Bias-Corrected Confidence Intervals for Rare Searches.

Wolfgang A. Rolke and Angel M. López

Abstract

We discuss the problem of selecting an optimal cut and then finding confidence limits based on that cut. Even when performing a blind analysis this may lead to a bias in the limits because the cut selected may be optimal for a real or simulated data set with limited statistics and thus may be sensitive to fluctuations in this data set. We propose to use a variant of the bootstrap to find the limits. A Monte Carlo study shows that these new limits have correct coverage.

1. Introduction

In recent years researchers in high energy physics have come to realize that some of the standard analysis techniques carry with them the danger of introducing a bias. One major example is the search for a cut combination which eliminates background events but retains as many signal events as possible. This can lead to using a cut combination which is optimal for the data set at hand, but not for the process that generated the data set, and that therefore depends too much on random fluctuations in the data. A good way to lower the risk of introducing this bias is by performing a blind analysis, whereby the choice of cut is solely based on the background sidebands or on simulated background. It should be clear, though, that this only alleviates half the problem: we no longer are subject to upfluctuations in the signal region, but we are still in danger of optimizing our cut on a downfluctuation in the sideband or simulated background.

The source of the remaining problem can be understood to be the fact that we are using the same data to do two things: to choose the cut set and to estimate the background level. One attempt at dealing with this is to use split samples: randomly divide the data into two parts, use one part to find the optimal cut combination, use the other to find the limits. This approach has two major flaws: first there is the question of what the relative size of the two parts should be, and then there is the problem that the parts have even lower statistics than the whole, thus worsening the effects of fluctuations. In this paper we propose to use a different method to minimize this bias. Our method is a variant of a well known technique in Statistics, the bootstrap. We will show that this method has neither of the two problems of split samples and that it leads to unbiased estimates of the branching ratio.

When we choose a cut combination we need an optimality criterion. In this work we use the experimental sensitivity, a variable which is derived from the sensitivity defined in Feldman and Cousins [1] and in Review of Particle Physics [2]. The experimental sensitivity is defined as the average of the upper limits that would be quoted for an ensemble of experiments with no true signal. It can be thought of as a measure for the size of an effect that could be discovered by a certain experiment. The smaller the experimental sensitivity of a cut set, the more likely we are to discover a signal that is truly present.

2. Why there might be a bias

The bias introduced by combining a minimization and an evaluation step into one procedure is actually quite common. As one example, consider the problem of fitting a parametric curve to a histogram. Here we usually start by estimating the parameters of the parametric function to be fit, for example by finding the estimates of the parameters that yield the lowest χ^2. Then we want to know whether our fit is sufficiently good, so we proceed to find the confidence level of the χ^2 statistic. But in fact the χ^2, and therefore the confidence level, will be biased because the parameter estimates were chosen to make the χ^2as small as possible. In the next section we will give the results of a Monte Carlo study that shows the presence of this type of bias in the search for small signals.

Of course we have known for almost a century how to adjust for this bias in the case of the χ^2, namely by adjusting the degrees of freedom of the χ^2 distribution. Unfortunately, in general it is very difficult to find this kind of an analytic correction.

One important question is whether this bias is large or small compared to the statistical error. In the analysis of small signals the end result is typically a confidence limit, either a two sided confidence interval or just an upper limit. Whether or not a method to compute confidence intervals works correctly has to be judged solely based on the true coverage rate of the limits. If two or more methods with correct coverage are available, then one may use other criteria to make the choice of method. For example, in physics one might prefer to use a method that never yields an empty interval, or one might prefer a method that yields on average the shortest intervals. Such a choice has to made before examining the data, of course.

We will use the method of Rolke and López [3] to compute the confidence intervals. This is the only method currently known that treats the uncertainty in the background rate as a statistical error. Feldman and Cousins [4] independently solved this problem, and proposed a modification to their solution. Like Feldman and Cousins [1] the Rolke-Lopez method solves the "flip-flop" problem, and it always results in physically meaningful limits. The problem described here as well as its solution, though, do not depend on what method of computation is used for either the sensitivity or the limits. As long as there is some uncertainty in the background rate the bias would be equally present if we had used for example Feldman and Cousins [1] or a Bayesian method.

To get an idea of the size of the bias we performed a Monte Carlo study of the analysis of the $D^0 \to \mu^+\mu^-$ decay using data from FOCUS [5]. One problem in doing this MC is obtaining a large sample of background events. In our study this sample was obtained by assuming the background was due to other particles being misidentified as muons.

Fake data sets were generated by randomly choosing M events from the simulated signal set and N events from the background set. For the purposes of this study we chose the number of background events N from a Poisson distribution with rate 16 as in the real dimuon data, and M was chosen from a Poisson distribution with rate λ, where λ was varied from 0 (meaning no signal was present) to 6. For each value of the rate λ we generated 5000 fake data sets in this manner. To each of these data sets we applied each of 13122 cuts. The cuts used for this simulation were the same cuts that had previously been chosen as appropriate for this analysis. The cut that had the lowest experimental sensitivity was then applied to the signal region and the Rolke-López method was used to find the corresponding confidence limits. Finally those 5000 confidence limits were used to calculate the true coverage rates. To make sure that any observed bias is really due to the minimization-estimation problem, we also randomly chose 9 individual cuts and always applied those same cuts to the fake data. Clearly in this case no minimization takes place, and so there should be no bias.

The results of this MC study are shown in figure 1. As expected the limits for the individual cuts have correct coverage, with the true coverage not dropping much below the nominal rate of 0.9. That a few of the coverage rates on the right side of the graph are below the 0.9 line is due to random fluctuations in the MC as well as the discrete nature of the Poisson distribution. The apparent drop in the coverage rates from the left to the right does not continue, with the rates for $\lambda's$ larger than 6 all just above 0.9. This was verified by running the MC for various values of λ up to $\lambda = 15$.

Correct coverage is not the only characteristic a good method should have. It is also important to obtain the strictest limits possible. That is what the minimum sensitivity cut methodology attempts to do but using the limits from this method results in true coverage rates well below the nominal rate. The graph is based on just 15 different values for λ, and it would be pure coincidence if the lowest true coverage were obtained for one of those values. Therefore the worst coverage should be expected to be well below the worst one observed of about 0.845. We can therefore conclude that we have a sizable bias in our confidence limits due to the cut selection bias.

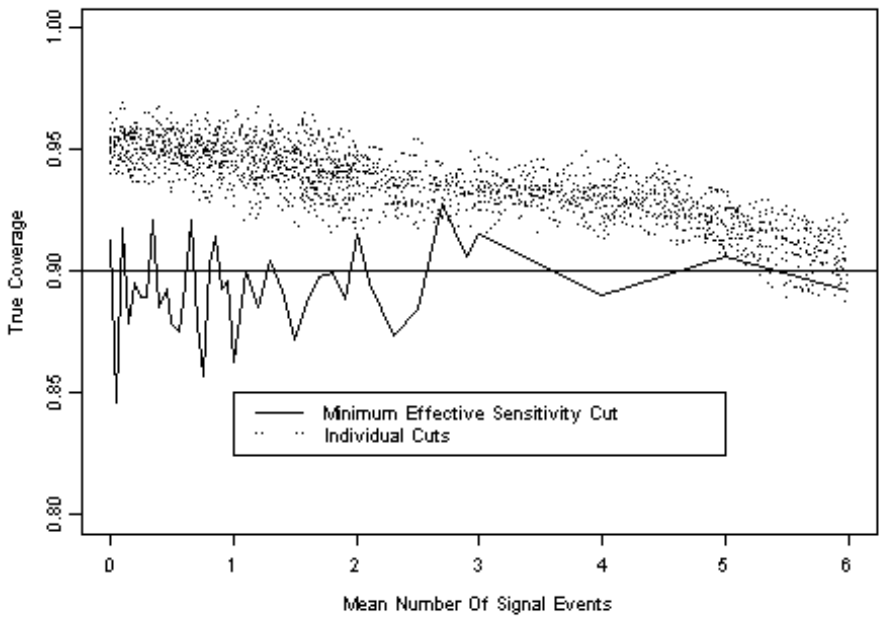

Fig. 1: True coverage rates for individual cuts (dotted line) and for minimum sensitivity cut (solid line). The rates for the individual cuts are always larger than the nominal rate as is desired, wheras the true coverage for the minimum sensitivity is often well below the nominal rate of 0.9.

3. The Bootstrap

The statistical bootstrap method is a non-parametric alternative for finding error and bias estimates in situations where the assumption of a Gaussian distribution is not satisfied and where it is difficult or even impossible to develop an analytic solution. In this section we will show the reasoning behind the bootstrap method and how it is applied in practice.

Let us assume we are interested in estimating a certain parameter θ such as the width of a signal or a branching ratio. Let us also assume that we have observations $X_1, .., X_n$ from a distribution F that depends on θ. Furthermore we have a method for finding an estimate $\widehat{\theta}$ of θ, say $\widehat{\theta} = T(X_1, .., X_n)$. The estimator T might be as simple as computing the mean of the observations or as complicated as fitting a Dalitz plot.

Now, in addition to $\widehat{\theta}$ we will also need an error estimate as well as an idea of the bias in the estimator T. If T is fairly simple we might be able to find its distribution and get an error and a bias estimate analytically. If the situation is more complicated we might instead try a Monte Carlo study. To do this we would simulate sampling from the distribution F, generating many (say k) independent samples of size n, apply the estimator T to each and thereby get a sample of estimators $\widehat{\theta}_1, .., \widehat{\theta}_k$. Then we can look at a histogram of the estimators, compute their standard deviation, and so on.

But what can we do if we do not know the distribution F? In that case the data $X_1, .., X_n$ is all we have, and any analysis has to be based on these observations. The best estimate of the distribution function $F(x)$ is the empirical distribution function $\widehat{F}(x)$ given by $\widehat{F}(x) = \frac{1}{n} \cdot$(number of observations $X_i \leq x$), $x \in \Re$, that is the percentage of events smaller than x. The basic idea of the bootstrap is to replace the distribution F in the MC study above by its empirical distribution function $\widehat{F}$.

It can be shown that sampling from the empirical distribution function means sampling *with re-*

placement from the observations $X_1, .., X_n$. A bootstrap sample has the same sample size as the original data. It is made up out of the original observations, some of which might appear more than once whereas others might not be included at all. As in the MC study, we will draw many (say B) of these bootstrap samples, apply the estimator T to each of them and thereby get bootstrap estimates $\widehat{\theta}^*_1, .., \widehat{\theta}^*_B$ of θ. We can then study these bootstrap estimates to get an idea of the error and the bias of T.

The bootstrap method as described above was first developed by B. Efron in [6]. Since then a great deal of theoretical work has been done to show why and when the bootstrap method works, see for example Hall [7], and it has been successfully used in a wide variety of areas. Previous applications of the bootstrap in High Energy Physics can be found in Hayes, Perl and Efron [8] and in Alfieri et al. [9]. For a very readable introduction to the subject see Efron and Tibshirani [10].

4. The Dual Bootstrap and Bias Corrected Limits

As discussed in the introduction, the usual approach for dealing with this bias, namely the split sample method, has some serious problems of its own. Instead we will use a variant of the bootstrap: we will draw one bootstrap sample from the data and find the cut with the smallest sensitivity for this bootstrap sample, then we will draw another bootstrap sample, independent from the first, to find the limits. This procedure will then be repeated B times, with a B of about 5000. In this manner we will get B lower and upper limits. Finally we will use the median of the lower and the median of the upper limits as our estimates. We use the median because it is less sensitive than the mean to a few unusually large observations. Also, in the case where the signal rate is zero, if even a few of the B bootstrap estimates of the lower limit are positive, the mean would also be positive (and wrong), whereas the median is still zero (and therefore correct).

In this way for each bootstrap sample we get a cut set that is optimal for the first bootstrap sample but not necessarily for the second, which is representative of the underlying distribution. We can therefore expect to get unbiased estimates for the limits or, in other words, limits with the correct coverage rate.

We repeated the MC study discussed in section 2, now using the dual bootstrap method. Figure 2 shows that the dual bootstrap method yields limits with the correct coverage, effectively removing the minimization-evaluation bias. Similar MC studies with different nominal coverage rates and different background rates, both smaller and larger than the rate of 16 shown here, have confirmed this conclusion.

5. Conclusion

We have shown that performing a blind analysis is not enough to eliminate the cut selection bias. A MC study of the $D^0 \rightarrow \mu^+\mu^-$ decay indicates that this bias is quite large. We have developed a method based on the bootstrap technique from Statistics that corrects for this type of bias. A MC study for the $D^0 \rightarrow \mu^+\mu^-$ decay shows that this new method performs very well.

FORTRAN routines for the dual bootstrap method as well as for computing the Rolke-López limits are available from the authors by sending an email to w_rolke@rumac.uprm.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. Caso *et al.* (Particle Data Group), Eur. Phys. J. C **3**, 1 (1998) 177.

[3] W.A. Rolke, A.M. López, "Confidence Intervals and Upper Bounds for Small Signals in the Presence of Background Noise", *Nucl. Inst. and Methods A458* (2001) 745-758

[4] G. Feldman, "*Multiple measurements and parameters in the unified approach*", talk at Fermilab Workshop on Confidence Limits 27-28 March, 2000, http://conferences.fnal.gov/cl2k/ , p.10-14.

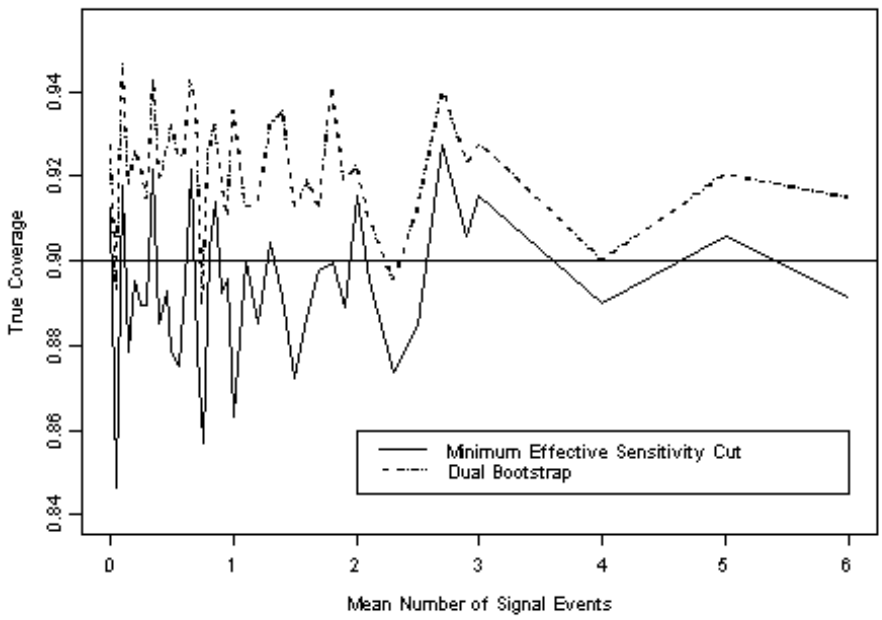

Fig. 2: True coverage rates for the limits of the signal rate based on the cut set with the smallest sensitivity (solid line) and for the dual bootstrap method. Clearly the dual bootstrap correctly adjusts the limits and yields correct limits with the correct coverage rate.

[5] J. M. Link et al., *Proceedings of Heavy Quarks at Fixed Target 1998, AIP Conf. Proc. 459,* eds. H. W. K. Cheung and J. N. Butler (1998) 261

[6] B. Efron, "Bootstrap methods: another look at the jacknife", *Ann. Statistics*, **7**, (1979) 1-26.

[7] P. Hall, *The Bootstrap and Edgeworth Expansion*, Springer Verlag, (1992)

[8] K.G. Hayes, M. L. Perl, B. Efron, "Application of the Bootstrap Statistical Method to the Tau Decay Mode Problem", *Phys.Rev*. **D39**: (1989) 274

[9] R. Alfieri *et al.*, "Understanding stochastic perturbation theory: toy models and statistical analysis", hep-lat/0002018 (2000)

[10] B. Efron, R.J. Tibshirani, *An Introduction to the Bootstrap*, Chapman & Hall, (1993)