CDF/EXOTIC/PUB/PUBLIC/5236

15 Feb 2000

SETTING LIMITS AND MAKING DISCOVERIES IN CDF

John Conway
Rutgers University

Abstract

This paper presents the statistical methods used in setting limits and discovery significances in the search for new particles in the CDF experiment at the Fermilab Tevatron. For single-channel counting experiments the collaboration employs the classical Helene formula, with Bayesian integration over systematic uncertainties in the signal acceptance and background. For more complex cases such as spectral fits and combining channels, likelihood-based methods are used. In the discoveries of the top quark and B_c meson, the significance was estimated from the probability of the null hypothesis, using toy Monte Carlo methods. Lastly, in the recent SUSY/Higgs Workshop, the Higgs Working Group used a method of combining channels and experiments based on the calculation of the joint likelihood for a particular experimental outcome, and averaging over all possible outcomes.

1. Introduction

In most new particle searches in high energy physics, one selects from a large number of recorded events those which bear characteristics of the new process while minimizing the retention of events from well-understood processes. This typically results in a small number of events passing the selection requirements, consistent with the expectation from a calculation of the expected background. At this stage one typically wishes to determine an upper limit on the number of signal events present in the sample, at some desired confidence level (usually 95%), employing a statistical method which allows one to take into account the systematic uncertainties in signal acceptance and expected background.

If, on the other hand, one observes an excess number of events passing the selection criteria, possibly consistent with the prediciton of an as-yet-unobserved new particle, one would like to estimate the statistical significance of the observation in order to decide if a statistical fluctuation in the number of background events is more likely the cause of the excess.

This note discusses the method used by the CDF Collaboration to determine upper limits on Poisson processes in the presence of uncertainties (both statistical and systematic) simultaneously in the acceptance and background, and the methods for determining the statistical significance of an excess. The collaboration employs rather different methods for single-channel and multi-channel (spectral) searches, in the latter case using a likelihood-based approach which can also be used to estimate experimental sensitivity or expected limits.

FERMILAB-Conf-00/048-E CDF May 2000

2. Single-Channel Limits without Uncertainties

Given n_0, the number of observed events, the probability P for observing *that number* depends on μ, the mean number of signal events expected, according to the Poisson distribution (assuming no background events are expected):

$$P(n_0|\mu) = \frac{\mu^{n_0} e^{-\mu}}{n_0!} \quad . \tag{1}$$

In new particle searches one wishes to determine the value of μ. We define the upper limit N on the number of expected events[1] as that value of μ for which there is some probability ϵ to observe n_0 or fewer events. The confidence level (CL) of the upper limit is then simply $1 - \epsilon$. One can calculate ϵ by summing over the Poisson probabilities:

$$\epsilon = \sum_{n=0}^{n_0} P(n|\mu) \quad . \tag{2}$$

In practice, then, to calculate N one varies μ until finding the value of ϵ corresponding to the desired CL; N is the resulting value of μ.

If one expects an average of μ_B background events among the n_0 observed, and if one knows μ_B precisely, then the method can be extended to calculating a Poisson upper limit N on the number of *signal* events present in the observation. The value of N represents that value of μ_S, the mean number of signal events expected, for which the probability is $1 - \epsilon$ that in a random experiment one would observe *more than* n_0 events *and* have $n_B \leq n_0$, where n_B is the number of background events present in the sample. This can be calculated as before by adjusting N until the relation

$$\epsilon = \frac{\sum_{n=0}^{n_0} P(n|\mu_B + N)}{\sum_{n=0}^{n_0} P(n|\mu_B)} \tag{3}$$

obtains.[1] Various editions of the PDG's Review of Particle Properties in the 1980's and 1990's point out that this results in a "conservative" upper limit in that for some true μ_S the probability of obtaining $N > \mu_S$ exceeds $1 - \epsilon$ on average. This statement means no more than that if the true μ_S exceeds N, then there is a probability smaller than ϵ that one would observe more than n_0 events and have $n_B \leq n_0$; clearly if $\mu_S = N$ the limit is exact.

Note that if one obtains a value of n_0 significantly lower than μ_B, the resulting limit is "better" in that it results in a lower value for N. This is viewed as a shortcoming by some authors,[2] though clearly on average the experiments with larger expected background will on average obtain "worse" (larger) limits on the signal.

The denominator on the right side of Equation 3 makes ϵ a conditional probability, and ensures that N remains positive. This is clearly a desirable feature, and although the method has a frequentist interpretation, this feature is Bayesian in spirit in that the non-physical values are excluded.

[1] Note that N is a real number, not an integer.

3. Single-Channel Upper Limits with Uncertainties

There is no generally accepted method in the high-energy physics community for the incorporation of systematic errors into upper limits on Poisson processes. CDF employs a method which is in essence a Bayesian-style integration over the uncertainties in the signal acceptance and expected background.

Suppose that one knows the value of μ_B to within an overall (statistical plus systematic) Gaussian uncertainty of σ_B, and the acceptance A to within an overall uncertainty of σ_A. In this case the relative uncertainty on μ_S is σ_A/A. One can define the Poisson upper limit N on μ_S as before: we seek that value of the true μ_S for which one would observe *more than* n_0 events *and* have $n_B \leq n_0$. In this case, however, one seeks the value of N such that

$$\epsilon = \frac{\sum_{n=0}^{n_0} \frac{1}{2\pi\sigma_N\sigma_B} \int_0^\infty \int_0^\infty P(n|\mu'_B + \mu'_S) e^{-\frac{(\mu_B-\mu'_B)^2}{2\sigma_B^2}} e^{-\frac{(N-\mu'_S)^2}{2\sigma_N^2}} d\mu'_B d\mu'_S}{\sum_{n=0}^{n_0} \int_0^\infty P(n|\mu_B) e^{-\frac{(\mu_B-\mu'_B)^2}{2\sigma_B^2}} d\mu'_B} \tag{4}$$

where we take $\sigma_N = N\sigma_A/A$. In this way one assumes an *a priori* Gaussian distribution of the true values of μ_S and μ_B about the values obtained in subsidiary studies, with width given by the uncertainties obtained in those studies.

One can perform the integral in Equation 4 by various numerical techniques. The method employed in CDF uses a Monte Carlo integration, rather than performing the integral directly. For each test value of N one generates a large ensemble of random pseudoexperiments, varying the expected number of signal and background about their nominal values according to a Gaussian distribution. In each experiment, the expected number of signal and background events are chosen from the Gaussians, and Poisson-distributed numbers of signal (n_S) and background (n_B) events are generated. For those trials where $n_B \leq n_0$, the fraction f in which $n_B + n_S > n_0$ is recorded. The confidence level for a given N is in fact equal to f; one must then simply vary N until the desired CL $(1 - \epsilon)$ is obtained.

4. Upper Limits with a Bayesian Method

One can also obtain upper limits on a Poisson process using a purely Bayesian approach, as discussed in the literature. [4] A Bayesian deems it sensible to treat the unknown expected number of signal events as a random variable, for which there is some "prior" probability density function (pdf) $\mathcal{P}(\mu_S)$. Given the observation of n_0 events, one can then construct a "posterior" pdf $\mathcal{P}(\mu_s|n_0)$ which depends on the likelihood $\mathcal{L}(n_0|\mu_s)$ for observing n_0 events given μ_S expected:

$$\mathcal{P}(\mu_S|n_0) = \frac{\mathcal{L}(n_0|\mu_S)\mathcal{P}(\mu_S)}{\int_0^\infty \mathcal{L}(n_0|\mu_S)\mathcal{P}(\mu_S)d\mu_S} \ . \tag{5}$$

One can set a Bayesian upper limit (or any other confidence interval) on the unknown parameter μ_S, then, simply from integration of $\mathcal{P}(\mu_S|n_0)$.

The values obtained depend, of course, on the choice of the prior $\mathcal{P}(\mu_S)$. In considering the results of a particular experiment, usually one usually uses an

"uninformed" prior pdf; that is, one wants to give no *a priori* bias to certain values of μ_S. This usually results, then, in choosing $\mathcal{P}(\mu_S)$ to be uniform for all physical values of μ_S: $\mathcal{P}(\mu_S)$ =const. for $\mu_S \geq 0$.[2]

Extension to the case where one expects μ_B background is straightforward:

$$\mathcal{P}(\mu_S|n_0,\mu_B) = \frac{\mathcal{L}(n_0|\mu_S+\mu_B)\mathcal{P}(\mu_S)}{\int_0^\infty \mathcal{L}(n_0|\mu_S+\mu_B)\mathcal{P}(\mu_S)d\mu_S} \ . \tag{6}$$

For uniform prior $\mathcal{P}(\mu_S)$ this reduces to

$$\mathcal{P}(\mu_S|n_0,\mu_B) = \frac{\mathcal{L}(n_0|\mu_S+\mu_B)}{\int_0^\infty \mathcal{L}(n_0|\mu_S+\mu_B)d\mu_S} \ . \tag{7}$$

Remarkably, as Cousins points out [4], the upper limits obtained with this expression match exactly those obtained with Equation 3. Note also that the denominator of Equation 6 can simply be regarded as a normalization constant whose value depends on n_0 and μ_B. Thus we see that

$$\mathcal{P}(\mu_S|n_0,\mu_B) \propto \mathcal{L}(n_0|\mu_S+\mu_B) \ . \tag{8}$$

To incorporate uncertainties on the signal and background one treats the expected background and signal as unknown parameters with uniform prior pdf, with Gaussian likelihood about the estimates from subsidiary studies, just as in the frequentist case. One thus obtains

$$\mathcal{P}(\mu_S|n_0,\mu_B) \propto \frac{1}{2\pi\sigma_B\sigma_S}\int_0^\infty\int_0^\infty \mathcal{L}(n_0|\mu'_S+\mu'_B)e^{-\frac{(\mu_B-\mu'_B)^2}{2\sigma_B^2}}e^{-\frac{(\mu_S-\mu'_S)^2}{2\sigma_S^2}}d\mu'_B d\mu'_S \ , \tag{9}$$

where $\sigma_S = (\sigma_A/A)\mu_S$ comes from the relative uncertainty on the acceptance.

To calculate upper limits, one can simply calculate the right hand side of Equation 9 for an appropriate range of μ_S, and then define the upper limit on μ_S as that value for which

$$\epsilon = \frac{\int_{\mu_S}^\infty \mathcal{P}(\mu_S|n_0,\mu_B)d\mu_S}{\int_0^\infty \mathcal{P}(\mu_S|n_0,\mu_B)d\mu_S} \tag{10}$$

obtains for some desired confidence level $1-\epsilon$.

In general the upper limits obtained using this method exceed those obtained with the frequentist version in Equation 4; that is the Bayes intervals "overcover" the frequentist (or more properly speaking, frequentist/Bayesian) ones. This is regarded as a shortcoming by some authors, and as laudably "conservative" by others. The difference lies, however, in the different meaning of the two statistics.

[2]Note that such a pdf is formally non-normalizable.

5. Discovery Significance: Two Examples

In searching for new particles the possibility exists that the result will be an excess of observed events in the selected sample. The standard in the community is to quote a significance for the excess in terms of the number of Gaussian sigma the result deviates from the null hypothesis. For Poisson processes with small numbers of events this is almost always based on the probability that the background alone can account for the observed number of events. Given n_0 observed events, with $B \pm \sigma_B$ expected background, one typically wishes to calculate the probability of observing n_0 or more, taking into account the uncertainties present. Then one relates this probability to the number of Gaussian standard deviations to quote a significance.

If the uncertainty in the expected number of background events is zero or negligible, then the calculation of the probability $\mathcal{P}_{null}$ of the null hypothesis is a straightforward sum over Poisson probabilities:

$$\mathcal{P}_{null} = \sum_{n=n_0}^{\infty} \frac{B^n e^{-B}}{n!} \quad . \tag{11}$$

To relate this probability to a Gaussian deviation (in units of sigma), one simply finds that value of x for which

$$\mathcal{P}_{null} = \sqrt{\frac{2}{\pi}} \int_x^{\infty} e^{-x'^2/2} dx' \tag{12}$$

obtains. Note that the normalization constant corresponds to finding that fraction of the integral over the *positive half* of the Gaussian lying beyond x. This effectively means that one is calculating the probability that, *for a positive fluctuation*, one would get x or larger in a Gaussian-distributed quantity. Such a convention is necessary to ensure consistency with the confidence intervals determines from the Helene equation (3) and the Bayesian equation (6).

When there is uncertainty in the background, and when there is more than one channel, calculating $\mathcal{P}_{null}$ becomes complicated. Typically in CDF a toy Monte Carlo is used to actually perform the calculation; two examples of actual new particle discoveries illustrate this, those of the B_c meson and the top quark.

In the case of the search for the B_c meson, one sought events where a $J/\psi \rightarrow \mu^+\mu^-$ decay from a secondary vertex was accompanied by an additional lepton (e or μ) from the same vertex, coming from the semileptonic decay of the b quark. The backgrounds were estimated from the sidebands of the J/ψ peak. Table 1 shows the results, the expected background, and the probability that the background alone could give the observed number of events or more in the electron and muon channels.

One might at this stage be tempted to simply quote the product of the two probabilities, or add the observed and expected numbers of events together and calculate a probability that way. But the collaboration first determined the number of signal events present in the sample by minimizing a complicated likelihood function which took into account systematic uncertainties and correlations in the expectation. This yielded a value of 20.4 signal events. To estimate the probability of the null hypothesis a toy Monte Carlo was used to generate over 350,000 pseudoexperiments in which the number of observed events was generated according to Poisson distributions of expected background events,

	$J/\psi + e$	$J/\psi + \mu$
observed	19	12
expected	5.0±1.1	7.1±1.5
probability	0.00002	0.084

Table 1: Results from the CDF search for the B_c meson.

putting in fluctuations and correlations as estimated in the experiment. For each pseudoexperiment the same likelihood fit was performed, and the fraction of such fits which yielded more than 20.4 events was determined from a fit to the shape of the distribution of number of signal events. This fraction, 6×10^{-7}, then, corresponded to a 4.8σ significance. However, this fraction included the results of those pseudoexperiments in which the fitted signal contribution was zero (negative values were not allowed). Thus, strictly speaking, the prescription of considering only positive fluctuations was not adhered to in this case; had it been, the resulting statistical significance would have been close to 4.2σ.

The case of the top quark discovery was more complicated in that there were three overlapping search channels involved, the so-called SVX, SLT, and DIL searches. In the SVX channel, events with a high-p_T lepton (e or μ) plus three or more jets were accepted, and at least one of the jets was required to have been tagged as a b jet with a reconstructed secondary vertex. In the SLT analysis, the same sample was selected, and one jet had to have been tagged as a b by the presence of a low-p_T lepton. In the DIL (dilepton) channel, events with two leptons, large missing E_T, and two or more jets were selected. Table 2 shows the observed number of events, the expected background, and the probability or that channel that the background alone could give rise to the observed number of events or more.

The acceptance for the SVX and SLT channels clearly overlap to a great extent; they are based on the same kinematic selection and only differ by the b-tagging algorithm. To take this into account, the probabilities in the table are calculated by considering the only that set of pseudoexperiments that give the same number of lepton plus jets events as were observed in the actual data sample before b tagging. The overlap in acceptance for the different tagging methods, as well as other uncertainties in the expected background, are modelled in each pseudoexperiment by appropriate Gaussian smearing of the parameters.

To determine the overall significance, the three resulting probabilities are multiplied together, yielding 3.6×10^{-9}. The probability of the null hypothesis is then taken to be the probability that the product of three random numbers, uniformly distributed in the range [0,1], is less than this value. This probability, in fact, can be calculated from a straightforward equation:

$$P(r_1 r_2 ... r_n < \epsilon) = \epsilon \sum_{i=0}^{n-1} \frac{-1^i (\ln \epsilon)^i}{i!} \tag{13}$$

This yields 10^{-6}, which was claimed to be equivalent to a 4.8σ Gaussian significance. However, this value would have been 4.9σ had only positive fluctuations been considered, as discussed above.

	SVX	SLT	DIL
observed	27	23	6
expected	6.7±2.1	15.4±2.0	1.3±0.3
probability	0.00002	0.06	0.003

Table 2: Results from the CDF search for the top quark.

6. Limits from Spectra and Combining Channels

Quite often, particularly in recent years, one uses fits to the spectra of kinematic variables in order to maximize the sensitivity to new particles. Such fits can be made in variables such as the new particle mass, other kinematic quantities which distinguish signal and background, or even the output value of a neural network trained to distinguish signal and background.

The Helene formula applies to only single-channel counting experiments, and thus cannot be used in this case. The natural practice in the case of fitting spectra is to perform a χ^2 or maximum-likelihood fit. For the likelihood, the Poisson probability of observing the number of events n_i in each bin, given the expected background B_i and signal S_i is multiplied together:

$$\mathcal{L} = \prod_{i=1}^{n_{bin}} \frac{\mu_i^n e^{-\mu_i}}{n_i!} \tag{14}$$

where $\mu_i = B_s + S_i$. The likelihood can be maximized (or, more usually, $-\ln \mathcal{L}$ is minimized) with respect to the normalization of the signal, or more generally calculated as a function of the signal normalization. This can be expressed as a variable f which multiplies the signal prediction, such that we have $\mu_i = B_i + fS_i$. Though it is not often made explicit, if one assumes a flat prior pdf in f, then the posterior pdf in f is, via Bayes' Theorem, proportional to the likelihood:

$$\mathcal{P}(f|n_i, S_i, B_i) \propto \mathcal{L}(n_i|B_i, fS_i) \tag{15}$$

One can then, by plotting the likelihood as a function of f, set confidence intervals on f, the signal normalization. For example, to set a 95% CL limit on the signal, one finds that value of f beyond which 5% of the total integral of the likeihood lies. If this value is less than $f = 1$, then one can conclude that the theoretical prediction is excluded at at least the 95% level. Stated more precisely, one can conclude that, if there is equal *a priori* probability that the signal could have any normalization from zero to infinity, then it is less than 5% probable that the true value is more than the theoretical value.

Such a technique has been applied in numerous searches in CDF, including the search for fourth-generation b' quarks decaying to bZ [5], the search for the Standard Model neutral Higgs [6], and other searches. In fact, in these cases, the likelihood is written in such a way as to take into account uncertainties in the signal and background, and correlations in these uncertainties, by integrating over them in the same way as described above for single channel counting experiments. Also, in these cases, there is more than one channel involved. This is handled by simply multiplying the likelihoods for the different channels.

This illustrates powerfully the flexibility inherent in likelihood-based methods: combining channels and taking into account uncertainties is a trivial extension

of the definition of the likelihood. The main difficulty lies in actually calculating the likelihood in cases where the correlations are complicated. This can be made tractable by Monte Carlo integration over these uncertainties.

7. Estimating Experimental Sesitivity

Often in new particle searches one wants to know the sensitivity of a particular analysis, to know how strong a limit can be set with a certain amount of integrated luminosity, or conversely how much integrated luminosity is needed to set a limit or, more optimistically, discover the new particle. This information can be used to optimize analyses, or to estimate the discovery reach of a new machine or detector.

Most often one finds a simple approach is used, in which the ratio of the signal to the square root of the background, $S/\sqrt{B}$ is used as the main indictor of experimental sensitivity. One can then estimate the integrated luminosity necessary for, say, a 5σ discovery by finding when $S/\sqrt{B} = 5$. A 95% CL limit would correspond to $S/\sqrt{B} = 1.96$, using the one-sided formulation discussed above. This procedure gives a reasonable estimate of the required integrated luminosity only when uncertainties are negligible, and the statistics are well in the Gaussian range. It is possible to consider combining single channel counting experiments this way, by adding the values of $S/\sqrt{B}$ in quadrature, but doing this procedure for spectral fits is not possible.

The most straightforward way to estimate experimental sensitivities is to use the likelihood as a function of the signal cross section (or cross section multiplier). This immediately allows for the possibility of incorporating systematic errors and correlations, combining channels, and using spectral fits, just as in the methods outlined in the previous sections.

The main new element in estimating experimental sensitivities is including the fact that there are many possible future experimental outcomes: how does one average over or otherwise take into account the relative probability for all the possible outcomes?

In the Tevatron Run 2 SUSY/Higgs Workshop [7], the Higgs Working Group adopted a statistical procedure based on the joint likelihood for all the various search channels. To take into account all possible future outcomes, the procedure generated large numbers of pseudoexperiments, and for each pseudoexperiment the same procedure which would be applied in a real experiment was applied to that particular outcome. In the case of no signal actually present, for example, the outcome would have only background events present, with Poisson fluctuations around the expected mean background. Then, the integral of the likelihood as a function of Higgs cross section was determined, and the 95% point compared with the theoretical value. To determine the integrated luminosity threshold, then, the integrated luminosity was increased or decreased until in 50% of the pseudoexperiments one could obtain a 95% CL limit. (This follows the convention set by the LEP-II Working Group.

In the case of determining discovery thresholds, again many pseudoexperiments were generated, this time with signal present at the appropriate rate, given the theoretical cross section. To determine whether the particular outcome represented a 5σ discovery, for example, the ratio of the maximum likelihood to the likelihood at zero cross section was used. If this ratio was greater

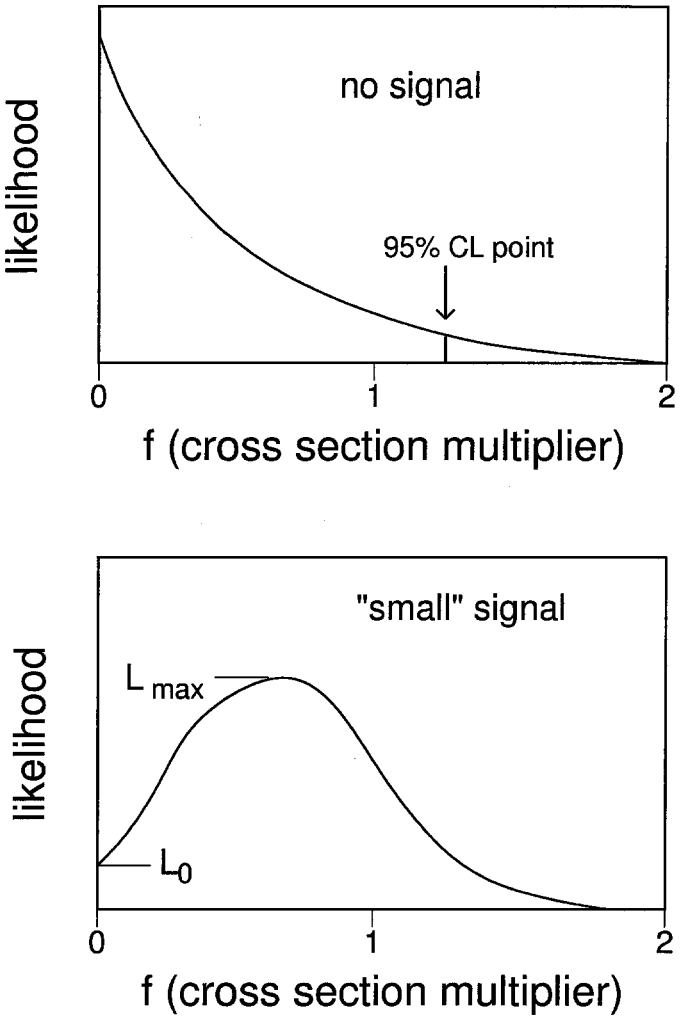

Fig. 1: Illustration of likelihood versus cross section multiplier for two cases in new particle searches, above where there is no signal, and below with a small signal present.

than the equivalent ratio for a Gaussian at 5σ, then the outcome was deemed a 5σ discovery. As in the case of setting a limit, if in 50% of the pseudoexperiments this was the case, then the threshold was said to be met. Figure 1 illustrates graphically the technique, as applied in both cases.

One could also imagine using more standard confidence interval definitions, such as the 68% central interval, to determine whether the pseudoexperiment represented a 5σ discovery. In the limit of Gaussian statistics, the methods should be equivalent. But in the case of a likelihood which is asymmetric about the maximum, there is no set convention for setting such confidence intervals anyway. The bottom line was that the likelihood ratio was much easier to calculate numerically, and with the integral over systematic errors, compute time was very limited.

8. Summary and Conclusions

The techniques in CDF for setting limits and discovery significances in new particle searches have evolved, beginning early on with the Helene formula,

extending the formula to include uncertainties on backgrounds and acceptance. In recent years the collaboration has shifted to likelihood-based methods, which allows the use of fits to spectra, and allows combining channels and the results from different experiments.

For discovery significances, typically CDF has used toy Monte Carlo techniques to estimate the probability of the null hypothesis, the probability that, in the case of no signal, the background alone could produce the observed number of events or more. But clearly this question as well can be addressed, in future analyses, using the same likelihood methods by which we would otherwise set limits, estimate experimental sensitivity and estimate integrated luminosity discovery thresholds.

A clear conclusion is thus that basing the estimates of limits, significances, and sensitivities on the likelihood offers the greatest hope of meeting the needs for incorporating uncertainties, fitting to spectra, and combining channels. Yet it leaves open many questions: Should the field abandon the frequentist view and adopt a purely Bayesian viewpoint? If so, what about the issue of the choice of prior pdf? If a frequentist approach is the goal, should the field adopt the Feldman-Cousins unified approach of likelihood ratio ordering or choose another statistic, such as in the LEP-II CL_s method? [8] Hopefully the field can overcome the present surfeit of methods and adopt a simply understood, explainiable, and meaningful method for making these statistical estimates.

References

[1] O. Helene, Nucl. Intrum. Methods Phys. Res. A 212, 319 (1983).

[2] G. Feldman and r. Cousins, Phys. Rev. D57, 3873 (1998).

[3] G. Zech, Nucl. Instrum. Methods Phys. Res., Sect. A 277, 608 (1989); T.M. Huber *et al.*, Phys. Rev. D 41, 2709 (1990).

[4] R. Cousins, Am. J. Phys. 63, 398 (1995) and references therein.

[5] F. Abe, *et al.*, Phys. Rev. Lett. 84, 216 (2000).

[6] F. Abe, *et al.*, Phys. Rev. Lett. 81, 5748 (1998); F. Abe, *et al.*, Phys. Rev. Lett. 79, 3819 (1997).

[7] See http://fnth37.fnal.gov/susy.html.

[8] See A. Read, these Proceedings.