DRAFT 1.1

20 May, 2012

Statistical tests with weighted Monte Carlo events

Glen Cowan

Physics Department, Royal Holloway, University of London, Egham, TW20 0EX, U.K.

Abstract

In particle physics a search for a new signal process often takes the form of a frequentist statistical test based on observed numbers of collision events. In general the test makes reference to the numbers of events expected under different hypotheses, e.g., that they were produced by certain background processes, and in some cases these numbers are estimated using Monte Carlo simulations. To take into account the statistical fluctuations that are due to the limited size of the simulated event sample, the usual procedure is to treat the number of Monte Carlo events found passing the required selection criteria as a binomially (or approximately Poisson) distributed quantity. In many analyses in particle physics, however, the Monte Carlo events are generated with an associated weight, such that the sum of the weights plays the role of the number of events found, and thus the simple binominal or Poisson models are no longer sufficient. This paper examines several methods for using weighted Monte Carlo events in statistical tests such that the test outcome properly reflects the statistical fluctuations of the simulated and real data. The properties of the different approaches are investigated with numerical examples.

1 Introduction

Although it is usually desirable in searches for new processes to base background estimates on control samples of real data, this is not always possible and for at least some background components one may need to rely on Monte Carlo (MC) predictions. In some cases, the generated MC events may come with associated weights, and this complicates the statistical analysis of the data. In this note the use of weighted Monte Carlo events together with real data in statistical tests is discussed.

Section 2 outlines the basic analysis treated in this paper, namely, a measurement where one counts a certain number of events, n, in a region where signal and background are potentially both present, and one uses Monte Carlo to constrain the expected background. Section 3 describes weighted MC events, and Section 4 illustrates how these may be used in statistical tests. In Section 5 the special case of finding zero MC events is discussed. Several examples are shown in Section 6. Finally in Section 7 the treatment is extended to the case where the MC events only carry weights of either $+1$ or -1. Conclusions are given in Section 8.

2 Use of Monte Carlo in place of a control measurement

In a search for a new physics process, one often counts events in a region where signal may be present. Suppose the number found, n, is modeled as following a Poisson distribution with an expectation value

$$E[n] = \mu s + b \,, \tag{1}$$

where s and b are the expected numbers of events from the signal and background processes, respectively. Here μ is a strength parameter defined such that $\mu = 0$ corresponds to the background-only hypothesis and $\mu = 1$ gives the nominal signal model. We treat here the case where there is only one background component, but the idea is easily generalized to multiple components.

Often a control measurement is carried out to constrain the expected background b. This counts events in a region where little or no signal is expected. Suppose the number of events, m, is modeled as following a Poisson distribution with a mean value

$$E[m] = \tau b \,, \tag{2}$$

where τ is a scale factor that relates the size of the control region to that of the search region where n is measured. Here we will assume that τ can be determined with negligible uncertainty.

The problem described above has been widely studied, e.g., in Refs. [1, 2, 3]. Using the measured quantities n and m one can test different values of μ using a test statistic such as the profile likelihood ratio. In this way one can find a p-value for a specified value of μ, and if this is found below a given threshold then the value of μ is rejected. In particular, rejecting the background-only ($\mu = 0$) hypothesis is the first step in establishing the discovery of the signal process.

In some searches it is not practical or possible to carry out a control measurement for every background component, and instead an estimate based on a Monte Carlo simulation

is used. The mathematical formalism described above, however, holds equally well when the number m, normally from the control measurement, is obtained from Monte Carlo. That is, by treating the number of events found as an effective measurement, the statistical uncertainty due to the limited size of the MC sample is incorporated into the statistical test. In this case, one runs the MC simulation for the background process and simply counts the number of events appearing in the search region. The scale factor τ is then the ratio of the effective integrated luminosity of the MC sample to that of the data:

$$\tau = \frac{L_{\rm MC}}{L_{\rm data}} \; . \tag{3}$$

In this note we only consider the statistical uncertainty in the use of the MC events due to the limited number of events generated. There are also in general systematic uncertainties due to the imperfect modeling of the background processes, but for the present discussion these are neglected.

3 Monte Carlo with weighted events

Some Monte Carlo generators produce events with associated weights, and it is the sum of the weights rather than the number of events found that should be taken as the model estimate. In this section we review how the weights arise and how they can be treated in a statistical analysis. A related discussion of weighted Monte Carlo events can be found in Ref. [4].

Suppose events are characterized by a kinematic variable x (in general x can be a vector), which follows a density $f(x)$. The purpose of a Monte Carlo calculation is to find the probability $P_f(x \in A)$ for x to be in a specified acceptance region A:

$$P_f(x \in A) = \int_A f(x)\, dx \; . \tag{4}$$

Equivalently one may want to estimate the expected number of events ν in the region A given a total number of N events, i.e., one wants to find

$$\nu = N P_f(x \in A) \; . \tag{5}$$

It may be, however, that one does not have an MC model capable of generating $x \sim f(x)$, but rather one can generate x according to a different density $g(x)$. (Note both $f(x)$ and $g(x)$ here are normalized pdfs.) The desired probability $P_f(x \in A)$ can be written

$$\begin{aligned} P_f(x \in A) &= \frac{\int_A \frac{f(x)}{g(x)}\, g(x)\, dx}{\int_A g(x)\, dx} \int_A g(x)\, dx \\ &= E_g[w(x)|x \in A]\, P_g(x \in A) \; , \end{aligned} \tag{6}$$

where

$$w(x) = \frac{f(x)}{g(x)} \tag{7}$$

is the *weight function* and

$$P_g(x \in A) = \int_A g(x)\, dx \tag{8}$$

is the probability to find $x \in A$ assuming $x \sim g(x)$. That is, $P_f(x \in A)$ is the conditional expectation value of $w(x)$ with respect to $g(x)$ given $x \in A$ multiplied by the probability to find $x \in A$ under assumption of $g(x)$.

Suppose N values of x are generated according to $g(x)$ and n of them are found in the region A. Then the probability to be in A for $x \sim g(x)$ can be estimated by n/N, and the expectation value above can be estimated using the arithmetic average of the weights in A. Therefore the desired probability $P_f(x \in A)$ can be estimated using

$$\hat{P}_f(x \in A) = \frac{1}{n}\sum_{i=1}^{n} w_i \times \frac{n}{N} = \frac{1}{N}\sum_{i=1}^{n} w_i \,, \tag{9}$$

where $w_i = w(x_i)$ and we denote estimators for parameters with hats. Equivalently, we can estimate ν using the sum of the weights for the events in the acceptance region:

$$\hat{\nu} = \sum_{i=1}^{n} w_i \,. \tag{10}$$

The conditional expectation value and variance of $\hat{\nu}$ given a value of n are

$$E_g[\hat{\nu}|n] = n\omega \,, \tag{11}$$

$$V_g[\hat{\nu}|n] = n\sigma^2 \,, \tag{12}$$

where

$$\omega = E_g[w|x \in A] = \frac{P_f(x \in A)}{P_g(x \in A)} \,, \tag{13}$$

$$\sigma^2 = V_g[w|x \in A] \,, \tag{14}$$

are the conditional expectation and variance of $w(x)$ with respect to $g(x)$ given that x is found in the acceptance region A. The final equality in Eq. (13) follows from Eq. (6). In the following we will refer to ω and σ^2 simply as the mean and variance of the weights, and the expectations and variances below all refer to the density $g(x)$.

Strictly speaking we should model n as being binomially distributed, but we may approximate this as a Poisson distribution for $n \ll N$, which we will assume to hold. Then the expectation value and variance of n are $E[n] = \nu/\omega$, $V[n] = \nu/\omega$. The variance of the estimator $\hat{\nu}$ is therefore

$$V[\hat{\nu}] = E\left[V[\hat{\nu}|n]\right] + V\left[E[\hat{\nu}|n]\right] = E[n\sigma^2] + V[n\omega] = \frac{\nu}{\omega}(\sigma^2 + \omega^2) \,. \tag{15}$$

To estimate the variance $V[\hat{\nu}]$ we can use the sum of the squares of the weights,

$$\hat{\sigma}^2_{\hat{\nu}} = \sum_{i=1}^{n} w_i^2 \,, \tag{16}$$

and we demonstrate below that this estimator is unbiased. Since $\sigma_w^2 = E[w^2] - \omega^2$, the conditional expectation of $\hat{\sigma}_{\hat{\nu}}^2$ given n is

$$E\left[\hat{\sigma}_{\hat{\nu}}^2 | n\right] = \sum_{i=1}^{n} E[w_i^2] = n(\sigma_w^2 + \omega^2) \; , \tag{17}$$

and therefore the expectation value of $\hat{\sigma}_{\hat{\nu}}^2$ is

$$E\left[\hat{\sigma}_{\hat{\nu}}^2\right] = E\left[E\left[\hat{\sigma}_{\hat{\nu}}^2 | n\right]\right] = \frac{\nu}{\omega}(\sigma_w^2 + \omega^2) \; . \tag{18}$$

This is the same as the true variance from Eq. (15), and therefore the sum of the squares of the weights (16) is an unbiased estimator for the variance of $\hat{\nu}$. That is, we can take the sum of the weights to estimate the number of events that would be present if all had a weight of unity, and we use the square root of the sum of the squares of weights for the corresponding standard deviation (i.e., the statistical error).

4 Using weighted MC events in a statistical test

To define a test of a hypothesized value of the strength parameter μ one usually defines a test statistic q_μ that reflects the level of agreement between the data and μ. Methods for doing this using the profile-likelihood ratio are described, for example, in Ref. [1]. For this or other methods for constructing a test, one requires a model for how the data are distributed. Therefore in this section we consider possible models and write down for them the corresponding likelihood functions.

Consider a search of the type described in Sec. 2 where we observe n events in a search region where signal may be present, and we model n as following a Poisson distribution with mean $\mu s + b$. We take s, the mean number of events predicted by the nominal signal model, as known, and the parameter of interest is the strength parameter μ. The expected number of background events b is a nuisance parameter, and here we will use a Monte Carlo simulation with weighted events to constrain its value. Thus the "measured outcomes" consist of n events from real data found in the search region, m events found in the same search region but from background Monte Carlo, and m weights: $w_1, \ldots, w_m$.

We assume that n and m are Poisson distributed,

$$n \sim \text{Poisson}(\mu s + b) \; , \tag{19}$$

$$m \sim \text{Poisson}(\tau b/\omega) \; , \tag{20}$$

where τ is the luminosity ratio from Eq. (3) and ω is the mean weight for events in the acceptance region. The weights themselves will follow a certain distribution, but in practice it will not be possible to write this down in closed form. If we wish to treat the weights as part of the measured outcome, however, we need to make some assumption for their distribution.

Below we explore a normal and log-normal pdf for the weights, and argue that the log-normal is in some ways better justified and in any case more conservative. In addition we consider the case where the weighted events are used to construct directly an estimate of the background parameter b, and this estimate is then modeled as following a Gaussian distribution.

4.1 Normal distribution of weights

To construct a statistical test of a hypothesized value of μ, we will need the likelihood function L. Suppose that the weights w follow a Gaussian distribution with mean ω and standard deviation σ_w. By combining this with the Poisson distributions for n and m, the full likelihood function can be written

$$L(\mu, b, \omega, \sigma_w) = \frac{(\mu s + b)^n}{n!} e^{-(\mu s + b)} \, \frac{(\tau b/\omega)^m}{m!} e^{-\tau b/\omega} \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi}\sigma_w} e^{(w_i - \omega)^2/2\sigma_w^2} \; . \tag{21}$$

The log-likelihood is therefore

$$\begin{aligned} \ln L(\mu, b, \omega, \sigma_w) &= n \ln(\mu s + b) - (\mu s + b) + m \ln(\tau b/\omega) - \tau b/\omega \\ &- m \ln \sigma_w - \frac{m\omega^2}{2\sigma_w^2} + \frac{\omega}{\sigma_w^2} \sum_{i=1}^{m} w_i - \frac{1}{2\sigma_w^2} \sum_{i=1}^{m} w_i^2 + C \; , \end{aligned} \tag{22}$$

where C represents terms that do not depend on the parameters and thus can be dropped. In this expression, the data only enter through n, m, and the sums of the weights and the weights squared:

$$S_1 = \sum_{i=1}^{m} w_i \; , \tag{23}$$

$$S_2 = \sum_{i=1}^{m} w_i^2 \; . \tag{24}$$

That is, n, m, S_1 and S_2 form a set of sufficient statistics, and thus if we know them it is not necessary to retain the values of every weight individually. One can show this is equivalent to regarding the data outcomes as n, m, S_1 and S_2, where S_1 follows a Gaussian distribution with a mean $m\omega$ and standard deviation $\sqrt{m}\sigma_w$, and the quantity

$$\frac{1}{\sigma_w^2} \left(S_2 - \frac{1}{m} S_1^2 \right) \tag{25}$$

follows a chi-square distribution for $m - 1$ degrees of freedom.

To define a test statistic based on the likelihood function, we will need to find the values of the parameters that maximize $\ln L$. If the measurement gives $m = 0$, however, then the terms involving the weights are all zero (there are no events that carry weights), and σ_w decouples from the problem completely. Furthermore, maximizing $\ln L$ with respect to ω in this case gives $\omega \to \infty$.

From Eq. (13) we see that ω is the ratio of probabilities to find $x \in A$ under assumption of the two densities $f(x)$ and $g(x)$, and in general this ratio will not be known. It may be possible in some cases to place an upper bound on ω such that it is known to be not greater than a given value $\omega_{\max}$. For purposes of defining the test statistic described in Sec. 4.4, and in the absence of any further information, for $m = 0$ we set $\omega = 1$. If a bound $\omega_{\max}$ is known one could use this value instead. This issue is discussed further in Sec. 5.

4.2 Log-normal distribution of weights

As defined in the present problem, the weight function $w(x) = f(x)/g(x)$ is strictly non-negative, and thus a Gaussian distribution for w is at best an approximation. Furthermore it may be in practical examples that the tails of the weight distribution are substantially longer than those of a Gaussian. As an alternative, therefore, we may take the logarithm of the weights as Gaussian distributed, which is to say that w follows a log-normal pdf.

To write down this model we define l as

$$l = \ln w \ , \tag{26}$$

and assume that l follows a Gaussian distribution with mean λ and standard deviation σ_l. The log-likelihood is then found to be (up to an additive constant)

$$\begin{aligned} \ln L(\mu, b, \lambda, \sigma_l) &= n \ln(\mu s + b) - (\mu s + b) + m \ln(\tau b/\omega) - \tau b/\omega \\ &- m \ln \sigma_l - \frac{m\lambda^2}{2\sigma_l^2} + \frac{\lambda}{\sigma_l^2} \sum_{i=1}^{m} l_i - \frac{1}{2\sigma_l^2} \sum_{i=1}^{m} l_i^2 \ . \end{aligned} \tag{27}$$

The quantity ω above is as before the expectation value of $w = e^l$. Because w follows a log-normal distribution, one can show ω is related to $\lambda = E[\ln w]$ by

$$\omega = \exp\left(\lambda + \tfrac{1}{2}\sigma_l^2\right) \ . \tag{28}$$

As in the case of Gaussian distributed weights, the likelihood function for the log-normal model depends only on four measured quantities: n, m, the sum of the weights and the sum of the squares of the weights. And as in Sec. 4.1 for the case of Gaussian weights, if $m = 0$ then for purposes of determining the test statistic described in Sec. 4.4, we take $\lambda = 0$ and $\sigma_l = 0$ (and therefore $\omega = 1$). This issue is discussed further in Sec. 5.

4.3 Normal distribution for estimate of b

We may use the Monte Carlo events to construct directly an estimator for the background parameter b using

$$\hat{b} = \frac{1}{\tau} \sum_{i=1}^{m} w_i \ , \tag{29}$$

and by making use of the results from Sec. 3, the variance of $\hat{b}$ can be estimated by

$$\hat{\sigma}_{\hat{b}}^2 = \frac{1}{\tau^2} \sum_{i=1}^{m} w_i^2 \ . \tag{30}$$

We can then regard the outcomes of the measurement to be n and $\hat{b}$. We model n as Poisson distributed with mean $\mu s + b$, and for sufficiently large m, because of the central limit theorem, $\hat{b}$ will follow a Gaussian distribution with mean b and a variance given by Eq. (30). The log-likelihood function is therefore found to be (up to an additive constant)

$$\ln L(\mu, b) = n \ln(\mu s + b) - (\mu s + b) - \frac{1}{2} \frac{(b - \hat{b})^2}{\hat{\sigma}_{\hat{b}}^2} \; . \tag{31}$$

This statistic may be used to define a test of a hypothesized value of μ, and for sufficiently large m the power of such a test should not differ substantially from a test based on the likelihoods of the previous sections. Because the denominator in the final term of (31) must not be zero, the statistic is only defined for $m > 0$, and in practice one would only use this approach in cases where the probability of $m = 0$ can be neglected. It provides the advantage, compared to the tests based on the normal or log-normal distribution of weights, that the likelihood depends only on the parameter of interest μ and the single nuisance parameter b.

4.4 Statistical test using the profile likelihood ratio

Using one of the likelihood functions given in the previous section, we can construct a statistic to test a hypothesized value of the strength parameter μ. Methods for doing this using the profile likelihood ratio are discussed in Ref. [1]. These methods assume that the likelihood function can be written as a function of the parameter of interest, μ, and a set of nuisance parameters, which we denote in this section as θ. For the likelihoods considered here, the nuisance parameters are b together with ω and σ_w for the normal-weight model and λ and σ_l for the log-normal model.

To define a test of μ, we define the profile likelihood ratio as

$$\lambda(\mu) = \frac{L(\mu, \hat{\hat{\theta}})}{L(\hat{\mu}, \hat{\theta})} \; . \tag{32}$$

(Note this $\lambda(\mu)$ is not to be confused with the parameter $\lambda = E[\ln w]$ used in the log-normal model.) The numerator of (32) is the profile likelihood, and $\hat{\hat{\theta}}$ denotes the value of θ that maximizes L for the given μ (thus $\hat{\hat{\theta}}$ is a function of μ). The denominator is the maximum of the likelihood function, i.e., the single hats denote the maximum-likelihood estimators for both μ and θ.

Often one considers signal models where the strength parameter μ must satisfy $\mu \geq 0$. Even if this holds, however, it is convenient to allow $\hat{\mu}$ to take on negative values, because then for a sufficiently large data sample it can be modeled as following a Gaussian distribution. Allowing this to be the case, we can define a statistic for a one-sided test of the $\mu = 0$ hypothesis as

$$q_0 = \begin{cases} -2 \ln \lambda(0) & \hat{\mu} \geq 0 \; , \\ 0 & \hat{\mu} < 0 \; , \end{cases} \tag{33}$$

where $\lambda(0)$ is the profile likelihood ratio for $\mu = 0$ as defined in Eq. (32). Large values of q_0 correspond to increasing disagreement with the $\mu = 0$ hypothesis, and thus for an observed value of the statistic $q_{0,\mathrm{obs}}$ the p-value of the $\mu = 0$ hypothesis is

$$p = \int_{q_{0,\mathrm{obs}}}^{\infty} f(q_0|0) \, dq_0 \; . \tag{34}$$

Often one converts the p-value to the equivalent Gaussian significance,

$$Z = \Phi^{-1}(1-p) \ , \tag{35}$$

where Φ^{-1} is the standard normal quantile (inverse of the standard normal cumulative distribution).

In this procedure for finding the p-value of significance, therefore, one makes an assumption about the distribution of the data, which here includes the distribution of the weights, in two distinct places. First, we need the likelihood function in order to write down the profile likelihood ratio, $\lambda(\mu)$, which is used to test different values of μ. If the assumed model is not correct, then the resulting test will not be optimal, in the sense that it will not necessarily have the maximal power with respect to a given alternative hypothesis. Nevertheless a test statistic based on an incorrect likelihood can be used to construct a valid statistical test.

Second, we need distribution of the data to determine $f(q_0|0)$ in order to compute the p-value using Eq. (34). But in general this distribution is depends on the nuisance parameters, e.g., $\theta = (b, \omega, \sigma_w)$, and should more correctly be written $f(q_0|0,\theta)$.

By constructing the test statistic from the profile likelihood ratio of Eq. (32), one can show that the distribution of $-2\ln\lambda(\mu)$ under assumption of μ approaches an asymptotic form related to a chi-square distribution that is independent of the nuisance parameters. For small data samples this form no longer holds, but in practical examples the dependence of the result on the nuisance parameters is often found to be weak (see, e.g., Ref. [1]).

For purposes of investigating tests that include weighted Monte Carlo events, we will use as an example the statistic q_0. Other statistics appropriate for different types of tests, e.g., upper limits and two-sided limits, are discussed in Ref. [1]. Under assumption of the large-sample distribution of q_0, the discovery significance, i.e., the significance Z obtained from the p-value of the $\mu = 0$ hypothesis, can be found from the simple formula

$$Z = \sqrt{q_0} \ . \tag{36}$$

For problems with very few events, however, it is recommended not to trust the asymptotic formulae and to determine $f(q_0|0,\theta)$ using Monte Carlo. This requires a choice for the nuisance parameters, and the validity of the p-values and significances rely on this choice. In principle one would like the p-value of μ to hold for any point in nuisance-parameter space, and thus to be conservative one would scan this space and use the largest p-value found. In practice this is not feasible and one is generally satisfied if the p-value would be valid if the true values of the nuisance parameters are within some specified region.

A good starting point when testing a hypothesized value of μ is to take the values of the nuisance parameters that maximize the likelihood function under assumption of μ, e.g., $\hat{\hat{b}}(\mu)$, $\hat{\hat{\omega}}(\mu)$, $\hat{\hat{\sigma}}_w(\mu)$, etc. This is the basis of the "profile construction" discussed in Ref. [2]. In most problems this will be sufficient, but in some cases it may be necessary to consider a larger region. For purposes of the present study we will suppose that a point in the nuisance-parameter space has somehow been chosen, and the p-values are valid under assumption of that choice.

5 The case of $m = 0$

A case of particular interest is when no Monte Carlo events are found, i.e., $m = 0$. Here the procedure of Sec. 4.3 is not applicable, as this requires an estimate of b and of its variance,

and there would be no information on which to base this. If, however, we use the Gaussian or log-normal distribution of weights as in Sections 4.1 or 4.2, then for both cases when $m = 0$ the log-likelihood function becomes

$$\ln L(\mu, b, \omega) = n \ln(\mu s + b) - (\mu s + b) - \frac{\tau b}{\omega} \; . \tag{37}$$

Consider first the case where the mean weight of the Monte Carlo events, ω, is known a priori. This includes the special case where all events have unit weight. Then the log-likelihood function (37) contains only the parameter of interest μ and the single nuisance parameter b. This can be used to construct the profile likelihood ratio as described in Sec. 4.4, which in turn can be used to test different values of μ.

As an example, consider using the statistic q_0 defined in Eq. (33) to test $\mu = 0$. For a sufficiently large data sample, one can show that the significance Z with which one rejects the $\mu = 0$ hypothesis obeys the simple formula (see, e.g., Ref. [1]),

$$Z = \sqrt{q_0} \; . \tag{38}$$

For $m = 0$ this reduces to

$$Z = \sqrt{2n \ln \left(1 + \frac{\tau}{\omega}\right)} \tag{39}$$

This significance is shown for $\omega = 1$ as a function of τ in Fig. 1.

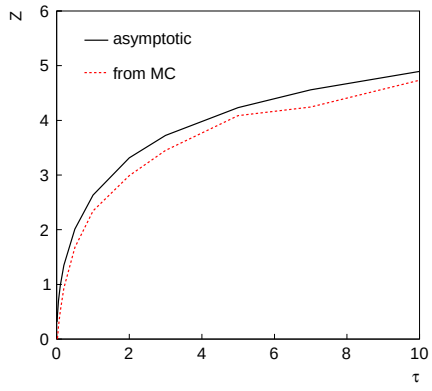

Figure 1: The significance Z versus the luminosity scale factor τ from the asymptotic formula (39) and from Monte Carlo for the example with $n = 5$, $m = 0$ (see text).

In the current example with $n = 5$ and $m = 0$, however, one would not necessarily expect the asymptotic formula to be accurate. Figure 1 shows the significance calculated from this formula compared to the value obtained from Monte Carlo, and in fact the agreement is found to be reasonably close. Here for each value of τ one must generate events according to the hypothesis of $\mu = 0$ using given values for the nuisance parameters. As recommended in Sec. 4.4, we choose these to be the values that maximize the likelihood for $\mu = 0$ (those written with a double hat). For the present example we find

$$\hat{\hat{b}}(0) = \frac{n}{1 + \tau/\omega} \; , \tag{40}$$

which means that we use a different value of b for each value of τ.

If we have observed $m = 0$, then there is no information on which to base an estimate of the mean or standard deviation of the weights. In the example shown in Fig. 1, we set $\omega = 1$ and $\sigma_w = 0$, which is equivalent to having all events carry unit weight.

In this procedure one obtains an estimate of b and one could also estimate its standard deviation, but this "error on b" does not enter explicitly into the inference about μ. Instead one proceeds directly from the likelihood function involving μ and b to a p-value (or significance) for μ.

If the mean weight ω is not known, then maximizing $\ln L$ from Eq. (37) gives $\omega \to \infty$, and the statistic q_0 is zero for all n. That is, if there is absolutely no constraint on ω, then if $m = 0$, any observed value of n could is fully compatible with the background-only hypothesis, and no discovery is possible.

If an upper bound $\omega_{\rm max}$ is available, then assumption of this value will overestimate the background and thus provide a conservative test of $\mu = 0$. If a lower bound $\omega_{\rm min}$ is known, then this will lead in general to an underestimate of the background and therefore a conservative upper limit on μ.

If an independent estimate $\hat{\omega}$ and an estimate of its standard deviation, $\hat{\sigma}_{\hat{\omega}}$ are available (perhaps from a separate Monte Carlo study) then these can be used in a manner analogous to how the estimate of b was used in Sec. 4.3. If this estimate is treated as a Gaussian distributed measurement, for example, then the likelihood function would be multiplied by an additional Gaussian term, and the log-likelihood function would have an addition term of the form

$$-\frac{1}{2}\frac{(\omega - \hat{\omega})^2}{\hat{\sigma}_{\hat{\omega}}^2} \; . \tag{41}$$

Whether done in this or some other manner, if $m = 0$ one must include some information about the possible values of the weights in order to make any inference about μ.

6 Examples

In order to investigate the properties of the statistical tests discussed above, suppose that we want to generate events each characterized by a value x that follows an exponential distribution truncated at $x = a$, i.e.,

$$f(x) = \frac{e^{-x/\xi}}{\xi(1 - e^{-a/\xi})} \; , \quad 0 \leq x \leq a \; . \tag{42}$$

Of course it is easy in this case to generate $x \sim f(x)$, e.g., using the transformation method (see, e.g., [5, 6]). But suppose for soem reason we were unable to do this and instead could only generate events following a uniform distribution with $0 \leq x \leq a$, i.e.,

$$g(x) = \frac{1}{a} \; , \quad 0 \leq x \leq a \; , \tag{43}$$

and for each event we also obtain the weight,

$$w(x) = \frac{f(x)}{g(x)} = \frac{a}{\xi}\frac{e^{-x/\xi}}{1 - e^{-a/\xi}} \; . \tag{44}$$

In this case it is easy to show that if $x \sim g(x)$, then the distribution of the weights is

$$p(w) = \frac{\xi}{aw} , \qquad w_{\min} \ \leq \ w \ \leq w_{\max} , \tag{45}$$

with $w_{\min}$ and $w_{\max}$ found from Eq. (44) for $x = a$ and $x = 0$, respectively. The distribution of the log of the weights is then uniform in the interval $[\ln(w_{\min}), \ln(w_{\max})]$. Here the ratio ξ/a determines the width of the weight distribution. If ξ/a is large, the weights are all close to unity, and if it is small, then the weights take on a broad range of values.

As a first example, suppose we take $a = 5$, $\xi = 25$, so the weights are all close to unity. Further by taking $b = 6$, $s = 10$ and $\tau = 1$, we can generate the data shown in Table 1.

Table 1: The weights and log-weights for a data set consisting of $m = 6$ weighted events, generated using $a = 25$, $\xi = 25$.

weight w	$\ln w$
0.9684	-0.0320
0.9217	-0.0816
1.0238	0.0235
1.0063	0.0063
0.9709	-0.0295
1.0813	0.0782

Suppose that in addition to the 6 events found in the weighted MC sample, $n = 17$ events were found in the search region, and we want to test the background-only hypothesis ($\mu = 0$). The three likelihoods described in Sec. 4 are used as the basis of the profile likelihood ratio, which determines the statistic q_0. To find the distribution $f(q_0|0)$, the weights are distributed according to the $1/w$ model described above, and also according to the model used to define the statistic, namely, normal or log-normal (for the statistic based on a normal distribution of $\hat{b}$, the normal distribution of w is used). For the normal and log-normal models, the mean and variance of the weights are computed so as to be the same as in the $1/w$ model. From the distributions of q_0 the p-value of the $\mu = 0$ hypothesis is found and this is converted into the equivalent Gaussian significance Z according to Eq. (35). These are shown in Table 2.

Table 2: The significance Z with which one would reject the $\mu = 0$ hypothesis given $n = 17$ and the Monte Carlo data from Table 1, which is based on $a = 5$, $\xi = 25$, and thus has a narrow distribution of weights.

Likelihood used to define q_0	Distribution of w for $f(q_0\|0)$	Significance Z to reject $\mu = 0$
$w \sim$ normal	normal	2.287
$w \sim$ normal	$1/w$	2.268
$w \sim$ log-normal	log-normal	2.301
$w \sim$ log-normal	$1/w$	2.267
$\hat{b} \sim$ normal	normal	2.289
$\hat{b} \sim$ normal	$1/w$	2.224

As seen from Table 2, the different approaches all give very similar answers. The reason is that all of the weights are close to unity, and the result is thus insensitive to exactly how they are distributed about this value.

If, however, we choose $a = 5$, $\xi = 1$, then the distribution of weights is much broader. Making this change in the input parameters results in the six events shown in Table 3. The distributions of the statistic q_0 for this case are shown in Fig. 2. Figure 2(a) for the case where the test statistic is based on the normal model for w, and in Fig. 2(b) q_0 is based on the log-normal model. The two distributions in each plot refer to the two different assumptions for the distribution of weights used to determine $f(q_0|0)$, namely, the same model as used to define the test statistic (normal or log-normal), and the $1/w$ model.

Table 3: The weights and log-weights for a data set consisting of $m = 6$ weighted events, generated using $a = 25$, $\xi = 1$.

weight w	$\ln w$
0.1934	-1.6429
0.0561	-2.8809
0.7750	-0.2548
0.5039	-0.6853
0.2059	-1.580
3.0404	1.1120

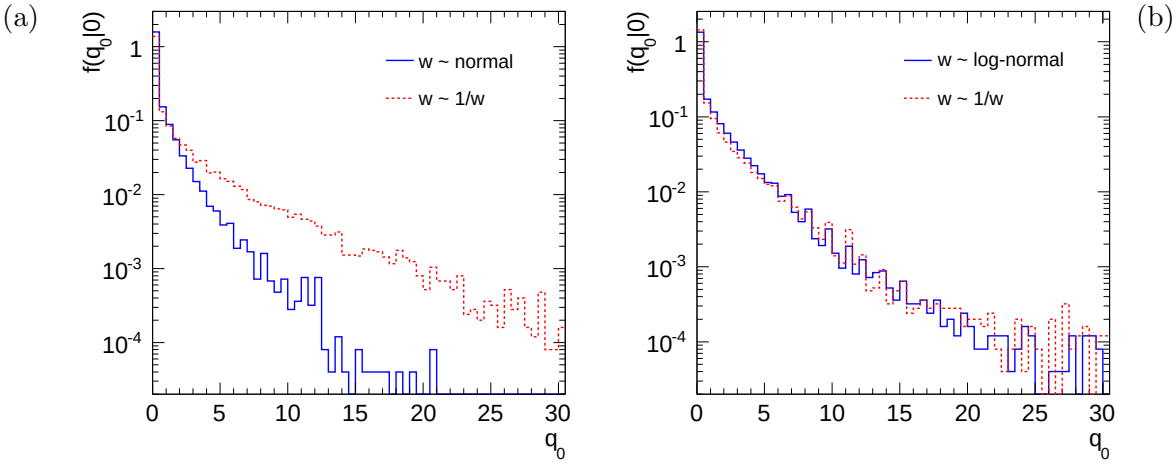

Figure 2: Distributions of the statistic q_0 based the profile likelihood using (a) a normal model for the weights and (b) on a log-normal model. In each plot the curves are shown representing two assumptions for the distribution of weights: the same as used to define q_0 (normal or log-normal) and the $1/w$ distribution.

These distributions result in the significance values for a test of the $\mu = 0$ hypothesis shown in Tab. 4. Looking at the values, one sees that if a test statistic based on the log-normal likelihood for w is used, then the discovery significance is substantially lower, and here both the log-normal and $1/w$ distributions used to determine $f(q_0|0)$ lead to similar results.

If, however, a test statistic based on a normal-distribution for w is assumed, then the Z value is substantially higher. Furthermore, when the data distribution used for $f(q_0|0)$ was also based on a normal distribution for w, then Z was substantially higher (2.163) than obtained if a $1/w$ distribution is used. Using the test statistic based on a normal distribution of $\hat{b}$ is qualitatively similar to using the one based on normally distributed weights.

These examples show that to decide how best to treat the weighted events, there are two important considerations. First, the outcome of the statistical test depends on the assumed

Table 4: The significance Z with which one would reject the $\mu = 0$ hypothesis given $n = 17$ and the Monte Carlo data from Table 1, which is based on $a = 5$, $\xi = 1$, and thus has a broad distribution of weights.

Likelihood used to define q_0	Distribution of w for $f(q_0\|0)$	Significance Z to reject $\mu = 0$
$w \sim$ normal	normal	2.163
$w \sim$ normal	$1/w$	1.308
$w \sim$ log-normal	log-normal	0.863
$w \sim$ log-normal	$1/w$	0.983
$\hat{b} \sim$ normal	normal	1.788
$\hat{b} \sim$ normal	$1/w$	1.387

distribution of the data. Here assumption of a normal distribution for the weights used to find $f(q_0|0)$ led to a stronger constraint on the expected background and thus to a higher value of Z for the test of $\mu = 0$. If the weights were to have a broader distribution (e.g., the $1/w$ model), then the background is less well constrained and the Z values are lower.

Second, if one wants to allow for the possibility that the weights could a distribution with long tails, then the test statistic used should be sensitive to the presence of such tails. In the example above when using the statistic q_0 defined with the Gaussian likelihood for the weights, even if the data were generated following the $1/w$ model, the significance Z was found to have a value of 1.308. With the same data distribution but when using a statistic based on the log-normal likelihood, one obtained $Z = 0.983$. If the distribution of the weights really does follow the $1/w$ model, then both values of Z are correct, but correspond to the results of different tests, each having different sensitivity to the tails of the weight distribution.

7 Weights of ± 1

When the weights of the Monte Carlo events are assigned according to the procedure described in Sec. 3, they are always non-negative. Some MC generators, however, such as the MC@NLO program [7] produce events with weights of either ± 1. These can be handled by a simple generalization of the ideas described above.

Suppose as previously that the number of events found in a search, n, is Poisson distributed with a mean of $\mu s + b$. Then, the Monte Carlo program yields m_+ events with weight $+1$ and m_- events with weight -1. These can each be treated as Poisson variables with mean values

$$E[m_+] = \tau b p_+ \,, \tag{46}$$

$$E[m_-] = \tau b (1 - p_+) \,. \tag{47}$$

Here p_+ is the probability for an MC event found in the acceptance region to have a weight of $+1$. The likelihood function for n, m_+ and m_- is therefore

$$L(\mu, b, p_+) = \frac{(\mu s + b)^n}{n!} e^{-(\mu+b)} \, \frac{(\tau b p_+)^{m_+}}{m_+!} e^{-\tau b p_+} \, \frac{(\tau b (1 - p_+))^{m_-}}{m_-!} e^{-\tau b (1-p_+)} \,. \tag{48}$$

This likelihood may then be used to construct the profile likelihood ratio according to Eq. (32), which is then used to test hypothesized values of μ.

In the previous discussion where the weight function $w(x) = f(x)/g(x)$ was non-negative but could take on a continuum of values, we needed some assumption about the distribution of weights in order to construct a test statistic, and this assumption contained additional parameters. Here we simply assume that both m_+ and m_- are Poisson distributed, and the only additional parameter is the probability p_+ for an accepted event to have weight of $+1$. Beyond that the procedures for carrying out tests of μ are the same as before.

8 Conclusions

A correct treatment of weighted Monte Carlo events in statistical tests requires an assumption for the distribution of the weights. Furthermore, optimal tests should be based on a statistic that incorporates this distribution, e.g., in a likelihood ratio.

If the distribution of weights is very broad, then the information that constrains the estimated number of events is relatively weak. A particular danger is if the distribution of weights is assumed to have tails that fall off very quickly, such as with a Gaussian, whereas in reality the weight distribution could have longer tails. In this case, one would overestimate the accuracy of the estimated number of background events, and could be led to an unjustifiably large estimate of discovery significance. A log-normal model for the weights has longer tails than a Gaussian and is thus more conservative.

Treatment of weights that only take on the values of ± 1 is straightforward and requires that one record the numbers of events with both positive and negative weights, not only their difference. A single additional nuisance parameter, namely, the probability for a positive weight, is introduced, and can be treated, e.g., using profile likelihood methods.

References

[1] Glen Cowan, Kyle Cranmer, Eilam Gross and Ofer Vitells, Eur. Phys. J. C 71 (2011) 1-19; arXiv:1007.1727.

[2] Kyle Cranmer, *Statistical Challenges for Searches for New Physics at the LHC*, proceedings of PhyStat2005, Oxford; arXiv:physics/0511028.

[3] Robert D. Cousins, James T. Linnemann and Jordan Tucker, NIM A 595 (2008) 480–501; arXiv:physics/0702156.

[4] Frederick James, *Statistical Methods in Experimental Physics, 2nd ed.*, World Scientific, 2006.

[5] G. Cowan, *Statistical Data Analysis*, Oxford University Press, 1998.

[6] K. Nakamura et al. (Particle Data Group), J. Phys. G 37, 075021 (2010). Monte Carlo techniques are described in Ch. 34.

[7] S. Frixione and B.R. Webber, *Matching NLO QCD computations and parton shower simulations*, JHEP 0206 (2002) 029 [hep-ph/0204244].