Comparison of maximum likelihood estimation and chi-square statistics applied to counting experiments

ARTICLE *in* NUCLEAR INSTRUMENTS AND METHODS IN PHYSICS RESEARCH SECTION A ACCELERATORS SPECTROMETERS DETECTORS AND ASSOCIATED EQUIPMENT · JANUARY 2001

Impact Factor: 1.22 · DOI: 10.1016/S0168-9002(00)00756-7

ELSEVIER

Nuclear Instruments and Methods in Physics Research A 457 (2001) 384–401

NUCLEAR INSTRUMENTS & METHODS IN PHYSICS RESEARCH Section A

www.elsevier.nl/locate/nima

Comparison of maximum likelihood estimation and chi-square statistics applied to counting experiments

T. Hauschild*, M. Jentschel

Institut Laue-Langevin, Rue Jules Horowitz, BP 156, F-38042 Grenoble, France

Received 25 April 2000; received in revised form 15 June 2000; accepted 29 June 2000

Abstract

Five different statistics are compared with respect to parameter, error, and goodness-of-fit estimation in the case of counting experiments. In particular, maximum likelihood approaches are opposed to chi-square techniques. It could be shown that the maximum likelihood estimation derived for Poisson distributed data (Poisson MLE) produces the best statistic in order to estimate parameters. If goodness-of-fit estimations are to be done, Pearson's chi-square should be used. It is the only statistic that leads to the correct expectation value for chi-square. All the other statistics do not follow a chi-square distribution. It is discussed that the chi-square per degree of freedom is not well suited for judging the consistency of a model and the data. When estimating the mean of Poisson distributed data or the area under a peak, Poisson MLE was shown to be the only statistic that comes to consistent and unbiased results, two other statistics give asymptotically consistent results. The widely used Neyman's chi-square fails in all cases. Further, artificial Poisson distributed data have been created on the basis of known model functions. It is shown and discussed in which cases chi-square techniques fail to extract the correct parameter values and where they still can be used. Special emphasis is put on the evaluation of Doppler-broadened gamma line shapes as they are measured in the Crystal-GRID technique.

PACS: 02.50

Keywords: Maximum likelihood estimation; Chi-square statistics; Goodness-of-fit estimation; Counting experiment; γ-Ray spectroscopy; Crystal-GRID

1. Introduction

In the past five years the Crystal-GRID technique has been developed as an experimental tool to study interatomic potentials in solids at energies of several hundreds of eV [1–3]. In this technique one measures Doppler-broadened line shapes of γ-rays emitted from moving atoms in the solid. These line shapes are compared to theoretical profiles, which were extracted from computer simulations of atomic motion. During the evaluation of the data the theoretical model underlying the computer simulations is modified such that the best agreement of experiment and theory is achieved. This procedure brings about the more general question of the reliability of extracted information and thus, the question of parameter estimation in counting experiments in general.

* Corresponding author.

E-mail addresses: hauschild@ill.fr (T. Hauschild), jentsch@ill.fr (M. Jentschel).

0168-9002/01/$ - see front matter
PII: S0168-9002(00)00756-7

Estimating parameters is one of the main duties when evaluating any experimental data. Models are used to interpret the experimental findings. If a model is free of parameters, the experiment can be directly used to check the accuracy of the model. Otherwise, the comparison of the parameter-dependent model and the experimental results may lead to a determination of these parameters. The model function is then called "fit function" and the result of the "fit" is the best value for the parameter(s). In order to judge which parameters are best, a so-called "statistic" is used. A statistic is a measure of the consistency of the model and the experimental data.

Most often least-squares statistics are applied, e.g. one of the well-known χ^2 methods. A general overview of these methods will be presented in Section 2.1. Problems concerning their applicability have been reported throughout the literature and some of them will be summarised in Section 4. In order to cope with these problems, a new χ^2 method has recently been developed [4].

A second approach is the use of maximum likelihood estimation (MLE). These statistics maximise the likelihood that the experimental data is a random sample, that could have been drawn from the model. They have been studied in varying contexts, but are still not commonly applied. An overview will be given in Sections 2.2 and 2.3. Maximum likelihood approaches are numerically more costly than χ^2 methods. With modern computers, however, this difference becomes meaningless.

Both approaches will be compared in this paper. The discussion will be focused on counting experiments. A priori, it is not clear, which approach is better suited in this case. The measured counts are Poisson distributed. This distribution differs significantly from the Gauss (normal) distribution, if the total number of counts per measurement is very small, as may be the case here. The χ^2 methods, however, are only valid for normally distributed data [5]. In Sections 4 and 5 it will be studied under which circumstances the often used χ^2 methods can be used.

In the particular case of the Crystal-GRID technique, the model does not allow an analytic expression of the fit function to be established. It is generated by Molecular Dynamics simulations instead. Consequently, the dependence on the parameters is not known analytically. The fit function can be described as a structured peak on a statistical background, where the background may be as low as 0.5 counts. Total counts per measurement of 0 can occur. The overall size of the function can be varied by one parameter. Often, especially in astronomy, one merely looks at the total peak intensity, which is present on the statistical background. In this work, however, the entire shape of the line profile is important. In Section 5 the different statistics will be applied to the Crystal-GRID method. It will be shown, under which conditions, which statistics lead to accurate estimates of the parameters under investigation.

1.1. Using statistics

The estimation of parameters is the main aim of fitting model functions to experimental data. It is called "point estimation" in statistics literature. Two more quantities can be estimated from the fit in order to get an idea about the reliability of the parameters. First, a measure of the errors or uncertainties needs to be determined using the so-called "interval estimation". A typical measure is the standard deviation σ of the parameter.

Point and interval estimation, respectively, parameter and error estimation are sufficient if the underlying model is known to be correct. Otherwise a statement on the validity of the model is needed, the so-called "goodness-of-fit" needs to be calculated. In statistics this is called the "test of hypothesis". It tells how probable it is that the data have occurred if the assumed model were true.

1.2. Definitions and notation

In the following, a counting experiment with a crystal spectrometer is considered, where incoming photons are counted as a function of the angular position x of a diffraction crystal corresponding to an energy $E(x)$. The expected counts depend on this position. The following notations are used:

c_i = number of counts in measurement i
$\boldsymbol{c}$ = $(c_1, c_2, \ldots, c_N)$ = set of measured count values

$N =$ number of measured values
$x_i =$ angular position of diffraction crystal in measurement i

The aim is to fit a theoretically predicted function to the data.

$m_{\boldsymbol{a}} =$ model function, numerically derived from the theory, depending on parameters $\boldsymbol{a}$
$m_{\boldsymbol{a}}(x_i) =$ predicted value at position x_i, i.e. in measurement i
$m'_i =$ parameter-independent model value in measurement i
$\boldsymbol{m}' = (m'_1, m'_2, \ldots, m'_N) =$ set of parameter-independent model values
$\sigma_i =$ standard deviation (error) of measurement i
$t =$ unknown true function
$a_j =$ parameter that model function depends on
$\boldsymbol{a} = (a_1, a_2, \ldots, a_J) =$ set of parameters
$J =$ number of parameters

If only a mean value is to be calculated, the model function is constant and given by

$\mu =$ mean value.

The model function includes the background, as information would be lost if trying to subtract the background from the measured counts. Both the background signal and the decay-related signal are Poisson distributed. The measured signal as the sum of both contributions is therefore also Poisson distributed [6, p. 71].

2. Parameter and error estimation

2.1. χ^2 statistics

The classical χ^2 statistics base on the χ^2 density which is defined as

$$\chi^2 = \sum_i \frac{(c_i - m_i)^2}{\sigma_i^2} \tag{1}$$

where $(c_i - m_i)$ is normally (Gauss) distributed with the known variance σ_i^2. When counting random events, however, the number of counts per given time is not Gauss but Poisson distributed around the true value $t(x_i)$. It is necessary to check whether χ^2 statistics can be used, anyhow.

For Poisson distributed data, it can easily be shown that the variance is equal to the true value, i.e. $\sigma_i^2 = t(x_i)$. As this true value is not known, it must be approximated. Depending on what approximation one uses, the statistic can be represented by one of the classical chi-square forms:

$$\chi^2_{\mathrm{P}} = \text{Pearson's } \chi^2 = \sum_i \frac{(c_i - m_{\boldsymbol{a}}(x_i))^2}{m_{\boldsymbol{a}}(x_i)} \tag{2}$$

$$\chi^2_{\mathrm{N}} = \text{modified Neyman's } \chi^2 = \sum_i \frac{(c_i - m_{\boldsymbol{a}}(x_i))^2}{\max(c_i, 1)}. \tag{3}$$

As discussed by Baker and Cousins, misunderstanding can occur when speaking about χ^2 statistics due to different names used [7]. In this paper, we apply the names given in Eqs. (2) and (3). Originally, Neyman's χ^2_{N} has been defined with c_i in the denominator. As this expression is undefined if any measured value c_i is zero, the so-called "modified Neyman's χ^2" has been widely used [8–10]. If all c_i are non-zero, both the original and the modified Neyman's χ^2 are identical. Consequently, we will only study the modified χ^2_{N}.

These statistics have problems estimating the area under a peak or simply the mean of Poisson distributed values. This will be discussed in Section 4. Different efforts are made to circumvent this problem. Wheaton, e.g., proposes different weights in the special case where $m_{\boldsymbol{a}} \ll 1$ [10]. This proposal will not be analysed in this paper. An even further modified statistic has been proposed recently by Mighell [4]. His χ^2_γ statistic has been designed in order to find the correct mean value of a Poisson distribution. It is defined as

$$\chi^2_\gamma = \sum_i \frac{(c_i + \min(c_i, 1) - m_{\boldsymbol{a}}(x_i))^2}{c_i + 1}. \tag{4}$$

This statistic can be easily applied with a code based on the modified Neyman's χ^2_{N} by replacing all counts c_i by $c_i + \min(c_i, 1)$ in the data.

However, as Stoneking and Den Hartog state clearly, "for non-normal uncertainty distributions such as the Poisson distribution, minimising χ^2 does not maximise the likelihood that the fitted

parameters reflect the data" [5]. Sections 2.2 and 2.3 will be dedicated to the more appropriate maximum likelihood method.

2.2. Maximum likelihood estimation (*MLE*) applied to Poisson distributed data

An intuitive way of extracting information from measured data is to ask the question whether it is probable that the data have occurred given a set of parameters and a model. The idea of maximum likelihood estimation (MLE) is to maximise this probability, called "likelihood". In its simplest form it can be used to determine the expectation value of an observable if the latter has been measured several times.

The likelihood L is given by the combined probability of measuring a set of c_i counts at positions x_i where the model predicts $m_{\boldsymbol{a}}(x_i)$ counts

$$L = \prod_i P(c_i;\, m_{\boldsymbol{a}}(x_i)) \tag{5}$$

$P(c_i;\, m_{\boldsymbol{a}}(x_i))$ is the probability that a measurement gives c_i counts if the true value is $m_{\boldsymbol{a}}(x_i)$. If $m_{\boldsymbol{a}}$ depends on the J parameters a_j, then the likelihood L needs to be maximised with respect to these parameters in order to find their optimum values.

If the measurements are Poisson distributed, the corresponding probability $P_{\mathrm{P}}(c_i;\, m_{\boldsymbol{a}}(x_i))$ must be used in Eq. (5). In this case the Poisson likelihood L_{P} becomes [9, p. 111]

$$L_{\mathrm{P}}(\boldsymbol{a}) = \prod_i \frac{[m_{\boldsymbol{a}}(x_i)]^{c_i}}{c_i!} \mathrm{e}^{-m_{\boldsymbol{a}}(x_i)}. \tag{6}$$

It is easier to use the logarithm of this expression. Minimising $-2\ln L_{\mathrm{P}}$ leads to the same parameters as maximising L_{P}.

$$\begin{aligned} -2\ln L_{\mathrm{P}}(\boldsymbol{a}) &= 2\sum_i\, [m_{\boldsymbol{a}}(x_i) - c_i \ln m_{\boldsymbol{a}}(x_i) + \ln c_i!] \\ &= 2\sum_i\, [m_{\boldsymbol{a}}(x_i) - c_i \ln m_{\boldsymbol{a}}(x_i)] + \text{const.} \end{aligned} \tag{7}$$

A similar equation has been derived several times. In 1979 Awaya presented it as a "new method" and called it G [8]. In the same year Cash published his C statistic [11]. Further formulations can be found in Refs. [12], [13], [14, p. 191].

If an analytic expression of the model function $m_{\boldsymbol{a}}$ exists, it may be possible to calculate the parameters by solving the set of equations

$$\frac{\mathrm{d}\ln L_{\mathrm{P}}(\boldsymbol{a})}{\mathrm{d}a_j} = 0 \quad (j = 1, \ldots, J). \tag{8}$$

This has been done for example in Refs. [5,15,16]. Otherwise numerical methods are used to find the minimum.

2.3. Maximum likelihood estimation (*MLE*) applied to normally (*Gauss*) distributed data

The Poisson probability P_{P} is only defined for positive integer values. In the Gauss case, however, p_{G} is a probability density function defined on all real numbers:

$$\begin{aligned} &p_{\mathrm{G}}(c_i;\, m_{\boldsymbol{a}}(x_i), \sigma_i)\,\mathrm{d}c_i \\ &\quad = \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left[-\frac{(c_i - m_{\boldsymbol{a}}(x_i))^2}{2\sigma_i^2} \right] \mathrm{d}c_i. \end{aligned} \tag{9}$$

Hence, it is not automatically suited for counting experiments. However, Gauss statistics are an acceptable approximation for Poisson distributed data only if the number of counts is larger than approximately 5. In this case, the integer Gauss distribution P_{G} can be approximated by the Gauss probability density function p_{G} taken at the integer positions. Gauss statistics should be avoided if low numbers of counts occur frequently.

The Gauss probability $P_{\mathrm{G}} \simeq p_{\mathrm{G}}$ can be used in the definition of the likelihood (Eq. (5)). The likelihood L_{G}, i.e. the combined probability for observing a set of c_i counts at positions x_i from a Gauss distribution with theoretical values $m_{\boldsymbol{a}}(x_i)$ and standard deviations σ_i is then [9, p. 102/144]

$$L_{\mathrm{G}}(\boldsymbol{a}) = (2\pi)^{-N/2} \prod_i\, (\sigma_i^2)^{-1/2} \exp\left[-\frac{(c_i - m_{\boldsymbol{a}}(x_i))^2}{2\sigma_i^2} \right], \tag{10}$$

$$-2\ln L_{\mathrm{G}}(\boldsymbol{a}) = \sum_i \left[\frac{(c_i - m_{\boldsymbol{a}}(x_i))^2}{\sigma_i^2} + \ln \sigma_i^2 \right] + N\ln(2\pi). \tag{11}$$

Most text books and articles implicitly assume that the standard deviations σ_i are known and, thus, do not depend on the parameters $\boldsymbol{a}$ (e.g. Refs. [9], [18, p. 169]). Under these assumptions the Gauss log-likelihood $-2\ln L_{\mathrm{G}}$ becomes equal to the general χ^2 except for additive constants.

$$-2\ln L_{\mathrm{G(c)}}(\boldsymbol{a}) = \sum_i \left[\frac{(c_i - m_{\boldsymbol{a}}(x_i))^2}{\sigma_i^2}\right] + \text{const.} = \chi^2 + \text{const.} \quad \text{if } \sigma_i^2 = \text{const.} \tag{12}$$

In counting experiments, however, this is not the case. The variance σ_i^2 is equal to the true number of counts which is best approximated by $m_{\boldsymbol{a}}(x_i)$. In this case, the Gauss log-likelihood $-2\ln L_{\mathrm{G}}$ is no longer equal to χ^2 but includes additional non-constant terms:

$$-2\ln L_{\mathrm{G(P)}}(\boldsymbol{a}) = \sum_i \left[\frac{(c_i - m_{\boldsymbol{a}}(x_i))^2}{m_{\boldsymbol{a}}(x_i)} + \ln m_{\boldsymbol{a}}(x_i)\right] + \text{const.} = \chi^2_{\mathrm{P}} + \sum_i \ln m_{\boldsymbol{a}}(x_i) + \text{const.} \quad \text{if } \sigma_i^2 = m_{\boldsymbol{a}}(x_i). \tag{13}$$

Instead, the variance is often approximated by the observed number of counts, i.e. $\sigma_i^2 = \max(c_i, 1)$. This choice is very convenient as the c_i are known. As in the modified Neyman's case, the variance has been set to the arbitrary value of 1 if $c_i = 0$.

$$-2\ln L_{\mathrm{G(N)}}(\boldsymbol{a}) = \sum_i \left[\frac{(c_i - m_{\boldsymbol{a}}(x_i))^2}{\max(c_i, 1)} + \ln(\max(c_i, 1))\right] + \text{const.} = \chi^2_{\mathrm{N}} + \text{const.} \quad \text{if } \sigma_i^2 = \max(c_i, 1). \tag{14}$$

It is a common statement that the maximum likelihood derivation in the Gauss case yields the same result as the χ^2 statistic [7,13,5]. As can be seen, this is only true if the standard deviations do not depend on the parameters and thus are not fitted. In the best possible approximation $\sigma_i^2 = m_{\boldsymbol{a}}(x_i)$ additional terms are present (Eq. (13)).

2.4. Error estimation

Estimating parameters is only useful if one can also state the error of the newly obtained value. It is well known that for χ^2 statistics the 1σ error of non-correlated parameters can be obtained by looking for their values where χ^2 is increased by 1.

For sufficiently large event samples, the likelihood (not the distribution of the single measurements) becomes Gaussian centred on those values a_j' that minimise its function [9, p. 144] (see also Refs. [14, p. 189], [18]). Near the minimum the likelihood defined in Eq. (5) can be approximated by

$$L(a_j) = A\exp\left[-\frac{1}{2}\left(\frac{a_j - a_j'}{\sigma_j}\right)^2\right] \tag{15}$$

where A is a function of the other parameters. Calculating the negative logarithm of Eq. (15) yields

$$-2\ln L(a_j) = \left(\frac{a_j - a_j'}{\sigma_j}\right)^2 + \text{const.} \tag{16}$$

showing that a variation of the parameter a_j by $n\sigma_j$ ($a_j = a_j' \pm n\sigma_j$) leads to an increase of the log-likelihood $-2\ln L(a_j)$ by n^2. In order to extract the 1σ error limit, one can vary the parameter until the log-likelihood has been increased by 1, just as one would do for χ^2 statistics.

If the second derivatives of L exist, it is possible to calculate the error directly. Gregorich discusses at length the topic of error limits [12]. He proposes to look either at the half-maximum limit or at the 68.3% confidence level limit. The latter approach corresponds to the 1σ estimation. In order to calculate the half-maximum, one looks at the parameter value where L has half its maximum value $L_{\max}$, i.e. where $-2\ln L = -2\ln(L_{\max}/2) = -2\ln L_{\max} + 2\ln 2$. The half-maximum error corresponds to a log-likelihood increased by approximately 1.4 or a 1.2σ error.

3. Goodness-of-fit estimation

Once a set of parameters has been obtained using a suitable statistic, one might tend to take these values as correct within the error margin. However, the underlying model might be wrong. This can easily be the case, if parameters have been implicitly set to fixed values during the evaluation. In order to estimate how accurately the model describes the data, a goodness-of-fit test should always be performed. This can be easily done by calculating the

probability $Q(\chi^2_{\min}, \nu)$ that an observed χ^2 exceeds the obtained value $\chi^2_{\min}$ by chance *even* for a correct model. Q is the incomplete gamma function and is defined as

$$Q(\chi^2_{\min}, \nu) = \frac{1}{\Gamma(\nu/2)} \int_{\chi^2_{\min}/2}^{\infty} \exp(-t) t^{\nu/2-1}\, \mathrm{d}t \qquad (17)$$

where Γ is the Gamma function. The number ν of degrees of freedom is equal to the number N of measuring points minus the number of fitted parameters. Q will be called "goodness-of-fit" in the following. If $Q > 0.1$ the model can be accepted, if $Q < 0.001$ the model is very likely to be wrong [19, pp. 160–165].

The general χ^2 statistic follows a χ^2 distribution, if all terms in the sum are independent and Gauss distributed with variance 1 [18, p. 64]. In this case, the goodness-of-fit Q is a correct probability and can be calculated using Eq. (17). By definition the classical χ^2 methods are χ^2 distributed if the numbers c_i are Gauss distributed. For counting experiments this is only approximately fulfilled.

The maximum likelihood methods need to be extended for this purpose. A common way is the so-called "likelihood ratio test" which will be discussed in the next section [20, Section 13] (see also Refs. [7], [11], [18, p. 230ff], [21, p. 189], [22, p. 228]). Section 4.3 will be dedicated to a systematic comparison of the goodness-of-fit estimations, based on the different statistics.

In order to describe the goodness-of-fit, always the value of χ^2 and the number ν of degrees of freedom should be given. Though a value of the so-called "reduced χ^2" χ^2/ν close to 1 indicates a good fit, the meaning of the term "close" depends strongly on the number ν of degrees of freedom.

3.1. Likelihood ratio test

The (maximum) likelihood ratio λ was first introduced by Neyman and Pearson [23] and defined as

$$\lambda = \max_{\boldsymbol{a}} \lambda^* = \frac{\max_{\boldsymbol{a}} L(\boldsymbol{c}\,|\,m_{\boldsymbol{a}})}{\max L(\boldsymbol{c}\,|\,\boldsymbol{m})}. \qquad (18)$$

The numerator gives the likelihood $L(\boldsymbol{c}\,|\,m_{\boldsymbol{a}})$ that the experimental data $\boldsymbol{c}$ occurred if the model $m_{\boldsymbol{a}}$ was true, and maximised with respect to the parameters $\boldsymbol{a}$. The denominator gives the likelihood $L(\boldsymbol{c}\,|\,\boldsymbol{m})$ that the experimental data $\boldsymbol{c}$ occurred, maximised without having any restriction on the model. As the denominator is constant and independent of the model, the quantity λ^* has been introduced. Parameters can be estimated by maximising its value. It is evident that the ratio λ, as well as λ^*, can only take on values between 0 and 1.

The likelihood in the numerator is identical to $L_{\mathrm{P}}(\boldsymbol{a})$ from Eq. (6) in the Poisson case or to $L_{\mathrm{G}}(\boldsymbol{a})$ from Eq. (10) in the Gauss case. The global maximum of L in the denominator is found by calculating

$$\frac{\partial}{\partial m'_j}[-2\ln L(\boldsymbol{c}\,|\,\boldsymbol{m})] = 0 \qquad (19)$$

where the parameter-independent m'_i replaces $m_{\boldsymbol{a}}(x_i)$ in the likelihood function L used (Eq. (6) or Eq. (10)).

One can now define a "maximum likelihood χ^2_λ" as

$$\chi^2_\lambda = -2\ln\lambda^* = 2\ln[\max L(\boldsymbol{c}\,|\,\boldsymbol{m})] - 2\ln L(\boldsymbol{c}\,|\,m_{\boldsymbol{a}}). \qquad (20)$$

As $0 < \lambda^* \leqslant 1$, the maximum likelihood χ^2_λ cannot become negative. A theorem of Wilks shows that $\min_{\boldsymbol{a}} \chi^2_\lambda$ asymptotically follows a χ^2 distribution with $N - J$ degrees of freedom [20, p. 419] except for a deviation varying as $1/N$ which is important if the number N of data points is small. A correction is possible to make χ^2_λ behave as a χ^2 statistic to order $1/N^2$ ([18, p. 237], [24, p. 233], [26]).

As the first term in Eq. (20) does not depend on the parameters to be estimated, minimising χ^2_λ with respect to one parameter is identical to maximising $L(\boldsymbol{a}) = L(\boldsymbol{c}\,|\,m_{\boldsymbol{a}})$. Thus, the parameter and error estimation are unchanged if using χ^2_λ instead of $-2\ln L$. In addition, however, χ^2_λ should be applicable to check the consistency of the model and the data by a goodness-of-fit test.

3.2. Application to Poisson distributed data

In the case of Poisson distributed data, Eq. (19) becomes

$$\frac{\partial}{\partial m'_j}[-2\ln L_{\mathrm{P}}|_{m_{\boldsymbol{a}}(x_i)=m'_i}] = 2\left[1 - \frac{c_j}{m'_j}\right] \overset{!}{=} 0 \qquad (21)$$

$$\Rightarrow\ m'_j = c_j \quad \forall j. \qquad (22)$$

The best-possible estimation of the counts $\boldsymbol{c}$ is the counts themselves, leading to $\max L(\boldsymbol{c}\,|\,\boldsymbol{m}) = L_{\mathrm{P}}|_{m_a(x_i)=c_i}$. Even though this sounds trivial, it is *not* a general result. For the case of multinomial distributions a similar derivation can be found in Ref. [18, p. 256].

Eq. (20) can then be written as

$$\begin{aligned}\chi^2_{\lambda,\mathrm{P}} &= 2\ln L_{\mathrm{P}}|_{m_a(x_i)=c_i} - 2\ln L_{\mathrm{P}}(\boldsymbol{c}\,|\,m_a)\\ &= 2\ln\left(\prod_i \frac{[c_i]^{c_i}}{c_i!}\mathrm{e}^{-c_i}\right)\\ &\quad - 2\ln\left(\prod_i \frac{[m_a(x_i)]^{c_i}}{c_i!}\mathrm{e}^{-m_a(x_i)}\right)\\ &= 2\sum_{\substack{i\\ c\neq 0}} [-c_i + c_i\ln c_i - \ln c_i!]\\ &\quad - 2\sum_i [-m_a(x_i) + c_i\ln m_a(x_i) - \ln c_i!]\\ &= 2\left[\sum_i (m_a(x_i) - c_i) - \sum_{\substack{i\\ c_i\neq 0}} c_i\ln\left(\frac{m_a(x_i)}{c_i}\right)\right]. \quad (23)\end{aligned}$$

This equation will be referred to as *Poisson MLE*.

3.3. Application to normally (Gauss) distributed data

What does the likelihood ratio test look like when applied to Gauss statistic? When replacing L in Eq. (20) by the Gauss likelihood from Eq. (10), one must again choose which approximation to use for the variance. The easiest approximation is $\sigma_i^2 \approx \max(c_i, 1)$, the modified Neyman case. The calculation of $\max L(\boldsymbol{c}\,|\,\boldsymbol{m})$ shows that Eq. (22) is also valid in this case, leading to

$$\chi^2_{\lambda,\mathrm{G(N)}} = \sum_i \frac{(c_i - m_a(x_i))^2}{\max(c_i, 1)} = \chi^2_{\mathrm{N}} \quad (24)$$

as all the other quantities cancel out. This approximation for the variances eliminates the difference between the maximum likelihood approach and the pure χ^2 method.

A better approximation is again to use $\sigma_i^2 = m_a(x_i)$, the Pearson case. Due to the additional term in the log-likelihood, Eq. (22) no longer holds true. We find instead

$$\begin{aligned}&\frac{\partial}{\partial m'_j}[-2\ln L_{\mathrm{G(P)}}|_{m_a(x_i)=m'_i}]\\ &\quad = \left[\frac{(m'_j)^2 + m'_j - c_j^2}{(m'_j)^2}\right] = 0 \quad (25)\\ &\Rightarrow\ m'_j = \sqrt{\tfrac{1}{4} + c_j^2} - \tfrac{1}{2} \quad \forall j. \quad (26)\end{aligned}$$

The best-possible estimates for the counts $\boldsymbol{c}$ are no longer the counts themselves, the denominator of the likelihood ratio can thus be written as

$$\max L(\boldsymbol{c}|\boldsymbol{m}) = L_{\mathrm{P}}|_{m_a(x_i)=c'_i} \quad \text{with } c'_i = \sqrt{\tfrac{1}{4} + c_i^2} - \tfrac{1}{2}. \quad (27)$$

This expression is undefined if any $c_i = 0$ (in this case $c'_i = 0$ as well) because the Gauss probability density $p_{\mathrm{G}}(0;0,0)$ becomes infinite. If all $c_i > 0$, one obtains

$$\begin{aligned}&\chi^2_{\lambda,\mathrm{G(P)}}\\ &= \sum_i \left[\frac{(c_i - m_a(x_i))^2}{m_a(x_i)} + \ln\frac{m_a(x_i)}{c'_i} - \frac{(c_i - c'_i)^2}{c'_i}\right]\\ &= \chi^2_{\mathrm{P}} + \sum_i \left[\ln\frac{m_a(x_i)}{c'_i} - \frac{(c_i - c'_i)^2}{c'_i}\right] \quad \text{if all } c_i > 0. \quad (28)\end{aligned}$$

This equation can only be used if all measured values c_i are non-zero. Otherwise it is not possible to perform a goodness-of-fit test using the maximum likelihood estimation and Gauss statistic in the approximation $P_{\mathrm{G}} \approx p_{\mathrm{G}}$. The estimation of parameters and errors is still possible when using Eq. (13) instead. In the following, Eq. (28), respectively, Eq. (13) will be referred to as "Gauss MLE". For large c_i the last term tends quickly towards 0 and the logarithmic term becomes negligible compared to the first, so that asymptotically $\chi^2_{\lambda,\mathrm{G(P)}} \simeq \chi^2_{\mathrm{P}}$.

4. Problems with area estimation or estimation of mean

A common problem when dealing with χ^2 statistics is that the area under the measured peak is not identical to the area under the fitted curve. This

problem does not arise with maximum likelihood methods. When using MLE, the areas are identical. As pointed out by Baker and Cousins, this is true as long as the overall scale in the fitting function can be freely varied during the fit, e.g. when estimating a constant background [7].

4.1. Asymptotic behaviour

The estimation of a mean is a special case of area determination. It is widely known that when trying to find the mean value of Poisson distributed data, the modified Neyman's χ^2_N statistic underestimates the true mean and Pearson's χ^2_P overestimates it. Similar results have been found or stated in several publications, e.g. Refs. [4,7,8,10,12,13,16]. In statistics, an estimator that does asymptotically lead to a wrong value is called "inconsistent" [18, p. 115]. The problem leading to the wrong values can be well explained qualitatively. Using $1/c_i$ as a weighting factor in Neyman's statistic, a stronger weight for all values $c_i < \mu$ will thus lead to an underestimation. On the other hand, using $1/\mu$ in Pearson's statistic, makes χ^2 smaller when μ is increased, thus leading to an overestimation.

The size of the inconsistency depends on the true mean of the Poisson distributed data. This can be seen in Figs. 1(a) and (b). A set of 500,000 Poisson

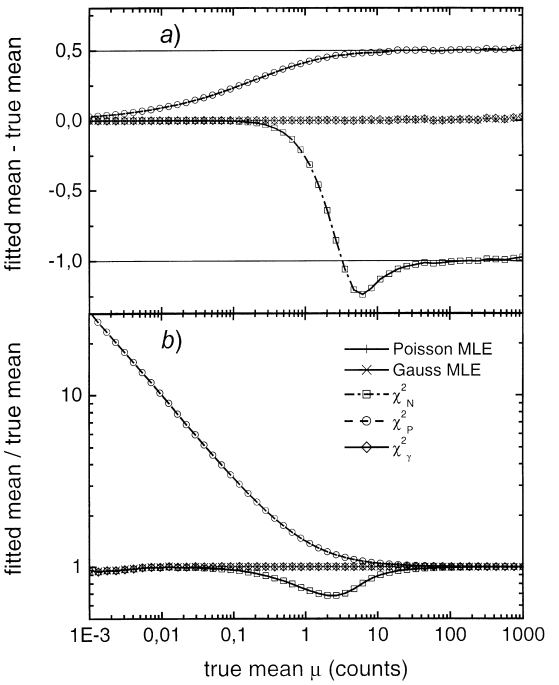

Fig. 1. Deviation of fitted mean from true mean value when using the five statistics under study. (a) For large μ Pearson's χ^2_P leads to an overestimation of 0.5 whereas Neyman's χ^2_N underestimates the true mean by -1. (b) Looking at the ratio of fitted and true mean, one clearly sees that using Pearson's χ^2_P, the fitted mean deviates extremely from the true one for very small μ.

Table 1
Relations giving the best estimate for the five statistics under study derived by setting $d\chi^2/d\mu = 0$. In the second column the general result is given. N' is the number of non-zero measurements $N' = \sum_{c_i \neq 0} 1$. The third column gives the special case of only one measurement. The last column gives the asymptotic mean in the case of many observations

Statistic	$\langle\mu\rangle$	$\langle\mu\rangle$ (only 1 datum)	Asymptotic mean
Poisson MLE	$\frac{\sum c_i}{N}$	c	μ [17]
Gauss MLE	$\sqrt{\frac{\sum c_i^2}{N} + \frac{1}{4}} - \frac{1}{2}$	$\sqrt{c^2 + \frac{1}{4}} - \frac{1}{2}$	μ [17]
χ^2_P	$\sqrt{\frac{\sum c_i^2}{N}}$	c	$\mu\sqrt{1 + \frac{1}{\mu}}$ [13]
χ^2_N	$N' \Big/ \sum \frac{1}{\max(c_i, 1)}$	c	$(e^\mu - 1) \Big/ \left(1 + \sum_{k=1}^{\infty} \frac{\mu^k}{k(k!)}\right)$ [4]
χ^2_γ	$N' \Big/ \sum \frac{1}{c_i + 1}$	$c + \min(c, 1)$	μ [4]

distributed numbers with mean μ has been generated. Using the five statistics described in Section 2, best estimates $\langle\mu\rangle$ for the mean have been calculated using the relations given in Table 1. This has been done for true mean values μ from 0.001 to 1000. The deviation from the true mean by -1 for χ^2_N, and 0.5 for χ^2_P, can be found for true mean values of approximately 10 or more. Using the χ^2_γ statistic or one of the two MLE approaches, the mean of Poisson distributed data is estimated correctly for any μ.

Baker and Cousins claim that the wrong "area", i.e. the wrong estimate of the mean, is due to the implicit assumption of Gauss statistic in the χ^2 methods [7]. However, the maximum likelihood approach "Gauss MLE", based on the assumption that data are normally distributed, also estimates the correct mean if the variance is approximated by the fitted value. The calculation of the mean fails if the normalisation factor $(2\pi\sigma^2)^{-1/2}$ of the Gauss distribution is neglected as in χ^2_P or artificially made independent of the mean to be fitted as in χ^2_N. The best-possible derivation coming from the maximum likelihood method leads to additional terms $\ln m_a(x_i)$ compared to Pearson's χ^2_P (see Eq. (13)) leading to a correct estimation of the mean even with Gauss statistic.

The correctness of the χ^2_γ statistic is not surprising as it has been constructed in order to estimate the correct mean.

For large mean values, the Poisson distribution can be approximated by a normal distribution to high precision. This could explain the correctness of the Gauss MLE approach in the high-μ part. For low averages, however, this is not true. The overall correctness is due to the large number of Poisson distributed numbers used. The central limit theorem states that a variable is normally distributed if it is the sum of many independent identically distributed random variables. Consequently, averaging (summing) over a large number of Poisson data, a correct estimation of the mean becomes possible even with Gauss statistic.

Similar investigations have shown that inconsistent results are also obtained when fitting some predefined curve (e.g. a Gaussian or Lorentzian) to a peak in order to estimate a total intensity. If information about intensities is needed, the classical χ^2 methods are clearly not suited.

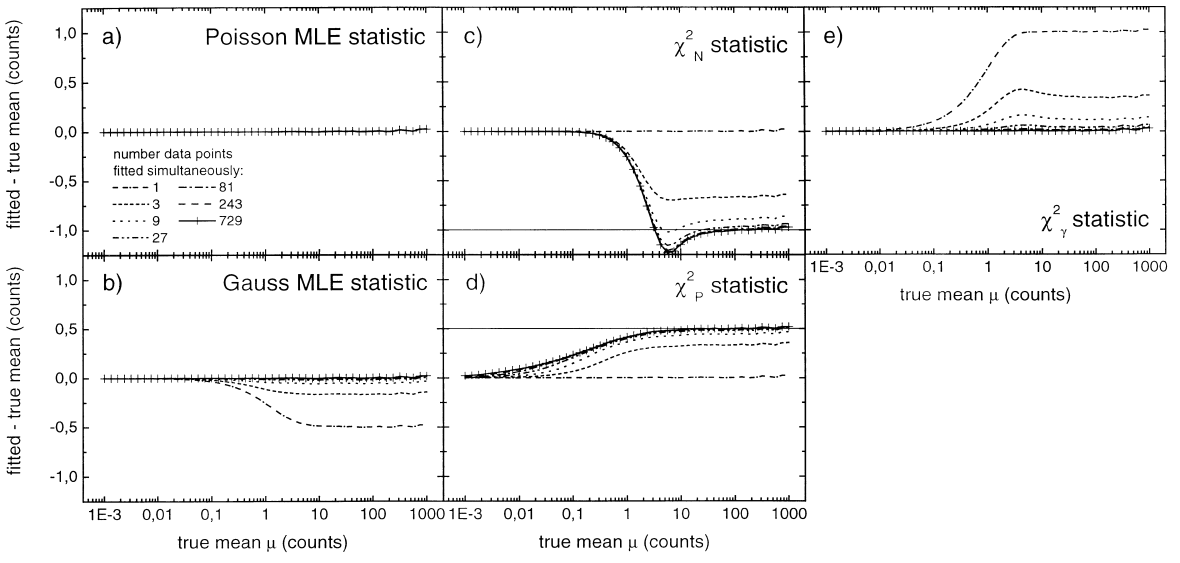

Fig. 2. Convergence test for the five statistics under study: (a) Poisson MLE, (b) Gauss MLE, (c) χ^2_N, (d) χ^2_P, (e) χ^2_γ. Displayed is the deviation of the fitted mean from the known true mean value. The fitted mean has been obtained by averaging the fitted mean values of groups of $1, 3, \ldots$, respectively, $3^6 = 729$ random numbers (calculated by the relations given in Table 1). Only the Poisson MLE approach finds the correct mean in all cases.

4.2. Convergence properties

In true experiments, only a limited number of measurements can be performed. Is it possible to estimate a correct mean when using only very few data points? The same 500,000 Poisson distributed numbers as used before, have been grouped by $1, 3, 9, \ldots, 3^6 = 729$ numbers. The estimate of every group can then be calculated by the equations given in Table 1. Every group leads to a different result, of course, due to the random nature of the input data. However, it is possible to compare the distributions of results via their mean values and their spread.

It is important to note that the average has to be calculated without any weighting factor, as all the random numbers have been taken from a Poisson distribution around the same true value μ.

$$\langle \mu \rangle_{\text{average}} = \sum_{k=1}^{K} \frac{\langle \mu \rangle_k}{K}. \tag{29}$$

The results of the calculations are displayed in Figs. 2(a)–(e). For the modified Neyman's χ^2_{N} and Pearson's χ^2_{P} statistics, on average, the correct mean value μ is found if every datum is "fitted" separately (Figs. 2(c) and (d)). This can be understood as, for a single datum, these statistics yield $\mu_i = c_i$. As soon as the inherent weighting of the statistics comes into play, i.e. as soon as more than one datum is fitted simultaneously, the correct value is no longer obtained on average. Combining many data finally leads to the deviation of -1, respectively, 0.5 for high values of μ. This proves that the failure of these two statistics can be understood as a problem with the weights used.

What about the asymptotically correct statistics? Using χ^2_γ or Gauss MLE statistics, an offset between the average of fitted results and the true value exists. It becomes smaller the more data points are simultaneously fitted. Only asymptotically are the correct values approached. These estimators are said to be "biased" [18, p. 115]. The convergence is faster for the Gauss MLE statistic, where approximately 10 data points are sufficient, whereas 100 data points are needed for χ^2_γ.

Only in the case of Poisson maximum likelihood estimation does the average of the fitted results coincide with the true mean value for all values of μ and for any number of simultaneously fitted data points (Fig. 2(a)).

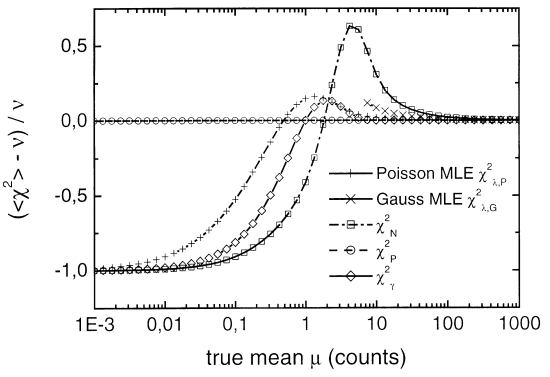

Fig. 3. Normalised expectation values for χ^2 calculated for different mean values of a Poisson distribution. If the statistic is distributed following a χ^2 distribution, the expectation value $\langle \chi^2 \rangle$ should be equal to the number ν of degrees of freedom, therefore $(\langle \chi^2 \rangle - \nu)/\nu)$ should vanish. Only Pearson's χ^2_{P} shows this behaviour for all mean values. A goodness-of-fit estimation based on the other statistics would lead to incorrect results for the consistency.

In any of the asymptotically correct statistics, the spread of the single results due to the randomness of the data is larger than the deviation of the fitted average of the mean. Therefore one might tend to ignore the slight difference. Nonetheless, it is important to be aware that the distribution of results is not centred around the true value when using χ^2_γ or Gauss MLE statistics.

4.3. Goodness of fit

After extracting parameters and errors (standard deviation) from the data, one should look at the goodness of fit Q. When estimating a mean value, a sufficiently high value of Q indicates that the measured data really belong to a Poisson distribution around one value. Whereas a low value is a sign for a systematic error. If the experimental conditions have changed during the measurement, e.g., the correct description might be a superposition of two Poisson distributions around different mean values.

Which statistic is suited to judge on the goodness of fit? If the statistic follows a χ^2 distribution, the

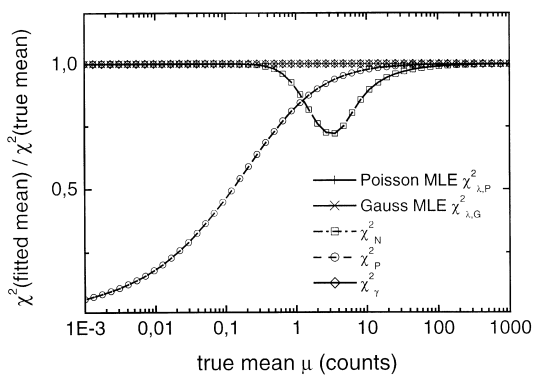

Fig. 4. Comparison of χ^2 calculated at the true and at the estimated mean of a Poisson distribution. The maximum likelihood statistics, as well as the χ^2_γ statistics, lead to correct estimates for the mean. Consequently, the two χ^2 values are identical. Both, Pearson's χ^2_P and Neyman's χ^2_N, however, yield wrong estimates. In some range of μ, the values of χ^2 calculated at these incorrect parameter values are significantly smaller than χ^2 at the true mean. If using these incorrect values for a goodness-of-fit estimation, the consistency of the model and the data is overestimated.

expectation value $\langle\chi^2\rangle$ should be equal to the number ν of degrees of freedom. This number is equal to the number N of measured points minus the number of fitted parameters. In the case of estimating a mean $\nu = N - 1$.

For the five statistics under study we calculated the expectation value of χ^2

$$\langle\chi^2\rangle = \sum_{c_i=0}^{\infty} P_{\mathrm{P}}(c_i;\mu)\cdot\chi^2(c_i|\mu). \tag{30}$$

The deviation from ν is shown in Fig. 3. It can be clearly seen that only Pearson's χ^2_P leads to the correct estimation. All the other statistics deviate from the theoretical expectation value at least for small values of μ. This behaviour has also been found by Mighell (see Fig. 2 in Ref. [4]), but he does not draw any conclusions from the disagreement.

The result is very interesting as the wide application of the χ^2 methods is often explained by the applicability to the goodness-of-fit estimation. It can be clearly seen, however, that the Neyman approximation does not give the correct expectation value, i.e. that it is not distributed following a χ^2 distribution. The new χ^2_γ method is also not suited for this purpose.

The maximum likelihood approaches, as well, lead to incorrect results. This is astonishing as the textbooks do not mention any deviation from the χ^2 distribution for small numbers. Deviations for small data sets exist and possible corrections are discussed for example in Ref. [24, p. 233] or originally by Lawley in Refs. [25,26]. He proposes to correct the obtained χ^2 by multiplying a factor $\nu/\langle\chi^2\rangle$.

To summarise, the estimation of a mean of a Poisson distribution should be done using the Poisson MLE statistic. The goodness-of-fit Q should be calculated from Pearson's χ^2_P with the previously estimated parameter. It is worth noting that Pearson's χ^2_P is not minimum and so the goodness-of-fit Q is not maximum at this parameter (cf. Section 4.1). This can be seen from Fig. 4. The value of χ^2_P calculated at the true position is larger, i.e. the goodness-of-fit Q is smaller, than its minimum value corresponding to the χ^2_P-fitted estimate of the mean.

4.4. Binned data – combining bins

Counting experiments often deal with binned data (histograms) and so do many publications. In this work the theoretical function is sampled by taking measurements at different values of x. This sequential way of measuring can be looked at as if all measurements were taken at the same time in different bins of identical size.

The χ^2 methods require, that the spread of the measured data is sufficiently well described by a normal distribution, i.e. that the number of counts in each bin is sufficiently large. Otherwise adjacent bins must be combined ("re-binning"). Many textbooks state that at least 5 counts must be in every bin (see e.g. Ref. [18, p. 257]). Combining bins, however, washes out the fine structure that may contain valuable information. Gumble already studied this phenomenon half a century ago using the χ^2_P statistic [27]. He finds that not only the χ^2 values but also fitted intensities and parameters may depend on the way of re-binning.

5. Fitting of artificial Crystal-GRID data

5.1. The idea of Gamma Ray Induced Doppler broadening (GRID) measurements

In a thermal neutron capture reaction the binding energy of a neutron is deposited in the nucleus. The de-excitation takes place via a γ cascade. The emission of a first, primary photon γ_1 entails a recoil of the still excited nucleus with an initial kinetic energy of several hundreds of eV. The direction of this initial recoil is arbitrary. The further motion of the recoiling nucleus is defined by the interatomic forces, i.e. by the interatomic potential in the solid.

At the time t' a second photon γ_2 is emitted and observed by the spectrometer if emitted along the direction of observation. If the nucleus, moving with velocity $\boldsymbol{v}(t)$, has not yet come to rest, the γ energy is Doppler shifted by

$$\Delta E_{\gamma_2} = E^0_{\gamma_2}\frac{\boldsymbol{v}(t')\cdot\boldsymbol{n}}{c} \tag{31}$$

where $E^0_{\gamma_2}$ is the un-shifted energy of the second photon, c the velocity of light and $\boldsymbol{n}$ the unit vector in the direction of observation. Instead of measuring the shifted energy of one photon, many photons are observed giving rise to a Doppler broadened photon energy spectrum, called "GRID line shape" [28]. The probability for a photon to be emitted at time t' is determined by the radioactive decay and depends on the lifetime τ of the intermediate nuclear level.

The motion of the recoiling atoms within a single crystal is highly anisotropic due to the existence of blocking and channelling directions. In the case of a single crystal as a target (Crystal-GRID), the anisotropy exists macroscopically and leads to a fine structure in the Doppler broadened line shape.

It is not possible to calculate the line shapes analytically. If reasonable assumptions on the interatomic potential and the nuclear level lifetime are made, however, the line shapes can be predicted numerically. Molecular dynamics (MD) simulations are used to calculate trajectories of the recoiling atom within the crystal. These can be converted to a probability distribution $P_{\boldsymbol{n}}(E,t')$, i.e., the probability for a decaying nucleus that a photon with Doppler shifted energy E is emitted in a given direction of observation $\boldsymbol{n}$ at time t'. Multiplying the radioactive decay law, the Doppler broadened line shape $P_{\boldsymbol{n},\tau}(E)$ is obtained as

$$P_{\boldsymbol{n},\tau}(E)\,\mathrm{d}E = \frac{1}{\tau}\int_0^\infty \exp\left(-\frac{t'}{\tau}\right)P_{\boldsymbol{n}}(E,t')\,\mathrm{d}t'\,\mathrm{d}E. \tag{32}$$

The predicted line shape depends on the interatomic potential chosen, on the direction of observation $\boldsymbol{n}$ in the experiment, and on the nuclear level lifetime τ that is to be obtained by the experiment.

A real measurement consists of a large number of scans, where "scan" means one recorded Doppler broadened energy spectrum. It is not possible to measure the spectrum in one scan with sufficient statistics, as the measuring time per scan is limited by the drift stability of the spectrometer. The Bragg angles involved in the double crystal reflection are very small. The positioning of the crystals is controlled via optical interferometers. Their intrinsic stability is affected by environmental influences, setting the limit for the duration of a scan to several hours. In this paper we assume that the

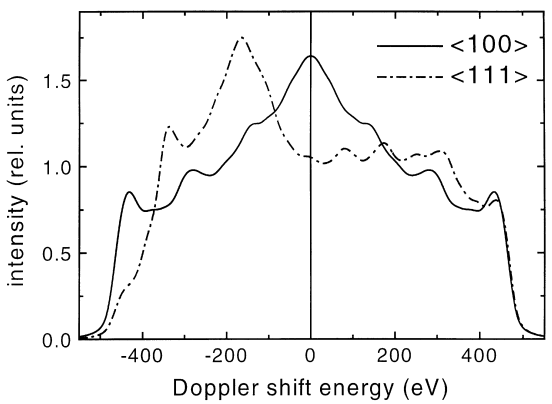

Fig. 5. Crystal-GRID line shapes for two directions of observation derived from MD simulations based on idealised assumptions about material properties. The nuclear level lifetime is taken to be $\tau = 8$ fs. These line shapes have been used in the creation of artificial Crystal-GRID data.

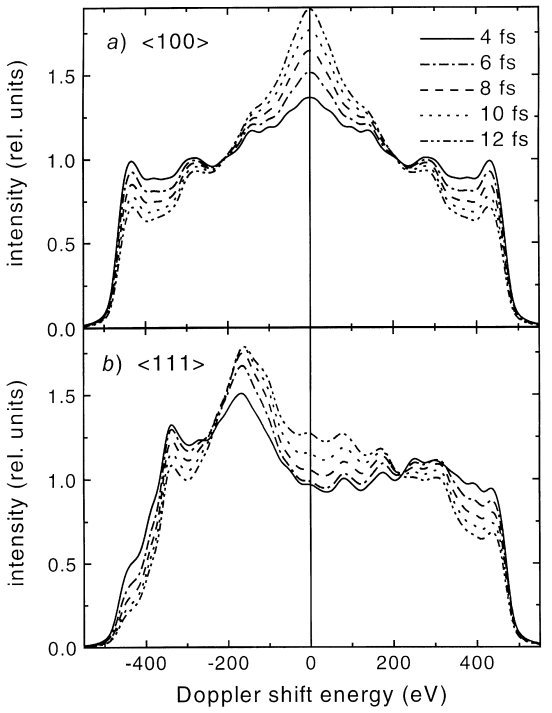

Fig. 6. Dependence of the two Crystal-GRID line shapes ((a) $\langle 1\,0\,0\rangle$, (b) $\langle 1\,1\,1\rangle$) on the studied parameter, the lifetime τ. While the total area under the peak remains constant, the shape of the line depends strongly on the lifetime.

deformations of the line shapes due to drift within the scans are negligible. However, a small drift in between the scans, which leads to a displacement of the entire scan along the scan axis, is taken into account. Due to these displacements, a summation of the single scans is only possible after fitting the calculated line shape to every scan separately. Thus, the peak positions are determined and the scans can be shifted to make the lines coincide.

The background level and the flux (related to the true peak height) are relatively stable. Nonetheless, it is not clear that they remain constant over a period of several weeks. This is especially true when taking measurements with different crystal orientations, as different crystals are used in this case and a few months can pass in between.

The influence of the drift within the scans and ways of correcting them will be discussed in a separate paper.

5.2. Artificial Crystal-GRID data

For the purpose of testing the different statistics, artificial data have been created using a Monte-Carlo approach. Theoretical line shapes have been calculated for a true material (GaP) and for a given interatomic potential. The decay cascade has been idealised, the details of which are of no importance to the following. Two directions of observation will be investigated, along the $\langle 1\,0\,0\rangle$ and $\langle 1\,1\,1\rangle$ crystal orientations, respectively. The major difference is related to the symmetry of the lines. The line shapes for the fixed lifetime of 8 fs are displayed in Fig. 5, their dependence on the lifetime τ is shown in Figs. 6(a) and (b).

The artificial data have been created by scaling the line shape so that the maximum value, called "peak height", is equal to a value μ (6, 10, 18, 30, 60, 100, 180, 300, 600, or 1000 counts) and afterwards adding a constant background of 2 counts. A total of 100 equidistant points (distance: 0.13 interferometer fringes ≈ 13.8 eV) have been chosen within the line shape as well as in the background region to the right and left of the line shape. At each of these points a number of counts has been arbitrarily chosen from a Poisson distribution around the true value, where the true value corresponds to the value of the model line shape with $\tau = 8$ fs (see Fig. 5).

As pointed out before, the duration of a single Crystal-GRID scan is limited to several hours. Therefore, a measurement has to be divided into many scans. For each peak height, 1000 artificial scans have been created, for low peak heights up to 5000 scans will be used, as convergence is not easily obtained. The model line shape is fitted to these scans in order to determine an estimate for the lifetime. Ideally the true value of 8 fs should be reproduced.

As the data are created artificially, the model is completely known and correct for the data. This differs from experiment where the true interatomic interactions are not known and need to be approximated.

5.3. Parameter estimation – asymptotic behaviour

In the fitting procedure, four different parameters are used: the lifetime, the peak position, the peak

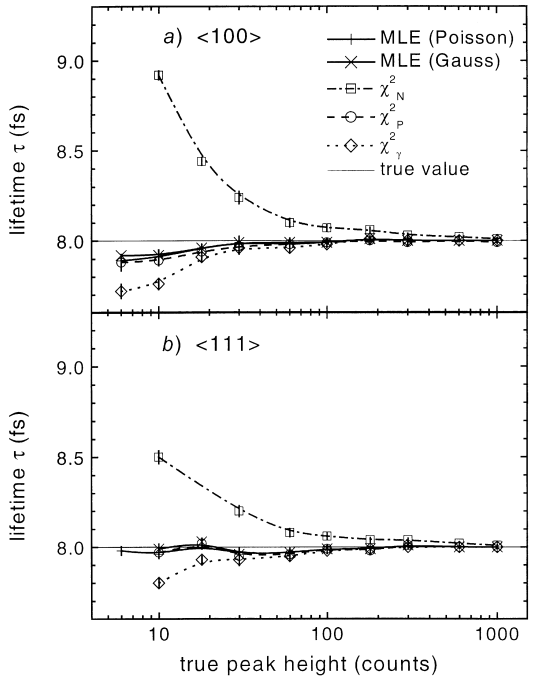

Fig. 7. Resulting lifetime when fitting the model line shape to artificial data created using the true lifetime of 8 fs. (a) The fit has been performed using scans for an observation in $\langle 1\,0\,0\rangle$ direction, and (b) in $\langle 1\,1\,1\rangle$ direction, respectively.

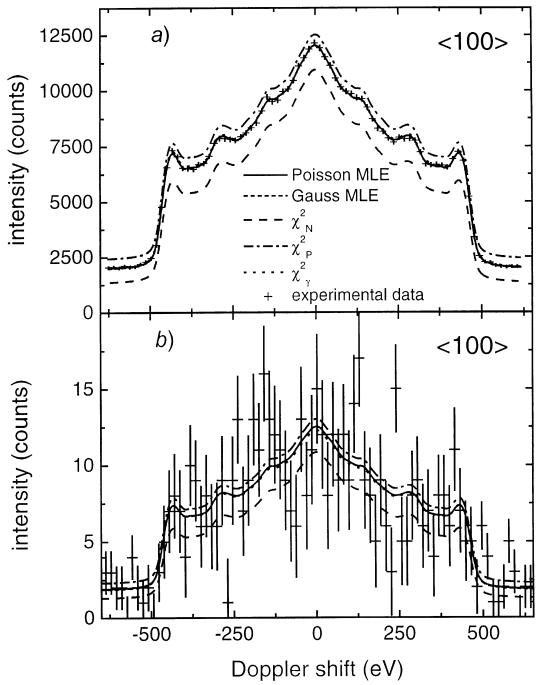

Fig. 8. Comparison of fitted theoretical line shapes and artificial experimental data (+). (a) The sum of 1000 artificial scans (peak height 10) is compared to the sum of the fitted lines. It is most evident, that Neyman's χ^2_N and Pearson's χ^2_P do not reproduce the data at all. (b) The inconsistency can *not* be seen in the plot of a single scan.

height and the background level. The first parameter is related to the radioactive decay of the recoiling nucleus, thus a physical constant. It is a "global" parameter, identical for all of the scans. As discussed before, the peak position, peak height and background can, however, slightly change from scan to scan. Therefore they are fitted as "local" parameters scan by scan.

Consequently, if looking at M scans, the total number of parameters to be fitted is $3M + 1$. The problem can be separated, as the local parameters are only related to the data within one scan. Fixing the lifetime to a trial value, the corresponding line shape is fitted to all the scans separately by varying the local parameters. For every scan an individual value χ^2_m is obtained. In the ideal case of a correct model, χ^2_m should follow a χ^2 distribution with $N - 3$ degrees of freedom, where $N = 100$ is the number of data points in the scan. This procedure is repeated for different values of the lifetime, thereby going towards the minimum of the global $\chi^2 = \sum_{m=1}^{M} \chi^2_m$ as a function of τ.

The result of the fit can be found in Figs. 7(a) and (b) where the values of the fitted lifetime corresponding to the minimum χ^2 are displayed versus the true peak height in a single scan. In the case of the modified Neyman's χ^2_N statistic the resulting lifetime is too high, especially at low peak heights. The new χ^2_γ statistic leads to accurate estimates for peak heights of more than approximately 50. For smaller peaks, however, the estimates are too small. The two MLE statistics, as well as Pearson's χ^2_P, lead to nearly identical results. They are consistent with the true value in the case of $\langle 1\,1\,1\rangle$ orientation, they are slightly too small for small peak heights in the $\langle 1\,0\,0\rangle$ orientation.

The incorrect fitting in Pearson's and Neyman's case becomes most evident when looking at the summed data compared to the summed fitted line shapes. The curves do not coincide at all, as can be

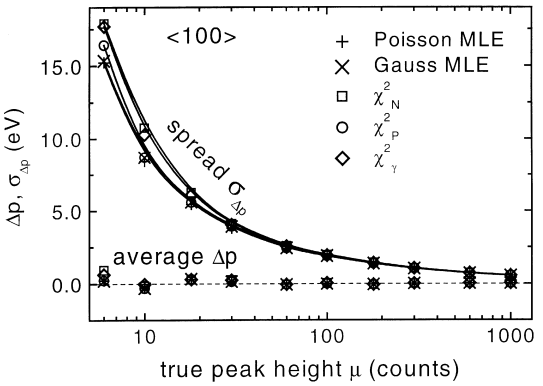

Fig. 9. Average deviation and spread of the peak positions when fitting a model line shape to a large number of scans for an observation in $\langle 100 \rangle$ direction. On average, the five statistics under investigation estimate the correct peak positions, the deviation from the true values is very small. However a large spread (standard deviation) exists, especially for peak heights up to 30. A similar result was found for the $\langle 111 \rangle$ orientation.

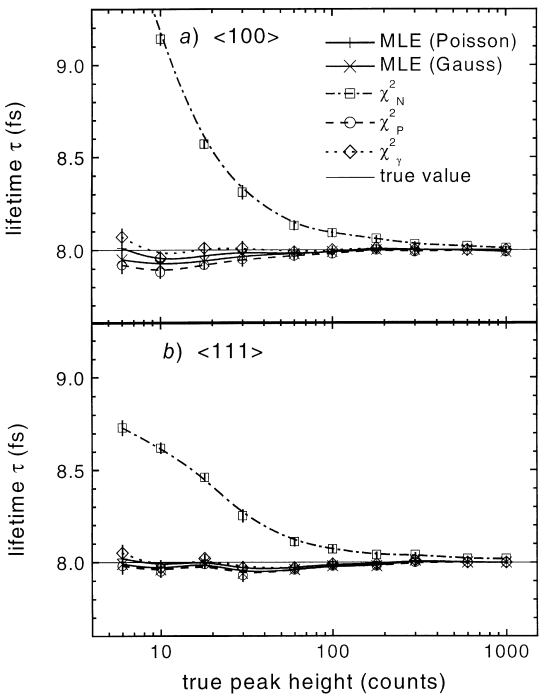

Fig. 10. Resulting lifetime when fitting the model line shape to artificial data created using the true lifetime of 8 fs with the peak positions fixed at their true values. (a) The fit has been performed using scans for an observation in $\langle 100 \rangle$ direction, and (b) $\langle 111 \rangle$ direction, respectively.

seen in Fig. 8(a). This obvious disagreement only becomes visible after the summation of many scans. For every single scan, all statistics lead to a fit that cannot be judged correct or incorrect at first sight (see Fig. 8(b)). The disagreement can be explained by the failure of the two statistics to estimate a mean value of Poisson distributed data, as seen in Fig. 1.

Why is it not always possible to find the correct value of the lifetime? This can be related to the spread of fitted peak positions. A peak position is fitted individually for every scan. This fitted position $p_{\mathrm{f},i}$ deviates from the known true one $p_{\mathrm{t},i}$ by $(\Delta p)_i = p_{\mathrm{f},i} - p_{\mathrm{t},i}$. Looking at all the scans, we can calculate the average of these deviations as well as their spread or standard deviation. The common finding for all of the statistics is that the average deviation is very small for any peak height, but there is a large spread, especially for small peak heights. This can be seen in Fig. 9.

If the peak height in the single scans is very low, a relatively large deviation exists in any single scan. For this reason, the model is not correctly fitted to the data, i.e. the determination of every $\chi^2_m(\tau)$ is slightly incorrect leading to an inaccurate estimate of the lifetime. It is possible to visualise this behaviour. If one summed up the single scans centred around the wrong peak positions, the obtained "experimental" sum scan is smeared out in comparison to the true one. Fitting the theory to this sum scan, is of course not the same as fitting it to the true sum.

This reasoning has been verified in a second step. The peak position in every scan has been fixed to the correct value, so that besides the lifetime only the peak heights and background levels have to be fitted as "local" parameters. The resulting lifetimes are plotted in Figs. 10(a) and (b).

In this case, the true value of the lifetime can be obtained by all statistics except for the modified Neyman's χ^2_{N}. For Pearson's χ^2_{P} the lifetime can be correctly estimated. The freely varying background term in the fit is able to include the inconsistency found in Section 4 (cf. Fig. 1) [13]. However, the free background term does not prevent the fitted

lifetime from being wrong in the case of Neyman's χ^2_N statistic.

Back to the first step. Why is the true value of the lifetime not at least within one or two standard deviations σ_τ of the estimated value? The fitting procedure separates the fits of the local parameters from the lifetime fit. Consequently, the statistical error σ_τ estimated from the curvature of $\chi^2(\tau)$ does not include the systematic error related to the spread of the peak positions. The statistical error σ_τ of the lifetime can be reduced to an arbitrarily small value by adding further and further scans. Yet, the spread of the peak positions, and thereby the systematic error, is not reduced. For low peak heights, the statistical error may become negligible compared to the systematic one. The precision of the lifetime measurement is then limited by the systematic error.

5.4. Parameter estimation – convergence properties

The estimated value of the lifetime is independent of the number of scans investigated. The model line shape was fitted separately to a large number of scans. Thereby a set of estimates τ_i for the lifetime was obtained. Within the error, the average of these estimates is identical to the lifetime obtained in a simultaneous fit of all the scans with the lifetime as global parameter. This could be shown for all statistics investigated. Of course, the more scans measured, the smaller the error on the parameter, and therefore the higher the probability that the fitted result will be near the true value.

This is true as long as the peak in every scan is high enough, to prevent systematic errors as stated in the paragraph before. Whether this condition is fulfilled in the case of an actual measurement, it should be checked by creating artificial scans corresponding to the material studied. In these scans the peak height should be varied in order to obtain the dependence of the fitted lifetime τ on the peak height. For peaks high enough, the fitted lifetime should reach the true value and become independent of the peak height. If possible, a sufficient peak height should be chosen. Otherwise, the deviation of the fitted value from the true one for the experimental peak height gives an idea of the additional systematic error of the fitted lifetime.

5.5. Error estimation for the lifetime

The error of a fitted parameter can be estimated by looking for the parameter values where the statistical variable χ^2 is increased by 1. This has been discussed in Section 2.4 and is true in the case where only one parameter is fitted. In our fitting procedure, the local parameters are optimised for every value of τ. Thereby the fitting problem is separated into M subproblems with 3 parameters, and one fit with 1 parameter, the lifetime. In this case it is correct to look for the increase of χ^2 by 1 to find σ_τ [9, p. 212].

As proposed in the previous section, the investigation of artificial data corresponding to a true measurement allows us to determine whether the peak height is sufficient in a single scan. It is further possible to investigate how many scans are needed in order to get a result to a certain precision.

5.6. How to optimise potential parameters

By varying the parameters in the interatomic potential and looking for a minimum value of the Poisson MLE $\chi^2_{\lambda,P}$ (Eq. (23)), it is possible to optimise the potential. If $\chi^2_{\lambda,P}$ does not change its value significantly when varying one parameter within reasonable limits, the measurement was not sensitive to this parameter. By repeating the same procedure with artificial data containing more scans or higher peaks, it is possible to estimate if the Crystal-GRID measurement with this material is not sensitive at all or if the non-sensitivity is only due to an insufficient total count rate. Once the optimisation is done, a goodness-of-fit test using Pearson's χ^2_P is essential. For an accurate potential, the goodness-of-fit estimator Q should be above 0.001 (see Section 3).

6. Conclusions

The evaluation of experimental data often includes a fit of theoretical predictions to the data. A statistic is used to find the best parameters and their standard deviations as well as to judge the consistency of the model and the data. In this paper, three different χ^2 statistics as well as two

maximum likelihood approaches were presented and compared in the case of counting experiments.

It was clearly shown that the maximum likelihood estimation (MLE) based on Poisson statistic (Eq. (23) or Eq. (7)) is the best possible approach in order to estimate parameters and errors when dealing with counting experiments, especially if small total numbers of counts may occur.

The maximum likelihood approach using Gauss statistic is identical to the classical χ^2 methods only if the standard deviations do not depend on the parameters. In the best approximation for the variances $\sigma_i^2 = m_a(x_i)$, however, this is not the case. Additional terms are present.

When estimating mean values or areas under a peak, consistent estimates of the true value are not found when applying Neyman's χ^2_N or Pearson's χ^2_P. This result was verified. A new χ^2_γ method (Eq. (4)), also studied in this work, is especially designed in order to cope with this problem. It is yet more convincing to apply the maximum likelihood method. Both derivations, based on Poisson and Gauss statistics, respectively, also come to the consistent estimates. The normalisation factor of the Gauss distribution leads to additional terms $\ln m_a(x_i)$ (Eq. (13)) in the Gauss MLE statistic compared to the classical χ^2 statistics. It can be concluded that the neglect of this factor – and not the use of Gauss statistic in general – is the reason for the failure when estimating an area or a mean.

It was further shown that the Gauss MLE and χ^2_γ statistics yield only asymptotically consistent results due to the central limit theorem, whereas in Poisson MLE (Eq. (23) or Eq. (7)) the true mean is always found. Consequently, in order to obtain a consistent and unbiased estimate of the mean value, the Poisson MLE approach should be used.

The consistency of the fitted values or function with the data should always be checked after estimating parameters. The wide application of the classical χ^2 statistics (Eq. (2) or (3)) can be attributed to their reputed applicability to goodness-of-fit tests. However, the MLE method can be extended to this type of application by means of the likelihood ratio test.

Our investigations show that only χ^2_P yields the correct expectation value for χ^2 which is the number ν of degrees of freedom, and thus gives the correct goodness-of-fit estimate. All the other statistics cannot be used for a goodness-of-fit estimation, even in the case of many data points. It should be pointed out that the widely used Neyman's χ^2 is neither good for parameter estimation nor for goodness-of-fit estimation. It should be avoided in the evaluation of counting experiments.

In order to judge the goodness-of-fit one should not look at the "reduced χ^2" χ^2/ν but always state χ^2_P and the number ν of degrees of freedom allowing to calculate the goodness-of-fit estimator Q (Eq. (17)).

The different statistics have been applied in our special case of Crystal-GRID experiments. Investigations using artificial data showed that the most significant parameter, a nuclear level lifetime, cannot be correctly estimated using Neyman's χ^2_N. The other statistics give consistent estimates if the peak positions can be precisely determined. Otherwise, a small systematic error may be introduced. This error is not included in the error estimated from the $\chi^2(\tau)$ curvature as the fit of the peak positions is separated from the fit of the lifetime.

Even though a large spread in the peak positions of repeated measurements exists, on average the correct values are found with any statistic under study.

When preparing a new experiment, artificial data should be created in order to estimate the peak height necessary for a consistent result of the lifetime. The artificial scans can further be used to estimate the number of measurements needed for a desired precision of the lifetime or in order to check whether an actual measurement is sensitive to a certain parameter of the interatomic potential.

Acknowledgements

We thank Erich Lohrmann, DESY and University Hamburg, Germany, for his helpful remarks to the manuscript.

References

[1] M. Jentschel, K.H. Heinig, H.G. Börner, J. Jolie, E.G. Kessler, Nucl. Instr. and Meth. B 115 (1996) 446.

[2] N. Stritt, J. Jolie, M. Jentschel, H.G. Börner, C. Doll, Phys. Rev. B 59 (1999) 6762.
[3] N. Stritt, J. Jolie, M. Jentschel, H.G. Börner, H. Lehmann, Phys. Rev. B 60 (1999) 6476.
[4] K.J. Mighell, Astrophys. J. 518 (1999) 380.
[5] M.R. Stoneking, D.J. Den Hartog, Rev. Sci. Instr. 68 (1997) 914.
[6] F.A. Haight, Handbook of the Poisson distribution, Publications in Operations Research, Vol. 11, Wiley, New York, 1967.
[7] S. Baker, R.D. Cousins, Nucl. Instr. and Meth. 221 (1984) 437.
[8] T. Awaya, Nucl. Instr. and Meth. 165 (1979) 317.
[9] P.R. Bevington, D.K. Robinson, Data Reduction and Error Analysis for the Physical Sciences, 2nd Edition, McGraw-Hill, New York, 1992.
[10] W.A. Wheaton, A.L. Dunklee, A.S. Jacobson, J.C. Ling, W.A. Mahoney, R.G. Radocinski, Astrophys. J. 438 (1995) 322.
[11] W. Cash, Astrophys. J. 228 (1979) 939.
[12] K.E. Gregorich, Nucl. Instr. and Meth. A 302 (1991) 135.
[13] Y. Jading, K. Riisager, Nucl. Instr. and Meth. A 372 (1996) 289.
[14] V. Blobel, E. Lohrmann, Statistische und numerische Methoden der Datenanalyse, Teubner, Stuttgart, 1998.
[15] T. Sandor, G.D. Wilson, Comp. Progr. Biomed. 2 (1972) 111.
[16] V.A. Muravsky, S.A. Tolstov, A.L. Kholmetskii, Nucl. Instr. and Meth. B 145 (1998) 573.
[17] M.D. Hannam, W.J. Thompson, Nucl. Instr. and Meth. A 431 (1999) 239.
[18] W.T. Eadie, D. Drijard, F.E. James, M. Roos, B. Sadoulet, Statistical Methods in Experimental Physics, 1st Edition, North-Holland Publishing, Amsterdam, 1971.
[19] W.H. Press, B.P. Flannery, S.A. Teukolsky, W.T. Vetterling, Numerical Recipes. The Art of Scientific Computing, Cambridge University Press, Cambridge, 1989.
[20] S.S. Wilks, Mathematical Statistics, 2nd Printing Edition, Wiley, New York, 1963.
[21] P.G. Hoel, Introduction to Mathematical Statistics, 2nd Edition, Wiley, New York, 1954.
[22] S. Brandt, Datenanalyse, 4th Edition, Spektrum, Akad. Verl., Heidelberg, 1999.
[23] J. Neyman, E.S. Pearson, Biometrika 20 A (1928) 263.
[24] M.G. Kendall, A. Stuart, The Advanced Theory of Statistics, 2nd Edition, Vol. 2, Charles Griffin & Company, London, 1967.
[25] D.N. Lawley, Biometrika 43 (1956) 128.
[26] D.N. Lawley, Biometrika 43 (1956) 295.
[27] E.J. Gumbel, Ann. Math. Statist. 14 (1943) 253.
[28] H.G. Börner, J. Jolie, J. Phys. G 19 (1993) 217.