Eur. Phys. J. C (2011) 71: 1554
DOI 10.1140/epjc/s10052-011-1554-0

THE EUROPEAN PHYSICAL JOURNAL C

Special Article - Tools for Experiment and Theory

Asymptotic formulae for likelihood-based tests of new physics

Glen Cowan[1], **Kyle Cranmer**[2], **Eilam Gross**[3], **Ofer Vitells**[3,a]

[1]Physics Department, Royal Holloway, University of London, Egham TW20 0EX, UK
[2]Physics Department, New York University, New York, NY 10003, USA
[3]Weizmann Institute of Science, Rehovot 76100, Israel

Received: 15 October 2010 / Revised: 6 January 2011 / Published online: 9 February 2011

Abstract We describe likelihood-based statistical tests for use in high energy physics for the discovery of new phenomena and for construction of confidence intervals on model parameters. We focus on the properties of the test procedures that allow one to account for systematic uncertainties. Explicit formulae for the asymptotic distributions of test statistics are derived using results of Wilks and Wald. We motivate and justify the use of a representative data set, called the "Asimov data set", which provides a simple method to obtain the median experimental sensitivity of a search or measurement as well as fluctuations about this expectation.

1 Introduction

In particle physics experiments one often searches for processes that have been predicted but not yet seen, such as production of a Higgs boson. The statistical significance of an observed signal can be quantified by means of a p-value or its equivalent Gaussian significance (discussed below). It is useful to characterize the sensitivity of an experiment by reporting the expected (e.g., mean or median) significance that one would obtain for a variety of signal hypotheses.

Finding both the significance for a specific data set and the expected significance can involve Monte Carlo calculations that are computationally expensive. In this paper we investigate approximate methods based on results due to Wilks [1] and Wald [2] by which one can obtain both the significance for given data as well as the full sampling distribution of the significance under the hypothesis of different signal models, all without recourse to Monte Carlo. In this way one can find, for example, the median significance and also a measure of how much one would expect this to vary as a result of statistical fluctuations in the data.

A useful element of the method involves estimation of the median significance by replacing the ensemble of simulated data sets by a single representative one, referred to here as the "Asimov" data set.[1] In the past, this method has been used and justified intuitively (e.g., [4, 5]). Here we provide a formal mathematical justification for the method, explore its limitations, and point out several additional aspects of its use.

The present paper extends what was shown in [5] by giving more accurate formulas for exclusion significance and also by providing a quantitative measure of the statistical fluctuations in discovery significance and exclusion limits. For completeness some of the background material from [5] is summarized here.

In Sect. 2 the formalism of a search as a statistical test is outlined and the concepts of statistical significance and sensitivity are given precise definitions. Several test statistics based on the profile likelihood ratio are defined.

In Sect. 3, we use the approximations due to Wilks and Wald to find the sampling distributions of the test statistics and from these find p-values and related quantities for a given data sample. In Sect. 4 we discuss how to determine the median significance that one would obtain for an assumed signal strength. Several example applications are shown in Sect. 5, and numerical implementation of the methods in the RooStats package is described in Sect. 6. Conclusions are given in Sect. 7.

2 Formalism of a search as a statistical test

In this section we outline the general procedure used to search for a new phenomenon in the context of a frequentist statistical test. For purposes of discovering a new signal process, one defines the null hypothesis, H_0, as describing only known processes, here designated as background. This

[a] e-mail: ofer.vitells@weizmann.ac.il

[1]The name of the Asimov data set is inspired by the short story *Franchise*, by Isaac Asimov [3]. In it, elections are held by selecting the single most representative voter to replace the entire electorate.

is to be tested against the alternative H_1, which includes both background as well as the sought after signal. When setting limits, the model with signal plus background plays the role of H_0, which is tested against the background-only hypothesis, H_1.

To summarize the outcome of such a search one quantifies the level of agreement of the observed data with a given hypothesis H by computing a p-value, i.e., a probability, under assumption of H, of finding data of equal or greater incompatibility with the predictions of H. The measure of incompatibility can be based, for example, on the number of events found in designated regions of certain distributions or on the corresponding likelihood ratio for signal and background. One can regard the hypothesis as excluded if its p-value is observed below a specified threshold.

In particle physics one usually converts the p-value into an equivalent significance, Z, defined such that a Gaussian distributed variable found Z standard deviations above[2] its mean has an upper-tail probability equal to p. That is,

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

where Φ^{-1} is the quantile (inverse of the cumulative distribution) of the standard Gaussian. For a signal process such as the Higgs boson, the particle physics community has tended to regard rejection of the background hypothesis with a significance of at least $Z = 5$ as an appropriate level to constitute a discovery. This corresponds to $p = 2.87 \times 10^{-7}$. For purposes of excluding a signal hypothesis, a threshold p-value of 0.05 (i.e., 95% confidence level) is often used, which corresponds to $Z = 1.64$.

It should be emphasized that in an actual scientific context, rejecting the background-only hypothesis in a statistical sense is only part of discovering a new phenomenon. One's degree of belief that a new process is present will depend in general on other factors as well, such as the plausibility of the new signal hypothesis and the degree to which it can describe the data. Here, however, we only consider the task of determining the p-value of the background-only hypothesis; if it is found below a specified threshold, we regard this as "discovery".

It is often useful to quantify the sensitivity of an experiment by reporting the expected significance one would obtain with a given measurement under the assumption of various hypotheses. For example, the sensitivity to discovery of a given signal process H_1 could be characterized by the expectation value, under the assumption of H_1, of the value of Z obtained from a test of H_0. This would not be the same as the Z obtained using (1) with the expectation of the p-value, however, because the relation between Z and p is nonlinear. The median Z and p will, however, satisfy (1) because this is a monotonic relation. Therefore in the following we will take the term 'expected significance' always to refer to the median.

A widely used procedure to establish discovery (or exclusion) in particle physics is based on a frequentist significance test using a likelihood ratio as a test statistic. In addition to parameters of interest such as the rate (cross section) of the signal process, the signal and background models will contain in general *nuisance parameters* whose values are not taken as known *a priori* but rather must be fitted from the data.

For the significances obtained to be valid it is necessary that the model predictions for data distributions represent accurately the underlying theory being tested. That is, any errors due to approximations, e.g., in detector modeling or in methods used to relate observable quantities to the fundamental theories, should be negligible for some point in the full parameter space. By including additional parameters in the model one can approach this ideal situation more closely. However, the additional flexibility introduced to parameterize systematic effects results, as it should, in a loss in sensitivity. To the degree that the model is not able to reflect the truth accurately, an additional systematic uncertainty will be present that is not quantified by the statistical method presented here.

To illustrate the use of the profile likelihood ratio, consider an experiment where for each selected event one measures the values of certain kinematic variables, and thus the resulting data can be represented as one or more histograms. Using the method in an unbinned analysis is a straightforward extension.

Suppose for each event in the signal sample one measures a variable x and uses these values to construct a histogram $\mathbf{n} = (n_1, \ldots, n_N)$. The expectation value of n_i can be written

$$E[n_i] = \mu s_i + b_i, \tag{2}$$

where the mean number of entries in the ith bin from signal and background are

$$s_i = s_{\text{tot}} \int_{\text{bin}\,i} f_s(x;\boldsymbol{\theta}_s)\,dx, \tag{3}$$

$$b_i = b_{\text{tot}} \int_{\text{bin}\,i} f_b(x;\boldsymbol{\theta}_b)\,dx. \tag{4}$$

Here the parameter μ determines the strength of the signal process, with $\mu = 0$ corresponding to the background-only hypothesis and $\mu = 1$ being the nominal signal hypothesis. The functions $f_s(x;\boldsymbol{\theta}_s)$ and $f_b(x;\boldsymbol{\theta}_b)$ are the probability density functions (pdfs) of the variable x for signal and background events, and $\boldsymbol{\theta}_s$ and $\boldsymbol{\theta}_b$ represent parameters that characterize the shapes of pdfs. The quantities s_{tot} and

[2]Some authors, e.g., [6], have defined this relation using a two-sided fluctuation of a Gaussian variable, with a 5σ significance corresponding to $p = 5.7 \times 10^{-7}$. We take the one-sided definition above as this gives $Z = 0$ for $p = 0.5$.

$b_{\rm tot}$ are the total mean numbers of signal and background events, and the integrals in (3) and (4) represent the probabilities for an event to be found in bin i. Below we use $\boldsymbol{\theta} = (\boldsymbol{\theta}_s, \boldsymbol{\theta}_b, b_{\rm tot})$ to denote all of the nuisance parameters. The signal normalization $s_{\rm tot}$ is not, however, an adjustable parameter but rather is fixed to the value predicted by the nominal signal model.

In addition to the measured histogram $\mathbf{n}$ one often makes further subsidiary measurements that help constrain the nuisance parameters. For example, one may select a control sample where one expects mainly background events and from them construct a histogram of some chosen kinematic variable. This then gives a set of values $\mathbf{m} = (m_1, \ldots, m_M)$ for the number of entries in each of the M bins. The expectation value of m_i can be written

$$E[m_i] = u_i(\boldsymbol{\theta}), \tag{5}$$

where the u_i are calculable quantities depending on the parameters $\boldsymbol{\theta}$. One often constructs this measurement so as to provide information on the background normalization parameter $b_{\rm tot}$ and also possibly on the signal and background shape parameters.

The likelihood function is the product of Poisson probabilities for all bins:

$$L(\mu, \boldsymbol{\theta}) = \prod_{j=1}^{N} \frac{(\mu s_j + b_j)^{n_j}}{n_j!} e^{-(\mu s_j + b_j)} \prod_{k=1}^{M} \frac{u_k^{m_k}}{m_k!} e^{-u_k}. \tag{6}$$

To test a hypothesized value of μ we consider the *profile likelihood* ratio

$$\lambda(\mu) = \frac{L(\mu, \hat{\hat{\boldsymbol{\theta}}})}{L(\hat{\mu}, \hat{\boldsymbol{\theta}})}. \tag{7}$$

The numerator of this ratio is the *profile likelihood function* (see, e.g., [7]). The quantity $\hat{\hat{\boldsymbol{\theta}}}$ denotes the value of $\boldsymbol{\theta}$ that maximizes L for the specified μ, i.e., it is the conditional maximum-likelihood (ML) estimator of $\boldsymbol{\theta}$ (and thus is a function of μ). The denominator is the maximized (unconditional) likelihood function, i.e., $\hat{\mu}$ and $\hat{\boldsymbol{\theta}}$ are their ML estimators. The presence of the nuisance parameters broadens the profile likelihood as a function of μ relative to what one would have if their values were fixed. This reflects the loss of information about μ due to the systematic uncertainties.

In many analyses, the contribution of the signal process to the mean number of events is assumed to be non-negative. This condition effectively implies that any physical estimator for μ must be non-negative. Even if we regard this to be the case, however, it is convenient to define an effective estimator $\hat{\mu}$ as the value of μ that maximizes the likelihood, even this gives $\hat{\mu} < 0$ (but providing that the Poisson mean values, $\mu s_i + b_i$, remain nonnegative). This will allow us in Sect. 3.1 to model $\hat{\mu}$ as a Gaussian distributed variable, and in this way we can determine the distributions of the test statistics that we consider. Therefore in the following we will always regard $\hat{\mu}$ as an effective estimator which is allowed to take on negative values.

2.1 Test statistic $t_\mu = -2\ln\lambda(\mu)$

From the definition of $\lambda(\mu)$ in (7), one can see that $0 \leq \lambda \leq 1$, with λ near 1 implying good agreement between the data and the hypothesized value of μ. Equivalently it is convenient to use the statistic

$$t_\mu = -2\ln\lambda(\mu) \tag{8}$$

as the basis of a statistical test. Higher values of t_μ thus correspond to increasing incompatibility between the data and μ.

We may define a test of a hypothesized value of μ by using the statistic t_μ directly as measure of discrepancy between the data and the hypothesis, with higher values of t_μ correspond to increasing disagreement. To quantify the level of disagreement we compute the p-value,

$$p_\mu = \int_{t_{\mu,\rm obs}}^{\infty} f(t_\mu|\mu)\, dt_\mu, \tag{9}$$

where $t_{\mu,\rm obs}$ is the value of the statistic t_μ observed from the data and $f(t_\mu|\mu)$ denotes the pdf of t_μ under the assumption of the signal strength μ. Useful approximations for this and other related pdfs are given in Sect. 3.3. The relation between the p-value and the observed t_μ and also with the significance Z are illustrated in Fig. 1.

When using the statistic t_μ, a data set may result in a low p-value in two distinct ways: the estimated signal strength $\hat{\mu}$ may be found greater or less than the hypothesized value μ. As a result, the set of μ values that are rejected because their p-values are found below a specified threshold α may lie to either side of those values not rejected, i.e., one may obtain a two-sided confidence interval for μ.

2.2 Test statistic $\tilde{t}_\mu$ for $\mu \geq 0$

Often one assumes that the presence of a new signal can only increase the mean event rate beyond what is expected from background alone. That is, the signal process necessarily has $\mu \geq 0$, and to take this into account we define an alternative test statistic below called $\tilde{t}_\mu$.

For a model where $\mu \geq 0$, if one finds data such that $\hat{\mu} < 0$, then the best level of agreement between the data and any physical value of μ occurs for $\mu = 0$. We therefore define

$$\tilde{\lambda}(\mu) = \begin{cases} \dfrac{L(\mu, \hat{\hat{\boldsymbol{\theta}}}(\mu))}{L(\hat{\mu}, \hat{\boldsymbol{\theta}})}, & \hat{\mu} \geq 0, \\ \dfrac{L(\mu, \hat{\hat{\boldsymbol{\theta}}}(\mu))}{L(0, \hat{\hat{\boldsymbol{\theta}}}(0))}, & \hat{\mu} < 0. \end{cases} \tag{10}$$

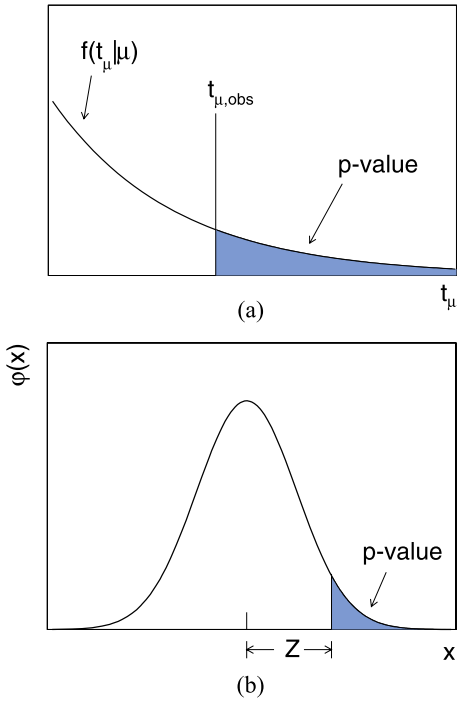

Fig. 1 (**a**) Illustration of the relation between the p-value obtained from an observed value of the test statistic t_μ. (**b**) The standard normal distribution $\varphi(x) = (1/\sqrt{2\pi})\exp(-x^2/2)$ showing the relation between the significance Z and the p-value

Here $\hat{\hat{\boldsymbol{\theta}}}(0)$ and $\hat{\hat{\boldsymbol{\theta}}}(\mu)$ refer to the conditional ML estimators of $\boldsymbol{\theta}$ given a strength parameter of 0 or μ, respectively.

The variable $\tilde{\lambda}(\mu)$ can be used instead of $\lambda(\mu)$ in (8) to obtain the corresponding test statistic, which we denote $\tilde{t}_\mu$. That is,

$$\tilde{t}_\mu = -2\ln\tilde{\lambda}(\mu) = \begin{cases} -2\ln\frac{L(\mu,\hat{\hat{\boldsymbol{\theta}}}(\mu))}{L(0,\hat{\hat{\theta}}(0))}, & \hat{\mu} < 0, \\ -2\ln\frac{L(\mu,\hat{\hat{\boldsymbol{\theta}}}(\mu))}{L(\hat{\mu},\hat{\boldsymbol{\theta}})}, & \hat{\mu} \geq 0. \end{cases} \tag{11}$$

As was done with the statistic t_μ, one can quantify the level of disagreement between the data and the hypothesized value of μ with the p-value, just as in (9). For this one needs the distribution of $\tilde{t}_\mu$, an approximation of which is given in Sect. 3.4.

Also similar to the case of t_μ, values of μ both above and below $\hat{\mu}$ may be excluded by a given data set, i.e., one may obtain either a one-sided or two-sided confidence interval for μ. The statistic defined here as $\tilde{t}_\mu$ is also discussed for the case without nuisance parameters in [8] and including nuisance parameters in [7]. In these references, however, the estimator of the parameter of interest is assumed to remain within the allowed range of the parameter. Here in contrast we treat $\hat{\mu}$ as an effective estimator that can be negative even if the physical model requires $\mu \geq 0$. As shown below in Sect. 3, this allows us to apply large-sample approximations to find the distribution of $\tilde{t}_\mu$.

2.3 Test statistic q_0 for discovery of a positive signal

An important special case of the statistic $\tilde{t}_\mu$ described above is used to test $\mu = 0$ in a class of model where we assume $\mu \geq 0$. Rejecting the $\mu = 0$ hypothesis effectively leads to the discovery of a new signal. For this important case we use the special notation $q_0 = \tilde{t}_0$. Using the definition (11) with $\mu = 0$ one finds

$$q_0 = \begin{cases} -2\ln\lambda(0), & \hat{\mu} \geq 0, \\ 0, & \hat{\mu} < 0, \end{cases} \tag{12}$$

where $\lambda(0)$ is the profile likelihood ratio for $\mu = 0$ as defined in (7).

We may contrast this to the statistic t_0, i.e., (8), used to test $\mu = 0$. In that case one may reject the $\mu = 0$ hypothesis for either an upward or downward fluctuation of the data. This is appropriate if the presence of a new phenomenon could lead to an increase or decrease in the number of events found. In an experiment looking for neutrino oscillations, for example, the signal hypothesis may predict a greater or lower event rate than the no-oscillation hypothesis.

When using q_0, however, we consider the data to show lack of agreement with the background-only hypothesis only if $\hat{\mu} > 0$. That is, a value of $\hat{\mu}$ much below zero may indeed constitute evidence against the background-only model, but this type of discrepancy does not show that the data contain signal events, but rather points to some other systematic error. For the present discussion, however, we assume that the systematic uncertainties are dealt with by the nuisance parameters $\boldsymbol{\theta}$.

If the data fluctuate such that one finds fewer events than even predicted by background processes alone, then $\hat{\mu} < 0$ and one has $q_0 = 0$. As the event yield increases above the expected background, i.e., for increasing $\hat{\mu}$, one finds increasingly large values of q_0, corresponding to an increasing level of incompatibility between the data and the $\mu = 0$ hypothesis.

To quantify the level of disagreement between the data and the hypothesis of $\mu = 0$ using the observed value of q_0 we compute the p-value in the same manner as done with t_μ, namely,

$$p_0 = \int_{q_{0,\mathrm{obs}}}^{\infty} f(q_0|0)\,dq_0. \tag{13}$$

Here $f(q_0|0)$ denotes the pdf of the statistic q_0 under assumption of the background-only ($\mu = 0$) hypothesis. An approximation for this and other related pdfs are given in Sect. 3.5.

2.4 Test statistic q_μ for upper limits

For purposes of establishing an upper limit on the strength parameter μ, we consider two closely related test statistics. First, we may define

$$q_\mu = \begin{cases} -2\ln\lambda(\mu), & \hat{\mu} \le \mu, \\ 0, & \hat{\mu} > \mu, \end{cases} \tag{14}$$

where $\lambda(\mu)$ is the profile likelihood ratio as defined in (7). The reason for setting $q_\mu = 0$ for $\hat{\mu} > \mu$ is that when setting an upper limit, one would not regard data with $\hat{\mu} > \mu$ as representing less compatibility with μ than the data obtained, and therefore this is not taken as part of the rejection region of the test. That is, the upper limit is obtained by testing μ against the alternative hypothesis consisting of lower values of μ. From the definition of the test statistic one sees that higher values of q_μ represent greater incompatibility between the data and the hypothesized value of μ.

One should note that q_0 is not simply a special case of q_μ with $\mu = 0$, but rather has a different definition (see (12) and (14)). That is, q_0 is zero if the data fluctuate downward ($\hat{\mu} < 0$), but q_μ is zero if the data fluctuate upward ($\hat{\mu} > \mu$). With that caveat in mind, we will often refer in the following to q_μ with the idea that this means either q_0 or q_μ as appropriate to the context.

As with the case of discovery, one quantifies the level of agreement between the data and hypothesized μ with p-value. For, e.g., an observed value $q_{\mu,\mathrm{obs}}$, one has

$$p_\mu = \int_{q_{\mu,\mathrm{obs}}}^{\infty} f(q_\mu|\mu)\, dq_\mu, \tag{15}$$

which can be expressed as a significance using (1). Here $f(q_\mu|\mu)$ is the pdf of q_μ assuming the hypothesis μ. In Sect. 3.6 we provide useful approximations for this and other related pdfs.

2.5 Alternative test statistic $\tilde{q}_\mu$ for upper limits

For the case where one considers models for which $\mu \ge 0$, the variable $\tilde{\lambda}(\mu)$ can be used instead of $\lambda(\mu)$ in (14) to obtain the corresponding test statistic, which we denote $\tilde{q}_\mu$. That is,

$$\begin{aligned} \tilde{q}_\mu &= \begin{cases} -2\ln\tilde{\lambda}(\mu), & \hat{\mu} \le \mu \\ 0, & \hat{\mu} > \mu \end{cases} \\ &= \begin{cases} -2\ln\frac{L(\mu,\hat{\hat{\boldsymbol{\theta}}}(\mu))}{L(0,\hat{\hat{\theta}}(0))}, & \hat{\mu} < 0, \\ -2\ln\frac{L(\mu,\hat{\hat{\boldsymbol{\theta}}}(\mu))}{L(\hat{\mu},\hat{\boldsymbol{\theta}})}, & 0 \le \hat{\mu} \le \mu, \\ 0, & \hat{\mu} > \mu. \end{cases} \end{aligned} \tag{16}$$

We give an approximation for the pdf $f(\tilde{q}_\mu|\mu')$ in Sect. 3.7.

In numerical examples we have found that the difference between the tests based on q_μ (14) and $\tilde{q}_\mu$ usually to be negligible, but use of q_μ leads to important simplifications. Furthermore, in the context of the approximation used in Sect. 3, the two statistics are equivalent. That is, assuming the approximations below, q_μ can be expressed as a monotonic function of $\tilde{q}_\mu$ and thus they lead to the same results.

3 Approximate sampling distributions

In order to find the p-value of a hypothesis using (13) or (15) we require the sampling distribution for the test statistic being used. In the case of discovery we are testing the background-only hypothesis ($\mu = 0$) and therefore we need $f(q_0|0)$, where q_0 is defined by (12). When testing a nonzero value of μ for purposes of finding an upper limit we need the distribution $f(q_\mu|\mu)$ where q_μ is defined by (14), or alternatively we require the pdf of the corresponding statistic $\tilde{q}_\mu$ as defined by (16). In this notation the subscript of q refers to the hypothesis being tested, and the second argument in $f(q_\mu|\mu)$ gives the value of μ assumed in the distribution of the data.

We also need the distribution $f(q_\mu|\mu')$ with $\mu \ne \mu'$ to find what significance to expect and how this is distributed if the data correspond to a strength parameter different from the one being tested. For example, it is useful to characterize the sensitivity of a planned experiment by quoting the median significance, assuming data distributed according to a specified signal model, with which one would expect to exclude the background-only hypothesis. For this one would need $f(q_0|\mu')$, usually with $\mu' = 1$. From this one can find the median q_0, and thus the median discovery significance. When considering upper limits, one would usually quote the value of μ for which the median p-value is equal to 0.05, as this gives the median upper limit on μ at 95% confidence level. In this case one would need $f(q_\mu|0)$ (or alternatively $f(\tilde{q}_\mu|0)$).

In Sect. 3.1 we present an approximation for the profile likelihood ratio, valid in the large sample limit. This allows one to obtain approximations for all of the required distributions, which are given in Sects. 3.3 through 3.6. The approximations become exact in the large sample limit and are in fact found to provide accurate results even for fairly small sample sizes. For very small data samples one always has the possibility of using Monte Carlo methods to determine the required distributions.

3.1 Approximate distribution of the profile likelihood ratio

Consider a test of the strength parameter μ, which here can either be zero (for discovery) or nonzero (for an upper limit), and suppose the data are distributed according to

a strength parameter μ'. The desired distribution $f(q_\mu|\mu')$ can be found using a result due to Wald [2], who showed that for the case of a single parameter of interest,

$$-2\ln\lambda(\mu) = \frac{(\mu-\hat{\mu})^2}{\sigma^2} + \mathcal{O}\big(1/\sqrt{N}\big). \tag{17}$$

Here $\hat{\mu}$ follows a Gaussian distribution with a mean μ' and standard deviation σ, and N represents the data sample size. The standard deviation σ of $\hat{\mu}$ is obtained from the covariance matrix of the estimators for all the parameters, $V_{ij} = \mathrm{cov}[\hat{\theta}_i, \hat{\theta}_j]$, where here the θ_i represent both μ as well as the nuisance parameters (e.g., take $\theta_0 = \mu$, so $\sigma^2 = V_{00}$). In the large-sample limit, the bias of ML estimators in general tend to zero, in which case we can write the inverse of the covariance matrix as

$$V_{ij}^{-1} = -E\left[\frac{\partial^2 \ln L}{\partial\theta_i \partial\theta_j}\right], \tag{18}$$

where the expectation value assumes a strength parameter μ'. The approximations presented here are valid to the extent that the $\mathcal{O}(1/\sqrt{N})$ term can be neglected, and the value of σ can be estimated, e.g., using (18). In Sect. 3.2 we present an alternative way to estimate σ which lends itself more directly to determination of the median significance.

If $\hat{\mu}$ is Gaussian distributed and we neglect the $\mathcal{O}(1/\sqrt{N})$ term in (17), then one can show that the statistic $t_\mu = -2\ln\lambda(\mu)$ follows a *noncentral chi-square* distribution for one degree of freedom (see, e.g., [10, 11]),

$$\begin{aligned} f(t_\mu;\Lambda) = {} & \frac{1}{2\sqrt{t_\mu}}\frac{1}{\sqrt{2\pi}} \\ & \times\left[\exp\left(-\frac{1}{2}\big(\sqrt{t_\mu}+\sqrt{\Lambda}\big)^2\right)\right. \\ & \left.+\exp\left(-\frac{1}{2}\big(\sqrt{t_\mu}-\sqrt{\Lambda}\big)^2\right)\right], \end{aligned} \tag{19}$$

where the noncentrality parameter Λ is

$$\Lambda = \frac{(\mu-\mu')^2}{\sigma^2}. \tag{20}$$

For the special case $\mu' = \mu$ one has $\Lambda = 0$ and $-2\ln\lambda(\mu)$ approaches a chi-square distribution for one degree of freedom, a result shown earlier by Wilks [1].

The results of Wilks and Wald generalize to more than one parameter of interest. If the parameters of interest can be explicitly identified with a subset of the parameters $\boldsymbol{\theta}_r = (\theta_1, \ldots, \theta_r)$, then the distribution of $-2\ln\lambda(\boldsymbol{\theta}_r)$ follows a noncentral chi-square distribution for r-degrees of freedom with noncentrality parameter

$$\Lambda_r = \sum_{i,j=1}^{r} (\theta_i - \theta_i')\,\tilde{V}_{ij}^{-1}\,(\theta_j - \theta_j'), \tag{21}$$

where $\tilde{V}_{ij}^{-1}$ is the inverse of the submatrix one obtains from restricting the full covariance matrix to the parameters of interest. The full covariance matrix is given from inverting (18), and we show an efficient way to calculate it in Sect. 3.2.

3.2 The Asimov data set and the variance of $\hat{\mu}$

Some of the formulae given require the standard deviation σ of $\hat{\mu}$, which is assumed to follow a Gaussian distribution with a mean of μ'. Below we show two ways of estimating σ, both of which are closely related to a special, artificial data set that we call the "Asimov data set".

We define the Asimov data set such that when one uses it to evaluate the estimators for all parameters, one obtains the true parameter values. Consider the likelihood function for the generic analysis given by (6). To simplify the notation in this section we define

$$\nu_i = \mu' s_i + b_i. \tag{22}$$

Further let $\theta_0 = \mu$ represent the strength parameter, so that here θ_i can stand for any of the parameters. The ML estimators for the parameters can be found by setting the derivatives of $\ln L$ with respect to all of the parameters equal to zero:

$$\frac{\partial \ln L}{\partial\theta_j} = \sum_{i=1}^{N}\left(\frac{n_i}{\nu_i} - 1\right)\frac{\partial\nu_i}{\partial\theta_j} + \sum_{i=1}^{M}\left(\frac{m_i}{u_i} - 1\right)\frac{\partial u_i}{\partial\theta_j} = 0. \tag{23}$$

This condition holds if the Asimov data, $n_{i,\mathrm{A}}$ and $m_{i,\mathrm{A}}$, are equal to their expectation values:

$$n_{i,\mathrm{A}} = E[n_i] = \nu_i = \mu' s_i(\boldsymbol{\theta}) + b_i(\boldsymbol{\theta}), \tag{24}$$

$$m_{i,\mathrm{A}} = E[m_i] = u_i(\boldsymbol{\theta}). \tag{25}$$

Here the parameter values represent those implied by the assumed distribution of the data. In practice, these are the values that would be estimated from the Monte Carlo model using a very large data sample.

We have not proved that the Asimov data set as defined above always exists, and there may be pathological cases where it does not. This issue is not expected to be relevant in practice, and indeed for an analysis of the binned type outlined above, the Asimov data set can always be found using the expectation values as in (24) and (25). Note also that an unbinned likelihood can always be interpreted as a limiting case of a binned likelihood when the bin size goes to zero, so (24) and (25) can be applied to such cases as well. Furthermore the Asimov data set is not strictly unique, and if this is the case then any one may be used.

We can use the Asimov data set to evaluate the "Asimov likelihood" L_A and the corresponding profile likelihood ra-

tio $\lambda_{\rm A}$. The use of non-integer values for the data is not a problem as the factorial terms in the Poisson likelihood represent constants that cancel when forming the likelihood ratio, and thus can be dropped. One finds

$$\lambda_{\rm A}(\mu) = \frac{L_{\rm A}(\mu, \hat{\hat{\boldsymbol{\theta}}})}{L_{\rm A}(\hat{\mu}, \hat{\boldsymbol{\theta}})} = \frac{L_{\rm A}(\mu, \hat{\hat{\boldsymbol{\theta}}})}{L_A(\mu', \boldsymbol{\theta})}, \tag{26}$$

where the final equality above exploits the fact that the estimators for the parameters are equal to their hypothesized values when the likelihood is evaluated with the Asimov data set.

A standard way to find σ is by estimating the matrix of second derivatives of the log-likelihood function (cf. (18)) to obtain the inverse covariance matrix V^{-1}, inverting to find V, and then extracting the element V_{00} corresponding to the variance of $\hat{\mu}$. The second derivative of $\ln L$ is

$$\frac{\partial^2 \ln L}{\partial\theta_j \partial\theta_k} = \sum_{i=1}^{N}\left[\left(\frac{n_i}{\nu_i} - 1\right)\frac{\partial^2\nu_i}{\partial\theta_j\partial\theta_k} - \frac{\partial\nu_i}{\partial\theta_j}\frac{\partial\nu_i}{\partial\theta_k}\frac{n_i}{\nu_i^2}\right] + \sum_{i=1}^{M}\left[\left(\frac{m_i}{u_i} - 1\right)\frac{\partial^2 u_i}{\partial\theta_j\partial\theta_k} - \frac{\partial u_i}{\partial\theta_j}\frac{\partial u_i}{\partial\theta_k}\frac{m_i}{u_i^2}\right]. \tag{27}$$

From (27) one sees that the second derivative of $\ln L$ is linear in the data values n_i and m_i. Thus its expectation value is found simply by evaluating with the expectation values of the data, which is the same as the Asimov data. One can therefore obtain the inverse covariance matrix from

$$V_{jk}^{-1} = -E\left[\frac{\partial^2 \ln L}{\partial\theta_j\partial\theta_k}\right] = -\frac{\partial^2 \ln L_{\rm A}}{\partial\theta_j\partial\theta_k} = \sum_{i=1}^{N}\frac{\partial\nu_i}{\partial\theta_j}\frac{\partial\nu_i}{\partial\theta_k}\frac{1}{\nu_i} + \sum_{i=1}^{M}\frac{\partial u_i}{\partial\theta_j}\frac{\partial u_i}{\partial\theta_k}\frac{1}{u_i}. \tag{28}$$

In practice one could, for example, evaluate the derivatives of $\ln L_{\rm A}$ numerically, use this to find the inverse covariance matrix, and then invert and extract the variance of $\hat{\mu}$. One can see directly from (28) that this variance depends on the parameter values assumed for the Asimov data set, in particular on the assumed strength parameter μ', which enters via (22).

Another method for estimating σ (denoted $\sigma_{\rm A}$ in this section to distinguish it from the approach above based on the second derivatives of $\ln L$) is to find the value that is necessary to recover the known properties of $-\lambda_{\rm A}(\mu)$. Because the Asimov data set corresponding to a strength μ' gives $\hat{\mu} = \mu'$, from (17) one finds

$$-2\ln\lambda_{\rm A}(\mu) \approx \frac{(\mu - \mu')^2}{\sigma^2} = \Lambda. \tag{29}$$

That is, from the Asimov data set one obtains an estimate of the noncentrality parameter Λ that characterizes the distribution $f(q_\mu|\mu')$. Equivalently, one can use (29) to obtain the variance σ^2 which characterizes the distribution of $\hat{\mu}$, namely,

$$\sigma_{\rm A}^2 = \frac{(\mu - \mu')^2}{q_{\mu,{\rm A}}}, \tag{30}$$

where $q_{\mu,{\rm A}} = -2\ln\lambda_{\rm A}(\mu)$. For the important case where one wants to find the median exclusion significance for the hypothesis μ assuming that there is no signal, then one has $\mu' = 0$ and therefore

$$\sigma_{\rm A}^2 = \frac{\mu^2}{q_{\mu,{\rm A}}}, \tag{31}$$

and for the modified statistic $\tilde{q}_\mu$ the analogous relation holds. For the case of discovery where one tests $\mu = 0$ one has

$$\sigma_{\rm A}^2 = \frac{\mu'^2}{q_{0,{\rm A}}}. \tag{32}$$

The two methods for obtaining σ and Λ—from the Fisher information matrix or from $q_{\mu,{\rm A}}$—are not identical, but were found to provide similar results in examples of practical interest. In several cases that we considered, the distribution based on $\sigma_{\rm A}$ provided a better approximation to the true sampling distribution than the standard approach based on the Fisher information matrix, leading to the conjecture that it may effectively incorporate some higher-order terms in (17).

This can be understood qualitatively by noting that under assumption of the Wald approximation, the test statistics q_0, q_μ and $\tilde{q}_\mu$ are monotonically related to $\hat{\mu}$, and therefore their median values can be found directly by using the median of $\hat{\mu}$, which is μ'. But monotonicity is a weaker condition than the full Wald approximation. That is, even if higher-order terms are present in (17), they will not alter the distribution's median as long as they do not break the monotonicity of the relation between the test statistic and $\hat{\mu}$. If one uses $\sigma_{\rm A}$ one obtains distributions with medians given by the corresponding Asimov values, $q_{0,{\rm A}}$ or $q_{\mu,{\rm A}}$, and these values will be correct to the extent that monotonicity holds.

3.3 Distribution of t_μ

Consider first using the statistic $t_\mu = -2\ln\lambda(\mu)$ of Sect. 2.1 as the basis of the statistical test of a hypothesized value of μ. This could be a test of $\mu = 0$ for purposes of establishing existence of a signal process, or non-zero values of μ for purposes of obtaining a confidence interval. To find

the p-value p_μ, we require the pdf $f(t_\mu|\mu)$, and to find the median p-value assuming a different strength parameter we will need $f(t_\mu|\mu')$.

The pdf $f(t_\mu|\mu')$ is given by (19), namely,

$$f(t_\mu|\mu') = \frac{1}{2\sqrt{t_\mu}}\frac{1}{\sqrt{2\pi}} \times \left[\exp\left(-\frac{1}{2}\left(\sqrt{t_\mu}+\frac{\mu-\mu'}{\sigma}\right)^2\right) + \exp\left(-\frac{1}{2}\left(\sqrt{t_\mu}-\frac{\mu-\mu'}{\sigma}\right)^2\right)\right]. \tag{33}$$

The special case $\mu=\mu'$ is simply a chi-square distribution for one degree of freedom:

$$f(t_\mu|\mu) = \frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{t_\mu}}e^{-t_\mu/2}. \tag{34}$$

The cumulative distribution of t_μ assuming μ' is

$$F(t_\mu|\mu') = \Phi\left(\sqrt{t_\mu}+\frac{\mu-\mu'}{\sigma}\right)+\Phi\left(\sqrt{t_\mu}-\frac{\mu-\mu'}{\sigma}\right)-1, \tag{35}$$

where Φ is the cumulative distribution of the standard (zero mean, unit variance) Gaussian. The special case $\mu=\mu'$ is therefore

$$F(t_\mu|\mu) = 2\Phi\left(\sqrt{t_\mu}\right)-1. \tag{36}$$

The p-value of a hypothesized value of μ for an observed value t_μ is therefore

$$p_\mu = 1-F(t_\mu|\mu) = 2\left(1-\Phi\left(\sqrt{t_\mu}\right)\right), \tag{37}$$

and the corresponding significance is

$$Z_\mu = \Phi^{-1}(1-p_\mu) = \Phi^{-1}\left(2\Phi\left(\sqrt{t_\mu}\right)-1\right). \tag{38}$$

If the p-value is found below a specified threshold α (often one takes $\alpha=0.05$), then the value of μ is said to be excluded at a confidence level (CL) of $1-\alpha$. The set of points not excluded form a confidence interval with CL $=1-\alpha$. Here the endpoints of the interval can be obtained simply by setting $p_\mu=\alpha$ and solving for μ. Assuming the Wald approximation (17) and using (37) one finds

$$\mu_{\text{up/lo}} = \hat{\mu} \pm \sigma\Phi^{-1}(1-\alpha/2). \tag{39}$$

One subtlety with this formula is that σ itself depends at some level on μ. In practice to find the upper and lower limits one can simply solve numerically to find those values of μ that satisfy $p_\mu=\alpha$.

3.4 Distribution of $\tilde{t}_\mu$

Assuming the Wald approximation, the statistic $\tilde{t}_\mu$ as defined by (11) can be written

$$\tilde{t}_\mu = \begin{cases} \frac{\mu^2}{\sigma^2}-\frac{2\mu\hat{\mu}}{\sigma^2}, & \hat{\mu}<0, \\ \frac{(\mu-\hat{\mu})^2}{\sigma^2}, & \hat{\mu}\geq 0. \end{cases} \tag{40}$$

From this the pdf $f(\tilde{t}_\mu|\mu')$ is found to be

$$f(\tilde{t}_\mu|\mu') = \frac{1}{2}\frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{\tilde{t}_\mu}}\exp\left[-\frac{1}{2}\left(\sqrt{\tilde{t}_\mu}+\frac{\mu-\mu'}{\sigma}\right)^2\right] + \begin{cases} \frac{1}{2}\frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{\tilde{t}_\mu}}\exp[-\frac{1}{2}(\sqrt{\tilde{t}_\mu}-\frac{\mu-\mu'}{\sigma})^2], \\ \quad \tilde{t}_\mu \leq \mu^2/\sigma^2, \\ \frac{1}{\sqrt{2\pi}(2\mu/\sigma)}\exp[-\frac{1}{2}\frac{(\tilde{t}_\mu-\frac{\mu^2-2\mu\mu'}{\sigma^2})^2}{(2\mu/\sigma)^2}], \\ \quad \tilde{t}_\mu > \mu^2/\sigma^2. \end{cases} \tag{41}$$

The special case $\mu=\mu'$ is therefore

$$f(\tilde{t}_\mu|\mu') = \begin{cases} \frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{\tilde{t}_\mu}}e^{-\tilde{t}_\mu/2}, & \tilde{t}_\mu \leq \mu^2/\sigma^2, \\ \frac{1}{2}\frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{\tilde{t}_\mu}}e^{-\tilde{t}_\mu/2} & \\ \quad + \frac{1}{\sqrt{2\pi}(2\mu/\sigma)}\exp[-\frac{1}{2}\frac{(\tilde{t}_\mu+\mu^2/\sigma^2)^2}{(2\mu/\sigma)^2}], & \\ \tilde{t}_\mu > \mu^2/\sigma^2. & \end{cases} \tag{42}$$

The corresponding cumulative distribution is

$$F(\tilde{t}_\mu|\mu') = \Phi\left(\sqrt{\tilde{t}_\mu}+\frac{\mu-\mu'}{\sigma}\right) + \begin{cases} \Phi(\sqrt{\tilde{t}_\mu}-\frac{\mu-\mu'}{\sigma})-1, & \tilde{t}_\mu \leq \mu^2/\sigma^2, \\ \Phi(\frac{\tilde{t}_\mu-(\mu^2-2\mu\mu')/\sigma^2}{2\mu/\sigma})-1, & \tilde{t}_\mu > \mu^2/\sigma^2. \end{cases} \tag{43}$$

For $\mu=\mu'$ this is

$$F(\tilde{t}_\mu|\mu) = \begin{cases} 2\Phi(\sqrt{\tilde{t}_\mu})-1, & \\ \quad \tilde{t}_\mu \leq \mu^2/\sigma^2, & \\ \Phi(\sqrt{\tilde{t}_\mu})+\Phi(\frac{\tilde{t}_\mu+\mu^2/\sigma^2}{2\mu/\sigma})-1, & \\ \quad \tilde{t}_\mu > \mu^2/\sigma^2. & \end{cases} \tag{44}$$

The p-value of the hypothesized μ is given by one minus the cumulative distribution, under assumption of the parameter μ,

$$p_\mu = 1-F(\tilde{t}_\mu|\mu). \tag{45}$$

The corresponding significance is $Z_\mu=\Phi^{-1}(1-p_\mu)$.

A confidence interval for μ at confidence level CL $= 1-\alpha$ can be constructed from the set μ values for which the p-value is not less than α. To find the endpoints of this interval, one can set p_μ from (45) equal to α and solve for μ. In general this must be done numerically. In the large sample limit, i.e., assuming the validity of the asymptotic approximations, these intervals correspond to the limits of Feldman and Cousins [8] for the case where physical range of the parameter μ is $\mu \geq 0$.

3.5 Distribution of q_0 (discovery)

Assuming the validity of the approximation (17), one has $-2\ln\lambda(0) = \hat{\mu}^2/\sigma^2$. From the definition (12) of q_0, we therefore have

$$q_0 = \begin{cases} \hat{\mu}^2/\sigma^2, & \hat{\mu} \geq 0, \\ 0, & \hat{\mu} < 0, \end{cases} \tag{46}$$

where $\hat{\mu}$ follows a Gaussian distribution with mean μ' and standard deviation σ. From this one can show that the pdf of q_0 has the form

$$f(q_0|\mu') = \left(1 - \Phi\left(\frac{\mu'}{\sigma}\right)\right)\delta(q_0) + \frac{1}{2}\frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{q_0}}\exp\left[-\frac{1}{2}\left(\sqrt{q_0} - \frac{\mu'}{\sigma}\right)^2\right]. \tag{47}$$

For the special case of $\mu' = 0$, this reduces to

$$f(q_0|0) = \frac{1}{2}\delta(q_0) + \frac{1}{2}\frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{q_0}}e^{-q_0/2}. \tag{48}$$

That is, one finds a mixture of a delta function at zero and a chi-square distribution for one degree of freedom, with each term having a weight of 1/2. This distribution is also found in [9]. In the following we will refer to this mixture as a half chi-square distribution or $\frac{1}{2}\chi^2_1$.

From (47) the corresponding cumulative distribution is found to be

$$F(q_0|\mu') = \Phi\left(\sqrt{q_0} - \frac{\mu'}{\sigma}\right). \tag{49}$$

The important special case $\mu' = 0$ is therefore simply

$$F(q_0|0) = \Phi\left(\sqrt{q_0}\right). \tag{50}$$

The p-value of the $\mu = 0$ hypothesis (see (13)) is

$$p_0 = 1 - F(q_0|0), \tag{51}$$

and therefore using (1) for the significance one obtains the simple formula

$$Z_0 = \Phi^{-1}(1 - p_0) = \sqrt{q_0}. \tag{52}$$

3.6 Distribution of q_μ (upper limits)

Assuming the validity of the Wald approximation, we can write the test statistic used for upper limits, (14) as

$$q_\mu = \begin{cases} \frac{(\mu-\hat{\mu})^2}{\sigma^2}, & \hat{\mu} < \mu, \\ 0, & \hat{\mu} > \mu, \end{cases} \tag{53}$$

where $\hat{\mu}$ as before follows a Gaussian centred about μ' with a standard deviation σ.

The pdf $f(q_\mu|\mu')$ is found to be

$$f(q_\mu|\mu') = \Phi\left(\frac{\mu'-\mu}{\sigma}\right)\delta(q_\mu) + \frac{1}{2}\frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{q_\mu}}\exp\left[-\frac{1}{2}\left(\sqrt{q_\mu} - \frac{\mu-\mu'}{\sigma}\right)^2\right], \tag{54}$$

so that the special case $\mu = \mu'$ is a half-chi-square distribution:

$$f(q_\mu|\mu) = \frac{1}{2}\delta(q_\mu) + \frac{1}{2}\frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{q_\mu}}e^{-q_\mu/2}. \tag{55}$$

The cumulative distribution is

$$F(q_\mu|\mu') = \Phi\left(\sqrt{q_\mu} - \frac{\mu-\mu'}{\sigma}\right), \tag{56}$$

and the corresponding special case $\mu' = \mu$ is thus the same as what was found for q_0, namely,

$$F(q_\mu|\mu) = \Phi\left(\sqrt{q_\mu}\right). \tag{57}$$

The p-value of the hypothesized μ is

$$p_\mu = 1 - F(q_\mu|\mu) = 1 - \Phi\left(\sqrt{q_\mu}\right) \tag{58}$$

and therefore the corresponding significance is

$$Z_\mu = \Phi^{-1}(1 - p_\mu) = \sqrt{q_\mu}. \tag{59}$$

As with the statistic t_μ above, if the p-value is found below a specified threshold α (often one takes $\alpha = 0.05$), then the value of μ is said to be excluded at a confidence level (CL) of $1-\alpha$. The upper limit on μ is the largest μ with $p_\mu \leq \alpha$. Here this can be obtained simply by setting $p_\mu = \alpha$ and solving for μ. Using (53) and (58) one finds

$$\mu_{\text{up}} = \hat{\mu} + \sigma\Phi^{-1}(1-\alpha). \tag{60}$$

For example, $\alpha = 0.05$ gives $\Phi^{-1}(1-\alpha) = 1.64$. Also as noted above, σ depends in general on the hypothesized μ. Thus in practice one may find the upper limit numerically as the value of μ for which $p_\mu = \alpha$.

3.7 Distribution of $\tilde{q}_\mu$ (upper limits)

Using the alternative statistic $\tilde{q}_\mu$ defined by (16) and assuming the Wald approximation we find

$$\tilde{q}_\mu = \begin{cases} \frac{\mu^2}{\sigma^2} - \frac{2\mu\hat{\mu}}{\sigma^2}, & \hat{\mu} < 0, \\ \frac{(\mu-\hat{\mu})^2}{\sigma^2}, & 0 \le \hat{\mu} \le \mu, \\ 0, & \hat{\mu} > \mu. \end{cases} \tag{61}$$

The pdf $f(\tilde{q}_\mu|\mu')$ is found to be

$$f(\tilde{q}_\mu|\mu') = \Phi\left(\frac{\mu'-\mu}{\sigma}\right)\delta(\tilde{q}_\mu) + \begin{cases} \frac{1}{2}\frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{\tilde{q}_\mu}}\exp[-\frac{1}{2}(\sqrt{\tilde{q}_\mu} - \frac{\mu-\mu'}{\sigma})^2], \\ \quad 0 < \tilde{q}_\mu \le \mu^2/\sigma^2, \\ \frac{1}{\sqrt{2\pi}(2\mu/\sigma)}\exp[-\frac{1}{2}\frac{(\tilde{q}_\mu-(\mu^2-2\mu\mu')/\sigma^2)^2}{(2\mu/\sigma)^2}], \\ \quad \tilde{q}_\mu > \mu^2/\sigma^2. \end{cases} \tag{62}$$

The special case $\mu = \mu'$ is therefore

$$f(\tilde{q}_\mu|\mu) = \frac{1}{2}\delta(\tilde{q}_\mu) + \begin{cases} \frac{1}{2}\frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{\tilde{q}_\mu}}e^{-\tilde{q}_\mu/2}, \\ \quad 0 < \tilde{q}_\mu \le \mu^2/\sigma^2, \\ \frac{1}{\sqrt{2\pi}(2\mu/\sigma)}\exp[-\frac{1}{2}\frac{(\tilde{q}_\mu+\mu^2/\sigma^2)^2}{(2\mu/\sigma)^2}], \\ \quad \tilde{q}_\mu > \mu^2/\sigma^2. \end{cases} \tag{63}$$

The corresponding cumulative distribution is

$$F(\tilde{q}_\mu|\mu') = \begin{cases} \Phi(\sqrt{\tilde{q}_\mu} - \frac{\mu-\mu'}{\sigma}), & 0 < \tilde{q}_\mu \le \mu^2/\sigma^2, \\ \Phi(\frac{\tilde{q}_\mu-(\mu^2-2\mu\mu')/\sigma^2}{2\mu/\sigma}), & \tilde{q}_\mu > \mu^2/\sigma^2. \end{cases} \tag{64}$$

The special case $\mu = \mu'$ is

$$F(\tilde{q}_\mu|\mu) = \begin{cases} \Phi(\sqrt{\tilde{q}_\mu}), & 0 < \tilde{q}_\mu \le \mu^2/\sigma^2, \\ \Phi(\frac{\tilde{q}_\mu+\mu^2/\sigma^2}{2\mu/\sigma}), & \tilde{q}_\mu > \mu^2/\sigma^2. \end{cases} \tag{65}$$

The p-value of the hypothesized μ is as before given by one minus the cumulative distribution,

$$p_\mu = 1 - F(\tilde{q}_\mu|\mu), \tag{66}$$

and therefore the corresponding significance is

$$Z_\mu = \begin{cases} \sqrt{\tilde{q}_\mu}, & 0 < \tilde{q}_\mu \le \mu^2/\sigma^2, \\ \frac{\tilde{q}_\mu+\mu^2/\sigma^2}{2\mu/\sigma}, & \tilde{q}_\mu > \mu^2/\sigma^2. \end{cases} \tag{67}$$

As when using q_μ, the upper limit on μ at confidence level $1-\alpha$ is found by setting $p_\mu = \alpha$ and solving for μ, which reduces to the same result as found when using q_μ, namely,

$$\mu_{\rm up} = \hat{\mu} + \sigma\Phi^{-1}(1-\alpha). \tag{68}$$

That is, to the extent that the Wald approximation holds, the two statistics q_μ and $\tilde{q}_\mu$ lead to identical upper limits.

3.8 Distribution of $-2\ln(L_{s+b}/L_b)$

Many analyses carried out at the Tevatron Collider (e.g., [12]) involving searches for a new signal process have been based on the statistic

$$q = -2\ln\frac{L_{s+b}}{L_b}, \tag{69}$$

where L_{s+b} is the likelihood of the nominal signal model and L_b is that of the background-only hypothesis. That is, the $s+b$ corresponds to having the strength parameter $\mu = 1$ and L_b refers to $\mu = 0$. The statistic q can therefore be written

$$q = -2\ln\frac{L(\mu=1,\hat{\hat{\boldsymbol{\theta}}}(1))}{L(\mu=0,\hat{\hat{\boldsymbol{\theta}}}(0))} = -2\ln\lambda(1) + 2\ln\lambda(0). \tag{70}$$

Assuming the validity of the Wald approximation (17), q is given by

$$q = \frac{(\hat{\mu}-1)^2}{\sigma^2} - \frac{\hat{\mu}^2}{\sigma^2} = \frac{1-2\hat{\mu}}{\sigma^2}, \tag{71}$$

where as previously σ^2 is the variance of $\hat{\mu}$. As $\hat{\mu}$ follows a Gaussian distribution, the distribution of q is also seen to be Gaussian, with a mean value of

$$E[q] = \frac{1-2\mu}{\sigma^2} \tag{72}$$

and a variance of

$$V[q] = \frac{4}{\sigma^2}. \tag{73}$$

That is, the standard deviation of q is $\sigma_q = 2/\sigma$, where the standard deviation of $\hat{\mu}$, σ, can be estimated, e.g., using the second derivatives of the log-likelihood function as described in Sect. 3.1 or with the methods discussed in Sect. 3.2. Recall that in general σ depends on the hypothesized value of μ; here we will refer to these as σ_b and σ_{s+b} for the $\mu = 0$ and $\mu = 1$ hypotheses, respectively.

From (72) one sees that for the $s+b$ hypothesis ($\mu = 1$) the values of q tend to be lower, and for the b hypothesis

($\mu = 0$) they are higher. Therefore we can find the p-values for the two hypothesis from

$$p_{s+b} = \int_{q_{\rm obs}}^{\infty} f(q|s+b)\,dq = 1 - \Phi\left(\frac{q_{\rm obs} + 1/\sigma_{s+b}}{2/\sigma_{s+b}}\right), \tag{74}$$

$$p_b = \int_{-\infty}^{q_{\rm obs}} f(q|b)\,dq = \Phi\left(\frac{q_{\rm obs} - 1/\sigma_b}{2/\sigma_b}\right), \tag{75}$$

where we have used (72) and (73) for the mean and variance of q under the b and $s+b$ hypotheses.

The p-values from (74) and (75) incorporate the effects of systematic uncertainties to the extent that these are connected to the nuisance parameters $\boldsymbol{\theta}$. In analyses done at the Tevatron such as in [12], these effects are incorporated into the distribution of q in a different but largely equivalent way. There, usually one treats the control measurements that constrain the nuisance parameters as fixed, and to determine the distribution of q one only generates the main search measurement (i.e., what corresponds in our generic analysis to the histogram $\mathbf{n}$). The effects of the systematic uncertainties are taken into account by using the control measurements as the basis of a Bayesian prior density $\pi(\boldsymbol{\theta})$, and the distribution of q is computed under assumption of the Bayesian model average

$$f(q) = \int f(q|\boldsymbol{\theta})\pi(\boldsymbol{\theta})\,d\boldsymbol{\theta}. \tag{76}$$

The prior pdf $\pi(\boldsymbol{\theta})$ used in (76) would be obtained from some measurements characterized by a likelihood function $L_{\boldsymbol{\theta}}(\boldsymbol{\theta})$, and then used to find the prior $\pi(\boldsymbol{\theta})$ using Bayes' theorem,

$$\pi(\boldsymbol{\theta}) \propto L_{\boldsymbol{\theta}}(\boldsymbol{\theta})\pi_0(\boldsymbol{\theta}). \tag{77}$$

Here $\pi_0(\boldsymbol{\theta})$ is the initial prior for $\boldsymbol{\theta}$ that reflected one's knowledge before carrying out the control measurements. In many cases this is simply take as a constant, in which case $\pi(\boldsymbol{\theta})$ is simply proportional to $L_{\boldsymbol{\theta}}(\boldsymbol{\theta})$.

In the approach of this paper, however, all measurements are regarded as part of the data, including control measurements that constrain nuisance parameters. That is, here to generate a data set by MC means, for a given assumed point in the model's parameter space, one simulates both the control measurements and the main measurement. Although this is done for a specific value of $\boldsymbol{\theta}$, in the asymptotic limit the distributions required for computing the p-values (74) and (75) are only weakly dependent on $\boldsymbol{\theta}$ to the extent that this can affect the standard deviation σ_q. By contrast, in the Tevatron approach one generates only the main measurement with data distributed according to the averaged model (76). In the case where the nuisance parameters are constrained by Gaussian distributed estimates and the initial prior $\pi_0(\boldsymbol{\theta})$ is taken to be constant, the two methods are essentially equivalent.

Assuming the Wald approximation holds, the statistic q as well as q_0 from (12), q_μ from (14) and $\tilde{q}_\mu$ from (16) are all monotonic functions of $\hat{\mu}$, and therefore all are equivalent to $\hat{\mu}$ in terms of yielding the same statistical test. If there are no nuisance parameters, then the Neyman–Pearson lemma (see, e.g., [7]) states that the likelihood ratio L_{s+b}/L_b (or equivalently q) is an optimal test statistic in the sense that it gives the maximum power for a test of the background-only hypothesis with respect to the alternative of signal plus background (and vice versa). But if the Wald approximation holds, then q_0 and q_μ lead to equivalent tests and are therefore also optimal in the Neyman–Pearson sense. If the nuisance parameters are well constrained by control measurements, then one expects this equivalence to remain approximately true.

Finally, note that in many analyses carried out at the Tevatron, hypothesized signal models are excluded based not on whether the p-value p_{s+b} from (74) is less than a given threshold α, but rather the ratio is compared to α. We do not consider this final step here; it is discussed in, e.g., [14, 15].

4 Experimental sensitivity

To characterize the sensitivity of an experiment, one is interested not in the significance obtained from a single data set, but rather in the expected (more precisely, median) significance with which one would be able to reject different values of μ. Specifically, for the case of discovery one would like to know the median, under the assumption of the nominal signal model ($\mu = 1$), with which one would reject the background-only ($\mu = 0$) hypothesis. And for the case of setting exclusion limits the sensitivity is characterized by the median significance, assuming data generated using the $\mu = 0$ hypothesis, with which one rejects a nonzero value of μ (usually $\mu = 1$ is of greatest interest).

The sensitivity of an experiment is illustrated in Fig. 2, which shows the pdf for q_μ assuming both a strength parameter μ and also assuming a different value μ'. The distribution $f(q_\mu|\mu')$ is shifted to higher value of q_μ, corresponding on average to lower p-values. The sensitivity of an experiment can be characterized by giving the p-value corresponding to the median q_μ assuming the alternative value μ'. As the p-value is a monotonic function of q_μ, this is equal to the median p-value assuming μ'.

In the rest of this section we describe the ingredients needed to determine the experimental sensitivity (median discovery or exclusion significance). In Sect. 3.2 we introduced the Asimov data set, in which all statistical fluctuations are suppressed. This will lead directly to estimates of the experimental sensitivity (Sect. 4.1) as well as providing an alternative estimate of the standard deviation σ of the estimator $\hat{\mu}$. In Sect. 4.2 we indicate how the procedure can be

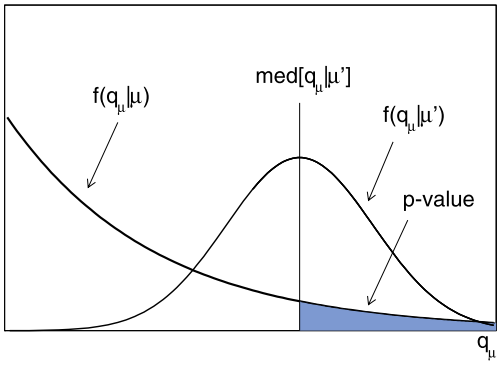

Fig. 2 Illustration of the p-value corresponding to the median of q_μ assuming a strength parameter μ' (see text)

extended to the case where several search channels are combined, and in Sect. 4.3 we describe how to give statistical error bands for the sensitivity.

4.1 The median significance from Asimov values of the test statistic

By using the Asimov data set one can easily obtain the median values of q_0, q_μ and $\tilde{q}_\mu$, and these lead to simple expressions for the corresponding median significance. From (52), (59) and (67) one sees that the significance Z is a monotonic function of q, and therefore the median Z is simply given by the corresponding function of the median of q, which is approximated by its Asimov value. For discovery using q_0 one wants the median discovery significance assuming a strength parameter μ' and for upper limits one is particularly interested in the median exclusion significance assuming $\mu' = 0$, med$[Z_\mu|0]$. For these one obtains

$$\text{med}[Z_0|\mu'] = \sqrt{q_{0,\text{A}}}, \tag{78}$$

$$\text{med}[Z_\mu|0] = \sqrt{q_{\mu,\text{A}}}. \tag{79}$$

When using $\tilde{q}_\mu$ for establishing upper limits, the general expression for the exclusion significance Z_μ is somewhat more complicated depending on μ', but is in any case found by substituting the appropriate values of $\tilde{q}_{\mu,\text{A}}$ and σ_A into (67). For the usual case where one wants the median significance for μ assuming data distributed according to the background-only hypothesis ($\mu' = 0$), (67) reduces in fact to a relation of the same form as (59), and therefore one finds

$$\text{med}[Z_\mu|0] = \sqrt{\tilde{q}_{\mu,\text{A}}}. \tag{80}$$

4.2 Combining multiple channels

In many analyses, there can be several search channels which need to be combined. For each channel i there is a likelihood function $L_i(\mu, \boldsymbol{\theta}_i)$, where $\boldsymbol{\theta}_i$ represents the set of nuisance parameters for the ith channel, some of which may be common between channels. Here the strength parameter μ is assumed to be the same for all channels. If the channels are statistically independent, as can usually be arranged, the full likelihood function is given by the product over all of the channels,

$$L(\mu, \boldsymbol{\theta}) = \prod_i L_i(\mu, \boldsymbol{\theta}_i), \tag{81}$$

where $\boldsymbol{\theta}$ represents the complete set of all nuisance parameters. The profile likelihood ratio $\lambda(\mu)$ is therefore

$$\lambda(\mu) = \frac{\prod_i L_i(\mu, \hat{\hat{\boldsymbol{\theta}}}_i)}{\prod_i L_i(\hat{\mu}, \hat{\boldsymbol{\theta}}_i)}. \tag{82}$$

Because the Asimov data contain no statistical fluctuations, one has $\hat{\mu} = \mu'$ for all channels. Furthermore any common components of $\boldsymbol{\theta}_i$ are the same for all channels. Therefore when using the Asimov data corresponding to a strength parameter μ' one finds

$$\lambda_\text{A}(\mu) = \frac{\prod_i L_i(\mu, \hat{\hat{\boldsymbol{\theta}}})}{\prod_i L_i(\mu', \boldsymbol{\theta})} = \prod_i \lambda_{\text{A},i}(\mu), \tag{83}$$

where $\lambda_{\text{A},i}(\mu)$ is the profile likelihood ratio for the ith channel alone.

Because of this, it is possible to determine the values of the profile likelihood ratio entering into (83) separately for each channel, which simplifies greatly the task of estimating the median significance that would result from the full combination. It should be emphasized, however, that to find the discovery significance or exclusion limits determined from real data, one needs to construct the full likelihood function containing a single parameter μ, and this must be used in a global fit to find the profile likelihood ratio.

4.3 Expected statistical variation (error bands)

By using the Asimov data set we can find the median, assuming some strength parameter μ' of the significance for rejecting a hypothesized value μ. Even if the hypothesized value μ' is correct, the actual data will contain statistical fluctuations and thus the observed significance is not in general equal to the median.

For example, if the signal is in fact absent but the number of background events fluctuates upward, then the observed upper limit on the parameter μ will be weaker than the median assuming background only. It is useful to know by how much the significance is expected to vary, given the expected fluctuations in the data. As we have formulae for all of the relevant sampling distributions, we can also predict how the significance is expected to vary under assumption of a given signal strength.

It is convenient to calculate error bands for the median significance corresponding to the $\pm N\sigma$ variation of $\hat{\mu}$. As $\hat{\mu}$ is Gaussian distributed, these error bands on the significance are simply the quantiles that map onto the variation of $\hat{\mu}$ of $\pm N\sigma$ about μ'.

For the case of discovery, i.e., a test of $\mu = 0$, one has from (46) and (52) that the significance Z_0 is

$$Z_0 = \begin{cases} \hat{\mu}/\sigma, & \hat{\mu} \geq 0, \\ 0, & \hat{\mu} < 0. \end{cases} \tag{84}$$

Furthermore the median significance is found from (78), so the significance values corresponding to $\mu' \pm N\sigma$ are therefore

$$Z_0(\mu' + N\sigma) = \text{med}[Z|\mu'] + N, \tag{85}$$

$$Z_0(\mu' - N\sigma) = \max\bigl[\text{med}[Z|\mu'] - N, 0\bigr]. \tag{86}$$

For the case of exclusion, when using both the statistic q_μ as well as $\tilde{q}_\mu$ one found the same expression for the upper limit at a confidence level of $1 - \alpha$, namely, (60). Therefore the median upper limit assuming a vstrength parameter μ' is found simply by substituting this for $\hat{\mu}$, and the $\pm N\sigma$ error bands are found similarly by substituting the corresponding values of $\mu' \pm N\sigma$. That is, the median upper limit is

$$\text{med}[\mu_{\text{up}}|\mu'] = \mu' + \sigma\Phi^{-1}(1 - \alpha), \tag{87}$$

and the $\pm N\sigma$ error band is given by

$$\text{band}_{N\sigma} = \mu' + \sigma\bigl(\Phi^{-1}(1 - \alpha) \pm N\bigr). \tag{88}$$

The standard deviation σ of $\hat{\mu}$ can be obtained from the Asimov value of the test statistic q_μ (or $\tilde{q}_\mu$) using (30).

5 Examples

In this section we describe two examples, both of which are special cases of the generic analysis described in Sect. 2. Here one has a histogram $\mathbf{n} = (n_1, \ldots, n_N)$ for the main measurement where signal events could be present and one may have another histogram $\mathbf{m} = (m_1, \ldots, m_M)$ as a control measurement, which helps constrain the nuisance parameters. In Sect. 5.1 we treat the simple case where each of these two measurements consists of a single Poisson distributed value, i.e., the histograms each have a single bin. We refer to this as a "counting experiment". In Sect. 5.2 we consider multiple bins for the main histogram, but without a control histogram; here the measured shape of the main histogram on either side of the signal peak is sufficient to constrain the background. We refer to this as a "shape analysis".

5.1 Counting experiment

Consider an experiment where one observes a number of events n, assumed to follow a Poisson distribution with an expectation value $E[n] = \mu s + b$. Here s represents the mean number of events from a signal model, which we take to be a known value; b is the expected number from background processes, and as usual μ is the strength parameter.

We will treat b as a nuisance parameter whose value is constrained by a control measurement. This measurement is also a single Poisson distributed value m with mean value $E[m] = \tau b$. That is, τb plays the role of the function u for the single bin of the control histogram in (5). In a real analysis, the value of the scale factor τ may have some uncertainty and could be itself treated as a nuisance parameter, but in this example we will take its value to be known. Related aspects of this type of analysis have been discussed in the literature, where it is sometimes referred to as the "on-off problem" (see, e.g., [13, 16]).

The data thus consist of two measured values: n and m. We have one parameter of interest, μ, and one nuisance parameter, b. The likelihood function for μ and b is the product of two Poisson terms:

$$L(\mu, b) = \frac{(\mu s + b)^n}{n!} e^{-(\mu s + b)} \frac{(\tau b)^m}{m!} e^{-\tau b}. \tag{89}$$

To find the test statistics q_0, q_μ and $\tilde{q}_\mu$, we require the ML estimators $\hat{\mu}$, $\hat{b}$ as well as the conditional ML estimator $\hat{\hat{b}}$ for a specified μ. These are found to be

$$\hat{\mu} = \frac{n - m/\tau}{s}, \tag{90}$$

$$\hat{b} = \frac{m}{\tau}, \tag{91}$$

$$\hat{\hat{b}} = \frac{n + m - (1 + \tau)\mu s}{2(1 + \tau)} + \left[\frac{(n + m - (1 + \tau)\mu s)^2 + 4(1 + \tau)m\mu s}{4(1 + \tau)^2}\right]^{1/2}. \tag{92}$$

Given measured values n and m, the estimators from (90), (91) and (92) can be used in the likelihood function (89) to find the values of the test statistics q_0, q_μ and $\tilde{q}_\mu$. By generating data values n and m by Monte Carlo we can compare the resulting distributions with the formulae from Sect. 3.

The pdf $f(q_0|0)$, i.e., the distribution of q_0 for under the assumption of $\mu = 0$, is shown in Fig. 3(a). The histograms show the result from Monte Carlo simulation based on several different values of the mean background b. The solid curve shows the prediction of (48), which is independent of the nuisance parameter b. The point at which one finds a significant departure between the histogram and the asymptotic

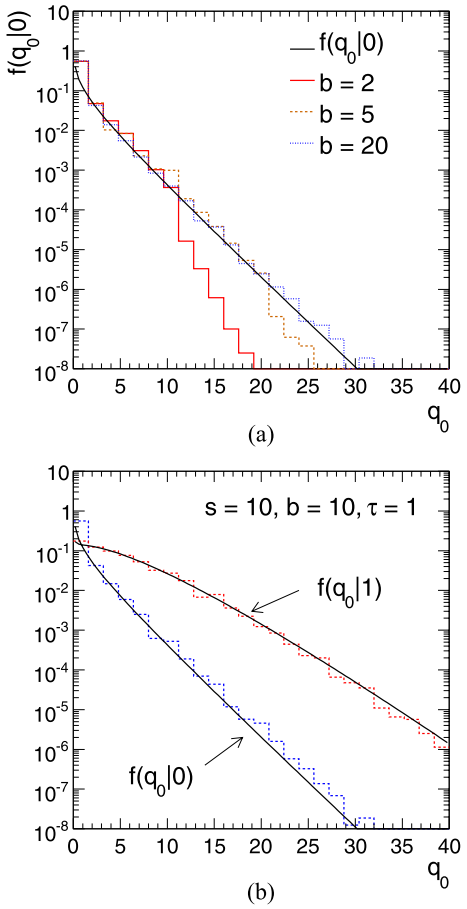

Fig. 3 (**a**) The pdf $f(q_0|0)$ for the counting experiment. The *solid curve* shows $f(q_0|0)$ from (48) and the histograms are from Monte Carlo simulation using different values of b (see text). (**b**) The distributions $f(q_0|0)$ and $f(q_0|1)$ from both the asymptotic formulae and Monte Carlo simulation based on $s = 10$, $b = 10$, $\tau = 1$

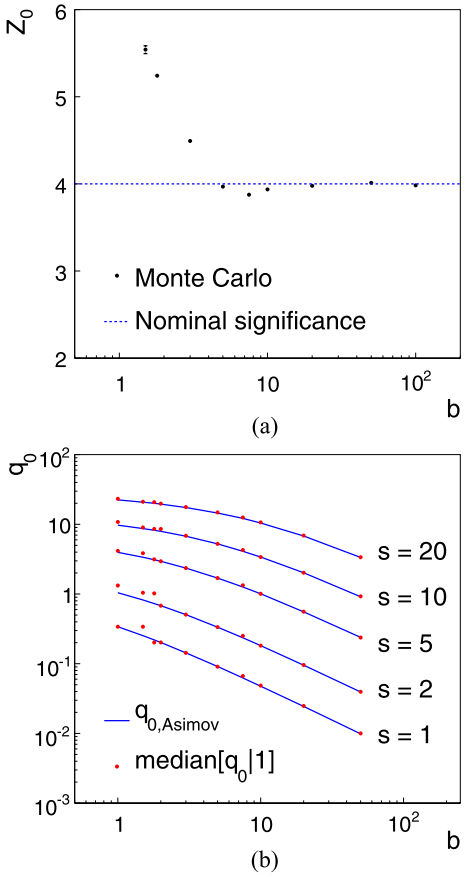

Fig. 4 (**a**) The discovery significance Z_0 obtained from Monte Carlo simulation (*points*) corresponding to a nominal value $Z_0 = \sqrt{q_0} = 4$ (*dashed line*) as a function of the expected number of background events b, in the counting analysis with a scale factor $\tau = 1$. (**b**) The median of q_0 assuming data distributed according to the nominal signal hypothesis from Monte Carlo simulation for different values of s and b (*points*) and the corresponding Asimov values (*curves*)

formula occurs at increasingly large q_0 for increasing b. For $b = 20$ the agreement is already quite accurate past $q_0 = 25$, corresponding to a significance of $Z = \sqrt{q_0} = 5$. Even for $b = 2$ there is good agreement out to $q_0 \approx 10$.

Figure 3(b) shows distributions of q_0 assuming a strength parameter μ' equal to 0 and 1. The histograms show the Monte Carlo simulation of the corresponding distributions using the parameters $s = 10$, $b = 10$, $\tau = 1$. For the distribution $f(q_0|1)$ from (47), one requires the value of σ, the standard deviation of $\hat{\mu}$ assuming a strength parameter $\mu' = 1$. Here this was determined from (32) using the Asimov value $q_{0,\mathrm{A}}$, i.e., the value obtained from the Asimov data set with $n \to \mu' s + b$ and $m \to \tau b$.

We can investigate the accuracy of the approximations used by comparing the discovery significance for a given observed value of q_0 from the approximate formula with the exact significance determined using a Monte Carlo calculation. Figure 4(a) shows the discovery significance that one finds from $q_0 = 16$. According to (52), this should give a nominal significance of $Z = \sqrt{q_0} = 4$, indicated in the figure by the horizontal line. The points show the exact significance for different values of the expected number of background events b in the counting analysis with a scale factor $\tau = 1$. As can be seen, the approximation underestimates the significance for very low b, but achieves an accuracy of better than 10% for b greater than around 4. It slightly overestimates for b greater than around 5. This phenomenon can be seen in the tail of $f(q_0|0)$ in Fig. 3(b), which uses $b = 10$. The accuracy then rapidly improves for increasing b.

Figure 4(b) shows the median value of the statistic q_0 assuming data distributed according to the nominal signal hypothesis from Monte Carlo simulation (points) and the value based on the Asimov data set as a function of b for different values of s, using a scale factor $\tau = 1$. One can see that the Asimov data set leads to an excellent approximation to the median, except at very low s and b.

Figure 5(a) shows the distribution of the test statistic q_1 for $s = 6$, $b = 9$, $\tau = 1$ for data corresponding to a strength parameter $\mu' = 1$ and also $\mu' = 0$. The vertical lines indicate

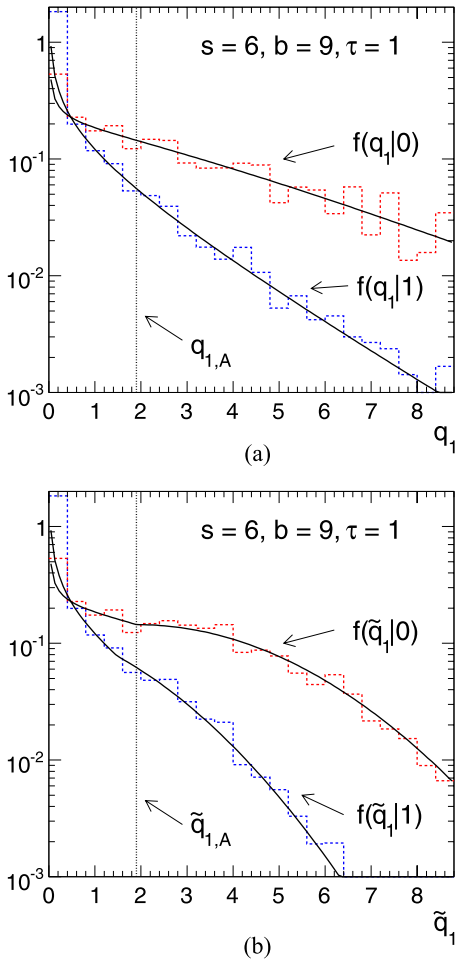

Fig. 5 (**a**) The pdfs $f(q_1|1)$ and $f(q_1|0)$ for the counting experiment. The *solid curves* show the formulae from the text, and the histograms are from Monte Carlo simulation using $s = 6$, $b = 9$, $\tau = 1$. (**b**) The same set of histograms with the alternative statistic $\tilde{q}_1$. The oscillatory structure evident in the histograms is a consequence of the discreteness of the data. The *vertical line* indicates the Asimov value of the test statistic corresponding to $\mu' = 0$

the Asimov values of q_1 and $\tilde{q}_1$ assuming a strength parameter $\mu' = 0$. These lines correspond to estimates of the median values of the test statistics assuming $\mu' = 0$. The areas under the curves $f(q_1|1)$ and $f(\tilde{q}_1|1)$ to the right of this line give the median p-values.

For the example described above we can also find the distribution of the statistic $q = -2\ln(L_{s+b}/L_b)$ as defined in Sect. 3.8. Figure 6 shows the distributions of q for the hypothesis of $\mu = 0$ (background only) and $\mu = 1$ (signal plus background) for the model described above using $b = 20$, $s = 10$ and $\tau = 1$. The histograms are from Monte Carlo simulation, and the solid curves are the predictions of the asymptotic formulae given in Sect. 3.8. Also shown are the p-values for the background-only and signal-plus-background hypotheses corresponding to a possible observed value of the statistic q_{obs}.

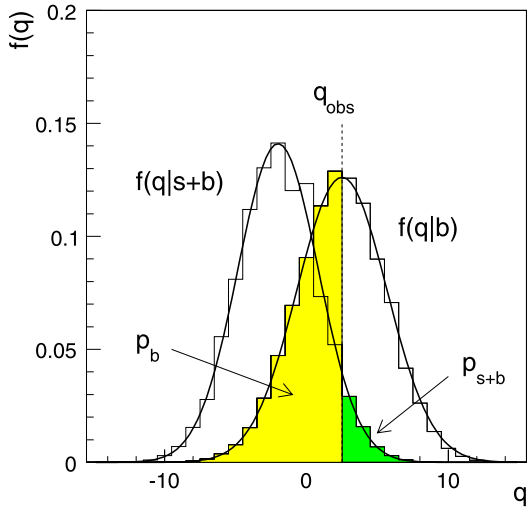

Fig. 6 The distribution of the statistic $q = -2\ln(L_{s+b}/L_b)$ under the hypotheses of $\mu = 0$ and $\mu = 1$ (see text)

5.1.1 *Counting experiment with known b*

An important special case of the counting experiment above is where the mean background b is known with negligible uncertainty and can be treated as a constant. This would correspond to having a very large value for the scale factor τ.

If we regard b as known, the data consist only of n and thus the likelihood function is

$$L(\mu) = \frac{(\mu s + b)^n}{n!} e^{-(\mu s + b)}. \tag{93}$$

The test statistic for discovery q_0 can be written

$$q_0 = \begin{cases} -2\ln\frac{L(0)}{L(\hat{\mu})}, & \hat{\mu} \geq 0, \\ 0, & \hat{\mu} < 0, \end{cases} \tag{94}$$

where $\hat{\mu} = n - b$. For sufficiently large b we can use the asymptotic formula (52) for the significance,

$$Z_0 = \sqrt{q_0} = \begin{cases} \sqrt{2(n\ln\frac{n}{b} + b - n)}, & \hat{\mu} \geq 0, \\ 0, & \hat{\mu} < 0. \end{cases} \tag{95}$$

To approximate the median significance assuming the nominal signal hypothesis ($\mu = 1$) we replace n by the Asimov value $s + b$ to obtain

$$\text{med}[Z_0|1] = \sqrt{q_{0,\text{A}}} = \sqrt{2\big((s+b)\ln(1+s/b) - s\big)}. \tag{96}$$

Expanding the logarithm in s/b one finds

$$\text{med}[Z_0|1] = \frac{s}{\sqrt{b}}\big(1 + \mathcal{O}(s/b)\big). \tag{97}$$

Although $Z_0 \approx s/\sqrt{b}$ has been widely used for cases where $s + b$ is large, one sees here that this final approximation is strictly valid only for $s \ll b$.

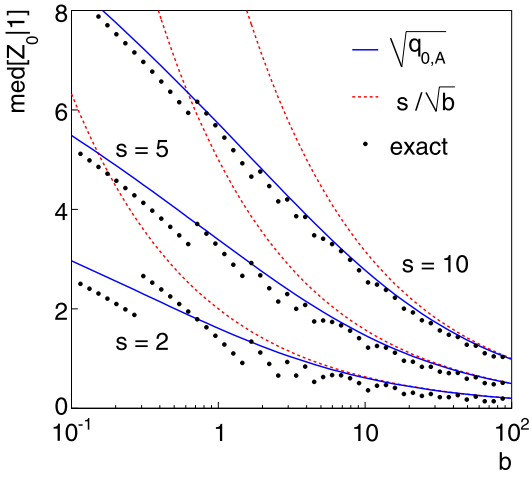

Fig. 7 The median, assuming $\mu = 1$, of the discovery significance Z_0 for different values of s and b (see text)

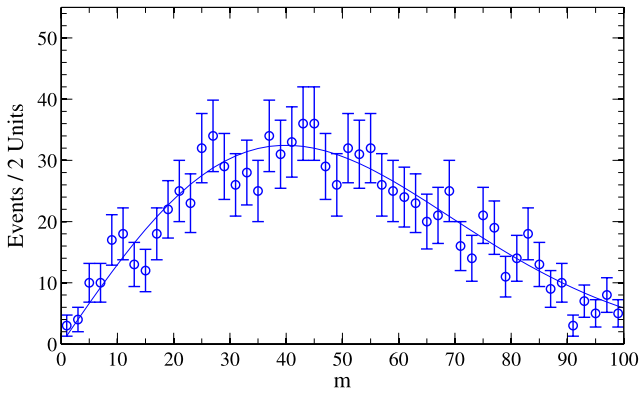

Fig. 8 The background mass distribution for the shape analysis (see text)

Median values, assuming $\mu = 1$, of Z_0 for different values of s and b are shown in Fig. 7. The solid curve shows (96), the dashed curve gives the approximation $s/\sqrt{b}$, and the points are the exact median values from Monte Carlo simulation. The structure seen in the points is due to the discrete nature of the data. One sees that (96) provides a much better approximation to the true median than does $s/\sqrt{b}$ in regions where s/b cannot be regarded as small.

5.2 Shape analysis

As a second example we consider the case where one is searching for a peak in an invariant mass distribution. The main histogram $\mathbf{n} = (n_1, \ldots, n_N)$ for background is shown in Fig. 8, which is here taken to be a Rayleigh distribution. The signal is modeled as a Gaussian of known width and mass (position). In this example there is no subsidiary histogram $(m_1, \ldots, m_M)$.

If, as is often the case, the position of the peak is not known a priori, then one will test all masses in a given

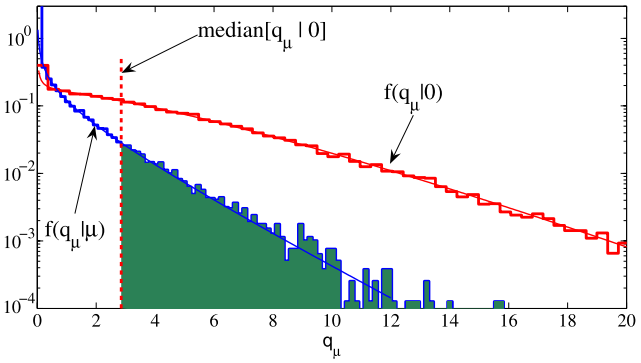

Fig. 9 (Color online) The distributions $f(q_\mu|0)$ (*red*) and $f(q_\mu|\mu)$ (*blue*) from both the asymptotic formulae and Monte Carlo histograms (see text)

range, and appearance of a signal-like peak anywhere could lead to rejection of the background-only hypothesis. In such an analysis, however, the discovery significance must take into account the fact that a fluctuation could occur at any mass within the range. This is often referred to as the "look-elsewhere effect", and is discussed further in [17].

In the example presented here, however, we will test all values of the mass and μ using the statistic q_μ for purposes of setting an upper limit on the signal strength. Here, each hypothesis of mass and signal strength is in effect tested individually, and thus the look-elsewhere effect does not come into play.

We assume that the signal and background distributions are known up to a scale factor. For the signal, this factor corresponds to the usual strength parameter μ; for the background, we introduce an analogous factor θ. That is, the mean value of the number of events in the ith bin is $E[n_i] = \mu s_i + b_i$, where μ is the signal strength parameter and the s_i are taken as known. We assume that the background terms b_i can be expressed as $b_i = \theta f_{\mathrm{b},i}$, where the probability to find a background event in bin i, $f_{\mathrm{b},i}$, is known, and θ is a nuisance parameter that gives the total expected number of background events. Therefore the likelihood function can be written

$$L(\mu, \theta) = \prod_{i=1}^{N} \frac{(\mu s_i + \theta f_{\mathrm{b},i})^{n_i}}{n_i!} e^{-(\mu s_i + \theta f_{\mathrm{b},i})}. \tag{98}$$

For a given data set $\mathbf{n} = (n_1, \ldots, n_N)$ one can evaluate the likelihood (98) and from this determine any of the test statistics discussed previously. Here we concentrate on the statistic q_μ used to set an upper limit on μ, and compare the distribution $f(q_\mu|\mu')$ from (47) with histograms generated by Monte Carlo simulation. Figure 9 shows $f(q_\mu|0)$ (red) and $f(q_\mu|\mu)$ (blue).

The vertical line in Fig. 9 gives the median value of q_μ assuming a strength parameter $\mu' = 0$. The area to the right of this line under the curve of $f(q_\mu|\mu)$ gives the p-value of

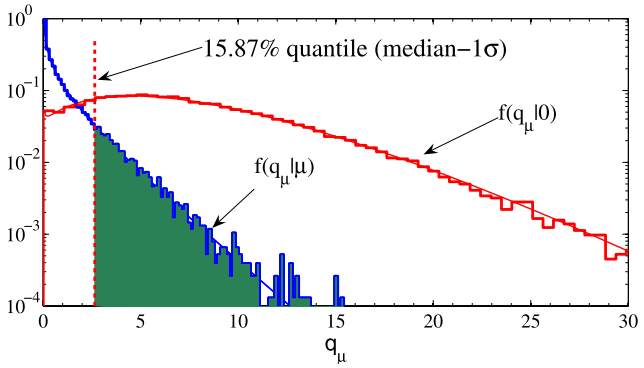

Fig. 10 (Color online) The distributions $f(q_\mu|0)$ (*red*) and $f(q_\mu|\mu)$ (*blue*) as in Fig. 9 and the 15.87% quantile of $f(q_\mu|0)$ (see text)

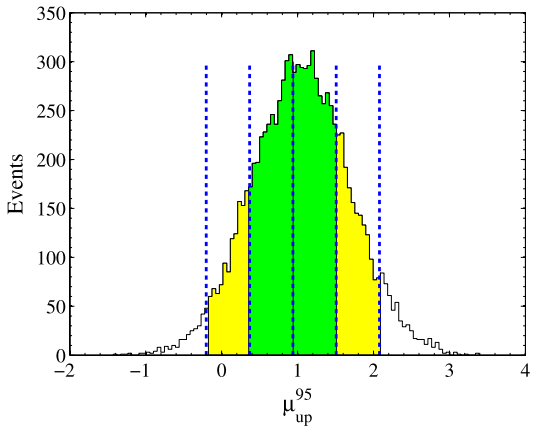

Fig. 11 Distribution of the upper limit on μ at 95% CL, assuming data corresponding to the background-only hypothesis (see text)

the hypothesized μ, as shown shaded in green. The upper limit on μ at a confidence level $\mathrm{CL} = 1 - \alpha$ is the value of μ for which the p-value is $p_\mu = \alpha$. Figure 9 shows the distributions for the value of μ that gave $p_\mu = 0.05$, corresponding to the 95% CL upper limit.

In addition to reporting the median limit, one would like to know how much it would vary for given statistical fluctuations in the data. This is illustrated in Fig. 10, which shows the same distributions as in Fig. 9, but here the vertical line indicates the 15.87% quantile of the distribution $f(q_\mu|0)$, corresponding to having $\hat{\mu}$ fluctuate downward one standard deviation below its median.

By simulating the experiment many times with Monte Carlo, we can obtain a histogram of the upper limits on μ at 95% CL, as shown in Fig. 11. The $\pm 1\sigma$ (green) and $\pm 2\sigma$ (yellow) error bands are obtained from the MC experiments. The vertical lines indicate the error bands as estimated directly (without Monte Carlo simulation) using (87) and (88). As can be seen from the plot, the agreement between the formulae and MC predictions is excellent.

Figures 9 through 11 correspond to finding an upper limit on μ for a specific value of the peak position (mass). In a search for a signal of unknown mass, the procedure would be repeated for all masses (in practice in small steps). Figure 12 shows the median upper limit at 95% CL as a function of mass. The median (central blue line) and error bands ($\pm 1\sigma$ in green, $\pm 2\sigma$ in yellow) are obtained using (87) and (88). The points and connecting curve correspond to the upper limit from a single arbitrary Monte Carlo data set, generated according to the background-only hypothesis. As can be seen, most of the plots lie as expected within the $\pm 1\sigma$ error band.

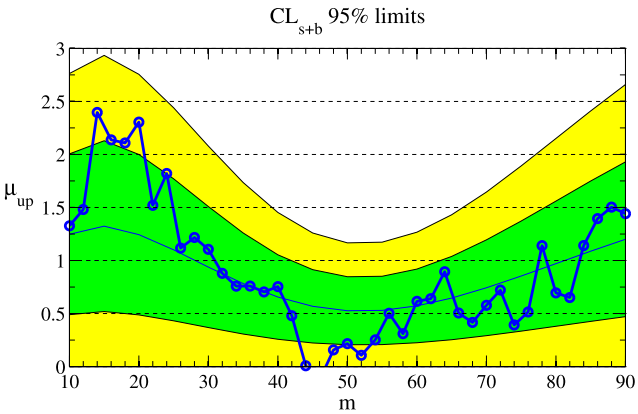

Fig. 12 (Color online) The median (*central blue line*) and error bands ($\pm 1\sigma$ in *green*, $\pm 2\sigma$ in *yellow*) for the 95% CL upper limit on the strength parameter μ (see text)

6 Implementation in RooStats

Many of the results presented above are implemented or are being implemented in the RooStats framework [18], which is a C++ class library based on the ROOT [19] and RooFit [20] packages. The tools in RooStats can be used to represent arbitrary probability density functions that inherit from `RooAbsPdf`, the abstract interfaces for probability density functions provided by RooFit.

The framework provides an interface with minimization packages such as `Minuit` [21]. This allows one to obtain the estimators required in the profile likelihood ratio: $\hat{\mu}$, $\hat{\boldsymbol{\theta}}$, and $\hat{\hat{\boldsymbol{\theta}}}$. The Asimov dataset defined in (24) can be determined for a probability density function by specifying the `ExpectedData()` command argument in a call to the `generateBinned` method. The Asimov data together with the standard `HESSE` covariance matrix provided by `Minuit` makes it is possible to determine the Fisher information matrix shown in (28), and thus obtain the related quantities such as the variance of $\hat{\mu}$ and the noncentrality parameter Λ, which enter into the formulae for a number of the distributions of the test statistics presented above.

The distributions of the various test statistics and the related formulae for p-values, sensitivities and confidence intervals as given in Sects. 2, 3 and 4 are being incorporated

as well. RooStats currently includes the test statistics t_μ, $\tilde{t}_\mu$, q_0, and q, q_μ, and $\tilde{q}_\mu$ as concrete implementations of the `TestStatistic` interface. Together with the Asimov data, this provides the ability to calculate the alternative estimate, σ_A, for the variance of $\hat{\mu}$ shown in (30). The noncentral chi-square distribution is being incorporated into both RooStats and ROOT's mathematics libraries for more general use. The various transformations of the noncentral chi-square used to obtain (33), (41), (47), (54), and (62) are also in development in the form of concrete implementations of the `SamplingDistribution` interface. Together, these new classes will allow one to reproduce the examples shown in Sect. 5 and to extend them to an arbitrary model within the RooStats framework.

7 Conclusions

Statistical tests are described for use in planning and carrying out a search for new phenomena. The formalism allows for the treatment of systematic uncertainties through use of the profile likelihood ratio. Here a systematic uncertainty is included to the extent that the model includes a sufficient number of nuisance parameters so that for at least some point in its parameter space it can be regarded as true.

Approximate formulae are given for the distributions of test statistics used to characterize the level of agreement between the data and the hypothesis being tested, as well as the related expressions for p-values and significances. The statistics are based on the profile likelihood ratio and can be used for a two-sided test of a strength parameter μ (t_μ), a one-sided test for discovery (q_0), and a one-sided test for finding an upper limit (q_μ and $\tilde{q}_\mu$). The statistic $\tilde{t}_\mu$ can be used to obtain a "unified" confidence interval, in the sense that it is one- or two-sided depending on the data outcome.

Formulae are also given that allow one to characterize the sensitivity of a planned experiment through the median significance of a given hypothesis under assumption of a different one, e.g., median significance with which one would reject the background-only hypothesis under assumption of a certain signal model. These exploit the use of an artificial data set, the "Asimov" data set, defined so as to make estimators for all parameters equal to their true values. Methods for finding the expected statistical variation in the sensitivity (error bands) are also given.

These tools free one from the need to carry out lengthy Monte Carlo calculations, which in the case of a discovery at 5σ significance could require simulation of around 10^8 measurements. They are particularly useful in cases where one needs to estimate experimental sensitivities for many points in a multidimensional parameter space (e.g., for models such as supersymmetry), which would require generating a large MC sample for each point.

The approximations used are valid in the limit of a large data sample. Tests with Monte Carlo simulation indicate, however, that the formulae are in fact reasonably accurate even for fairly small samples, and thus can have a wide range of practical applicability. For very small samples and in cases where high accuracy is crucial, one is always free to validate the approximations with Monte Carlo simulation.

Acknowledgements The authors would like to thank Louis Fayard, Nancy Andari, Francesco Polci and Marumi Kado for fruitful discussions. We received useful feedback at the Banff International Research Station, specifically from Richard Lockhart and Earl Lawrence. One of us (E.G.) is obliged to the Benoziyo Center for High Energy Physics, to the Israeli Science Foundation (ISF), the Minerva Gesellschaft and the German Israeli Foundation (GIF) for supporting this work. K.C. is supported by US National Science Foundation grant PHY-0854724. G.C. thanks the UK Science and Technology Facilities Council as well as the Einstein Center at the Weizmann Institute of Science, where part of his work on this paper was done.

References

1. S.S. Wilks, The large-sample distribution of the likelihood ratio for testing composite hypotheses. Ann. Math. Stat. **9**, 60–62 (1938)
2. A. Wald, Tests of statistical hypotheses concerning several parameters when the number of observations is large. Trans. Am. Math. Soc. **54**(3), 426–482 (1943)
3. I. Asimov, Franchise, in *Isaac Asimov: The Complete Stories*, vol. 1 (Broadway Books, New York, 1990)
4. V. Bartsch, G. Quast, Expected signal observability at future experiments, CMS Note 2005/004 (2003), (available on CMS information server)
5. ATLAS Collaboration, Expected performance of the ATLAS experiment, detector, trigger and physics. CERN-OPEN-2008-020, Geneva (2008). e-print: arXiv:0901.0512
6. ALEPH, DELPHI, and L3 and OPAL Collaborations, Search for the standard model higgs boson at LEP. Phys. Lett. B **565**, 61–75 (2003). CERN-EP/2003-011
7. A. Stuart, J.K. Ord, S. Arnold, *Kendall's Advanced Theory of Statistics, Classical Inference and the Linear Model*, vol. 2A, 6th edn. (Oxford University Press, London, 1999), and earlier editions by Kendall and Stuart
8. R.D. Cousins, G.J. Feldman, Phys. Rev. D **57**, 3873 (1998)
9. H. Chernoff, On the distribution of the likelihood ratio. Ann. Math. Stat. **25**, 573–578 (1954)
10. M. Abramowitz, I.A. Stegun, *Handbook of Mathematical Functions* (Dover, New York, 1972). Sect. 26.4.25
11. Wikipedia, The Free Encyclopedia, Noncentral chi-square distribution. Wikimedia Foundation, Inc., 6 July 2010
12. T. Aaltonen et al., Phys. Rev. Lett. **104**, 061802 (2010). e-Print: arXiv:1001.4162 [hep-ex]
13. K. Cranmer, Frequentist hypothesis testing with background uncertainty, in *Proceedings of PHYSTAT 2003, SLAC*, ed. by L. Lyons et al., Stanford, California, 8–11 September 2003, pp. 261–264
14. T. Junk, Nucl. Instrum. Methods Phys. Res., Sect. A **434**, 435 (1999)

15. A.L. Read, J. Phys. G **28**, 2693 (2002)
16. R.D. Cousins, J.T. Linnemann, J. Tucker, Nucl. Instrum. Methods Phys. Res., Sect. A **595**, 480–501 (2008). e-Print: arXiv:physics/0702156v4 [physics.data-an]
17. E. Gross, O. Vitells, Trial factors or the look elsewhere effect in high energy physics. arXiv:1005.1891 [physics.data-an] (2010)
18. L. Moneta, K. Belasco, K. Cranmer et al., The RooStats project, in *Proceedings of ACAT*, Jaipur, India (2010). arXiv:1009.1003 [physics.data-an]. https://twiki.cern.ch/twiki/bin/view/RooStats/
19. R. Brun, F. Rademakers, ROOT: An object oriented data analysis framework,. Nucl. Instrum. Methods A **389**, 81–86 (1997)
20. W. Verkerke, D.P. Kirkby, The RooFit toolkit for data modeling, in *Proceedings for CHEP03* (2003). physics/0306116
21. F. James, M. Roos, Minuit: a system for function minimization and analysis of the parameter errors and correlations. Comput. Phys. Commun. **10**, 343–367 (1975)