THE ESTIMATED INTENSITY VARIANCE AS A MEAN OF EXPECTED AND SAMPLE VARIANCES

WALTER GONSCHOREK (*)
Departamento de Física da Universidade de Coimbra
3000 Coimbra, Portugal

(Received 19 January 1981)

ABSTRACT — Two estimates for the variance σ^2 (I) of repeatedly measured integrated intensities I are given, one based on the sample variance the other one based on Poisson statistics. The weighted mean of both is taken as the final estimate of σ^2 (I). The ratio of this final estimate to the «sample variance» of symmetry dependent intensities plotted once against I and once against sin ϑ/λ and alternatively a χ^2-test can help to detect systematic errors inherent in the intensities or errors of the final estimates of σ^2 (I)

1 — INTRODUCTION

Weighting, in least squares procedures, can considerably influence the refined parameters. Given a set of observations x_i to be compared with calculated values x_{ci}, the function to be minimized is [1]

$$Q = \Sigma (x_i - x_{ci})^2 \sigma_i^{-2} \qquad (1)$$

σ_i^2 is the variance of x_i. Its reciprocal σ_i^{-2} is called the weight of the observation x_i. Repeated observations x_i of the same quantity under equal conditions have the same variance σ_i^2. Therefore in (1) they can be replaced by their mean with the variance of the mean σ_i^2/n if n values x_i are observations of the same quantity [2]. This can be understood without applying the rules of

(*) On leave from Institut für Kristallographie, T. H. Aachen, Germany.

statistics by simply considering the normal equations to be derived from (1) [3]. The problem is how to determine the variances σ_i^2.

Undoubtedly the processes of generating and diffracting X-rays or neutrons obey Poisson statistics so that the variance of an integrated intensity I is known in principle [4]. In practice, however, sample variances $s^2(I)$, which in statistics serve as estimates for the population variances $\sigma^2(I)$, often are considerably larger than $\sigma_P^2(I)$ of Poisson distributions [5] [6] (*). McCandlish, Stout & Andrews [6] describe a procedure by which, through the use of somewhat modified sample variances derived from repeatedly-measured standard intensities, two correction terms are added to the variance obtained from the Poisson distribution. Abrahams [7] emphasizes the experiment as the best proof for any error estimation. He gives a list of some presumed error sources and proposes to estimate the contribution of each individual error source to the total error if it is not accessible either to experiment or to theory, as for example intensity drift of the primary beam or statistical variances are. On the other side Hamilton [8] (p. 148) strictly gives preference to theoretical variances over sample variances at least for small samples just as Schulz & Schwarz [9] practise.

In electron density work the reliability of the variances $\sigma^2(I)$ (or their estimates) seems to be as important as the reliability of the integrated intensities I themselves. The variances $\sigma^2(I)$ in least squares refinements besides their action through weights via the goodness of fit parameters serve as an indicator if either the model is inadequate or the data are burdened by some hidden errors or if both situations occur. In Fourier methods reliable estimates of $\sigma^2(I)$ are indispensable to decide whether the experimentally determined electron densities are significant or not.

Regarding the «instability constant» [6] it must be noticed that in X-ray diffraction the strongest reflections are low order reflections and these contain most of the information on bonding electrons. If therefore the deformation of the atoms or ions caused by bonding effects are not taken into account the differences $\Delta = |I_{obs} - I_{calc}|$ naturally will be larger for strong reflections than

(*) Schulz in his expression for the variance of the corrected intensity already took into account a term representing the uncertainty of the scaling parameter deduced from control or standard reflections.

for weak ones. Furthermore strong reflections are most affected by extinction and counting losses through dead time of the counting devices. The correction of these effects introduces additional errors whose magnitudes must be estimated and taken into account.

If, however, the estimation of $\sigma^2(I)$ through sample variances reveals a linear dependence of $\sigma^2(I)$ upon I^2 this error can be kept small using small crystals or weaker primary intensities. Some other error sources which preferably affect strong reflections like weakening filters or movable β - filters can be easily avoided.

In this paper it is assumed that all intensities have been corrected for time drifts according to McCandlish et al. [6] if this turned out to be necessary. The variances $\sigma^2_{P_{ji}}$ are thought of to have the form $\sigma^2_{P_{ji}} = \sigma^2_P + \sigma^2_K$ where σ^2_P is the variance derived from Poisson statistics and σ^2_K is the variance which takes into account the uncertainty of the scale parameter.

In the following two estimates of the intensity variances shall be given, the first one based on sample variances according to (8), the second one derived from Poisson variances. The weighted mean of both shall be taken as the final estimate of $\sigma^2(I)$.

2 — CONFIDENCE LIMITS AND WEIGHTS

Suppose (with close reference to Hamilton [8] (p. 40)) that m samples each with n_j observations all having the same population mean μ led to m means $\bar{x}_j$ with variances $\sigma^2(\bar{x}_j)$. Then it is reasonable to take the weighted mean $\bar{x}$:

$$\bar{x} = \sum_{j=1}^{m} w_j \bar{x}_j \tag{2}$$

$$w_j = \sigma^{-2}(\bar{x}_j) \Big/ \sum_{i}^{m} \sigma^{-2}(\bar{x}_j) \tag{3}$$

In w_j, the quantity $\sigma(\bar{x}_j)$, considered as a confidence limit, guarantees a certain probability P_j for the interval $(\bar{x}_j - \sigma(\bar{x}_j), \bar{x}_j + \sigma(\bar{x}_j))$ to include the population mean μ. For normally distributed observations this probability has the value $P_G = 0.682689$.

Therefore it is proposed to replace the $\sigma(\bar{x}_j)$ in (3) by confidence limits s_{tj} for the probability P_G if the $\sigma(\bar{x}_j)$ are unknown :

$$s_{tj} = t_\alpha (n_j - 1)\, s_j / n_j^{1/2} \tag{4}$$

$t_\alpha(\nu)$ = fractile (percentage point) of Student's t-distribution for the probability $\alpha = 1 - P_G$ and the degree of freedom ν, s_j^2 = sample variance of the j-th sample. Table I gives the fractiles $t_\alpha(\nu)$.

TABLE I — Fractiles of Student's t-distribution for the two-tailed probability $\alpha = 1 - P_G$, $P_G = 0.682689$ (*). ν is the number of degrees of freedom.

ν	$t_\alpha(\nu)$	ν	$t_\alpha(\nu)$	ν	$t_\alpha(\nu)$
1	1.837	7	1.077	13	1.040
2	1.321	8	1.067	14	1.037
3	1.197	9	1.059	15	1.034
4	1.142	10	1.053	20	1.026
5	1.111	11	1.048	25	1.020
6	1.091	12	1.043	30	1.017

(*) P_G is the value of the integral taken over the Gaussian density function from $-\sigma$ to $+\sigma$, $\mu = 0$. The integration was carried out by the quadrature method of Gauss once with 24 and once with 26 grid points for the integration from 0 to σ. Both results (with 24 and with 26 grid points) agreed to within the 10th digit after the decimal point. The fractiles were obtained by integration of Student's t-distribution function again using the quadrature method of Gauss with 26 grid points. The limits of integration were varied until the value of the integral deviated less than 4.10^{-6} from $P_G = 0.6826894921$. That last limit of integration was taken as $t_\alpha(\nu)$.

3 — TWO ESTIMATES OF $\sigma^2(\bar{I}_j)$

Suppose the integrated intensity of the j-th reflection repeatedly has been measured n_j times yielding the integrated intensities I_{ji} and variances $\sigma^2_{P_{ji}} = \sigma^2_P + \sigma^2_K$: The weighted mean

$$\bar{I}_j = \sum_i^{n_j} w_i I_{ji} \tag{5}$$

is taken as the integrated intensity of that reflection. Here the weights w_i have the form

$$w_i = \sigma_{P\,ji}^{-2} \Big/ \sum_{k}^{n_j} \sigma_{P\,jk}^{-2} \tag{6}$$

3.1 — *Estimate of* $\sigma^2(\bar{I}_j)$ *through the sample variance*

According to (4) the first estimate of $\sigma^2(\bar{I}_j)$ takes the form

$$s_{tj}^2 = t_\alpha^2 (n_j - 1)\, s_j^2 / n_j \tag{7}$$

with $\alpha = 1 - P_G$ and

$$s_j^2 = (n_j - 1)^{-1} \sum_{i}^{n_j} w_i (I_{ji} - \bar{I}_j)^2 \tag{8}$$

with w_i as defined in (6).

The variance of s_j^2 is derived in Appendix A. With that result the variance of s_{tj}^2 is

$$\sigma^2(s_{tj}^2) = t_\alpha^4 (n_j - 1)\; 2(n_j - 1)^{-1} \Big/ \Big(\sum_{i}^{n_j} \sigma_{ji}^{-2}\Big)^2 \tag{9}$$

σ_{ji}^2 is the (unknown) variance of I_{ji}. If in (9) σ_{ji}^2 is replaced by σ_{Pji}^2 and if the definition (12) is used, $\sigma^2(s_{tj}^2)$ takes the form

$$\sigma_P^2(s_{tj}^2) = t_\alpha^4 (n_j - 1)\; 2(n_j - 1)^{-1}\; \sigma_P^4(\bar{I}_j) \tag{10}$$

3.2 — *Estimate of* $\sigma^2(\bar{I}_j)$ *through Poisson statistics*

For m repeatedly measured reference reflections, the weighted means according to (5) and (6) and the quantities s_i^2 (eq. (8)) $j = 1, \ldots m$ are taken. The variances of I_j derived from (5) are

$$\sigma^2(\bar{I}_j) = \sum_{i}^{n_j} w_i^2\; \sigma_{ji}^2 \tag{11}$$

Here σ^2_{ji} is replaced by σ^2_{Pji}. This leads to

$$\sigma^2_P(\bar{I}_j) = (\sum_i^{n_j} \sigma^{-2}_{Pji})^{-1} \tag{12}$$

Now the ratios v_j are taken

$$v_j = s^2_{tj} / \sigma^2_P(\bar{I}_j) \tag{13}$$

Their weighted mean is

$$\bar{v} = \sum_j^m s_t^{-2}(v_j)\, v_j / \sum_k^m s_t^{-2}(v_k) \tag{14}$$

with $s^2_t(v_j)$ as derived in Appendix B.

Then as the second estimate of $\sigma^2(\bar{I}_j)$ for each repeatedly measured intensity the quantity s''^2_j is taken:

$$s''^2_j = \bar{v}\ \sigma^2_P(\bar{I}_j) \tag{15}$$

4 — THE FINAL ESTIMATE OF $\sigma^2(\bar{I}_j)$

The final estimate of $\sigma^2(\bar{I}_j)$ is

$$s'^2(\bar{I}_j) = (\sigma_P^{-2}(s^2_{tj})\, s^2_{tj} + s_t^{-2}(s'^2_j)\, s'^2_j) / (\sigma_P^{-2}(s^2_{tj}) + s_t^{-2}(s''^2)) \tag{16}$$

with $\sigma^2_P(s^2_{tj})$ as defined in (10) and with $s^2_t(s'^2_j)$ as derived in Appendix C.

An estimate of the variance of $s'^2(\bar{I}_j)$ is

$$s^2(s'^2(\bar{I}_j)) = (\sigma_p^{-2}(s^2_{tj}) + s_t^{-2}(s''^2))^{-1} \tag{17}$$

If for some j-th reflection there exists only one single measurement I_{ji} its variance is estimated as

$$s'^2(I_{ji}) = \bar{v}\, \sigma^2_{Pji} \tag{18}$$

An estimate of the variance of $s'^2(I_{ji})$ is

$$s^2(s'^2(I_{ji})) = s^2_t(\bar{v})\, \sigma^4_{Pji} \tag{19}$$

with $s^2_t(\bar{v})$ as defined in (C6).

Often $\bar{I}_j$ is corrected by some factor c_j (e. g. absorption): $\bar{I}_{jc} = c_j \bar{I}_j$. In the following $s'^2(\bar{I}_j)$ must then be replaced by the appropriate estimate of $\sigma^2(c_j \bar{I}_j)$.

5 — SYMMETRY DEPENDENT INTENSITIES

For symmetry dependent intensities $\bar{I}_j$, $j = 1, \ldots n_D$ the weighted mean is taken:

$$\bar{I} = \sum_j^{n_D} w_j \bar{I}_j \tag{20}$$

with weights

$$w_j = s'^{-2}(\bar{I}_j) \; / \sum_k^{n_D} s'^{-2}(\bar{I}_k) \tag{21}$$

An estimate of the variance of $\bar{I}$ deduced from (20) is

$$s'^2(\bar{I}) = (\sum_j^{n_D} s'^{-2}(\bar{I}_j))^{-1} \tag{22}$$

and this should be taken as the estimate of $\sigma^2(\bar{I})$. The estimated variance $s'^2(\bar{I})$ can be compared with the weighted scatter of $\bar{I}_j$:

$$s^2(\bar{I}) = (n_D - 1)^{-1} \sum_j^{n_D} w_j (\bar{I}_j - \bar{I})^2 \tag{23}$$

The ratios of $s_t^2(\bar{I}) = t_a^2 (n_D - 1) s^2(\bar{I})$ to $s'^2(\bar{I})$ plotted once against $\bar{I}$ and once against $\sin\vartheta/\lambda$ may help to reveal systematic dependences if they exist. Alternatively a χ^2 - test according to van der Waerden ([1] p. 222) can be applied. This will be done in a paper which is in preparation.

A considerable part of this work was done at the Institut für Kristallographie der T. H. Aachen. The author, therefore, is deeply indebted to Prof. Th. Hahn. Thanks are due to Prof. Alte da Veiga who gave the opportunity to complete this work. The grant of a research and teaching fellowship by Deutscher Akademischer Austauschdienst is highly appreciated.

APPENDIX A

The variance of the quantity s_j^2 (8) is to be derived. For that purpose the intensities I_{ji}, with constant j, are assumed to be normally and independently distributed with variances σ_{ji}^2 and all with the same (population) mean μ. Then the weights w_i instead of (6) have the form

$$w_i = \sigma_{ji}^{-2} \; / \; \sum_k^{n_j} \sigma_{jk}^{-2} \tag{A1}$$

Correspondingly the variances of $\bar{I}_j$ instead of (12) are

$$\sigma^2(\bar{I}_j) = (\sum_i^{n_j} \sigma_{ji}^{-2})^{-1} \tag{A2}$$

Now the quantity

$$\chi^2 = (n_j - 1)\; s_j^2 \; / \; \sigma^2(\bar{I}_j) \tag{A3}$$

shall be shown to be χ^2 — distributed with $(n_j - 1)$ degrees of freedom. For that purpose according to van der Waerden [1], (p. 111) I_{ji} is replaced by

$$x_i = (I_{ji} - \mu) \; / \; \sigma_{ji}\,; \tag{A4}$$

x_i is normally distributed with mean zero and unit variance. Equation (8) can be rewritten as

$$s_j^2 = (n_j - 1)^{-1} \left[\sum_i^{n_j} w_i I_{ji}^2 - (\sum_i^{n_j} w_i I_{ji})^2 \right] \tag{A5}$$

or, taking (A4) into account,

$$s_j^2 = (n_j - 1)^{-1} \left[\sum_i^{n_j} w_i (\sigma_{ji} x_i + \mu)^2 - (\sigma^2(\bar{I}_j) \sum_i^{n_j} \sigma_{ji}^{-1} x_i + \mu)^2 \right] \tag{A6}$$

This leads to

$$s_j^2 = \sigma^2(\bar{I}_j)\,(n_j - 1)^{-1} \left[\sum_i^{n_j} x_i^2 - (\sum_i^{n_j} w_i^{1/2}\; x_i)^2 \right] \tag{A7}$$

The quantity in square brackets in (A7) is identical with χ^2 defined by (A3). Now the transformation

$$y_1 = \sum_i^{n_j} w_i^{1/2} \, x_i \tag{A8}$$

is considered. The sum of the squares of its coefficients is equal to unity. Therefore this equation can replace the first row of the linear equations in van der Waerden ([1], p. 112) or the last row of the corresponding equations in Martin ([10], p. 59), and the conclusions drawn in these works concerning χ^2 as defined there apply also to χ^2 defined in (A3). Therefore χ^2 (A3) assumes a χ^2 - distribution with $(n_j - 1)$ degrees of freedom and the variance of s_j^2 is

$$\sigma^2 (s_j^2) = 2 (n_j - 1)^{-1} \, \sigma^4 (\bar{I}_j) \tag{A9}$$

APPENDIX B

The variance $\sigma^2 (v_j)$ with v_j as defined in (13) according to Hamilton ([8] p. 32) is given by

$$\sigma^2 (v_j) = \sigma^2 (s_{tj}^2) \, \sigma^2 (\sigma_P^{-2} (\bar{I}_j)) + \sigma_P^{-4} (\bar{I}_j) \, \sigma^2 (s_{tj}^2) + s_{tj}^4 \, \sigma^2 (\sigma_P^{-2} (\bar{I}_j)) \tag{B1}$$

Here it is assumed that s_{tj}^2 and $\sigma_P^{-2} (\bar{I}_j)$ are statistically independent. $\sigma^2 (s_{tj}^2)$ is estimated through (10). $\sigma^2 (\sigma_P^{-2} (\bar{I}_j))$ is found using a formula again given by Hamilton ([8], p. 32):

$$\sigma^2 (\sigma_P^{-2} (\bar{I}_j)) \approx \sigma^2 (\sigma_P^2 (\bar{I}_j)) \; / \; \sigma_P^8 (\bar{I}_j) \tag{B2}$$

To find an estimate of $\sigma^2 (\sigma_P^2 (\bar{I}_j))$ the weighted mean of $\sigma_{P\,ji}^2$ as defined in (6) is considered

$$\sigma_{Pj}^2 = \sum_i^{n_j} w_i \, \sigma_{P\,ji}^2 = n \; / \; \sum_i^{n_j} \sigma_{P\,ji}^{-2} \tag{B3}$$

$\sigma^2_{P\,j}$ is an estimate of the variance of the observations I_{ji} for constant j. Therefore an estimate of the variance of $\bar{I}_j$ is $\sigma^2_{P\,j}/n_i$ which is identical with (12):

$$\sigma^2_P(\bar{I}_j) = \sigma^2_{P\,j}/n_j \tag{B4}$$

Now $\sigma^2_{P_j}$ (B3) being the weighted mean of $\sigma^2_{P_{ji}}$ an estimate of its variance is given by

$$s^2(\sigma^2_{P\,j}) = (n_j - 1)^{-1} \sum_i^{n_j} w_i (\sigma^2_{P\,ji} - \sigma^2_{P_j})^2 \tag{B5}$$

With (B4) this leads to an estimate of $\sigma^2_P(\bar{I}_j)$):

$$s^2(\sigma^2_P(\bar{I}_j)) = s^2(\sigma^2_{P_j}) \;/\; n^2 \tag{B6}$$

Now an estimate of $\sigma^2(v_j)$ (B1) can be given. First in (B2) $\sigma^2(\sigma^2_P(\bar{I}_j))$ is replaced by (B6). $s^2(\sigma^2_P(\bar{I}_j))$ (B6) according to (4) is multiplied by $t^2_\alpha(n_j - 1)$. s^4_{tj} in the last term of (B1) according to (13) is replaced by $v_j^2\,\sigma^4_P(\bar{I}_j)$. Then the estimate of $\sigma^2(v_j)$ is

$$s^2_t(v_j) = t^2_\alpha(n_j - 1)\Big[s^2(\sigma^2_{P\,j})\,n_j^{-2}\,\sigma^{-4}_P(\bar{I}_j)$$

$$(2t^4_\alpha(n_j - 1)/(n_j - 1) + v_j^2) + 2t^2_\alpha(n_j - 1)/(n_j - 1)\Big] \tag{B7}$$

with $s^2(\sigma^2_{P_j})$ as defined in (B5) and $\sigma^2_P(\bar{I}_j)$ as defined in (12).

APPENDIX C

The variance $\sigma^2(s''^2_j)$ according to Hamilton ([8], p. 32) has the form

$$\sigma^2(s''^2_j) = \sigma^2(\bar{v})\,\sigma^2(\sigma^2_P(\bar{I}_j)) + \sigma^4_P(\bar{I}_j)\,\sigma^2(\bar{v}) + \bar{v}^2\,\sigma^2(\sigma^2_P(\bar{I}_j)) \tag{C1}$$

Here $\bar{v}$ and $\sigma^2_P(\bar{I}_j)$ are considered to be statistically independent. An estimate of $\sigma^2(\sigma^2_P(\bar{I}_j))$ is given through (B6). For $\sigma^2(\bar{v})$ two estimates can be derived:

1. $\bar{v}$ as defined in (14) has the variance

$$\sigma^2(\bar{v}) = \sum_j^m w_j^2 \sigma^2(v_j) \tag{C2}$$

with

$$w_j = s_t^{-2}(v_j) \; / \; \sum_k^m s_t^{-2}(v_k) \tag{C3}$$

In (C2) $\sigma^2(v_j)$ is replaced by $s_t^2(v_j)$ (B7) yielding

$$\sigma_t^2(\bar{v}) = (\sum_j^m s_t^{-2}(v_j))^{-1} \tag{C4}$$

2. An alternative estimate of $\sigma^2(\bar{v})$ is the weighted scatter of v_j

$$s^2(\bar{v}) = (m-1)^{-1} \sum_j^m w_j (v_j - \bar{v})^2 \tag{C5}$$

with weights as defined in (C3). As the final estimate of $\sigma^2(\bar{v})$ the maximum

$$s_t^2(\bar{v}) = \text{Max}\ (\sigma_t^2(\bar{v});\ t_\alpha^2(m-1)\, s^2(\bar{v})) \tag{C6}$$

shall be taken. Finally the estimate of $\sigma^2(s_j'^2)$ (C1) to be used in (16) is

$$s_t^2(s_j'^2) = t_\alpha^2(n_j - 1)\, s^2(\sigma_{P_j}^2)\, n_j^{-2} \left[s_t^2(\bar{v}) + \bar{v}^2 \right] + \sigma_P^4(\bar{I}_j)\, s_t^2(\bar{v}) \tag{C7}$$

REFERENCES

[1] VAN DER WAERDEN, B. L. (1965). *Mathematische Statistik*. Berlin: Springer.

[2] CRUICKSHANK, D. W. J. (1965). *Computing Methods in Crystallography*, edited by J. S. Rollett, pp. 99-106. Oxford: Pergamon Press.

[3] ROLLETT, J. S. (1970). *Crystallographic computing*, edited by F. R. Ahmed, pp. 167-181.

[4] SCHULZ, H. & HUBER, P. H. (1971). *Acta Cryst.* **A27**, 536-539.

[5] SCHULZ, H. (1971). *Acta Cryst.* **A27**, 540-544.

[6] MCCANDLISH, L. E., STOUT, G. H. & ANDREWS, L. C. (1975). *Acta Cryst.* **A31**, 245-249.
[7] ABRAHAMS, S. C. (1969). *Acta Cryst.* **A25**, 165-173.
[8] HAMILTON, C. H. (1964). *Statistics in Physical Science.* New York: Ronald Press.
[9] SCHULZ, H. & SCHWARZ, K. H. (1978). *Acta Cryst.* **A34**, 999-1005.
[10] MARTIN, B. R. (1971). *Statistics for Physicists.* London: Academic Press.