Likelihood description for comparing data to simulation of limited statistics

Dmitry Chirkin

1Dept. of Physics and WIPAC, University of Wisconsin, Madison, WI 53706, USA
dima@icecube.wisc.edu

Abstract: It is often not possible to construct a probability density function that describes the data. This can happen if there is no analytic description, and the number of parameters is too large so that it is impossible to simulate and tabulate all combinations. In these situations it is still interesting to rank simulation sets performed with different parameters in how well they compare to data. We propose a solution that appears to be better suited to this task than some of the obvious alternatives.

Keywords: likelihood, limited simulation statistics

1 Introduction

It is often the case that the mean rate of counts in a bin is not known exactly but rather approximated with simulation. The simulation can be repeated many times, obtaining a total number of counts of \( s \) in \( n_s \) trials, and the expected rate of counts is often approximated as \( \mu = s/n_s \). For the sake of generality let’s assume that we repeat the experiment \( n_d \) times and collect a total of \( d \) counts. In order to fit for some unknown property of the experiment one often maximizes the likelihood, and for convenience that is usually done by minimizing the minus log likelihood, \(- \ln \mathcal{L}\). The minus log likelihood based on the Poisson probability of our observation is given by

\[- \ln \mathcal{L} = \ln d! + n_d \mu - d \cdot \ln(n_d \mu).\]

Unfortunately this expression is only an approximation as the quantity \( \mu \) is not known precisely but was calculated from simulation and is known within statistical uncertainties corresponding to the total number of simulated counts \( s \) in our bin.

When the counts \( s \) in simulation and \( d \) in data are large, one minimizes the \( \chi^2 \):

\[\chi^2 = \frac{(s/n_s - d/n_d)^2}{s/n_s + d/n_d},\]

where the total uncertainty in the denominator is computed as the square root of the sum of squares of the mutually independent statistical uncertainties of \( s/n_s \) and \( d/n_d \).

One may approach this problem from the Bayesian point of view: the counts in simulation are distributed with a Poisson probability around some unknown value of the true rate \( \mu \), so we convolve that probability (treating it as the likelihood) with the probability to observe the \( d \) counts in data, and with some prior (taken as \( \mu \)\(^{-1}\) in the following expression):

\[- \ln \left( \int_0^\infty \frac{(n_d \mu)^d e^{-n_d \mu}}{d!} \cdot \frac{(n_s \mu)^s e^{-n_s \mu}}{s!} \mu^{-5} d\mu \right)\]

For our example of section 1 we chose \( z = -1 \), corresponding to a non-normalizable prior of \( \mu^{-1}\).

In the next section we introduce a possible new treatment that considers statistical uncertainties in both data and simulation, and appears to perform better than the alternatives listed above, as demonstrated later in section 3.

2 Likelihood description with statistical uncertainties only

Consider a repeatable experiment that is performed \( n_d \) times to collect a total of \( d \) counts with a per-event expectation of \( \mu_d \) (we call a single instance of this experiment an “event”). We predict the result of the experiment with the simulation, which collects \( s \) counts in \( n_s \) simulated events and a per-event expectation of \( \mu_s \).

Given that the total count in the combined set of simulation and data is \( s + d \), the conditional probability distribution function of observing \( s \) simulation and \( d \) data counts is

\[P(\mu_s, \mu_d; s, d | s + d) = \frac{(s + d)!}{s! \cdot d!} \cdot \left( \frac{n_s \mu_s}{s + d} \right)^s \cdot \left( \frac{n_d \mu_d}{s + d} \right)^d.\]

An obvious constraint that is implied here is \( n_s \mu_s + n_d \mu_d = s + d \), which can be derived from the normalization requirement

\[\sum_{s,d} P(\mu_s, \mu_d; s, d | s + d) = \left( \frac{n_s \mu_s}{s + d} + \frac{n_d \mu_d}{s + d} \right)^{s+d} = 1.\]

If the data and simulation are completely unrelated the best possible estimates of \( \mu_s \) and \( \mu_d \) are determined by maximization of the probability function given above with the constraint \( n_s \mu_s + n_d \mu_d = s + d \), which results in the estimates

\[\mu_s = \frac{s}{n_s}, \quad \mu_d = \frac{d}{n_d}.\]

Now, the alternative hypothesis that we could try to verify is that data and simulation counts are described by the same process, i.e., come with the same mean \( \mu = \mu_s = \mu_d \). This identity together with the constraint \( n_s \mu_s + n_d \mu_d = s + d \) uniquely determines the values of

\[\mu = \mu_s = \mu_d = \frac{s + d}{n_s + n_d}.\]

We can now compare the probabilities of the two of the above hypotheses by forming a likelihood ratio

\[\frac{P(\text{same process})}{P(\text{independent processes})} = \left( \frac{\mu}{s/n_s} \right)^s \cdot \left( \frac{\mu}{d/n_d} \right)^d.\]
This is the expression that we propose for comparison of different simulation sets with data. It can obviously also be derived starting with the Poisson probability

\[ P(\mu_s, \mu_d; s, d) = \frac{(n_s \mu_s)^{s} e^{-n_s \mu_s}}{s!} \cdot \frac{(n_d \mu_d)^{d} e^{-n_d \mu_d}}{d!}. \]

### 3 Generalization to many bins

If there are several bins \( \{i\} \) in which simulation and data counts are compared, the conditional probability can be written as

\[ P(\{\mu_i^s\}, \{\mu_i^d\}; \{s_i\}, \{d_i\}|S + D) = \frac{(S + D)!}{\prod_i s_i! \cdot \prod_i d_i!} \prod_i \left( \frac{n_s \mu_i^s}{S + D} \right)^{s_i} \prod_i \left( \frac{n_d \mu_i^d}{S + D} \right)^{d_i}. \]

Here we use notations \( S = \sum_i s_i, D = \sum_i d_i \). The probability sum of 1 requires \( \sum_i (n_s \mu_i^s + n_d \mu_i^d) = S + D \). Taking the negative logarithm, losing constant terms, and introducing a Lagrange multiplier term for this constraint (with a new unknown \( \zeta \)), this becomes:

\[ F = -\sum_i s_i \ln(n_s \mu_i^s) - \sum_i d_i \ln(n_d \mu_i^d) + \zeta \cdot (\sum_i n_s \mu_i^s + \sum_i n_d \mu_i^d) - S - D. \]

If data and simulation are independent, this expression is minimized for each \( \mu_i^s, \mu_i^d \) independently:

\[ \frac{\partial F}{\partial \mu_i} = -\frac{s_i}{\mu_i^s} \cdot \frac{1}{\zeta} + \zeta \cdot n_s = 0, \quad \frac{\partial F}{\partial \mu_i^d} = -\frac{d_i}{\mu_i^d} + \zeta \cdot n_d = 0 \Rightarrow \mu_i^s = \frac{s_i}{\zeta \cdot n_s}, \quad \mu_i^d = \frac{d_i}{\zeta \cdot n_d}. \]

Plugging this back into the constraint equation (which we also get back by setting \( \partial F / \partial \zeta = 0 \)), we get

\[ \sum_i n_s \frac{s_i}{\zeta \cdot n_s} + n_d \frac{d_i}{\zeta \cdot n_d} = \sum_i n_s + n_d = S + D \Rightarrow \zeta = 1. \]

If data and simulation come from the same distribution we require \( \mu_i = \mu_i^s = \mu_i^d \), and minimize against \( \mu^t \):

\[ \frac{\partial F}{\partial \mu^t} = -\frac{s_i + d_i}{\mu^t} + \zeta \cdot (n_s + n_d) = 0 \Rightarrow \mu^t = \frac{s_i + d_i}{\zeta \cdot (n_s + n_d)}. \]

Once again, plugging this back into the constraint relation we get

\[ \sum_i (n_s + n_d) \frac{s_i + d_i}{\zeta \cdot (n_s + n_d)} = \sum_i (n_s + n_d) = S + D \Rightarrow \zeta = 1. \]

Thus, the expressions derived in the previous section for 1-bin situation are valid per-bin when there are more than one bin, and we get back the likelihood ratio formula

\[ \frac{P(\text{same process})}{P(\text{independent processes})} = \prod_i \left( \frac{\mu_i}{s_i/n_s} \right)^{s_i} \cdot \prod_i \left( \frac{\mu_i}{d_i/n_d} \right)^{d_i}, \quad \text{with } \mu_i = \frac{s_i + d_i}{n_s + n_d}. \]

We compare the performance of reconstruction using this formula with that using likelihood expressions listed in the introduction in section [6].

### 4 Likelihood description: adding model errors

The error in describing data with simulation (i.e., describing \( \mu_d \) with \( \mu_s \)) is often non-zero. In such a case one may quantify the amount of disagreement between data and simulation with a \( \chi^2 \):

\[ \chi^2 = \frac{(\ln \mu_d - \ln \mu_s)^2}{\sigma^2}. \]

Instead of setting \( \mu_s = \mu_d \) as in the previous sections we assume that a difference between \( \mu_s \) and \( \mu_d \) can exist due to this systematic error and is modeled with a likelihood penalty term

\[ \exp \left( -\frac{\ln^2(\mu_d/\mu_s)}{2\sigma^2} \right). \]

The likelihood ratio is therefore determined as

\[ \frac{P(\text{same process})}{P(\text{independent processes})} = \frac{(\mu_s/s/n_s)^s \cdot (\mu_d/d/n_d)^d \cdot \exp \left( \frac{-\ln^2(\mu_d/\mu_s)}{2\sigma^2} \right)}{(s + d)! \cdot \left( \frac{n_s \mu_s}{S + D} \right)^{s} \cdot \left( \frac{n_d \mu_d}{S + D} \right)^{d}.} \]

with the constraint \( n_s \mu_s + n_d \mu_d = s + d \). Taking the negative logarithm, losing constant terms, and introducing a Lagrange multiplier term for this constraint (with a new unknown \( \zeta \)), this becomes:

\[ -s \ln(\mu_s/n_s) - d \ln(\mu_d/n_d) + \frac{1}{2\sigma^2} \ln^2 \mu_d/\mu_s + \zeta \cdot (n_s \mu_s + n_d \mu_d - s - d) \equiv F. \]

The function \( F(\mu_s, \mu_d) \) can be easily minimized against \( \mu_s \) and \( \mu_d \), yielding estimates of these quantities. To demonstrate this, first the derivatives of \( F \) are calculated and set to 0:

\[ \mu_s \frac{\partial F}{\partial \mu_s} = \zeta \mu_s n_s - s - \frac{1}{\sigma^2} \ln \mu_d/\mu_s = 0, \]

\[ \mu_d \frac{\partial F}{\partial \mu_d} = \zeta \mu_d n_d - d - \frac{1}{\sigma^2} \ln \mu_s/n_s = 0. \]

The sum of these, \( \zeta \cdot (n_s \mu_s + n_d \mu_d) = s + d \), results in the value for \( \zeta = 1 \). The derivative of \( F \) with respect to \( \zeta \) gives back the constraint \( n_s \mu_s + n_d \mu_d = s + d \), which yields an expression of \( \mu_d \) as a function of \( \mu_s \). Plugging it into the first of the above two equations one gets

\[ f = \mu_s \frac{\partial F}{\partial \mu_s}(\mu_s, \mu_d(\mu_s)) = n_s n_s - s - \frac{1}{\sigma^2} \ln \mu_d/\mu_s = 0. \]

This equation can be solved with a few iterations of the Newton’s root finding method starting with a solution to

\[ \mu_s = \mu_d(\mu_s); \quad \mu_s = \mu_d = \frac{s + d}{n_s + n_d}. \]
At each iteration the value of \( \mu_s \) is adjusted by \(-f/f'\), where the derivative is evaluated as

\[
f' = n_s \left( 1 + \frac{1}{\sigma^2} \left( \frac{1}{\mu_s n_s} + \frac{1}{\mu_d n_d} \right) \right).
\]

Once the likelihood function is solved for the best values of \( \mu_s \) and \( \mu_d \), these can be plugged into the likelihood ratio given above. One can now write the likelihood ratio as a sum over all bins:

\[
\sum_k \left[ s_k \ln \frac{s_k/n_s}{\mu_s^k} + d_k \ln \frac{d_k/n_d}{\mu_d^k} + \frac{1}{2\sigma^2} \ln \frac{s_k}{\mu_s^k} + \frac{1}{2\sigma^2} \ln \frac{d_k}{\mu_d^k} \right].
\]

This is an improved expression compared to the one used in [1]. The probability \( P(\text{same process}) \) can also be thought of as a convolution of the binomial probability part of the expression with the penalty term. Solving the convolution integral approximately with the Laplace’s method results (up to a constant term) in an expression for \( P(\text{same process}) \) given above.

5 Likelihood description of data with weighted simulation

One can apply the method for calculating the likelihood ratio of the previous section to a situation that is common when the number of data counts \( d_k \) in bin \( k \) measured during time \( t_d \) is fitted with a number of simulation counts \( s_k \), each representing a possible (different for simulation) time \( t_k \) (usually related to the event weight \( w_k \)). Although we can assume that all \( s_k \) is 1 without the loss of generality, we continue with the notation \( s_k \).

The combined number of events in data and simulation is then \( S + D \), where \( S = \sum_k s_k, s_k = \sum_k s_k, D = \sum_k d_k \). The expression for the conditional probability is now

\[
P(\{\mu^{(k)}_s\}, \{\mu^{(k)}_d\}; \{s_k\}, \{d_k\}; S + D) = \frac{(S + D)!}{\prod_k s_k! \prod_k d_k!} \prod_k \left( \frac{t_k \mu^{(k)}_s}{S + D} \right)^{s_k} \prod_k \left( \frac{t_d \mu^{(k)}_d}{S + D} \right)^{d_k}.
\]

The probability sum of 1 necessitates the constraint

\[
\sum_k t_k \mu^{(k)}_s + \sum_k t_d \mu^{(k)}_d = S + D.
\]

Taking the negative logarithm of \( P \), losing constant terms, and introducing a Lagrange multiplier term for this constraint (with a new unknown \( \zeta \)), we get:

\[
F = -\sum_k s_k \ln (t_k \mu^{(k)}_s) - \sum_k d_k \ln (t_d \mu^{(k)}_d) + \\
\zeta \cdot \left( \sum_k t_k \mu^{(k)}_s + \sum_k t_d \mu^{(k)}_d - S - D \right)
\]

If data and simulation are independent, this expression can be minimized for each \( \mu^{(k)}_s \) independently:

\[
\frac{\partial F}{\partial \mu^{(k)}_s} = -\frac{s_k}{\mu^{(k)}_s} + \zeta \cdot t_k = 0, \quad \frac{\partial F}{\partial \mu^{(k)}_d} = -\frac{d_k}{\mu^{(k)}_d} + \zeta \cdot t_d = 0 \quad \Rightarrow \quad \mu^{(k)}_s = \frac{s_k}{\zeta \cdot t_k}, \quad \mu^{(k)}_d = \frac{d_k}{\zeta \cdot t_d}.
\]

Plugging this back into the constraint equation (which we also get back by setting \( \partial F/\partial \zeta = 0 \)), we get

\[
\sum_k t_k \frac{s_k}{\zeta \cdot t_k} + \sum_k t_d \frac{d_k}{\zeta \cdot t_d} = \sum_k \frac{s_k}{\zeta} + \frac{d_k}{\zeta} = S + D \Rightarrow \zeta = 1.
\]

If data and simulation come from the same distribution we require \( \mu^{(k)}_s = \sum \mu^{(k)}_s \) for each \( k \). These conditions can be introduced into the above expression for \( F \) as additional terms (with new unknowns \( \zeta_k \)):

\[
F = -\sum_k s_k \ln (t_k \mu^{(k)}_s) - \sum_k d_k \ln (t_d \mu^{(k)}_d) + \\
\zeta_k \cdot \left( \sum_k t_k \mu^{(k)}_s + \sum_k t_d \mu^{(k)}_d - S - D \right) + \sum_k \zeta_k \cdot \left( \sum_i \mu^{(k)}_s - \mu^{(k)}_d \right).
\]

Derivatives with respect to \( \zeta \) and \( \zeta_k \) give back the constraint equations. The other derivatives are:

\[
\frac{\partial F}{\partial \mu^{(k)}_s} = -\frac{s_k}{\mu^{(k)}_s} + \zeta_k \cdot t_k + \xi_k = 0, \\
\frac{\partial F}{\partial \mu^{(k)}_d} = -\frac{d_k}{\mu^{(k)}_d} + \zeta_k \cdot t_d - \xi_k = 0.
\]

Multiplying the first equation by \( t_k \mu^{(k)}_s \), the second by \( d_k \), and summing them together, we get

\[
0 = -\sum_k s_k + \sum_k t_d \mu^{(k)}_d + \zeta_k \cdot \left( \sum_k t_k \mu^{(k)}_s + \sum_k t_d \mu^{(k)}_d \right) + \\
\sum_k \xi_k \cdot \left( \sum_i \mu^{(k)}_s - \mu^{(k)}_d \right) = -S - D + \zeta_k \cdot (S + D) + \sum_k \xi_k \cdot 0.
\]

Therefore \( \zeta = 1 \). To find \( \xi \) we substitute the expressions for \( \mu^{(k)}_s \) and \( \mu^{(k)}_d \) into constraints for \( \xi_k \):

\[
\mu^{(k)}_s = \frac{s_k}{t_k + \xi_k}, \quad \mu^{(k)}_d = \frac{d_k}{t_d - \xi_k} \Rightarrow \sum_k \frac{s_k}{t_k + \xi_k} = \frac{d_k}{t_d - \xi_k}.
\]

Therefore, the likelihood ratio is

\[
P(\text{same process}) \quad \frac{P(\text{independent processes})}{\prod_k \left( \frac{t_k}{t_k + \xi_k} \right)^{s_k} \prod_k \left( \frac{t_d}{t_d - \xi_k} \right)^{d_k}}.
\]

The equation for \( \xi_k \) is similar to equation 15 of [1]. As suggested there, we solve them for each \( k \) by Newton’s method starting with \( \xi_k = 0 \), ensuring that \( -\sum_i \{s_k > 0\} \cdot t_k < \xi_k < t_d \):

\[
f_k = 1/\left[ \sum_i \frac{s_k}{t_k + \xi_k} \cdot \frac{t_d - \xi_k}{d_k} \right],
\]

\[
\frac{d f_k}{d \xi_k} = \sum_i \frac{s_k}{(t_k + \xi_k)^2} \left[ \sum_i \frac{s_k}{t_k + \xi_k} \right]^2 + \frac{1}{d_k} \Rightarrow \xi_k(\text{next}) = \xi_k - \frac{f_k}{d f_k/d \xi_k}.
\]
The particular form of function $f_k$ above (inverted compared to the original equation for $\xi_k$) was chosen to linearize the problem in simple cases (e.g., all simulated events having the same weight). After the first iteration (starting with $\xi_k = 0$) we get

$$\xi_k \approx \frac{t_d/d_k - 1}{1/d_k + \sum s_k / t_k^2} = \frac{t_d}{1 + d_k / \sigma_k^2} \approx \frac{t_d}{1 - d_k/m_k},$$

with $m_k = \sum s_k w_{ki}$, $\sigma_k^2 = \sum s_k w_{ki}^2$, and $w_{ki} = t_d/d_k$.

Here $m_k$ is the total simulation in bin $k$ evaluated for time $t_d$ (i.e., is a prediction of data in bin $k$); $\xi_k$ is a statistical uncertainty of the value $m_k$. If $m_k$ and $d_k$ are not too far from each other, and the statistical uncertainty of the simulation is much smaller than that of data, i.e., $\xi_k \ll \sqrt{d_k}$, then $\xi_k \approx t_d(1 - d_k/m_k)$. This holds in the limit of infinite simulation statistics, when all $t_{k_i} \to \infty$, and thus, $w_{ki} \to 0$. Continuing with this approximation,

$$P(\text{same process}) = \frac{\prod_{k_i} \left[ \frac{1}{1 + w_{ki} \cdot (1 - d_k/m_k)} \right]^{t_{k_i}} \prod_{k} \left[ \frac{1}{1 - (1 - d_k/m_k)} \right]^{d_k}}{\prod_{k} \left[ \exp(d_k - m_k) \cdot \left( \frac{m_k}{d_k} \right)^{d_k} \right]} \approx \exp \left[ -\sum s_{ki} w_{ki} \cdot (1 - d_k/m_k) \right] \prod_{k} \left( \frac{m_k}{d_k} \right)^{d_k}$$

Up to a constant term this is a product of the usual expressions for the Poisson likelihood,

$$p(m;d) = \frac{m^d}{d!} \exp(-m).$$

However, the expression above is also precisely identical to the product of ratios $p(m;d)/p(d;d)$, where $m = d$ in the denominator is the value at which the Poisson likelihood $p(m;d)$ achieves its maximum.

We would like to emphasize that using the usual Poisson expression for the likelihood fit of data to simulation is, strictly speaking, only correct in the limit of counts of the simulation being infinite in all of the bins used in the likelihood fit, and all of the simulation event weights being infinitesimally small. An uneven distribution of simulation counts in bins (e.g., energy bins) may create a bias in the resulting fit (i.e., bias in energy). In this section we advocate using the exact expression for the likelihood ratio given above. We are aware of several recent works (e.g., [4]) trying the new expression and finding that it works better than the usual Poisson likelihood.

7 Conclusion

In this paper we consider a problem of describing data with limited-statistics simulation sets. After a brief review of the usual methods we presented an alternative approach to the likelihood-based comparison of data with multiple simulation sets. In our tests the new approach appears to improve the results of model parameter fits, mainly reducing the bias and uncertainty of the fitted parameters. We remark that depending on the statistics of the data and available simulation, and on the nature of the experiment other approaches may occasionally produce a better result, so we encourage testing of several likelihood expressions as we did in section 6.

The method of this paper as described in sections 2-4 was developed as part of calibration work reported in [3]. The extension to weighted simulation described in section 5 was proposed for use in several analyses of IceCube, and was shown to yield slightly better results (e.g., in [6]: better limits, due to narrower distributions of test statistic).

Acknowledgment

This work was supported by the U.S. National Science Foundation-Office of Polar Programs and the U.S. National Science Foundation-Physics Division.

References