

Maximum-Likelihood Fits to Histograms for Improved Parameter Estimation

J. W. Fowler

Received: 24 July 2013 / Accepted: 18 January 2014 / Published online: 7 February 2014
© Springer Science+Business Media New York 2014

Abstract Straightforward methods for adapting the familiar χ^2 statistic to histograms of discrete events and other Poisson distributed data generally yield biased estimates of the parameters of a model. The bias can be important even when the total number of events is large. For the case of estimating a microcalorimeter's energy resolution at 6 keV from the observed shape of the Mn $K\alpha$ fluorescence spectrum, a poor choice of χ^2 can lead to biases of at least 10 % in the estimated resolution when up to thousands of photons are observed. The best remedy is a Poisson maximum-likelihood fit, through a simple modification of the standard Levenberg-Marquardt algorithm for χ^2 minimization. Where the modification is not possible, another approach allows iterative approximation of the maximum-likelihood fit.

Keywords Energy resolution · Histogram fitting · Maximum likelihood

1 Introduction

One of the most important tasks in experimental physics is estimating the parameters of a model from data. Often, the data are summarized in the form of a histogram; each bin counts independent, discrete events, and its contents therefore follow the Poisson distribution [1]. It is all too easy to make biased parameter estimates by inappropriate use of techniques adapted to counting data from the χ^2 statistic designed for Gaussian distributions. Fortunately, it is also nearly as easy to use maximum-

Official contribution of NIST, not subject to copyright in the United States.

J. W. Fowler (✉)
National Institute of Standards and Technology, 325 Broadway MS 686.02, Boulder, CO 80305, USA
e-mail: joe.fowler@nist.gov

likelihood estimators, which have substantially less bias—and for certain parameters, none at all.

In the low-temperature detector community, we often assess the performance of a sensor or a multiplexing readout by estimating the energy resolution from measurements of a gamma-ray line (such as the 97 and 103 keV lines of ^{153}Gd) or of a fluorescence line (such as the Mn $K\alpha$ line at 5.9 keV). This assessment requires fitting a histogram of energies to find the energy resolution, among other parameters. Clearly, any bias in estimating the resolution can produce misleading results.

Many methods appear in the literature for fitting models to histograms. Several articles explore their biases in the special case where N independent observations are assumed to measure the same underlying Poisson parameter μ . Fitting models such as Gaussians to observations is different from fitting a single unknown constant, however, and it is hard to generalize from the published results to the biases on estimates of energy resolution. The goal of this article is twofold: to establish that maximum-likelihood estimators are the best choice for unbiased estimates of spectral resolution, and to show that such estimators are not inconsistent with fast and convenient computation.

2 Possible Cost Functions for Fitting Models to Histograms

When estimating (“fitting”) parameters of a model from a data set, one generally chooses a cost function $C(\mathbf{p}; \mathbf{x})$ that depends on the model parameters \mathbf{p} and the observations \mathbf{x} . The “best” parameters $\hat{\mathbf{p}}$ are those which minimize C given the data \mathbf{x} . For measurements normally distributed with known variances about the (unknown) true values, the standard χ^2 statistic is the appropriate cost function. If the variances are uncorrelated with the true values, then minimizing χ^2 also maximizes the likelihood function.

The situation is less clear when the measured data are Poisson distributed, such as the bin contents in a histogram. The usual χ^2 statistic is a weighted sum of the squared difference between measured and modeled values, with weights equal to the inverse variance of each value. In Poisson distributed data, the variance equals the expected value in each measurement, so should the inverse weight be set equal to the measured or the modeled number of events in each bin? In fact, either choice produces parameter biases.

Consider a set of N measurements, the contents of the N bins in a histogram. Let c_i denote the measured counts in bin i . The parameterized model for the i th value is $y(x_i; \mathbf{p})$, or y_i for short. If the \mathbf{c} were normally distributed (which they are not), then the appropriate cost function would be $\chi_{\text{normal}}^2 \equiv \sum_{i=1}^N w_i (c_i - y_i)^2$ with each weight w_i equal to the inverse variance of measurement c_i . Following the convention of Baker and Cousins [2], the respective choices of $1/c_i$ or $1/y_i$ as the weights for Poisson data are called Neyman’s χ_{N}^2 and Pearson’s χ_{P}^2 (Eqs. 1 and 2). In the former case, the weight must be modified to $1/\max(1, c_i)$ to avoid divergences if any bins are empty (any $c_i = 0$). Neyman’s χ_{N}^2 is the simplest statistic to use in a general-purpose least-squares fitting algorithm, because its weights are independent of the model. Unfortunately, χ_{N}^2 is also the most biased.

Table 1 Five functions of both data and model that can be minimized with respect to model parameters to yield fits to a histogram. The last, χ_{MLE}^2 , is best for Poisson c_i . It derives from the Poisson-data likelihood function $\mathcal{L}_P = \prod_{i=1}^N e^{-y_i} y_i^{c_i}/c_i!$, with constant factors (i.e., factors independent of the model y) omitted

| | |
|--|-----|
| $\chi_N^2 \equiv \sum_{i=1}^N \frac{(c_i - y_i)^2}{\max(c_i, 1)}$ | (1) |
| $\chi_P^2 \equiv \sum_{i=1}^N \frac{(c_i - y_i)^2}{y_i}$ | (2) |
| $\chi_Y^2 \equiv \sum_{i=1}^N \frac{[c_i + \min(c_i, 1) - y_i]^2}{c_i + 1}$ | (3) |
| $\chi_{\text{GMLE}}^2 \equiv \chi_P^2 + \sum_{i=1}^N \log y_i$ | (4) |
| $\chi_{\text{MLE}}^2 \equiv -2 \log \mathcal{L}_P = 2 \sum_{i=1}^N (y_i - c_i \log y_i)$ | (5) |

Mighell [3] advocates another statistic with the practical advantages of χ_N^2 , yet constructed specifically to avoid bias at any expected number of counts: χ_Y^2 (Eq. 3). Hauschild and Jentschel [4], however, point out that χ_Y^2 estimators are unbiased only in the asymptotic limit of many observations N but are biased for finite N . They consider (without advocating) another cost function, the ‘‘Gauss max-likelihood’’ (Eq. 4) which improves on the χ_P^2 in accounting for an additional factor in the Gaussian likelihood function.

The cost function that introduces the least bias is the negative logarithm of the likelihood function for Poisson-distributed data (Table 1; Eq. 5). This quantity, χ_{MLE}^2 , is also known in x-ray astronomy as the Cash C -statistic [5]. If an additional term is added, a sum over nonzero observations, then the resulting $\chi_\lambda^2 \equiv \chi_{\text{MLE}}^2 + 2 \sum_{i=0, c_i > 0} (c_i \log c_i - c_i)$ is asymptotically distributed [2] as χ^2 for large N . This term, being independent of the model y , does not affect parameter fits. There is some disagreement in the literature over the best statistic to use in goodness-of-fit tests, with some authors[4] arguing for χ_P^2 and others [2,6–8] favoring χ_λ^2 .

3 Biases on the Histogram Area and Width

The simplest model to study is that of an N -bin histogram with an unknown, constant level: $y_i = \mu$. In this case, the best-fit estimate $\hat{\mu}$ for all of the cost functions can be expressed in closed form. The MLE method is the only one that yields the data mean, $\hat{\mu} = (\sum c_i)/N$, which is an unbiased estimator of μ . For the other cost functions, the estimators are: $\hat{\mu}_N = N'/\sum 1/\max(c_i, 1)$; $\hat{\mu}_P = \sqrt{\sum c_i^2/N}$; $\hat{\mu}_Y = N'/\sum 1/(c_i + 1)$; and $\hat{\mu}_{\text{GMLE}} = \sqrt{\hat{\mu}_P + \frac{1}{4} - \frac{1}{2}}$. Here, N' signifies how many c_i (out of N) are non-zero. From these expressions, it is simple to compute the estimators for simulated Poisson data; the results are shown in Fig. 1. With $N \gtrsim 5$ (left and center panels), the Pearson $\hat{\mu}_P$ has a positive bias as large as +0.5 counts when the true value $\mu \gtrsim 1$, while the Neyman $\hat{\mu}_N$ has a negative bias of approximately -1 count for $\mu \gtrsim 2$. In the asymptotic limit $N \rightarrow \infty$, the other three estimators are unbiased. In the opposite limit, where only one value is used to estimate μ , $\hat{\mu}_P$ and $\hat{\mu}_N$ are unbiased, while for $\mu \gtrsim 1$, $\hat{\mu}_{\text{GMLE}}$ and $\hat{\mu}_Y$ have biases of +1 and -0.5 counts.

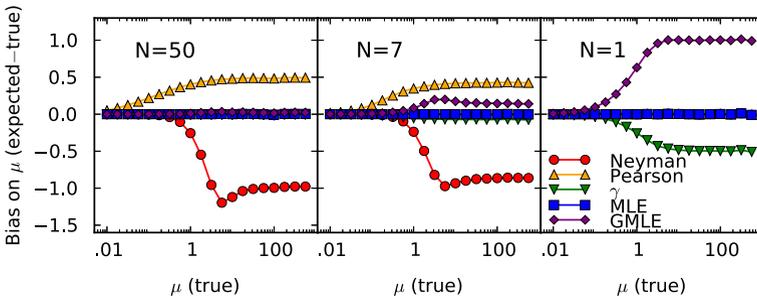


Fig. 1 The expected additive bias on μ (expected–true) for five different cost functions. The model is the trivial case where all data estimate the same constant value of μ . The *three panels* have the indicated number N of Poisson variates. The $N = 50$ case is not visibly different from the $N \rightarrow \infty$ limit (Color figure online)

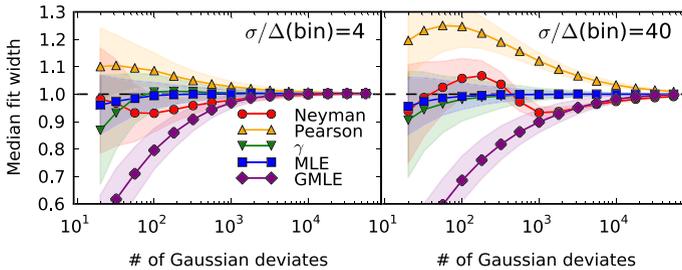


Fig. 2 Median width in fit from unit-variance Gaussian deviates histogrammed with wide and narrow bin widths $w = \sigma/4$ and $w = \sigma/40$. *Shaded bands* contain the middle two quartiles of the 8,000 fits per point. The simulated data have no constant background, and none is fit. In all cases, the bias is largest if the total number of events N_e is small and if narrow bins (*right panel*) produce a small number of events per bin. The Poisson Maximum Likelihood (“MLE”) fit is the least biased of the methods, though it shows a small bias at $\lesssim 100$ events. Mighell’s χ^2_γ is also acceptable for $\gtrsim 10$ events per bin at the peak (Color figure online)

For more complicated model functions, these results generalize readily if the model contains a parameter (or combination of them) that allows the entire model to scale freely by a factor. The new statement is that χ^2_N and χ^2_P produce biased estimators of a histogram’s area when more than a few bins are used, and χ^2_γ and χ^2_{GMLE} produce biased area estimators when not many bins are used. Only the MLE value is unbiased for any area and all N . In some circumstances, biases on the fitted area might be unimportant. Still, using the one estimator mathematically guaranteed not to bias the area seems prudent in any conditions.

The width of a spectral peak is another parameter often estimated from Poisson-distributed histogram data. Figure 2 shows the median $\hat{\sigma}$ fit to simulated Gaussian deviates with 4 or 40 bins per σ . The most important result is that the χ^2_γ and χ^2_{MLE} estimators are the only ones without significant bias on $\hat{\sigma}$ for fits to $\gtrsim 100$ Gaussian deviates. Below 100 events (in either binning), a small underestimate in width is apparent. In such conditions of small peaks, Monte Carlo modeling of the fit is probably necessary.

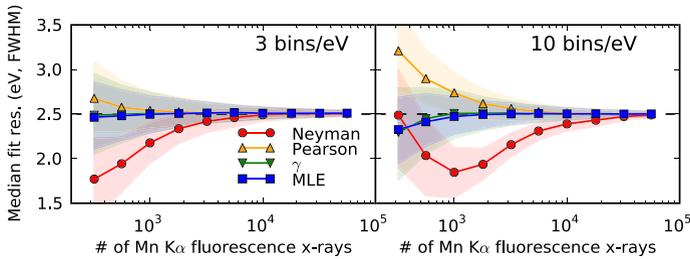


Fig. 3 The median fit result for energy resolution using the Mn $K\alpha$ complex, with true energy resolution 2.5 eV (FWHM). Shaded bands contain the middle two quartiles of the 3,000 fits per point. The data are binned with 3 and 10 bins per eV, putting approximately 4 or 1 % of all values in the histogram's peak bin. A constant background per bin varies freely in the fit, but the true value is 0. The usual Neyman χ^2_N severely underestimates the energy resolution, while Pearson's overestimates it. Achieving low bias is possible with the other χ^2_γ and MLE cost functions, given $\gtrsim 10$ counts per bin at the peak (Color figure online)

Figure 3 shows the median Gaussian energy resolution (FWHM) estimated by fitting the Mn $K\alpha$ x-ray fluorescence spectrum [9] to simulated energies in the range 5,860–5,930 eV. Each point is the result of fitting 3,000 simulations. The true resolution is 2.5 eV. As with fits to the width of Gaussian distributions, the χ^2_N and χ^2_P estimators exhibit substantial bias and cannot be recommended. (The same is true of χ^2_{GMLE} though its minimizer often failed to converge, and it is not shown here.) In this estimation problem, both χ^2_{MLE} and χ^2_γ lead to minimal bias, though care is again required with spectra containing fewer than $\sim 1,000$ photons.

In these tests of fitting parameters related to the area under a histogram peak and to its width, the biases of the earliest and simplest methods (χ^2_N and χ^2_P) are manifest. The χ^2_γ approach seems to have low bias on the width parameters, but the area bias in Figure 1 is a cause for concern. Only the maximum-likelihood estimators perform well on all bias tests.

4 Efficient Maximum-Likelihood Parameter Estimation

If the maximum-likelihood estimators are preferred, how can they be computed efficiently? Some software computes them without modification; for example, the CERN library MINUIT. More often, software libraries or packages offer only the extremes of fully general nonlinear minimization (which can be slow or fragile, or both) and “least-squares minimization.” Routines of the latter sort can find the minimum of a χ^2 with constant (i.e., model-independent) weights, as in Eq. 1 or 3. Examples include `gsl_multifit_fdfsolver_lmshd` in `gsl`, `Fitmrq` in *Numerical Recipes* [10], `scipy.optimize.leastsq` in Python, and `leasqr` in Octave. They assume that the nonlinear function $\chi^2(y(\mathbf{p}))$ is strictly quadratic in the set of predicted values $\{y_i\}$, while $y_i(\mathbf{p})$ may be a general nonlinear function. An effective algorithm for least-squares fitting is the Levenberg-Marquardt (LM) method [11, 12], which can outperform general nonlinear minimizers by exploiting its knowledge that the target function is a sum of squares. A least-squares fitter is not appropriate for minimizing a function like χ^2_{MLE} , which is not quadratic in y_i . Fortunately, it is possible

to find maximum-likelihood solutions either by using a slightly modified LM-based fitter or by iterating the results generated by a standard least-squares fitter.

4.1 Modified Levenberg-Marquardt Method

The LM algorithm is commonly identified as a least-squares fitter because it is so widely used for this one purpose. In fact, LM is readily adapted to minimize any composite function $f(y(\mathbf{p}))$ even when f is not quadratic in y , provided that we can compute the first and second derivatives of f . Using the labels and scaling of *Numerical Recipes*, for the Poisson maximum-likelihood $f = \chi_{MLE}^2$, these are:

$$\beta_j = -\frac{1}{2} \frac{\partial f}{\partial p_j} = \sum_{i=1}^N \left(\frac{c_i}{y_i} - 1 \right) \frac{\partial y_i}{\partial p_j}, \tag{6}$$

$$\alpha_{jk} = \frac{1}{2} \frac{\partial^2 f}{\partial p_j \partial p_k} \approx \sum_{i=1}^N \left(\frac{c_i}{y_i^2} \right) \frac{\partial y_i}{\partial p_j} \frac{\partial y_i}{\partial p_k}. \tag{7}$$

The equality is only approximate because terms containing second derivatives of y_i are dropped; they generally have a destabilizing or minimal effect on the LM performance. Adapting the standard LM to achieve a Poisson maximum-likelihood fit instead requires only three changes: using the parenthesized factors in Eqs. 6 and 7 in place of $[w_i(c_i - y_i)]$ and $[w_i]$ respectively in the standard χ^2 minimization, and computing χ_{MLE}^2 (Eq. 5) in place of χ^2 . When starting from source code that implements the basic LM least-squares algorithm, these replacements should be straightforward [13]. This is the approach used for the fits summarized in Sect. 3.

Incidentally, by setting $\beta_0 = 0$ in Eq. 6, it is easy to show that a model of the form $y = p_0 f_i(p_1, \dots, p_m)$ guarantees $\sum_i c_i = \sum_i y_i$. That is, a model having an overall scale factor as one parameter is guaranteed to have equal observed and predicted total counts in a maximum-likelihood fit [2].

4.2 Iterative Least-Squares Method

It is also possible to approach the Poisson maximum-likelihood fit using only the framework of a standard LM least-squares fitter by employing an additional “outer iteration” over least-squares fits. Compare the desired MLE solution (set Eq. 6 to zero) with

$$0 = \frac{\partial \chi_{normal}^2}{\partial p_j} = \sum_{i=1}^N w_i (c_i - y_i) \frac{\partial y_i}{\partial p_j}. \tag{8}$$

If the predicted counts \hat{y}_i arise from the maximum-likelihood model, then by definition they solve Eq. 6. The same model solves Eq. 8 if the weights are taken to be $w_i \equiv 1/\hat{y}_i$. This suggests an iterative approach to solving the desired Eq. 6 employing a least-squares solver capable only of solving Eq. 8. Start by minimizing χ_N^2 , i.e., solve

Eq. 8 with $w_i = \max(1, c_i)$, to obtain an initial guess at the model parameters \mathbf{p}_0 . Use constant weights $w_i = 1/y_i(\mathbf{p}_0)$ to solve the least-squares equation again. This produces an improved set of parameters \mathbf{p}_1 and weights, from which another iteration can begin. This method appears to be discussed first by Wheaton [14]. In at least some conditions, two rounds of least-squares fits suffice to remove the bias introduced in the initial fit that minimized χ_N^2 .

5 Conclusion

Two practical approaches were given to maximize the Poisson likelihood and minimize χ_{MLE}^2 . It is important to do this rather than to minimize Neyman's χ_N^2 when estimating energy resolution from a spectrum, or the estimate will give an overly optimistic view of detector performance, with a bias comparable to the uncertainty on the resolution. Minimizing Pearson's χ_P^2 has the opposite effect. Estimates of shape parameters from a data histogram are best found through maximum-likelihood fits, which have lower bias than other approaches and are not difficult to perform.

Acknowledgments The author was supported by an American Recovery and Reinvestment Act senior fellowship and by the NIST Innovations in Measurement Science program. The author thanks J. Ullom for encouragement and many helpful discussions and C. Pryke for debates on the topic long ago.

References

1. S.D. Poisson, *Recherches sur la Probabilité des Jugement en Matière Criminelle et en Matière Civile* (Bachelier, Paris, 1837)
2. S. Baker, R.D. Cousins, Nucl. Instrum. Methods **221**, 437 (1984)
3. K.J. Mighell, Astrophys. J. **518**, 380 (1999)
4. T. Hauschild, M. Jentschel, Nucl. Instrum. Methods A **457**, 384 (2001)
5. W. Cash, Astrophys. J. **228**, 939 (1979)
6. J. Neyman, E.S. Pearson, Biometrika **20A**, 175 (1928)
7. S.S. Wilks, Ann. Math. Stat. **6**, 190 (1935)
8. R.A. Fisher, Biometrics **6**, 17 (1950)
9. G. Hölzer et al., Phys. Rev. A **56**, 4554 (1997)
10. W.H. Press, S.A. Teukolsky, W.T. Vetterling, B.P. Flannery, *Numerical recipes: the art of scientific computing*, 3rd edn. (Cambridge University Press, New York, 2007)
11. K. Levenberg, Q. Appl. Math. **2**, 164 (1944)
12. D.W. Marquardt, SIAM J. Appl. Math. **11**, 431 (1963)
13. T.A. Laurence, B.A. Chromy, Nature Methods **7**, 338 (2010)
14. W.A. Wheaton et al., Astrophys. J. **438**, 322 (1995)