Confidence Intervals and Upper Bounds for Small Signals in the Presence of Background Noise

Wolfgang A. Rolke

Department of Mathematics, University of Puerto Rico - Mayagüez, Mayagüez, PR 00681, USA,
Postal Address: PO Box 5959, Mayagüez, PR 00681,
Tel: (787) 851-6212, Email: w_rolke@rumac.upr.clu.edu

Angel M. López

Department of Physics, University of Puerto Rico - Mayagüez, Mayagüez, PR 00681, USA

We discuss a new method for setting limits on small signals in the presence of background noise. The method is based on a combination of a two dimensional confidence region and the large sample approximation to the likelihood ratio test statistic. It automatically quotes upper limits for small signals and two-sided confidence intervals for larger samples. We show that this method gives the correct coverage and also has good power.

Key words: Maximum likelihood, profile likelihood, confidence regions, coverage, Monte Carlo

Preprint submitted to Elsevier Preprint 20 May 2000
1 Introduction

Finding confidence intervals or upper limits for small signals has recently attracted a great deal of attention. The paper by Cousins and Feldman [1] rekindled the interest in this area by providing a unified approach, that is, it based the choice of quoting an upper limit or a two-sided confidence interval on the data alone, without the experimenter having to make this decision. The unified approach uses the Neyman construction together with a novel ordering principle. Subsequent papers such as Giunti [2] and Roe and Woodroofe [3] gave variations of this method, basing the ordering on other quantities. Common to all these methods is the need to have a fairly precise knowledge of the background rate, for example, from Monte Carlo simulations. Unfortunately, as we will see in section 4, these methods can fail when used in the presence of background uncertainty. A possible remedy is discussed in Cousins and Highland [4] where it is suggested to treat the background uncertainty as a systematic error. We will instead treat the background uncertainty as a statistical error and develop a method that is suitable in this situation. Our method is based on the likelihood ratio test statistic, together with an adjustment for those situations where there is very little (or no) signal observed. We will show that this method yields the correct coverage rate and that it has

\[1 \text{ This work was supported by the High Energy Physics Office (Grant DE-FG-97ER41045) and the EPSCoR Program (Grant DE-FG-94ER75764) of the US Department of Energy and the University of Puerto Rico} \]
good power.

The background rate has to be estimated either from the data or through Monte Carlo. Both of those methods have their strengths and their weaknesses. Using data requires choosing sidebands, and implicitly makes the assumption that the density generating the background events is the same in the signal region as it is in the sidebands. This leads to a quandary: If we choose a small sideband, this assumption seems more reasonable, but then we will also see fewer background events and therefore have a higher statistical uncertainty in the estimate of the background rate. Choosing a large region might yield higher statistics but makes the assumption of a linear background more tenuous. An alternative way to estimate the background rate is by Monte Carlo. One problem with this approach is that a good Monte Carlo is often difficult to do because we can never be completely sure that we have modeled all the relevant effects correctly. This is particularly true when searching for small signals since the efforts to reduce the background often mean one is probing the tails of distributions which may be difficult to model. Also, in High Energy Physics today running a complete Monte Carlo simulation of an experiment can be a formidable task from a computational point of view, and it might not be possible to run enough Monte Carlo to effectively eliminate the uncertainty in the background estimate. For these reasons, it is in many cases not possible to ignore the uncertainty in the background rate. Our method is well equipped to deal with the uncertainty that comes from limited statistics in both situations,
those where the background is estimated from the data as well as those where
the background is estimated using Monte Carlo.

In some cases there are two different sources of background. It turns out that
our method can be extended fairly easily to this situation also, regardless of
the method of estimation used.

2 A Description of the Method

In this section we will outline the basic ideas of this new method. We will
need the following notation. Assume that we observe \( x \) events in a suitably
chosen signal region and a total of \( y \) events in the background region. Here
the background region can be chosen fairly freely and need not be contiguous.
Furthermore, we assume that the ratio of the size of the background region
to the size of the signal region is \( \tau \). For example, if we use two background
regions of the same size as the signal region we get \( \tau = 2 \). Then a probability
model for the data is given by

\[
X \sim \text{Pois}(\mu + b), \quad Y \sim \text{Pois}(\tau b)
\]
where $\mu$ is the signal rate, $b$ is the background rate and $Pois$ is the usual Poisson distribution. We can assume $X$ and $Y$ to be independent and so

$$P_{\mu,b}(X = x, Y = y) = (\mu + b)^x x! e^{-(\mu + b)} \cdot (\tau b)^y y! e^{-\tau b}$$

2.1 A Confidence Region for $\mu$ and $b$

A common technique for constructing confidence intervals is to find a corresponding hypothesis test and then to invert the test. This is also the approach taken in Cousins and Feldman [1]. We will start with a simultaneous hypothesis test for $\mu$ and $b$ with the null hypothesis $H_0 : \mu = \mu_0, b = b_0$. The steps are as follows:

(i) List (almost) all possible observations $(u_i, v_i), i = 1, ..., K$ assuming $H_0$ is true

(ii) Find the probability for each of these observations by

$$P_{\mu_0,b_0}(X = u_i, Y = v_i) = (\mu_0 + b_0)^{u_i} u_i! e^{-(\mu_0 + b_0)} \cdot (\tau b_0)^{v_i} v_i! e^{-\tau b_0}$$

(iii) Sort all observations from the most likely to the most unlikely

(iv) Find the partial sums from the largest to the $k^{th}$ observation until you reach $1 - \alpha$ if the desired level of the test is $\alpha$

(v) If the observed $(x, y)$ appears in the list of possible observations before
1 − α is reached, accept the null hypothesis, otherwise reject it.

This is in effect a Neyman construction like the one used in Cousins and Feldman [1], only we use the probabilities as the ordering quantities. The test simply checks whether or not the observation \((x, y)\) is compatible with the rates \(\mu_0\) and \(b_0\) specified in the null hypothesis.

The inversion of the hypothesis test then involves a search through all pairs \((\mu, b)\). If a certain pair leads to the acceptance of the null hypothesis, we add it to the confidence region, otherwise we do not. As an example we have figure 1 where we show the confidence regions obtained for three observations. The boundaries of the confidence regions are somewhat ragged due to the discrete nature of the Poisson random variable.

2.2 A Confidence Interval for \(\mu\)

Again we will start with a hypothesis test, this time though we will only fix the signal rate \(\mu\). The null hypothesis then becomes \(H_0: \mu = \mu_0\).

A popular test in Statistics for any kind of hypothesis test is the likelihood ratio test which is based on the likelihood ratio test statistic \(\Lambda\) given in our problem by:

\[
\Lambda(\mu_0; x, y) = \max \{l(\mu_0, b; x, y) : b \geq 0\} \max \{l(\mu, b; x, y) : \mu \geq 0, b \geq 0\}
\]
Here \( l(\mu, b; x, y) = P_{\mu,b}(X = x, Y = y) \) is the likelihood function of \( \mu \) and \( b \) given the observation \((x, y)\). This test statistic can be thought of as the ratio of the best explanation for the data if \( H_0 \) is true and the best explanation for the data if no assumption is made on \( \mu \). The denominator is simply the likelihood function evaluated at the usual maximum likelihood estimator. To find the numerator we have to find the maximum likelihood estimator of the background rate \( b \) assuming that signal rate is known to be \( \mu_0 \).

\[
\frac{\partial}{\partial b} \log l(\mu_0, b; x, y) = x\mu_0 + b - 1 + yb - \tau \overset{!}{=} 0
\]

\[
\hat{b} = x + y - (1 + \tau)\mu_0 + \sqrt{(x + y - (1 + \tau)\mu_0)^2 + 4(1 + \tau)y\mu_0^2(1 + \tau)}
\]

\( l(\mu, \hat{b}; x, y) \) is called the profile likelihood function of \( \mu \). The usefulness of the likelihood ratio test statistic lies in the fact that approximately we have

\[
-2 \log \Lambda(\mu_0; x, y) \sim \chi^2(d)
\]

that is \(-2 \log \Lambda\) has an approximate Chi-Square distribution with \( d \) degrees of freedom, where \( d \) is the number of parameters in the model minus the number of parameters specified in the null hypothesis. Here we have \( d = 1 \). For more details on the likelihood ratio test statistic see Casella and Berger [5]. For
information on the profile likelihood see Bartlett [6] and Lawley [7]. In figure 2 we have the profile likelihood function for the case $x = 6$, $y = 2$, $\tau = 2$.

To find a $(1 - \alpha) \cdot 100\%$ confidence interval we start at the minimum, which of course is at the usual maximum likelihood estimator, and then move to the left and to the right until the function rises by the $\alpha$ percentile of a $\chi^2$ distribution with 1 degree of freedom.

The method here uses an approximation to the profile likelihood function by a quadratic function. Unfortunately, in cases where the number of observations in the signal region is small compared to the number of background events, the profile likelihood function becomes almost linear and this approximation does not work. For those cases we will use the following method:

We will overlay the confidence region described previously with the profile likelihood curve $(\mu, \hat{b})$. Then we find the smallest value of $\mu$ that is on this curve but not in the confidence region. Figure 3 illustrates this method. Clearly, in the case of fewer observations in the signal region than in the background region only an upper bound will be quoted. We will use this second method whenever the profile likelihood function has a positive derivative at $\mu = 0$. It turns out that the limits obtained by these two methods are compatible. Figure 4 shows the upper bound for a number of cases with the limits obtained by the first method drawn as circles and the limits using the second method as squares. The transition from one method to the other is quite smooth.
3 Extensions of the Method

3.1 Estimating background from Monte Carlo

Our method can also be applied when a Monte Carlo with limited statistics is used to estimate the background rate. Assume we run the Monte Carlo $n$ times and observe a total of $y$ events. In the data we see $x$ events in the signal region. Then a probability model for this situation is given by:

$$X \sim Pois(\mu + b), Y \sim Pois(nb)$$

We notice of course that this is actually the same model as the one previously used, only with an $n$ instead of a $\tau$. Therefore, we can use our method for this situation without any changes.

3.2 Include a second background source

Sometimes there is a second source of background present in the data, this one characterized by the fact that it only appears in the signal region. An example is an invariant mass histogram where some of the background comes from the misidentification of pions with muons. Say we run a Monte Carlo $n$ times and observe a total of $z$ events of this type. In the data we have $x$ observations in the signal region and $y$ observations in a suitably chosen background region.
The probability model for this case is given by:

(1)

\[ X \sim Pois(\mu + b + \eta), Y \sim Pois(\tau b) \]

\[ Z \sim Pois(n \eta) \]

where \( \mu \) is the signal rate, \( b \) is the rate of the first background source and \( \eta \) is the rate of the second background source. Then:

\[
P_{\mu,b,\eta}(X = x, Y = y, Z = z) = (\mu + b + \eta)^x x! e^{-(\mu+b+\eta)} (\tau b)^y y! e^{-\tau b (n \eta)} z! e^{-n \eta}
\]

We can extend our method to this situation in a fairly straightforward manner. First we need to find the profile likelihood function, which leads to the following nonlinear system of equations:

\[
\begin{align*}
\partial \log l / \partial b &= x \mu + b + \eta - 1 + y b - \tau = 0 \\
\partial \log l / \partial \eta &= x \mu + b + \eta - 1 + z \eta - n = 0
\end{align*}
\]

It turns out that \( \hat{\eta} \) can be found as the second largest root of the cubic equation
\[ ax^3 + bx^2 + cx + d = 0 \text{ with} \]

\[
a = (1 + n)(n - \tau)
\]

\[
b = (x + z - (1 + n)\mu) (\tau - n) - (1 + n)(z + y)
\]

\[
c = xz + (\tau - n)z\mu + z^2 + yz - (1 + n)z\mu
\]

\[
d = \mu z^2
\]

and then \( \hat{b} \) is given by

\[
\hat{b} = x + y - (1 + \tau)(\mu + \hat{\eta}) + \sqrt{(x + y - (1 + \tau)(\mu + \hat{\eta}))^2 + 4(1 + \tau)y(\mu + \hat{\eta})^2(1 + \tau)}
\]

The same problem as before arises again in this situation: in the case of few events in the signal region the profile likelihood function is nearly linear. We can use the same remedy as before by instead finding the intersection of the boundary of the three-dimensional confidence region in \((\mu, b, \eta)\) space with the profile likelihood curve \((\mu, \hat{b}, \hat{\eta})\).
In this section we will study the true coverage and the power of this method. For comparison we will use the unified approach of Cousins and Feldman [1]; although, of course that method was not designed to deal with uncertainty in the background. The coverage rates shown here were obtained through actual computation, not by Monte Carlo. In figure 5 we have the case of one background region of equal size to the signal region, and finding 90% confidence intervals. The true background rates are $b = 1, 3$ and $5$ and the signal rates go from $0$ to $5$. Clearly our method has much better coverage than Cousins and Feldman [1]; although, due to the approximation used in the method, the coverage is sometimes slightly worse than the nominal one. The ragged appearance of the graph is due to the discrete nature of the Poisson distribution and in general is unavoidable. Figure 6 has the case of two background regions and a 99% confidence interval. Again the new method performs very well whereas Cousins and Feldman [1] does not achieve the nominal coverage.

Next we illustrate the performance of this method when the background rate is estimated by Monte Carlo. In figure 7 we compute the coverage rates for the cases $\mu = 0$ and $b = 0.5, 1.5$ and $2.5$. The Monte Carlo is run $n$ times were we let $n$ range from $1$ to $10$. Our method performs satisfactorily for all cases. This plot is also interesting because it gives some insight into the number of runs of the Monte Carlo needed before one can assume that the background
rate is known.

The performance of the extension of our method to the case of two different background sources discussed in section 3.2 has to be studied using mini Monte Carlo because the number of possible observations \((x, y, z)\) becomes quite large. We have run a variety of these mini Monte Carlo studies. In Figure 8 we show the results for the case \(\tau = 2\), \(n = 10\) and a 90% confidence interval. The true coverage rates of our method appear to be in line with the nominal ones, again with a few cases where the coverage is slightly worse due to the approximation used in the method. Due to the discreteness of the Poisson distribution the method is also quite conservative for many situations.

5 Conclusion

The methods for quoting limits for rare decays, mainly Cousins and Feldman [1] and its variants, all suffer from the requirement that the background source be known with a high degree of precision. We have described a new method based on the likelihood ratio test which treats the background uncertainty as a statistical error. The performance of the method is shown to be quite good, with the true coverage rates close to the nominal ones and good power. The method can be used in situations where the background rate has been estimated from data sidebands as well as where it has been obtained from Monte Carlo. The method can also be extended to cases where a second background
source solely appearing in the signal region is present.

In the appendix we provide tables for the confidence intervals in the cases \( \tau = 1, 2 \) and \( \alpha = 0.9, 0.99 \). A FORTRAN program for the computation of the limits for any other case as well as for the extension discussed in section 4.2. can be obtained by writing to w_rolke@rumac.upr.clu.edu.

References


