

Recommendation of the Ad-Hoc Committee on Limit-Setting Procedures to be Used by DØ in Run II

V. Buescher, J.-F. Grivaz, J. Hobbs, A. Kharchilava,
G. Landsberg, J. Linnemann, H. Prosper, and
S. Söldner-Rembold

Final version, October 20, 2004

1 Introduction

Establishing proper statistical procedure for limit-setting and quoting potential signal significance is an important requirement for every experiment involved in searches for physics beyond the standard model. Various prescriptions exist and yet a number of papers appear every year claiming that some of them are more applicable to a particular case than the other. These papers often contradict each other and stem from the fact that several prescriptions that exist use different definition of confidence interval and deal with the physical boundary problem in different ways.

About a decade ago the Particle Data Group (PDG) has attempted to clarify the situation and came up with the “PDG Prescription” on setting limits in astro-particle experiments. Over the years this prescription evolved, and currently the following methods are discussed by the PDG as the most robust: the Bayesian method [1] and two variations of the Frequentist method

based on the ratio of likelihoods: the Feldman-Cousins [2] and CL_s [3] methods. Each of the methods has its own advantages and drawbacks, as documented in [4]. Fortunately, despite different philosophy and sometimes even definition of the confidence interval used in these three approaches, in most of the cases when the observed number of events is reasonably consistent with the background expectation numerical results of all three methods agree well, although no strict mathematical proof of this fact exists. The difference between the results given by the various methods becomes pronounced either when a large negative fluctuation of the background is seen (e.g., if it is more than two standard deviations for the case when one is interested in 95%, or $\approx 2\sigma$ confidence level limit on signal hypothesis), or if the uncertainties in background and acceptances are so large that a significant part of the resulting distribution extends into non-physical regions of negative acceptance and/or background expectation. It is not even clear that in these “pathological” cases proper confidence level limit can be derived unambiguously, so we choose to leave the discussion of the “pathological” cases out of the scope of this note.

In Run I the $D\bar{O}$ experiment established the Statistics Working Group that analyzed various existing methods and recommended Bayesian prescription for limit-setting [5]. CDF generally used a Frequentist approach in Run I, but has switched to the Bayesian prescription since [6]; however they are considering using the CL_s method as well. LEP II experiments used the CL_s method in the Higgs searches [7]. Given the goal that our new results can be combined with Run I results, old or new CDF results, and the LEP results, depending on a particular analysis and channel, we felt that forcing $D\bar{O}$ to use a single limit setting method would be counterproductive. Consequently, this group **recommends that either the $D\bar{O}$ Run I Bayesian prescription or the LEP CL_s method are used for limit-setting in Run II**. Both prescriptions are well-documented and have been implemented in various tools. The following sections provide an overview of the two recommended techniques and their implementation.

Although two different prescriptions are recommended, there are some guidelines for choosing which one to use. When comparing new $D\bar{O}$ results to existing results, the comparison is simpler if both the old and new results use the same scheme. If the previously existing limits are Run I Tevatron results,

then the official DØ Run I prescription [5] is the recommended choice. If the previously existing limits use the CL_s prescription [3] (or similar methods) favored by the LEP working groups, then it should be used when computing the limits for new DØ analyses. If neither was used, and there are no other issues either of the two prescriptions can be used, depending on the authors' preference and availability of tools most appropriate for the analysis at hand.

2 DØ Bayesian Method

Inferences about a set of parameters (σ, λ) are made using Bayes' theorem [1]:

$$p(\sigma, \lambda | \mathbf{x}) = \frac{\mathbf{p}(\mathbf{x} | \sigma, \lambda) \pi(\sigma, \lambda)}{\int \int \mathbf{p}(\mathbf{x} | \sigma, \lambda) \pi(\sigma, \lambda) \mathbf{d}\lambda \mathbf{d}\sigma}, \quad (1)$$

where σ is the *parameter of interest*, for example a cross section, and λ represents all other parameters such as acceptances and backgrounds, referred to collectively as *nuisance parameters*. The functions $\pi(\sigma, \lambda)$, $p(\mathbf{x} | \sigma, \lambda)$ and $p(\sigma, \lambda | \mathbf{x})$ are the *prior*, *model* and *posterior* densities, respectively. The canonical model density for a cross-section measurement is

$$p(\mathbf{n} | \sigma, \lambda) = \prod_{i=1}^{\mathbf{K}} \text{Poisson}(\mathbf{n}_i, \mathbf{a}_i \sigma + \mathbf{b}_i), \quad (2)$$

for K channels, each with n_i observed events, a signal acceptance a_i and a background b_i ; $\lambda \equiv \mathbf{a}, \mathbf{b}$. The prior density $\pi(\sigma, \lambda)$ can be factorized as follows:

$$\pi(\sigma, \lambda) = \pi(\lambda | \sigma) \pi(\sigma), \quad (3)$$

into a prior $\pi(\sigma)$ that involves the cross-section only and one that depends on the nuisance parameters conditional on the value of the cross-section. Usually, we *assume* $\pi(\lambda | \sigma) = \pi(\lambda)$. The prior $\pi(\lambda)$ is typically modelled as a multivariate Gaussian with known mean $\mathbf{m} = \hat{\lambda}$, where $\hat{\lambda}$ represents estimates of the nuisance parameters λ , and a covariance matrix Σ describing how well we know these parameters. The Run I Statistics Working Group [5] suggested, as a *matter of convention*, a flat prior for the cross-section in some

interval $[0, \sigma_{\max}]$. Given the posterior density $p(\sigma|\mathbf{n})$, found by integrating over the nuisance parameters, an upper limit σ^u is obtained by solving

$$CL = \int_0^{\sigma^u} p(\sigma|\mathbf{n}) \mathbf{d}\sigma \quad (4)$$

for σ^u , where CL is the desired confidence level.

Whatever prior is used, a Bayesian upper limit on the cross-section will be most sensitive to the choice of prior precisely in “pathological” circumstances in which the reporting of an upper limit is not a well-established procedure. The upper limit can change by as much as 30% [8] over a plausible class of priors when the data are insufficient to justify a definitive statement. Also, any prior for a scale parameter a , such as an acceptance, that is non-zero at $a = 0$ yields a posterior density for the cross-section that is singular at $\sigma = 0$ when such a prior is used in conjunction with the conventional choice of a flat prior in cross-section. In particular, the common practice of using a Gaussian prior for scale factors suffers from this problem. However, a careful consideration of how an acceptance is arrived at shows that, in fact, one expects its prior to go to zero at $a = 0$. For such priors the singularity does not arise [8].

The Web-based readily available code [9] for the $D\bar{O}$ prescription works only for single-channel counting experiments with uncorrelated signal and background uncertainties. Nevertheless, it’s a powerful tool used in a number of $D\bar{O}$ analyses and publications. There are natural generalizations of the method to multi-channel counting experiments and to analyses which use shape to separate signal and background. However, most of these generalizations have been done via private codes. There is not yet easily usable code which has the extension(s). Because of this, if shapes (e.g., mass distribution of signal and all backgrounds) are explicitly used to determine the limit, either a designated Bayesian program or the CL_s -based MCLIMIT [3] or TLimit [10] codes can be used.

The single top group has implementations [11] of both CL_s and Bayesian methods that handle correlations between parameters (acceptances and backgrounds). This is done by assuming one can model the prior density by a multivariate Gaussian. (If this assumption is unsatisfactory, the user can

supply her or his own prior as a “swarm” of points. That part is still under development.) The code has been developed as a general utility, which is intended to make use of shape information (i.e., histograms).

3 The CL_s Method

The CL_s method uses the estimated signal, s_i , background, b_i , and the number of candidates, n_i , in each bin in the calculation of confidence levels. The description of the CL_s method follows closely the description in [7]. More information can be found on [12].

Confidence levels are computed by comparing the observed data configuration to the expectations for two hypotheses. In the background hypothesis, only the SM background processes contribute to the accepted event rate, while in the signal+background hypothesis the signal from some form of new physics (e.g., leptoquarks, Higgs, SUSY) adds to the background. Each assumed test-variable (e.g. leptoquark mass, SUSY scale parameter Λ) corresponds to a separate signal+background hypothesis.

In order to test the signal+background and background hypotheses optimally with the data, a *test statistic* is defined which summarises the results of the experiment with expectations of the signal+background and background hypotheses maximally different. An optimal choice [13] is the likelihood ratio of Poisson probabilities.

$$Q = \frac{P_{\text{poiss}}(\text{data}|\text{signal} + \text{background})}{P_{\text{poiss}}(\text{data}|\text{background})}, \quad (5)$$

where

$$P_{\text{poiss}}(\text{data}|\text{signal} + \text{background}) = \prod_{i=1}^{n_{\text{bins}}} \frac{(s_i + b_i)^{n_i} e^{-(s_i + b_i)}}{n_i!}, \quad (6)$$

and

$$P_{\text{poiss}}(\text{data}|\text{background}) = \prod_{i=1}^{n_{\text{bins}}} \frac{(b_i)^{n_i} e^{-b_i}}{n_i!}. \quad (7)$$

The products runs over all bins of all distributions to be combined. The signal estimation, s_i , depends on the expected signal cross-section, the decay branching ratios, the integrated luminosity and the detection efficiency for the signal. The background estimation, b_i , depends on the SM background cross-sections, the integrated luminosity, and selection efficiencies. The number of observed events in bin i is n_i . The test statistic is more conveniently expressed in the logarithmic form:

$$-2 \ln Q = 2 \sum_{i=1}^{n_{\text{bins}}} s_i - 2 \sum_{i=1}^{n_{\text{bins}}} n_i \ln(1 + s_i/b_i), \quad (8)$$

which reduces to a sum of event weights, $w = \ln(1 + s_i/b_i)$, depending on the local s_i/b_i for each candidate event observed and on the test-variable. For a given problem the ratio s_i/b_i should be kept finite either by generating enough Monte Carlo statistics for signal and background or by rebinning or smoothing. In this procedure an event-weight is assigned to each event. These weights depend on the test-variable.

To test the consistency of the data with the background hypothesis, the confidence level $1 - \text{CL}_b$ is defined as

$$1 - \text{CL}_b = P(Q \geq Q_{\text{obs}} | \text{background}), \quad (9)$$

the fraction of experiments in a large ensemble of background-only experiments which would produce results at least as background-like as the observed data.

To test the consistency of the data with the signal+background hypothesis, the confidence level CL_{s+b} is defined as

$$\text{CL}_{s+b} = P(Q \leq Q_{\text{obs}} | \text{signal} + \text{background}), \quad (10)$$

the fraction of experiments in a large ensemble of signal+background experiments which would produce results less signal-like than the observed data. By definition a signal+ background hypothesis is excluded at the 95% confidence level if $\text{CL}_{s+b} < 0.05$.

Statistical downward fluctuations in the background can lead to deficits of observed events which are inconsistent with the expected background and

this can cause the signal+background hypothesis to be excluded even if the expected signal is so small that there is little or no experimental sensitivity to it. The confidence level CL_s is defined to regulate this behaviour of CL_{s+b} :

$$CL_s = CL_{s+b}/CL_b. \quad (11)$$

There is some loss of sensitivity by using CL_s rather than CL_{s+b} , but in no case is a limit more restrictive than the one obtained by using CL_{s+b} . We therefore consider a signal hypothesis to be excluded at the 95% CL if $CL_s < 0.05$. This is sometimes referred to as the ‘Modified Frequentist Model’, since it reduces the dependence on the signal distribution.

Because all of the s_i , the b_i (in general), and the candidates in each bin depend on the test-variable, CL_b , CL_{s+b} , and CL_s all depend on the test-variable. For the example of leptoquark production the limit on the leptoquark mass is the smallest test-mass m_{LQ} such that $CL_s(m_{LQ}) \geq 0.05$.

The sensitivity of the analysis can be expressed by the median CL_s in an ensemble of background-only experiments. It is used as the figure of merit to optimise the analysis.

Several programs are using the CL_s method to calculate limits. Two FORTRAN programs, CONFL10 [14] and MCLIMIT [3] have been tested using Run II data and have been found to give consistent results [15]. A limit setting program similar to MCLIMIT can also be accessed directly within ROOT [10].

Systematic uncertainties are taken into account using a generalisation of the method by Cousins and Highland [16]. Systematic uncertainties are incorporated into the confidence level calculations by averaging over possible values of the signal and background given by their systematic uncertainty probability distribution. The probability distributions are assumed to be Gaussian-distributed with a cut off so that negative s or b are not allowed [3]. Correlations between systematic uncertainties are taken into account. Results of different channels and different experiments (which are just treated as different channels) can easily be combined.

4 Conclusions

To conclude, we recommend that DØ analyses use either the Run I Bayesian limit-setting procedure [5], or the CL_s [3] approach, depending on the aspects of a particular analysis and the methods used in related analyses, to be combined with the one at hand. Various tools for calculating confidence intervals in both approaches exist; several of them are briefly reviewed in this note and recommended for use in DØ. This does not imply that other implementations of the DØ Run I or CL_s prescriptions can not be used; however any non-standard code used in the analysis must be properly documented.

References

- [1] R.T. Cox, Am. J. Phys. **14**, 1 (1946); H. Jeffreys, “*Theory of Probability*,” 3rd edition, Oxford University Press, (1961); E.T. Jaynes and L. Bretthorst, “*Probability Theory, the Logic of Science*,” Oxford, 2003; A. O’Hagan, “*Kendall’s Advanced Theory of Statistics, Volume 2B: Bayesian Inference*,” Oxford (1994).
- [2] G.J. Feldman and R.D. Cousins, Phys. Rev. D **57**, 3873 (1998).
- [3] T. Junk, Nucl. Instrum. Meth. A **434**, 435 (1999); A.L. Read, CERN Yellow Report 2000-005.
- [4] Particle Data Group, S. Eidelman *et al.*, Phys. Lett. B **592**, 283–288 (2004).
- [5] I. Bertram *et al.*, DØNote 3476 (1998); preprint Fermilab TM-2104 (2000), URL: <http://library.fnal.gov/archive/test-tm/2000/fermilab-tm-2104.pdf>.
- [6] CDF Memo 7117 (2004),
http://www-cdf.fnal.gov/publications/cdf7117_bayesianlimit.pdf;
CDF Statistics Committee Recommendations, URL:
<http://www-cdf.fnal.gov/physics/statistics/recommendations/limits.html>.

- [7] OPAL Collaboration, G. Abbiendi et al., Eur. Phys. J. C **26**, 479 (2003); ALEPH, DELPHI, L3 and OPAL Collaborations and the LEP Working Group for Higgs Boson Searches, Phys. Lett. B **565**, 61 (2003).
- [8] J. Linnemann, “*Sensitivity of Bayesian upper limits to signal and nuisance priors,*” *Workshop on Confidence Limits*, 27-28 March, 2000; URL: <http://conferences.fnal.gov/cl2k>.
- [9] J. Hobbs and G. Landsberg,
URL: http://www-d0.fnal.gov/~hobbs/limit_calc.html.
- [10] URL: <http://root.cern.ch/root/roottalk/roottalk02/3753.html>;
URL: <http://aleph-proj-alphapp.web.cern.ch/aleph-proj-alphapp/doc/tlimit.html>.
- [11] URL:
http://www-d0.fnal.gov/Run2Physics/top/d0_private/wg/singletop/runI_packages/limits_package.
- [12] URL: <http://www-clued0.fnal.gov/~soldner/limit/>.
- [13] R. Barlow, “*Statistics: a guide to the use of statistical methods in the physical sciences,*” John Wiley & Sons Ltd., West Sussex (1989).
- [14] URL: <http://www-clued0.fnal.gov/~soldner/limit/confprimer.ps>.
- [15] DØ Collaboration, V.M. Abazov et al., Phys. Rev. Lett. **93**, 141801 (2004).
- [16] R.D.Cousins and V.L. Highland, Nucl. Instrum. Meth. A **320**, 331 (1992).