Comparison of experimental data to Monte Carlo simulation—Parameter estimation and goodness-of-fit testing with weighted events

G. Bohm a, G. Zech b,*

a Deutsches Elektronen-Synchrotron DESY, D-15738 Zeuthen, Germany
b Universit"at Siegen, D-57068 Siegen, Germany

Available online 26 June 2012

A R T I C L E   I N F O

Keywords:
Parameter estimation
Goodness-of-fit test
Monte Carlo simulation
Least square fit
Maximum likelihood fit
Weighted events

A B S T R A C T

Relations are derived that can be used to infer parameters in situations where theoretical predictions can be compared to experimental data only indirectly via Monte Carlo simulation. We consider least square and likelihood ratio fits. Parameter changes in the fitting procedures are performed by re-weighting Monte Carlo events. Formulas for goodness-of-fit tests based on the $\chi^2$ and the likelihood ratio statistic are presented.

© 2012 Elsevier B.V. All rights reserved.

1. Introduction

This article follows to a large extend a section in the textbook [1] which in turn is based on an older manuscript [2]. The content has been extended, a numerical example has been added and two truncated formulas in Ref. [1] have been corrected. The article is motivated by a recent series of publications [3], the results of which rely on the false assumption that the median of a sample of $\chi^2$ distributed random numbers again follows a $\chi^2$ distribution. As a consequence, independent of the amount of data, the corresponding results are biased. In Ref. [3] $\chi^2$ goodness-of-fit testing and parameter estimation based on the $\chi^2$ statistic of weighted events are discussed. In addition to these topics we treat the likelihood ratio test and parameter estimation with the maximum likelihood method.

In most experiments acceptance losses and resolution effects have to be simulated by Monte Carlo programs. Therefore, the measurements cannot be described by analytic probability density functions (p.d.fs). We compare our data in the form of histograms to the simulation for different values of the parameter, to re-weight the Monte Carlo events by $f(x|\theta)f(x|\theta_0)$ and then compare the re-weighted sample to the data. Therefore, we have to remember for each simulated event the true value $x$ of the variable. Histograms are formed with the measured variable $x'$ and a test statistic based on a $\chi^2$ approximation or on the Poisson likelihood of the number of events in a bin is formed. There are also situations where the experimental events are weighted, for instance to correct for acceptance losses. However, usually we prefer not to weight the data but to simulate the losses.

In Ref. [3] the weights are restricted to functions of $x$ but the results have been applied to situations where this condition is violated, i.e. where resolution effects have to be considered.

When we estimate parameters and their errors in a least square fit, we usually rely on the validity of the approximation that the test statistic can be described by a $\chi^2$ distribution. This assumption is based upon the following conditions: (1) The entries in all bins of the histogram are normally distributed. (2) The expected number of entries depends linearly on the free parameters in the considered parameter range. (An indication for a non-linearity is asymmetric errors of the adjusted parameters.) (3) The estimated uncertainties in the denominators of the summands of $\chi^2$ are independent of the parameters. These conditions are also important if we apply a $\chi^2$ goodness-of-fit test of a composite hypothesis.

In the following we will assume that these conditions are satisfied, even though they are rarely realized exactly. Therefore, we have to check our least square fit results. For a Poisson likelihood fit, the requirements are more relaxed.
There are many powerful goodness-of-fit tests on the market. We restrict ourselves to the $\chi^2$ test and the likelihood ratio test but our approach can be generalized to other tests as well. For goodness-of-fit tests, the $\chi^2$ approximation is not always essential. Often the distribution of the test statistic can be evaluated through a Monte Carlo simulation. This is especially important for small samples and the evaluation of small $p$-values.

The $p$-value for the $\chi^2$ statistic is defined by the relation

$$ p = \int_{\chi^2}^{\infty} f_{\chi^2}(u) \, du $$

where $\chi^2$ is the observed value of the statistic and $f_{\chi^2}(u)$ is the $\chi^2$ distribution for $n$ degrees of freedom. The $p$-value is the fraction of experiments that observe a value $u$ larger than $\chi^2$ in an infinite number of experiments that are correctly described by the theoretical prediction of the distribution of $x$. Measurements that are not compatible with the prediction are supposed to lead to large values of $\chi^2$ and low $p$-values. Rejecting a prediction with $p \leq 0.05$ would reject in 5% of all measurements a correct prediction. Because we do not want to reject a valid hypothesis, mainly small $p$-values are of interest which are evaluated from the tails of the $\chi^2$ distribution which may be sensitive to approximations that are used to estimate $\chi^2$.

In the following section we compare measured histograms to theoretical predictions. In Section 3 we turn to the case where the prediction is given in the form of a Monte Carlo simulation neglecting the statistical uncertainty of the simulation. In Section 4 the fluctuations of the simulated numbers are included. A special way to speed-up the parameter estimation problem is presented in Section 5. Finally, we summarize the results and give recommendations for their application.

2. Comparing a histogram to a theoretical prediction

2.1. Unweighted events

In the following we assume that the number of entries in the histogram follow independent Poisson distributions or a multinomial distribution. In the first case the probability $W$ to observe $n_i$ entries in bin $i$ in a histogram with $B$ bins is

$$ W(n_1, \ldots, n_B | t_1, \ldots, t_B) = \prod_{i=1}^{B} e^{-t_i} \frac{t_i^{n_i}}{n_i!}. $$

In the second case, in this relation the constraint $N = \sum_{i=1}^{B} n_i = \sum_{i=1}^{B} t_i = t$ has to be fulfilled. Thus the normalized multi-Poisson distribution is the multinomial distribution with probability $p_i = t_i / t$ for an event to fall into bin $i$. We assume that the numbers $t_i$ are predicted. They can be functions of parameters. If we have an absolute prediction, we apply (2) directly, if only the shape of the histogram is predicted, we normalize the prediction in (2).

To simplify the writing, we omit occasionally in the formulas the limits of sums where they are obvious. To start with, we repeat some simple relations which are common knowledge and which we need later.

The log-likelihood for the predictions $t_i$ for a histogram filled with Poisson distributed event numbers is

$$ \ln L = \sum_{i=1}^{B} [n_i \ln t_i - t_i - \ln n_i!]. $$

Here and in all the following formulas of $\ln L$ the term $\ln n_i!$ can be dropped because it does not depend on $t_i$.

The likelihood ratio statistic $V$ for a histogram filled with Poisson numbers $n_i$ with predictions $t_i$ is the ratio of the maximal likelihood under the condition that the $n_i$ are drawn from Poisson distributions $P(n_i | t_i)$ with means $t_i$ to the unconditional maximum of the likelihood $P(n_i | n_i)$ for the observations. We use the statistic

$$ V = \ln V' : $$

$$ V = \sum_{i=1}^{B} [n_i \ln t_i - t_i - \ln n_i + n_i]. $$

In the limit where we can approximate the Poisson distribution by a normal distribution under the null hypothesis $H_0$ that the numbers $n_i$ are described by Poisson distributions with mean values $t_i$, the statistic $\chi^2$, defined as

$$ \chi^2 = \sum_{i=1}^{B} \frac{(n_i - t_i)^2}{t_i} $$

follows by definition a $\chi^2$ distribution with $B$ degrees of freedom. The corresponding log-likelihood in the Gaussian approximation omitting the term not depending on $t_i$ is

$$ \ln L = \sum_{i=1}^{B} - \frac{(n_i - t_i)^2}{2 t_i} - \frac{1}{2} \ln t_i $$

$$ \ln L = -\frac{\chi^2}{2} + \frac{1}{2} B \sum_{i=1}^{B} \ln t_i. $$

For large values $t_i$ where the Gaussian approximation holds, the logarithmic term is an extremely slowly varying function of $t_i$. It can safely be neglected if we are interested in parameter inference. We get $\chi^2 = 2 \ln L$. In this approximation the results of a likelihood analysis and a least square fit of a parameter are equivalent. Also the likelihood ratio statistic $V$ (4) approaches asymptotically $-\chi^2/2$.

If only the shape of the histogram is predicted, we normalize the prediction and have $B - 1$ degrees of freedom. We have to replace in all relations $t_i$ by $c t_i$ with the normalization constant $c = \Sigma n_i / 2 \Sigma t_i$. The probability to fall into bin $i$ is $p_i = t_i / 2 \Sigma t_i$ and we arrive at Pearson’s $\chi^2$ statistic

$$ \chi^2 = \sum_{i=1}^{B} \frac{(n_i - N p_i)^2}{N p_i}, \quad N = \sum_{i=1}^{B} n_i. $$

The Poisson approach where the prediction is normalized to the total number of entries and Pearson’s multinomial approach are of course equivalent (see also [4]). The former is more general in that we can apply it also to absolute predictions ($B$ degrees of freedom), we can fit the normalization or we can apply the normalization that corresponds to the multinomial distribution. In both cases we have $B - 1$ degrees of freedom. If we estimate $J$ parameters $\theta_j$, $j = 1, \ldots, J$, and replace $t_i$ by the estimates $\hat{t}_i$, another $J$ degrees of freedom are lost. The sum in Eq. (5) can be described approximately by a $\chi^2$ distribution of $B - 1 - J$ degrees of freedom. The conditions for the validity of this approximation are discussed in Ref. [5].

2.2. Weighted events

We consider the distribution

$$ f(v_1, \ldots, v_K) = P(K | \kappa) \prod_{k=1}^{K} g(v_k) $$

where $g(v)$ is a continuous p.d.f. for the weight of an event and $P(K | \kappa)$ the Poisson distribution of $K$ with mean $\kappa$. We observe $K$ events each with a random weight drawn from $g(v)$. The expectation values $E(v)$ and $E(v^2)$ exist and the ratio $E(v^2) / E(v)$ is of the order of one in our applications. We are interested in the sum of the weights $N = \sum_{k=1}^{K} v_k$. The expected value and variance of $N$ are $E(N) = \kappa E(v)$ and $\text{Var}(N) = \kappa E(v^2)$. To simplify the writing of some
formulas, we introduce the concept of equivalent number of unweighted events \( \hat{n} \). The variable \( \hat{n} = n \bar{E}(v)/\bar{E}(v^2) \) has the property \( \bar{E}(n) = \sqrt{\bar{E}(\hat{n})} \) which is characteristic for the Poisson distribution. The name equivalent number of unweighted events indicates that the \( n \) weighted events have the same statistical significance as \( \hat{n} \) unweighted events. As in our case events are generated by a very complex random process, we neither know \( \kappa \) nor the p.d.f. of the expected number of events, \( E(n) \), but only \( K \) and the \( K \) weights \( v_k \). We have to approximate the expected values by its empirical estimates \( n = \sum_{k=1}^{K} v_k \) and \( \bar{E}(n) = \sum_{k=1}^{K} 1 v_k^2 \) and

\[
\hat{n} \approx \left( \sum_{k=1}^{K} v_k \right)^2 / \sum_{k=1}^{K} v_k^2.
\] (10)

In the following we form a \( \chi^2 \) expression where the Poisson distribution with mean \( \kappa \) is approximated by a Gaussian with variance \( \kappa \). In the asymptotic limit of \( K \) also the distribution of \( \hat{n} \) will according to the central limit theorem be described asymptotically by a normal distribution with variance \( \bar{E}(\hat{n}) \). This is the essential property required to form the \( \chi^2 \) expressions. We have to replace the Poisson random variables by the equivalent numbers of unweighted events.

The treatment of weighted events is now straightforward, we just replace in all formulas the Poisson random variables by the equivalent numbers of unweighted events. For example, if the experimental events are weighted and there are predictions \( t_i \) for the sums \( n_i \), we get

\[
\chi^2 = \sum_{i=1}^{K} \frac{\left( \bar{E}(n_i)/\bar{E}(v_i) - c \right)^2}{c t_i}
\] (11)

where the numbers \( c_i \) are chosen such that the predictions are correctly normalized:

\[
c_i = \frac{n_i}{\bar{E}(n_i)} \quad c = \frac{\sum n_i}{\sum t_i}.
\] (12)

Here \( c \) normalizes the weighted sums \( n_i \) to the predictions \( t_i \), as previously and \( c_i \) includes the scaling of \( n \) to \( \hat{n} \). Again, we had to replace the expected value \( \bar{E}(n_i)/\bar{E}(v_i) \) by its actually observed value \( n_i/n_i \). With the initial variables we obtain:

\[
\chi^2 = \sum_{i=1}^{K} \frac{\left( \bar{E}(n_i)/\bar{E}(v_i) - c \right)^2}{c t_i n_i}.
\] (13)

The relation (13) can of course also be derived directly without the detour that we have taken. The denominator is the estimated variance of the root of the numerator under \( H_0 \) because the expected value of \( n_i \) is \( c t_i \), and the variance \( c t_i \bar{E}(v^2) \).

The validity of our approximations depends on how well the \( \hat{n} \) are described by the normal distribution just as for \( n_i \) in the situation where all events have weights equal to one. Clearly, the approximations improve with the increasing number of events. However, the minimum number of events required per bin cannot be fixed for all applications. For goodness-of-fit tests with low \( p \)-values the requirements are rather stringent if we rely on the \( \chi^2 \) distribution. If we determine the distribution of the statistic by a simulation, then we can relax our requirements. If we estimate parameters and determine the parameter errors from the anticipated \( \chi^2 \) dependence on the parameters, we should be rather careful. The uncertainties introduced by the approximation should be small compared to uncertainties attributed to the fitted parameters. It is recommended to verify the validity of the obtained results by simulations, variation of the amount of data or by bootstrap techniques.

Empirically, the distribution of \( \hat{n} \) follows to a rather good approximation a Poisson distribution [1] for weight distributions with not too large variations of the weights, \( \bar{E}(v^2)/\bar{E}(v^2) < 2 \). For a ML fit of a histogram of weighted events we do not want to rely on the Poisson approximation but for testing, we can build a test statistic analogous to the likelihood ratio statistic. There the precise form of the distribution of \( \hat{n} \) is not essential because we determine the distribution of the test statistic by simulation.

3. Monte Carlo prediction with negligible uncertainty

Now we treat the case where the prediction is given in the form of a sample of Monte Carlo events. When we perform a quantitative comparison of an experimental histogram containing \( n_i \) events in bin \( i \) with a simulated one, the situation is relatively simple if we can neglect the statistical error of the simulation. The statistics of weighted events does not enter. We simply replace in Eqs. (3) and (5) the prediction \( t_i \) by the normalized number of simulated entries \( m_i \):

\[
\ln L = \sum_{i=1}^{B} \left[ n_i \ln(c_m)m_i - c_m \ln(n_i) \right]
\] (14)

\[
\chi^2 = \sum_{i=1}^{B} \frac{(n_i-c_m)^2}{c_m n_i}.
\] (15)

For an absolute prediction, \( c \) is given (\( B \) degrees of freedom). If only the shape is predicted, then \( c \) can be adjusted in the fit or fixed by \( c = \sum n_i/\sum m_i \) (\( B-1 \) degrees of freedom). If \( J \) parameters have been adjusted in addition, then we have \( B-J-1 \) degrees of freedom.

3.1. Goodness-of-fit tests

The expression (15) can be used for a \( \chi^2 \) goodness-of-fit test. For the likelihood ratio test of the histogram, we insert the Monte Carlo prediction into Eq. (4) and obtain the test statistic \( V \):

\[
V = \sum_{i=1}^{B} \left[ n_i \ln c_m m_i - c_m \ln(n_i + n_i) \right]
\] (16)

The null hypothesis \( H_0 \) is that the data consisting of \( N \) events are drawn from \( f(x|\hat{\theta}_0) \). The test proceeds as follows: The quantity \( V \) is computed. The distribution of the test statistic \( V \) has to be generated numerically: \( K \) times an event sample following \( f(x|\hat{\theta}_0) \) of \( N \) events and a Monte Carlo sample are generated. Each time a test statistic \( V_k \) is computed. For a test of size \( \alpha \), \( H_0 \) is rejected if \( V \) is larger than the fraction \( \alpha \) of the values \( V_c \). The test is exact in the limit \( K \to \infty \). For practical applications we require \( 1/\sqrt{2K} < 1 \).

If parameters have been adjusted, we have a composite hypothesis. The null hypothesis is that the data are drawn from \( f(x|\hat{\theta}) \), \( \hat{\theta} \) being confined to some physical range. \( V \) is computed from the data sample and the weighted Monte Carlo sample with adjusted weights, i.e. with the fitted parameters. The experiment is simulated \( K \) times. For each couple of event sample generated with the fitted parameters and Monte Carlo sample the parameters are estimated and the statistic \( V_k \) is computed. A test is applied as in the situation where the parameters are fixed.

3.2. Parameter estimation

Relations (14) and (15) can be used for parameter estimation. The Poisson likelihood fit where \( \ln L(\hat{\theta}) \) is maximized is to be preferred if the numbers \( n_i \) are small. To change parameters in the fit, we have to modify the Monte Carlo predictions \( m_i(\hat{\theta}) \) for each

---

1 To avoid the use of too many variables, we do not distinguish between a random variable and its realization.
optimization step. As explained in the Introduction, to modify the prediction, we weight the Monte Carlo event \( k \) by \( w_k = f(x_k | b) / f(x_k | b_0) \). Here \( x \) represents the set of the true kinematic variables, while the histogram is formed with the measured variables \( x' \) which determine the bin index \( i \). The parameters are adjusted such that the histogram of the weighted events describes the experimental histogram. The prediction for bin \( i \) is \( c_i = c \sum_k w_k (b) \), where the sum runs over all \( K \) events in bin \( i \).

Re-weighting does not only avoid the repetition of the full simulation with every fitting step. Apart from the fact that we want to avoid the computational effort of the detector simulation, there is another more important reason: With the \( \chi^2 \) fit we find the standard error interval by letting vary \( \chi^2 \) by one unit or the log-likelihood by half a unit. On the other hand, when we compare the experimental data with an optimal simulation, we expect a contribution to \( \chi^2 \) from the simulation of the order of \( \sqrt{2BN/M} \) for \( B \) histogram bins, \( N \) experimental events and \( M \) simulated events. Even with a simulation sample which is a hundred times larger than the data sample this value is of the order of one. This means that a repetition of the simulation causes considerable fluctuations of the \( \chi^2 \) value which have nothing to do with parameter changes. These fluctuations can only be reduced if the same Monte Carlo sample is used for all parameter values. We have to adjust the simulation to the modified parameters by weighting the simulated events. Also re-weighting produces additional fluctuations which may become important for histograms with a large number of bins and large variations of the weights within the error interval around the fitted parameter value. (Condition 3 stated in the Introduction is violated.) This effect should be tolerable in the majority of cases but should be estimated by varying the number of experimental events and simulated events. In the following section we take into account the statistical fluctuations of the simulation. This does not eliminate the problem related to the change of weights discussed above. In Ref. [3] these uncertainties are ignored.

4. Including the simulation uncertainty

4.1. Comparison of two Poisson numbers with different normalization

We compare two Poisson numbers \( n, m \) with different normalization constants \( c_n, c_m \), such that \( c_n n \) should be compatible with \( c_m m \). The null hypothesis \( H_0 \) is that \( n \) is drawn from a Poisson distribution with mean \( \lambda / c_n \) and \( m \) from a Poisson distribution with mean \( \lambda / c_m \). We form a \( \chi^2 \) expression

\[
\chi^2 = \frac{(c_n n - c_m m)^2}{\lambda (c_n + c_m)}
\]

(17)

where the denominator is the variance of the bracket in the numerator under \( H_0 \), where the expected value of the numerator is zero.

On the other hand, the log-likelihood of \( \lambda \) is according to Eq. (3)

\[
\ln L(\lambda) = n \ln \frac{\lambda}{c_n} + \frac{\lambda}{c_n} + m \ln \frac{\lambda}{c_m} + \frac{\lambda}{c_m} + \text{const}
\]

(18)

with the ML estimate

\[
\hat{\lambda} = \frac{c_n c_m}{c_n + c_m} (n + m).
\]

Replacing \( \hat{\lambda} \) by its estimate and inserting the result into Eq. (17), we obtain

\[
\chi^2 = \frac{1}{c_n c_m} \frac{(c_n n - c_m m)^2}{n + m}.
\]

(20)

The expected variance of \( c_n n - c_m m \) is proportional to the sum of the Poisson numbers \( n \) and \( m \). To compute the logarithm of the likelihood ratio, we need the full likelihood similar to Eq. (3)

\[
\ln L = n \ln \frac{\lambda}{c_n} - \frac{\lambda}{c_n} - \ln n! + m \ln \frac{\lambda}{c_m} - \frac{\lambda}{c_m} - \ln m!
\]

(21)

To simplify the writing, we have replaced \( \hat{\lambda} \) again by \( \lambda \).

For a pair of Poisson numbers \( n, m \) the likelihood ratio test statistic \( V \) is the logarithm of the ratio of the maximal likelihood under the condition that the two numbers are drawn from the same distribution to the unconditional maximum of the likelihood for the observations. With the predictions \( \lambda / c_n \) and \( \lambda / c_m \), respectively, the sum of the contributions from the experimental data and from the simulation to our test statistic \( V \) is analogous to Eq. (4):

\[
V = n \ln \frac{\lambda}{c_n} - \frac{\lambda}{c_n} - n \ln n! + m \ln \frac{\lambda}{c_m} - \frac{\lambda}{c_m} - m \ln m!.
\]

(22)

4.2. Comparison of two histograms

We add the \( \chi^2 \) contributions of all bins:

\[
\chi^2 = \sum_{i=1}^{B} \frac{1}{c_n c_m} \frac{(c_n n_i - c_m m_i)^2}{n_i + m_i}.
\]

(23)

The ratio \( c_n / c_m \) is either given (\( B \) degrees of freedom) or the histograms are normalized to the total number of entries \( N = \sum n_i, M = \sum m_i \) of the two histograms (\( B-1 \) degrees of freedom). Then we can set \( c_n = M, c_m = N \). We may also leave \( c_n / c_m \) as a free parameter and determine it in a fit. Notice that the normalization constants are defined up to a common factor, only the relative normalization \( c_n / c_m \) is relevant.

With \( \lambda_i = (n_i + m_i) / (1 / c_n + 1 / c_m) \). The log-likelihood is

\[
\ln L = \sum_{i=1}^{B} \left[ n_i \ln \frac{\lambda_i}{c_n} - \frac{\lambda_i}{c_n} - n_i \ln n_i! + m_i \ln \frac{\lambda_i}{c_m} - \frac{\lambda_i}{c_m} - m_i \ln m_i! \right].
\]

(24)

The corresponding relation for the likelihood ratio statistic \( V \) is

\[
V = \sum_{i=1}^{B} \left[ n_i \ln \frac{\lambda_i}{c_n} - \frac{\lambda_i}{c_n} - n_i \ln n_i + m_i \ln \frac{\lambda_i}{c_m} - \frac{\lambda_i}{c_m} - m_i \ln m_i + m_i \right].
\]

(25)

Relations (23) and (25) can be used to test the compatibility of the simulation with the experimental data. For parameter estimation we need to compare weighted histograms.

4.3. Comparison of weighted sums

When we compare the experimental data to a Monte Carlo simulation in order to determine parameters, the simulated events have to be weighted. We generalize our result (23) to the situation where both numbers \( n, m \) consist of a sum of individual events with weights, \( n = \sum_{k=1}^{N} w_k n_k, m = \sum_{k=1}^{M} w_k m_k \). We replace \( n, m \) by the equivalent numbers of unweighted events \( \hat{n} \) and \( \hat{m} \) and get for a single bin

\[
\chi^2 = \frac{1}{c_n c_m} \frac{(c_n \hat{n} - c_m \hat{m})^2}{\hat{n} + \hat{m}}.
\]

(26)

where now \( c_n, c_m \) are the relative normalization constants for the equivalent numbers of unweighted events, and as before \( c_n n \) is
supposed to agree within the statistical fluctuations with \( c_m \). With \( c_n = c_m = c_m \), we have

\[

c_n = c_n = \frac{\sum v_i^2}{\sum v_i}, \quad c_m = c_m = \frac{\sum w_i^2}{\sum w_i}
\]

(27)

**Remark.** In the previous subsection we have estimated \( \lambda \) using the maximum likelihood of the Poisson distributions. Asymptotically, we can replace the Poisson distribution by a normal distribution and vice versa. Therefore, the result for \( \lambda \) must hold also in the limit where \( n \) and \( m \) are normally distributed.

### 4.4. Comparison of two weighted histograms

We have to evaluate the expression (26) for each bin and sum over all \( B \) bins

\[
\chi^2 = \sum_{i=1}^{B} \left[ \frac{1}{c_n} \left( \frac{c_n \hat{n} - c_m \hat{m}}{n + m} \right)^2 \right]_i
\]

(28)

where the prescription indicated by the index \( i \) means that all quantities in the bracket have to be evaluated for bin \( i \). In case the entries are not weighted the tilde is obsolete. The overall constants \( c_n, c_m \) are equal for all bins of the histograms. If the histograms are normalized with respect to each other, we have as before \( c_n = 2\pi n_i, m_i = \Sigma \pi w_i \) and \( c_m = 2\pi n_i = \Sigma \pi v_i \).

In most applications where we compare experimental data to Monte Carlo simulations, only the latter are weighted and setting \( c_n = 1 \), relation (28) simplifies to

\[
\chi^2 = \sum_{i=1}^{B} \left[ \frac{1}{c_m} \left( \frac{n - \hat{m}}{n + m} \right)^2 \right]_i
\]

(29)

where \( c_m \) either is given, set to \( \Sigma n_i/\Sigma n, \) or left as a free parameter in the case of parameter estimation. As before, we have either \( B \) or \( B - 1 \) degrees of freedom. In Ref. [3] one additional degree of freedom is given away. As a consequence information is lost.

We can write down a pseudo log-likelihood

\[
\ln \hat{L} = \sum_{i=1}^{B} \left[ n \ln \frac{\hat{\lambda} - \lambda}{\hat{\lambda} - \ln n + \hat{m} \ln \frac{\hat{\lambda} - \lambda}{\hat{\lambda} - \ln \hat{m}} \right]
\]

(30)

The parameter \( \hat{\lambda} = (n + m \hat{m})/(1 + 1/c_m) \) is obtained from the general relation \( \hat{\lambda} = (\hat{n} + \hat{m})/(1/c_n + 1/c_m) \) with the mentioned simplifications and \( \hat{m} \) is to be evaluated using the Gamma function, \( \hat{m} = \Gamma(\hat{m} + 1) \). In Eq. (30) we have used the Poisson distribution to describe weighted events. This approximation is not adequate for ML fits but can be used for a modified likelihood ratio test.

### 4.5. \( \chi^2 \) Goodness-of-fit test

The expression (28) can be used for \( \chi^2 \) goodness-of-fit tests in the same way as (23).

With the same replacements as before, we obtain an approximate likelihood ratio test statistic for weighted events:

\[
V = \sum_{i=1}^{B} \left[ n \ln \frac{\hat{\lambda} - \lambda}{\hat{\lambda} - \frac{\hat{\lambda} - \lambda}{\hat{\lambda} - \ln n + \hat{m} \ln \frac{\hat{\lambda} - \lambda}{\hat{\lambda} - \ln \hat{m} + \hat{m}}} \right]_i
\]

(31)

The distribution of the test statistic under \( H_0 \) has to be obtained by simulation, as in the unweighted case. Again, the normalization constants are either given or derived from the normalization as above. If only the simulated events are weighted, we get with \( \hat{\lambda}_i = (n_i + \hat{m}_i)/(1 + 1/c_m) \) the test statistic

\[
V = \sum_{i=1}^{B} \left[ n \ln \frac{\hat{\lambda}_i - \lambda}{\hat{\lambda}_i - \frac{\hat{\lambda}_i - \lambda}{\hat{\lambda}_i - \ln n_i + \hat{m}_i \ln \frac{\hat{\lambda}_i - \lambda}{\hat{\lambda}_i - \ln \hat{m}_i + \hat{m}_i}} \right]_i
\]

(32)

As \( \hat{n} \) and \( \hat{m} \) approximately follow Poisson distributions, the power of a test based on Eq. (32) is expected to be similar to the likelihood ratio test with numbers \( n \) and \( m \).

### 4.6. Parameter estimation

When we compare experimental data to a parameter dependent Monte Carlo simulation, as before, the parameter change during the fitting is performed by re-weighting the Monte Carlo events.

**Remark.** In preceding formulas, we have replaced \( \lambda \) by its estimate. It is not correct to replace a nuisance parameter by its unconditional estimate, i.e. to fix \( \lambda \) independent of \( \hat{\theta} \). In this case the uncertainties of \( \hat{\theta} \) would be underestimated. In particle physics the error interval of a parameter is usually taken as enclosed by the tangent planes to the error ellipsoid of constant likelihood perpendicular to the axis of the parameter. (This is done for example in the program package MINUIT.) The limits correspond to those derived from the profile likelihood \( \lambda(p) = \lambda(\hat{\theta} + \hat{\theta}(\hat{\theta}) \text{data}) \), where \( \hat{\lambda}(\hat{\theta}) \) is the ML estimate of \( \lambda \) given \( \hat{\theta} \). The dependence \( \lambda(\hat{\theta}) \) is automatically included in our procedure.

We have to minimize the \( \chi^2 \) expression (29) with respect to the parameters.

We have applied Eq. (29) to a simple example: The location of a truncated normal distribution with given standard deviation \( \sigma = 0.4 \) measured with resolution \( \sigma = 0.4 \) has been estimated. We generated in the interval \( \sigma = 0.1 \) 1000 “experimental” and 1000 Monte Carlo events, centered at 0.6 and 0.5, respectively. Events accepted after smearing in the same interval were distributed into 20 histogram bins. We chose an unrealistically low number of Monte Carlo events and a large smearing to enhance the effects caused by our approximations. The experiment was repeated 10,000 times. The mean of the fitted values is 0.600 and the r.m.s. deviation of the fitted results was 0.070. The mean error of the location parameter extracted from the curvature of the \( \chi^2 \) functions \( \delta = 0.100 \) underestimated the uncertainty by 10%. If we increase the number of simulated events to 40,000, the r.m.s. deviation is reduced to 0.070. The same result is obtained with a Poisson likelihood fit using relation (14) where the statistical errors of the simulation are neglected. In addition to the location, the normalization was fitted or estimated directly\(^2\) leading to 18 degrees of freedom. In Fig. 1 we compare the \( p \)-value as a function of \( \chi^2 \) of the experimental distribution of this statistic to that of the \( \chi^2 \) distribution for 18 degrees of freedom. The agreement is reasonable, but not perfect: The observed mean value \( \bar{\chi}^2 = 17.86 \) and the width of the distribution of 5.58 differ slightly from the nominal values of 18 and \( \sqrt{\chi^2} = 4.3 \). If we increase the observed and the simulated event numbers to 10,000 the mean errors agree within 1%, the mean value \( \bar{\chi}^2 = 17.99 \) and the width of the experimental \( \chi^2 \) distribution of 6.02 are well compatible within their statistical uncertainties with the nominal values.

\(^2\) The results are the same independent of the way the histograms are normalized.
5. Parameter estimation in experiments with a large number of events

Statistical problems decrease with increasing event numbers, but computational requirements may increase. The numerical minimum search that is required to estimate the wanted parameters can become quite slow. It may happen that we have the order of $10^6$ or more simulated events. This means that, for say 10 changes of a parameter value during the extremum search, that $10^6$ weights have to be computed. This is feasible, but we may want to speed-up the fitting procedure. This can be achieved in situations where the Monte Carlo uncertainties can be neglected. We represent the prediction by a superposition of Monte Carlo histograms with factors that depend on the parameters. To this end, it is useful to expand the p.d.f. $f(x|\theta)$ in a Taylor expansion with respect to the parameter at some preliminary estimate $\theta_0$:

$$f(x|\theta) = f(x|\theta_0) + \Delta \theta \frac{df(x|\theta)}{d\theta} \bigg|_{\theta_0} + \frac{\Delta \theta^2}{2!} \frac{d^2f(x|\theta)}{d\theta^2} \bigg|_{\theta_0} + \cdots$$

(33)

$$f(x|\theta) = f(x|\theta_0) \left\{ 1 + \Delta \theta \frac{1}{f_0} \frac{df(x|\theta)}{d\theta} \bigg|_{\theta_0} + \frac{\Delta \theta^2}{2!} \frac{1}{f_0} \frac{d^2f(x|\theta)}{d\theta^2} \bigg|_{\theta_0} + \cdots \right\}.$$  

(34)

We generate events according to $f_0(x) = f(x|\theta_0)$ and obtain simulated events with the observed kinematic variable $x$. We histogram $x$ and obtain the histogram $m_0$. Weighting each event by $\omega_1(x)$, we obtain the histogram $m_1$, and weighting by $\omega_2(x)$ the histogram $m_2$, with the weights

$$\omega_1(x) = \frac{1}{f_0} \frac{df(x|\theta)}{d\theta} \bigg|_{\theta_0}$$  

(35)

$$\omega_2(x) = \frac{1}{2f_0} \frac{d^2f(x|\theta)}{d\theta^2} \bigg|_{\theta_0}.  

(36)

The parameter inference of $\Delta \theta$ is performed by comparing $m = (m_0 + \Delta \theta m_1 + (\Delta \theta)^2 m_2)$ with the experimental histogram $n_i$ as explained in Section 2:

$$\chi^2 = \sum_{i=1}^{B} \frac{(n_i - cm_i)^2}{cm_i}.  

(37)

In many cases the quadratic term can be omitted. In other situations it might be necessary to iterate the procedure.

This method works only if the number of Monte Carlo events is high enough to neglect its uncertainty with respect to that of the experimental data. To illustrate the method, we consider a lifetime measurement. We expand the p.d.f.

$$f(t|\gamma) = \gamma e^{-\gamma t}$$

into a Taylor expansion at $\gamma_0$ which is a first guess of the decay rate $\gamma$:

$$f(t|\gamma) = \gamma_0 e^{-\gamma_0 t} \left\{ 1 + \frac{\Delta \gamma}{\gamma_0} (1 - \gamma_0 t) + \left( \frac{\Delta \gamma}{\gamma_0} \right)^2 \left( -\gamma_0 t + \frac{\gamma_0^2 t^2}{2} \right) + \cdots \right\}.  

(39)

The Monte Carlo simulation follows the distribution $f_0 = \gamma_0 e^{-\gamma_0 t}$. Weighting the events by $(1/\gamma_0 - t)$ and $(-t/\gamma_0 + t^2/2)$, we obtain the distributions $f_1 = (1 - \gamma_0 t)e^{-\gamma_0 t}$, $f_2 = (-t + \gamma_0 t^2/2)e^{-\gamma_0 t}$ and $f(t|\gamma) = f_0(t) + (\Delta \gamma)^2 f_2(t) + \cdots$.  

(40)

If it is justified to neglect the higher powers of $\Delta \gamma/\gamma_0$, we can again describe our experimental distribution this time by a superposition of three distributions $f_0(t)$, $f_1(t)$, $f_2(t)$ which are the distorted versions of $f_0(t)$, $f_1(t)$, $f_2(t)$. The parameter $\Delta \gamma$ is determined by a $\chi^2$ or likelihood fit. In our special case it is even simpler to weight $f_0$ by $t$, and $t^2$, respectively, and to superpose the corresponding distributions $f_0$, $g_1 = tf_0$, $g_2 = t^2f_0$ with the factors given in the following expression:

$$f(t|\gamma) \approx f_0(t) \left( 1 + \frac{\Delta \gamma}{\gamma_0} \right) - \gamma_0 g_1(t) \left( \frac{\Delta \gamma}{\gamma_0} \right)^2 + \frac{1}{2} g_2(t) \left( \frac{\Delta \gamma}{\gamma_0} \right)^2.  

(41)
The parameter $\Delta p$ is then modified until the correspondingly weighted sum of the distorted histograms agrees optimally with the data. Fig. 2 shows an example. (The data of the example do not suffer from experimental distortions. We have set $x' = x$ to illustrate better the weighting effect.) In case the quadratic term can be neglected, two histograms are sufficient.

6. Conclusions and recommendations

We have treated the comparison of experimental histograms with Monte Carlo predictions that depend on a parameter. The parameter dependence is implemented by weighting the simulated events. We have applied Poisson statistics and have included the constraint from the normalization of the data to the simulation in the fit. We have derived several formulas that apply for different conditions. Not all of them are equally important. Let us consider the following situations:

1. We estimate parameters from a small number of experimental events. Then we can afford enough events in the simulation such that the contribution to the error from the simulation can be neglected. We should apply a likelihood fit or, if the number of events per bin is large enough, a least square fit. We apply relations (14) or (15) and if we are not sure that the Monte Carlo fluctuation can be neglected, we can check the fit with relation (29). A goodness-of-fit test can be performed comparing the $\chi^2$ of the fit (15) to the $\chi^2$ distribution. For a precise evaluation of the $p$-value, we have to simulate the event statistic many times to generate the distribution of the test statistic. With very low event numbers, we recommend to use the likelihood ratio test with the test statistic (16). Bins with few events should not be eliminated, because this biases the result. It can be avoided by generating enough Monte Carlo events and by combining bins.

2. In situations where we cannot afford to generate enough Monte Carlo events, we have to consider the uncertainties from the simulation. As the distribution of weighted events can only be described approximately by a Poisson distribution, a least square fit based on Eq. (29) is to be applied. Since the event number per bin will be relatively large, it will produce reasonable results. Also $p$-values derived from the $\chi^2$ distribution will mostly be adequate. Very low $p$-values are related to the tails of the $\chi^2$ distribution and have to be considered with care.

3. If we have a simple hypothesis, i.e. no parameters to be estimated, we can avoid weighting. Apart from the $\chi^2$ test we can also apply empirical distribution free tests like the Kolmogorov–Smirnov or the Anderson–Darling tests which often are more powerful than the $\chi^2$ test [6,1]. The simulation uncertainties can be included in these tests which, however, are restricted to one-dimensional histograms.

In any case, the validity of the applied approximations has to be checked.

References