A binned likelihood for stochastic models

C.A. Argüelles, A. Schneider, T. Yuan

Dept. of Physics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
Dept. of Physics and Wisconsin IceCube Particle Astrophysics Center, University of Wisconsin, Madison, WI 53706, USA
E-mail: caad@mit.edu, aschneider@icecube.wisc.edu, tyuan@icecube.wisc.edu

ABSTRACT: Metrics of model goodness-of-fit, model comparison, and model parameter estimation are the main categories of statistical problems in science. Bayesian and frequentist methods that address these questions often rely on a likelihood function, which describes the plausibility of model parameters given observed data. In some complex systems or experimental setups predicting the outcome of a model cannot be done analytically and Monte Carlo techniques are used. In this paper, we present a new analytic likelihood that takes into account Monte Carlo uncertainties, appropriate for use in large or small statistics regimes. Our formulation has better performance than semi-analytic methods, prevents strong claims on biased statements, and results in better coverage properties than available methods.

KEYWORDS: Likelihood, Monte Carlo, Poisson distribution

ORCID: 0000-0003-4186-4182
ORCID: 0000-0002-0895-3477
ORCID: 0000-0002-7041-5872
1 Introduction

The use of Monte Carlo (MC) techniques to calculate non-trivial theoretical quantities and expectations in complex experimental settings is common practice in particle physics. A MC event is a single representation of what can be detected in data, and is typically generated from a single realization of the underlying physics parameters, $\vec{\theta}$. These events are often binned in some observable space and compared with the data. Since the generation process is stochastic, a particular $\vec{\theta}$ used for generating the MC can lead to different outputs. This stochasticity introduces an uncertainty in the MC distributions. Furthermore, as production of large-statistics MC is often time consuming, reweighting is used to move from one hypothesis to another. In reweighting, a MC event with previously assigned weight, $w(\vec{\theta})$, is assigned a new weight, $w(\vec{\theta}')$ [1]. It follows that MC uncertainties will be hypothesis dependent; thus, to do hypothesis testing, it is important to account for them. This is especially important for small-signal searches, performed far away from the Gaussian limit, where a modified-$\chi^2$ may not be suitable [2]. A Poisson likelihood is a more appropriate statistical description of event counts, but there a proper treatment of MC statistical uncertainties is less straightforward. Solutions to this problem have been discussed in the literature in the context of frequentist statistics by adding nuisance parameters [3–5]. A detailed probabilistic treatment of MC weights was discussed in [6]. However, [3, 5, 6] add additional time complexity, and [4] does not provide a full exposition on how to incorporate weighted MC. We present a new treatment that is valid in the large and small statistics limits of data and MC, suited for frequentist and Bayesian analyses,
based on the Poisson likelihood. Our likelihood accounts for statistical uncertainties due to MC, converges appropriately in the high-statistics limit, allows for arbitrary event-by-event reweighting, has better coverage properties, and is computationally efficient. An implementation of the likelihood described in this work can be found in [7].

The paper is organized as follows. In Sec. 2 we briefly review two common treatments available in the literature to account for MC statistical uncertainty. In Sec. 3 we define and discuss our new likelihood. In Sec. 4 we study the performance of the likelihood in an example and compare it to other treatments in the literature. Finally, in Sec. 5 we conclude.

2 The Poisson likelihood and previous work

In order to compare MC with data, events are often binned into distributions across a set of observables. For simplicity we focus on a single-bin; in the absence of cross-bin-correlated systematic uncertainties the generalization to multiple bins is trivial. It is well known that the count of independent, rare natural processes can be described by the Poisson likelihood [8], given by

$$L(\vec{\theta}|k) = \text{Pois}(k; \lambda(\vec{\theta})) = \frac{\lambda(\vec{\theta})^k e^{-\lambda(\vec{\theta})}}{k!},$$

(2.1)

where $\lambda(\vec{\theta})$ is the expected bin count for a hypothesis, which depends on parameters $\vec{\theta}$, and $k$ is the number of observed data events. For weighted MC, often a direct substitution of $\lambda(\vec{\theta}) = \sum_i w_i(\vec{\theta})$ is used, where $w_i$ are the weights of each of the MC events. Then Eq. (2.1) can be rewritten as

$$L_{\text{Poisson}}(\vec{\theta}|k) = \left(\sum_i w_i(\vec{\theta})\right)^k e^{-(\sum_i w_i(\vec{\theta}))} \frac{1}{k!},$$

(2.2)

where the sum over $i$ is performed across all MC events in the bin.

To treat MC statistical uncertainties in the small sample regime a modification of the Poisson likelihood was introduced in [3], which is briefly covered below. First, note that the expectation in a single bin is given by contributions from different physical processes, which we index by $s$. Then, the number of expected events can be written as

$$\lambda(\vec{\theta}) = \sum_s \bar{N}_s(\vec{\theta}),$$

(2.3)

where $\bar{N}_s$ is the expected number of MC events from process $s$ that fall in the bin and the sum runs $s = 1, \ldots, n_s$ where $n_s$ is the total number of relevant processes. This quantity can be compared with the data by using the likelihood in Eq. (2.1). The MC generation process can be modeled as having drawn $N_s$ events out of a random process that simulates the physical process. Thus $N_s$ is an estimator of the expected number of events from source $s$, $\bar{N}_s$. The MC generation is a binomial process where we generate an arbitrary and fixed number of events per source of physical process, $N_{s,\text{gen}}$, and accept them into the bin of interest with probability $\beta_s(\vec{\theta})$, such that $\bar{N}_s(\vec{\theta}) = \beta_s(\vec{\theta}) N_{s,\text{gen}}$. In the
regime of rare processes, \( \beta_s(\vec{\theta}) \ll 1 \), and large number of generated events, \( N_{s,\text{gen}} \gg 1 \), the expected counts in the bin can be approximated as Poisson distributed with parameter
\[
\lambda(\vec{\theta}) = \sum_s \beta_s(\vec{\theta}) N_{s,\text{gen}} = \sum_s \bar{N}_s(\vec{\theta}).
\]
This is the case of interest, as it is when MC generation is expensive, so we can state that \( N_s \) is drawn from a Poisson process with mean \( \bar{N}_s(\vec{\theta}) \).

Profiling on the true number of MC events per process in the bin results in the Barlow-Beeston (BB) likelihood, given by [3]
\[
L_{\text{BB}}(\vec{\theta}|k) = \max_{\{\bar{N}_s\}} \frac{\lambda(\vec{\theta})^k e^{-\lambda(\vec{\theta})}}{k!} \prod_s \frac{\bar{N}_s^{N_s} e^{-\bar{N}_s}}{N_s!},
\]
where \( \lambda(\vec{\theta}) \) is given by Eq. (2.3), \( N_s \) and \( \bar{N}_s \) are the estimated and true MC counts in the bin respectively, and we have profiled over \( n_s \) nuisance parameters denoted by \( \{\bar{N}_s\}_{s=1}^{n_s} \).

In the above formalism we have produced the MC at the natural rate, but this is not the case in weighted MC. The prescription is given by replacing Eq. (2.3) with
\[
\lambda(\vec{\theta}) = \sum_s w_s(\vec{\theta}) \bar{N}_s,
\]
where \( w_s(\vec{\theta}) \) are the average weights of MC events of source \( s \) that account for the differences in the MC generation and target hypothesis of interest. In this case, the likelihood definition is still given by Eq. (2.4).

In the large-statistics regime the Gaussian distribution is an appropriate description of the observed data. In this limit, the use of the Pearson-\( \chi^2 \) as a test-statistic [9] is common practice. For a single analysis bin the Pearson-\( \chi^2 \) is defined as
\[
\chi^2(\vec{\theta}) = \frac{(k - \lambda(\vec{\theta}))^2}{\lambda(\vec{\theta})},
\]
where \( \lambda(\vec{\theta}) = \sum_i w_i(\vec{\theta}) \) and \( w_i \) are the weights of each of the MC events. The form of the Pearson-\( \chi^2 \) arises from the fact that the Gaussian distribution of \( k \) is the large sample limit of a Poisson distribution for which the expected statistical variance of the observation is given by \( \lambda(\vec{\theta}) \). Systematic uncertainties, under the assumption that they follow a Gaussian distribution and are independent between bins, can be included as,
\[
\chi^2_{\text{mod}}(\vec{\theta}) = \frac{(k - \lambda(\vec{\theta}))^2}{\lambda(\vec{\theta})} + \sigma^2_{\text{syst}} + \sigma^2_{\text{mc}},
\]
though this method of incorporating systematic uncertainties tends to overestimate them in shape-only analyses; see [10] for a recent discussion in the context of reactor neutrino anomalies. Similarly, one can include uncertainties to account statistical fluctuations of the MC in the test statistic. In doing so the Gaussian behaviour is implicit and the modified-\( \chi^2 \) reads
\[
\chi^2_{\text{mod}}(\vec{\theta}) = \frac{(k - \lambda(\vec{\theta}))^2}{\lambda(\vec{\theta}) + \sigma^2_{\text{syst}} + \sigma^2_{\text{mc}}}.
\]
where $\sigma_{\text{mc}}^2$ is the MC statistical uncertainty in the bin given by

$$
\sigma_{\text{mc}}^2(\vec{\theta}) \equiv \sum_{i=1}^{m} w_i(\vec{\theta})^2.
$$

(2.9)

Note that this test statistic definition is not appropriate in the small statistics regime, as the data is no longer well described by a Gaussian distribution. If one uses a $\chi^2$ test-statistic in the small sample regime, one ought to calculate the test-statistic distribution properly to achieve appropriate coverage [11].

3 Generalization of the Poisson likelihood

Using MC we can compute the expected event count, $\lambda(\vec{\theta})$, by means of adding the MC event weights as discussed in Sec. 2. Due to the finite nature of MC we can never recover exactly the intrinsic parameter, $\lambda(\vec{\theta})$, from which the MC events are drawn, but only construct estimators; the same is true for any observable of interest. As such, $\mathcal{L}_{\text{Poisson}}$ written in Eq. (2.2) is an approximation that presumes the MC realization matches the expectation. In general, we should rewrite $\mathcal{L}_{\text{Poisson}}$ as

$$
\mathcal{L}_{\text{General}}(\vec{\theta}|k) = \int_{0}^{\infty} \frac{\lambda^k e^{-\lambda}}{k!} \mathcal{P}(\lambda|\vec{w}(\vec{\theta})) \, d\lambda,
$$

(3.1)

where the distribution of $\lambda$, $\mathcal{P}(\lambda|\vec{w}(\vec{\theta}))$, is inferred from the MC. The likelihood, $\mathcal{L}_{\text{Poisson}}$, in Eq. (2.2) is recovered when $\mathcal{P}(\lambda|\vec{w}(\vec{\theta})) = \delta(\lambda - \sum_i w_i(\vec{\theta}))$, but clearly this is an unrealistic assumption as it presumes perfect knowledge of the parameter $\lambda(\vec{\theta})$ from a finite number of realizations. Instead, it is more appropriate to construct $\mathcal{P}(\lambda|\vec{w}(\vec{\theta}))$ based on the MC realization. This is given by

$$
\mathcal{P}(\lambda|\vec{w}(\vec{\theta})) = \frac{\mathcal{L}(\lambda|\vec{w}(\vec{\theta})) \mathcal{P}(\lambda)}{\int_{0}^{\infty} \mathcal{L}(\lambda'|\vec{w}(\vec{\theta})) \mathcal{P}(\lambda') \, d\lambda'},
$$

(3.2)

where $\mathcal{P}(\lambda)$ is a prior on $\lambda$ that must be chosen appropriately and $\mathcal{L}(\lambda|\vec{w}(\vec{\theta}))$ is the likelihood of $\lambda$ given $\vec{w}(\vec{\theta})$. This is similar to [3, 4], but instead of fitting $\lambda$ as a nuisance parameter as in $\mathcal{L}_{\text{BB}}$ Eq. (2.4), we marginalize over it as informed by the MC weights.

3.1 Likelihood construction

In this section we first construct $\mathcal{L}(\lambda|\vec{w}(\vec{\theta}))$ and then motivate an appropriate choice of $\mathcal{P}(\lambda)$. We will show that the likelihood can be written in terms of two quantities,

$$
\mu \equiv \sum_{i=1}^{m} w_i \quad \text{and} \quad \sigma^2 \equiv \sum_{i=1}^{m} w_i^2,
$$

(3.3)

for a bin with $m$ MC events.
Assume that $m$ is the outcome of sampling a Poisson distributed random variable $M$ with probability mass function

$$\text{Pois}(M = m; \bar{m}) = \frac{e^{-\bar{m}\bar{m}^m}}{m!},$$

where $\bar{m}$ is the mean of the distribution. We will motivate a probability distribution of the expected number of data events, $\lambda$, by studying the case of identical weights. For identical weights, $w \equiv w_i \ \forall i$, the following equalities hold:

$$\mu = w\bar{m}, \sigma^2 = w^2\bar{m}, \ \text{and} \ \bar{m} = \lambda/w.$$  (3.5)

We can interpret $\text{Pois}(M = m; \bar{m})$ as a likelihood function of $\lambda$

$$L(\lambda|\bar{w}(\bar{\theta})) = L(\lambda|\mu, \sigma) = \frac{e^{-\lambda\mu/\sigma^2} (\lambda\mu/\sigma^2)^{\mu^2/\sigma^2}}{(\mu^2/\sigma^2)!},$$

as $\mu$ and $\sigma$ fully specify $\bar{w}(\bar{\theta})$ for identical weights. Assuming a uniform $\mathcal{P}(\lambda)$ and substituting Eq. (3.6) for $L(\lambda|\bar{w}(\bar{\theta}))$ in Eq. (3.2), we obtain

$$\mathcal{P}(\lambda|\bar{w}(\bar{\theta})) = \beta \frac{(\beta\lambda)^{\alpha-1}e^{-\beta\lambda}}{\Gamma(\alpha)} = \frac{\lambda^{\alpha-1}\beta^\alpha e^{-\beta\lambda}}{\Gamma(\alpha)} = \mathcal{G}(\lambda; \alpha, \beta),$$

where $\Gamma$ is the gamma function and $\mathcal{G}$ the gamma distribution with shape parameter $\alpha$ and inverse-scale parameter $\beta$. These parameters are given in terms of $\mu$ and $\sigma$ as

$$\alpha = \frac{\mu^2}{\sigma^2} + 1 \text{ and } \beta = \frac{\mu}{\sigma^2}.$$  (3.8)

With this choice of $L(\lambda|\bar{w}(\bar{\theta}))$ and $\mathcal{P}(\lambda)$, we can rewrite $L_{\text{General}}$ from Eq. (3.1) as

$$L_{\text{Eff}}(\bar{\theta}|k) = \int_0^\infty \frac{\lambda^k e^{-\lambda}}{k!} \mathcal{G}(\lambda; \alpha, \beta) \ d\lambda = \frac{\beta^\alpha \Gamma(k + \alpha)}{k! (1 + \beta)^{k+\alpha} \Gamma(\alpha)},$$

where $\alpha$ and $\beta$ depend on $\bar{\theta}$ through $\bar{w}$.

The derivation above assumed identical weights. For arbitrary weights, $\mu$ is an outcome sampled from a compound Poisson distribution, which can be approximated by a scaled Poisson distribution by matching their first and second moments as outlined in [12]. Equation (3.3) can be written

$$\mu = \mu_{\text{Eff}} m_{\text{Eff}} \text{ and } \sigma^2 = \sigma_{\text{Eff}}^2 m_{\text{Eff}},$$

where $m_{\text{Eff}}$ is an effective number MC events and $w_{\text{Eff}}$ an effective weight. Both are given in terms of $\mu$ and $\sigma$ by Eq. (3.5) as $m_{\text{Eff}} = \mu^2/\sigma^2$ and $w_{\text{Eff}} = \sigma^2/\mu$. They can be directly substituted into the derivation for identical weights, where $\mu$ is described by $\text{Pois}(M =$ – 5 –
\( m_{\text{Eff}}; \tilde{m} \) = Pois\((\mu/w_{\text{Eff}}; \lambda/w_{\text{Eff}})\), corresponding to the scaled Poisson distribution described in [12]. Since \( m_{\text{Eff}} \) need not be an integer, we replace the factorial in Eq. (3.6) with a gamma function, as was done in Eq. (3.7). The derivation of \( L_{\text{Eff}} \) remains the same and is a valid approximation for arbitrary weights.

It is possible to generalize the choice of \( \alpha \) and \( \beta \) in Eq. (3.8) by choosing a particular form of \( P(\lambda) \). Since the distribution of interest is Poisson, a well-motivated choice of \( P(\lambda) \) is its conjugate prior the gamma distribution [13]; also see [6] for a recent discussion. Thus we set \( P(\lambda) = G(\lambda; a, b) \), where \( a \) and \( b \) are the shape and inverse-scale parameters of the gamma distribution respectively. These hyper-parameters dictate the distribution of the Poisson parameter \( \lambda \) [14]. In line with our previous discussion, the gamma distribution prior implies that Eq. (3.8) becomes

\[
\alpha = \frac{\mu^2}{\sigma^2} + a \quad \text{and} \quad \beta = \frac{\mu}{\sigma^2} + b.
\]  

The rest of the likelihood derivation remains the same. This allows the choice of specific values for \( a \) and \( b \) to satisfy certain properties. Equation (3.8) is obtained with \( a = 1 \) and \( b = 0 \), corresponding to the uniform prior discussed above. Another interesting choice is to require that the mean and variance of \( P(\lambda|\mu, \sigma) \) matches \( \mu \) and \( \sigma^2 \) respectively. This can be achieved by setting \( a = b = 0 \) and we refer to this parameter assignment as \( L_{\text{Mean}} \).

In the case of identical weights, \( L_{\text{Mean}} \) is equivalent to Eq. (20) in [6]. Both choices are improper priors, as technically they are limiting cases of the gamma distribution. However, we can use them to obtain proper \( P(\lambda|\mu, \sigma) \) distributions.

In [6] a convolutional approach is suggested for handling arbitrary weights. Each weighted MC event has \( P(\lambda_i|w_i) = G(\lambda_i; 1, 1/w_i) \), corresponding to the prior \( P(\lambda_i) = G(\lambda_i; 0, 0) \), such that \( \lambda = \sum_i^m \lambda_i \). The likelihood \( L_{\text{Mean}} \) is a good analytic approximation of the more computationally expensive calculation given in [6]. The latter has time complexity \( O(k^2m) \) where \( k \) and \( m \) are the number of data and MC events in the bin respectively. When assuming uniform priors, the convolutional approach does not recover Eq. (3.7) for identical weights, so it cannot be used as a generalization of \( L_{\text{Eff}} \).

3.2 Likelihood convergence

In this section we will show that, if the relative uncertainty of the bin content vanishes as MC statistics increases, \( L_{\text{Eff}} \) and \( L_{\text{Mean}} \) both converge to \( L_{\text{Poisson}} \).

For positive weights \( w_i \), the relative uncertainty \( \sigma/\mu \) is bounded between zero and one. Uncertainty as large as the estimated quantity, \( \sigma/\mu = 1 \), occurs if and only if \( m = 1 \). In the limit that \( \sigma/\mu \) goes to zero, Eq. (3.7) converges to \( \delta(\lambda - \mu) \) and \( L_{\text{Eff}} \) and \( L_{\text{Mean}} \) both go to \( L_{\text{Poisson}} \). We can see this by noting that the shape parameter, \( \alpha \), goes to infinity as the MC relative uncertainty goes to zero, turning the gamma distribution to a Gaussian distribution of mean \( \alpha/\beta \) and variance \( \alpha/\beta^2 \). This Gaussian converges to \( \delta(\lambda - \mu) \) in the limit of vanishing \( \sigma/\mu \). Substituting into Eq. (3.1), we recover \( L_{\text{Poisson}} \).

It remains to be shown that the relative uncertainty of the bin content vanishes as MC statistics increase. For identical weights,

\[
\lim_{m \to \infty} \frac{\sigma_{\text{identical}}}{\mu_{\text{identical}}} = \lim_{m \to \infty} \frac{1}{\sqrt{m}} = 0.
\]  

\[ -6 - \]
For arbitrary weights, the limit can be written in terms of the running average of $w_i$ and $w_i^2$ as

$$
\lim_{m \to \infty} \frac{\sigma}{\mu} = \lim_{m \to \infty} \frac{\sqrt{\langle w^2 \rangle_m}}{\langle w \rangle_m \sqrt{m}},
$$

(3.13)

where $\langle w \rangle_m$ is the average over $w_i$ and $\langle w^2 \rangle_m$ the average over $w_i^2$ for $i \leq m$. This shows that as long as $\langle w^2 \rangle_m$ does not grow much faster than $\langle w \rangle_m^2$, the limit will converge to zero. For most practical weight distributions, this should be the case.

3.3 Likelihood behavior

![Figure 1. Likelihood behavior. $l_{\text{Eff}}(k = 100)$ minimized over $\mu$ for fixed values of $\sigma$ (left) and fixed values of $\sigma/\mu$ (right). The shaded regions indicate where $l(\mu, \sigma | k) - l(\hat{\mu}, \sigma | k) < 1$. For the fixed $\sigma$ plot, the shaded regions are calculated without requiring that $\mu \geq \sigma$, which would be the case for Eq. (3.3). As $\sigma$ goes to zero, the Poisson best-fit $\hat{\mu}_{\text{Poisson}} = 100$ is obtained. For fixed $\sigma/\mu$, $\hat{\mu}_{\text{Eff}}$ deviates from $\hat{\mu}_{\text{Poisson}}$ as $\sigma/\mu$ increases.

It is instructive to examine the behavior of $\mathcal{L}_{\text{Eff}}$ for the simple case of a single-bin fit with $k = 100$. As the general behavior of $\mu$ and $\sigma$ can be complicated for arbitrary weights, we minimize $l_{\text{Eff}} \equiv -2 \ln \mathcal{L}_{\text{Eff}}$ over $\mu$ for two possibilities: fixed $\sigma$ and fixed $\sigma/\mu$. In terms of Eq. (3.3), a sufficient but not necessary condition for constant $\sigma/\mu$ with varying $\mu$ is equal weights, and a necessary but not sufficient condition for constant $\sigma$ with varying $\mu$ is $m \geq 2$. Furthermore, a constant $\sigma$ implies that $\mu \leq \sigma \sqrt{m}$. For arbitrary weights the correlation between $\mu$ and $\sigma$ should lie somewhere in between the two cases.

For a pure Poisson likelihood, $\hat{\mu}_{\text{Poisson}} \equiv \min_{\mu} l_{\text{Poisson}}(\mu | k = 100)$ as well as the region where $l(\mu, \sigma | k) - l(\hat{\mu}, \sigma | k) < 1$ for fixed $\sigma$ (left) and fixed $\sigma/\mu$ (right). As $\sigma$ goes to zero the Poisson best-fit and $1\sigma$ interval is recovered. As $\sigma$ or $\sigma/\mu$ increases, the shaded region becomes wider as expected. For fixed $\sigma$, $\hat{\mu}_{\text{Eff}}$
does not deviate much from $\hat{\mu}_{\text{Poisson}}$, while for fixed $\sigma/\mu$, $\hat{\mu}_{\text{Eff}}$ deviates from $\hat{\mu}_{\text{Poisson}}$ as $\sigma/\mu$ increases. However, since $0 < \sigma/\mu \leq 1$, $l_{\text{Eff}}$ covers $\hat{\mu}_{\text{Poisson}}$ for all possible values of $\sigma/\mu$.

The shaded regions correspond to the 1σ interval assuming Wilks’ theorem and gives a sense of the shape of $L_{\text{Eff}}$ projected onto 1D slices.

4 Example and performance

In practice, likelihood constructions such as those discussed above are used to estimate physical parameters from data. As discussed in Sec. 1, weighted MC is often used to compute the likelihood of a particular physical scenario given the observed data. Statements are then made about the physical scenarios either by maximizing the likelihood, or examining the posterior distribution assuming some priors. We examine a toy experiment where we measure the mode, $\Omega$, and normalization, $\Phi$, of a signal gaussian against a steeply falling inverse power-law background. The performance of $L_{\text{Eff}}$ is evaluated and compared against other likelihoods.

For our toy experiment, we generate the true energies, $E_t$, of fake data events from a background falling as $(E_t/100\text{GeV})^{-3.07}$ and a Gaussian signal centered at $\Omega = 125 \text{ GeV}$ with width 2 GeV and normalization $\Phi = 5013$ for 1 arb. unit of livetime. Our imaginary detector is sensitive in the range from 100–160 GeV. To simulate the effect of a real detector, 5% (background) and 3% (signal) smearings are applied to $E_t$ in order to obtain event-by-event reconstructed energies $E_r$. The MC generation is performed for signal and background separately, assuming inverse powerlaw distributions of $(E_t/100\text{GeV})^{-1}$ for signal and $(E_t/100\text{GeV})^{-2}$ for background. Reweighting of the MC can then be performed as a function of $E_t$ and forward-folded onto distributions in $E_r$ over which the events are histogrammed and likelihoods evaluated. For all toy analyses, the background component and signal width are kept fixed to their true values. Only the signal peak $\Omega$ and normalization $\Phi$ are treated as free parameters.

4.1 Point estimation

![Figure 2. Benchmark scenario.](image)

Leftmost panel: the underlying distribution that data is drawn from. Right three panels: observed data (black), $L_{\text{Eff}}$ best-fit MC distributions (blue), and MC uncertainties (blue band), with increasing MC statistics from left to right. The background component and signal width are fixed, while the mode and normalization of the signal peak are fit by maximizing the likelihood with respect to those parameters.
Figure 2 shows the expectation in $E_t$ as well as the data and $\mathcal{L}_{\text{Eff}}$ best-fit distributions in $E_r$. The left-most panel shows the expectation for both signal and background assuming no smearing in $E_t$. The three other panels show the smeared, $E_r$, distribution for data (black) and the best-fit result from $\mathcal{L}_{\text{Eff}}$ for three different MC datasets (blue) of varying statistics. The smeared shape of the signal peak is clearly visible in data, but not in the lowest statistics MC. As the MC increases in statistics, the best-fit MC can be seen to converge to data.

The best-fit values for the example shown in Fig. 2 are given in Table 1 for $\mathcal{L}_{\text{Poisson}}$ and $\mathcal{L}_{\text{Eff}}$. As point estimators, both likelihoods return similar values. This is driven by the fact that the same underlying MC distribution is used to fit to the data. The effect of convoluting $P(\lambda|\vec{w}(\hat{\theta}))$ mostly serves to broaden the likelihood space, while preserving the maximum within the constraints described in Sec. 3.3. In the high-statistics limit, both likelihoods can be used for unbiased point estimation, provided that the likelihood-space is smooth enough to for standard minimization techniques to probe the global minima.

| Likelihood   | $N_{MC} = 10^4$       | $10^5$       | $10^6$       |
|--------------|-----------------------|--------------|--------------|
| $\mathcal{L}_{\text{Poisson}}$ | (127.0, 6368.0)       | (124.7, 5655.7) | (125.1, 4888.5) |
| $\mathcal{L}_{\text{Eff}}$    | (127.1, 6077.1)       | (124.7, 5576.0) | (125.1, 4889.4) |

Table 1. **Best-fit parameters** for the toy experiment shown in Fig. 2 using $\mathcal{L}_{\text{Poisson}}$ and $\mathcal{L}_{\text{Eff}}$. The columns in the table are for the different MC statistics. The two numbers in parenthesis in each entry correspond to $(\Omega, \Phi)$.

4.2 Coverage

Due to the higher computational cost of generating frequentist confidence intervals by generating pseudodata to compute the test-statistic ($TS$) distribution, it is common to estimate frequentist coverage assuming Wilks’ theorem. While this is usually valid in the high-statistics regime, in the case of finite MC, a likelihood description based on MC expectations may lead to undercoverage even in the case of large data statistics. In this section, we will use $TS = \Delta l$. We evaluated the coverage properties, assuming Wilk’s theorem, of the two-dimensional fit over $(\Omega, \Phi)$ for several likelihood constructions. These include the modified-$\chi^2$, $\mathcal{L}_{\text{Poisson}}$, $\mathcal{L}_{\text{BB}}$, $\mathcal{L}_{\text{Mean}}$, and $\mathcal{L}_{\text{Eff}}$. These five likelihoods were chosen on the basis of their computation speed and as tests of different approaches towards the treatment of weighted MC.

Several configurations were tested, all under the assumptions of the toy experiment described in Sec. 4.1. The data had livetime set to 1 arb. unit. The MC was generated for two different settings of the total number of events: $10^3$ and $10^6$. For each setting, 500 toy experiments were generated, their best-fits found, and their $\Delta l = l(\hat{\theta}_{\text{true}}) - l(\hat{\theta})$ evaluated, where $\hat{\theta}_{\text{true}}$ and $\hat{\theta}$ corresponds to the true and best-fit $(\Omega, \Phi)$ respectively. Each toy experiment was classified as covering $\hat{\theta}_{\text{true}}$ at a specified level $p$ if $\Delta l < I(p; 2)$, where $I$ is the inverse of the $\chi^2$ cumulative density function and the 2 indicates the number of degrees of freedom.
Figure 3. Coverage properties. True coverage of the Wilks’ confidence interval for two sets of toy experiments with different amounts of MC statistics: $10^3$ events (left) and $10^6$ events (right). The Poisson likelihood severely undercovers. The modified-$\chi^2$, $L_{BB}$, and $L_{Mean}$ also undercover for low MC statistics. The effective likelihood derived in Sec. 3.1 performs best.

Figure 3 shows the percentage of times the true parameters were within the confidence intervals at level $p$ as a function of the estimated coverage percentile for that level. The test was performed for two settings of MC statistics, $N_{MC} = 10^3$ (left panel) and $N_{MC} = 10^6$ (right panel), using the different likelihoods. First note that, as expected, the true coverage is highly dependent on MC statistics, with higher MC statistics leading towards improved agreement. In the case of $N_{MC} = 10^3$, $L_{BB}$, $L_{Mean}$, modified-$\chi^2$, and $L_{Poisson}$ all undercover to varying degrees of severeness. For $N_{MC} = 10^6$, $L_{Poisson}$ still undercovers but the other finite-MC likelihoods exhibit good agreement. In this test, $L_{Eff}$ exhibits the best coverage properties.

4.3 Posterior distributions

It is also possible to use $L_{Eff}$ in a Bayesian approach. Using Bayes’ theorem, the posterior is proportional to

$$P(\bar{\theta} | k) \propto L(\bar{\theta} | k)\pi(\bar{\theta}),$$

(4.1)

where $\pi(\bar{\theta})$ is a prior on the parameters. As evaluation of the normalization factor can be challenging, $P(\bar{\theta} | k)$ can be approximated using a Markov Chain Monte Carlo (MCMC). For our toy example, we used emcee [15] to sample $P(\bar{\theta} | k)$ under a uniform box prior for two different likelihood functions: $L_{Eff}$ and $L_{Poisson}$. The sampling was performed using the data and MC sets described in Sec. 4.1.

Figure 4 shows the posterior distributions of $\Omega$ and $\Phi$ of the signal peak. For each comparison, $L_{Eff}$ (blue) and $L_{Poisson}$ (orange) were sampled under a uniform prior using the same underlying data and MC. We used 20 walkers with 300 burn-in steps followed by
Figure 4. **Likelihood scans and observable space.** Comparison of $P(\vec{\theta}|k)$ for $L_{\text{Eff}}$ (blue) and $L_{\text{Poisson}}$ (orange). Each horizontal row above uses a different MC set size, with $N_{\text{MC}} = 10^4$, $10^5$, and $10^6$ from top to bottom. The left and center column show the marginal posterior distribution for the mass, $\Omega$, and normalization, $\Phi$, respectively. The true value is indicated by the dashed, vertical line. The rightmost column shows the joint posterior distribution with 68% (solid) and 95% (dashed) contours. The truth is indicated by the star.

1000 steps as settings for emcee. The left and center column show the marginal posterior distribution for the mass, $\Omega$, and normalization, $\Phi$, respectively. The true value is indicated by the dashed, vertical line. The rightmost column shows the joint posterior distribution with 68% (solid) and 95% (dashed) contours. The truth is indicated by the star. With $L_{\text{Poisson}}$, the true value of the parameter is highly improbable for the lower MC-statistics cases of the top and middle rows. In contrast, the posterior evaluated using $L_{\text{Eff}}$ has
increased width due to the reduced MC statistics. Even for $N_{MC} = 10^6$ (bottom row) the shape of the posterior evaluated using $\mathcal{L}_{\text{Poisson}}$ is narrower than that using $\mathcal{L}_{\text{Eff}}$. Credible intervals estimated using $\mathcal{L}_{\text{Poisson}}$ would bias the result.

5 Conclusion

The use of MC to estimate expected outcomes of physical processes is nowadays standard practice. By construction, MC distributions are sample observations and subject to statistical fluctuations. MC events are also typically weighted to a particular physics model, and these weights may not be uniform across all events in a observable bin. A direct comparison of MC distributions to data is typically performed using Poisson or $\chi^2$ minimization, where the expectation from MC is computed as a sum over weights in a particular observable bin, may be inappropriate in the case of low MC statistics. Such likelihoods neglect the intrinsic MC fluctuations and may lead to vastly underestimated parameter uncertainties. A better approach is to construct a likelihood that accounts for MC statistical uncertainties.

The main result of this work is an analytic extension of the Poisson likelihood that accounts for MC statistical uncertainty. This new $\mathcal{L}_{\text{Eff}}$ is motivated by treating the MC realization as an observation of a Poisson random variate, computing the likelihood of the expectation using the MC, and convoluting with the Poisson probability of observed data. Our construction is computationally efficient, exhibits proper limiting behavior in both the high-statistics and equal-weights limits, and has excellent coverage properties. In our tests, it out-performs other treatments of MC statistical uncertainty.

Acknowledgements

We thank Thorsten Glüsenkamp for useful discussions. CAA is supported by U.S. National Science Foundation (NSF) grant PHY-1505858. AS and TY are supported in part by NSF grant PHY-1607644 and by the University of Wisconsin Research Committee with funds granted by the Wisconsin Alumni Research Foundation.

References

[1] J. S. Gainer, J. Lykken, K. T. Matchev, S. Mrenna and M. Park, Exploring Theory Space with Monte Carlo Reweighting, JHEP 10 (2014) 078 [1404.7129].
[2] L. Lyons, STATISTICS FOR NUCLEAR AND PARTICLE PHYSICISTS. 1986.
[3] R. J. Barlow and C. Beeston, Fitting using finite Monte Carlo samples, Comput. Phys. Commun. 77 (1993) 219.
[4] K. Cranmer, G. Lewis, L. Moneta, A. Shibata and W. Verkerke, HistFactory: A tool for creating statistical models for use with RooFit and RooStats,.
[5] D. Chirkin, Likelihood description for comparing data with simulation of limited statistics, 1304.0735.
[6] T. Glüsenkamp, Probabilistic treatment of the uncertainty from the finite size of weighted Monte Carlo data, Eur. Phys. J. Plus 133 (2018) 218 [1712.01293].
[7] C. Argüelles, A. Schneider and T. Yuan, “MCLLH.”
https://github.com/austinschneider/MCLLH, 2019.

[8] S. D. Poisson, Recherches sur la probabilité des jugements en matière criminelle et en matière civile précédées des règles générales du calcul des probabilités. Bachelier, 1837.

[9] K. Pearson, On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling, The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science 50 (1900) 157.

[10] B. K. Cogswell, D. J. Ernst, K. T. L. Ufheil, J. T. Gaglione and J. M. Malave, Neutrino oscillations: ILL experiment revisited, 1802.07763.

[11] G. Cowan, Statistical data analysis. Oxford university press, 1998.

[12] G. Bohm and G. Zech, Statistics of weighted Poisson events and its applications, Nucl. Instrum. Meth. A748 (2014) 1 [1309.1287].

[13] D. Fink, A compendium of conjugate priors, 1997.

[14] J. Bernardo and A. Smith, Bayesian Theory, Wiley Series in Probability and Statistics. Wiley, 2009.

[15] D. Foreman-Mackey, D. W. Hogg, D. Lang and J. Goodman, emcee: The MCMC Hammer, Publ. Astron. Soc. Pac. 125 (2013) 306 [1202.3665].