International Tables for Crystallography (2006). Vol. B, Section 2.1.6.2, pp. 197–198.

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

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

For the centric distribution, then

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

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

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 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 Σ, 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, of 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

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

$$p(u)\,\mathrm{d}u = \beta_2[nY/(mZ);\, n, m]\,\mathrm{d}[nY/(mZ)], \qquad (2.1.6.10)$$

where β_2 is a beta distribution of the second kind, Y is given by equation (2.1.6.2) and Z by

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

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

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

with variance

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

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

$$p(u)\,\mathrm{d}u = \beta_2[nY/(mZ);\, n/2, m/2]\,\mathrm{d}[nY/(mZ)] \qquad (2.1.6.14)$$

with the expected value of Y/Z equal to

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

and with its variance equal to

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

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, 1978*c*) that is not so easily removed. Wilson (1986*a*) has given reasons for supposing that the bias of the ratio (2.1.6.7) approximates to

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

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

$$p(I/\langle I\rangle)\,\mathrm{d}(I/\langle I\rangle) = \beta_1(I/n\langle I\rangle;\, 1, n-1)\,\mathrm{d}(I/n\langle I\rangle) \qquad (2.1.6.18)$$

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

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

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

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

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

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

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

Similar considerations apply to intensities normalized to Σ 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

references