Setting Limits for Poisson Rates in the Presence of Noise [1]

Wolfgang A. Rolke

Abstract

We discuss the recent progress in the treatment of the rate of a Poisson random variable when the data is also contaminated by a background source. We discuss the method by Feldman and Cousins and some of its variants, and we describe a new method for setting limits in the situations where the background rate is unknown. This new 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.

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

[1] 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.

1 Introduction

High Energy Physics experiments provide statistical problems of a wide variety, and over many magnitudes of data size. For example, the FOCUS group experiment E831 at the Fermi National Accelerator Laboratory has yielded a data set of approximately 20 terabytes, and yet some of the analysis of rare decays is in the end based on only a few, or sometimes even zero, events. In this paper we are concerned with the problem of setting limits on the rates at which such rare events are generated, a problem made more difficult by the fact that not all observed events are of the desired kind but are contaminated by background.

Finding confidence intervals or upper limits for small signals has recently attracted a great deal of attention. The paper by Feldman and Cousins [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 an ordering principle based on likelihood ratios. Subsequent papers such as Giunti A [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, these methods can fail when used in the presence

of background uncertainty. Possible remedies are discussed in Cousins and Highland [4] where it is suggested to treat the background uncertainty as a systematic error, and in Giunti B [5] ,where the background is assumed to have a Gaussian distribution. 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 have previously shown in Rolke and Lopez [6] that this method yields the correct coverage rate and that it has 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 Feldman and Cousins and Variants

Consider the following situation: we observe n events. Some of these may have been generated by the mechanism of interest, which we will call the signal, others may come from background contamination. Both the signal and the background are generated according to the Poisson distribution, with rate μ for the signal and rate b for the background. This leads to the following

probability model:

$$P_{\mu,b}(N = n) = \frac{(\mu + b)^n}{n!} e^{-(\mu+b)}$$

Here we will assume that the background rate b is known, and we want to find a confidence interval for the signal rate μ.

Until recently the commonly accepted methods for quoting limits in this situation were based on Bayesian statistics, with the usual difficulties of agreeing on a suitable prior distribution. In order to remedy this problem, Feldman and Cousins [1] developed a purely Frequentist method. Their method has the desirable property that it always gives confidence intervals in the physical region, that is, it does not give negative answers for parameters that are clearly non-negative. In addition they solved the problem of "flip-flopping", that is the practice of deciding to quote a two-sided confidence interval or mearly an upper limit after seeing the limits. Flip-flopping is known to effect the actual coverage rate, and is therefore undesirable.

Feldman and Cousins [1] method is based on the Neyman construction of confidence intervals, which has been a mainstay of Statistics for a long time. The ordering of the outcomes is done according to the likelihood ratio given by

$$R = \frac{P_{\mu,b}(n)}{P_{\widehat{\mu},b}(n)}$$

where $\widehat{\mu}$ is the the maximum likelihood estimator of μ given by $\widehat{\mu} = \max(0, n-$

b).

As Feldman and Cousins [1] noted, there has been some criticism of their limits in cases where an experiment finds fewer events than are expected from the background rate. For example, an experiment with an expected background rate of $b = 0.5$ and no observed events would quote a Feldman-Cousins upper limit of 1.94; whereas an experiment with an expected background rate of $b = 1.5$ and no observed events would quote a Feldman-Cousins upper limit of 1.33. It is somewhat counterintuitive that an experiment with a higher background level should be able to quote a lower limit. This apparent paradox stems from a misinterpretation of frequentist confidence intervals as Bayesian credible intervals, that is, as statements about $P\,(\text{Theory}\,|\text{Data}\,)$ rather than what they are, namely statements about $P\,(\text{Data}\,|\text{Theory}\,)$. Some papers have tried to correct this perceived problem, for example Giunti A[2] and Roe and Woodroofe [3].

Feldman and Cousins [1] is the standard method used in calculating confidence intervals in High Energy Physics at this time as recommended by the Review of Particle Physics [7].

3 A Description of our new method

In this section we will outline the basic ideas of our 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 τ. 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

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

where μ 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

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

As part of our method we construct a two-dimensional confidence region for μ and b simultaneously by inverting the hypothesis test for μ and b with the null hypothesis $H_0 : \mu = \mu_0, b = b_0$. This is in effect a Neyman construction like the one used in Feldman and Cousins [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 μ_0 and b_0 specified in the null hypothesis.

Using the likelihood ratios as the ordering quantity in a way similar to Feldman and Cousins [1] would in principle be superior, but unfortunately does not yield a viable method here because the list of "likely" observations is infinite.

In figure 1 where we show the confidence regions obtained for three observa-

tions. The boundaries of the confidence regions are somewhat ragged due to the discrete nature of the Poisson random variable.

Next we will find a confidence interval for the signal parameter μ alone, again by inverting a hypothesis test. The null hypothesis of the test is $H_0 : \mu = \mu_0$. As the test statistic for the test we will use the likelihood ratio test statistic Λ given in our problem by:

$$\Lambda(\mu_0; x,y) = \frac{\max\{l(\mu_0, b; x, y) : b \geq 0\}}{\max\{l(\mu, b; x, y) : \mu \geq 0, b \geq 0\}}$$

Here $l(\mu, b; x, y) = P_{\mu,b}(X = x, Y = y)$ is the likelihood function of μ 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 μ. 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 the signal rate is known to be μ_0:

$$\frac{\partial}{\partial b} \log l(\mu_0, b; x, y) = \frac{x}{\mu_0 + b} - 1 + \frac{y}{b} - \tau \doteq 0$$

$$\widehat{b} = \frac{x + y - (1+\tau)\mu_0 + \sqrt{(x + y - (1+\tau)\mu_0)^2 + 4(1+\tau)y\mu_0}}{2(1+\tau)}$$

$l(\mu, \widehat{b}; x, y)$ is called the profile likelihood function of μ. 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 $d = 1$. Because the profile likelihood l differs from the likelihood ratio Λ only by a constant independent of μ_0, we can concentrate on the profile likelihood. For more details on the likelihood ratio test statistic see Casella and Berger [8]. For information on the profile likelihood see Bartlett [9], Lawley [10] and Murphy and Van Der Vaart [11]. 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 α percentile of a χ^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 overlay the confidence region described previously with the profile likelihood curve $\left(\mu, \widehat{b}\right)$. Then we find the smallest value of μ 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 $\mu = 0$.

4 Extensions of the Method

4.1 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 τ. Therefore, we can use our method for this situation without any changes.

4.2 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 :

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

$$Z \sim Pois(n\eta)$$

where μ is the signal rate, b is the rate of the first background source and η is the rate of the second background source. We can extend our method to this situation in a fairly straightforward manner by again using the profile likelihood to eliminate the nuisance parameters b and η. 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 (μ, b, η) space with the profile likelihood curve $(\mu, \widehat{b}, \widehat{\eta})$.

A more detailed description as well as results of computations of the actual coverage rates of our method for a variety of situations can be found in Rolke and Lopez [6]

5 Conclusion

The number of new methods for quoting limits for rare decays have appeared in the last two years, mainly Feldman and Cousins [1] and its variants. All of

these methods require a precise knowledge of the background rate. We have added to this field a new method based on the likelihood ratio test which treats the background uncertainty as a statistical error. The performance of the method was 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.

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

[1] R.D. Cousins, G.J. Feldman, "A Unified Approach to the Classical Statistical Analysis of Small Signals", *Phys. Rev*, **D57**, (1998) 3873.

[2] C. Giunti, "A new ordering principle for the classical statistical analysis of Poisson processes with background" , *Phys. Rev* **D59**, 053001 (1999).

[3] B.P. Roe, M.B. Woodroofe, "Improved Probability Method for Estimating Signal in the Presence of Background", *Phys. Rev* **D60** 053009 (1999).

[4] R.D. Cousins, V.L.Highland, "Incorporating systematic uncertainties into

an upper limit", *Nucl. Inst. and Methods A320*, (1992) 331-335.

[5] C. Giunti, "Treatment of the background error in the statistical analysis of Poisson processes", *Phys. Rev* **D59**, 113009 (1999).

[6] W.A. Rolke and A.M. Lopez, "Confidence Intervals and Upper Bounds for Small Signals in the Presence of Background Noise", to appear in *Nuclear Instruments and Methods A*, hep-ph/0005187.

[7] C. Caso *et al.* (Particle Data Group), Eur. Phys. J. C **3**, 1 (1998) 177.

[8] G. Casella, R.L. Berger, *Statistical Inference*, Duxburry Press, (1990) 346.

[9] M.S. Bartlett, "Approximate Confidence Intervals, Part II: More than one Unknown Parameter", *Biometrica* **Vol. 40**, (1953) 306-317.

[10] D.N. Lawley, "A General Method For Approximating To The Distribution Of The Likelihood Ratio Criteria", *Biometrica* **Vol. 43**, (1956) 295-303.

[11] S.A. Murphy, A.W. Van Der Vaart, "On Profile Likelihood", *Journal of the American Statistical Association*, **Vol**. **95**, (2000), 449-485.

6 Appendix

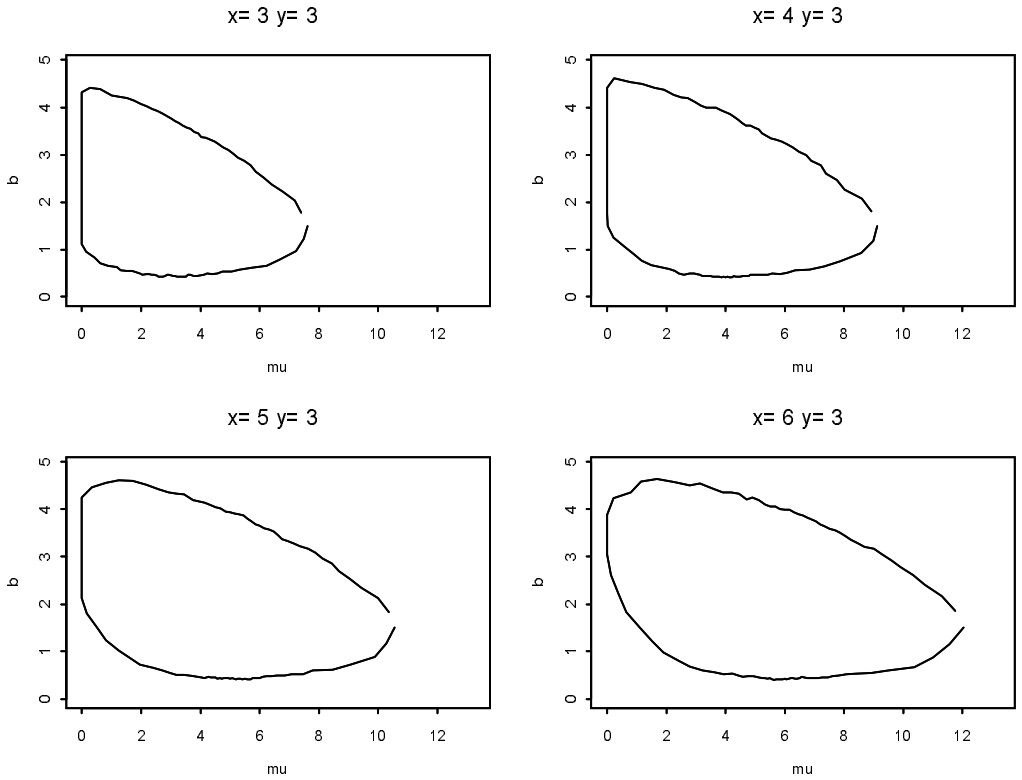

Figure 1: Two dimensional confidence regions for $x = 3, 4, 5$ and 6 signal events, and $y = 3$ background events. The background region is twice the size of the signal region, and a 90% confidence level was used.

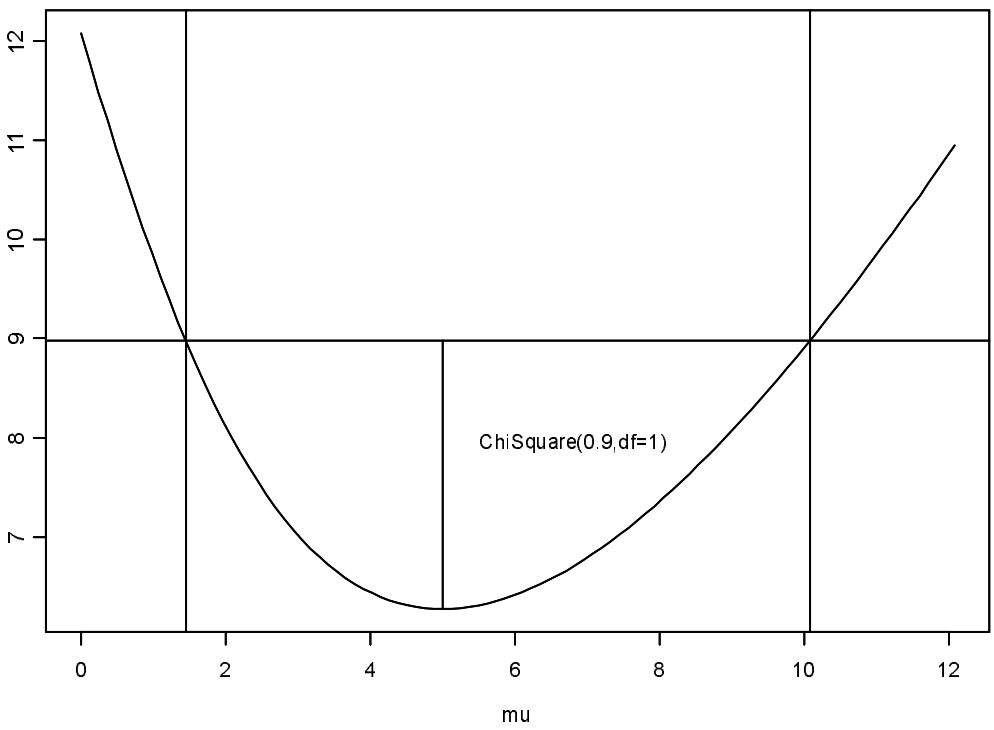

Figure 2: Profile likelihood function with two-sided limits for the case of $x = 6$ events in the signal region and $y = 2$ events in the background region. The background region is twice the size of the signal region ($\tau = 2$). The nominal coverage probability is 0.9.

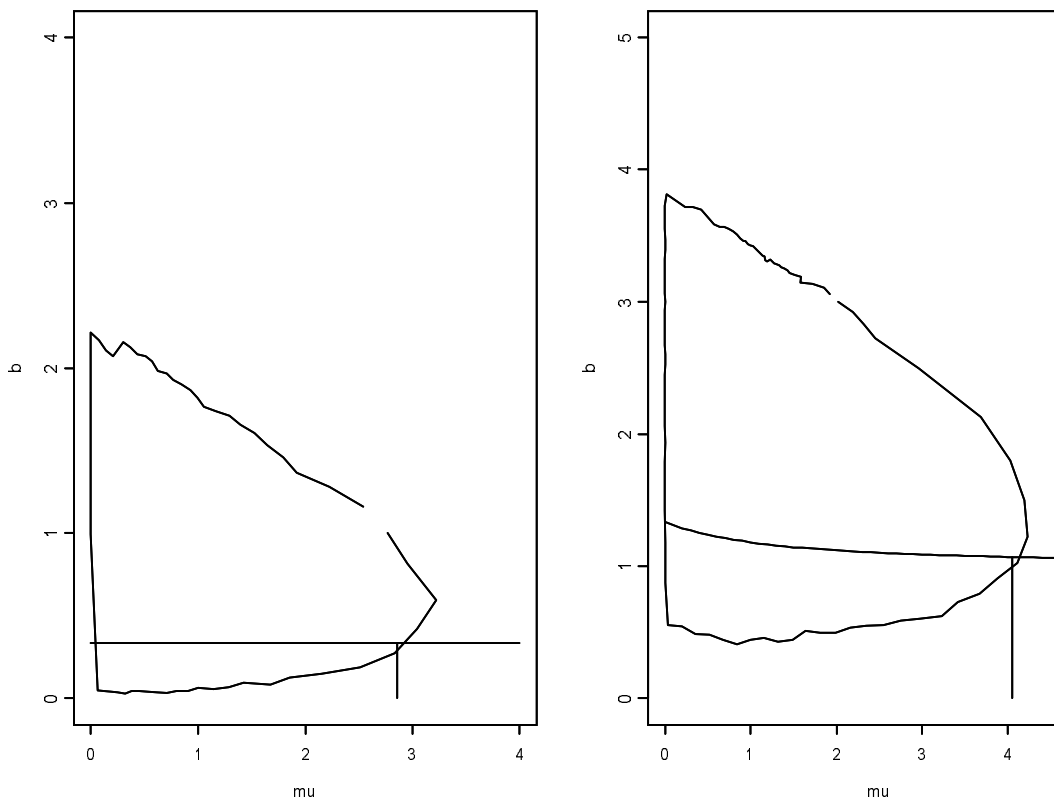

Figure 3: Confidence region with profile likelihood curve and upper bounds for the cases $x = 0, y = 1$ and $x = 1, y = 3$.