Use of the profile likelihood function in searches for new physics

Glen Cowan
Physics Department, Royal Holloway, University of London, Egham TW20 0EX, UK

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. Explicit formulae for the asymptotic distributions of test statistics based on the profile likelihood ratio 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

This paper summarizes results recently published in Ref. [1]. These allow one to carry out statistical tests in 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. The approximate methods reported here are based on results due to Wilks [2] and Wald [3], which allow one to 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 Sec. 2 the formalism of a search as a statistical test is outlined. Several test statistics based on the profile likelihood ratio are defined in Sec. 3 that can be used for establishing a discovery or setting upper limits. Example applications are shown in Sec. 4, and conclusions are given in Sec. 5.

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 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. 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 its mean has an upper-tail probability equal to p. That is, $Z = \Phi^{-1}(1-p)$, where Φ^{-1} is the quantile (inverse of the cumulative distribution) of the standard Gaussian.

It is often useful to quantify the sensitivity of an experiment by reporting the expected (or more precisely, the median) 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 median value, under the assumption of H_1, of the value of Z obtained from a test of H_0.

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 $\vec{n} = (n_1, \ldots, n_N)$. The expectation value of n_i can be written $E[n_i] = \mu s_i + b_i$, where the mean number of entries in the ith bin from signal and background are

$$s_i = s_{\rm tot} \int_{{\rm bin}\,i} f_s(x; \vec{\theta}_s)\, dx \;, \qquad b_i = b_{\rm tot} \int_{{\rm bin}\,i} f_b(x; \vec{\theta}_b)\, dx \,. \tag{1}$$

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 $f_s(x; \vec{\theta}_s)$ and $f_b(x; \vec{\theta}_b)$ are the probability density functions (pdfs) of the variable x for signal and background events, and $\vec{\theta}_s$ and $\vec{\theta}_b$ represent parameters that characterize the shapes of pdfs. The quantities $s_{\rm tot}$ and $b_{\rm tot}$ are the total mean numbers of signal and background events. Below we will use $\vec{\theta} = (\vec{\theta}_s, \vec{\theta}_b, b_{\rm tot})$ 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 $\vec{n}$ one often makes 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 $\vec{m} = (m_1, \ldots, m_M)$ for the number of entries in each of the M bins. The expectation value of m_i can be written $E[m_i] = u_i(\vec{\theta})$, where the u_i are calculable quantities depending on the parameters $\vec{\theta}$. 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:

$$L(\mu, \vec{\theta}) = \prod_{j=1}^{N} \frac{(\mu s_j + b_j)^{n_j}}{n_j!} e^{-(\mu s_j + b_j)} \;\; \prod_{k=1}^{M} \frac{u_k^{m_k}}{m_k!}\, e^{-u_k} \;. \tag{2}$$

To test a hypothesized value of μ we consider the profile likelihood ratio (see, e.g., [4]),

$$\lambda(\mu) = \frac{L(\mu, \hat{\hat{\vec{\theta}}})}{L(\hat{\mu}, \hat{\vec{\theta}})} \;. \tag{3}$$

Here $\hat{\hat{\vec{\theta}}}$ in the numerator denotes the value of $\vec{\theta}$ that maximizes L for the specified μ, i.e., it is the conditional maximum-likelihood (ML) estimator of $\vec{\theta}$ (and thus is a function of μ). The denominator is the maximized (unconditional) likelihood function, i.e., $\hat{\mu}$ and $\hat{\vec{\theta}}$ are the ML estimators.

In many analyses, the contribution of the signal process to the mean number of events is assumed to be nonnegative, which is to say that any physical estimator for μ must be nonnegative. 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 Sec. 3 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.

3 Test statistics for discovery and upper limits

In this section we present test statistics based on the profile likelihood ratio. To compute p-values and sensitivities one requires the sampling distributions of these statistics. These are given below in an approximate form valid in the large-sample limit. More details can be found in Ref. [1].

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

From the definition of $\lambda(\mu)$ in Eq. (3), one can see that $0 \le \lambda \le 1$, with λ near 1 implying better agreement between the data and the hypothesized value of μ. Equivalently it is convenient to use the statistic $t_\mu = -2\ln\lambda(\mu)$ as the basis of a statistical test. Higher values of t_μ thus correspond to increasing incompatibility between the data and μ. To quantify the level of disagreement we compute the p-value,

$$p_\mu = \int_{t_{\mu,\mathrm{obs}}}^{\infty} f(t_\mu|\mu)\, dt_\mu \ , \tag{4}$$

where $t_{\mu,\mathrm{obs}}$ 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. The p-values for all of the statistics considered here are obtained in an analogous fashion.

To find the distribution of t_μ as well as that of other related statistics, we can use a relation due to Wald [3], who showed that for the case of a single parameter of interest,

$$-2\ln\lambda(\mu) = \frac{(\mu-\hat{\mu})^2}{\sigma^2} + \mathcal{O}(1/\sqrt{N}) \ . \tag{5}$$

Here $\hat{\mu}$ follows a Gaussian distribution with a mean μ' and standard deviation σ, and N represents the data sample size. The approximations presented here are valid to the extent that the $\mathcal{O}(1/\sqrt{N})$ term can be neglected.

If $\hat{\mu}$ is Gaussian distributed and we neglect the $\mathcal{O}(1/\sqrt{N})$ term in Eq. (5), then one can show that the statistic $t_\mu = -2\ln\lambda(\mu)$ follows a *noncentral chi-square* distribution for one degree of freedom,

$$f(t_\mu;\Lambda) = \frac{1}{2\sqrt{t_\mu}}\frac{1}{\sqrt{2\pi}}\left[\exp\left(-\frac{1}{2}\left(\sqrt{t_\mu}+\sqrt{\Lambda}\right)^2\right) + \exp\left(-\frac{1}{2}\left(\sqrt{t_\mu}-\sqrt{\Lambda}\right)^2\right)\right] \ , \tag{6}$$

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

3.2 The statistic q_0 for discovery

Often one wishes to test $\mu = 0$ in a class of models where we assume $\mu \ge 0$. Rejecting $\mu = 0$ amounts to discovering a new (positive) signal. In such a case one can define the test such that the data are only regarded as discrepant with the hypothesis of $\mu = 0$ if one observes an excess of events, i.e., one finds $\hat{\mu} > 0$. That is, we define the statistic

$$q_0 = \begin{cases} -2\ln\lambda(0) & \hat{\mu} \ge 0 \ , \\ 0 & \hat{\mu} < 0 \ , \end{cases} \tag{7}$$

where $\lambda(0)$ is the profile likelihood ratio for $\mu = 0$ as defined in Eq. (3).

Assuming the validity of the Wald approximation, one can show that the pdf of q_0 has the form

$$f(q_0|\mu') = \left(1 - \Phi\left(\frac{\mu'}{\sigma}\right)\right)\delta(q_0) + \frac{1}{2}\frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{q_0}}\exp\left[-\frac{1}{2}\left(\sqrt{q_0} - \frac{\mu'}{\sigma}\right)^2\right] . \tag{8}$$

From Eq. (8) the corresponding cumulative distribution is found to be

$$F(q_0|\mu') = \Phi\left(\sqrt{q_0} - \frac{\mu'}{\sigma}\right) . \tag{9}$$

The p-value of the hypothesis $\mu = 0$, p_0, is obtained from these distributions by using $\mu' = 0$. For the significance one finds the simple formula

$$Z_0 = \Phi^{-1}(1 - p_0) = \sqrt{q_0} . \tag{10}$$

3.3 The statistic q_μ for upper limits

For purposes of establishing an upper limit on the strength parameter μ, one can define

$$q_\mu = \begin{cases} -2\ln\lambda(\mu) & \hat{\mu} \le \mu , \\ 0 & \hat{\mu} > \mu , \end{cases} \tag{11}$$

where $\lambda(\mu)$ is the profile likelihood ratio from Eq. (3). 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. 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. A closely related statistic, which we call $\tilde{q}_\mu$, is discussed in Ref. [1]. In the large-sample limit they are equivalent.

Assuming the validity of the Wald approximation, the pdf $f(q_\mu|\mu')$ is found to be

$$f(q_\mu|\mu') = \Phi\left(\frac{\mu' - \mu}{\sigma}\right)\delta(q_\mu) + \frac{1}{2}\frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{q_\mu}}\exp\left[-\frac{1}{2}\left(\sqrt{q_\mu} - \frac{\mu - \mu'}{\sigma}\right)^2\right] , \tag{12}$$

and the cumulative distribution is

$$F(q_\mu|\mu') = \Phi\left(\sqrt{q_\mu} - \frac{\mu - \mu'}{\sigma}\right) . \tag{13}$$

Using these ingredients with $\mu' = 0$, one can obtain the p-value of a hypothesized value of μ, p_μ, and the corresponding significance, Z_μ, which is found to be

$$Z_\mu = \Phi^{-1}(1 - p_\mu) = \sqrt{q_\mu} . \tag{14}$$

3.4 Asimov data set, variance of $\hat{\mu}$, median significance

Some of the formulae above require the standard deviation σ of $\hat{\mu}$. A useful way of estimating σ involves a special, artificial data set that we call the "Asimov data set". This is defined such that when it is used to evaluate the estimators for all parameters, one obtains the true parameter values. One can show that under conditions generally satisfied in practice, this amounts to setting the Poisson data values equal to their expectation values, which can be estimated using a very large Monte Carlo data sample. That is, the Asimov values for the measured histograms $\vec{n}$ and $\vec{m}$ are $n_{i,\mathrm{A}} = E[n_i] = \mu' s_i(\vec{\theta}) + b_i(\vec{\theta})$ and $m_{i,\mathrm{A}} = E[m_i] = u_i(\vec{\theta})$.

We can use the Asimov data set to evaluate the "Asimov likelihood" $L_{\rm A}$ and the corresponding profile likelihood ratio $\lambda_{\rm A}$. Because the Asimov data set corresponding to a strength μ' gives $\hat{\mu} = \mu'$, from Eq. (5) one finds

$$-2\ln\lambda_{\rm A}(\mu) \approx \frac{(\mu-\mu')^2}{\sigma^2} = \Lambda \; . \tag{15}$$

That is, the Asimov data set provides an estimate of the noncentrality parameter Λ that characterizes the distribution of $-2\ln\lambda(\mu)$. Equivalently, one can use Eq. (15) to obtain the variance of $\hat{\mu}$, σ^2.

When the statistics q_0 and q_μ are evaluated with an Asimov data set (denoted $q_{0,{\rm A}}$ and $q_{\mu,{\rm A}}$) one obtains good estimates for their median values, and these lead to simple expressions for the corresponding median significance. From Eqs. (10) and (14) 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. 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$[Z_\mu|0]$. Using the corresponding Asimov data set for each case, one finds

$$\text{med}[Z_0|\mu'] \;=\; \sqrt{q_{0,{\rm A}}} \; , \tag{16}$$

$$\text{med}[Z_\mu|0] \;=\; \sqrt{q_{\mu,{\rm A}}} \; . \tag{17}$$

4 Tests of asymptotic formulae

Several tests of the validity of the asymptotic formulae given above are described in Ref. [1]. Here as an example we consider a measurement consisting of a number of events n assumed to be Poisson distributed with a mean $\mu s+b$, and a control measurement m modeled as following a Poisson distribution with mean τb. Here s and τ are treated as known with $\tau = 1$, b is a nuisance parameter and μ is the parameter of interest. Figure 1(a) shows the distributions from the asymptotic formula as well as histograms from Monte Carlo using different values of b. One can see that even for b as low as 2, the asymptotic curve agrees out to $q_0 \approx 10$, corresponding to a discovery significance of $Z_0 \approx \sqrt{10}$. To establish a 5σ effect one needs to model the distribution beyond $q_0 = 25$, which is achieved reasonably well here for $b = 20$.

As a second example, Fig. 1(b) shows the median discovery significance with which one would reject $\mu = 0$ assuming data distributed according to $\mu = 1$ in an experiment where n is Poisson distributed with mean $\mu s + b$, but here b is known exactly and there is no control measurement. The exact values shown as points are determined from Monte Carlo, and the jumps are a consequence of the discreteness of the data. Using the Asimov data value $s + b$ to approximate the median significance, one finds

$$\text{med}[Z_0|1] = \sqrt{q_{0,{\rm A}}} = \sqrt{2\left((s+b)\ln(1+s/b) - s\right)} \; . \tag{18}$$

Expanding the logarithm to second order in s/b one finds med$[Z_0|1] = (s/\sqrt{b})\,(1 + \mathcal{O}(s/b))$. Although $Z_0 \approx s/\sqrt{b}$ has been widely used for cases where $s + b$ is large, this final approximation is strictly valid only for $s \ll b$, as can be seen in Fig. 1(b).

5 Conclusions

Statistical tests are described for use in planning and carrying out a search for new phenomena; further details can be found in Ref. [1]. 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 formulae are implemented in the `RooStats` software package [5].

(a) (b)

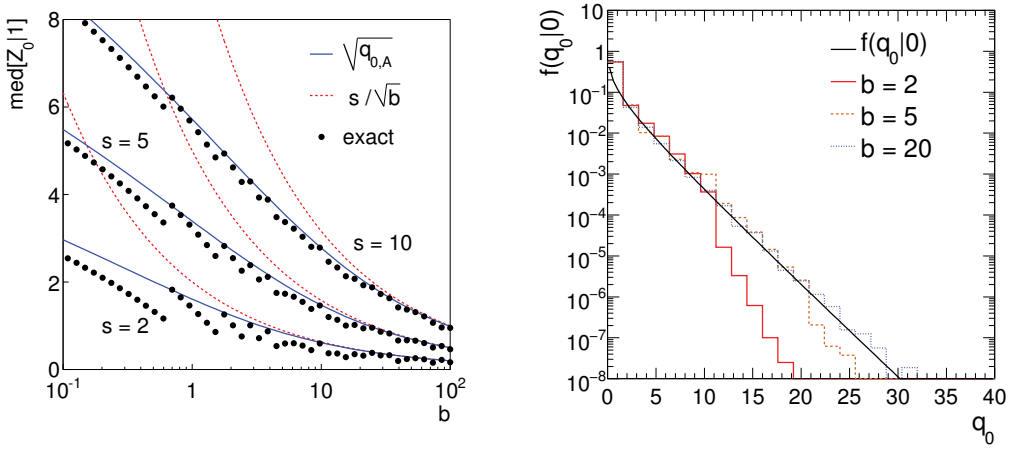

Fig. 1: (a) The pdf $f(q_0|0)$ for the counting experiment. The solid curve shows $f(q_0|0)$ from the asymptotic formula and the histograms are from Monte Carlo using different values of b (see text). (b) The median, assuming $\mu = 1$, of the discovery significance Z_0 for different values of s and b (the plot shown here corrects a minor numerical error in Fig. 7 of Ref. [1]).

The asymptotic formulae 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. The approximations used are valid in the limit of a large data sample. Tests with Monte Carlo 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.

Acknowledgements

I thank my coauthors in Ref. [1], Eilam Gross, Kyle Cranmer and Ofer Vitells; this report is presented here on all of our behalf. I also thank the organizers of PHYSTAT 2011, especially Louis Lyons and Harrison Prosper, for an enjoyable and stimulating meeting.

References

[1] G. Cowan, K. Cranmer, E. Gross and O. Vitells, Eur. Phys. J. C (2011) 71:1554; arXiv:1007.1727 [physics.data-an].

[2] S.S. Wilks, *The large-sample distribution of the likelihood ratio for testing composite hypotheses*, Ann. Math. Statist. **9** (1938) 60-2.

[3] A. Wald, *Tests of Statistical Hypotheses Concerning Several Parameters When the Number of Observations is Large*, Transactions of the American Mathematical Society, Vol. **54**, No. 3 (Nov., 1943), pp. 426-482.

[4] A. Stuart, J.K. Ord, and S. Arnold, *Kendall's Advanced Theory of Statistics*, Vol. 2A: *Classical Inference and the Linear Model* 6th Ed., Oxford Univ. Press (1999), and earlier editions by Kendall and Stuart.

[5] L. Moneta, K. Belasco, K. Cranmer *et al.*, *The RooStats Project*, proceedings of ACAT, 2010, Jaipur, India; arXiv:1009.1003 [physics.data-an].