LIDIN, S., ANDERSSON, S., BOVIN, J.-O., MALM, J.-O. & TERASAKI, O. (1989). *Acta Cryst.* A**45**, *FC*33–*FC*36.
NINHAM, B. W. (1991). *Acta Chem. Scand.* **45**, 775–780.
NINHAM, B. W., HUGHES, B. D., FRANKEL, N. E. & GLASSER, M. L. (1992). *J. Phys. A.* In the press.
PENROSE, R. (1974). *Bull. Inst. Math. Appl.* **10**, 266–271.
ROBINSON, R. M. (1975). Mimeographed notes as quoted by Grünbaum & Shephard (1987).
SHLESINGER, M. F. & HUGHES, B. D. (1981). *Physica (Utrecht)*, A**109**, 597–608.
THOMPSON, D. W. (1942). *On Growth and Form.* Cambridge Univ. Press.

SHORT COMMUNICATIONS

Contributions intended for publication under this heading should be expressly so marked; they should not exceed about 1000 *words; they should be forwarded in the usual way to the appropriate Co-editor; they will be published as speedily as possible.*

Acta Cryst. (1992). A**48**, 649–650

Largest likely *R* factors for normal distributions.

By R. P. MILLANE, *Whistler Center for Carbohydrate Research, Purdue University, West Lafayette, Indiana* 47907-1160, *USA*

(*Received* 10 *December* 1991; *accepted* 3 *March* 1992)

Abstract

An expression is obtained for the largest likely R factor for data that are normally distributed. For zero-mean data, the largest likely R factor is $2^{1/2}$ and, for positive data ($\mu \gg \sigma$), it is equal to $2\sigma/(\mu\pi^{1/2})$. These results are applied to fiber diffraction and other possible applications in crystallography are discussed.

R factors are used in a variety of areas in crystallography as a measure of the similarity between two sets of parameters or data. Some applications are in assessing accuracies of structures, errors in scaling, effectiveness of derivatization and lack of phase closure in isomorphous replacement. Evaluating the significance of a particular-valued R factor is aided by comparison with the largest likely R factor; that which would be obtained if the two sets of parameters or data were unrelated or uncorrelated. The largest likely R factor depends on the statistical distribution of the data. Largest likely R factors have been derived for structures determined by crystallography (Wilson, 1950) and by fiber diffraction (Stubbs, 1989; Millane 1989*a*, *b*, 1990*a*, 1992). Largest likely R factors are derived here for data that are normally distributed. Applications to fiber diffraction are described and other possible applications are discussed.

Consider two sets of data x and y (not necessarily positive) that are compared by calculating the R factor

$$R = \sum_i |x_i - y_i| \Big/ \sum_i |x_i| = \langle\delta\rangle/\langle|x|\rangle, \tag{1}$$

where $\delta = |x - y|$ and $\langle\ \rangle$ denotes the average. From Wilson (1950), the probability density for δ, $Q(\delta)$, is given by

$$Q(\delta) = \int_{-\infty}^{\infty} P(x)P(x+\delta)\,dx \tag{2}$$

and $G(x)$ is defined by

$$G(x) = \int_{-\infty}^{x} x'P(x')\,dx'. \tag{3}$$

Using these equations shows that

$$\langle\delta\rangle = 2[\langle x\rangle - 2\langle G(x)\rangle] \tag{4}$$

so that the largest likely R factor is given by

$$R = [2\langle x\rangle - 4\langle G(x)\rangle]/\langle|x|\rangle. \tag{5}$$

Equation (5) is a general result for any distribution of x, and reduces to equation (6) of Wilson (1950) if $x \geq 0$.

If the random variables x and y are identically normally distributed with mean μ and variance σ^2, *i.e.*

$$P(x) = (2\pi)^{-1/2}\sigma^{-1}\exp[-(x-\mu)^2/2\sigma^2], \tag{6}$$

then $\langle x\rangle = \mu$ and $\langle|x|\rangle$ is given by

$$\langle|x|\rangle = \int_0^{\infty} x[P(x) + P(-x)]\,dx \tag{7}$$

so that

$$\langle|x|\rangle = (2/\pi)^{1/2}\sigma\exp(-\mu^2/2\sigma^2) + \mu\,\mathrm{erf}(\mu/2^{1/2}\sigma), \tag{8}$$

where erf$(\cdot)$ denotes the error function. Note that, for $\mu/\sigma \to \infty$, $\langle|x|\rangle \to \mu$ (as it must, since when $\mu \gg \sigma$ most values of x will be positive) and that, for $\mu = 0$, $\langle|x|\rangle = (2/\pi)^{1/2}\sigma$. Substituting (6) into (3) shows that

$$G(x) = -(2\pi)^{-1/2}\sigma\exp[-(x-\mu)^2/2\sigma^2] + (\mu/2)\{1 + \mathrm{erf}[(x-\mu)/2^{1/2}\sigma]\} \tag{9}$$

and evaluating the mean gives

$$\langle G(x)\rangle = \mu/2 - \sigma/(2\pi^{1/2}). \tag{10}$$

Substituting (10) into (5) gives

$$R = 2\sigma/(\pi^{1/2}\langle|x|\rangle), \tag{11}$$

where $\langle|x|\rangle$ is given by (8), which is the desired result. Note that, for zero-mean data, the largest likely R factor is

$$R = 2^{1/2}, \qquad \text{for } \mu = 0. \tag{12}$$

It is instructive to examine the dependence of $\langle|x|\rangle$ on μ, shown as the solid line in Fig. 1. The approximation

$$\langle|x|\rangle \simeq \mu \tag{13}$$

is the first term in the asymptotic expansion for $\langle|x|\rangle$ as $\mu/\sigma \to \infty$, and is quite accurate for $\mu/\sigma \gtrsim 1.5$ (broken line in Fig. 1). For small μ, the power-series expansions for the

0108-7673/92/040649-02$06.00

error function (Abramowitz & Stegun, 1972) and the exponential function can be used in (8), giving

$$\langle|x|\rangle \simeq (2/\pi)^{1/2}\sigma(1+\mu^2/2\sigma^2), \qquad \mu/\sigma \to 0. \quad (14)$$

This approximation is shown as the dotted line in Fig. 1 and is seen to be quite accurate for $\mu/\sigma \lesssim 1.5$. Substituting (13) into (11) gives

$$R = 2\sigma/(\mu\pi^{1/2}), \qquad \text{for } \mu/\sigma \to \infty, \quad (15)$$

which is quite accurate for $\mu \gtrsim 1.5\sigma$. For positive data that are (approximately) normally distributed, $\mu \gg \sigma$, so that the simple expression (15) is very accurate.

Applications of these results to fiber diffraction are now described. The cylindrical averaging of fiber diffraction patterns means that, as a result of the central limit theorem, the amplitudes are approximately normally distributed, the approximation improving with an increasing number, m, of contributing Fourier-Bessel structure factors (Millane, 1990*b*). This description applies to a noncrystalline fiber, but the same is true for a polycrystalline fiber where the cylindrical averaging results in the superposition of different reflections. The mean and variance of the amplitude distribution at a point on a fiber diffraction pattern where m terms contribute are given by (Millane, 1990*b*)

$$\mu_m = \varepsilon^{1/2}\Gamma(m/2+1/2)/\Gamma(m/2) \quad (16)$$

and

$$\sigma_m^2 = \varepsilon m/2 - \mu_m^2, \quad (17)$$

where $\Gamma(\cdot)$ is the gamma function and ε is defined in Millane (1990*b*). Asymptotic approximations for large m (which are quite accurate) are given by (Millane, 1990*b*)

$$\mu_m \simeq (\varepsilon m/2)^{1/2}, \quad m \to \infty \quad (18)$$

and

$$\sigma_m^2 \simeq \varepsilon/4, \quad m \to \infty. \quad (19)$$

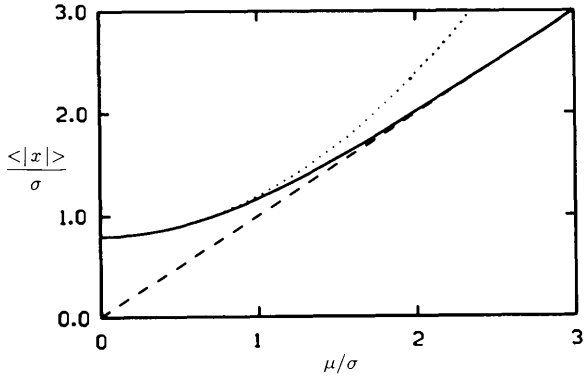

Fig. 1. $\langle|x|\rangle/\sigma$ as a function of μ/σ. The different curves are the exact values using (8) (——) and approximations using (13) (- - -) and (14) (· · ·).

Table 1. *Exact (R_m) and approximate (R'_m and R''_m) largest likely R factors in fiber diffraction for m overlapping (real and imaginary) Fourier-Bessel terms*

m	R_m	R'_m	R''_m
1	0.828	0.798	0.853
2	0.586	0.564	0.590
3	0.475	0.461	0.476
4	0.409	0.399	0.410
5	0.364	0.357	0.365
6	0.332	0.326	0.332
7	0.306	0.302	0.307
8	0.286	0.282	0.286
9	0.269	0.266	0.269
10	0.255	0.252	0.255

R_m is calculated using equation (12) of Millane (1989*a*) and R'_m and R''_m using (20) and (21), respectively, of this paper.

Substituting (18) and (19) into (15) gives the approximation

$$R_m \simeq (2/\pi)^{1/2} m^{-1/2}, \quad m \to \infty \quad (20)$$

for the largest likely R factor. This agrees with the result derived by Millane (1990*a*), using a rather tedious asymptotic analysis of the incomplete beta function, and it was shown to be quite accurate. Equation (20) can be used to calculate quite accurate largest likely R factors for a whole fiber diffraction pattern, both numerically (Millane, 1990*a*) and analytically (Millane, 1992). A more accurate approximation can be obtained by simply substituting (16) and (17) into (15), giving

$$R_m \simeq \left\{\frac{2m[\Gamma(m/2)]^2}{\pi[\Gamma(m/2+1/2)]^2} - \frac{4}{\pi}\right\}^{1/2}. \quad (21)$$

This is considerably simpler than the exact expression [equation (12) of Millane (1989*a*)], but is exceptionally accurate as can be seen from Table 1, where it is compared with the exact values and with the approximation (20).

In conclusion, (11) gives the largest likely R factor for comparison of two sets of normally distributed parameters or data. It varies between $2^{1/2}$ for zero-mean data and $2\sigma/(\mu\pi^{1/2})$ for $\mu \gtrsim 1.5\sigma$. These results can be used to obtain approximate largest likely R factors in fiber diffraction very easily. The results could also be applied in powder diffraction (where the amplitudes are also approximately normally distributed) and other applications may exist.

I am grateful to the US National Science Foundation for support (DMB-8916477) and to Pat Aveline and Cheryl Ralston for word processing.

References

ABRAMOWITZ, M. & STEGUN, I. A. (1972). *Handbook of Mathematical Functions*, p. 297. New York: Dover.
MILLANE, R. P. (1989*a*). *Acta Cryst.* **A45**, 258-260.
MILLANE, R. P. (1989*b*). *Acta Cryst.* **A45**, 573-576.
MILLANE, R. P. (1990*a*). *Acta Cryst.* **A46**, 68-72.
MILLANE, R. P. (1990*b*). *Acta Cryst.* **A46**, 552-559.
MILLANE, R. P. (1992). *Acta Cryst.* **A48**, 209-215.
STUBBS, G. (1989). *Acta Cryst.* **A45**, 254-258.
WILSON, A. J. C. (1950). *Acta Cryst.* **3**, 397-399.