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

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

so that

$$p(|E|)\ \mathrm{d}|E| = 2|E| \exp(-|E|^2)\ \mathrm{d}|E| \qquad (2.1.5.8)$$

is the normalized-structure-factor version of (2.1.5.6).

Distributions resulting from noncentrosymmetric crystals are known as *acentric* distributions; those arising from centrosymmetric crystals are known as *centric*. These adjectives are used to describe *distributions*, not crystal symmetry.

2.1.5.2. *Ideal centric distributions*

When a non-dispersive crystal is centrosymmetric, and the space-group origin is chosen at a crystallographic centre of symmetry, the imaginary part B of its structure amplitude is zero. In the simplest case, space group $P\bar{1}$, the contribution of the jth atom plus its centrosymmetric counterpart is $2f_j \cos\vartheta_j$. The calculation of $p(A)$ goes through as before, with allowance for the fact that there are $N/2$ pairs instead of N independent atoms, giving

$$p(A)\ \mathrm{d}A = (2\pi\Sigma)^{-1/2} \exp[-A^2/(2\Sigma)]\ \mathrm{d}A \qquad (2.1.5.9)$$

or equivalently

$$p(|F|)\ \mathrm{d}|F| = [2/(\pi\Sigma)]^{1/2} \exp[-|F|^2/(2\Sigma)]\ \mathrm{d}|F| \qquad (2.1.5.10)$$

or

$$p(|E|)\ \mathrm{d}|E| = (2/\pi)^{1/2} \exp(-|E|^2/2)\ \mathrm{d}|E|. \qquad (2.1.5.11)$$

2.1.5.3. *Effect of other symmetry elements on the ideal acentric and centric distributions*

Additional crystallographic symmetry elements do not produce any essential alterations in the ideal centric or acentric distribution; their main effect is to replace the parameter Σ by a 'distribution parameter', called S by Wilson (1950) and Rogers (1950), in certain groups of reflections. In addition, in noncentrosymmetric space groups, the distribution of certain groups of reflections becomes centric, though the general reflections remain acentric. The changes are summarized in Tables 2.1.3.1 and 2.1.3.2. The values of S are integers for lattice centring, glide planes and those screw axes that produce absences, and approximate integers for rotation axes and mirror planes; the modulations of the average intensity in reciprocal space outlined in Section 2.1.3.2 apply.

It should be noted that if intensities are normalized to the average of the group to which they belong, rather than to the general average, the distributions given in equations (2.1.5.8) and (2.1.5.11) are not affected.

2.1.5.4. *Other ideal distributions*

The distributions just derived are asymptotic, as they are limiting values for large N. They are the only ideal distributions, in this sense, when there is only strict crystallographic symmetry and no dispersion. However, other ideal (asymptotic) distributions arise when there is noncrystallographic symmetry, or if there is dispersion. The *subcentric* distribution,

$$p(|E|)\ \mathrm{d}|E| = \frac{2|E|}{(1-k^2)^{1/2}} \exp[-|E|^2/(1-k^2)] \times I_0\left(\frac{k|E|^2}{1-k^2}\right)\ \mathrm{d}|E|, \qquad (2.1.5.12)$$

where $I_0(x)$ is a modified Bessel function of the first kind and k is the ratio of the scattering from the centrosymmetric part to the total scattering, arises when a noncentrosymmetric crystal contains centrosymmetric parts or when dispersion introduces effective noncentrosymmetry into the scattering from a centrosymmetric crystal (Srinivasan & Parthasarathy, 1976, ch. III; Wilson, 1980*a*,*b*; Shmueli & Wilson, 1983). The *bicentric* distribution

$$p(|E|)\ \mathrm{d}|E| = \pi^{-3/2} \exp(-|E|^2/8) K_0(|E|^2/8)\ \mathrm{d}|E| \qquad (2.1.5.13)$$

arises, for example, when the 'asymmetric unit in a centrosymmetric crystal is a centrosymmetric molecule' (Lipson & Woolfson, 1952); $K_0(x)$ is a modified Bessel function of the second kind. There are higher hypercentric, hyperparallel and sesquicentric analogues (Wilson, 1952; Rogers & Wilson, 1953; Wilson, 1956). The ideal subcentric and bicentric distributions are expressed in terms of known functions, but the higher hypercentric and the sesquicentric distributions have so far been studied only through their moments and integral representations. Certain hypersymmetric distributions can be expressed in terms of Meijer's G functions (Wilson, 1987*b*).

2.1.5.5. *Relation to distributions of I*

When only the intrinsic probability distributions are being considered, it does not greatly matter whether the variable chosen is the intensity of reflection (I), or its positive square root, the modulus of the structure factor ($|F|$), since both are necessarily real and non-negative. In an obvious notation, the relation between the intensity distribution and the structure-factor distribution is

$$p_I(I) = (1/2) I^{-1/2} p_{|F|}(I^{1/2}) \qquad (2.1.5.14)$$

or

$$p_{|F|}(|F|) = 2|F| p_I(|F|^2). \qquad (2.1.5.15)$$

Statistical fluctuations in counting rates, however, introduce a small but finite probability of negative observed intensities (Wilson, 1978*a*, 1980*a*) and thus of imaginary structure factors. This practical complication is treated in *IT* C (1999, Parts 7 and 8).

Both the ideal centric and acentric distributions are simple members of the family of gamma distributions, defined by

$$\gamma_n(x)\ \mathrm{d}x = [\Gamma(n)]^{-1} x^{n-1} \exp(-x)\ \mathrm{d}x, \qquad (2.1.5.16)$$

where n is a parameter, not necessarily integral, and $\Gamma(n)$ is the gamma function. Thus the ideal acentric intensity distribution is

$$p(I)\ \mathrm{d}I = \exp(-I/\Sigma)\ \mathrm{d}(I/\Sigma) \qquad (2.1.5.17)$$

$$= \gamma_1(I/\Sigma)\ \mathrm{d}(I/\Sigma) \qquad (2.1.5.18)$$

and the ideal centric intensity distribution is

$$p(I)\ \mathrm{d}I = (2\Sigma/\pi)^{1/2} \exp[-I/(2\Sigma)]\ \mathrm{d}[I/(2\Sigma)] \qquad (2.1.5.19)$$

$$= \gamma_{1/2}[I/(2\Sigma)]\ \mathrm{d}[I/(2\Sigma)]. \qquad (2.1.5.20)$$

The properties of gamma distributions and of the related beta distributions, summarized in Table 2.1.5.1, are used in Section 2.1.6 to derive the probability density functions of sums and of ratios of intensities drawn from one of the ideal distributions.

2.1.5.6. *Cumulative distribution functions*

The integral of the probability density function $f(x)$ from the lower end of its range up to an arbitrary value x is called the cumulative probability distribution, or simply the distribution function, $F(x)$, of x. It can always be written

$$F(x) = \int_{-\infty}^{x} f(u)\ \mathrm{d}u; \qquad (2.1.5.21)$$

if the lower end of its range is not actually $-\infty$ one takes $f(x)$ as identically zero between $-\infty$ and the lower end of its range. For the distribution of A [equation (2.1.5.4) or (2.1.5.9)] the lower limit is in

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

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

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

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

$$F(x) = \int_0^x f(x)\, \mathrm{d}x. \tag{2.1.5.22}$$

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:

$$N(|E|) = 2 \int_0^{|E|} y \exp(-y^2)\, \mathrm{d}y = 1 - \exp(-|E|^2). \tag{2.1.5.23}$$

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 erf(x), defined by

$$\mathrm{erf}(x) = (2/\pi^{1/2}) \int_0^x \exp(-t^2)\, \mathrm{d}t. \tag{2.1.5.24}$$

For the centric distribution, then

$$N(|E|) = (2/\pi)^{1/2} \int_0^{|E|} y \exp(-y^2/2)\, \mathrm{d}y \tag{2.1.5.25}$$

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

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 Σ [equation (2.1.2.4)]. When the number of reflections is not 'sufficient', their mean value will show statistical fluctuations about Σ; 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

$$J_n = \sum_{i=1}^{n} G_i, \tag{2.1.6.1}$$

and averages like

$$Y = J_n/n, \tag{2.1.6.2}$$

where G_i is the intensity of the ith reflection. The probability density distributions are easily obtained from a property of gamma distributions: If $x_1, x_2, \ldots, x_n$ are independent gamma-distributed variables with parameters $p_1, p_2, \ldots, p_n$, 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

$$p(J_n)\, \mathrm{d}J_n = \gamma_n(J_n/\Sigma)\, \mathrm{d}(J_n/\Sigma); \tag{2.1.6.3}$$

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

$$p(J_n)\, \mathrm{d}J_n = \gamma_{n/2}[J_n/(2\Sigma)]\, \mathrm{d}[J_n/(2\Sigma)]; \tag{2.1.6.4}$$

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

$$p(Y)\, \mathrm{d}Y = \gamma_n(nY/\Sigma)\, \mathrm{d}(nY/\Sigma) \tag{2.1.6.5}$$

for the acentric case, and

$$p(Y)\, \mathrm{d}Y = \gamma_{n/2}[nY/(2\Sigma)]\, \mathrm{d}[nY/(2\Sigma)] \tag{2.1.6.6}$$

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

2.1.6.2. *Distribution of ratios*

Ratios like

$$S_{n,m} = J_n/K_m, \tag{2.1.6.7}$$

where J_n is given by equation (2.1.6.1),

$$K_m = \sum_{j=1}^{m} H_j, \tag{2.1.6.8}$$

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