

Statistical Issues in Particle Physics

Louis Lyons

*Particle Physics, Keble Rd, Oxford OX1 3RH, UK and
Blackett Lab, Imperial College, Prince Consort Rd, London SW7 2BW, UK*

e-mail: l.lyons@physics.ox.ac.uk

Abstract

Many statistical issues arise in the analysis of Particle Physics experiments. This review covers such questions as: Do we really need coverage? Should we insist on at least a 5σ effect to claim a discovery? How should p -values be combined? If two different models both have respectable χ^2 probabilities, can it be reasonable to reject one in favour of the other? Are there different possibilities for quoting the sensitivity of a search? How do upper limits change as the number of observed events becomes smaller and smaller than the predicted background? Is it possible to combine 1 ± 10 and 2.0 ± 7.5 (two measurements of the same parameter) to obtain a result of 5 ± 1 ? What is the Punzi effect and how can it be understood?

1 Introduction

Analyses of experimental data in Particle Physics have, perhaps not surprisingly, tended to use statistical methods that have been described by other Particle Physicists. The articles by Solmitz [1] and by Orear[2] have provided an introduction to statistics for many of us, and there are several books written on the subject by Particle Physicists[3].

In recent years there has been a growing awareness by Particle Physicists of the desirability of using good statistical practice. This is because the accelerator and detector facilities have become so complex and expensive, and involve so much physicist effort to be built, tested and run, that it is clearly important to treat the data with respect, and to extract the maximum information from them. The PHYSTAT series of Workshops and Conferences[4] – [10] has been devoted specifically to statistical issues in Particle Physics and neighbouring fields, and many interesting articles can be found in the relevant Proceedings. These meetings have benefited enormously from the involvement of professional statisticians, who have been able to provide specific advice as well as pointing us to some techniques which had not yet filtered down to Particle Physics analyses.

Another source of useful information is provided by the statistics committees set up by some of the large collaborations (see, for example, refs. [11] – [14]).

1.1 Types of statistical analysis

There are several different types of statistical procedures employed by Particle Physicists:

- Separating signal from background: Almost every Particle Physics analysis uses some method to enhance the possible signal of interest with respect to uninteresting background.
- Parameter determination: Many analyses make use of some theoretical or empirical model, and use the data to determine values of parameters, and their uncertainties and possible correlations.
- Goodness of fit: Here the data are compared with a particular hypothesis, often involving free parameters, to check their degree of consistency.
- Comparing hypotheses: The data are used to see which of two hypotheses is favoured. These could be the Standard Model (SM), and some specific version of new physics, such as the existence of the Higgs boson.
- Decision making: Based on what we believe about the current state of physics, the likelihood of possible discoveries and estimates of how difficult it would be to carry out future experiments, we could decide what should be thrust of our future research. This subject is beyond the scope of this review.

1.2 Statistical and systematic errors

In general any attempt to measure a physics parameter will be affected by statistical and by systematic errors. The former are such that, if the experiment were to be repeated, random effects would result in a distribution of results being obtained. These can include effects due to the limited accuracy of the measurement devices and/or the experimentalist; and also from the inherent Poisson variability of observing a number of counts n . On the other hand, there can be effects that shift the measurements from their true values, and which need to be corrected for; uncertainties in these corrections contribute to the systematic error. Another systematic could arise from uncertainties in theoretical models which are used to interpret the data. Our systematics are often ‘nuisance parameters’ for Statisticians.

Thus we could consider an experiment designed to measure the temperature at the centre of the sun by measuring the flux of solar neutrinos on earth. The main statistical error might well be that due to the limited number of neutrino interactions observed in the detector. On the other hand, there are likely to be systematic uncertainties from limited knowledge of neutrino cross-sections in the detector material, the energy calibration of the detector, neutrino oscillation parameters, models of energy convection in the sun, etc. If some calibration measurement or subsidiary experiment can be performed, this effectively converts a systematic error into a statistical one. Whether this source of uncertainty

is quoted as statistical or systematic is not crucial; what is important is that possible sources of correlation between error contributions here and in other measurements (in this or in other experiments) are well understood.

Assessing the magnitude of systematic effects in a parameter-determination situation often requires producing simulations of the data with different values of the nuisance parameter(s), and seeing how much the result changes¹ when the nuisance parameter value is varied by its uncertainty (compare Sections 4.5 and 6.5 for ways of incorporating nuisance parameters in upper limit and in p -value calculations respectively). When several nuisance parameters are involved, there is the question of whether separate simulations should be produced, in each of which only one of the nuisance parameters is changed from its optimal value by its uncertainty; or whether it is better to generate simulations in each of which all nuisance parameters are simultaneously changed from their optimal values according to their expected (possibly correlated) multivariate distribution. The two methods are sometimes referred to as unisim (or OFAT = **O**ne **F**actor **A**t a **T**ime) and multisim respectively. Roe has discussed which method requires less computing time to achieve the same accuracy for the systematic error[15].

How to assess systematics was much discussed at the Banff meeting[9] and at PHYSTAT-LHC[16, 17, 18]. Many reviews of this complex subject exist and can be traced back via ref. [19].

1.3 Bayes and Frequentism

These are two fundamental approaches to making inferences about parameters or whether data support particular hypotheses. There are also other methods which do not correspond to either of these philosophies; the use of χ^2 or the likelihood are examples.

Particle Physicists tend to favour a frequentist method. This is because in many cases we really believe that our data are representative as samples drawn according to the model we are using (decay time distributions often are exponential; the counts in repeated time intervals do follow a Poisson distribution; etc), and hence we want to use a statistical approach that allows the data “to speak for themselves”, rather than our analysis being dominated by our assumptions and beliefs, as embodied in the assumed Bayesian priors. Bayesians would counter this by remarking that frequentist inference can depend on the reference ensemble, the ordering rule, the stopping rule, etc.

With enough data, the results of Bayesian and frequentist approaches tend to agree.

1.3.1 Probability

There are at least three different approaches to the question of what probability is. The first is the mathematical one, which is based on axioms e.g. it must lie in the range 0 to 1; the probabilities of an event occurring and of it not

¹If the simulation yields a change in the result of $a \pm b$, there is much discussion about how the contribution to the systematic error should be assessed in terms of a and b – see ref. [19].

occurring add up to 1; etc. It does not give much feeling for what probability is, but it does provide the underpinning for the next two methods.

Frequentists, not surprisingly, define probability in terms of frequencies in a long series of essentially identical repetitions² of the relevant procedure. Thus the probability of the number 5 being uppermost in throws of a die is 1/6, because that is the fraction of times we expect (or approximately observe) it to happen. This implies that probability cannot be defined for a specific event (Will the first astronaut who lands on Mars return to earth alive?) or for the value of a physical constant (Is the fraction of dark matter in the Universe larger than 28%?).

In contrast, Bayesians define probability in terms of degree of belief. Thus it can be used for unique events or for the values of physical constants. It can also vary from person to person, because my information may differ from yours. The numerical value of the probability to be assigned to a particular statement is determined by the concept of a ‘fair bet’; if I think the probability (or ‘Bayesian credibility’) of the statement being true is 20%, then I must offer odds of 4-to-1, and allow you to bet in either direction.

This difference in approach to probability affects the way Bayesians and frequentists deal with statistical procedures. We illustrate this below by considering parameter determination.

1.3.2 Bayesian approach

The Bayesian approach makes use of Bayes’ Theorem:

$$p(A|B) = p(B|A) \times p(A)/p(B), \quad (1)$$

where $p(A)$ is the probability or probability density of A , and $p(A|B)$ is the conditional probability for A , given that B has happened. This formula is acceptable to frequentists, provided the probabilities are acceptable frequentist probabilities. However Bayesians use it with $A = \text{parameter}$ (or hypothesis) and $B = \text{data}$. Then

$$p(\text{parameter}|\text{data}) \sim p(\text{data}|\text{parameter}) \times p(\text{parameter}), \quad (2)$$

where the three terms are respectively the Bayesian posterior, the likelihood function and the Bayesian prior. Thus Bayes’ theorem enables us to use the data (as encapsulated in the likelihood) to update our prior knowledge ($p(\text{parameter})$); the combined information is given by the posterior.

Frequentists object to the use of probability for physical parameters. Even Bayesians agree that it is often hard to specify a sensible prior. For a parameter which has been well determined in the past, a prior might be a Gaussian distribution of appropriate central value and width, but for the case where no useful information is available the choice is not so clear; it may be problematic to try to quantify prior ignorance. The ‘obvious’ choice of a uniform distribution has

²Bayesians attack this concept of ‘essentially identical trials’, claiming that it is hard to define it without using the concept of probability, thus making the definition circular.

the problem of being not unique (Should our lack of knowledge concerning, for example, the mass of a neutrino m_ν be parametrised by a uniform prior for m_ν , or for m_ν^2 or for $\log m_\nu$, etc?). Also a uniform prior over an infinite parameter range cannot be normalised. For situations involving several parameters, the choice of prior becomes even more problematic.

It is important to check that conclusions about possible parameter ranges are not dominated by the choice of prior. This can be achieved by changing to other ‘reasonable’ priors (sensitivity analysis); or by looking at the posterior when the data has been removed.

1.3.3 Frequentist approach: Neyman construction

The frequentist way of constructing intervals completely eliminates the need for a prior, and avoids considering probability distributions for parameters. Consider a measurement x which provides information concerning a parameter μ . For example, we could use a month’s data from a large solar neutrino detector (x) to estimate the temperature at the centre of the sun (μ). It is assumed that enough is known about solar physics, fusion reactions, neutrino properties, the behaviour of the detector, etc. that, for any given value of μ , the probability density for every x is calculable. Then for that μ , we can select a region in x which contains, say, 90% of this probability. If we do this for every μ , we obtain a 90% confidence band; it shows the values of x which are likely results of the experiment for any μ , assuming the theory is correct (see fig. 1). Then if the actual experiment gives a measurement x_2 , it is merely necessary to find the values of μ for which x_2 is in the confidence band. This is the Neyman construction.

Of course, the choice of a region in x to contain 90% of the probability is not unique. The one shown in fig. 1 is a central one, with 5% of the probability on either side of the selected region. Another possibility would be to have a region with 10% of the probability to the left, and then the region in x extends up to infinity. This choice would be appropriate if we always wanted to quote upper limits on μ . Other choices of ‘ordering rule’ are also possible (see, for example, Section 4.2).

The Neyman construction can be extended to more parameters and measurements, but in practice it is very hard to use it when more than two or three parameters are involved; software to perform a Neyman construction efficiently in several dimensions would be very welcome. The choice of ordering rule is also very important. Thus from a pragmatic point of view, even ardent frequentists are prepared to use Bayesian techniques. They would, however, like to ensure that the technique they use provides parameter intervals with reasonable frequentist coverage.

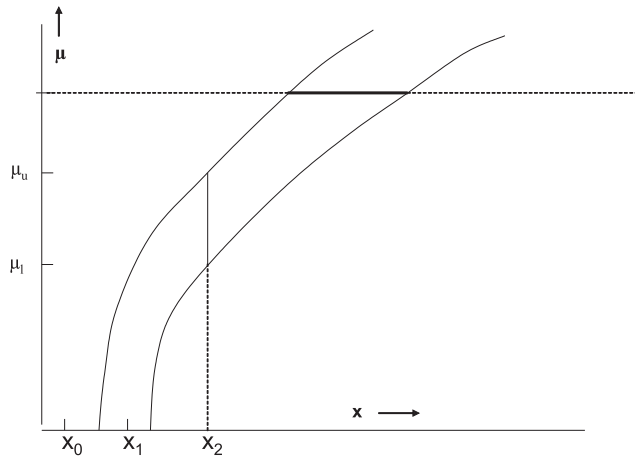


Figure 1: The Neyman construction for setting a confidence range on a parameter μ . At any value of μ , it is assumed that we know the probability density for obtaining a measured value x . We can then choose a region in x which contains, say, 90% of the probability; this is denoted by the solid part of the horizontal line. By repeating this procedure for all possible μ , the band between the curved lines is constructed. This confidence band contains the likely values of x for any μ . For a particular measured value x_2 , the confidence interval from μ_l to μ_u gives the range of parameter values for which that measured value was likely. For x_2 , this interval would be two-sided, while for a lower value x_1 , an upper limit would be obtained. In contrast, there are no parameter values for which x_0 is likely, and for that measured value the confidence interval would be empty.

1.3.4 Coverage

One of the major advantages of the frequentist method is that it guarantees coverage. This is a property of a statistical technique³ for calculating intervals, and specifies how often the interval contains the true value μ_t of the parameter. This can vary with μ_t . Coverage is not guaranteed for non-frequentist methods (see Section 2.1). Interesting plots of coverage as a function of the parameter value for the simple case of a Poisson counting experiment can be found in ref. [39].

³It is important to realise that coverage is a property of the **method**, and not of an **individual measurement**.

1.3.5 Likelihoods

The likelihood approach makes use of the probability density function (*pdf*) for observing the data, evaluated for the data actually observed⁴. It is a function of any parameters, although it does not behave like a probability density for them. It provides a method for determining values of parameters. These include point estimates for the ‘best’ values, and ranges (or contours in multi-parameter situations) to characterise the uncertainties. It usually has good properties asymptotically, but a major use is with sparse multi-dimensional data.

The likelihood method is neither frequentist nor Bayesian. It thus does not guarantee frequentist coverage or Bayesian credibility. It does, however, play a central role in the Bayesian approach, which obtains the posterior probability density by multiplying the likelihood by the prior. The Bayesian approach thus obeys the likelihood principle, which states that the only way the experimental data affects inference is via the likelihood function. In contrast, the Neyman construction requires not only the likelihood for the actual data, but also for all possible data that might have been observed.

Because it is not a probability density, it does not transform like one. Thus the value of the likelihood for a parameter μ_0 is identical to that for $\lambda_0 = 1/\mu_0$. This means that ratios of likelihoods (or differences in their logarithms) are useful to consider, but that the integration of tails of likelihoods is not a recognised statistical procedure.

A longer account of the Bayesian and frequentist approaches can be found in ref. [20]. Ref. [21] provides a very readable account for a Poisson counting experiment.

2 Likelihood issues

In this section, we discuss some potential misunderstandings of likelihoods.

2.1 $\Delta(\ln L) = 0.5$ rule

In the maximum likelihood approach to parameter determination, the best value λ_0 of a parameter is determined by finding where the likelihood maximises; and its error is estimated by finding how much the parameter must be changed⁵ in order for the logarithm of the likelihood to decrease by 0.5 as compared with

⁴The *pdf* $f(x, \mu_0)$ gives the probability density for obtaining various data x when the parameter has some specified value μ_0 . The likelihood is the same function of two variables $f(x_0, \mu)$, but now with x_0 fixed at the data actually obtained, and μ regarded as the variable.

⁵If there are more than just one parameter, the likelihood must of course be remaximised with respect to all the other parameters when looking for the $\Delta(\ln L) = 0.5$ points. Alternatively, a region in multi-parameter space can be selected by finding the contour at which $\Delta(\ln L)$ decreases from its maximum by an amount which depends on the number of parameters.

the maximum⁶. From a frequentist viewpoint, this should ideally result in the parameter range having 68% coverage. That is, in repeated use of this procedure to estimate the parameter, 68% of the intervals should contain the true value of the parameter, whatever its true value happens to be.

If the measurement is distributed about the true value as a Gaussian with constant width, the likelihood approach will yield exact coverage, but in general this is not so. For example, Heinrich[39] has investigated the properties of the likelihood approach (and other methods too) to estimate μ , the mean of a Poisson, when n_{obs} events are observed. Because n_{obs} is a discrete variable, the coverage is a discontinuous function of μ , and varies from 100% at $\mu = 0$ down to 30% at $\mu \approx 0.5$.⁷

2.2 Unbinned maximum likelihood and goodness of fit

With sparse data, the unbinned likelihood method is a good one for estimating parameters of a model. In order to understand whether these estimates of the parameters are meaningful, we need to know whether the model provides an adequate description of the data. Unfortunately, as emphasised by Heinrich[23], the magnitude of the unbinned maximum likelihood is often insensitive to whether or not the data agree with the model. He illustrates this by the example of the determination of the lifetime τ of a particle whose decay distribution is $(1/\tau) \exp(-t/\tau)$. For a set of observed times t_i , the maximum likelihood L_{max} depends on the data t_i only through their average value \bar{t} . Thus any data distributions with the same \bar{t} would give identical L_{max} , which demonstrates that, at least in this case, L_{max} gives no discrimination about whether the data are consistent with the expected distribution.

Another example is fitting an expected distribution $(1 + \alpha \cos^2 \theta)/(1 + \alpha/3)$ to data θ_i on the decay angle of some particle, to determine α . According to the expected distribution, the data should be symmetrically distributed about $\cos \theta = 0$. However, the likelihood depends only on the **square** of $\cos \theta$, and so would be insensitive to all the data having $\cos \theta_i$ negative.

2.3 Punzi effect

Sometimes we have two or more nearby peaks, and we try to fit our data in order to determine the fractions of each peak. Punzi[24] has pointed out that it is very easy to write down a plausible but incorrect likelihood function that gives a biased result. This occurs in situations where the events have experimental resolutions σ in the observable x that vary event-by-event; and the distributions of σ are different for the two peaks.

For a set of observations x_i , it is tempting but wrong to write the unbinned

⁶This (like several other methods) can give rise to asymmetric errors. Techniques for dealing with such errors have been discussed by Barlow[22].

⁷It is of course not surprising that methods that are expected to have good asymptotic behaviour may not display optimal properties for $\mu \approx 0$.

likelihood as

$$L(f)_{wrong} = \Pi\{f * G(x_i, 0.0, \sigma_i) + (1 - f) * G(x_i, 1.0, \sigma_i)\} \quad (3)$$

where f is the fraction of the first peak (labelled A below) which is parametrised as $G(x_i, 0.0, \sigma_i)$, a Gaussian in x_i , centred on zero, and with width σ_i , and i is the label for the i^{th} event; and similarly for the second peak (labelled B), except that it is centred at unity.

Application of the rules of conditional probability shows that the correct likelihood is

$$L(f)_{right} = \Pi\{f * G(x_i, 0.0, \sigma_i) * p(\sigma_i|A) + (1 - f) * G(x_i, 1.0, \sigma_i) * p(\sigma_i|B)\} \quad (4)$$

where $p(\sigma_i|A)$ and $p(\sigma_i|B)$ are the probability densities for the resolution being σ_i for the A and B peaks respectively. We then see that $L(f)_{wrong}$ and $L(f)_{right}$ give identical values for f , provided that $p(\sigma_i|A) = p(\sigma_i|B)$. If however, the distributions of the resolution differ, $L(f)_{wrong}$ will in general give a biased estimate.

Punzi investigated the extent of this bias in a simple Monte Carlo simulation, and it turns out to be surprisingly large. For example, with $f = 1/3$, and $p(\sigma_A)$ and $p(\sigma_B)$ being δ -functions at 1.0 and 2.0 respectively, the fitted value of f turned out to be 0.65. Given that f is confined to the range from zero to unity, this is an enormous bias.

The way the bias arises can be understood as follows: The fraction f of the events that are really A have relatively good resolution, and so the fit to them alone would assign essentially all of them as belonging to A i.e. these events alone would give $f \approx 1$ with a small error. In contrast the $1 - f$ of the events that are B have poor resolution, so for them the fit does not mind too much what is the value of f . But the fit uses all the events together, and so assigns a single f to the complete sample; this will be some sort of weighted average of f of the values for the A and the B events. Because the A events result in a more accurate determination of f than do the B events, the fitted f will be biased upwards (i.e. it will over-estimate the fraction of events corresponding to the peak with the better resolution).

The Punzi effect can also appear in other situations, such as particle identification. Different particle types (e.g. pions and kaons) would appear as different peaks in the relevant particle-identification variable e.g. time of flight, rate of energy loss dE/dx , angle of Cerenkov radiation, etc. The separation of these peaks for the different particle types depends on the momentum of the particles (see fig. 2). So here the Punzi bias can arise even with constant resolution, because the momentum spectra of pions and kaons can be different. To avoid the bias, the likelihood needs to incorporate information on the different momentum distributions of pions and kaons. If these momentum distributions are different enough from each other and are specified more precisely than is justified by the available information, it could be that the likelihood function bases its separation of the different particle types on the momenta of the particles rather than on the data from the detector's particle identifier. Catastini and Punzi[25] avoid

this by using parametric forms for the momentum distributions of the particles, with the parameters being determined by the data being analysed.

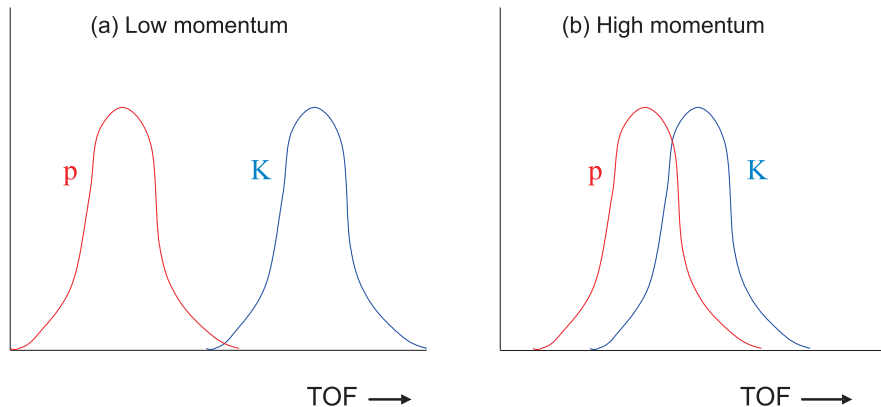


Figure 2: The Punzi effect in particle identification. The diagrams show the expected (normalised) distributions of the output signal from a particle identifier, for pions and for kaons (a) at low momentum where separation is easier, and (b) at high momentum where the distributions overlap. Because kaons are heavier than pions, they tend to have larger momenta. The likelihood function does not care very much whether high momentum tracks are classified as pions or kaons, and hence the fraction of high momentum tracks classified as kaons will have a large uncertainty. In contrast, low momentum tracks will be correctly identified. Thus if the plausible but incorrect likelihood function that ignores the momentum distributions is used to determine the overall fraction of kaons, it will be biased downwards towards the fraction of low momentum particles that are kaons.

The common feature potentially leading to bias in these two examples is that the ratio of peak separation to resolution is different for the two types of objects. For the first example of separating the two peaks, it was the denominators that were different, while in the particle identification problem it was the numerators.

The Punzi bias may thus occur in situations where the templates in a multi-component fit depend on additional observations whose distributions are not explicitly included in the likelihood.

3 Separating Signal from Background

Almost every Particle Physics analysis uses some technique for separating signal from background. First some simple ‘cuts’ are applied; these are generally loose selections on single variables, which are designed to remove a large fraction of the background while barely reducing the signal. Then to obtain a better separation of signal from background in the multi-dimensional space of the event

observables, methods like Fisher discriminants, decision trees, artificial neural networks (including Bayesian nets), support vector machines, etc. are used[26, 27]. More recently, bagging, boosting and random forests have been used to achieve improved performance of the separation as seen on a plot of signal efficiency against background mis-acceptance rate. A description of the software available for implementing some of these techniques can be found in the talks by Narsky[28] and by Tegenfeldt[29] at the PHYSTAT-LHC Workshop.

The signal-to-background ratio before this multi-variate stage can vary widely, as can the signal purity after it. If some large statistics study is being performed (e.g. to use a large sample of events to perform an accurate measurement of the lifetime of some particle), then it is not a disaster if there is some level of background in the finally selected events, provided that it can be accurately assessed and allowed for in the subsequent analysis. At the other extreme, the separation technique may be used to see if there is any evidence for the existence of some hypothesised particle (the potential signal), in the presence of background from well-known sources. Then the actual data may in fact contain no observable signal.

These techniques are usually ‘taught’ to recognise signal and background by being given examples consisting of large numbers of events of each type. These may be produced by Monte Carlo simulation, but then there is a problem of trying to verify that the simulation is a sufficiently accurate representation of reality. It is better to use real data for this, but the difficulty then is to obtain sufficiently pure samples of background and signal. Indeed, for the search for a new particle, true data examples do not exist. However, it is the accurate representation of background that is likely to pose a more serious problem.

The way that, for example, neural networks are trained is to present the software with approximately equal numbers of signal and background events⁸ and then to minimise the cost function C for the network. This is defined as $C = \Sigma(z_i - t_i)^2$, where z_i is the trained network’s output for the i^{th} event; t_i is the target output, usually chosen as 1 for signal and zero for background; and the summation is over all testing events presented to the network. The problem with this is that C is only loosely related to what we really want to optimise. For a search for a new particle this could be the sensitivity of the experimental upper limit in the absence of signal, while for a high statistics analysis measuring the properties (such as mass or lifetime) of some well-established particle, we would be interested in minimising the error (including systematic effects) on the result.

Some open questions are:

- How can we check that our multi-dimensional training samples for signal and background are reliable descriptions of reality?
- How should the ratio of the numbers of signal and background training

⁸For searches for rare processes, it is clearly inappropriate to use the actual fractions expected in the data to determine the ratio of signal to background Monte Carlo events to be used as the training sample, because the network could then achieve an excellent score simply by classifying everything as background.

events be chosen, especially when there are several different sources of background?

- What is the best way of allowing for nuisance parameters in the models of the signal and/or background?[17, 30]
- Are there useful and easy ways of optimising on what is really of interest?[31]

4 Upper Limits

Almost all searches for new phenomena have not found any evidence for exciting new physics. Recent examples from Particle Physics include searches for the Higgs boson, supersymmetric particles, dark matter, etc; attempts to find sub-structure of quarks or leptons; looking for extra spatial dimensions; measuring the mass of a neutrino; etc. Rather than just saying that nothing was found, it is more useful to quote an upper limit on the sought-for effect, as this could be useful in ruling out some theories. For example in 1887, Michelson and Morley attempted to measure the speed of the Earth with respect to the aether. No effect was seen, but the experiment was sensitive enough to lead to the demise of the aether theory.

A simple scenario is a counting experiment where a background b is expected from conventional sources, together with the possibility of an interesting signal s . The number of counts n observed is expected to be Poisson distributed with a mean $\mu = \epsilon * s + b$, where ϵ is a factor for converting the basic physics parameter s into the number of signal events expected in our particular experiment; it thus allows for experimental inefficiency, the experiment's running time; etc. Then given a value of n which is comparable to the expected background, what can we say about s ? The true value of the parameter s is constrained to be non-negative. The problem is interesting enough if b and ϵ are known exactly; it becomes more complicated when only estimates with uncertainties σ_b and σ_ϵ are available.

Even without the nuisance parameters, a variety of methods is available. These include likelihood, χ^2 , Bayesian with various priors for s , frequentist Neyman constructions with a variety of ordering rules for n , and various *ad hoc* approaches. The methods give different upper limits for the same data⁹. A comparison of several methods can be found in ref. [32]. The largest discrepancies arise when the observed n is less than the expected background b , presumably because of a downward statistical fluctuation. The following different behaviours of the limit (when $n < b$) can be obtained:

- Frequentist methods can give **empty** intervals for s i.e. there are no values of s for which the data are likely. Particle Physicists tend to be unhappy when their years of work result in an empty interval for the parameter of

⁹By coincidence, the values obtained by the Bayesian approach with an (improper) flat prior for s and by the Neyman construction for upper limits agree when $b = 0$.

interest, and it is little consolation to hear that frequentist statisticians are satisfied with this feature, as it does not lead to undercoverage.

When n is not quite small enough to result in an empty interval, the upper limit might be **very small**¹⁰. This could confuse people into thinking that the experiment was much more sensitive than it really was.

- The Feldman-Cousins frequentist method[33] (see Section 4.2) that employs a likelihood-ratio ordering rule gives upper limits which **decrease** as n gets smaller at constant b . A related effect is the growth of the limit as b decreases at constant n – this can also occur in other frequentist approaches. Thus if no events are observed ($n = 0$), the upper limit of a 90% Feldman-Cousins interval is 1.08 for $b = 3.0$, but 2.44 for $b = 0$. This is sometimes presented as a paradox, in that if a bright graduate student worked hard and discovered how to eliminate the expected background without much reduction in signal efficiency, they would be ‘rewarded’ by obtaining a weaker upper limit¹¹. An answer is that although the actual limit had increased, the sensitivity of the experiment with the smaller background was better. There are other situations - for example, variants of the random choice of voltmeter (compare ref. [34]) - where a measurement with better sensitivity can on occasion give a less precise result.
- In the Bayesian approach, the dependence of the limit on b is **weaker**. Indeed when $n = 0$, the limit does not depend on b .
- Sen *et al* [35] consider a related problem, of a physical non-negative parameter λ producing a measurement x , which is distributed about λ as a Gaussian of variance σ^2 . As the observable x becomes more and more negative, the upper limit on λ **increases**, because it is deduced that σ must in fact be larger than its quoted value.

In trying to assess which of the methods is best, one first needs a list of desirable properties. These include:

- Coverage: Even though coverage is a frequentist concept, most Bayesian Particle Physicists would like the coverage of their intervals to match their reported credibility, at least approximately.

Because the data in counting experiments is discrete, it is impossible in any sensible way to achieve exact coverage for all μ . However, it is not completely obvious that even Frequentists need coverage for every possible

¹⁰Bayesian methods that use priors with part of the probability density being a δ -function at $s = 0$ can result in a posterior with an enhanced δ -function at zero, such that the upper limit contains only the single point $s = 0$.

¹¹The $n = 0$ situation is perhaps a special case, as the number of observed events cannot decrease as further selections are imposed to reduce the expected background. For non-zero observed events, if n decreases with the tighter cuts (as expected for reduced background), the upper limit is likely to go down, in agreement with intuition. But if n stays constant, that could be because the observed events contain signal, so it is perhaps not surprising that the upper limit increases.

value of μ , since different experiments will have different values of b and of ϵ . Thus even for a constant value of the physical parameter s , different experiments will have different $\mu = \epsilon * s + b$. Thus it would appear that, if coverage in some average (over μ) sense were satisfactory, the frequentist requirement for intervals to contain the true value at the requisite rate would be maintained. This, however, is not the generally accepted view by Particle Physicists, who would like not to undercover for **any** μ .

- Not too much overcoverage: Because coverage varies with μ , for methods that aim not to undercover anywhere, some overcoverage is inevitable. This corresponds to having some upper limits which are high, and this leads to undesirable loss of power in rejecting alternative hypotheses about the parameter's value.
- Short and empty intervals: These can be obtained for certain values of the observable, without resulting in undercoverage. They are generally regarded as undesirable for the reasons explained above.

It is not obvious how to incorporate the above desiderata on interval length into an algorithm that would be useful for choosing between different methods for setting limits.

4.1 Two-sided intervals

An alternative to giving upper limits is to quote two-sided intervals. For example, a 68% confidence interval for the mass of the top quark might be 169 to 173 GeV/ c^2 , as opposed to its 90% upper limit being 174 GeV/ c^2 . Most of the difficulties and ambiguities mentioned above apply in this case too, together with some extra possibilities. Thus, while it is clear which of two possible upper limits is tighter, this is not necessarily so for two-sided intervals, where which is shorter may be metric dependent; the first of two intervals for a particle's lifetime τ may be shorter, but the second may be shorter when the ranges are quoted for its decay rate ($= 1/\tau$). Also there is more scope for choice of ordering rule for the frequentist Neyman construction, or for choosing the interval from the Bayesian posterior probability density¹².

4.2 Feldman-Cousins approach

Feldman and Cousins' fully frequentist approach[33] exploits the freedom available in the Neyman construction of how to choose an interval in the data that contains a given fraction α of the probability, by using their 'ordering rule'. This

¹²A Bayesian statistician would be happy with the posterior as the final result. Particle Physicists like to quote an interval as a convenient summary. For a parameter that cannot be negative and for which the exclusion of zero is interesting (e.g. testing whether the production rate of some hypothesised particle is non-zero), an upper limit would always include zero, a lower limit would exclude it and a maximum probability density one would not be invariant with respect to changes in the functional form of the parameter.

is based on the likelihood ratio $L(x, \mu)/L(x, \mu_{best})$, where μ_{best} is the physically-allowed value of μ which gives the largest value of L for that particular x . For values of μ far from a physical boundary, this makes little difference from the standard central Neyman construction, but near a boundary the region is altered in such a way as to make it unlikely that there will be an empty intervals for the parameter μ ; these can occur in the standard approach (see fig. 3).

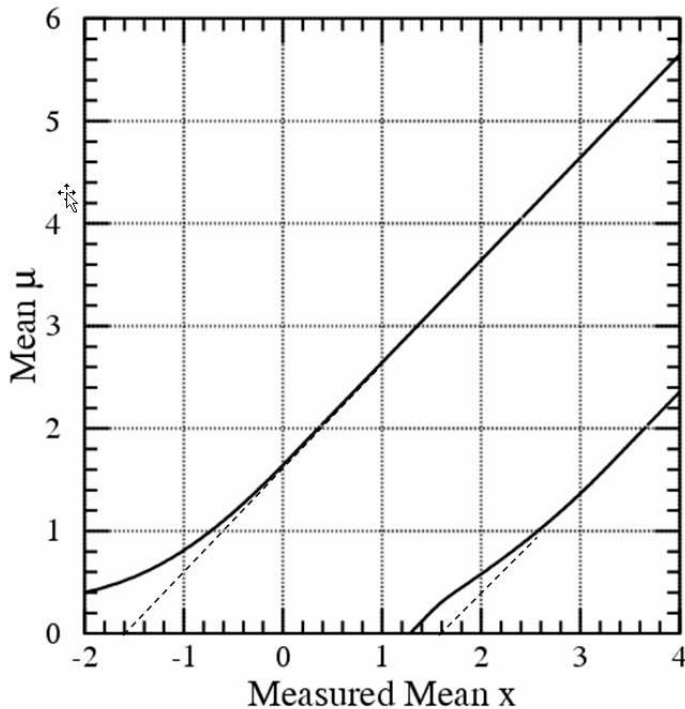


Figure 3: The Feldman-Cousins confidence band (solid curves) for the mean μ of a Gaussian probability density function of unit variance for a measurement x . The straight dashed lines show the confidence band for the central Neyman construction. The Feldman-Cousins ordering rule pulls the interval to the left at small μ , and hence even for negative observed x , the μ interval is not empty.

Feldman and Cousins also point out that an apparently innocuous procedure for choosing what result to quote may lead to undercoverage. Many physicists would quote an upper limit on any possible signal if their observation was not more than 3 standard deviations above the expected background, but a two-sided interval if their result was above this. With each type of interval constructed to give 90% coverage, there are some values of the parameter for which the coverage for this mixed procedure drops to 85%; Feldman and Cousins refer to this as ‘flip-flop’. Their ‘unified’ approach circumvents this problem, as it

automatically yields upper limits for small values of the data, but two-sided intervals for larger measurements, while maintaining correct coverage for all possible true values of the signal.

4.3 Sensitivity

We have already mentioned the idea of quoting the sensitivity of a procedure, as well as the actual upper limit as derived from the observed data¹³. For upper limits or for uncertainties on measurements, this can be defined as the median value that would be obtained if the procedure was repeated a large number of times. Using the median is preferable to the mean because (a) it is metric independent (i.e. the median lifetime upper limit would be the reciprocal of the median decay rate lower limit); and (b) it is much less sensitive to a few anomalously large upper limits or error estimates.

Punzi[36] has drawn attention to the fact that this choice of definition for sensitivity has some undesirable features. Thus designing an analysis procedure to minimise the median upper limit for a search in the absence of a signal provides a different optimisation from maximising the median number of standard deviations for the significance of a discovery when the signal is present. Also there is only a 50% chance of achieving the median result or better. Instead, for pre-defined levels α and CL , Punzi determines at what signal strength there is a probability of at least CL for establishing a discovery at a significance level α . This is what he quotes as the sensitivity, and is the signal strength at which we are sure to be able either to claim a discovery or to exclude its existence. Below this, the presence or otherwise of a signal makes too little difference, and we may remain uncertain (see fig. 4).

4.4 CL_s

This is a technique[37] which is used for situations in which a discovery is not made, and instead various parameter values are excluded. For example the failure to observe the SM Higgs boson can be converted into a mass range for the Higgs which is excluded (at some confidence level).

Fig. 5 illustrates the expected distributions for some suitably chosen statistic under two different hypotheses: the null H_0 in which there is only standard known physics, and H_1 which also includes some specific new particle, such as the Higgs boson. In the simplest case, the statistic could be simply the observed number of events n in some selected region. In fig. 5(c), the new particle is produced prolifically, and an experimental observation of n should fall in one peak or the other, and easily distinguish between the two hypotheses. In contrast, fig. 5(a) corresponds to very weak production of the new particle and it is almost impossible to know whether the new particle is being produced or not.

The conventional method of claiming new particle production would be if n fell well above the main peak of the H_0 distribution; typically a p_0 value

¹³The sensitivity on its own will not do, because it is independent of the data.

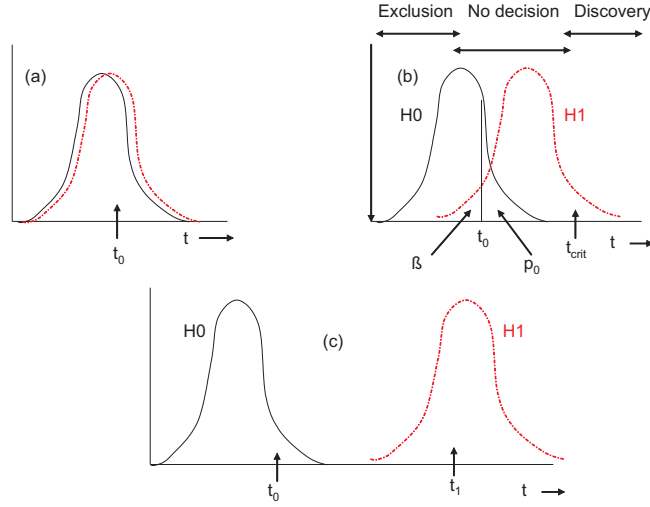


Figure 4: Punzi definition of sensitivity. Expected distributions for a statistic t (which in simple cases could be simply the observed number of events n), for $H_0 =$ background only (solid curves) and for $H_1 =$ background plus signal (dashed curves). In (a), the signal strength is very weak, and it is impossible to choose between H_0 and H_1 . As shown in (b), which is for moderate signal strength, p_0 is the probability according to H_0 of t being equal to or larger than the observed t_0 . To claim a discovery, p_0 should be smaller than some pre-set level α , usually taken to correspond to 5σ ; t_{crit} is the minimum value of t for this to be so. Then the power function $1 - \beta$ (equivalent to p_1 in fig. 5(b)) is the probability according to the alternative hypothesis that t will exceed t_{crit} . According to Punzi, the sensitivity should be defined as the expected production strength of the signal such that $1 - \beta$ exceeds another predefined level CL e.g. 95%. The exclusion region in (b) corresponds to t_0 in the 5% lower tail of H_1 , while the discovery region has t_0 in the 5σ upper tail of H_0 ; there is a “No decision” region in between, as the signal strength in (b) is below the sensitivity value. The sensitivity is thus the signal strength above which there is a 95% chance of making a 5σ discovery. i.e. The distributions for H_0 and H_1 are sufficiently separated that, apart possibly for the 5σ upper tail of H_0 and the 5% lower tail of H_1 , they do not overlap. In (c) the signal strength is so large that there is no ambiguity in choosing between the hypotheses.

corresponding to 5σ would be required (see Section 6.6). In a similar way, new particle production would be excluded if n were below the main part of the H_1 distribution. Typically a 95% exclusion region would be chosen (i.e. $1 - p_1 \leq .05$). The CL_s method aims to provide protection against a downward fluctuation of n in fig 5(a) resulting in a claim of exclusion in a situation where the experiment has no sensitivity to the production of the new particle; this

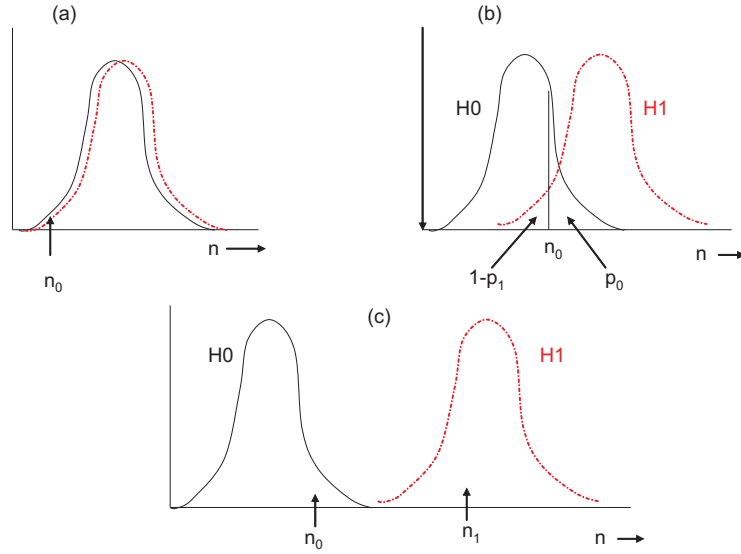


Figure 5: The CL_s method. The expected distributions for a data statistic n are shown (i) for the null hypothesis H_0 of background only (solid curve): and (ii) for H_1 (dashed curve), where there is also some exciting new physics, which tends to result in larger n . In (b), the tail areas of H_0 above the observed n_0 and of H_1 below n_0 are indicated by arrows; they correspond to probabilities p_0 and $1 - p_1$ respectively. Fig. (c) shows a situation where the new physics is strongly produced, and H_0 and H_1 are well separated. Thus n_0 would result in H_1 being excluded, while n_1 would be taken as evidence in favour of new physics. In (a), production is very weak, and the H_0 and H_1 curves are barely distinguishable. In order to protect against a downward fluctuation (statistic = n_0) in a situation like (a) resulting in an exclusion of H_1 when the curves are essentially identical, CL_s is defined as $(1 - p_1)/(1 - p_0)$.

could happen in 5% of experiments. It achieves this by defining¹⁴

$$CL_s = (1 - p_1)/(1 - p_0), \quad (5)$$

and requiring CL_s to be below 0.05. From the definition, it is clear that CL_s cannot be smaller than $1 - p_1$, and hence is a conservative version of the frequentist quantity $1 - p_1$. It tends to $1 - p_1$ when n lies above the H_0 distribution, and to unity when the H_0 and H_1 distributions are very similar.

It is deemed not to be necessary to protect against statistical fluctuations giving rise to discovery claims in situations with no sensitivity, because that

¹⁴Given the fact that CL_s is essentially the ratio of two p-values, the choice of symbol CL_s (standing for ‘confidence level of signal’) is not optimal. Another source of confusion is that in the definition of CL_s , the way the p-values are defined varies, so the formulae can look different but the underlying concept is the same.

should happen only at the $3 * 10^{-7}$ rate.

Most statisticians are appalled by the use of CL_s . This is because they consider that it is meaningless to take the ratio of two p -values. Its appeal to Particle Physicists is the protection it provides against excluding particles from data which have no sensitivity to them. We thus regard it as a conservative frequentist approach.

Other approaches to this problem are mentioned in Section 6.3.

4.5 Nuisance parameters

For calculating upper limits in the simple counting experiment described in Section 4, the nuisance parameters arise from the uncertainties in the background rate b and the acceptance ϵ . These uncertainties are usually quoted as σ_b and σ_ϵ (e.g. $b = 3.1 \pm 0.5$), and the question arises of what these errors mean. Sometimes they encapsulate the results of a subsidiary measurement, performed to estimate b or ϵ , and then they would express the width of the Bayesian posterior or of the frequentist interval obtained for the nuisance parameters. However, in many situations, the errors may be based on a series of subsidiary measurements; they may involve Monte Carlo simulations, which have systematic uncertainties (e.g. related to how well the simulation describes the real data) as well as statistical errors; or they may reflect uncertainties or ambiguities in theoretical calculations required to derive b and/or ϵ . In the absence of further information the posterior is often assumed to be a Gaussian, usually truncated so as to exclude unphysical (e.g. negative) values. This may be at best only approximately true, and deviations are likely to be most serious in the tails of the distribution.

There are many methods for incorporating nuisance parameters in upper limit calculations. These include:

- Profile likelihood

The likelihood, based on the data from the main and from the subsidiary measurements, is a function of the parameter of interest s and of the nuisance parameters. The profile likelihood $L_{prof}(s)$ is simply the full likelihood $L(s, b_{best}(s), \epsilon_{best}(s))$, evaluated at the values of the nuisance parameters that maximise the likelihood at each s . Then the profile likelihood is simply used to extract the limits on s , much as the ordinary likelihood could be used for the case when there are no nuisance parameters.

Rolke *et al* [38] have studied the behaviour of the profile likelihood method for limits. Heinrich[39] had shown that the likelihood approach for estimating a Poisson parameter (in the absence of both background and of nuisance parameters) can have poor coverage at low values of the Poisson parameter. However, the profile likelihood seems to do better, probably because the nuisance parameters have the effect of smoothing away the fluctuating coverage observed by Heinrich.

- Full Bayes

When there is a subsidiary measurement for a nuisance parameter, a prior is chosen for b (or ϵ), the data are used to extract the likelihood, and then Bayes' Theorem is used to deduce the posterior for the nuisance parameter. This posterior from the subsidiary measurement is then used as the prior for the nuisance parameter in the main measurement (this prior could alternatively come from information other than a subsidiary measurement); with the prior for s and the likelihood for the main measurement, the overall joint posterior for s and the nuisance parameter(s) is derived¹⁵. This is then integrated over the nuisance parameter(s) to determine the posterior for s , from which an upper limit can be derived.

Table 1: Bayesian 90% confidence level upper limits for the production rate s as a function of n , the observed number of events. The Poisson parameter $\mu = \epsilon*s+b$, where the expected background b is either 0.0 or 3.0, and is precisely known; and ϵ , whose true values is 1.0, is estimated in a subsidiary measurement with 10% accuracy. The numbers in brackets are the corresponding upper limits when ϵ is known precisely. At large n , the limits for $b = 3.0$ are 3 units lower than those for $b = 0.0$; the latter are approximately $n + 1.28\sqrt{n}$ at large n . The effect of the uncertainty in ϵ is to increase the limits, and by a larger amount at large n . For $n = 0$, these Bayesian limits are independent of the expected background b .

n	$b = 0.0$	$b = 3.0$
0	2.35 (2.30)	2.35 (2.30)
3	6.87 (6.68)	4.46 (4.36)
6	10.88 (10.53)	7.80 (7.60)
9	14.71 (14.21)	11.56 (11.21)
20	28.27 (27.05)	25.05 (24.05)

Numerical examples of upper limits can be found in ref. [40], where a method is discussed in detail. Thus for precisely determined backgrounds, the effect of a 10% uncertainty in ϵ can be seen for various measured values of n in Table 1. A plot of the coverage when the uncertainty in ϵ is 20% is reproduced in fig. 6.

It is not universally appreciated that the choice for the main measurement of a truncated Gaussian prior for ϵ and an (improper) constant prior for non-negative s results in a posterior for s which diverges[41]. Thus numerical estimates of the relevant integrals are meaningless. Another problem comes from the difficulty of choosing sensible multi-dimensional priors. Heinrich has pointed out the problems that can arise for the above

¹⁵This is usually equivalent to starting with a prior for s and the nuisance parameters, and the likelihood for the data from the main and the subsidiary experiments together, to obtain the joint posterior.

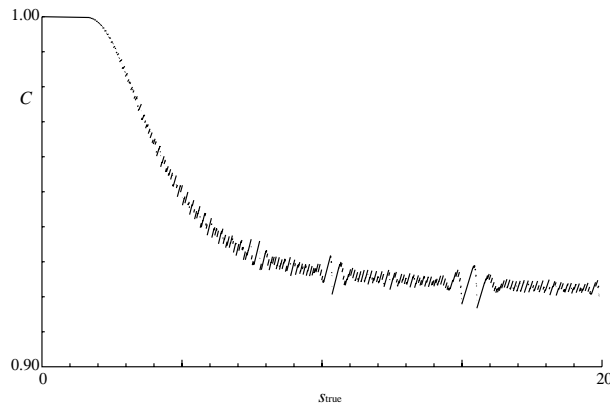


Figure 6: The coverage C for the estimated 90% confidence level upper limit as a function of the true parameter s_{true} . The background $b = 3.0$ is assumed to be known exactly, while the subsidiary measurement for ϵ gives a 20% accuracy. The discontinuities are a result of the discrete (integer) nature of the measurements. There is no undercoverage.

Poisson counting experiment, when it is extended to deal with several data channels simultaneously[42].

- Fully frequentist

In principle, the fully frequentist approach to setting limits when provided with data from the main and from subsidiary measurements is straightforward: the Neyman construction is performed in the multidimensional space where the parameters are s and the nuisance parameters, and the data are from all the relevant measurements. Then the region in parameter space for which the observed data was likely is projected onto the s -axis, to obtain the confidence region for s .

In practice there are severe difficulties in writing a program to do this in a reasonable amount of time. To date, the largest number of parameters used is three[43]. Another problem is that, unless a clever ordering rule is used for producing the acceptance region in data space for fixed values of the parameters, the projection phase leads to overcoverage, which can become larger as the number of nuisance parameters increases. Good ordering rules have been found for a version of the Poisson counting experiment[44], and for the ratio of Poisson means[45], where the confidence intervals are tighter than those obtained by conditioning on the sum of the numbers of counts in the two observations.

For the fully frequentist method, it is guaranteed that there will be no undercoverage for any combination of parameter true values. This is not so for any other method, and so most Particle Physicists would like assurance

that the technique used does indeed provide reasonable coverage, at least for s . There is usually lively debate between frequentist and Bayesians as to whether coverage is desirable for all values of the nuisance parameter(s), or whether one should be happy with no or little undercoverage when experiments are averaged over the nuisance parameter true values.

- **Mixed**

Because of the difficulty of performing a fully frequentist analysis in all but the simplest problems, an alternative approach[46] is to use Bayesian averaging over the nuisance parameters, but then to employ a frequentist approach for s . The hope is that for most experiments setting upper limits, the statistical errors on the low n data are relatively large and so, provided the uncertainties in the nuisance parameters are not too large, the effect of the systematics on the upper limits will not be too large, and so an approximate method of dealing with them may be reasonable.

Although such an approach cannot be justified from fundamentals, it provides a practical method whose properties can be checked, and are often satisfactory.

4.6 Banff challenge

Given the large number of techniques available for extracting upper limits from data, especially in the presence of nuisance parameters, it was decided at the Banff meeting[9] that it would be useful to compare the properties of the different approaches under comparable conditions. This led to the setting up of the ‘Banff Challenge’, which consisted of providing common data sets for anyone to calculate their upper limits. This was organised by Joel Heinrich, who reported on the performance of the various methods at the PHYSTAT-LHC meeting[47].

4.7 Recommendations

It would be incorrect to say that there is one method that must be used. What is important is that the procedure should be fully defined before the data are analysed; and that when the experimental result and the sensitivity of the search are reported, the method used should be fully explained.

The CDF Statistics Committee [48] also suggests that it is useful to use a technique that has been employed by other experiments studying the same phenomenon; this makes for easier comparison. They tend to favour a Bayesian approach, chiefly because of the ease of incorporating nuisance parameters.

5 Goodness of fit

5.1 Sparse multi-dimensional data

The standard method loved by Particle Physicists is χ^2 . This, however, is only applicable to binned data (i.e. in a one or more dimensional histogram). Fur-

thermore it loses its attractive feature that its distribution is model-independent when there is not enough data, which is likely to be so in the multi-dimensional case.

Although the maximum likelihood method is very useful for parameter determination with unbinned data, the value of L_{max} usually does not provide a measure of goodness of fit (see Section 2.2).

An alternative that is used for sparse one-dimensional data is the Kolmogorov-Smirnov (KS) approach, or one of its variants. However, in the presence of fitted parameters, simulation is again required to determine the expected distribution of the KS-distance. Also because of the problem of how to order the data, the way to use it in multi-dimensional situations is not unique.

Aslan and Zech[49] have described a method that can be used with sparse multi-dimensional data¹⁶. It compares two separate sets of events, which could be data and simulations based on a theoretical model; or two sets of data taken under slightly different conditions; etc. The first set of points are assigned positive electric charges, and the second set negative ones, and then the “electrostatic energy” of the system is calculated as $E = \sum \sum q_i * q_j * f(d_{ij})$, where the summation extends over all pairs of observations; q_i is the charge of the i^{th} observation; and $f(d_{ij})$ is a function of the distance d_{ij} between observations i and j . For real electrostatics in 3 dimensions, $f(d)$ is proportional to $1/d$, but here it can be chosen to give desirable behaviour; Aslan and Zech favour $-\ln(d + \epsilon)$, where ϵ is a small constant to avoid problems as d tends to zero. This method requires the choice of a metric for each of the observables, and it also needs simulation to determine the expected distribution of E assuming the two distributions are identical. Aslan and Zech find that their method compares favourably with other approaches (e.g χ^2 , KS and its variants, etc.) in rejecting alternative hypotheses in various one-dimensional problems.

5.2 Number of degrees of freedom

If we construct the weighted sum of squares S between a predicted theoretical curve and some data in the form of a histogram, provided the Poisson distribution of the bin contents can be approximated by a Gaussian (and the theory is correct, the data are unbiased, the error estimates are correct, etc.), **asymptotically**¹⁷ S will be distributed as χ^2 with the number of degrees of freedom $\nu = n - f$, where n is the number of data points and f is the number of free parameters whose values are determined by minimising S .

The relevance of the asymptotic requirement can be seen by imagining fitting a more or less flat distribution by the expression $N(1 + 10^{-6} \cos(x - x_0))$, where the free parameters are the normalisation N and the phase x_0 . It is clear that, although x_0 is left free in the fit, because of the 10^{-6} factor, it will have a negligible effect on the fitted curve, and hence will not result in the typical reduction in S associated with having an extra free parameter. Of course,

¹⁶A similar approach can be found in the statistics literature[50].

¹⁷The examples in this section go beyond the requirement that we need enough events for the Poisson distribution to be well approximated by a Gaussian.

with an enormous amount of data, we would have sensitivity to x_0 , and so asymptotically it does reduce ν by one unit, but not for smaller amounts of data.

Another example involves the search for neutrino oscillations. The neutrino energy spectrum is fitted by a survival probability P of the form

$$P = 1 - \sin^2 2\theta \sin^2(C * \Delta m^2), \quad (6)$$

where C is a known function of the neutrino energy and the length of its flight path, θ is the neutrino mixing angle, and Δm^2 is the difference in mass squared of the relevant neutrino species. For small values of $C * \Delta m^2$,

$$P \approx 1 - \sin^2 2\theta (C * \Delta m^2)^2 \quad (7)$$

Thus the survival probability depends only on the two parameters in the combination $\sin 2\theta \Delta m^2$. Because this combination is all that we can hope to determine, we effectively have only one free parameter rather than two. Of course, an enormous amount of data can manage to distinguish between $\sin(C * \Delta m^2)$ and $C * \Delta m^2$, and so asymptotically we have two free parameters as expected.

6 Discovery Issues

Searches for new particles are an exciting endeavour, and will play an even bigger role with the start-up of the LHC at CERN, expected in 2008. The 2007 PHYSTAT Workshop at CERN[10] was devoted to statistical issues that arise in discovery-oriented analyses.

6.1 H_0 , or H_0 versus H_1 ?

In looking for new physics, there are two distinct types of approach. We can compare our data just with the null hypothesis H_0 , the SM of Particle Physics; alternatively we can see whether our data are more consistent with H_0 or with an alternative hypothesis H_1 , some specific manifestation of new physics, such as a particular form of quark and/or lepton substructure. The former is known as ‘goodness of fit’, while the term ‘hypothesis testing’ is often reserved for the latter.

Each of these approaches has its own advantage. By not specifying a specific alternative¹⁸, the goodness of fit test may be capable of detecting any form of deviation from the SM. On the other hand, if we are searching for some specific new effect, a comparison of H_0 and H_1 is likely to be a more sensitive way for that particular alternative. Also, the ‘hypothesis testing’ approach is less likely to give a false discovery claim if the assumed form of H_0 has been slightly mis-modelled.

¹⁸Even a test of the null hypothesis may not be completely independent of ideas about alternatives. Thus in an event counting experiment, new physics usually results in an **increase** in rate, unless we are looking for neutrino oscillations, in which case a **decrease** would be significant. Also, sometimes the statistic used for a goodness of fit test of H_0 may be the likelihood ratio for H_0 as compared with a specific alternative H_1 .

6.2 p -values

In order to quantify the chance of the observed effect being due to an uninteresting statistical fluctuation, some statistic is chosen for the data. The simplest case would be the observed number n_0 of interesting events. Then the p -value is calculated, which is simply the probability that, given the expected background rate b from known sources, the observed number of events would fluctuate up to n_0 or larger. A small value of p indicates that the data is not very compatible with the theory (which may be because we do not understand our detector or the background, rather than the theory being wrong).

Particle physicists usually convert p into the number of standard deviations σ of a Gaussian distribution, beyond which the one-sided tail area corresponds to p . Thus 5σ corresponds to a p -value of $3 * 10^{-7}$. This is done simply because it provides a number which is easier to remember, and not because Gaussians are relevant for every situation.

Unfortunately, p -values are often misinterpreted as the probability of the theory being true, given the data. It sometimes helps colleagues clarify the difference between $p(A|B)$ and $p(B|A)$ by reminding them that the probability of being pregnant, given the fact that you are female, is considerably smaller than the probability of being female, given the fact that you are pregnant.

6.3 Comparing two hypotheses via χ^2

Assume we have a histogram with 100 bins, and that we are using a χ^2 method for fitting it with a function with one free parameter. We expect to obtain a χ^2 value of 99 ± 14 . Thus if p_0 , the best value of the parameter, yields a χ^2 of 85, we would regard that as very satisfactory. However, a theoretical colleague has a model which predicts that the parameter should have a different value p_1 , and wants to know what the data have to say about that. We test this by calculating the χ^2 for that p_1 and obtain a value of 110. We appear to have two contradictory conclusions:

- p_1 is satisfactory: This is based on the fact that the relevant χ^2 of 110 is well within the expected range of 99 ± 14 .
- p_1 is ruled out: The uncertainty on p is estimated by seeing how much it must change from its optimum value in order to make χ^2 increase by 1. For this data, $\chi^2(p_1)$ is **25** units larger than $\chi^2(p_0)$, and so, assuming that the behaviour of χ^2 in the neighbourhood of the minimum is parabolic, p_1 is ruled out at the ~ 5 standard deviation level.

Unfortunately, many physicists, over-impressed by the fact that $\chi^2(p_1)$ appears to be satisfactory, are reluctant to accept that p_0 is strongly favoured by the data.

A similar argument applies to comparing a given set of data with 2 separate hypotheses e.g. fitting a histogram with an exponential or a straight line. Again the **difference** between the χ^2 quantities provides better discrimination

between the hypotheses than does the **individual** χ^2 [51]. Another example of using the difference in χ^2 's is given in the next Section.

There are of course other ways available for comparing two hypotheses. e.g. likelihood ratio, Bayes factor¹⁹, Bayesian information criterion, etc. A discussion of their application in cosmology can be found, for example, in ref. [52].

6.4 Peak above smooth background

When comparing two hypotheses with our data, we can use the numerical values of the two χ^2 quantities with a view to making some decision about the hypotheses. For example, we may be fitting a smooth distribution by a power series, and wonder whether we need a quadratic term, or whether a linear expression would suffice. Alternatively we may want to assess whether a mass spectrum favours the existence of a peak on top of a smooth background, as compared with just the smooth background. Qualitatively, if the extra term(s) are unnecessary, they will result in a relatively small reduction in χ^2 , while if they really are required, the reduction could be larger.

It is sometimes possible to be quantitative about the expected reduction when the extra terms are not needed[53]. If we are in the asymptotic regime, and if the hypotheses are nested²⁰, and if the extra parameters of the larger hypothesis are defined under the smaller one, and in that case do not lie on the boundary of their allowed region, then the difference in χ^2 should itself be distributed as a χ^2 , with the number of degrees of freedom equal to the number of extra parameters.

An example that satisfies this is provided by the different order polynomials. The hypotheses are nested, in that the linear situation is a special case of a quadratic, where the coefficient of the quadratic term is zero. Thus the extra parameter is defined and within the (infinite) allowed range. Then, provided we have a large amount of data, we expect the difference in χ^2 to have one degree of freedom, so a value larger than around 5 would be unlikely.

A contrast is provided by a smooth background $C(x)$ compared with a background plus peak, $C(x) + A \exp[-0.5 * (x - x_0)^2 / \sigma^2]$. The extra parameters for the peak are its amplitude, position and width: A , x_0 and σ respectively. Again the hypotheses are nested, in that $C(x)$ is just a special case of the peak plus background, with $A = 0$. However, although A is defined in the background only case, x_0 and σ are not, as their values become completely irrelevant when $A = 0$. Furthermore, unless the peak plus background fit allows A to be negative, zero is on the boundary of its allowed region. We thus should not expect the difference of the χ^2 quantities itself to be distributed as a χ^2 [58, 54, 55]. To assess the significance of a particular χ^2 difference, this unfortunately means that we have to obtain its distribution ourselves, presumably by Monte Carlo. If we want to find out probabilities of statistical fluctuations at the 10^{-6} level,

¹⁹For two hypotheses H_1 and H_2 with parameters θ_i , this is the ratio of the marginalised likelihoods $\int p(x|\theta_i, H_i) p(\theta_i|H_i) d\theta_i$ for the two hypotheses, where x are the data

²⁰This means that for suitable values of the parameters the larger hypothesis reduces to the smaller one.

this requires a lot of simulation, and probably needs us to use something better than brute force.

The problem of non-standard limiting distributions for χ^2 tests has a substantial statistical literature (see, for example, refs. [56] and [57].)

6.5 Incorporating nuisance parameters

The calculation of p -values is complicated in practice by the existence of nuisance parameters. (For the simple situation described in Section 6.2, there could be some uncertainty in the estimated background.) There are numerous ways of incorporating them. These include:

- Conditioning: For example, with a single nuisance parameter, it may be possible to condition on the sum of the number of counts in the main and the subsidiary experiments, and then to use the binomial distribution to obtain the p -value.
- Plug-in p -value: The best estimate of the nuisance parameter under the null hypothesis is used to calculate p .
- Prior predictive p -value: The p -values are averaged over the nuisance parameters, weighted by their prior distributions.
- Posterior predictive p -value: This time, the posterior distributions of the nuisance parameters are used for weighting.
- Supremum p -value: The largest p -value for any possible value of the nuisance parameter is used. This is likely to be useful only when the nuisance parameter is forced to be within some range; or when there is only a small number of possible alternative theoretical interpretations.
- Confidence interval: A region of frequentist confidence $1 - \gamma$ is used for the nuisance parameter(s), and then the adjusted p -value is $p_{max} + \gamma$, where p_{max} is the largest p -value as the nuisance parameters are varied over their confidence region. Clearly if it is desired to establish a discovery from p -values around 10^{-7} or smaller, then γ should be chosen at least an order of magnitude below this.

The properties of these and other methods are compared by Demortier [58], while Cranmer [59] has discussed some of them in the context of searches at the LHC, where the distributions in the tails of the probability distributions for data can be very relevant.

The role of systematic effects is likely to be more serious here than for upper limits discussed in Section 4.5. This is because in upper limit situations the number of events is usually small, and so statistical errors dominate. In contrast, discovery claims have p -values of 3×10^{-7} or smaller, and so tails of distributions are likely to be important.

6.6 Why 5σ ?

Unfortunately the usually accepted criterion for claiming a discovery in Particle Physics is that p should correspond to at least 5σ . Statisticians almost invariably ask why we use such a stringent level. One answer is past experience: we have all too often seen interesting effects at the 3σ or 4σ level go away as more data are collected. Another is the multiple comparison problem, or “look elsewhere” effect. While the chance of obtaining a 5σ effect in one bin of a particular histogram is really small, it is to be remembered that histograms have many bins²¹, they could be plotted with different selection criteria and different binning²², and there are very many other histograms that were or could have been looked at in the course of the experiment²³. Thus the chance of a 5σ fluctuation occurring somewhere in the data is much larger than might at first appear. Finally, physicists subconsciously incorporate Bayes’ priors in assessing how likely they feel that they have discovered something new, and hence whether they should claim a discovery. Thus, in deciding between the possibilities of a new discovery or of an undetected systematic effect, our priors might favour the latter, and hence strong evidence for discovery is required from the data²⁴.

However it is not necessarily equitable to use a uniform standard for large general-purpose experiments and for small ones with a specific aim; or for looking for a process which is expected, as compared with a very speculative search. But physicists and journal editors seem to like a defined rule rather than a flexible criterion, so this bolsters the 5σ standard. In any case, it is largely a semantic issue, in that physicists finding a 4.5σ effect would clearly report it, using judiciously chosen wording to describe the status of their observation.

Statisticians also ask whether we really believe our models out into the extreme tails of the distributions. In general, this may be so – counting experiments are expected to follow Poisson distributions, with small corrections for possible long time-scale drifts in detector calibrations; and particle decays usually are described by exponential distributions in time. However, the situation is much less clear for nuisance parameters, where error estimates may be less rigorous, and their distribution is often assumed to be Gaussian (or truncated Gaussian) by default. The effect of these uncertainties on very small p -values

²¹In calculating a p -value in such a case, it is very desirable to take into account the number of chances for a statistical fluctuation to occur anywhere in the histogram. At very least, it should be made clear what the basis of the calculated p -value is.

²²If a blind analysis is performed, such decisions are made before looking at the data, and so this aspect of the “look elsewhere” effect is reduced.

²³The extent to which other people’s searches should be included in an allowance for the “look elsewhere” effect depend subtly on the implied question being addressed. Thus are we considering the chance of obtaining a statistical fluctuation in any of the analyses we have performed; or by anyone analysing data in our experiment; or by any Particle Physicist this year? Anyone observing a possible Higgs signal at the LHC would be very unhappy about having to reduce the significance of their result because of the statistical fluctuations that could occur in speculative searches performed elsewhere.

²⁴If I were performing an experiment to look for violations of energy conservation, I would require more than 5σ , because my prior for energy being conserved is very large.

needs to be investigated case-by-case.

We also have to remember that p -values merely test the null hypothesis. There are more sensitive ways of looking for new physics when we have a specific alternative in mind. Thus a very small p -value on its own is usually not enough to make a convincing case for discovery.

6.7 Repetitions in time

A typical experiment at a large accelerator may collect data over 10 to 15 years. The same search for a new effect will typically be repeated once or twice each year as more data is collected. Does this constitute another factor of ~ 20 in the number of opportunities for a statistical fluctuation to appear? Our reply is “No”. If there had been a 6σ signal with half the data (which resulted in a claim for discovery), which had then become only 3σ with more data, this would be grounds for downplaying the earlier discovery claim. Thus at any time, there is only one set of data (everything) that is relevant.

6.8 Combining p -values

In looking for a given new effect, there may be several separate and uncorrelated analyses which are relevant. These could correspond to different decay modes for the new particle; or different experiments looking for the same signal. Thus, if the p -values for the null hypothesis (i.e. no new physics) for the separate analyses were 10^{-6} and 0.1 , what is the corresponding p -value for the pair of results?²⁵

The unambiguous answer is that there is no unique recipe for combining them[60, 61]. There is no single way of taking a uniform distribution in two variables, and finding a transformation $p_{comb}(p_1, p_2)$ that converts it into a uniform distribution of the single variable p_{comb} .

Two popular recipes involve asking what is the probability that the smaller p -value will be 10^{-6} or smaller; or that the product is below $p_1 * p_2 = 10^{-7}$. None of the possible methods has the property that in combining 3 p -values, the same answer is obtained if p_1 is first combined with p_2 , and then the result is combined with p_3 ; or whether some different ordering is used.

Another problem is the lack of other information that might be relevant. For example, the p -values might arise from χ^2 's with different numbers of degrees of freedom ν e.g. $\chi_1^2 = 90$ for 100 degrees of freedom, and $\chi_2^2 = 20$ for $\nu = 1$. The second has a very small p -value, so many combination methods (including the two mentioned above) would conclude that overall the data do not look consistent with the null hypothesis. However, another plausible-sounding method is to add the separate χ^2 values and also the individual ν ²⁶, to obtain a total

²⁵Rather than combining p -values, it is of course better to use the complete sets of original data (if available) for obtaining the combined result.

²⁶The method described earlier involving the product of the p -values is equivalent to converting each p to a χ^2 , assuming that $\nu = 2$, regardless of whether this was the actual number of degrees of freedom, and then adding the χ^2 and also the ν .

$\chi^2 = 110$ for $\nu = 101$, which sounds perfectly satisfactory. The resolution of this discrepancy of interpretation depends on the nature of the two tests. If the second analysis with $\chi^2 = 20$ corresponded to just one extra measurement like the previous 100, then it seems reasonable to combine the χ^2 values and the ν , and to conclude that overall there is indeed nothing surprising. But on the other hand, if the second measurement was genuinely different, and an alternative way of looking for some discrepancy, then it may be more appropriate to combine the p -values by one of the earlier methods, which suggest that the overall consistency with theory is not good. It is this extra information about the nature of the two tests that determines which combination method might be appropriate.

It is clearly important to decide in advance what combination method should be used, without reference to the specific data.

7 Blind analyses

These are becoming increasingly popular in Particle Physics, as a means of avoiding personal bias affecting the result. They involve keeping part of the data unseen by the physicists, until the data selection procedure and the analysis method have been completely defined, all correction procedures specified, etc.

The original suggestion to use a blind analysis for a Particle Physics experiment was due to Luis Alvarez. An experiment at Stanford had looked for quarks, by measuring the residual charge on small spheres that were levitated in a superconducting magnet. If a single free quark was present in a sphere, the residual charge would be a third or two-thirds of the electron's charge. Several of the balls tested indeed yielded such values[62]. A potential problem was that large corrections had to be applied to the raw data in order to extract the final result for the charge. The suspicion was that maybe the experimenters were (subconsciously) applying corrections until the value turned out to be 'satisfactory'. The blind approach would involve the computer adding a random number to the raw value of the charge, which would then be corrected until the experimentalists were satisfied, and only then would the computer subtract the random number to reveal the final answer for that sphere²⁷.

There are various methods of performing blind analyses[63] most of which aim to allow the experimentalists to look at some of the real data, in order to perform checks that nothing is terribly wrong. Some of these are:

- The computer adds a random number to the data, which is only subtracted after all corrections are applied. This was the method suggested by Alvarez.
- Use only Monte Carlo to define the procedure. This completely avoids the danger of allowing the data to determine the procedure to be used,

²⁷This suggestion was implemented, but in fact no subsequent results were published. The current consensus is that this 'discovery' of free quarks is probably spurious.

but suffers from the drawback that the data cannot be compared with the Monte Carlo, to check that the latter is reasonable.

- Use only a fraction of the data for defining the procedure. Then this is held fixed for the remainder of the data. In principle, an optimisation can be employed to determine the fraction to be kept open, but in practice this is often decided by choosing a semi-arbitrary time after which the future data is kept blind.
- The signal region is defined by a certain part of multi-dimensional space, and this is kept hidden, but all other regions, including those adjacent to the signal, are available for inspection.
- Keep the Monte Carlo parameters hidden. This is a technique used by the TWIST experiment in their high statistics precision determination of parameters associated with muon decay. The procedure involves comparing the data with various simulated sets, generated with a series of different parameter values. The data and the simulations are both visible, but the parameter values used to generate the simulations are kept hidden.
- Keep visible only a fraction of the the contents of each bin of a histogram. This is used by the MINOS experiment searching for neutrino oscillations; these would affect the energy distribution of the observed events. By keeping visible different unknown fractions of the data in each bin, the energy spectral shape cannot be determined from the visible part of the data.

If several different groups within the same collaboration are performing similar analyses for extracting some specific parameter, then it is desirable to fix the procedure for selecting which result to present, or alternatively how to combine the separate results. This should be done before the results are seen, and is worth doing even if the individual analyses were not “blind”.

A question that arises with blind analyses is whether it should be permitted to modify the analysis after the data had been unblinded. It is generally agreed that this should not be done unless everyone would regard it as ridiculous not to do so. For example, if a search for rare events yielded 10 candidates over the course of a year’s run, all of which occurred on Sunday mornings at precisely 1.17 a.m., it would be prudent to do some further investigation before publishing. If ‘post-unblinding’ modification of the procedure is performed, this should be made clear in any publication.

8 Combining data

We start with a question for the reader. Is it possible to combine two measurements of a single quantity, each with uncertainty ± 10 , such that the error on the combined best estimate is ± 1 ? The answer can be deduced later.

We often want to combine N different uncorrelated measurements $a_i \pm \sigma_i$ of the same physical quantity a . When the measurements are believed to be Gaussian distributed about the true value a_{true} , the well-known result is that the best estimate $a_{comb} \pm \sigma_{comb}$ is given by

$$a_{comb} = \Sigma(a_i * w_i) / \Sigma w_i, \quad \sigma_{comb} = 1 / \sqrt{\Sigma w_i}, \quad (8)$$

where the weights are defined as $w_i = 1/\sigma_i^2$. This is readily derived from minimising with respect to a a weighted sum of squared deviations.

$$S(a) = \Sigma(a_i - a)^2 / \sigma_i^2 \quad (9)$$

The extension to the case where the individual measurements are correlated (as is often the case for analyses using different techniques on the same data) is straightforward: S becomes $\Sigma \Sigma (a_i - a) * H_{ij} * (a_j - a)$, where H is the inverse error matrix. It provides **Best Linear Unbiased Estimates**[64].

There are, however, practical details that complicate its application. For example, in the above formula, the σ_i are supposed to be the **true** accuracies of the measurements. Often, all that we have available are **estimates** of their values. Problems arise in situations where the error estimate depends on the measured value a_i . For example, in counting experiments with Poisson statistics, it is typical to set the error as the square root of the observed number. Then a downward fluctuation in the observation results in an overestimated weight, and a_{comb} is biased downwards. If instead the error is estimated as the square root of the expected number a , the combined result is biased upwards – the increased error reduces S at larger a . A way round this difficulty has been suggested by Lyons et al[65].

Another problem arises when the individual measurements are very correlated. When the correlation coefficient of two uncertainties is larger than σ_1/σ_2 (where σ_1 is the smaller error), a_{comb} lies outside the range of the two measurements. As the correlation coefficient tends to $+1$, the extrapolation becomes larger, and is very sensitive to the exact value assumed for the covariance. The situation is aggravated by the fact that σ_{comb} tends to zero. This is usually dealt with by selecting one of the two analyses, rather than trying to combine them. However, if the estimated error increases with the estimated value, choosing the result with the smallest **estimated** error can again produce a downward bias. On the other hand, using the smallest **expected** error can cause us to ignore an analysis which had a particularly favourable statistical fluctuation, which produced a result that was genuinely more precise than expected²⁸. How to deal with this situation in general is an open question. It has features in common with the problem (inspired by ref. [34]) of measuring a voltage by choosing at random a voltmeter from a cupboard containing meters of different sensitivities.

Another extension of this procedure is for combining N pairs of correlated measurements (e.g. the gradient and intercept of a straight line fit to several

²⁸For example, the ALEPH experiment at LEP produced a tighter-than-expected upper limit on the mass of ν_τ because they happened to observe a τ -decay configuration producing ν_τ which was particularly sensitive for determining its mass.

sets of data). For several pairs of values (a_i, b_i) with inverse error matrices \mathbf{M}_i , the best combined values (a_{comb}, b_{comb}) have as their inverse error matrix $\mathbf{M} = \Sigma \mathbf{M}_i$. This means that, if the error matrix correlation coefficients ρ_i of the different measurements are very different from each other, the uncertainty on a_{comb} can be very much smaller than that for any single measurement. This situation applies for track fitting to hits in a series of groups of tracking chambers, where each set of close chambers provides a very poor determination of the track; but the combination involves widely spaced chambers and determines the track well. It is also relevant for the determination of the amount of dark energy in the Universe from various cosmological data.

8.1 Data consistency

The standard procedure for combining data pays no attention to whether or not the data are consistent. If they are clearly inconsistent, then they should not all be combined. When they are somewhat inconsistent, the procedure adopted by the Particle Data Group[66] is to increase all the errors by a common factor such that the overall χ^2 per degree of freedom equals unity²⁹.

The Particle Data Group prescription for expanding errors in the case of discrepant data sets has complications when the data sets each consists of two or more parameters, such as discussed at the end of Section 8[67].

A somewhat similar situation applies to fits of parton distribution functions to various sets of deep inelastic scattering data. The problem is that predictions of quantities of physical interest from the individual data sets used separately are not very consistent with each other, even though the overall χ^2 per degree of freedom is not unreasonable. (This could arise from the inconsistency of the separate data sets being compensated by the χ^2 -values for internal fits to individual data sets being smaller than expected[68].) Some groups performing these analyses try to deal with these inconsistencies by using an enlarged value of $\Delta\chi^2$ to determine the uncertainties on predictions[69]. A problem with this is that it might be hard to assess the significance of an observed discrepancy of LHC data with the SM prediction, if the latter is dominated by uncertainties from parton distribution functions; or alternatively a genuine signature of new physics might be missed[70].

9 Topics that deserve more attention

9.1 Statistical software

Particle physicists tend to write their own software for performing statistical computations. Although this has educational merits, it is inefficient use of our time. Jim Linnemann has recently set up at Fermilab a repository[71] for such software, while there are several useful statistical routines implemented in the

²⁹This is somewhat conservative, in that even if there are no problems, about half the data sets would be expected to have this larger than unity.

data-manipulation program ROOT[72]. A variety of goodness of fit techniques are in ref. [73], and tools exist for implementing many methods for separating signal from background[28, 29].

9.2 Deconvoluting data or smearing theory?

Our recorded experimental distributions are almost always smeared versions of ‘the true distributions of Nature’. In Particle Physics it is usual to compare theory and data by smearing the theory, rather than trying to deconvolute the experimental effects from the data, as the latter is a less stable procedure and also introduces correlations among the bins of the unfolded distribution. Other fields tend to favour deconvolution; this is partly because it is rarer for them to have a dominant theoretical model with which the data is to be compared. Deconvolution does have the advantage that it provides an estimate of the ‘true’ distribution, with which any future theory can be compared.

9.3 Visualisation

Particle Physicists very often study class separation (signal versus background) with multi-dimensional data, but tend to rely on some computational method to perform the separation, without attempting to use human eyes and brains to inspect some visual representation of the data.

9.4 Non-parametric methods

These are so unknown to most Particle Physicists that they are usually unaware when they are using them.

9.5 Collaboration with statisticians

Other scientists seem to be better than Particle Physicists about involving Statisticians in the analysis of their data. This is partly due to the fact that we like to try out statistical techniques ourselves; that we consider our data is too complicated for other people to deal with; and that we are somewhat over-protective of our data, and are reluctant to share it with others. None of this is over-convincing, and it is clear that we would benefit from the involvement of professional statisticians. The advantages of having them participating in the recent PHYSTAT meetings has been obvious.

In the past, Particle Physicists have on occasion asked rather specific questions to Statisticians they happened to know. Statisticians prefer to be much more directly involved with the data itself. With analyses becoming more and more complex, it will be highly desirable for them to be affiliated with experimental Collaborations.

10 Conclusion

It is clear that there are many practical statistical issues to be resolved in Particle Physics. It is hoped that more active collaboration with Statisticians will result in a better understanding of difficulties and improved analyses in the future.

I wish to acknowledge the patience and expertise of David Cox Brad Efron and Jerry Friedman, and also of other Statisticians too numerous to list, in explaining statistical issues to me; the ones who have contributed to the PHYSTAT meetings have been particularly helpful. My understanding of the practical application of statistical techniques has improved considerably as a result of discussions with many experimental Particle Physics colleagues, and in particular with the members of the CDF Statistics Committee. I particularly wish to thank Bob Cousins, Luc Demortier and Joel Heinrich for their careful reading and valuable comments on this article. To all of you, I am most grateful.

The Leverhulme Foundation kindly provide a grant which partially supported this work.

References

- [1] F. Solmitz, *Ann. Rev. Nucl. Sci.* **14** (1964) 375.
- [2] J. Orear, “Notes on statistics for Physicists”, UCRL 8417 (1958); revised version (1982) <http://nedwww.ipac.caltech.edu/level5/Sept01/Orear/Orear.html>
- [3] Roger Barlow, “Statistics: a Guide to the Use of Statistical Methods in the Physical Sciences”, Wiley (1989).
Glen Cowan, “Statistical Data Analysis”, Oxford University Press (1998).
W. T. Eadie, D. Drijard, F. E. James, M. Roos and B. Sadoulet, “Statistical Methods in Experimental Physics”, American Elsevier, (1971). Recently updated by Fred James, World Scientific Publishing Co (2007).
W. T. Frodesen, O. Skjeggstad and H. Tofte, “Probability and Statistics in Particle Physics”, Universitetsforlaget (1979).
Louis Lyons, “Statistics for Nuclear and Particle Physics”, Cambridge University Press (1986).
Byron Roe, “Probability and Statistics in Experimental Physics”, Springer Verlag (1991).
- [4] Workshop on Confidence Limits, CERN Yellow Report 2000-05.
- [5] FNAL Confidence Limits Workshop (2000) <http://conferences.fnal.gov/CLW/>
- [6] Advanced Statistical Techniques in Particle Physics, Durham IPPP/02/39.
- [7] Proceedings of PHYSTAT2003, eConf C030908, SLAC-R-703.

- [8] “PHYSTAT05: Statistical Problems in Particle Physics, Astrophysics and Cosmology”, Imperial College Press (2006) <http://www.physics.ox.ac.uk/phystat05/> .
- [9] BIRS Workshop on “Statistical inference Problems in High Energy Physics and Astronomy”, Banff 2006 http://www.birs.ca/birspages.php?task=displayevent&event_id=06w5054 .
- [10] PHYSTAT-LHC Workshop on “Statistical Issues for LHC Physics” (2007), <http://phystat-lhc.web.cern.ch/phystat-lhc/2008-001.pdf>
- [11] BaBar Statistics Working Group <http://www.slac.stanford.edu/BFROOT/www/Statistics/>
- [12] CDF Statistics Committee http://www-cdf.fnal.gov/physics/statistics/statistics_home.html
- [13] ATLAS Statistics Forum https://twiki.cern.ch/twiki/bin/view/Atlas/StatisticsTools#Statistics_Forum
- [14] CMS Statistics Committee <https://twiki.cern.ch/twiki/bin/view/CMS/StatisticsCommittee>
- [15] B. P. Roe, Nuclear Instruments and Methods **A570** (2007) 159.
- [16] N. Read, “Some aspects of design of experiments”, ref. [10], page 94.
- [17] R. Neal, “Computing likelihood functions when distributions are defined by simulations with nuisance parameters”, ref. [10], page 101; and in ref. [9].
- [18] J. Linnemann, “A pitfall in estimating systematic errors”, ref. [10], page 94.
- [19] J. Heinrich and L. Lyons, Annual Reviews of Nuclear and Particle Science **57** (2007) 145.
- [20] L. Lyons, “Bayes or Frequentism?”, http://www-cdf.fnal.gov/physics/statistics/statistics_recommendations.html
- [21] R. D. Cousins, Am. J. Phys. **63** (1995) 398.
- [22] R. Barlow, “Asymmetric errors”, ref. [7], page 250; and ref. [8], page 56.
- [23] J. Heinrich, “Pitfalls of Goodness-of-Fit from Likelihood”, ref. [7], page 52.
- [24] G. Punzi, “Comments on likelihood fits with variable resolution”, ref. [7], page 235.
- [25] P. Catastini and G. Punzi, “Bias-free estimation of multicomponent maximum likelihood fits with component-dependent templates”, ref. [8], page 60.

- [26] H. B. Prosper, “Multivariate methods: a unified perspective”, ref. [6], page 91.
- [27] J. H. Friedman, “Recent advances in predictive (machine) learning”, ref. [7], page 196; and “Separating signal from background using ensembles of rules”, ref. [8], page 127.
- [28] I. Narsky, “StatPatternRecognition in analysis of HEP and Astrophysics data”, ref. [10], page 188.
- [29] A. Hocker et al, “TMVA, Toolkit for Multi-Variate data Analysis with ROOT”, ref. [10], page 184.
- [30] L. Lyons, Nucl. Inst. Meth. **A324** (1993) 565.
- [31] S. Whiteson and D. Whiteson, “Stochastic Optimization for Collision Selection in High Energy Physics. IAAI 2007: Proceedings of the Nineteenth Annual Innovative Applications of Artificial Intelligence Conference (July 2007) 1819.
- [32] I. Narsky, “Comparison of upper limits”, in ref. [5].
- [33] G. J. Feldman and R. D. Cousins, Phys. Rev. **D57** (1998) 3873.
- [34] D. R. Cox (1958), “Some problems connected with statistical inference”, Annals of Mathematical Statistics **29** (1958) 357.
- [35] B. Sen, M. Walker and M. Woodroffe, “On the Unified Method with Nuisance Parameters”, to appear in Statistica Sin.
- [36] G. Punzi, “Sensitivity of searches for new signals and its optimisation”, ref. [7], page 235.
- [37] A. L. Read, “Modified frequentist analysis if search results”, ref. [4], page 81; “Presentation of search results - the CL_s method”, ref. [6], page 11.
- [38] W. A. Rolke, A. M. Lopez and J. Conrad, Nuclear Instruments and Methods **A551** (2005) 493.
- [39] J. Heinrich, “Coverage of error bars for Poisson data ”, http://www-cdf.fnal.gov/publications/cdf6438_coverage.pdf
- [40] J. Heinrich et al. “Interval estimation in the presence of nuisance parameters. 1. Bayesian approach”, CDF note 7117 (2004).
- [41] L. Demortier, “A fully Bayesian computation of upper limits for Poisson processes”, CDF note 5928 (2004).
- [42] J. Heinrich, “The Bayesian approach to setting limits: what to avoid”, ref. [8], p 98.

- [43] D. Nicolo and G. Signorelli, “Application of strong confidence to the CHOOZ experiment with frequentist inclusion of nuisance parameters”, ref. [6], page 152.
- [44] G. Punzi, “Ordering algorithms and confidence intervals in the presence of nuisance parameters”, ref. [8], page 88.
- [45] R. Cousins, Nuclear Instruments and Methods **A417** (1998) 391.
- [46] R. D. Cousins and V. L. Highland, Nuclear Instruments and Methods **A320** (1992) 331.
- [47] J. Heinrich, “Review of Banff challenge on upper limits”, ref. [10], page 125.
- [48] <http://www-cdf.fnal.gov/physics/statistics/recommendations/limits.txt>
- [49] B. Aslan and G. Zech, J. Stat. Comp. Simul. **75** (2004) 109; and Nuclear Instruments and Methods **A537** (2005) 626.
- [50] C. M. Cuadras, J. Fortiana and F. Oliva, J. of Classification **14** (1997) 117; C. M. Cuadras and J. Fortiana, “Distance-based multivariate two sample tests,” University of Barcelona Institute of Mathematics preprint No. 334 (June 2003), <http://www.imub.ub.es/publications/preprints/pdf/Cuadras-Fortiana.334.pdf>
- [51] L. Lyons, “Comparing two hypotheses”, http://www-cdf.fnal.gov/physics/statistics/statistics_recommendations.html
- [52] R. Trotta, “Bayes in the sky: Bayesian inference and model selection in cosmology”, to appear in Contemporary Physics.
- [53] S. S. Wilks, “The large-sample distribution of the likelihood ratio for testing composite hypotheses”, Annals of Math. Stat. **9** (1938) 60.
- [54] R. Protassov et al., “Statistics: Handle with care. Detecting multiple model components with the likelihood ratio test”, Astrophysics Journal **571** (2002) 545.
- [55] L. Demortier, “Setting the scene for p-values”, http://birs.pims.math.ca/~06w5054/Luc_Demortier.pdf.
- [56] S. G. Self and K. Y. Liang, JASA **82** (1987) 605.
- [57] M. Drton, “Likelihood ratio tests and singularities”, <http://front.math.ucdavis.edu/0703.5360>.
- [58] L. Demortier, “p-values and nuisance parameters”, ref [10], page 23.
- [59] K. Cranmer, “Statistics for LHC: progress, challenges and future”, ref. [10], page 47.

- [60] CDF Statistics Committee, “Frequently asked questions”, http://www-cdf.fnal.gov/physics/statistics/statistics_faq.html#iptn4
- [61] R. Cousins, “Annotated bibliography on some papers on combining significances or p -values”, arXiv:0705.2209 (2007)
- [62] G. S. LaRue, J. D. Phillips and W. M. Fairbank, Phys. Rev. Lett. **46** (1981) 967.
- [63] J. R. Klein and A. Roodman, Annual Review of Nuclear and Particle Physics **55** (2005) 141.
- [64] L. Lyons, D. Gibaut and P. Clifford, Nuclear Instr. Meth. **270** (1988) 210.
- [65] L. Lyons, A. Martin and D. Saxon, Phys Rev **D41** (1990) 982.
- [66] Particle Data Group, Journal of Physics G: Nuclear and Particle Physics **33** (2006) 1 (see page 14).
- [67] T. Trippe and Particle Data Group, private communication.
- [68] R. S. Thorne, private communication.
- [69] R. S. Thorne, “Role of uncertainties in parton distribution functions”, ref. [10], page 141.
- [70] R. D. Cousins, private communication.
- [71] PHYSTAT statistical software repository: www.phystat.org
- [72] L. Moneta, “ROOT statistical software”, ref. [10], page 179.
- [73] G. A. P. Cirrone et al., “A Goodness-of-Fit Statistical Toolkit”, IEEE Trans. Nucl. Sci. **51**, no. 5, (2004) 2056;
B. Mascialino, A. Pfeiffer, M. G. Pia, A. Ribon and P. Viarengo, “New developments of the Goodness-of-Fit Statistical Toolkit”, IEEE Trans. Nucl. Sci. **53**, no. 6 (2006) 3834.