CDF/MEMO/STATISTICS/PUBLIC/7739
Version 1.0

Recommendations Concerning Limits

The CDF Statistics Committee: Craig Blocker[a], John Conway[b], Luc Demortier[c],
Joel Heinrich[d], Tom Junk[e], Louis Lyons[f], Giovanni Punzi[g]

[a]*Brandeis University*
[b]*University of California, Davis*
[c]*Rockefeller University*
[d]*University of Pennsylvania*
[e]*University of Illinois at Urbana-Champaign*
[f]*Oxford University*
[g]*INFN, University and Scuola Normale Superiore of Pisa*

May 26, 2005

Abstract

Recommendations are given concerning setting limits on the production cross section of physics processes in analyses which count events or consist of several channels or histograms of counted events. A Bayesian technique is recommended, although other valid techniques are not ruled out. Software implementing these recommendations is available.

1 Foreword

We regard it very important in a search experiment to have in advance

(a) a strategy for deciding whether to quote limits and/or the significance of any possible signal;

(b) the technique to be used for quoting a limit; and

(c) the method for setting the significance of a signal. In this set of recommendations, however, we concentrate just on the issue of limits. We hope to produce corresponding recommendations for discovery significance shortly.

Simplest scenario: Determination of the upper limit on a physics production rate parameter s, where the number of observed events n is Poisson distributed with mean $s\epsilon+b$, where ϵ is the acceptance multiplied by the integrated luminosity, and b is the predicted background. Both ϵ and b may have uncertainties associated with them from the relevant subsidiary sources of information.

Other cases: Our comments and recommendations generalize fairly readily to cases where the data consist of the numbers of events in various channels, and where the acceptances in these channels may be partially or completely correlated (and similarly for the backgrounds). There is a further extension to cases where the data consist of one or more experimental distributions, rather than just numbers of events.

A discussion of these situations is provided in [1].

2 General Remarks

1. Of the methods that we have investigated so far, we find that the Bayesian technique is the most practical method currently available, especially for situations involving nuisance parameters (i.e., those connected with systematic effects). It is also the method already used in many CDF analyses. More specific recommendations are listed in Section 3.
2. We do not want to rule out the use of other valid methods.
3. It may be worth employing the same method as used in the corresponding Run 1 analysis. This would facilitate assessing the improvement in the quality of the data, without having to consider the effect of using a different technique for extracting the limit. (We do not find this an overriding argument.)
4. It may also be a good idea to use a similar technique to that used for the same physics parameter in other experiments. It would simplify the comparison and combination of results, although again we do not find this an overriding argument.
5. Decide in advance of looking at the data exactly what technique you are going to use. If the analysis is being performed blind, the method for extracting limits should be completely

specified before unblinding. Optimisation should be based on expectation derived, for example, from simulation, rather than on the data.

3 Specific Recommendations for Implementing the Bayesian Approach

A) CHOOSING THE METHOD

In advance of looking at the data, it should be decided what Bayesian priors to use, and how the limit will be extracted from the posterior.

B) BAYESIAN PRIORS

The Bayesian technique requires the use of prior probability density distributions for all parameters, except for those that are exactly known. We studied the use of the following:

B1. For the acceptance multiplied by the integrated luminosity ($\epsilon = A \int \mathcal{L} dt$):

The exact choice could be based on your detailed knowledge of how ϵ was estimated. If your knowledge is not very detailed (e.g., it consists of nothing more than estimates of the central value and uncertainty) then you could try a gamma, log-normal, beta or truncated Gaussian distribution. Whichever distribution is chosen, a robustness analysis must then be performed (see item E below), and care must be used in selecting the reference prior for the signal strength s (see item B3 below).

B2. For background (b):

Again the choice should depend on detailed knowledge of how b is estimated, and the same suggestions apply as for the ϵ prior. The choice of the functional form for this background prior is less critical than that for ϵ. This is essentially due to the fact that b contributes additively to the Poisson mean, as opposed to ϵ, which contributes multiplicatively.

B3. For the signal strength (s):

In principle one would like to choose a prior that will have as small an effect as possible on the final result (so that the latter will be dominated by the information in the data rather than by prior belief). For simplicity one can choose the prior for s to be 1 for $s \geq 0$ and zero otherwise; this will yield meaningful results as long as the ϵ prior probability density tends to zero as ϵ tends to zero, which is true for the gamma and log-normal distributions.

If the ϵ prior does NOT tend to zero as ϵ tends to zero (e.g., it is a truncated Gaussian), then the constant-s prior will yield a generally meaningless posterior and infinite upper limits. One solution in this case is to choose a prior that goes as $1/\sqrt{s}$ instead of being constant. A different solution is to choose a prior that is constant in the number of signal events $s\epsilon$ rather than in s. (For details please consult [1].) In general the choice of s prior can be made independently of the choice of b prior; this is again due to the fact that b contributes additively to the Poisson mean.

C) STATEMENT OF METHOD

State very clearly in your publication how you extracted the limit, including information about priors, etc. A statement that 'The Bayesian limit at the 90% credibility level is ...' is not sufficient.

D) POSTERIOR PROBABILITY DENSITY

In any publication, show the posterior probability density for s as derived from your data and your choice of priors. And of course, state what priors you have used.

E) ROBUSTNESS ANALYSIS

Check the robustness of your method with respect to changes in the functional form of the priors. Other possibilities for the priors are:

For s: Use a cut-off at large s, such as 1000 pb. Reference [1] gives examples of the effect of the choice of cut-off in s when the prior for ϵ is a truncated Gaussian.

Using Luc Demortier's different factorisation of the prior in s and ϵ could also provide a test of robustness.

For ϵ: Since this is a positive quantity, its prior is likely to be asymmetric, with its mean, median, and mode being different. Check whether it makes a difference which one of these quantities is set equal to your determination of the central value of ϵ. Also try changing the functional form of the prior to gamma, log-normal, truncated Gaussian,... Remember though that if the ϵ prior does not go to zero as ϵ goes to zero, care must be taken with the choice of prior for s (see B3 above).

For b: Similar to ϵ, although we do not think that this is so critical.

We suggest reporting the results of the robustness analysis with a statement such as: 'The results of this measurement do not vary by more than $XX\%$ when reasonable alternatives are tried for the priors.' Of course it may happen that the analysis results are in fact sensitive to the robustness checks. This usually means that there is not enough prior information and/or data to produce a stable limit, and this fact should be reported.

F) COVERAGE

It is desirable to quote the coverage of the method as a function of s, for at least one and preferably a few fixed values of the nuisance parameters. If this is unfeasible (see Appendix below), then the coverage could be calculated as a function of s, after averaging over the relevant nuisance parameters. Consult the References below for existing studies of coverage. Existing studies may mean that it may not be necessary to calculate coverage specifically again for very standard analyses.

G) SENSITIVITY

Quote the sensitivity of the method. This can consist of information about the expected limit in the absence of signal (i.e., median limit, mean limit, or distribution of limits). Alternatively, or additionally, following the suggestion in [2], one can quote the "worst-case limit", that is the loosest possible limit the procedure can produce in case of non-discovery. This metric-independent quantity is also a useful pre-unblind information.

G) AVAILABLE SOFTWARE

Software implementing the Bayesian procedures recommended here is described in [4] and [5].

References

[1] L. Demortier, CDF note 5928, 'Objective Bayesian Upper Limits for Poisson Processes' (2005). This note describes the Bayes approach with flat prior for $s\epsilon$.

[2] G. Punzi, 'Sensitivity of searches for new signals and its optimization', `arXiv:physics/0308063` (2003) and `http://www.slac.stanford.edu/econf/C030908/papers/MODT002.pdf`

[3] CDF Statistics Committee, J. Heinrich *et al.*, CDF note 7117, 'Interval estimation in the presence of nuisance parameters. 1. Bayesian approach' (2004). This describes the Bayes approach with flat prior for s.

[4] J. Heinrich, CDF note 7232 'User Guide to Bayesian-Limit Software Package' (2004), and `http://www-cdf.fnal.gov/physics/statistics/statistics_software.html`

[5] J. Heinrich, CDF note 7587, 'Bayesian limit software: multichannel with correlated backgrounds and efficiencies' (2005).

[6] J. Heinrich, CDF note 6438, 'Coverage of error bars for Poisson data' (2003).

Appendix A: Coverage

Coverage is defined as the fraction of repetitions of the experiment in which the upper limit on s is at least as large as $s_{\rm true}$. The repetitions are assumed to constitute an ensemble, for which the true value of s and of any nuisance parameters are kept fixed, but the measurements are allowed to fluctuate according according to defined probability functions. The coverage C is then

$$C(s_{\rm true}) = \sum P(\text{measurements}|s_{\rm true}, \text{other parameters})$$

where $P(....)$ gives the probabilities of obtaining any set of measured values, and the sum extends only over those measured values for which the upper limit on s is at least as large as $s_{\rm true}$. In our case, C is given by

$$C(s_{\rm true}) = \sum P(n_{\rm obs}|s_{\rm true}, \epsilon_{\rm meas}, b_{\rm meas})$$

In each of these cases, $C(s_{\rm true})$ is also a function of the values of the nuisance parameters.

The repetitions of the experiment are performed at fixed values of s and of the nuisance parameters. For each repetition, not only the observed number(s) of events are allowed to fluctuate, but so are the experimental results which provide information on the nuisance parameters. This enables the coverage to be determined for that value of s and for the chosen values of the nuisance parameters. The coverage C in general is a function of the true values of s, ϵ and b. It is conventional to plot C as a function of s, and we suggest that this is done for some representative fixed values of ϵ and b (e.g., their best estimates, and for each nuisance parameter being moved up and/or down by one error).

There are situations in which it may be difficult to follow this procedure. One would be when there is a large number of nuisance parameters, and this would involve a large number of coverage calculations. The coverage could be calculated as a function of s, but averaged over the relevant nuisance parameters, where the average is weighted by the priors for those parameters. This clearly provides a less stringent test of whether the procedure achieves strict frequentist coverage for all possible parameter values.

Another is when the 'subsidiary measurement' does not exist as such, and for example only a central value and error is known. Then the ensemble of subsidiary measurements may not be defined, even if the prior distribution for $\epsilon_{\rm true}$ has been specified for the extraction of the limit from the data. Even then, it may be feasible to 'invent' an equivalent subsidiary experiment, but this may perhaps be deemed to be too artificial.

In other cases, there may only be limits on a nuisance parameter, for example from theory. It would then be more appropriate to evaluate the coverage (and indeed the limit from the data too) for several values of the nuisance parameter within its range.

In some cases, there are fast analytic methods for evaluating C. Otherwise a slower numerical method is required.

From a strictly frequentist standpoint, coverage should never be less than the nominal confidence/credible level for the limit (i.e., 90% or some other chosen value). In a purely Bayesian approach, average coverage is achieved when the coverage is averaged over all parameters, each weighted by the prior used in the calculation.

It should be noted that coverage is a frequentist construction. Many subjective Bayesians would not accept the concept of "the possible results of an experiment, if it were to be repeated under essentially identical conditions." On the other hand, objective Bayesians may require quoting coverage for parameters with objective priors, but not for those with subjective ones (See Sections 2 and 6.2.2 of [1]) However, most physicists would consider information about coverage as interesting even for a Bayesian method. In a similar vein, they would regard a frequentist method that yielded very small or empty confidence intervals as unsatisfactory because the Bayesian credibility of such an interval is too low.

Coverage is also described in [3]. An educational example can be found in [6].