Table 2.1.3.2. *Intensity-distribution effects of symmetry elements not causing systematic absences*

Abbreviations and orientation of axes: A = acentric distribution, C = centric distribution, S = distribution parameter, $\langle I \rangle$ = average intensity. Axes are parallel to **c**, planes are perpendicular to **c**.

Element	Reflections	Distribution	$S/\Sigma = \langle I \rangle/\Sigma$
1	All	A	1
$\bar{1}$	All	C	1
2	hkl	A	1
	$hk0$	C	1
	$00l$	A	2
$\bar{2} = m$	hkl	A	1
	$hk0$	A	2
	$00l$	C	1
3	hkl	A	1
	$hk0$	A	1
	$00l$	A	3
$\bar{3}$	hkl	C	1
	$hk0$	C	1
	$00l$	C	3
4	hkl	A	1
	$hk0$	C	1
	$00l$	A	4
$\bar{4}$	hkl	A	1
	$hk0$	C	1
	$00l$	C	2
6	hkl	A	1
	$hk0$	C	1
	$00l$	A	6
$\bar{6} = 3/m$	hkl	A	1
	$hk0$	A	2
	$00l$	C	3

mechanism for compensation for the reflections with enhanced intensity is obvious.

2.1.3.2. *Symmetry elements not producing systematic absences*

Certain symmetry elements not producing absences (mirror planes and rotation axes) cause equivalent atoms to coincide in a plane or a line projection and hence produce a zone or row in reciprocal space for which the average intensity is an integral multiple of the general average (Wilson, 1950); the effects of single such symmetry elements are given in Table 2.1.3.2. There is, however, no obvious mechanism for compensation for this enhancement. When reflections are few this may be an important matter in assigning an approximate absolute scale by comparing observed and calculated intensities. Wilson (1964), Nigam (1972) and Nigam & Wilson (1980), noting that in such cases the finite size of atoms results in forbidden ranges of positional parameters, have shown that there is a diminution of the intensity of layers (rows) in the immediate neighbourhood of the enhanced zones (rows), just sufficient to compensate for the enhancement. In forming general averages, therefore, reflections from enhanced zones or rows should be included at their full intensity, not divided by the multiplier; the matter is discussed in more detail by Wilson (1987*a*). It should be noted, however, that organic structures containing molecules related by rotation axes are rare, and such structures related by mirror planes are even rarer (Wilson, 1993).

2.1.3.3. *More than one symmetry element*

Further alterations of the intensities occur if two or more such symmetry elements are present in the space group. The effects were treated in detail by Rogers (1950), who used them to construct a table for the determination of space groups by supplementing the usual knowledge of Laue group with statistical information. Only two pairs of space groups, the orthorhombic $I222$ and $I2_12_12_1$, and their cubic supergroups $I23$ and $I2_13_1$, remained unresolved. Examination of this table shows that what statistical information does is to resolve the Laue group into point groups; the further resolution into space groups is equivalent to the use of Table 3.2 in *IT* A (1983). The statistical consequences of each point group, as given by Rogers, are reproduced in Table 2.1.3.3.

2.1.4. Probability density distributions – mathematical preliminaries

For the purpose of this chapter, 'ideal' probability distributions or probability density functions are the asymptotic forms obtained by the use of the central-limit theorem when the number of atoms in the unit cell, N, is sufficiently large. In order to derive them it is necessary to outline the properties of characteristic functions and to state alternative conditions for the validity of the central-limit theorem; the distributions themselves are derived in Section 2.1.5.

2.1.4.1. *Characteristic functions*

The average value of $\exp(itx)$ is very important in probability theory; it is called the characteristic function of the distribution $f(x)$ and is denoted by $C_x(t)$ or, when no confusion can arise, by $C(t)$. It exists for all legitimate distributions, whether discrete or continuous. In the continuous case it is given by

$$C(t) = \int_{-\infty}^{\infty} \exp(itx) f(x)\, \mathrm{d}x, \tag{2.1.4.1}$$

and is thus the Fourier transform of $f(x)$. In many cases it can be obtained from known integrals. For example, for the Cauchy distribution,

$$C(t) = \frac{a}{\pi} \int_{-\infty}^{\infty} \frac{\exp(itx)}{a^2 + x^2}\, \mathrm{d}x \tag{2.1.4.2}$$

$$= \exp(-a|t|), \tag{2.1.4.3}$$

and for the normal distribution,

$$C(t) = (2\pi\sigma^2)^{-1/2} \int_{-\infty}^{\infty} \exp\left(\frac{-(x-m)^2}{2\sigma^2}\right) \exp(itx)\, \mathrm{d}x \tag{2.1.4.4}$$

$$= \exp\left(imt - \frac{\sigma^2 t^2}{2}\right). \tag{2.1.4.5}$$

Since the characteristic function is the Fourier transform of the distribution function, the converse is true, and if the characteristic function is known the probability distribution function can be obtained by the use of Fourier inversion theorem,

$$f(x) = (1/2\pi) \int_{-\infty}^{\infty} \exp(-itx) C(t)\, \mathrm{d}t. \tag{2.1.4.6}$$

Table 2.1.3.3. *Average multiples for the 32 point groups (modified from Rogers, 1950).*

The multiple gives S/Σ for the row and zone corresponding to the principal axis of the point-group symbol; those for the secondary and tertiary axes are given when the symbol contains such axes.

	Principal		Secondary		Tertiary	
Point group	Row	Zone	Row	Zone	Row	Zone
1	1	1				
$\bar{1}$	1	1				
2	2	1				
m	1	2				
$2/m$	2	2				
222	2	1	2	1	2	1
$mm2$	2	2	2	2	4	1
mmm	4	2	4	2	4	2
4	4	1				
$\bar{4}$	2	1				
$4/m$	4	2				
422	4	1*	2	1	2	1
$4mm$	8	1	2	2	2	2
$\bar{4}2m$	4	1	2	1	2	2
$4/mmm$	8	2	4	2	4	2
3	3	1				
$\bar{3}$	3	1				
321	3	1	2	1	1	1
$3m1$	6	1	1	2	2	
$31m$	6	1	2	2	2	1
6	6	1				
$\bar{6}$	3	2				
$6/m$	6	2				
622	6	1	2	1	2	1
$6mm$	12	1	2	2	2	2
$\bar{6}m2$	6	2	2	2	4	1
$6/mmm$	12	2	4	2	4	2
231	2	1	3	1	1	1
$m\bar{3}1$	4	2	3	1	1	1
432	4	1	3	1	2	1
$\bar{4}3m$	4	1	6	1	2	2
$m\bar{3}m$	8	2	6	2	4	2

Note. The pairs of point groups, 1 and $\bar{1}$ and 3 and $\bar{3}$, not distinguished by average multiples, may be distinguished by their centric and acentric probability density functions.

* The entry for the principal zone for the point group 422 was given incorrectly as 2 in the first edition of this volume.

An alternative approach to the derivation of the distribution from a known characteristic function will be discussed below.

The most important property of characteristic functions in crystallography is the following: if x and y are independent random variables with characteristic functions $C_x(t)$ and $C_y(t)$, the characteristic function of their sum

$$z = x + y \tag{2.1.4.7}$$

is the product

$$C_z(t) = C_x(t)C_y(t). \tag{2.1.4.8}$$

Obviously this can be extended to any number of independent random variables.

When the moments exist, the characteristic function can be expanded in a power series in which the kth term is $m_k(it)^k/k!$. If the power series

$$\exp(itx) = 1 + itx + \frac{(it)^2x^2}{2!} + \frac{(it)^3x^3}{3!} + \ldots \tag{2.1.4.9}$$

is substituted in equation (2.1.4.1), one obtains

$$C(t) = 1 + itm_1' + \frac{(it)^2m_2'}{2!} + \frac{(it)^3m_3'}{3!} + \ldots. \tag{2.1.4.10}$$

The moments are written with primes in order to indicate that equation (2.1.4.10) is valid for moments about an arbitrary origin as well as for moments about the mean. If the random variable is transformed by a change of origin and scale, say

$$y = \frac{x - a}{b}, \tag{2.1.4.11}$$

the characteristic function for y becomes

$$C_y(t) = b\exp(-iat/b)C_x(t). \tag{2.1.4.12}$$

2.1.4.2. *The cumulant-generating function*

A function that is often more useful than the characteristic function is its logarithm, the cumulant-generating function:

$$K(t) = \log C(t) = k_1 + \frac{k_2(it)^2}{2!} + \frac{k_3(it)^3}{3!} + \ldots, \tag{2.1.4.13}$$

where the k's are called the *cumulants* and may be regarded as being defined by the equation. They can be evaluated in terms of the moments by combining the series (2.1.4.10) for $C(t)$ with the ordinary series for the logarithm and equating the coefficients of t^r. In most cases the process as described is tedious, but it can be shortened by use of a general method [Stuart & Ord (1994), Section 3.14, pp. 87–88; Exercise 3.19, p. 119]. Obviously, the cumulants exist only if the moments exist. The first few relations are

$$\begin{aligned} k_0 &= 0 \\ k_1 &= m_1' \\ k_2 &= m_2 = m_2' - (m_1')^2 \\ k_3 &= m_3 = m_3' - 3m_2'm_1' + 2(m_1')^2 \\ k_4 &= m_4 - 3(m_2)^2 \\ &= m_4' - m_3'm_1' - 3(m_2')^2 + 12m_2'(m_1)^2 - 6(m_1')^4. \end{aligned} \tag{2.1.4.14}$$

Such expressions and their converses up to k_{10} are given by Stuart & Ord (1994, pp. 88–91). Since all the cumulants except k_1 can be expressed in terms of the central moments only (*i.e.*, those unprimed), only k_1 is changed by a change of the origin. Because of this property, they are sometimes called the semi-invariants (or seminvariants) of the distribution. Since addition of random variables is equivalent to the multiplication of their characteristic functions [equation (2.1.4.8)] and multiplication of functions is equivalent to the addition of their logarithms, each cumulant of the distribution of the sum of a number of random variables is equal to the sum of the cumulants of the distribution functions of the individual variables – hence the name cumulants. Although the cumulants (except k_1) are independent of a change of origin, they are not independent of a change of scale. As for the moments, a

change of scale simply multiplies them by a power of the scale factor; if $y = x/b$

$$(k_y)_r = (k_x)_r/b^r. \tag{2.1.4.15}$$

The cumulants of the normal distribution are particularly simple. From equation (2.1.4.5), the cumulant-generating function of a normal distribution is

$$K(t) = imt - \sigma^2 t^2/2 \tag{2.1.4.16}$$

$$k_1 = m \tag{2.1.4.17}$$

$$k_2 = \sigma^2, \tag{2.1.4.18}$$

all cumulants with $r > 2$ are identically zero.

2.1.4.3. *The central-limit theorem*

A simple form of this important theorem can be stated as follows:

> If $x_1, x_2, \ldots, x_n$ are independent and identically distributed random variables, each of them having the same mean m and variance σ^2, then the sum
>
> $$S_n = \sum_{j=1}^{n} x_j \tag{2.1.4.19}$$
>
> tends to be normally distributed – independently of the distribution(s) of the individual random variables – with mean nm and variance $n\sigma^2$, provided n is sufficiently large.

In order to prove this theorem, let us define a standardized random variable corresponding to the sum S_n, *i.e.*, such that its mean is zero and its variance is unity:

$$\hat{S}_n = \frac{S_n - nm}{\sigma\sqrt{n}} = \frac{\sum_{j=1}^{n}(x_j - m)}{\sigma\sqrt{n}} \equiv \sum_{j=1}^{n} \frac{W_j}{\sqrt{n}}, \tag{2.1.4.20}$$

where $W_j = (x_j - m)/\sigma$ is a standardized single random variable. The characteristic function of $\hat{S}_n$ is therefore given by

$$C_n(\hat{S}_n, t) = \langle \exp(it\hat{S}_n) \rangle$$

$$= \left\langle \exp\left[it \sum_{j=1}^{n} \frac{W_j}{\sqrt{n}} \right] \right\rangle \tag{2.1.4.21}$$

$$= \prod_{j=1}^{n} \left\langle \exp\left[it \frac{W_j}{\sqrt{n}} \right] \right\rangle \tag{2.1.4.22}$$

$$= \left\{ \left\langle \exp\left[it \frac{W_1}{\sqrt{n}} \right] \right\rangle \right\}^n, \tag{2.1.4.23}$$

where the brackets $\langle\ \rangle$ denote the operation of averaging with respect to the appropriate probability density function (p.d.f.) [*cf.* equation (2.1.4.1)]. Equation (2.1.4.22) follows from equation (2.1.4.21) by the assumption of independence, while the assumption of identically distributed variables leads to the identity of the characteristic functions of the individual variables – as seen in equation (2.1.4.23).

On the assumption that moments of all the orders exist – a most plausible assumption in situations usually encountered in structure-factor statistics – we can now expand the characteristic function of a single variable in a power series [*cf.* equation (2.1.4.10)]:

$$\left\langle \exp\left[it \frac{W_1}{\sqrt{n}} \right] \right\rangle = \left\langle \sum_{r=0}^{\infty} \frac{(it)^r}{r!} \frac{W_1^r}{n^{r/2}} \right\rangle$$

$$= \sum_{r=0}^{\infty} \frac{(it)^r}{r!} \frac{\langle W_1^r \rangle}{n^{r/2}}$$

$$\equiv 1 - \frac{t^2}{2n} + \frac{\zeta(t,n)}{n}, \tag{2.1.4.24}$$

since $\langle W_1 \rangle = 0$, $\langle W_1^2 \rangle = 1$, and the quantity denoted by $\zeta(t,n)$ in (2.1.4.24) is given by

$$\zeta(t,n) = \sum_{r=3}^{\infty} \frac{(it)^r}{r!} \frac{\langle W_1^r \rangle}{n^{(r/2)-1}}. \tag{2.1.4.25}$$

The characteristic function of $\hat{S}_n$ is therefore

$$\langle \exp(it\hat{S}_n) \rangle = \left[1 - \frac{t^2}{2n} + \frac{\zeta(t,n)}{n} \right]^n. \tag{2.1.4.26}$$

Now, as is seen from (2.1.4.25), for every fixed t the quantity $\zeta(t,n)$ tends to zero as n tends to infinity. The cumulant-generating function of the standardized sum then becomes

$$\log C_n(\hat{S}_n, t) = n \log\left[1 - \frac{1}{n}\left(\frac{t^2}{2} - \zeta(t,n) \right) \right] \tag{2.1.4.27}$$

and the logarithm on the right-hand side of equation (2.1.4.27) has the form $\log(1 - z)$ with $|z| \to 0$ as $n \to \infty$. We may therefore use the expansion

$$\log(1 - z) = -\left(z + \frac{z^2}{2} + \frac{z^3}{3} + \ldots \right),$$

which is valid for $|z| < 1$. We then obtain

$$\log C_n(\hat{S}_n, t) = -n\left[\frac{1}{n}\left(\frac{t^2}{2} - \zeta(t,n) \right) + \frac{1}{2n^2}\left(\frac{t^2}{2} - \zeta(t,n) \right)^2 + \frac{1}{3n^3}\left(\frac{t^2}{2} - \zeta(t,n) \right)^3 + \cdots \right]$$

$$= -\frac{t^2}{2} + \zeta(t,n) - \frac{1}{2n}\left(\frac{t^2}{2} - \zeta(t,n) \right)^2 - \frac{1}{3n^2}\left(\frac{t^2}{2} - \zeta(t,n) \right)^3 - \cdots$$

and finally, for every fixed t,

$$\lim_{n\to\infty} \log C_n(\hat{S}_n, t) = -\frac{t^2}{2}. \tag{2.1.4.28}$$

Since the logarithm is a continuous function of t, it follows directly that

$$\lim_{n\to\infty} C_n(\hat{S}_n, t) = \exp\left(-\frac{t^2}{2} \right). \tag{2.1.4.29}$$

The right-hand side of (2.1.4.29) is just the characteristic function of a standardized normal p.d.f., *i.e.*, a normal p.d.f. with zero mean and unit variance [*cf.* equation (2.1.4.5)]. The asymptotic expression for the p.d.f. of the standardized sum is therefore obtained as

$$p(\hat{S}) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\hat{S}^2}{2} \right),$$

which proves the above version of the central-limit theorem.

Surprisingly, this theorem has a very wide applicability and values of n as low as 30 are often large enough for the theorem to be useful. Situations in which the normal p.d.f. must be modified or replaced by an altogether different one are dealt with in Sections 2.1.7 and 2.1.8 of this chapter.

2.1.4.4. *Conditions of validity*

The above outline of a proof of the central-limit theorem depended on the existence of moments of all orders. The components of structure factors always possess finite moments of all orders, but the existence of moments beyond the second is not necessary for the validity of the theorem and it can be proved under much less stringent conditions. In fact, if all the random variables in equation (2.1.4.19) have the *same* distribution – as in a homoatomic structure – the only requirement is that the second moments of the distributions should exist [the Lindeberg–Lévy theorem (*e.g.* Cramér, 1951)]. If the distributions are not the same – as in a heteroatomic structure – some further condition is necessary to ensure that no individual random variable dominates the sum. The Liapounoff proof requires the existence of third absolute moments, but this is regarded as aesthetically displeasing; a theorem that ultimately involves only means and variances should require only means and variances in the proof. The Lindeberg–Cramér conditions meet this aesthetic criterion. Roughly, the conditions are that S^2, the variance of the sum, should tend to infinity and σ_j^2/S^2, where σ_j^2 is the variance of the jth random variable, should tend to zero for all j as n tends to infinity. The precise formulation is quoted by Kendall & Stuart (1977, p. 207).

2.1.4.5. *Non-independent variables*

The central-limit theorem, under certain conditions, remains valid even when the variables summed in equation (2.1.4.19) are not independent. The conditions have been investigated by Bernstein (1922, 1927); roughly they amount to requiring that the variables should not be too closely correlated. The theorem applies, in particular, when each x_r is related to a finite number, $f(n)$, of its neighbours, when the x's are said to be $f(n)$ dependent. The $f(n)$ dependence seems plausible for crystallographic applications, since the positions of atoms close together in a 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. Harker's (1953) idea of 'globs' seems equivalent to $f(n)$ dependence. Long-range stereochemical effects, as in pseudo-graphitic aromatic hydrocarbons, would presumably produce long-range correlations and make $f(n)$ dependence less plausible. If Bernstein's conditions are satisfied, the central-limit theorem would apply, but the actual value of $\langle x^2\rangle - \langle x\rangle^2$ would have to be used for the variance, instead of the sum of the variances of the random variables in (2.1.4.19). Because of the correlations the two values are no longer equal.

French & Wilson (1978) seem to have been the first to appeal explicitly to the central-limit theorem extended to non-independent variables, but many previous workers [for typical references, see Wilson (1981)] tacitly made the replacement – in the X-ray case substituting the local mean intensity for the sum of the squares of the atomic scattering factors.

2.1.5. Ideal probability density distributions

In applications of the central-limit theorem, and its extensions, to intensity statistics the x_j's of equation (2.1.4.19) have the form (atomic scattering factor of the jth atom) times (a trigonometric expression characteristic of the space group and Wyckoff position; also known as the trigonometric structure factor). These trigonometric expressions for all the space groups, and general Wyckoff positions, are given in Tables A1.4.3.1 through A1.4.3.7, and their first few even moments (fixed-index averaging) are given in Table 2.1.7.1. One cannot, of course, conclude that the magnitudes of the structure factor always have a normal distribution – even if the structure is homoatomic; one must look at each problem and see what components of the structure factor can be put in the form (2.1.4.19), deduce the m and σ^2 to be used for each, and combine the components to obtain the asymptotic (large N, not large x) expression for the problem in question. Ordinarily the components are the real and the imaginary parts of the structure factor; the structure factor is purely real only if the structure is centrosymmetric, the space-group origin is chosen at a crystallographic centre and the atoms are non-dispersive.

2.1.5.1. *Ideal acentric distributions*

The ideal acentric distributions are obtained by applying the central-limit theorem to the real and the imaginary parts of the structure factor, as given by equation (2.1.1.1). Consider first a crystal with no rotational symmetry (space group $P1$). The real part, A, of the structure factor is then given by

$$A = \sum_{j=1}^{N} f_j \cos\vartheta_j, \tag{2.1.5.1}$$

where N is the number of atoms in the unit cell and ϑ_j is the phase angle of the jth atom. The central-limit theorem then states that A tends to be normally distributed about its mean value with variance equal to its mean-square deviation from its mean. Under the assumption that the phase angles ϑ_j are uniformly distributed on the 0–2π range, the mean value of each cosine is zero, so that its variance is

$$\sigma^2 = \sum_{j=1}^{N} f_j^2 \langle\cos^2\vartheta_j\rangle. \tag{2.1.5.2}$$

Under the same assumption, the mean value of each $\cos^2\vartheta$ is one-half, so that the variance becomes

$$\sigma^2 = (1/2)\sum_{j=1}^{N} f_j^2 = (1/2)\Sigma, \tag{2.1.5.3}$$

where Σ is the sum of the squares of the atomic scattering factors [*cf.* equation (2.1.2.4)]. The asymptotic form of the distribution of A is therefore given by

$$p(A)\ \mathrm{d}A = (\pi\Sigma)^{-1/2}\exp(-A^2/\Sigma)\ \mathrm{d}A. \tag{2.1.5.4}$$

A similar calculation, with sines instead of cosines, gives an analogous distribution for the imaginary part B, so that the joint probability of the real and imaginary parts of F is

$$p(A,B)\ \mathrm{d}A\ \mathrm{d}B = (\pi\Sigma)^{-1}\exp[-(A^2+B^2)/\Sigma]\ \mathrm{d}A\ \mathrm{d}B. \tag{2.1.5.5}$$

Ordinarily, however, we are more interested in the distribution of the magnitude, $|F|$, of the structure factor than in the distribution of A and B. Using polar coordinates in equation (2.1.5.5) [$A = |F|\cos\phi$, $B = |F|\sin\phi$] and integrating over the angle ϕ gives

$$p(|F|)\ \mathrm{d}|F| = (2|F|/\Sigma)\exp(-|F|^2/\Sigma)\ \mathrm{d}|F|. \tag{2.1.5.6}$$

It is usually convenient, in structure-factor and intensity statistics, to express the results in terms of the normalized structure factor E and its magnitude $|E|$. If $|F|$ has been put on an absolute scale (see Section 2.2.4.3), we have

$$E = \frac{F}{\sqrt{\Sigma}} \quad \text{and} \quad |E| = \frac{|F|}{\sqrt{\Sigma}}, \tag{2.1.5.7}$$

references