FOUNDATIONS ADVANCES

ISSN 2053-2733

Statistical tests against systematic errors in data sets based on the equality of residual means and variances from control samples: theory and applications

Julian Henn[a]* and Kathrin Meindl[b]

[a]Emil-Warburg Weg 6, 95447 Bayreuth, Germany, and [b]Instituto de Biología Molecular de Barcelona (IBMB-CSIC), Barcelona Science Park, Baldiri Reixach 15, 08028 Barcelona, Spain. *Correspondence e-mail: julian.henn@uni-bayreuth.de

Received 8 September 2014
Accepted 15 December 2014

Keywords: fit-quality indicators; statistical tests; residuals; least-squares refinement.

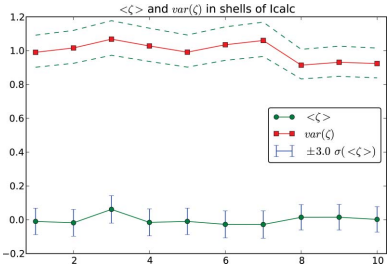
Supporting information: this article has supporting information at journals.iucr.org/a

Statistical tests are applied for the detection of systematic errors in data sets from least-squares refinements or other residual-based reconstruction processes. Samples of the residuals of the data are tested against the hypothesis that they belong to the same distribution. For this it is necessary that they show the same mean values and variances within the limits given by statistical fluctuations. When the samples differ significantly from each other, they are not from the same distribution within the limits set by the significance level. Therefore they cannot originate from a single Gaussian function in this case. It is shown that a significance cutoff results in exactly this case. Significance cutoffs are still frequently used in charge-density studies. The tests are applied to artificial data with and without systematic errors and to experimental data from the literature.

1. Introduction

Systematic errors in diffraction data are an important and widely neglected topic. There already exist a variety of standard tools for the detection of systematic errors such as the goodness of fit (GoF), normal probability plots (n.p.p.'s) (Abrahams & Keve, 1971) or plots of observed *versus* calculated intensities and of the residuals against the resolution; however, not even these simple and quick-to-apply tools are published frequently in charge-density studies. A high quality of the data/model is easy to prove with these tools, yet these are not often made accessible. Why is that? Publication of standard structures and charge-density studies with large GoF values and bent n.p.p.'s seems to be perfectly acceptable in the community. The appearance of large GoF values induces not even a discussion of the underlying causes.

In order to draw the community's attention more to this important topic of systematic errors, further tools have been developed and applied. It is important to develop graphical representations of the data that are conclusive with respect to the underlying error. The developed tools comprise real-space methods, like the fractal dimension plots together with the local and global descriptors net and gross residual electrons (Meindl & Henn, 2008), and the 'percentage of features' (POF) (Meindl & Henn, 2012), that is derived from the fractal dimension. With the help of the fractal dimension plot and artificial data it was shown that neglect of third-order anharmonic motion leads to a characteristic (shashlik-like) pattern in the residual density (Meindl *et al.*, 2010). This imprint was also found in experimental data (Henn *et al.*, 2010; Herbst-



Irmer *et al.*, 2013) and confirmed by other scientists [see, for example, Paul *et al.* (2011) and Poulain *et al.* (2014)].

Reciprocal-space methods were also developed, like the theoretical *R* values (Henn & Schönleber, 2013; Henn & Meindl, 2014*a*). These predict the agreement factor from the diffraction data under the assumption of absence of systematic errors and can be applied in a similar way to the free *R* value (Brünger, 1992). An important result from the theoretical *R* values is that when no systematic errors apply, the actual *R* value from the refinement is basically the square root of a constant that takes the degrees of freedom into account multiplied by the inverse mean-squared significance of the data. As a consequence, every action that increases the significance of the data leads to a reduction of the agreement factor. Among these actions leading to reduced *R* values are valid ones like increasing the redundancy of the data and invalid ones that only formally increase the average significance of the data set. These are, for example, application of a significance cutoff or underestimation of the standard uncertainties (s.u.'s) of the strong reflections. It appears that all of these valid and invalid methods are applied in current charge-density studies (Henn & Meindl, 2014*a*,*b*).

Additional descriptors were derived from the theoretical *R* values like the meta residual value $R^{\rm meta}$. The theoretical *R* values and the meta residual value, however, both assume a Gaussian distribution of residuals. This assumption holds neither in standard structure determination (Henn & Schönleber, 2013) nor in charge-density studies (Henn & Meindl, 2014*a*,*b*).

Statistical methods for the analysis of the distribution of residuals under certain limiting assumptions were developed, too, and extended to more general cases (Henn & Meindl, 2010, 2012). Recently, a statistical analysis method based on the uniform distribution of residuals with respect to different properties like resolution or intensity was developed and implemented with the help of a χ^2 test (Henn & Meindl, 2014*b*). A visual representation of these distributions that correspond to conditional probabilities was developed, too (BayCoN plots). These are expected to be of help in identifying sources of systematic errors. In the present work we complement the selection of tools by a statistical test of residuals against equality of residual mean values and variances and by analysis methods based on the distribution of rare events.

2. Theory

When the residuals $\zeta = (I_{\rm o} - I_{\rm c})/[\sigma(I_{\rm o})]$ all stem from one and the same distribution, which need not be a Gaussian distribution, different subsets of the residuals, *e.g.* from resolution or calculated intensity shells, will show similar mean values and variances within the limits given by statistical fluctuations.[1] Only if the residuals stem from one and the same distribution may they belong to a Gaussian distribution.

For a sample of n values taken from a continuous random variable X from a normal distribution with mean value μ and variance σ^2, the mean values $\overline{X}$ are distributed according to a normal distribution with the same mean value μ and reduced variance σ^2/n. This holds according to the central limit theorem approximately also in the case when X follows any arbitrary distribution with mean value μ and variance σ^2 (Semendjajew *et al.*, 2012).

2.1. Mean values of binned residuals

For applications to experimental data it is convenient to calculate the mean values of the residuals for bins of approximately the same number of data points n (strength) and to equip these with error bars. The bin mean values should be consistent with each other, *i.e.* in a plot of the mean values for different bins it should be possible to draw a line such that each error bar is crossed once. The mean values should also be consistent with zero, *i.e.* the zero line should cross all error bars once. In order to be tolerant of outliers the error bar might be chosen to be as large as $\pm 3(\sigma_i^2/n)^{1/2}$, where the variance σ_i^2 is the population variance from the ith bin.

The bins of equal strength may be taken randomly from the data or from the data sorted according to the resolution, the calculated intensity, or the experimental s.u.'s. In these cases the expected residual bin mean values are zero and their variances are one. Each way of sorting the data might reveal new dependencies of the mean values. Two ways of sorting deserve a short discussion, beginning with when the data are sorted according to the residuals themselves, for example in increasing order from the negative to the positive ones. It is clear that this cannot generally lead to similar mean residual values, as the data are sorted with respect to this property. It should, however, lead to similar variances for all bins but those of the most extreme values. This is because sorting residual data from a normal distribution in increasing order results in an 's' curve with approximately constant slope in the middle part. The exact value of the variance depends on the total number of data, the number of bins and the number of data points in each bin. For the extreme case of only two data points in each bin and an infinite large data set it is obvious that the expected variance for each bin approaches zero. When, however, for the same infinite large data set only two bins are chosen, one bin will contain the negative residuals and the other one the positive residuals leading to identical variances well below one in each bin.

Secondly, when the data are sorted according to squared residuals, this cannot generally lead to similar variances, but it should lead to similar mean values, as still all residual mean values should be consistent with the value zero.

This will become clearer when the respective plots are discussed in §4.

[1] We will not further distinguish here between residuals based on the bare s.u. from the reflection file and those derived from error models. In all cases the residuals are based on that entity which was used in the least-squares refinement.

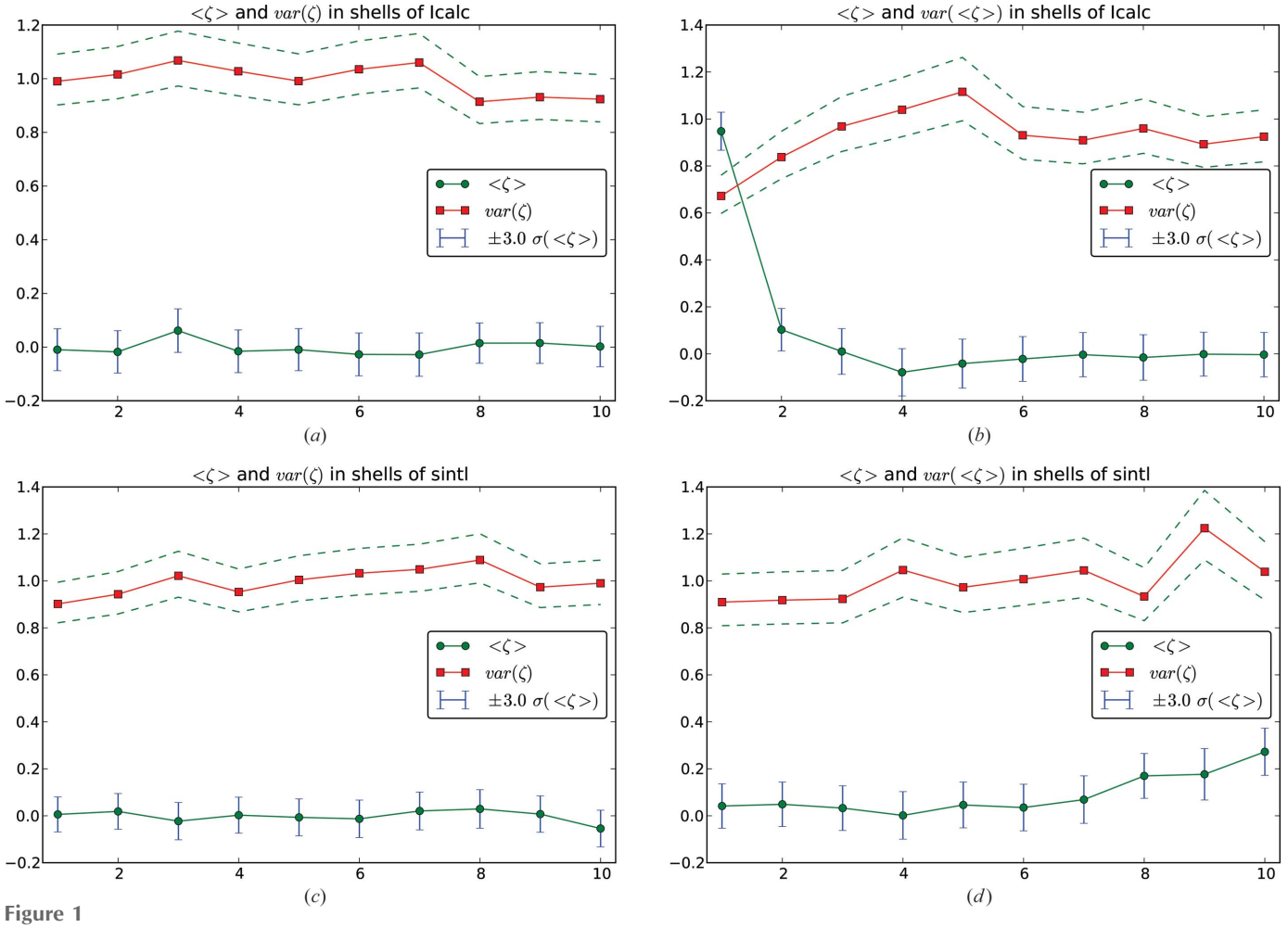


Figure 1
Mean values (green line with blue error bars) and variances (red line with dashed confidence interval according to $\alpha = 0.01$) of residuals $\zeta = (I_o - I_c)/[\sigma(I_o)]$ in bins of equal strength sorted by the calculated intensity (top) and by $\sin\theta/\lambda$ (bottom) of artificial data without (left) and with significance cutoff $I_o/\sigma(I_o) \geq 3$ (right). The same data set was used in both cases.

2.2. Variances of binned residuals

For X being normally distributed with parameters μ and σ the new variable $\chi^2 = (n-1)(\sigma_i^2/\sigma^2)$ with population variance σ_i^2 of the ith bin with strength n follows a χ^2 distribution with $m = n - 1$ degrees of freedom. For a given level of significance α a lower threshold value χ_l^2 and a higher threshold value χ_h^2 is given within which the value χ^2 is found with probability $1 - \alpha$: $\chi_l^2 = \chi_{1-\alpha/2;n-1}^2$ and $\chi_h^2 = \chi_{\alpha/2;n-1}^2$. The confidence interval is given by (Semendjajew *et al.*, 2012)

$$\frac{(n-1)\sigma_i^2}{\chi_{\alpha/2;n-1}^2} \leq \sigma^2 \leq \frac{(n-1)\sigma_i^2}{\chi_{1-\alpha/2;n-1}^2}. \qquad (1)$$

For applications it is convenient to plot the bin variances together with the confidence interval after specification of α. All bin variances should be consistent with the assumption of having the same population variance, *i.e.* it should be possible to draw a constant value that lies completely in the confidence interval. To be tolerant of outliers, α should be chosen small. For a start, we use $\alpha = 0.01$.

2.3. Validity of the refinements

From the point of view of the present statistical analysis method, the refinements are invalidated the more residual bin mean values are inconsistent with the value zero, the more residual bin mean values are inconsistent with each other, and the more residual bin variances are not consistent with each other. This is because these events make it unlikely that all residuals belong to the same population, which is characterized by only one mean value (hopefully close to zero) and only one variance (preferably close to one). When the residual bin variances are approximately the same around a value different from one this would mean that all s.u.'s are too small or too large and this could be resolved by scaling the s.u.'s. It should be stressed, however, that distorted residual statistics need not lead to invalid model parameter estimates; this has to be investigated in each individual case.

2.4. Cumulative plots of rare events

The events in the tails of the normal distribution are considered to be rare events, because they appear with a low frequency. It is convenient to measure the distance in multi-

ples of the standard uncertainty. Events $|\zeta| > 3$ have a probability of 0.27%, events $|\zeta| > 4$ of only 0.006%. It is also convenient to count the total number of rare events. When the data are sorted according to intensity or resolution and the total number of rare events up to a given resolution or to a given intensity are counted, the resulting plot is a (non-normalized) cumulative plot of rare events. This should result in a linearly increasing line, as there should be an equal probability for each reflection to result in a rare event, when the model is adequate and no systematic errors apply. A convex plot indicates that rare events are more frequently connected to the reflections in the beginning of the plot, *e.g.* low-resolution reflections. A concave shape indicates that the rare events originate more frequently from the reflections in the end of the plot, *e.g.* high-intensity reflections. Steps in the plots indicate a region with a very high frequency of rare events. Cumulative distributions are widely used in stochastics, for example as an antiderivative of probability density functions, and in the Kolmogorov–Smirnov test [see, for example, Von Mises (1964)].

3. Application to artificial data

3.1. Effects of a significance cutoff

The artificial data set 24 from Henn & Meindl (2014*b*) serves as a reference for a data set without any systematic errors (no cutoff applied, left column of Fig. 1) and for a data set with a well known single systematic error [application of cutoff $I_o/\sigma(I_o) \geq 3$ in least-squares refinement, right column of Fig. 1]. The respective BayCoN plots are in the cited literature, too (see Figs. 2 and 4 in the cited literature). The data were sorted according to increasing calculated intensities (top) and resolution (bottom). The level of significance for the calculation of the confidence interval of the population variances is $\alpha = 0.01$ in all cases. The left column shows roughly constant residual mean values and variances as expected for a data set without systematic errors. The mean values of the residuals in each bin are consistent with each other bin and with the value zero at a 3σ level. The variances of the residuals in each bin are consistent with each other and with the value one at an $\alpha = 0.01$ level of significance. At the given level of significance it is therefore justified to assume that the residuals of all bins follow the same distribution in Figs. 1(*a*) and 1(*c*).

Application of a significance cutoff $I_o/\sigma(I_o) \geq 3$ changes the situation (Fig. 1, right column). The mean values of the residuals are not consistent with each other any more, nor are all of them consistent with the value zero. This is particularly the case for the lowest calculated intensities (left part of Fig. 1*b*) and for the largest resolution shells (right part of Fig. 1*d*), which are increasingly composed of weak reflections. The cutoff criterion allows for those weak reflections that have accidentally a large I_o value, which is also correlated with a positive residual ζ, and rejects those weak reflections which accidentally have an I_o value not much larger than the respective I_c value. The situation is sketched in Fig. 1 of Hirshfeld & Rabinovich (1973). The cutoff leads to a significantly positive population of residuals for the weakest calculated intensities and to a significantly reduced variance. That model parameters are affected by a significance cutoff in real data sets was also shown by Arnberg *et al.* (1979) and is discussed in more detail in Watkin (2008).

The assumption that the residuals from all bins belong to the same distribution must be rejected at the given level of significance. Therefore, the totality of residuals (at the given level of significance) does not follow a distribution with equal probabilities for positive and negative deviations from the mean value in all resolution and intensity regions any more, *i.e.* the mean value of the residuals ceases to be unbiased on the true intensity. This systematic error may lead to model parameter estimates that are far from the unbiased value and far outside the range given by the calculated model parameter errors. A further regrettable effect of the application of the significance cutoff is that the detection of other systematic errors leading to increased mean values of residuals for the weak intensities and for high resolution may be obstructed.

The application to artificial data not suffering from systematic errors shows that the concept of equal mean values and variances for residuals applies and that the equality of mean values and variances within statistical fluctuations is easily destroyed by application of a significance criterion in the least-squares refinement. This effect is seen in the experimental data also, as will be discussed later.

Table 1
Original and modified values of five individually manipulated reflections for the simulation of measurement errors.

	h	k	l	I_o	$\sigma(I_o)$	$\sin\theta/\lambda$
Original	2	0	0	12788.78	127.24	0.0976
Modified	2	0	0	6788.78	127.24	0.0976
Original	2	1	0	2756.80	29.11	0.1136
Modified	2	1	0	2256.80	29.11	0.1136
Original	3	0	0	7518.50	76.95	0.1465
Modified	3	0	0	5518.50	76.95	0.1465
Original	1	1	0	4142.62	42.82	0.0758
Modified	1	1	0	1642.62	42.82	0.0758
Original	2	3	0	2341.25	25.28	0.1995
Modified	2	3	0	2041.25	25.28	0.1995

3.2. Effects of distorted individual reflections

Modifications of data set 24 from Henn & Meindl (2014*b*) were taken for studying well known systematic errors with the help of artificial data. In each case the data were modified in order to introduce a systematic error and the modified data were taken as observed intensities for a least-squares refinement. In the first case, five reflections from the lowest resolution shell were modified according to the list in Table 1 to simulate systematic measurement errors in the data.

The manipulated reflections lead to increased variances of the residuals (see right column in Fig. 2) in the respective bins and to $R^{\rm meta} = 17.2\%$; however, the n.p.p. does not differ significantly from that of the error-free reference data (for the n.p.p. of the error-free data see Fig. 3 K3 of the supporting

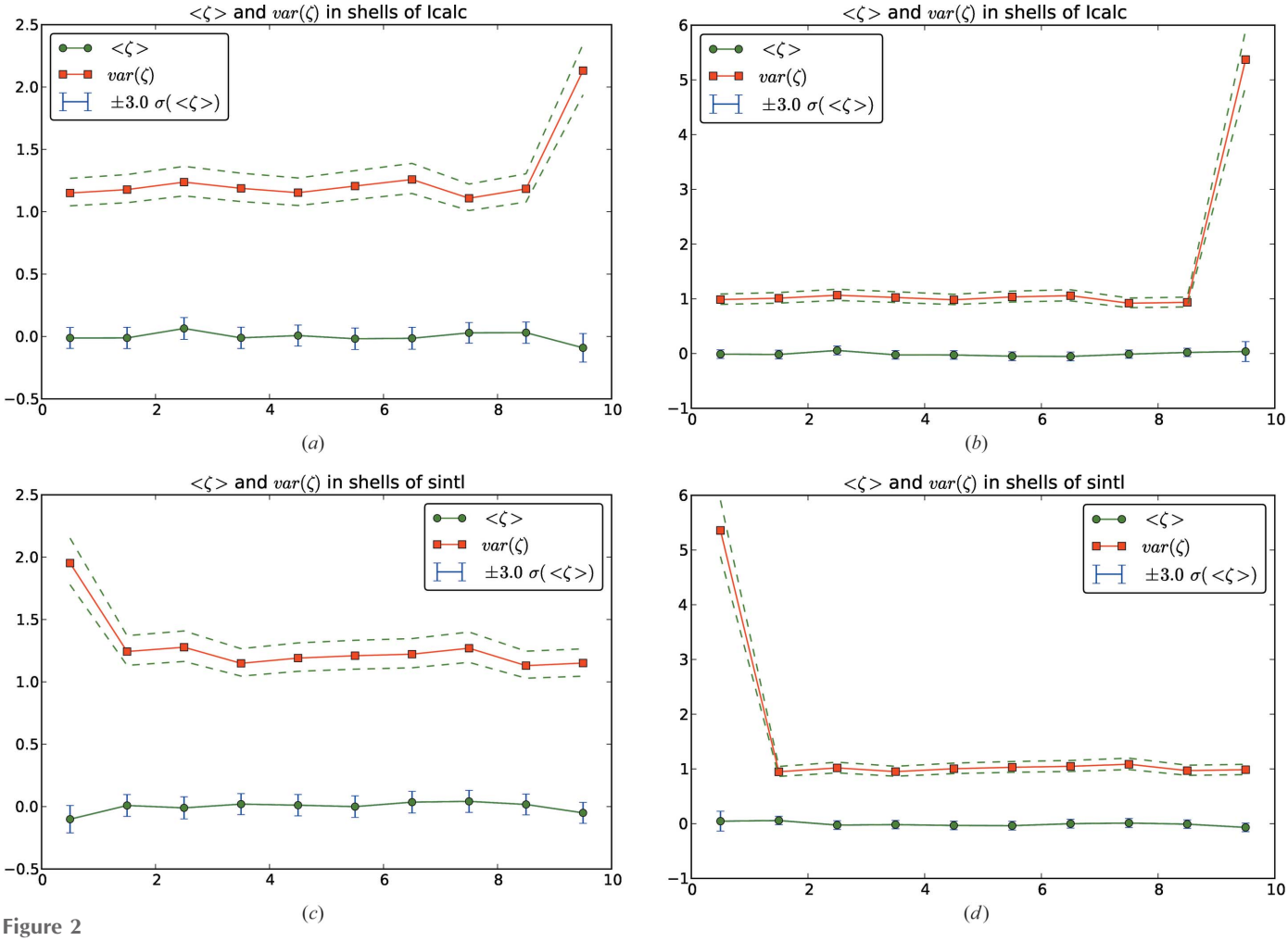


Figure 2
Effects of systematic errors. Mean values (green line with blue error bars) and variances (red line with dashed confidence interval according to $\alpha = 0.01$) of residuals $\zeta = (I_o - I_c)/[\sigma(I_o)]$ in bins of equal strength sorted by the calculated intensity (top) and by $\sin\theta/\lambda$ (bottom) of artificial data with intentionally flawed s.u. values according to equation (2) (left column) and artificial data with five intentionally flawed intensity observations according to Table 1 (right column).

information, for the n.p.p. of the data set with five intentionally flawed reflections see Fig. 4 K2 of the supporting information).

3.3. Effects of distorted s.u. values

To study the effect of distorted s.u. values on the distribution of residuals, the intensity data of data set 24 were used and the corresponding s.u. values were changed according to

$$\sigma \rightarrow \frac{\sigma}{p_1\,\sigma + 1} \qquad (2)$$

with $p_1 = 0.01$. A similar but much stronger transformation with $p_1 = 0.5$ was studied in Henn & Meindl (2014*b*). The transformation applied here is modest as the abundant weak s.u.'s remain virtually unchanged; only the largest s.u. values are affected. The transformation leads to a systematically increasing underestimation of the large s.u.'s, the larger these are. The effects on the mean values and variances of residuals (see Fig. 2, left column) are qualitatively similar to those of distorted reflections (see Fig. 2, right column): in the respective bins of intensities and resolution, the variances of residuals increase. A side effect of the distortion of s.u. values as given by equation (2) is again [as in Henn & Meindl (2014*b*)] that it reduces the actual as well as the predicted R values: whereas the noise present in the I_o is equivalent to $wR(F^2) = 0.036$, the actual R value becomes 0.025 and the prediction based on the transformed s.u.'s is 0.022. This is an important point, because it shows that by underestimating the large s.u. values lower R values can be achieved as is expected from the appearance of the squared significance in the denominator for the predicted R values. This artificial lowering of R values is at the expense of distorted model parameter estimates, flawed model parameter s.u.'s and a distorted distribution of residuals. It is rather delicate in this context that the underestimated s.u. values force the I_c in the proximity of the I_o such that an I_o *versus* I_c plot appears to show a very good fit. That this fit is not based on a Gaussian distribution is seen from the normal probability plot, which now deviates from the expected distribution in the periphery

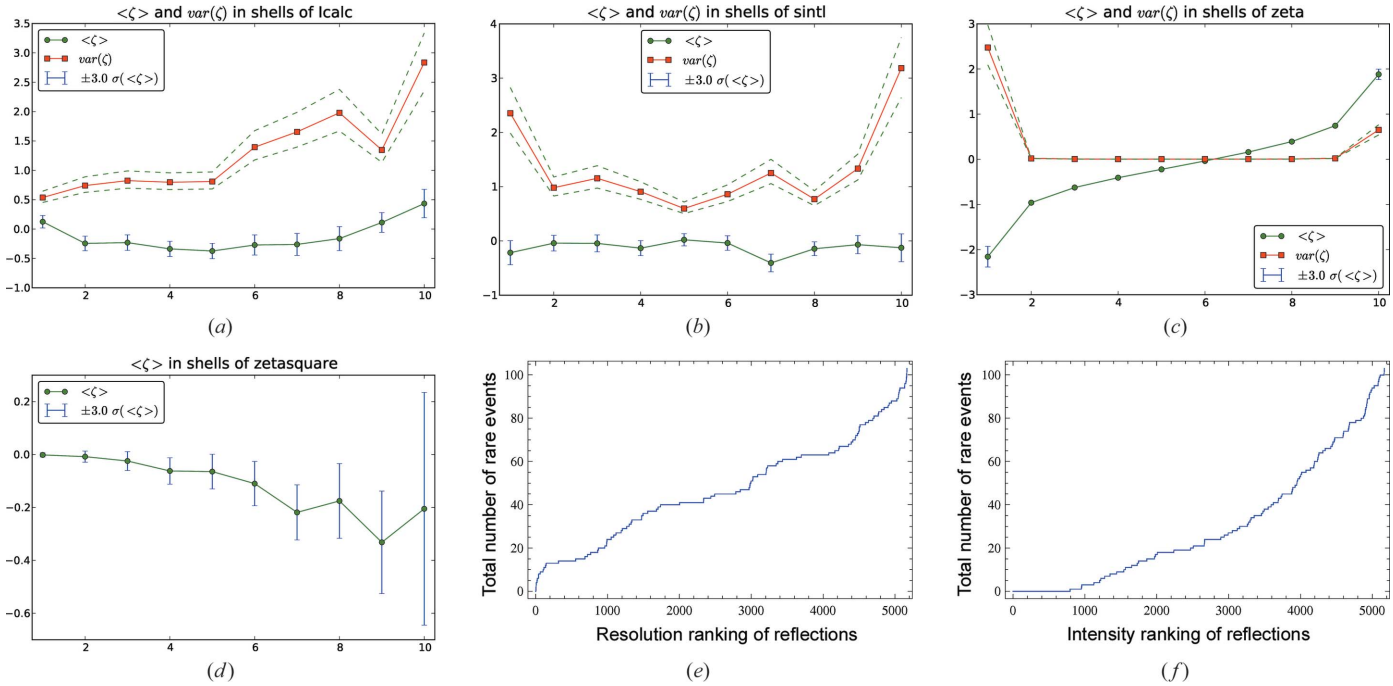


Figure 3
Effects of systematic errors. Mean values (green line with blue error bars) and variances (red line with dashed confidence interval according to $\alpha = 0.01$) of residuals $\zeta = (I_o - I_c)/[\sigma(I_o)]$ in bins of approximately equal strength sorted by increasing I_c (*a*), $\sin\theta/\lambda$ (*b*), residuals ζ (*c*) and squared residuals ζ^2 (*d*) of data set 5. Cumulative rare events $|\zeta| > 3$ sorted according to increasing $\sin\theta/\lambda$ (*e*) and I_c (*f*).

(see Fig. 4 K in the supporting information) resulting in too many rare events. Normal probability plots with deviations in the periphery are very frequently observed in crystallographic applications. The underlying reasons for these outliers should be identified. A plot of residuals *versus* resolution shows a broadening of the graph for low resolution in the case of underestimated large s.u.'s (see Fig. 5*c* in the supporting information). These are responsible for the large variances of residuals depicted in Fig. 2 in the respective bins. Outliers in plots of I_o *versus* I_c are not detected in this case (Fig. 5*d* in the supporting information), whereas in the case of correct s.u. values and individual artificially flawed intensity observations, these appear as outliers in plots of residuals *versus* resolution (Fig. 5*e* in the supporting information) *and* in plots of I_o *versus* I_c (Fig. 5*f* in the supporting information), at least the strong ones. The appearance of outliers in plots of residuals *versus* resolution *not* accompanied by outliers in plots of I_o *versus* I_c may therefore be taken as a characteristic of systematic errors caused by distorted s.u.'s.

4. Application to experimental data

If not indicated otherwise, we refer to the data sets as given in Henn & Meindl (2014*a*) with their respective data-set numbers. These data sets have been analysed by different methods in the literature (Henn & Meindl, 2014*a*,*b*). Normal probability plots, plots of observed *versus* calculated intensities and of residuals *versus* resolution as well as BayCoN plots are found in the cited literature and the respective supporting information.

4.1. A refinement with systematic errors

As a first example, data set 5 is discussed as an experimental data set known to be contaminated with systematic errors. The reflections $01\bar{1}$ and $\bar{2}00$ are identified as outliers in an I_o *versus* I_c plot (data not shown). Additionally, no weighting scheme was used such that this data set most likely suffers from underestimated large s.u. values.

Some of the residual mean values are not consistent with each other and most are additionally not consistent with the value zero at the 3σ level of significance (see Fig. 3*a* and Fig. 1, first column, in the supporting information). The assumption of underestimated large s.u. values is consistent with the increasing variance of the residuals sorted by the calculated intensity. The variances consequently differ significantly from each other at the $\alpha = 0.01$ level of significance.

When the data are sorted according to increasing resolution (Fig. 3*b*), the residual mean value of bin 7 is inconsistent with the assumption of a zero mean value at a 3σ level of significance. The variance of residuals shows distinct changes with the resolution ('u' shape). The large variance in the lowest resolution shell is influenced by the outliers; omitting the outliers $01\bar{1}$ and $\bar{2}00$ leads to a reduction (compare C1 and C2 in Fig. 1 of the supporting information). The bin values of the variances are not consistent with each other and those from bins 1, 5, 7, 8 and 10 are not consistent with the value one.

Sorting the residuals in increasing order (of the residuals) shows that the variances of the extreme residuals are significantly larger than those of the remaining residuals (Fig. 3*c*) and that this is asymmetric with respect to positive and negative residuals.

Sorting the residuals according to their squared value shows that the mean values of the residuals in bins 4, 6, 7, 8 and 9 are not consistent with a zero mean value at the 3σ level of significance. Instead, the residuals systematically tend to negative values (Fig. 3*d*).

The cumulative number of rare events $|\zeta| > 3$ plotted against the reflections sorted according to increasing resolution (Fig. 3*e*) shows an initial step for the lowest resolution range and a final step for the highest resolution range. Steps in these plots indicate a concentration of rare events, which should be distributed uniformly over the whole resolution range. There are 103 rare events $|\zeta| > 3$ for 5136 reflections, a fraction of 2%.

That rare events $|\zeta| > 3$ are dependent on the intensity is demonstrated by a plot of cumulative rare events plotted against the reflections sorted according to increasing calculated intensities (Fig. 3*f*). The slope of this plot tends to increase with increasing intensity, *i.e.* the larger the intensity the more frequently rare events appear. This behaviour is again in accordance with the assumption of too small large s.u. values.

These characteristics found for data set 5 appear similarly in data set 6 and, with modifications, in data set 7. Regarding the mean values of residuals, in all three cases the following observations apply:

(i) One or more bins with residual mean values are inconsistent with each other at a 3σ level of significance, when sorted against resolution.

(ii) One or more bins with distinctly negative residual values are inconsistent with the assumption of a value zero at a 3σ level of significance, when sorted against resolution.

(iii) Many bin mean values of residuals are inconsistent with each other when sorted against increasing calculated intensity.

(iv) Many bin mean values of residuals are inconsistent with the value zero when sorted against increasing calculated intensity.

(v) A characteristic 'u' shape of bin mean values of residuals appears, when sorted against the calculated intensity.

(vi) Residual bin mean values inconsistent with each other are observed when sorted in ascending order of ζ^2.

(vii) Residual bin mean values tend to negative values when sorted in ascending order of ζ^2.

Regarding the variances of residuals:

(i) Variances of residuals are inconsistent with each other in shells of resolution.

(ii) Variances are significantly increased for the lowest resolution bin.

(iii) Variances of residuals tend to increase with the calculated intensity.

Regarding rare events $|\zeta| > 3$:

(i) Rare events emerge with a higher frequency from low-resolution parts as indicated by steps in the cumulative rare events plots against resolution.

(ii) Rare events tend to originate more frequently from higher intensities as indicated by the increasing slopes of cumulative rare event plots against intensity.

These are only some common features of all three data sets; each point indicates that the residuals of each individual data set are not consistent with the other residuals of the same data set.

4.2. A refinement with few systematic errors

For an example of a data set with few systematic errors, data set 8 is discussed in analogy to the section before. This data set corresponds to an anharmonic motion model refinement of the explosive RDX at 20 K (Zhurov *et al.*, 2011). With increasing data quality, anharmonic motion models become more and more important in charge-density studies (see, for example, Meindl *et al.*, 2010; Paul *et al.*, 2011; Herbst-Irmer *et al.*, 2013; Poulain *et al.*, 2014; Jarzembska *et al.*, 2014; Pinkerton *et al.*, 2014, and many more).

Some of the residual mean values of residuals ordered according to increasing calculated intensities are not consistent with each other but most show an overlap and most are consistent with the value zero at the 3σ level of significance (see Fig. 4*a* and Fig. 2, first column, in the supporting information). The increased residual mean value for the lowest calculated intensities is presumably a consequence of the significance cutoff. Also the initial increasing variance of the residuals sorted by the calculated intensity may be influenced by the significance cutoff (compare with Fig. 1*b*). Some of the variances differ significantly from each other at the $\alpha = 0.01$ level of significance; however, the whole range of variances lies between approximately 0.7 and 1.5, *i.e.* the range of variances is much smaller as compared to Fig. 3(*a*).

When the data are sorted according to increasing resolution (Fig. 4*b*), all but the residual mean value of bin 1 are consistent with the assumption of a zero mean value at a 3σ level of significance. The variance of residuals shows a tendency to increase with the resolution, but all variances are consistent with the value one.

Sorting the residuals in increasing order (of the residuals) shows that the variances of the extreme residuals are slightly larger than those of the remaining residuals (Fig. 4*c*) and that all values are well below one. The increase is symmetric with respect to positive and negative residuals. This effect of slightly increased variances for the extreme residuals is also observed for artificial data (see Fig. 3, column 3, in the supporting information) and is dependent on the bin width.

Sorting the residuals according to their squared value shows that the mean values of all residual bins but No. 7 are consistent with a zero mean value at the 3σ level of significance. There is also a tendency to positive values for larger absolute residuals (Fig. 4*d*).

The cumulative number of rare events $|\zeta| > 3$ plotted against the resolution (Fig. 3*e*) also shows a little initial step for the lowest resolution range. There are in total 53 rare events for 8057 reflections (0.7%). Steps in these plots indicate a concentration of rare events, which should be distributed uniformly over the whole resolution range. A large fraction of those rare events $|\zeta| > 3$ originate from reflections at positions 4500–5500, when sorted in increasing order (Fig. 3*f*). This

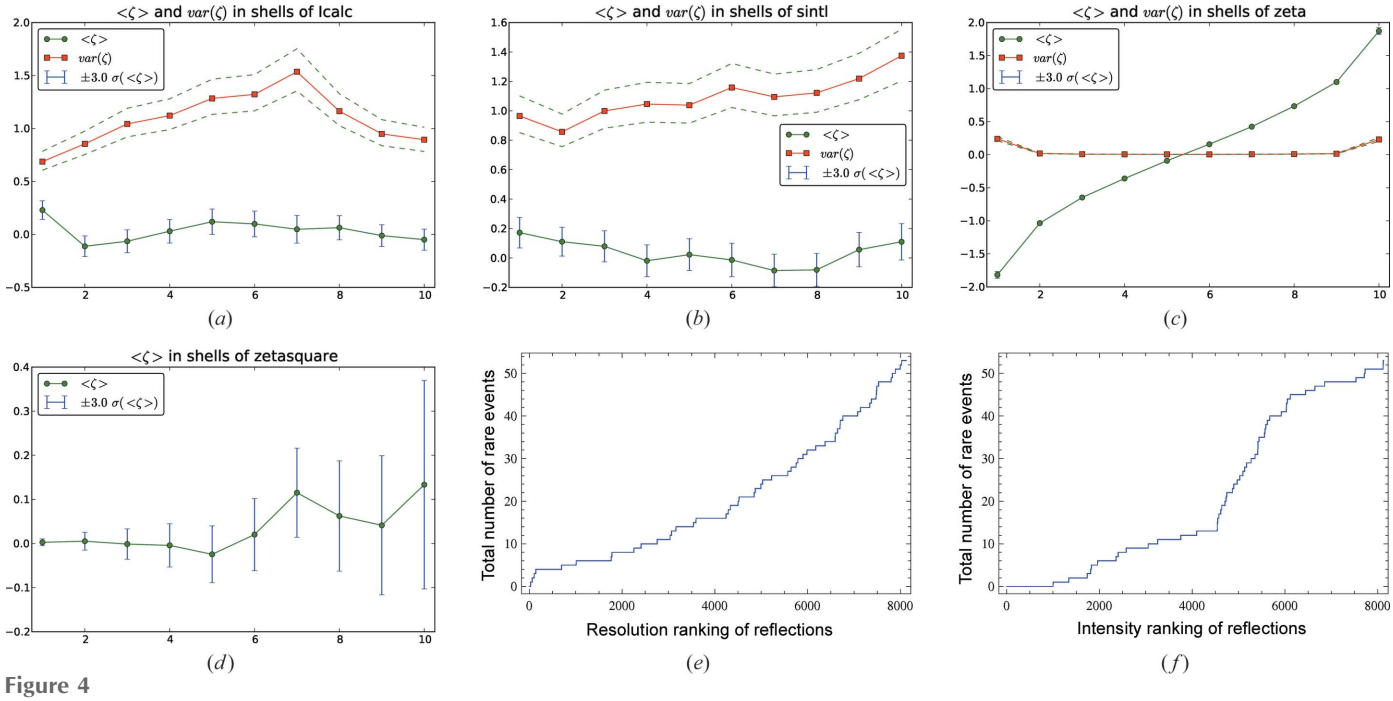


Figure 4
Effects of systematic errors. Mean values (green line with blue error bars) and variances (red line with dashed confidence interval according to $\alpha = 0.01$) of residuals $\zeta = (I_o - I_c)/[\sigma(I_o)]$ in bins of approximately equal strength sorted by increasing I_c (*a*), $\sin\theta/\lambda$ (*b*), residuals ζ (*c*) and squared residuals ζ^2 (*d*) of data set 8. Cumulative rare events $|\zeta| > 3$ sorted according to increasing $\sin\theta/\lambda$ (*e*) and I_c (*f*).

corresponds to intensity values 90–150 and to the whole resolution range. There are many similarities and only subtle differences in the diagnosis plots for data sets 8 and 9, which correspond to the same experimental data and different thermal nuclear motion models as well as different error models, as can be seen from columns 1 and 2 in Fig. 2 of the supporting information. A remarkable difference between the models is that the high frequency of rare events at very high and at very *low* resolution is significantly reduced by the anharmonic motion model (data set 8) as compared to the harmonic nuclear motion model (data set 9). Anharmonic motion models are expected to have a strong impact on high-resolution data (Kuhs, 1988, 1992), although it has been shown recently that data of considerably lower resolution (0.6 Å^{-1}) are affected, too (Herbst-Irmer *et al.*, 2013). The present case may indicate that the effects of anharmonic nuclear motion are detectable by cumulative rare event plots even in the lowest resolution shells. Clarification of this important point needs more systematic investigations in the future.

4.3. Comparison between the two refinements

When *e.g.* the residual mean values are inconsistent with each other despite a significance level that is tolerant of outliers, one could use the term that they are self-contradicting. In this sense, data set 5 is highly self-contradicting, whereas data set 8 is less self-contradicting, because from the ten bins of residual mean values sorted by the calculated intensities for data set 5, only bins 8 and 9 and probably bin 1 are in accordance with a value zero (Fig. 3*a*) whereas for data set 8 only one bin is clearly inconsistent with the value zero (Fig. 4*a*) and this inconsistency is likely to be connected to the applied significance cutoff.

5. Summary, outlook and conclusion

Statistical tests based on the equality of sample means and variances of residuals have been applied to artificial and experimental diffraction data from the literature. The application to artificial data without systematic errors proves the applicability of the concepts. Application to artificial data with significance cutoff showed how the significance cutoff leads to residuals significantly shifted to positive values and to a reduced variance of residuals in the bins of lowest intensities. These shifts necessarily destroy the normal distribution that might or might not be present without application of a significance cutoff. It was also shown, with the help of artificial data, how intentionally flawed reflections increase the variance of residuals in the respective bins, similar to the case of underestimated large s.u. values. An important difference between these two cases is that in the latter the normal probability plot shows deviations in the periphery and that the former shows outliers in plots of both I_o *versus* I_c and ζ *versus* $\sin\theta/\lambda$, whereas the latter only shows outliers in ζ *versus* $\sin\theta/\lambda$ plots. Plots of cumulative rare events, that ideally show a constant slope, proved to be useful to identify ranges of resolution or intensities with increased and with reduced probabilities of generating rare events. Application to experimental data showed in one case strong variations of residual means and variances with respect to intensities and resolution, distinctly and unsymmetrically increased variances for the extreme residuals, steps in the cumulative rare event

plot sorted by resolution, and a concave shape of the cumulative rare event plot sorted by intensity. A part of this can be explained by underestimated s.u. values; however, it is likely that more than one source of systematic errors applies here. It was stressed that data sets of the same structure measured at different temperatures show striking similarities. This may be expected from the structures; however, it is assumed that these similarities are also caused by similar data-processing steps. The application to experimental data in another case showed a behaviour of the descriptors much closer to the expected ideal behaviour; however, also in this case the residuals tend to slightly increased variances for increasing resolution, though on a distinctly reduced scale. Also the systematic error caused by application of a significance cutoff is clearly observed. A step in the cumulative rare event plot sorted with respect to the resolution is observed for very low resolution. This step is larger for the harmonic motion model. The developed methods can easily be generalized to include plots of residuals *versus* s.u. values or significance, *i.e.* there is potential for future developments.

There are now plenty of tools and techniques available for the detection and identification of systematic errors. We hope that this topic will attract more attention in the future. All of these old and new tools are useless when they are not applied.

References

Abrahams, S. C. & Keve, E. T. (1971). *Acta Cryst.* A**27**, 157–165.
Arnberg, L., Hovmöller, S. & Westman, S. (1979). *Acta Cryst.* A**35**, 497–499.
Brünger, A. T. (1992). *Nature (London)*, **355**, 472–475.
Henn, J. & Meindl, K. (2010). *Acta Cryst.* A**66**, 676–684.
Henn, J. & Meindl, K. (2012). *Acta Cryst.* A**68**, 304.
Henn, J. & Meindl, K. (2014*a*). *Acta Cryst.* A**70**, 248–256.
Henn, J. & Meindl, K. (2014*b*). *Acta Cryst.* A**70**, 499–513.
Henn, J., Meindl, K., Oechsner, A., Schwab, G., Koritsanszky, T. & Stalke, D. (2010). *Angew. Chem.* **122**, 2472–2476.
Henn, J. & Schönleber, A. (2013). *Acta Cryst.* A**69**, 549–558.
Herbst-Irmer, R., Henn, J., Holstein, J. J., Hübschle, C. B., Dittrich, B., Stern, D., Kratzert, D. & Stalke, D. (2013). *J. Phys. Chem. A*, **117**, 633–641.
Hirshfeld, F. L. & Rabinovich, D. (1973). *Acta Cryst.* A**29**, 510–513.
Jarzembska, K. N., Kamiński, R., Dobrzycki, Ł. & Cyrański, M. K. (2014). *Acta Cryst.* B**70**, 847–855.
Kuhs, W. F. (1988). *Aust. J. Phys.* **41**, 369–382.
Kuhs, W. F. (1992). *Acta Cryst.* A**48**, 80–98.
Meindl, K. & Henn, J. (2008). *Acta Cryst.* A**64**, 404–418.
Meindl, K. & Henn, J. (2012). *Electron Density and Chemical Bonding II*, edited by D. Stalke, Vol. 147 of *Structure and Bonding*, pp. 143–192. Berlin: Springer.
Meindl, K., Herbst-Irmer, R. & Henn, J. (2010). *Acta Cryst.* A**66**, 362–371.
Paul, A., Kubicki, M., Jelsch, C., Durand, P. & Lecomte, C. (2011). *Acta Cryst.* B**67**, 365–378.
Pinkerton, A., Zhurov, V. & Zhurova, E. (2014). *Acta Cryst.* A**70**, C1550.
Poulain, A., Wenger, E., Durand, P., Jarzembska, K. N., Kamiński, R., Fertey, P., Kubicki, M. & Lecomte, C. (2014). *IUCrJ*, **1**, 110–118.
Semendjajew, K. A., Bronstein, I. N., Musiol, G. & Mühlig, H. (2012). *Taschenbuch der Mathematik*. Frankfurt am Main: Harri Deutsch.
Von Mises, R. (1964). *Mathematical Theory of Probability and Statistics*. New York: Academic Press.
Watkin, D. (2008). *J. Appl. Cryst.* **41**, 491–522.
Zhurov, V. V., Zhurova, E. A., Stash, A. I. & Pinkerton, A. A. (2011). *Acta Cryst.* A**67**, 160–173.