A Calculator for Confidence Intervals

Roger Barlow

Department of Physics
Manchester University
England

Abstract

A calculator program has been written to give confidence intervals on branching ratios for rare decay modes (or similar quantities) calculated from the number of events observed, the acceptance factor, the background estimate and the associated errors. Results from different experiments (or different channels from the same experiment) can be combined. The calculator is available in `http://www.slac.stanford.edu/~barlow/limits.html`

PACS code 13.85.Rm

Keywords: Limits. Systematic errors. Combination of results. Rare processes.

1. Introduction

Many particle physics experiments measure a number n of events of a particular type, and use it to set limits on a physical quantity R, typically a cross section or branching ratio. R and n are related by various factors such as the efficiency, dead-time, backgrounds, beam luminosity, and duration of the experiment.

For a Poisson process, the production of confidence limits on the mean μ from the observed n has been well studied [1,2]. Tables exist, or standard numerical techniques can be used, to solve exactly the desired probability equalities

$$\sum_{r=0}^{n} P(r; \mu_+) = \alpha \qquad \sum_{r=0}^{n-1} P(r; \mu_-) = 1 - \alpha \tag{1}$$

for the upper limit μ_+ and the lower limit μ_- at the $(1 - \alpha)$ confidence level. If the factor for efficiency etc., collectively known as the sensitivity S, is known exactly, then in the absence of background these limits translate directly into limits on R using

$$\mu = RS \tag{2}.$$

If the expected background b is known exactly then it can be subtracted and the confidence limits on R are given by

$$R_\pm = \frac{1}{S}(\mu_\pm - b) \tag{3}.$$

This note considers cases where there are uncertainties σ_S and σ_b in S and b. It follows the approach of Cousins and Highland [3]. Exact calculation is not possible, but a Monte Carlo technique can readily be used. A trial value for the signal is taken, then repeatedly it has a background mean drawn from a Gaussian of mean b and standard deviation σ_b added to it, is multiplied by a Sensitivity drawn from a Gaussian of mean S and standard deviation σ_S, and is then used as the mean of a Poisson distribution from which a number of events r is generated. By counting the fraction of cases in which r is less than (or equal to) the observed number n, the probabilities $1 - \alpha$ (or α) of Equation 1 are given.

2. Philosophy

This is basically a conventional frequentist limit. There are no prior assumptions about R. However uncertainties in S and b introduce a Bayesian viewpoint if (as they often do) they include factors like the accuracy of a theoretical model. Such errors are often admittedly not objectively defined: the experimenter makes statements like 'we believe this is good to 10%'. Then the statement "$S = 0.50 \pm 0.05$", if it is taken to mean that the sensitivity has a 68% probability of lying between 0.45 and 0.55, is 'unscientific' from a strict frequentist viewpoint [3].

This frequentist/Bayesian admixture is general and inevitable. It is defensible on the grounds that the main source of uncertainty – the Poisson statistics of the number of events – is treated in an objective frequentist fashion. The errors on the factors are small and have small consequences. The ambiguity due to prior distributions are merely one more uncertainty among many.

3. Results from differing priors

In interpreting a result one naturally writes

$$R = An \tag{4}$$

where the 'appropriate' factor A is the inverse of the sensitivity S. This was done by the BaBar Statistics Working Group in their report for the collaboration [5].

It was found that probabilities obtained using the formalism of equation (2) and equation (4), smearing the factor S or A by the same relative amount, give slightly different results. Although one is merely the inverse of the other, the relationship is nontrivial. A Gaussian distribution for A is not equivalent to a Gaussian distribution for S. The experimenter may have an *a priori* reason for believing that one or the other has a Gaussian uncertainty, however this is not the case in general and the choice is usually merely one of convenience.

A statistician would say [6] that the appropriate technique is to use an 'invariant' prior according to the suggestion of Jeffreys, i.e. to work with a uniform prior in the quantity for which the Fisher information is constant. For a scale factor A this means a prior uniform in $\ln A$ or, equivalently, $\ln S = -\ln A$.

For example, with no error on the sensitivity the probability of a signal 5.0 giving a result of 3 events or less is 26.5%. With increasing uncertainty the probability rises, but by slightly different amounts.

% error on S	Cousins & Highland	Jeffreys	BaBar SWG
0	0.265	0.265	0.265
10%	0.272	0.269	0.266
20%	0.291	0.277	0.267
30%	0.316	0.291	0.273

Table 1: Probability for $n \leq 3$ with increasing uncertainty in the acceptance

Conversely, with 30% uncertainty the limit of 5.00 gives a 29.1% probability with a Jeffreys prior; to get the same probability with the Cousins and Highland hypothesis the limit has to be raised to 5.22, whereas for the BaBar SWG hypothesis it falls to 4.86. These differences are not large (and 30% is somewhat extreme) but this ambiguity does affect values at the level of precision with which they are generally quoted, so it cannot be ignored. All three values are given by the program, so the user can choose which to use (indeed, they are forced to be aware of the choice).

4. Combining results

When different experiments, or different channels in the same experiment, give limits, it is desirable to combine them. There is a possible ambiguity in the procedure. For, say, an upper limit from a single experiment one determines a value R_+ such that the probability of observing n events or less is equal to the desired (small) level, as in Equation 1. But

2

for a double channel, with results (n_1, n_2), the expression 'less' is undefined: is $(4, 1)$ less than $(3, 3)$?

It could be suggested that 'less than' for a pair signify that both values should be less, or either value. However these are clearly unworkable: if the two experiments are identical (in luminosity, efficiency, and background) then there is clearly no information obtainable from the distribution between the two: $n_1 + n_2$ is a sufficient statistic for (n_1, n_2), and 'less than' can be defined with reference to the sum of the two results. Any definition of 'less than' must satisfy this.

We propose that 'less than' be defined for a multiple result using the value of the Branching ratio that one would deduce from it. This is given by a maximum likelihood estimator as ([7], Equation 129)

$$\sum_i \frac{n_i S_i}{S_i R + b_i} - S_i = 0 \qquad (5)$$

where S_i is the sensitivity factor for experiment i, and b_i is the expected background. Sets of results can be ranked according to the value of R given by this equation. Thus to run the simulation for several experiments one can first solve Equation (5) iteratively to find R_{data} for the data values, and then for a given limit R_+ repeatedly generate random sets of result values $\{n_i\}$, calculate the resulting R, and count the fraction of cases that $R \leq R_{data}$. In practice this can be achieved without an iterative solution for each result, as the quantity in Equation 5 is a monotonically decreasing function of R, so having found R_{data} then for each set of results one merely calculates the quantity

$$Q = \sum_i \frac{n_i S_i}{S_i R_{data} + b_i} - S_i \qquad (6)$$

and the sign of Q gives the sign of $R - R_{data}$.

This goes over to the conventional case if all experiments have the same S_i and b_i.

5. Usage

The user is initally presented with a screen for a single experiment (Figure 1.) The number of observed events is entered, as are the expected background and its error (which, like all errors, is given as an absolute value), the overall sensitivity S with its error, the number of Monte Carlo events to be used in the evaluation, and a guess for the limit. A button is then pressed to run the Monte Carlo simulation, and evaluate the probability for either an upper or a lower limit. (The difference is merely that one includes the $r = n$ result in the probability and the other does not, according to the two parts of Equation 1.)

Figure 1: Example of use with a single result

The probability for this limit, under the three different assumptions about priors, is displayed. The user can then iterate towards whatever limit type they choose, e.g. for a 90% upper limit they adjust the limit guess until the probability is shown as 0.10; for a 90% lower limit they would aim for 0.90. Only the 'Limit guess' data entry field need be modified for each trial, although as the limit is approached greater accuracy may be desired and the number of Monte Carlo events increased accordingly. (The calculation takes a few seconds for 1,000,000 events). In the example shown, 3 events were seen with zero background and no sensitivity uncertainty. With 10,000 Monte Carlo events the limit is quickly established at around 6.7. Increasing the number to 1,000,000 establishes it as 6.68 (which could have been obtained from tables anyway).

It is helpful to remember that the Poisson mean used is the product of the limit guess and the (smeared) sensitivity factor. It may be convenient to work in 'event-count' units, taking the sensitivity as unity (with an error which is the the proportional error). The resulting limit is equivalent to a number of events and the sensitivity factor is applied afterwards: if in the above example the data sample corresponds to $20fb^{-1}$ and the efficiency is 100% this establishes the limit as $6.68 \div 20 = 0.334fb$. Alternatively the Sensitivity factor could have been entered directly as 20, and the limit guess found to be 0.334.

If the efficiency were, say, $50 \pm 5\%$ this could be accommodated either by setting the sensitivity error to 0.1 and dividing the resulting limit (risen to 6.81, using the Cousins and Highland prior) by a further 0.5, to give $0.681fb$, or by changing the sensitivity to 10 with an error of 1 and extracting a limit of 0.681 directly.

If the background is estimated as 0.5 ± 0 then the limit falls from 6.81 to to 6.31. For a background of 0.5 ± 0.2 it falls to 6.29.

If further experiments are desired (up to a limit of 10) there is a button to add data fields for them. There is now also a specific factor and error for each experiment, and this is entered in a further two columns, as shown in Figure 2.
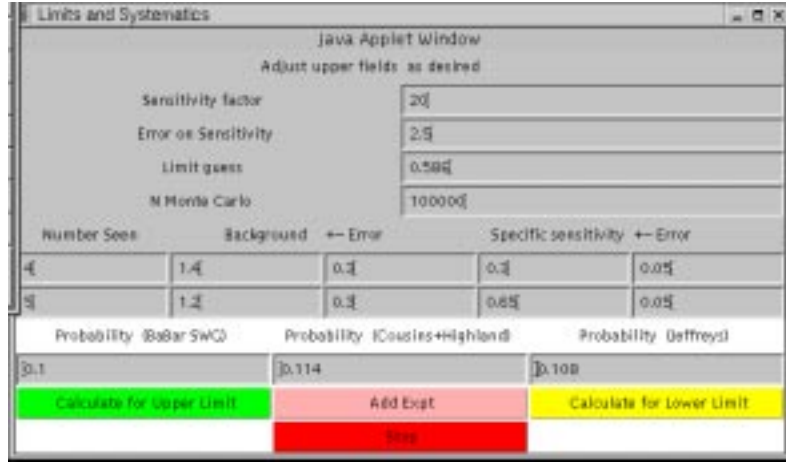
4

Limits and Systematics

Java Applet Window

Adjust upper fields as desired

Sensitivity factor	20
Error on Sensitivity	2.5
Limit guess	0.586
N Monte Carlo	100000

Number Seen	Background	←Error	Specific sensitivity	←Error
4	1.4	0.1	0.1	0.05
5	1.2	0.1	0.85	0.05

Probability (BaBar SWG)	Probability (Cousins+Highland)	Probability (Jeffreys)
0.1	0.114	0.108

Calculate for Upper Limit	Add Expt	Calculate for Lower Limit
	Stop	

Figure 2: Example of use with two results

For multiple results 'event count' units are not workable. The sensitivity is split into an overall value and a specific factor for each channel: the Poisson mean for an experiment is the product of the three values entered for the limit, the smeared overall sensitivity, and the smeared specific sensitivity.

The ambiguity in the factors (if the overall sensitivity is scaled up and each channel's fraction scaled down by the same amount this makes no difference) is there for convenience in the handling of errors. The common acceptance is varied together for all channels, while the specific factors vary independently. For example, if a single experiment is measuring the rate of K_s^0 production, using both the $\pi^+\pi^-$ and $\pi^0\pi^0$ decay modes, then the overall luminosity figure and global trigger efficiency (with associated errors) form the common value. The two specific factors are formed from the two branching ratios and the individual reconstruction efficiencies for making K^0 particles from tracks and from clusters. On the other hand, if two separate experiments were measuring the rate of Z production through a decay mode to $\mu^+\mu^-$, the common acceptance would contain the branching ratio $Z \rightarrow \mu^+\mu^-$ (and associated error) and the specific factors contain the luminosities etc of the two experiments.

6. Application to a related problem

A different problem arises where an overall limit is required from two (or more) channels for which the efficiencies are known, but the branching ratio is not. For example, suppose the X particle is hypothesised to decay to $\mu^+\mu^-$ and e^+e^- in an unknown ratio, but with (different) efficiencies for detecting $X \rightarrow \mu^+\mu^-$ and $X \rightarrow e^+e^-$ which are credibly estimated from Monte Carlo. From the observed numbers N_μ and N_e what can be said about the upper limit for X production?

This has been surprisingly difficult to answer. However we can now do so, at least in principle. There are two parameters: N_X, the number of X particles, and p, the probability of decaying into muons; the probability of decaying into electrons is $1-p$. We use the fact that in the full definition of confidence levels, the two parts of Equation 1 are inequalities: for the upper limit N_+ on N_X the probability of obtaining this result, or worse, is α, *or less*, whatever the value of p. For a given N_X the probability must be maximised by

varying p,

$$P(N_X) = Max(P(N_X, p))$$

and then N_X can be varied until $P(N_X) = \alpha$. It may be that there is a simple way of finding the value of p which maximises $P(N_X, p)$, though this is complicated as when p changes not only does the probability for each (n_1, n_2) value change, but the region which counts as 'less than' the data result changes discontinuously. However in default of a solution, the limit can be found using this calculator.

7. Possible future developments

It would be possible to relieve the user of the need to guess limits by doing the iteration within the program. This would require a more complicated front interface (specifying the confidence level required and the interval type: upper, lower, central, shortest-distance...) and an automatic scanning strategy to locate the limit to some desired accuracy.

The calculator does not consider the problems that arise when the signal n is similar to, or smaller than, the background b, i.e. when there has manifestly been a fluctuation downwards in the background process. This cannot be taken into account in the standard frequentist framework. For such cases extension of the calculation to use either a flat-prior Bayesian formula formula [8] or the frequentist Feldman-Cousins technique [9] would be possible.

The expected background contributions are specified in terms of numbers of observed events. The model assumes that the error on this quantity is independent of the errors on acceptance. This corresponds to cases where backgrounds are estimated, for example, from observed sidebands. If one has a background from a known physics process affected by a factor in common with the signal (e.g. background from Drell-Yan muon pairs and a serious uncertainty in the muon detection efficiency) then the model could be adapted to do this.

8. References

[1] E. L. Crow and R. S. Gardner, Biometrika **46** 441 (1959)

[2] A. Stuart and J. K. Ord, *Kendall's Advanced Theory of Statistics* 5th Edition, Volume 2, Oxford University Press (1991)

[3] R. D. Cousins and V. L. Highland, *Nucl. Instr & Meth.* **A320** 331 (1992)

[4] R. von Mises, *Probability, Statistics and Truth*, Dover Publications (1981)

[5] The Babar Statistics Working Group, *Recommended Statistical Procedures for Babar*, BaBar Analysis Document # 318. (Version of October 18, 2001) http://www.slac.stanford.edu/BFROOT/www/Statistics/Report/report.ps

[6] D. R. Cox and D. V. Hinkley *Theoretical Statistics*, p 378, Chapman and Hall, London (1974)
   D. S. Sivia *Data Analysis: a Bayesian Tutorial*, p 113, Clarendon Press, Oxford(1996)

[7] G. Cowan, Journal of Physics G: Nucl. Part. Phys. **27**, p 1375 (2001)

[8] O. Helene, *Nucl. Instr & Meth.* **A212** 319 (1983)

[9] G. J. Feldman and R. D. Cousins, *Phys. Rev.* **D57** 3873 (1998)