FULL TITLE
*ASP Conference Series, Vol. **VOLUME**, **YEAR OF PUBLICATION***
NAMES OF EDITORS

The small-n problem in High Energy Physics

Glen Cowan

Physics Department, Royal Holloway, University of London, Egham, Surrey, TW20 0EX, UK

Abstract. In High Energy Physics one often counts events that could come from signal or background processes, where the total number seen is a few or less. One must decide if and when to claim evidence for observation of the signal, and non-observation should be translated into constraints on model parameters. Some of the methods commonly used in Particle Physics are presented and their limitations are discussed.

1. Introduction

In High Energy Physics (HEP) one searches for evidence of new phenomena by looking for particle collisions with particular properties. The particles produced in each collision are characterized by a set of measured quantities (particle energies, momenta, etc.) and the reactions being sought (signal) are in general not fully distinguishable from at least some types of background. In HEP the statistical approaches to this problem have relied mainly on frequentist methods. For example, in the absence of a clear discovery one may try to place some sort of limits on models that predict the signal process, often using a confidence interval. This seemingly straightforward exercise becomes less so when trying to combine searches from different experiments, incorporate systematic uncertainties, and interpret confidence intervals whose size may be surprisingly small (or zero). Both frequentist and Bayesian approaches to the problem will be discussed.

2. The Standard Model and beyond

High Energy Physics has a very well defined "null hypothesis": the Standard Model. This is a quantum field theory of quarks and leptons and their interactions, which proceed through the exchange of vector bosons (photon, $\mathrm{W}^{\pm}$, Z, gluon). To preserve the fundamental gauge symmetry of the theory with massive particles one expects there to exist at least one as yet experimentally unseen Higgs boson. Since the discovery of neutrino oscillations, one has extended the Standard Model to allow nonzero neutrino masses. This version of the theory contains 25 free parameters, which correspond to particle masses, coupling strengths and mixing angles. Some of these parameters have been determined very accurately, e.g., the mass of the Z boson is measured to be $M_{\mathrm{Z}} = 91.1876 \pm 0.0021$ GeV (Eidelman et al. 2004). Others are very poorly known. Indirect information on the Higgs boson mass constrains it to lie in the range $114 < m_{\mathrm{H}} < 207$ GeV (LEP 2006).

Despite a quarter century of detailed tests, there is currently no important disagreement between the predictions of the Standard Model and experimental observation. The places where one can find a several standard-deviation discrepancy are few and most easily interpreted as reflecting an incomplete understanding of complicated systematic effects.

3. Particle physics data

In High Energy Physics the basic unit of collected data is an "event", which usually refers to the result of an individual particle collision or in other cases a single particle decay. For example, a single electron-positron collision at high energy, as carried out at the Large Electron Positron (LEP) Collider at CERN through the 1990s, could result in the production of several dozen particles: pions, kaons, etc. The Large Hadron Collider (LHC) is a proton-proton collider currently under construction in the same underground tunnel and will become operational in 2007. A single collision there can result in the production of several hundred particles or more.

The experiments often consist of large (house-size) particle detectors that surround the collision region and measure the momentum vectors of almost all of the particles produced in the event. In addition they can usually make some attempt to identify the type of particle. The detectors are not perfect and some particles are missed, e.g., those that are emitted at low angles relative to the incoming and outgoing particle beams. Some particles such as neutrinos interact so weakly that they are not detected. Often the particle identification capabilities of detectors is imperfect to the extent where one cannot tell with greater than a certain probability, say, 50 or 75%, whether a given particle was a pion or a kaon.

At the $\mathrm{e}^+\mathrm{e}^-$ collider LEP, the rate of interesting events was at the 1 Hertz level or less. Many analyses focused on events where hadrons were produced, perhaps two dozen per event. There were roughly 1 000 000 hadronic events collected by each of four experiments in operation.

At the LHC, the data rate will be vastly higher: the rate of inelastic proton-proton collisions will approach the 10^9 per second, with the data volume of a single event in the megabyte range. By far most of these collisions are, however, uninteresting from a physics standpoint and we attempt to remove them from the data stream as soon as possible. Events deemed to be sufficiently interesting for further study will be recorded at a rate of roughly 200 Hz. In a nominal year of 10^7 seconds, 10^{16} proton-proton events will occur, two billion of which will be recorded, giving a data volume of roughly two petabytes per year.

Given these very high data rates, it is perhaps not clear why any sort of small-n problem should arise. Many of phenomena one hopes to see, however, are expected to be very rare, if they exist at all, and will be hidden among a huge number of events from known Standard Model processes.

4. Statistical problems in HEP

The nature of the data analysis problem in HEP thus consists of counting the number of events of different types, i.e., different numbers of particles, configuration of momentum vectors, etc. Everything is fundamentally based on counting, and this usually boils down to finding the numbers of entries in the bins of a one or two-dimensional histogram. These numbers can often be modelled as either Poisson or multinomial random variables.

One of the most important goals in an HEP analysis is to look for new phenomena which go beyond the Standard Model, including evidence for supersymmetry, additional gauge bosons or extra space-time dimensions. A great deal of attention has been paid to supersymmetry (SUSY), as this class of model solves a number of theoretical issues as well as supplying a natural candidate for Dark Matter: the weakly interacting neutralino.

A large part of the experimentalists' activities consists of testing these new models and constraining their parameters. One may feel uneasy that we may therefore only find new processes if a theorist has been clever enough to propose the corresponding theory ahead of time. In fact the phenomenology of the models proposed is sufficiently broad that it should encompass a very wide variety of candidate theories. One can imagine that something like SUSY may at first be seen at the LHC, which then subsequently turns out not to be supersymmetry at all but has a very different explanation. The HEP community would be happy to work through this sort of confusion. The immediate focus is on seeing a disagreement with the Standard Model.

5. Claiming discovery of a new effect

Often in Particle Physics a new effect will manifest itself as the observation of events with properties that depart in some clear way from those expected by the Standard Model. For example, if Nature is supersymmetric then we may see events with missing energy, i.e., the total observable energy of the final state particles is found to be less than the initial centre-of-mass energy, since some energy might be carried away by weakly interacting neutralinos. Of course there are Standard Model processes that also result in missing energy, e.g., from neutrinos, and one must also deal with the limited resolution with which the final state energy can be measured.

So the general situation is that we may have events that look like something new, but which cannot in general be distinguished from Standard Model (background) processes. If the number of such events observed far exceeds the background expectation, then clearly a new process has been discovered. In making the transition from unknown to well established, however, the new process will at some point manifest itself at some marginally significant level. The question is then how to quantify this significance and when to decide that it merits a claim for a new discovery.

Frequently one has a specific idea of what a new signal may look like, and candidate events are selected by requiring that their properties satisfy certain criteria or *cuts*. For example, a candidate particle decay $\tau^- \to \mathrm{e}^- \gamma$ (essentially forbidden in the Standard Model but allowed in some supersymmetric models) would have particles identified as an electron and photon whose invariant mass is equal, within some measurement tolerance, to the known mass of the tau lepton.

The number of events n selected will follow a Poisson distribution with mean $b + s$, where b is the expected number due to Standard Model or 'background' processes and s is the number expected from 'signal' or the new process:

$$P(n; s, b) = \frac{(s+b)^n}{n!} e^{-(s+b)} . \tag{1}$$

The expected number of background events, b, is often estimated by using Monte Carlo models. These are of course approximations and the systematic uncertainty in b needs to be taken into consideration when making inferences about the parameter of interest, s. For the moment let as assume that b has been determined with negligible uncertainty.

In this situation the usual approach is to calculate the p-value of the null hypothesis, i.e., the probability, under the assumption of background processes only, of obtaining data as signal-like as, or more signal-like than, the data actually observed. Since the number of events n satisfying the selection criteria for signal events should be large in the presence of the new process, we have for the p-value,

$$p = \sum_{n=n_{\mathrm{obs}}}^{\infty} P(n; 0, b) = 1 - \sum_{n=0}^{n_{\mathrm{obs}}-1} \frac{b^n}{n!} e^{-b} . \tag{2}$$

To claim a new discovery, High Energy Physics folklore dictates a p-value corresponding to a 5σ fluctuation of a Gaussian variable, i.e., $p \approx 2.85 \times 10^{-7}$. Of course the p-value at which one actually believes the null hypothesis to be disproved is subjective and the problem seems to be crying out for a Bayesian approach; more later on why this is problematic as well. But a sort of Bayesian thinking enters even with p-values, since each user combines this result with his or her own prior beliefs. If the alternative to the null hypothesis seems in some sense reasonable or natural, then one might be comfortable announcing a discovery with $p = 0.05$ or less. Indeed the recent announcement of the oscillations of $\mathrm{B}_{\mathrm{s}}^0$ mesons, widely expected to be found, rejected the null hypothesis with $p = 0.038$ (Abazov et al. 2006). Other effects would be more surprising and require correspondingly stronger evidence.

There are a number of well-known dangers with the p-value approach. For example, the new type of event will not, in general, look exactly like the thing we were looking for. Perhaps we see an excess in the search for events with missing energy that gives a p-value of 10^{-3}. We then observe that if we modified the selection criteria somewhat, e.g., we require not 30 GeV of missing energy but rather 20 GeV, then the p-value drops to 10^{-4}, and if we require that there be at least one muon in the event, then it drops to 10^{-5}, and so forth. A problem arises of course if we make these changes to the selection criteria after seeing the data. One can always define a set of cuts that enclose the observed events in such a way as to make them appear highly significant. The best one can do here is to freeze the selection criteria and take more data, but this is expensive and not always an option.

In addition to counting events one often measures one or more quantities that characterize the events, e.g., invariant masses or energies of various particles or jets of particles. The single number of observed events n is then replaced by a set of numbers $\vec{n} = (n_1, \ldots n_N)$ of entries seen in a histogram. Defining the p-value of the null hypothesis requires one to define the region of the data space with equal or lesser compatibility with the data actually observed.

If the expected histogram is roughly flat but one observes a large peak in a couple of neighbouring bins, then there is a temptation to focus on these bins only and to ask for the probability to find, under the assumption of background only, as many events as were actually seen there or more. The problem could be, of course, that we may not have known a priori where the new peak would appear, indeed whether the new effect would be a peak, or an anomalous slope, or whatever. Again the only meaningful solution from the frequentist standpoint would be to fix the critical region and take more data.

Up to now we have assumed that the expected background b was known, but there can be various sorts of uncertainty in its value. It may be calculated from Monte Carlo model with a limited amount of simulated data, the modelling of the detector's response is imperfect, and even the Standard Model prediction, often computed using perturbation theory at a fixed order, is only approximate. Effectively the $s = 0$ hypothesis is no longer simple but characterized by a potentially large family of nuisance parameters. One approach is to report the corresponding range of p-values for different values of the nuisance parameters. Often to be conservative one will report the largest plausible p-value, i.e., the weakest evidence for a new discovery. In other cases one may wish to regard b as a random quantity, i.e., it effectively plays the role of an estimate $\hat{b}$ of the true (and unknown) expectation value for the number of background events. In principle one can remain within the frequentist framework where now the outcome of the experiment contains the number of events n and the value $\hat{b}$. We will return to this approach when we discuss setting limits on the signal parameters.

Finally, why don't particle physicists simply compute the posterior probability of the null hypothesis in the Bayesian framework? Here the problem lies primarily in assigning a meaningful prior probability to any of the hypotheses involved. Most physicists are relatively certain that the Standard Model is 'wrong' in that it is probably not Nature's final theory of particle interactions. Having said that, no one knows how to enumerate the alternatives in a meaningful way and to assign to them probabilities. If the final result of an analysis was announced to be that the probability of the Standard Model is X and the probability of such-and-such a supersymmetric theory was Y, then the average reader of this result would simply not know whether to be convinced or not. The reader's prior beliefs, to the extent that they were ever thought through at all, almost certainly disagree with those of the paper's authors. When one takes into account that HEP papers are written collectively by hundreds or even thousands (!) of authors, such an approach becomes highly problematic.

6. Setting limits

If a new phenomenon is sought but not found, one usually wants to at least constrain the parameters of the model in question. In the first instance this often means placing an upper limit on the the expected number of signal events,

s, which is related to the cross section σ by $s = \sigma L \varepsilon$, where L is the integrated luminosity and ε is the efficiency (probability for an event to be observed). The cross section σ will be predicted by the model as a function of its fundamental parameters, and often one translates the limit on the cross section into limits on these parameters. In other cases the measured quantity may follow other distributions, e.g., multinomial or Gaussian, but the general goal is still to quantify in some way the level of agreement between the observed data and the predictions represented by different regions of parameter space of the candidate models.

6.1. Setting limits with frequentist confidence intervals

The most widely used tool for setting limits in HEP is the standard frequentist confidence interval. A confidence interval for a parameter can be obtained by a test of the hypotheses corresponding to each possible parameter value. To carry out a test, one defines a *critical region* in the data space which is disfavoured by the hypothesis, such that there is a prespecified probability γ (the *size* or *significance level* of the test) for the data to be observed in the critical region. If the data are discrete such that one cannot achieve this condition exactly, then the probability to be in the critical region should be as close as possible to but not greater than γ. The significance level γ is often chosen to be a small value such as 0.1 or 0.05. There is a clearly a degree of arbitrariness in the definition of the critical region and the choice will depend on which alternative hypotheses we are most interesting in rejecting.

To obtain a confidence region one inverts this procedure. The confidence region at confidence level $1-\gamma$ is the set of parameter values which would not be rejected by a test of size γ. Construction of a confidence interval through inversion of a test is equivalent to the construction of a *confidence belt*, which is a graphical depiction of the test's acceptance region as a function of the parameter. By construction the confidence region will contain the true value of the parameter with a probability greater than or equal to $1-\gamma$.

6.2. Poisson data with signal and background

An important class of measurements consists of counting a certain number of events n, where n can be modeled as a Poisson variable with expectation value $E[n] = s + b$. Here s represents the contributions from the new (signal) process and b is the expected number from background processes. The goal is to constrain s.

Now if we want to set an upper limit on s, we can invert a test of the hypothesis that its true value is greater than or equal to s. The critical region for this hypothesis consists of low values of n. So for each value of s we define a critical region by the lowest n such that their total probability does not exceed a given value β, where $1-\beta$ is the desired confidence level of the upper limit. That is, we require

$$\beta = P(m \leq n; s, b) = \sum_{m=0}^{n} \frac{(s+b)^m}{m!} e^{-(s+b)} \,. \qquad (3)$$

The solution to (3) is the upper limit s_{up}. Similarly for a lower limit at confidence level $1-\alpha$, the parameter values we want to exclude are low, and therefore the critical region should consist of the highest values of n. So we require

$$\alpha = P(m \geq n; s, b) = 1 - \sum_{m=0}^{n-1} \frac{(s+b)^m}{m!} e^{-(s+b)} \ . \tag{4}$$

The solution to (4) is the lower limit s_{lo}. The sums of Poisson probabilities can be carried out easily by exploiting their relation to the χ^2 distribution. One finds

$$s_{\mathrm{lo}} = \frac{1}{2} F_{\chi^2}^{-1}(\alpha; 2n) - b \ , \tag{5}$$

$$s_{\mathrm{up}} = \frac{1}{2} F_{\chi^2}^{-1}(1-\beta; 2(n+1)) - b \ , \tag{6}$$

where $F_{\chi^2}^{-1}(\alpha; n_{\mathrm{d}})$ is the inverse of the cumulative distribution function (quantile) for the χ^2 distribution for n_{d} degrees of freedom.

There are a number of subtle difficulties with confidence limits obtained from this procedure. If, for example, the number of events observed n is small compared to the expected background b, then equation (6) can yield a negative number. In this case the confidence interval contains no physical values of the parameter; it is the empty set.

If one obtains an empty interval, from the statistician's standpoint, nothing has 'gone wrong' – the interval will by construction only contain the true parameter's value with a probability $1-\gamma$. Clearly if the interval is empty then we have encountered a case where we don't cover the true value. Now to report such a result is unsatisfying to say the least, and it does not summarize the outcome of the experiment in a meaningful way. The root of the problem lies in the physicist's desire to see the confidence interval as a region of parameter space where there is a high probability for the true value to lie, but of course in the purely frequentist approach one does not associate a probability with parameter values, only with data.

Because of these difficulties with small or zero size intervals that can result from a downward fluctuation in the number of background events, one should also quote the *sensitivity*, which can be defined as the expectation value of the upper limit under the assumption of background only. Alternatively the median value can be used, which has the advantage of being invariant under a reparameterization of the problem. To determine the sensitivity one could, for example, simulate the experiment many times and look at the distribution of resulting limits.

Several problems with the standard upper limits, including the issue with empty intervals mentioned above, can be mitigated using a procedure rediscovered for HEP by Feldman & Cousins (1998). To construct the interval they base the corresponding test on the likelihood ratio

$$l(s) = \frac{L(n|s,b)}{L(n|\hat{s},b)} \ . \tag{7}$$

Here in $\hat{s}$ is the Maximum Likelihood (ML) estimator for s, which in this case is

$$\hat{s} = \begin{cases} n - b & n \geq b, \\ 0 & \text{otherwise.} \end{cases} \tag{8}$$

(Note that $\hat{s} = n - b$ gives an *unbiassed* estimator but for $n < b$ goes outside the allowed parameter space.) The critical region of the test consists of those values of n with the lowest value of the likelihood ratio. The upper limits obtained from the standard (one-sided), Feldman-Cousins and Bayesian procedures (see below) are shown in Fig. 1.

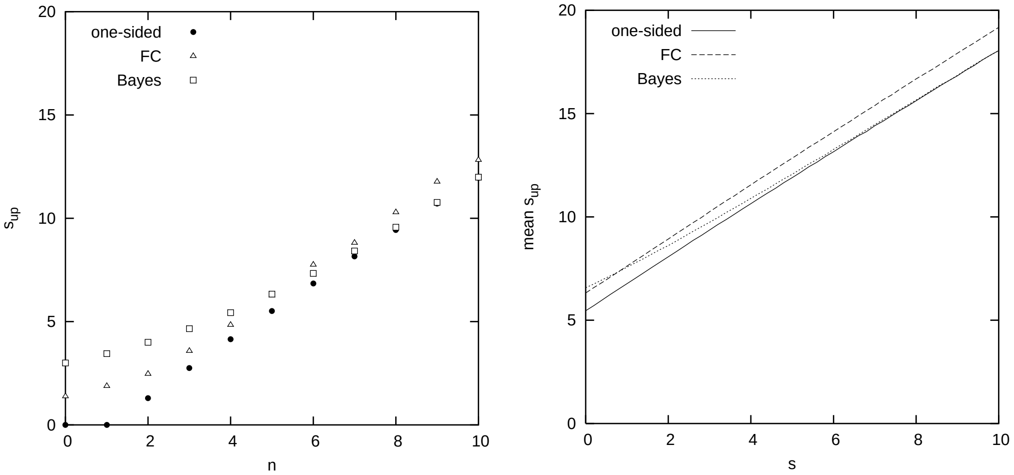

Figure 1. *Left:* Upper limits as a function of the observed number of events n assuming an expected background of $b = 5.0$. *Right:* The mean upper limit as a function of s for $b = 5.0$.

Now unless n comes out small compared to b, the upper limits from inversion of the likelihood-ratio test (Feldman-Cousins) are higher than those from the procedure for the one-sided upper limit, so naively one could think that the likelihood ratio is not providing limits that are as stringent. But with the Feldman-Cousins limits the critical region of the test generally contains both low and high values of n. Depending on the data observed, the resulting interval may be one- or two-sided. Feldman and Cousins therefore call these intervals 'unified', in the sense that there is a smooth transition between one- and two-sided. For n much greater than b, the Feldman-Cousin interval is two-sided, and the probability $1 - \gamma$ for the interval to miss the true value is shared above and below the two limits.

Feldman and Cousins have pointed out that if one decides whether to quote an upper limit or a two-sided interval only after seeing the outcome of the experiment, then the coverage probability of the interval is no longer guaranteed to be be greater than or equal to the nominal confidence level $1 - \gamma$. This 'flip-flopping' problem is effectively cured, however, by the unified interval.

Although the Feldman-Cousins intervals are never strictly empty, if $b \ll n$ they can be arbitrarily small, easily smaller than the sensitivity as defined above. So regardless of the type of interval reported one is always encouraged to report the expected (mean or median) limit under the absence of signal.

The desirability of unified intervals versus upper limits may depend on the type of question one wants to answer. The purpose of the limit is presumably to give guidance on what regions of parameter space are disfavoured by the data, for example so that one can proceed to plan further experiments to continue the search for the phenomenon. Now if a unified limit results in, say, $0.1 < s < 5.0$ at 90% confidence level, then one does not really have significant evidence for a discovery. The p-value of $s = 0$ may still be several percent.

The important result here is that $s > 5.0$ is now disfavoured, and one should move on and design a new experiment to be sensitive to s values significantly less than 5.0. With a one-sided upper limit one would have a more stringent limit, but would remain open to the criticism of 'flip-flopping'. One could always counter that regardless of the outcome of the experiment, even in the case of an obviously significant discovery, one will always quote an upper limit. Furthermore, to actually claim a discovery one wants an interval that excludes the null hypothesis at a much higher confidence level, e.g., 99.99%, not just 90%. If one decides the confidence level only after looking at the data, one winds up with another sort of flip-flopping. This type of discussion is still ongoing in the HEP community.

6.3. Combining results

Often in particle physics one has a number of measurements, all of which provide information on the same parameter. A simple example of this could be independent Poisson measurements looking for events of the same type but in two different experiments. Suppose one measurement counts n_1 events and the other n_2. The expectation values are $s_i = \varepsilon_i L_i \sigma$ $(i = 1, 2)$, i.e., the efficiencies and luminosities of the two measurements may in general differ but they are related to the same cross section σ. The goal is to place a limit on σ.

The example of two measurements is a special case of multivariate data $\vec{x} = (x_1, \ldots, x_n)$ whose joint distribution depends on a set of parameters $\vec{\theta} = (\theta_1, \ldots, \theta_m)$. Here there can be nontrivial correlations between the variables as they may not represent simply the outcomes of independent experiments, but rather sets of kinematic variables within the same event. As previously we can construct the likelihood ratio,

$$l(\vec{\theta}) = \frac{L(\vec{x}|\vec{\theta})}{L(\vec{x}|\hat{\vec{\theta}})} , \tag{9}$$

where $\hat{\vec{\theta}}$ is the ML estimator. And as before we can define a test of the point in parameter space $\vec{\theta}$ with a critical region consisting of that part of $\vec{x}$-space with the lowest values of the likelihood ratio. The confidence region at confidence level $1 - \gamma$ is defined by the set of points in parameter space that would not be rejected in a test of significance level γ.

Now in practice, this type of approach can be difficult because it is not always simple to determine the full joint distribution of the data for all points in parameter space. This procedure has been successfully applied, however, to combine the results of the the four LEP experiments' searches for the Higgs boson (see e.g. the contribution by Cranmer in Lyons & Karagöz Ünel (2005)).

6.4. Bayesian solutions

Instead of using frequentist confidence intervals, one can try to constrain the parameter in the Bayesian framework by assigning to it a prior probability density $\pi(s)$ and combining this with the likelihood $L(n|s)$ to find the posterior density from Bayes theorem,

$$p(s|n) \propto L(n|s)\pi(s) \ . \tag{10}$$

Here the likelihood function $L(n|s)$ is simply the Poisson probability (1). The difficulty with this approach is in assigning a meaningful prior $\pi(s)$. In the absence of detailed prior information, a common choice is

$$\pi(s) = \begin{cases} 0 & s < 0 \\ 1 & s \geq 0. \end{cases} \tag{11}$$

This prior probably does not reflect anyone's true prior belief about s but can be regarded as a basis for comparison. Reference priors have been widely discussed in the literature, including among particle physicists (see e.g. the contribution by Demortier in Lyons & Karagöz Ünel (2005)). Bayesian credible intervals can be found by integrating the posterior pdf to contain any desired probability. For example, for an upper limit $s_{\rm up}$ at with a probability content of $1-\gamma$ one solves

$$1 - \gamma = \int_0^{s_{\rm up}} p(s|n)\, ds \ . \tag{12}$$

For the uniform prior (11), one at least has a convenient point of contact with the frequentist confidence interval, namely, if the background is zero, then the upper Bayesian upper limit from (12) coincides with the frequentist upper limit from (6). In cases with $b \approx 0$ the author can evade having to commit to either approach.

A problem with using a uniform reference prior for s is that one might just as easily formulated the problem using a different parameter. For example, the expected number of Higgs boson events can be predicted as a function of the particle's mass, $m_{\rm H}$, and if we had taken a uniform prior in $m_{\rm H}$ rather than s then we would obtain a different result. A uniform prior in s translates into a non-uniform prior for a nonlinear function of s. If the prior really reflects our subjective degree of belief about where its value lies, then this problem does not arise. The transformed prior for the new parameter then correctly reflects our degree of belief. But if we simply take a certain function form (e.g., uniform) as a reference prior, then the end result will depend on our choice of parameter.

A strong selling point of the Bayesian approach, however, is the ease with which different measurements can be combined. In the general case with multivariate data $\vec{x}$ and possibly as well a multidimensional parameter space $\vec{\theta}$, we simply find the joint posterior pdf from Bayes theorem,

$$p(\vec{\theta}|\vec{x}) \propto L(\vec{x}|\vec{\theta})\pi(\vec{\theta}) \ , \tag{13}$$

where $L(\vec{x}|\vec{\theta})$ is the likelihood function and $\pi(\vec{\theta})$ is the joint prior. If we are only interested in a subset of the parameters, we simply integrate over those that are unwanted. This integral may indeed be nontrivial to carry out and could require, for example, Markov Chain Monte Carlo techniques. But for many problems, especially combining independent Poisson measurements related to the same cross section, the Bayesian framework provides a relatively simple solution.

A further advantage of the Bayesian approach is of course that we can treat systematic uncertainties, which are not easily thought of as a variation of an observation upon repetition of the experiment. We examine this point further in the next section.

7. Including systematic uncertainties

The full model that one tests generally contains not only the fundamental parameters of its underlying theory but various nuisance parameters. An example from the Poisson problem above is the parameter b, the expected number of events from background processes. Up to now we have treated it as a known constant but in practice it will have some uncertainty. The question is how to incorporate this uncertainty into the limits on s. We will use the parameter b as an example but the extension to other nuisance parameters, e.g., selection efficiencies or detector calibration constants, is straightforward.

In the Bayesian approach any uncertainty in the nuisance parameters is written down in the prior pdf. In our Poisson case we would have a prior $\pi(s, b)$. This will often factorize as $\pi(s, b) = \pi_s(s)\pi_b(b)$, and $\pi_b(b)$ could be the posterior from a subsidiary measurement of b. Bayes' theorem gives the joint pdf for s and b,

$$p(s, b|n) \propto L(n|s, b)\pi(s, b) \ . \tag{14}$$

To find the marginal pdf for s, we simply integrate over b,

$$p(s|n) = \int p(s, b|n)\, db \ . \tag{15}$$

With frequentist confidence intervals the problem is more difficult. One must extend the confidence intervals to a potentially large multiparameter confidence region that now includes all of the nuisance parameters as well. In the large sample limit the problem can be solved using the *profile likelihood.* This is given by

$$L_{\mathrm{p}}(s) = L(n|s, \hat{\hat{b}}) \ . \tag{16}$$

Here $\hat{\hat{b}}$ is found from the maximum $L(s, b)$ for each value of s. One then defines a test for the parameter value s using the ratio of profile likelihoods,

$$l_{\mathrm{p}}(s) = \frac{L_{\mathrm{p}}(s)}{L_{\mathrm{p}}(\hat{s})} \ . \tag{17}$$

The critical region of the test is defined by the values of n that give the lowest values of $l_{\mathrm{p}}(s)$, and the test is inverted to obtain a confidence interval for s. This procedure boils down to finding approximate confidence intervals from the tangent (hyper-)planes in parameter space to contours of constant log-likelihood. For the small-n problem the profile-likelihood recipe can be taken to provide an approximate confidence interval. These ideas are discussed further in the papers by Punzi, Cranmer, Rolke, Reid and Feldman in Lyons & Karagöz Ünel (2005).

One of the methods for incorporating systematic uncertainties into limits most frequently used in HEP was proposed by Cousins and Highland (Cousins & Highland 1992). Again for purposes of example let us assume the nuisance parameter in question is the expected number of background events b. Let us assume that the uncertainty in b can be described by a prior density $\pi(b)$. If, for example, we only had an estimate $\hat{b}$, then $\pi(b)$ might be a Gaussian centred about $\hat{b}$ with an appropriate standard deviation $\sigma_{\hat{b}}$. One then computes the probability law for n as

$$P(n;s) = \int P(n;s,b)\, \pi(b)\, db\ . \tag{18}$$

In a strictly frequentist framework this would be the probability for n if b were not constant but rather sampled from $\pi(b)$ upon repetition of the experiment. One can then use this $P(n;s)$ in equations (3) and (4) and solve numerically for the limits. A calculator has been provided by Barlow (Barlow 2002).

This approach can be extended easily to limits based on a likelihood-ratio test. Suppose we can characterize the uncertainty in b with a prior pdf $\pi(b)$. We define the 'integrated likelihood' (also called 'modified profile likelihood', in any case not a real likelihood) as

$$L'(n|s) = \int L(n|s,b)\, \pi(b)\, db \tag{19}$$

Now we can use this to construct the corresponding ratio,

$$l'(s) = \frac{L'(n|s)}{L'(n|\hat{s})}\ , \tag{20}$$

where here $\hat{s}$ is the estimator obtained from the maximum of $L'(n|s)$. The critical region for the test of s then consists of those values of n with the lowest value of $l'(s)$, in close analogy with the Feldman-Cousins procedure using equation (7). We then invert this test to obtain the confidence interval. This must be done numerically; calculators have been provided by Conrad and Tegenfeldt (Lyons & Karagöz Ünel 2005; Conrad et al. 2003).

8. Summary

The seemingly simple problem of placing limits on a parameter has been discussed at length by the Particle Physics community and there is still no clear winning method. Among the things that particle physicists agree on is the importance of including in the result either the likelihood function or an appropriate summary of it. In this way the combined likelihood from two independent measurements can easily be obtained, and furthermore the likelihood can be combined with any prior for use in a Bayesian analysis.

Frequentist confidence intervals have played a major role in HEP for many years and for purposes of setting limits it appears this situation will continue for some time. Bayesian methods using reference priors have also found some popularity, but here the attitude has generally been to use the result as a recipe to set the limit and then to study its frequentist properties. Subjective Bayesian methods have started to attract some attention, especially for the problem of systematic uncertainties. The hope is of course that in coming years, especially at the Large Hadron Collider, there will real new phenomena to discover and people will be less concerned with setting limits.

Acknowledgements

It is my pleasure to thank the organizers of the conference, especially Jogesh Babu and Eric Feigelson, for an extremely interesting and fruitful meeting. I thank Michael Woodroofe, Louis Lyons and Roger Barlow for their valuable comments.

References

Abazov, V. et al. (D0 Collaboration) 2006, hep-ex/0603029, Phys. Rev. Lett. 97:021802

Barlow, Roger 2002, Comput. Phys. Commun. 149, 97-102. Programs avaiable from `www.slac.stanford.edu/~barlow/statistics.html`

Conrad, J. et al. 2003, Phys. Rev. D67, 012002; see also Conrad, J. & Tegenfeldt, F. 2005, *Likelihood ratio intervals with Bayesian treatment of uncertainties*, physics/0511055

Cousins, R.D. and Highland, V.L. 1992, Nucl. Inst. Meth. A320, 331

Eidelman, S. et al. 2004, Physics Letters B592, 1

Feldman, G.J. & Cousins, R.D. 1998, Phys. Rev. D57, 3873

Heinrich, Joel et al. 2004, *Interval estimation in the presence of nuisance parameters. 1. Bayesian approach*, CDF Collaboration Note 7117, physics/0409129

LEP Electroweak Working Group 2006, update from `lepewwg.web.cern.ch/LEPEWWG/`

Lyons, L. & Karagöz Ünel, M. (eds.) 2005, proceedings of PHYSTAT05: Statistical Problems in Particle Physics, Astrophysics and Cosmology, Oxford, 12-15 September, 2005