>\langle x \rangle = p/(q-1)</math>;</p> <p>variance: <math>\langle (x - \langle x \rangle)^2 \rangle = p(p+q-1)/[(q-1)(q-2)]</math>.</p>

fact  $-\infty$ ; for the distribution of  $|F|$ ,  $|E|$ ,  $I$  and  $I/\Sigma$  the lower end of the range is zero. In such cases, equation (2.1.5.21) becomes

$$F(x) = \int_0^x f(x) dx. \quad (2.1.5.22)$$

In crystallographic applications the cumulative distribution is usually denoted by  $N(x)$ , rather than by the capital letter corresponding to the probability density function designation. The cumulative forms of the ideal acentric and centric distributions (Howells *et al.*, 1950) have found many applications. For the acentric distribution of  $|E|$  [equation (2.1.5.8)] the integration is readily carried out:

$$N(|E|) = 2 \int_0^{|E|} y \exp(-y^2) dy = 1 - \exp(-|E|^2). \quad (2.1.5.23)$$

The integral for the centric distribution of  $|E|$  [equation (2.1.5.11)] cannot be expressed in terms of elementary functions, but the integral required has so many important applications in statistics that it has been given a special name and symbol, the error function  $\text{erf}(x)$ , defined by

$$\text{erf}(x) = (2/\pi^{1/2}) \int_0^x \exp(-t^2) dt. \quad (2.1.5.24)$$

For the centric distribution, then

$$N(|E|) = (2/\pi)^{1/2} \int_0^{|E|} y \exp(-y^2/2) dy \quad (2.1.5.25)$$

$$= \text{erf}(|E|/2^{1/2}). \quad (2.1.5.26)$$

The error function is extensively tabulated [see *e.g.* Abramowitz & Stegun (1972), pp. 310–311, and a closely related function on pp. 966–973].

2.1.6. Distributions of sums, averages and ratios

2.1.6.1. Distributions of sums and averages

In Section 2.1.2.1, it was shown that the average intensity of a sufficient number of reflections is  $\Sigma$  [equation (2.1.2.4)]. When the number of reflections is not ‘sufficient’, their mean value will show statistical fluctuations about  $\Sigma$ ; such statistical fluctuations are in addition to any systematic variation resulting from non-independence of atomic positions, as discussed in Sections 2.1.2.1–2.1.2.3. We thus need to consider the probability density functions of sums like

$$J_n = \sum_{i=1}^n G_i, \quad (2.1.6.1)$$

and averages like

$$Y = J_n/n, \quad (2.1.6.2)$$

where  $G_i$  is the intensity of the  $i$ th reflection. The probability density distributions are easily obtained from a property of gamma distributions: If  $x_1, x_2, \dots, x_n$  are independent gamma-distributed variables with parameters  $p_1, p_2, \dots, p_n$ , their sum is a gamma-distributed variable with parameter  $p$  equal to the sum of the parameters. The sum of  $n$  intensities drawn from an acentric distribution thus has the distribution

$$p(J_n) dJ_n = \gamma_n(J_n/\Sigma) d(J_n/\Sigma); \quad (2.1.6.3)$$

the parameters of the variables added are all equal to unity, so that their sum is  $p$ . Similarly, the sum of  $n$  intensities drawn from a centric distribution has the distribution

$$p(J_n) dJ_n = \gamma_{n/2}[J_n/(2\Sigma)] d[J_n/(2\Sigma)]; \quad (2.1.6.4)$$

each parameter has the value of one-half. The corresponding distributions of the averages of  $n$  intensities are then

$$p(Y) dY = \gamma_n(nY/\Sigma) d(nY/\Sigma) \quad (2.1.6.5)$$

for the acentric case, and

$$p(Y) dY = \gamma_{n/2}[nY/(2\Sigma)] d[nY/(2\Sigma)] \quad (2.1.6.6)$$

for the centric. In both cases the expected value of  $Y$  is  $\Sigma$  and the variances are  $\Sigma^2/n$  and  $2\Sigma^2/n$ , respectively, just as would be expected.

2.1.6.2. Distribution of ratios

Ratios like

$$S_{n,m} = J_n/K_m, \quad (2.1.6.7)$$

where  $J_n$  is given by equation (2.1.6.1),

$$K_m = \sum_{j=1}^m H_j, \quad (2.1.6.8)$$

and the  $H_j$ 's are the intensities of a set of reflections (which may or may not overlap with those included in  $J_n$ ), are used in correlating intensities measured under different conditions. They arise in

## 2. RECIPROCAL SPACE IN CRYSTAL-STRUCTURE DETERMINATION

correlating reflections on different layer lines from the same or different specimens, in correlating the same reflections from different crystals, in normalizing intensities to the local average or to  $\Sigma$ , and in certain systematic trial-and-error methods of structure determination (see Rabinovich & Shakked, 1984, and references therein). There are three main cases:

(i)  $G_i$  and  $H_i$  refer to the *same* reflection; for example, they might be the observed and calculated quantities for the  $hkl$  reflection measured under different conditions or for different crystals of the same substance; or

(ii)  $G_i$  and  $H_i$  are *unrelated*; for example, the observed and calculated values for the  $hkl$  reflection for a completely wrong trial structure, or values for entirely different reflections, as in reducing photographic measurements on different layer lines to the same scale; or

(iii) the  $G_i$ 's are a subset of the  $H_i$ 's, so that  $G_i = H_i$  for  $i < n$  and  $m > n$ .

Aside from the scale factor, in case (i)  $G_i$  and  $H_i$  will differ chiefly through relatively small statistical fluctuations and uncorrected systematic errors, whereas in case (ii) the differences will be relatively large because of the inherent differences in the intensities. Here we are concerned only with cases (ii) and (iii); the practical problems of case (i) are postponed to *IT C* (1999).

There is little in the crystallographic literature concerning the probability distribution of sums like (2.1.6.1) or ratios like (2.1.6.7); certain results are reviewed by Srinivasan & Parthasarathy (1976, ch. 5), but with a bias toward partially related structures that makes it difficult to apply them to the immediate problem.

In case (ii) ( $G_i$  and  $H_i$  independent), acentric distribution, Table 2.1.5.1 gives the distribution of the ratio

$$u = nY/(mZ) \quad (2.1.6.9)$$

$$p(u) du = \beta_2[nY/(mZ); n, m] d[nY/(mZ)], \quad (2.1.6.10)$$

where  $\beta_2$  is a beta distribution of the second kind,  $Y$  is given by equation (2.1.6.2) and  $Z$  by

$$Z = K_m/m, \quad (2.1.6.11)$$

where  $n$  is the number of intensities included in the numerator and  $m$  is the number in the denominator. The expected value of  $Y/Z$  is then

$$\langle Y/Z \rangle = \frac{m}{m-1} = 1 + \frac{1}{m} + \dots \quad (2.1.6.12)$$

with variance

$$\sigma^2 = \frac{(n+m-1)m^2}{(m-1)^2(m-2)n}. \quad (2.1.6.13)$$

One sees that  $Y/Z$  is a biased estimate of the scaling factor between two sets of intensities and the bias, of the order of  $m^{-1}$ , depends only on the number of intensities averaged in the denominator. This may seem odd at first sight, but it becomes plausible when one remembers that the mean of a quantity is an unbiased estimator of itself, but the reciprocal of a mean is not an unbiased estimator of the mean of a reciprocal. The mean exists only if  $m > 1$  and the variance only for  $m > 2$ .

In the centric case, the expression for the distribution of the ratio of the two means  $Y$  and  $Z$  becomes

$$p(u) du = \beta_2[nY/(mZ); n/2, m/2] d[nY/(mZ)] \quad (2.1.6.14)$$

with the expected value of  $Y/Z$  equal to

$$\langle Y/Z \rangle = \frac{m}{m-2} = 1 + \frac{2}{m} + \dots \quad (2.1.6.15)$$

and with its variance equal to

$$\sigma^2 = \frac{2(n+m-2)m^2}{(m-2)^2(m-4)n}. \quad (2.1.6.16)$$

For the same number of reflections, the bias in  $\langle Y/Z \rangle$  and the variance for the centric distribution are considerably larger than for the acentric. For both distributions the variance of the scaling factor approaches zero when  $n$  and  $m$  become large. The variances are large for  $m$  small, in fact 'infinite' if the number of terms averaged in the denominator is sufficiently small. These biases are readily removed by multiplying  $Y/Z$  by  $(m-1)/m$  or  $(m-2)/m$ . Many methods of estimating scaling factors – perhaps most – also introduce bias (Wilson, 1975; Lomer & Wilson, 1975; Wilson, 1976, 1978c) that is not so easily removed. Wilson (1986a) has given reasons for supposing that the bias of the ratio (2.1.6.7) approximates to

$$1 + \frac{\sigma^2(I)}{m\langle I \rangle^2}, \quad (2.1.6.17)$$

whatever the intensity distribution. Equations (2.1.6.12) and (2.1.6.15) are consistent with this.

### 2.1.6.3. Intensities scaled to the local average

When the  $G_i$ 's are a subset of the  $H_i$ 's, the beta distributions of the second kind are replaced by beta distributions of the first kind, with means and variances readily found from Table 2.1.5.1. The distribution of such a ratio is chiefly of interest when  $Y$  relates to a single reflection and  $Z$  relates to a group of  $m$  intensities including  $Y$ . This corresponds to normalizing intensities to the local average. Its distribution is

$$p(I/\langle I \rangle) d(I/\langle I \rangle) = \beta_1(I/n\langle I \rangle; 1, n-1) d(I/n\langle I \rangle) \quad (2.1.6.18)$$

in the acentric case, with an expected value of  $I/\langle I \rangle$  of unity; there is no bias, as is obvious *a priori*. The variance of  $I/\langle I \rangle$  is

$$\sigma^2 = \frac{n-1}{n+1}, \quad (2.1.6.19)$$

which is less than the variance of the intensities normalized to an 'infinite' population by a fraction of the order of  $2/n$ . Unlike the variance of the scaling factor, the variance of the normalized intensity approaches unity as  $n$  becomes large. For intensities having a centric distribution, the distribution normalized to the local average is given by

$$p(I/\langle I \rangle) d(I/\langle I \rangle) = \beta_1[I/n\langle I \rangle; 1/2, (n-1)/2] d(I/n\langle I \rangle), \quad (2.1.6.20)$$

with an expected value of  $I/\langle I \rangle$  of unity and with variance

$$\sigma^2 = \frac{2(n-1)}{n+2}, \quad (2.1.6.21)$$

less than that for an 'infinite' population by a fraction of about  $3/n$ .

Similar considerations apply to intensities normalized to  $\Sigma$  in the usual way, since they are equal to those normalized to  $\langle I \rangle$  multiplied by  $\langle I \rangle/\Sigma$ .

### 2.1.6.4. The use of normal approximations

Since  $J_n$  and  $K_m$  [equations (2.1.6.1) and (2.1.6.8)] are sums of identically distributed variables conforming to the conditions of the central-limit theorem, it is tempting to approximate their distributions by normal distributions with the correct mean and variance. This would be reasonably satisfactory for the distributions of  $J_n$  and  $K_m$  themselves for quite small values of  $n$  and  $m$ , but unsatisfactory for the distribution of their ratio for any values of  $n$  and  $m$ , even large. The ratio of two variables with normal distributions is

notorious for its rather indeterminate mean and 'infinite' variance, resulting from the 'tail' of the denominator distributions extending through zero to negative values. The leading terms of the ratio distribution are given by Kendall & Stuart (1977, p. 288).

### 2.1.7. Non-ideal distributions: the correction-factor approach

#### 2.1.7.1. Introduction

The probability density functions (p.d.f.'s) of the magnitude of the structure factor, presented in Section 2.1.5, are based on the central-limit theorem discussed above. In particular, the centric and acentric p.d.f.'s given by equations (2.1.5.11) and (2.1.5.8), respectively, are expected to account for the statistical properties of diffraction patterns obtained from crystals consisting of nearly equal atoms, which obey the fundamental assumptions of uniformity and independence of the atomic contributions and are not affected by noncrystallographic symmetry and dispersion. It is also assumed there that the number of atoms in the asymmetric unit is large. Distributions of structure-factor magnitudes which are based on the central-limit theorem, and thus obey the above assumptions, have been termed 'ideal', and the subjects of the following sections are those distributions for which some of the above assumptions/restrictions are not fulfilled; the latter distributions will be called 'non-ideal'.

We recall that the assumption of uniformity consists of the requirement that the fractional part of the scalar product  $hx + ky + lz$  be uniformly distributed over the  $[0, 1]$  interval, which holds well if  $x, y, z$  are rationally independent (Hauptman & Karle, 1953), and permits one to regard the atomic contribution to the structure factor as a random variable. This is of course a necessary requirement for any statistical treatment. If, however, the atomic composition of the asymmetric unit is widely heterogeneous, the structure factor is then a sum of unequally distributed random variables and the Lindeberg-Lévy version of the central-limit theorem (cf. Section 2.1.4.4) cannot be expected to apply. Other versions of this theorem might still predict a normal p.d.f. of the sum, but at the expense of a correspondingly large number of terms/atoms. It is well known that atomic heterogeneity gives rise to severe deviations from ideal behaviour (e.g. Howells *et al.*, 1950) and one of the aims of crystallographic statistics has been the introduction of a correct dependence on the atomic composition into the non-ideal p.d.f.'s [for a review of the early work on non-ideal distributions see Srinivasan & Parthasarathy (1976)]. A somewhat less well known fact is that the dependence of the p.d.f.'s of  $|E|$  on space-group symmetry becomes more conspicuous as the composition becomes more heterogeneous (e.g. Shmueli, 1979; Shmueli & Wilson, 1981). Hence both the composition and the symmetry dependence of the intensity statistics are of interest. Other problems, which likewise give rise to non-ideal p.d.f.'s, are the presence of heavy atoms in (variable) special positions, heterogeneous structures with complete or partial noncrystallographic symmetry, and the presence of outstandingly heavy dispersive scatterers.

The need for theoretical representations of non-ideal p.d.f.'s is exemplified in Fig. 2.1.7.1(a), which shows the ideal centric and acentric p.d.f.'s together with a frequency histogram of  $|E|$  values, recalculated for a centrosymmetric structure containing a platinum atom in the asymmetric unit of  $P\bar{1}$  (Faggiani *et al.*, 1980). Clearly, the deviation from the Gaussian p.d.f., predicted by the central-limit theorem, is here very large and a comparison with the possible ideal distributions can (in this case) lead to wrong conclusions.

Two general approaches have so far been employed in derivations of non-ideal p.d.f.'s which account for the above-mentioned problems: the correction-factor approach, to be dealt

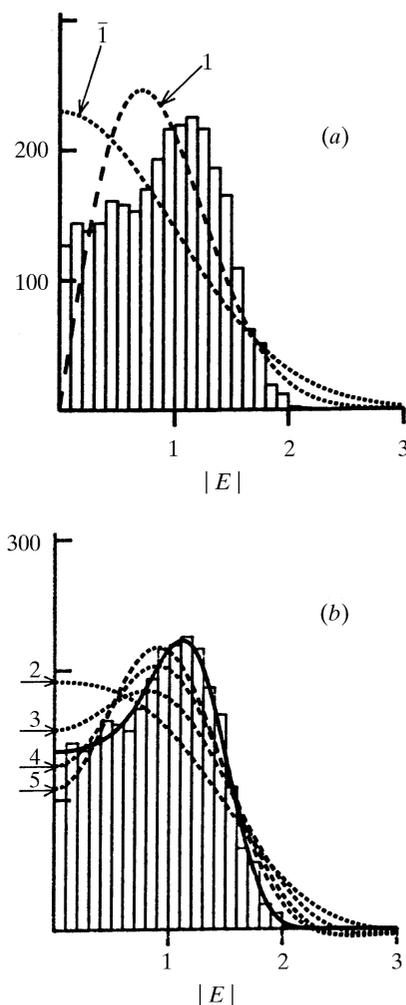


Fig. 2.1.7.1. Atomic heterogeneity and intensity statistics. The histogram appearing in (a) and (b) was constructed from  $|E|$  values which were recalculated from atomic parameters published for the centrosymmetric structure of  $C_6H_{18}Cl_2N_4O_4Pt$  (Faggiani *et al.*, 1980). The space group of the crystal is  $P\bar{1}$ ,  $Z = 2$ , i.e. all the atoms are located in general positions. (a) A comparison of the recalculated distribution of  $|E|$  with the ideal centric [equation (2.1.5.11)] and acentric [equation (2.1.5.8)] p.d.f.'s, denoted by  $\bar{1}$  and  $1$ , respectively. (b) The same recalculated histogram along with the centric correction-factor p.d.f. [equation (2.1.7.5)], truncated after two, three, four and five terms (dashed lines), and with that accurately computed for the correct space-group Fourier p.d.f. [equations (2.1.8.5) and (2.1.8.22)] (solid line).

with in the following sections, and the more recently introduced Fourier method, to which Section 2.1.8 is dedicated. In what follows, we introduce briefly the mathematical background of the correction-factor approach, apply this formalism to centric and acentric non-ideal p.d.f.'s, and present the numerical values of the moments of the trigonometric structure factor which permit an approximate evaluation of such p.d.f.'s for all the three-dimensional space groups.

#### 2.1.7.2. Mathematical background

Suppose that  $p(x)$  is a p.d.f. which accurately describes the experimental distribution of the random variable  $x$ , where  $x$  is related to a sum of random variables and can be assumed to obey (to some approximation) an ideal p.d.f., say  $p^{(0)}(x)$ , based on the central-limit theorem. In the correction-factor approach we seek to represent  $p(x)$  as