April 22, 2009

Methods for comparing two hypotheses

Louis Lyons
Blackett Lab, Imperial College, Kensington, London SW7 2BW
and
Particle Physics, Keble Rd, Oxford OX1 3RH, UK
e-mail: l.lyons@physics.ox.ac.uk

Abstract

In attempts to discover exciting new phenomena, it is useful to compare our data with (at least) two different hypotheses. These are H_0, the null hypothesis, corresponding to just known physics and nothing new; and H_1, a specific version of new physics, perhaps involving free parameters, such as the mass of a new particle. Methods of performing this comparison, with a view to claiming a discovery, are reviewed.

1 Introduction

It is exciting to make a new discovery. In Particle Physics, current knowledge is encapsulated in the Standard Model (SM), but there are plenty of candidates for new physics e.g. supersymmetry, extra dimensions, quark and/or lepton substructure, etc. There is also the possibility of discovering something completely unexpected.

From a statistical viewpoint, there are two types of approach that can be made towards claiming a discovery. The first is to test whether our data is consistent with the null hypothesis of there being nothing new (in our case, H_0 is the SM[1]). This is typically called "Goodness of Fit". The second consists of comparing the data with both H_0 and with a specific alternative H_1; this is "Hypothesis Testing". In this article we concentrate on the latter, although Section 4 contains a brief comparison of the two approaches. Apart from statistical issues, there are also (and usually more importantly) physics considerations, as discussed briefly in Section 2.

In the context of trying to make a discovery, Hypothesis Testing is essentially applied to the end-result of the experiment. It can also be used earlier in the analysis chain, where we are trying to select wanted signal events (which is now H_0) while rejecting unwanted background (H_1). The desirable features for a Hypothesis Testing procedure may be different for these two situations. This is discussed in Sections 2 and 3. In this note, we concentrate largely on the former.

Any Goodness of Fit or Hypothesis Testing situation provides a number on which a decision must be based. We then still have to make a decision about

[1] Although the Higgs boson is part of the SM, as at the time of writing it has not yet been discovered. For the purpose of this discussion, we classify Higgs searches as being part of new physics. i.e. we do not include H^0 in H_0.

what action to take. For Goodness of Fit, this is whether to reject or accept H_0; in Hypothesis Testing, it is whether to prefer one of the hypotheses to the other, or perhaps to make no choice. Section 5 discusses p-values (the probability of obtaining a result as extreme as ours, or more so), both for Goodness of Fit and for Hypothesis Testing situations.

In either situation, our decision may be incorrect. This may consist in rejecting H_0 when it is true; these are "Errors of the first kind". "Errors of the second kind" are when H_0 is false, but we accept it. For hypothesis tests where we allow 'no decision' as an outcome, a measure of the efficiency of the test is also desirable (see Section 6.2).

Section 6 considers p-values and likelihood ratios, while the much-maligned CL_s method is discussed in Section 7. Various Bayesian-based methods are described in Section 8, and some comparisons of the various approaches are to be found in Section 9. Finally different ways of incorporating systematics are mentioned (see Section 10); this topic is likely to be very important at new accelerators running at higher energies, where backgrounds and detectors may be poorly understood.

It is always crucial to remember that the statistical technique we are using will be completely insensitive to any undetected faults in our analysis. For example, the theory or software for estimating some background may provide a poor description of reality; our understanding of the detector may be inadequate; etc. Another important point is that the statistical method provides some numerical output only to the particular question being addressed. Thus a poor Goodness of Fit statistic could be due to many large random fluctuations or to some underestimated systematic, rather than to the presence of new physics. In a similar vein, consistency between the data and the alternative H_1 is not necessarily implied by a large likelihood ratio between H_1 and H_0; it is merely that H_1 is better than H_0 for describing the data.

Selecting between alternative hypotheses is a subject with a vast statistical literature; refs. [1] - [3] are just a very small sample of what is available. A variety of papers related to discovery issues can be found in ref.[4]. Demortier[5] has recently produced recommendations for these issues.

2 Physics Considerations

When we are considering claiming a new discovery, we want to ensure that the apparent evidence is very unlikely to be due to a statistical fluctuation, or to a mistaken analysis. This note deals predominantly with the statistical issues, but here we make a few remarks relevant to the physics.

If we are about to perform a Goodness of Fit test looking for any new phenomenon, we still have to decide which of the infinity of possible distributions of variables to examine. For example, mass plots with possible interesting peaks in them are likely to be more interesting than the instantaneous accelerator beam luminosity for the observed events. This already involves an implicit specification of the sort of new physics we might hope to find. In the other

situation of testing a specific alternative hypothesis (Hypothesis Testing), it should be clearer which are the relevant variables which might be most sensitive to our alternative hypothesis.

In either case, we still have to ensure that any apparently statistically unlikely occurrence for the null hypothesis is not produced by some bug. This could be due to one or more events accidentally appearing several times in the analysis; an incorrect statistical procedure, perhaps failing to recognise that the method was not blind, and hence the physicist could (subconsciously) work to enhance the effect by some (in)judicious selections; a poor description of the expected backgrounds in the selected event sample; not allowing for uncertainties in parameters associated with the simulation of H_0; incorrect description of the detector effects; etc. Thus an underestimate of systematics associated with the knowledge of parton distributions used to calculate expectations from H_0 for jet production at large p_T could lead to apparently statistically significant deviations in the data, which could be misinterpreted as a sign of the discovery of quark substructure.

If, in so far as is feasible, the possibility of a mistaken analysis is excluded, then it is important to consider whether further investigations can be performed on the possible discovery. For example, it may be possible to check background estimates in channels which are like the one used for the hoped-for discovery, but which should not contain the signal. Another check could be that if the suspected signal involved W bosons decaying to a muon or electron plus a neutrino, it might be possible to find confirmatory evidence from W hadronic decays.

As pointed out later, in deciding to make some claim for the result of the experiment (e.g. "We have excluded the SM Higgs boson for masses up to 750 GeV", or "Our results show conclusive existence of extra dimensions"), it is necessary to adopt some rule for using the value of the statistic (i.e. the p-value, likelihood ratio, etc.) for making a decision. In deciding how strong the evidence should be, a convention in Particle Physics has been that a p-value should correspond to a 5σ upper tail of a Gaussian, equivalent to a probability of $3 * 10^{-7}$ for a statistical fluctuation. Bob Cousins has argued strongly that, rather than adopting a single standard, the degree of surprise of the result should be taken into account. Thus if we were testing for the violation of the conservation of energy, we should require more than 5σ. In contrast, if we are looking for evidence for the production of single top quarks (rather than their observed pair production), we should be satisfied with less than 5σ since the top quark exists, and its mass and its expected production rate are known; indeed it would be surprising if the process did not exist. This allowance for the degree of unexpectedness of the result is essentially (or explicitly) taking into account our Bayesian priors for the various hypotheses.

Statisticians are always surprised to hear that anyone requires a level of 5σ before claiming a discovery. This is not only because they regard this as extremely stringent, but also because they find it hard to believe that *pdf*s (probability density functions) are known accurately enough, especially in the tails of their distributions, to make such numbers meaningful. This may be a very relevant comment, especially concerning nuisance parameters, whose estimation

is often more vague than for statistical errors.

Although the discovery threshold (i.e rejection of H_0) tends to be set very stringently, exclusion of H_1 is usually taken at the 95% level. This is because claiming an exclusion region slightly larger than it ought to be is regarded as not as serious as having made a false discovery claim. Also, as expressed by Cowan[6], "If you have lost your car keys and have looked hard in the kitchen so you are 95% certain they are not there, it is sensible to continue the search elsewhere."

An alternative use of hypothesis testing is at an earlier stage of most analyses, where an attempt is made to select wanted events for further analysis, while having a strong rejection of the usually more copious background. Here the relevant hypotheses are 'signal' as H_0 and 'background' as H_1. In this case, there is no option of not making a decision; the events are either accepted or rejected. A multivariate method (e.g. boosted decision trees, neural networks, support vector machines) is usually adopted for this. In order to achieve a good efficiency for signal (small error rate of the first kind), it is almost inevitable that there will be some level of backgound among the accepted events (errors of the second kind). Unlike the situation where our hypothesis testing was for the result of the experiment, this is no longer a disaster, but corresponds to a situation we can live with, provided we can make a reliable estimate of the level of background and correctly allow for it in the analysis. This will inevitably increase the statistical error on the result of the experiment as compared with an ideal situation with no background. The dangers are that this will be dominated by systematic errors associated with uncertainties in the background; or, even worse, that these systematics will be underestimated.

Multivariate methods usually allow the signal acceptance efficiency to be varied; thus with neural networks, the acceptance cut on the neural network output can be altered. As the efficiency is increased, the background generally rises, so some compromise is needed in deciding where to make the cut. This is usually done by optimising the expected accuracy of the result (although some approaches build this optimisation into the multivariate method itself). Estimating the background is generally much more difficult than obtaining the signal efficiency, because the background could come from many sources, and it is often the hard-to-simulate tails of distributions that creep into the acceptance region.

That is almost all for Physics; the rest of this note deals predominantly with statistical issues.

3 Examples of Hypotheses

A possible example of a hypothesis could be that the value of a parameter takes on a special value; for example, that the magnetic moment of the muon (μ_μ) should be as predicted by SM. Then it is possible to test this hypothesis by establishing a confidence or Bayesian credible range for μ_μ at some specified level, and seeing if this includes the predicted value. In general, testing whether

the distribution of the data agrees with the model is a better way of performing hypothesis testing, as it is possible to obtain seemingly acceptable parameter values even when the data do not fit the model well.

In some cases, one or both of our hypotheses may be completely specified, but more commonly the hypotheses may involve free parameters. The former are termed 'simple hypotheses', while the latter are 'composite'. Such free parameters could be experimental, such as the correction factor for the energy scale of our calorimeters; or they could be theoretical, like the mass of a particle we are trying to discover. Even though the latter is a quantity of physical interest, it is regarded as a 'nuisance parameter' for the specific hypothesis test we are trying to perform. It is of course necessary to take into account the uncertain values of nuisance parameters in performing statistical tests. In particular, performing Monte Carlo simulations for producing the expected distributions for our test statistic is more complicated when there are nuisance parameters, and frequentist and Bayesians have different approaches on how to do this. The former would generate several different sets of simulated data, each with fixed values of the nuisance parameters. In contrast, Bayesians might generate a single simulation, in which the nuisance parameters are varied randomly from event to event according to the assumed (possibly correlated) priors for the parameters.

Another categorisation of our hypotheses H_0 and H_1 is whether they are 'nested'. If so, the hypothesis with more free parameters (usually H_1) includes the other hypothesis as a special example for specific values of the parameters. Thus if we are fitting some data on the length L of an object as a function of temperature T, we might try a straight line as H_0 ($L = a + bT$) or a quadratic as H_1 ($L = a + bT + cT^2$); then H_1 reduces to H_0 if $c = 0$.

In contrast we could try to fit the data by the same H_0 but with H_1 as an exponential $L = L_0 e^{dT}$, where L_0 and d are parameters. Then there are no values of L_0 and d (or for a and b) for which the two hypotheses become identical. Thus H_0 and the new H_1 are not nested.

As implied above, when dealing with nested hypotheses, it is usual to take as H_0 the one with fewer parameters. e.g. a smooth background as opposed to 'background plus peak'. In particle physics situations, where the SM is being compared with new physics, the former is the usual H_0; this is likely to coincide with the convention of fewer parameters.

The reason for not choosing new physics as H_0 (especially in a p-value orientated method) is that with data that is not decisive in distinguishing the hypotheses, we would end up stating that we had not rejected the new physics hypothesis. While this is a true statement, it is in danger of being confused with the idea that our data supports the idea of new physics. With the SM as the null hypothesis, we would merely say that we had not ruled out the SM. Bayesian methods tend to deal more symmetrically with the two hypotheses.

Demortier has pointed out another reason for selecting SM as H_0. In the standard frequentist approach, we choose the value of α in advance, and this controls errors of the first kind. i.e. false rejections of H_0. The rate for errors of the second kind (failure to reject H_0 when the alternative is true) then depends on how well the hypotheses' *pdf*s are separated. Because in this context errors of

the first kind are usually regarded as more important, this too argues in favour of having SM as H_0.

4 Comparison of "Goodness of Fit" with "Hypothesis Testing"

Typical Goodness of Fit tests are χ^2, Kolmogorov-Smirnov and its variants, etc., which result in p-values. Methods for performing Hypothesis Testing include likelihood ratio, difference in χ^2, p-values, Bayes factor, Bayes information criterion, etc. The advantage of a goodness of fit test is that it looks for discrepancies between the data and H_0, without the need to specify any alternative. Thus in principle, **any** deviation from our null hypothesis can be detected, and we can be sensitive to discovering surprising forms of new physics. However, if we know what we are looking for (e.g. a leptoquark of specific mass, etc.), a comparison of the two hypotheses is likely to be more sensitive to the new phenomenon. On the other hand, this hypothesis comparison may fail to alert us to the fact that the data contains some other type of new physics; having H_1 as a leptoquark may not easily allow us to discover a 4^{th} generation top quark.

Even Goodness of Fit tests, however, are not completely free of the concept of an alternative hypothesis. For example, the choice of test statistic depends somewhat on what we expect the other possibilities to be; different statistics are likely to have different sensitivities for ruling out H_0. Similarly, in situations where we have a single statistic, we might regard a large value, or a small one or either as a sign of a discrepancy. Thus if we thought that the production of a new particle was a likely possibility, an **excess** of events could be significant. In searching for neutrino oscillations, a **deficit** might be interesting; while for a general search for new physics **any discrepancy** could be noteworthy.

Another example is of using χ^2 to test whether a student has invented some experimental results, rather than actually taking the measurements. Usually H_0 is rejected when the value of χ^2 is large, but in this case it is a very small value that would be suspicious[2].

5 p-values

For our null hypotheses, H_0, we construct some statistic t; one example could be simply the observed number of events in some pre-defined region. The expected distribution of t under the null hypothesis is $f_0(t)$. Then for a given observed value t_{obs}, p is the fractional area in the tail of $f_0(t)$ for t greater than or equal to t_{obs}. For definiteness we consider the single-sided upper tail (corresponding to other hypotheses tending to yield larger values of t_{obs}), but as mentioned above lower or 2-sided tails could be appropriate in other cases.

[2]Alternatively, if we thought the possibly-cheating student was very bright, we might see whether the value of χ^2 was too close to the number of degrees of freedom.

For the common case of a Poisson distributed number of events, the Poisson approximates to a Gaussian distribution for large values of the Poisson parameter μ. It is worth noting, however, that this approximation is poorer in the tails than for the bulk of the distribution. Thus the Poisson probability $P(t_{obs} = 50 \mid \mu = 100) = 0.12 * 10^{-7}$ as compared with $P(t_{obs} = 150 \mid \mu = 100) = 6.5 * 10^{-7}$, whereas the Gaussian approximation would suggest they were equal.

When the data are summarised by more than one variable, the definition of what is 'more extreme than the observed values' may well not be unique. It is convenient if the data can be compactified into a single variable (e.g. the output from a neural network; an 'optimum variable'; etc.) Sometimes a likelihood ratio is used; this is the likelihood of the observed data, assuming some specific alternative hypothesis, as compared with its likelihood assuming H_0. This statistic no longer depends solely on H_0, but p_0 will still be the probability of observing an extreme value of the statistic, assuming H_0.

Section 10 lists ways of calculating p-values in the presence of systematics.

5.1 What p is not

It is extremely important to realise that a p-value is the probability of observing data like that observed or more extreme, assuming the hypothesis is correct. It is **not** the probability of the hypothesis being true, given the data. These are not the same. For example, the probability of being pregnant, assuming you are female, is very much smaller than the probability of being female, given that you are pregnant.

A small p-value is an indication that the data are not very consistent with the hypothesis. Apart from the possibility that the cause of the discrepancy is new physics, it could be due to an unlikely statistical fluctuation, an incorrect implementation of the hypothesis being tested, an inaccurate allowance for detector effects, etc.

As more and more data is acquired, it becomes more likely that a small (and not physically significant) deviation from the tested null hypothesis could result in the p_0 becoming small as the data become sensitive to the small deviation. For example, a set of particle decays may be expected to have an exponential decay, but there might be a small background characterised by decays at very short times, and which is not allowed for in the analysis. A small amount of data might be insensitive to this background, whereas a large amount of data might give a very small p-value for a test of exponential decay, even though the background is fairly insignificant. The possibility of a statistically significant but physically unimportant deviation has been mentioned by Cox[7].

Many of the negative comments about p-values are based on the ease of misinterpreting p-values. Thus it is possible to find statements that of all experiments quoting p-values below 5%, and which thus reject H_0, many more than 5% are wrong (i.e H_0 is actually true). In fact, the expected fraction of these experiments for which H_0 is true depends on other factors, and could

take on any value between zero and unity, without affecting the validity of the p-value calculation.

5.2 p-values for two hypotheses

With two hypotheses H_0 and H_1, we can define a p-value for each of them. We adopt the convention that H_1 results in larger values for the statistic t than does H_0. Then p_0 is defined as the upper tail of $f_0(t)$, the *pdf* for observing a measured value t when H_0 is true. It is usual to define p_1 by the area in the lower tail of $f_1(t)$ (i.e towards the H_0 distribution – see Fig. 1).

As explained earlier, there are now several situations possible (see Table 1):

- p_1 is small, but p_0 acceptable. Then we accept H_0 and reject H_1. As applied to the result of the experiment, this means that we exclude the alternative hypothesis.
- p_0 is very small, and p_1 acceptable. Then we accept H_1 and reject H_0. This corresponds to claiming a discovery.
- Both p_0 and p_1 are acceptable. The data are compatible with both hypotheses, and for the case where the hypotheses relate to the result of the experiment, we make no decision. Alternatively in using the test as an event selector, the event could be assigned to one of the hypotheses on the basis, for example, of the larger p-value; it could be kept as an example of both hypotheses; or rejected as ambiguous.
- Both p_0 and p_1 are small. The choice of decision is not obvious, but basically both hypotheses should be rejected.

Table 1: Decisions based on p_0 and p_1.

p_0	p_1	Decision for conclusion	Decision for event selector	If H_0 true
Very small	O.K.	Discovery	Accept H_1	Error of 1^{st} kind
O.K.	Small	Exclude H_1	Accept H_0	Correct decision
O.K.	O.K.	No decision	Choose by larger p?	Loss of efficiency
Very small	Small	?	Reject event	?

A difference compared with a Goodness of Fit test based just on p_0 is that we now have the possibility of 'no choice'. As a consequence, the categories of 'errors of the first or second kind' are no longer sufficient to characterise the performance of the test. A method could have very low error rates, but at the price of not making a decision for the majority of cases. Table 2 shows the probabilities of the test statistic t falling in the different regions of fig. 1, for either H_0 or H_1 being true. We see that the probability of falsely excluding H_1

when it is true is less than $\beta(\alpha)$. This is a result of the protection built in by the 'no decision' region. The price to pay for this protection is that the probability of excluding H_1 when H_0 is true is reduced. For given *pdf*s of H_0 and H_1, these probabilities can be determined, for example by simulation.

Table 2: Probabilities for the data statistic t falling in the different regions of fig. 1, when the *pdf*s of H_0 and H_1 overlap; this corresponds to the top 3 lines of Table 1. (The probabilities are slightly different when the *pdf*s are very well separated, and the data can be such that both p_0 and p_1 are small - see the bottom line of Table 1.) α is the very small probability cut-off imposed for rejecting H_0, and $1-\beta(\alpha)$ is the power for accepting H_1 when it is true. The procedure is characterised by α, β, and the efficiencies ϵ_0 and ϵ_1 for making a decision.

Data region	Exclusion of H_1	No decision	Discovery of H_1
H_0 true	Less than $1-\alpha$	$1-\epsilon_0$	α
H_1 true	Less than $\beta(\alpha)$	$1-\epsilon_1$	$1-\beta(\alpha)$

The procedure as just described does not implement the protection provided by the CL_s method (see Section 7) against the possibility of rejecting H_1 when the *pdf*s show the analysis has little or no sensitivity to distinguishing between H_0 and H_1. This can be incorporated in the procedure if desired.

There are different ways in which the above procedures are implemented. Some of these are:

- A simple approach would impose cuts on p_0 and p_1. These might typically be $3*10^{-7}$ and 0.05 respectively (See fig. 2 (b))
- The CL_s method is used only for possible exclusion of H_1. It thus refers to the first and third items in the previous list. The decision line is that H_1 is rejected if $p_1/(1-p_0)$ is less than some pre-assigned value, like 0.05.
- In comparing separate fits to a histogram of the same data, the χ^2 values for the two fits can be evaluated. In selecting between the hypotheses, the **difference** in χ^2 values is more sensitive than the individual values[18]. The no-decision region in now reduced as compared with the simple method of using the tail probabilities of the individual χ^2 values (see Section 9.1).
- Demortier has described several **conditional** frequentist approaches[8]. These calculate p-values, subject to the conditioning statistic, and use it in making a decision concerning the two hypotheses. These methods sometimes also have a Bayesian interpretation. As far as I am aware, they have not yet been employed in Particle Physics analyses.

Fig 2 illustrates the (p_0, p_1) plot for defining various decision regions.

6 p-values, likelihood ratios, and all that

Some elementary ideas concerning p-values were discussed in Section 5. Here we discuss them further, especially in situations where we are comparing two (or more) hypotheses.

6.1 α, β and power

Fig. 1 shows the probability densities for obtaining a value t of a data statistic, for hypotheses H_0 and H_1. For a specific value t_{obs}, the p-values p_0 and p_1 correspond to the upper tail area above t_{obs} for H_0, and below t_{obs} for H_1, respectively.[3] Then t_{crit} is the critical value of t such that its p_0 value is equal to a preset level α for rejecting the null hypothesis.

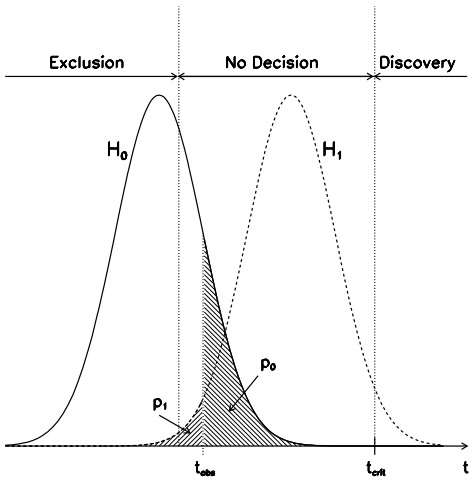

Figure 1: Schematic diagram of probability density functions for the observable t for two hypotheses H_0 and H_1. For a particular value t_{obs}, the p-values p_0 and p_1 are conventionally defined as the single-tail areas in the direction of the other hypothesis. (This definition is not always unique.) For rejecting H_0 (i.e. 'Discovery'), a typical choice is that p_0 should be smaller than $3 * 10^{-7}$, while exclusion of H_1 might require p_1 to be less than 0.05. In the diagrams in this note, they are shown at larger values simply for clarity. The pre-set level α for rejecting H_0 determines the critical value t_{crit} for the observable t. Then β is the p_1-value when $t = t_{crit}$, and the power of the test is $1 - \beta$ (see Section 6.1).

[3] If t is a discrete variable, such as a number of events, then 'above' is replaced by 'greater than or equal to', and correspondingly for 'below'.

Having chosen a value for α, then β is defined as the value of p_1 when $t = t_{crit}$. Clearly it depends on the *pdf* for H_1. The **power** of the test is defined as $1 - \beta$, and is the probability of rejecting H_0, assuming that H_1 is true. Its value depends on the separation of the *pdf*s for H_0 and H_1.

The above discussion implicitly assumes that H_0 and H_1 are simple hypotheses i.e. do not involve arbitrary parameters. While typically H_0 is the SM and may not involve arbitrary parameters, H_1 may do (e.g. the mass of the SM Higgs boson; some supersymmetry parameters; neutrino mixing angles; etc.) Then we can regard the *pdf* for H_1 as being for some specific values of the parameters, and we could produce a plot of the power of the statistical test for H_0 as a function of these parameters. In situations where H_1 reduces to H_0 for particular values of the parameters (e.g. the mass of the SM Higgs is very large), the ideal situation is where the power equals α when H_1 reduces to H_0, and rapidly rises towards unity for other values.

6.2 Likelihood ratio

Rather than calculating p-values for the various hypotheses, we could use the *pdf*s of Fig. 1 to evaluate their likelihoods L_0 and L_1. While p-values use tail areas beyond the observed statistic, the likelihood is simply the height of the *pdf* at t_{obs}.

In comparing two simple hypotheses, the Neyman–Pearson theorem deals with the best way of choosing a region in data space for accepting H_0 at a given probability level, while minimising the contamination from H_1. It says that the region should be such as to contain the largest values of L_0/L_1. The theorem applies only to comparisons of two hypotheses, both of which are simple. However, also in other situations the likelihood ratio may well be a suitable statistic for summarising the data and for helping choose among hypotheses. In general, it will be necessary to generate the expected distributions of the likelihood ratio according to the hypotheses H_0 and H_1, in order to make some deduction based on the observed likelihood ratio; for composite hypotheses there are of course the complications caused by the nuisance parameters. The decision process may well be based on the p-values p_0 and p_1 for the two hypotheses (see fig. 2). In that case, the procedure can be regarded as either a likelihood ratio approach, with the p-values simply providing a way of making a decision from the likelihood ratio; or as a p-value method, with the likelihood ratio merely being a convenient statistic.

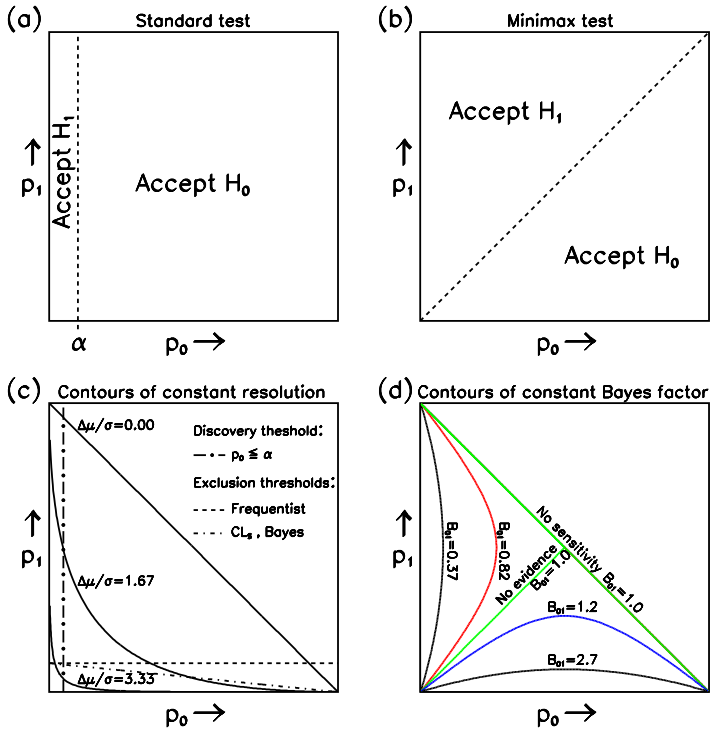

Figure 2: Plots of p_0 versus p_1. for some data x compared with two hypotheses H_0 and H_1. (a) The standard method of simply rejecting the null hypothesis if p_0 is below a specified level leads to an acceptance region to the right of the vertical dashed line. (b) The minimax approach accepts the hypothesis with the larger p-value. (c) With each hypothesis rejected if its p-value is below a specified value, the plot is divided into four regions. The largest rectangle is where both p-values are large and it may be decided to make no decision. The rectangle near the origin is where both p-values are small, and maybe both hypotheses are rejected. In the remaining two rectangles, one hypothesis is preferred over the other. For pdfs of a given separation, the (p_0, p_1) values lie on a curve; the ones shown are for Gaussian pdfs. As the separation of the pdfs increases, the curve moves towards the axes. The Punzi definition of sensitivity of a search corresponds to having enough data (equivalent to a large separation of the pdfs) such that the curve does not enter the 'no decision' region. The CL_s procedure for excluding H_1 uses the dashed diagonal line, rather than the standard horizontal one. (d) Contours of constant Bayes factor B_{01} for Gaussian pdfs. The upper right region is inaccessible; the diagonal line from (0,1) to (1,0) corresponds to the pdfs lying on top of each other i.e. no sensitivity. The diagonal through the origin is when x_{obs} is mid-way between the two pdfs. With larger separation of the Gaussian pdfs and for constant p_0 the Bayes factor B_{01} increases.

For a Gaussian *pdf*, as the value of t_{obs} rises more and more above the mean value, p_0 and the likelihood both decrease monotonically. Such a simple relationship does not always apply. For example, Fig. 3(a) shows a flat *pdf*; the value of the likelihood is independent of the value of t within its physically allowed range, and while the t_{obs} shown has a small p-value, its likelihood is not unusual. In contrast, the *pdf* of Fig. 3(b) has a dip in the middle, but t_{obs} corresponds to a large p-value; this suggests that it is consistent with the hypothesis while the *pdf* and the likelihood show that the t_{obs} cannot be produced if the hypothesis is true. A third example is a comparison of two hypotheses whose *pdf*'s are both Gaussians centred at zero, but with width 1 for H_0 and 2 for H_1. For any positive t, the one-sided p-value for H_0 is always smaller than that for H_1, but for t less than 1.67 the likelihood ratio L_0/L_1 is greater than unity. A comparison of the likelihood ratio and the difference in χ^2s is made 3 paragraphs below.

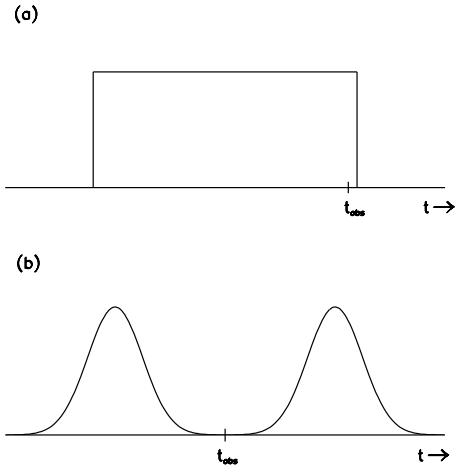

Figure 3: Examples of *pdf*s for which p-values and likelihoods may give rise to different conclusions. The flat distribution in (a) is such that the likelihood is independent of the observable t, and so even though t_{obs} has a small p-value, its likelihood is the same as for any other allowed t. In (b), if t_{obs} falls in the dip in the middle of the *pdf*, the likelihood will be zero, while the p-value will be close to 50%.

It is important to realise that the magnitude of the likelihood is not a measure of how well the data are consistent with the hypothesis[9][4]. However, for

[4]A simple illustration of this is provided by the likelihoods for the Poisson parameter μ when the observed number of events is μ. The likelihood for $\mu = 1$ is 0.37, as compared with 0.08 for $\mu = 25$, even though both are perfect fits.

comparing a histogram with a theory, a binned likelihood ratio[5] can be used. It is the ratio of the likelihoods of the data for the hypothesis, and of the data with the best possible theory which predicted exactly the observed numbers of events in each bin[10]. For data more or less consistent with the theory, the logarithm of this likelihood ratio is asymptotically approximately -0.5 times the χ^2 for the fit.

For the simple case of a Gaussian *pdf* $1/(\sqrt{2\pi}\sigma)\exp(-0.5(t-t_{pred})^2/\sigma^2)$, with observed value t_{obs}, the logarithm of the likelihood is

$$\begin{aligned}\log L &= -0.5*(t_{obs}-t_{pred})^2/\sigma^2 - \log(\sigma) - 0.5*\log(2\pi) \\ &= -0.5*\chi^2 - \log(\sigma) - 0.5*\log(2\pi) \end{aligned} \quad (1)$$

where χ^2 measures the degree of fit between $t_{obs} \pm \sigma$ and t_{pred}. Thus up to an additive constant, the chi-squared and -2 times the log of the likelihood agree. In the Baker and Cousins approach the likelihood is compared with its largest possible value L_{best} when $t_{pred} = t_{obs}$, in which case $L_{best} = -\log(\sigma) - 0.5 * \log(2\pi)$. Then $-2*\log(L/L_{best}) = \chi^2$, as stated in the previous paragraph. For non-Gaussian *pdf*s, the equality does not hold (but then $(t_{obs}-t_{pred})^2/\sigma^2$ is not distributed like χ^2 anyway).

Now let us assume we are comparing the data with two different hypotheses, H_0 and H_1, each described by Gaussians. The likelihood functions are

$$L_i = 1/(\sqrt{2\pi}\sigma_i)\exp(-0.5(t-t_i)^2/\sigma_i^2) \quad (2)$$

and the likelihood ratio is now given by

$$-2*\log(L_0/L_1) = \chi_0^2 - \chi_1^2 + \log(\sigma_0/\sigma_1) \quad (3)$$

Thus when the widths of the Gaussians for the two hypotheses are not equal, there is an offset between $-2*\log(L_0/L_1)$ and the difference in χs for the two fits. If these variables are being used to select between the two hypotheses, it would be sensible to choose different cuts for them.

6.3 When neither H_0 or H_1 is true

It may well be that neither H_0 nor H_1 is true. With no more information available, it is of course impossible to say what we expect for the distribution of our test statistic t. On the plot of fig. 2(b), our data may fall in the small rectangle next to the origin. It is certainly not true that a small value for p_0 implies that H_1 is correct, although for small enough p_0, ruling out H_0 is a possibility.

David Cox[7] has drawn attention to the situation where H_0 is nearly correct. A small amount of data may well appear to be consistent with H_0, but with more data, the discrepancy may become apparent. Then p_0 would be small (but L_0/L_1 may well still favour H_0 – see the second paragraph of Section 9.2 and

[5]It is to be noted that this likelihood ratio involves only a **single** hypothesis H_0, while the Neyman–Pearson theorem uses the likelihood ratio for two **different** hypotheses H_0 and H_1.

fig. 2(d). Cox comments that the discrepancy could be significant statistically, but insignificant physically.

With enough data, we may be able to include physically motivated corrections to our naive H_0.

7 CL_s

The CL_s method[11, 12] has been used at the LEP experiments at CERN (and to some extent at the Fermilab Tevatron) in searches for new particles. When evidence for such a particle is not found, the traditional frequentist approach is to exclude its production if p_1 is smaller than some preset level γ, which is typically set at 5%. However, there is then a 5% probability that H_1 could be excluded, even if the experiment was such that the H_0 and H_1 *pdf*'s lay on top of each other i.e. there was no sensitivity to the production of the new phenomenon. To protect against this, the decision to exclude H_1 is based on $p_1/(1-p_0)$, known as CL_s[6]. It is thus the ratio of the left hand tails of the *pdf*s for H_1 and H_0. Fig 2(c) shows a (p_0, p_1) region for which H_1 is excluded by CL_s. The fact that it is clearly smaller than for the standard frequentist exclusion region is the price one has to pay for the protection it provides against excluding H_1 when an experiment has no sensitivity to it. We regard it as conservative frequentist.

It is interesting to note that the CL_s exclusion line in fig. 2(c) for the case of two Gaussians is identical to that obtained by a Bayesian procedure for determining the upper limit on μ_1 when the latter is restricted to positive values, and with a uniform prior for μ_1. In a similar manner, the standard frequentist procedure agrees with the Bayesian upper limit when the restriction of μ_1 being positive is removed.

In principle, similar protection against discovery claims when the experiment has no sensitivity could be employed, but it is deemed not to be necessary because of the different levels used for discovery or exclusion of H_1 (typically $3 * 10^{-7}$ and 0.05 respectively).

8 Bayesian Methods

The Bayesian approach is more naturally suited to making statements about what we believe about two (or more) hypotheses in the light of our data. This contrasts with Goodness of Fit, which involves considering other possible data outcomes, but focusses on just one hypothesis.

The complications of applying Bayesian methods to model selection in practice are due to the choices for appropriate priors. This is particularly so for those parameters which occur in one model but not in the other(s).

[6]This stands for 'confidence level of signal', but it is a poor notation, as CL_s is really the ratio of p-values, which is itself not even a p-value, let alone a confidence level.

Confusion is also caused by apparently different definitions of CL_s. These are related to different conventions about which way p-values are defined.

Loredo[13] and Trotta[14] have provided reviews of the application of Bayesian techniques in Astrophysics and Cosmology, where their use is more common than in Particle Physics.

8.1 Likelihood ratio

All Bayesian methods for choosing between hypotheses involve the likelihood ratio. For simple hypotheses, this is just $L_0(x)/L_1(x)$, where $L_i(x) = p(x|H_i)$, the probability (density) for observing data x for the hypothesis H_i. The issue is going to be how nuisance parameters[7] μ are dealt with for non-simple hypotheses. For the likelihood approach (as opposed to the Bayesian one, which also require priors), it is usual to profile them i.e. L_i now becomes $p_i(x_i|H_i, \mu_{best})$, where μ_{best} is the set of parameters which maximise L. The profile likelihood approach is a popular method in Particle Physics for incorporating systematics in parameter determination problems.

8.2 Bayesian posterior odds ratio

When there are no nuisance parameters involved, the posterior odds for H_i are $p_{post}(H_0|x)/p_{post}(H_1|x)$, where

$$p_{post}(H_i|x) = L_i(x)\ \pi_i, \tag{4}$$

and π_i is the assigned prior probability for hypothesis i. For example, the hypothesis of there being a Higgs boson of mass 110 GeV might well be assigned a small prior, in view of the exclusion limits from LEP.

With nuisance parameters, the posterior probabilities become

$$p_{post}(H_i|x) = \int L_i(x|\mu)\ \pi_i(\mu)\ \pi_i\ d\mu \tag{5}$$

where $\pi_i(\mu)$ is the joint prior for the nuisance parameters of hypothesis i. i.e. we now have **integrated** over the nuisance parameters. This contrasts with the likelihood method, where **maximisation** with respect to them is more usual. Even with $\pi_i(\mu)$ being a constant, integration and maximisation can select different regions of parameter space. An example of this would be a likelihood function that has a large narrow spike at small μ, and a broad but lower enhancement at large μ.

In relation to all Bayesian methods, it is to be emphasised that the choice of a constant prior, especially for multi-dimensional μ, is by no means obvious. Very often, there are several possible choices of variable for the nuisance parameters, with none of them being obviously more natural or appropriate that the others. Thus a point in 2-dimensional space could be written as Cartesian (x, y) or polar (r, θ); constant priors in the two sets of variables are different. Similarly

[7] For the purpose of model comparison, any parameters are considered as nuisance parameters, even if they are physically meaningful. e.g. the parameters of a straight line fit, the mass of the Higgs boson, etc.

in fitting data by a straight line $y = a + b * x$, using a seemingly innocuous flat prior for $b = \tan\theta$ results in the undesirable feature that angles θ in the range 0° to 89° have the same prior probability as those in the range 89.98° to 89.99°.

It should be realised that the results for Hypothesis Testing are more sensitive to the choice of prior than in parameter determination. Thus in parameter determination, sometimes a prior is used which is constant over a wide range of μ, and zero outside it. The resulting range for the parameter, as deduced from its posterior, may well be insensitive to the range used, provided it includes the region where the likelihood $L(\mu)$ is significant. For comparing hypotheses, however, there can be parameters which occur in one hypothesis but not the other. (An example of this is where H_1 corresponds to smooth background plus a peak, while H_0 is just smooth background.) The width of their priors affects their normalisation, and hence affects the Bayes factor (see next Section) directly. Bayesian statisticians admit that this is a problem.

8.3 Bayes factor

For each hypothesis we define $R_i = p_{post}/\pi$, where as usual p_{post} and π are respectively the posterior and prior probabilities for hypothesis i. Thus R is just the ratio of posterior and prior probabilities. Then the Bayes factor for the two hypothesis H_0 and H_1 is $B_{01} = R_0/R_1$. If the two hypotheses are both simple, then this is just the likelihood ratio. If either is composite, the relevant integrals are required for p_{post}. A small value of B_{01} thus favours H_1.

As already mentioned, the priors for these nuisance paramters can have a strong effect on the Bayes factor. This has been investigated for a toy problem by Heinrich[15], who found it was extremely difficult to find satisfactory functional forms for the priors which produced reasonable behaviour for the result of the hypothesis comparison.

Demortier[16] has drawn attention to the fact that it can be useful to calculate the **minimum** Bayes factor[17]. This is defined as above, but with the extra nuisance parameters of H_1 set at values that minimise B_{01}, i.e. they are as favourable as possible for H_1. If even this value of B_{01} suggests that H_1 is not to be preferred, then it is a waste of time to investigate further since any choice of priors for the extra parameters cannot make B_{01} smaller.

8.4 BIC and AIC

These are variants of the Bayes factor approach, where the statistic on which a decision is made is $-L_{max} + C$, where L_{max} is the likelihood maximised with respect to the free parameters, and $C = k \log n$ for BIC (the Bayes Information Criterion) and $C = 2k$ for AIC (the Akaike Information Criterion). Here n is the sample size and k is the number of free parameters. In either approach, the hypothesis for which the statistic is smallest is selected. Thus they appear to have the attractive feature of doing away with the need for priors. Although this is true, priors are implicitly there, in that the motivation for the factors C is based on specific choices for the priors.

There is little or no experience of using these criteria in High Energy Physics. Until serious study is made of their properties, their use is not recommended.

9 Comparison of different methods

9.1 χ^2 and $\Delta\chi^2$

A paradox that appears in different guises is as follows.

A fit with one free parameter p is being performed to a histogram of 100 bins. Under the usual asymptotic assumptions, we expect the weighted sum of squared deviations $S(p)$ to be such that $S(p_0)$, the value of S for the best value p_0 of the parameter, is distributed like χ^2 for 99 degrees of freedom. This implies that its expected value is 99 ± 14. Thus a value of $S(p_0) = 85$ would not be unusual. Now a colleague has a theory which predicts a value of p_{th}, and wants to know whether the data confirms this. We again calculate the weighted sum of squares and find $S(p_{th}) = 110$. There appear to be two possible answers.

The first relies on the fact that the probability of obtaining a value of 110 or larger for S is better than 20%, and so we accept the value p_{th} of the parameter. The second approach uses the fact that to calculate the uncertainty on our best fit estimate p_0, we find how much we must change p in order for $S(p)$ to increase by 1 unit. Now $S(p_{th})$ is 25 units above the minimum, and so in the approximation that $S(p)$ is parabolic near its minimum, p_{th} is 5 standard deviations away from our best estimate, and hence is totally unacceptable.

Ref. [18] explains that using the difference in χ^2 values gives better discrimination between the hypotheses. It can be used in the following situations:

- Different fits are being performed to a set of data, involving functional forms like a straight line (H_0) or a higher order polynomial such as a cubic.
- A mass histogram is being compared with a smooth background (H_0), or with a smooth background plus a peak of specified shape, but whose position, width and amplitude are regarded as fit parameters.
- Different fits are being performed to a set of data, involving functional forms like a straight line of negative gradient (H_0), or a decreasing exponential.

In the first two examples, the fit with a larger number of parameters will be such that its weighted sum of squares S_1 will never be larger than S_0, that for the simpler hypothesis H_0. However, if H_0 is true, we do not expect the difference ΔS to be large. We need to quantify what constitutes a large difference, in which case we would tend to accept the alternative hypothesis. In the first example, this is easiest as asymptotically and with H_0 being true, we expect ΔS to follow a χ^2 distribution[19] with the number of degrees of freedom equal to the number of extra free parameters (2 in this case). This is because this example satisfies the necessary conditions: the hypotheses must be nested, and

the parameters of the larger model must all be defined and away from their physical boundaries under H_0.

Even though the hypotheses are nested in the second example above, the other conditions are not satisfied and ΔS thus does not follow a χ^2 distribution. This means that in order to interpret the value of ΔS for our data, we need to determine its expected distribution for H_0 ourselves (i.e. by simulation), which is a big nuisance.

The two hypotheses in the third case involve the same number of parameters, and are not nested. This is a more obvious example of where we need to calculate the expected distribution of ΔS ourselves.

9.2 Why p is not equal to B

In the scientific literature there is sometimes discussion of why a Bayes' factor approach (essentially the likelihood ratio, provided there are no nuisance parameters) can give a very different numerical answer to a p-value calculation. A reason some agreement might be expected is that they are both addressing the question of whether there is evidence in the data for new physics.

In fact they measure very different things. Thus p_0 simply measures the consistency with the null hypothesis, without any regard to the degree of agreement with the alternative, while the the likelihood ratio takes the alternative into account. More specifically, p_0 is the probability of obtaining data at least as discrepant with the H_0 expectation, assuming H_0 is true. In contrast the Bayesian odds give the relative (Bayesian) probabilities of H_0 or H_1 being true. There is thus no reason to expect them to bear any particular relationship to each other. This can be illustrated by contour plots of constant values of the Bayes factor B_{01} on a p_0 versus p_1 plot (see fig. 2(d)). The figure is constructed by assuming that the *pdf*'s for the two hypotheses H_0 and H_1 are given by Gaussian distributions of equal widths. Then at constant p_0, it is seen that B_{01} can take a range of values, corresponding to the Gaussians having different separations. Thus with the Gaussian for H_0 centred at zero and with unit width, a measured value of 5.0 yields a p_0-value of $3 * 10^{-7}$, regardless of the position of the H_1 Gaussian. Such a small p-value is usually taken as sufficient to reject H_0. As the centre of the H_1 Gaussian starts at $\mu_1 = 0$ (i.e. the two Gaussians at this stage are identical) and steadily moves to larger values, p_0 of course remains constant, but B_{01} rises to unity when $\mu_1 = 10$ and from then on decreases to arbitrary small values. At that stage, the data are more in agreement with H_0 than with H_1.

Simulations of repeated measurements would be incapable of enabling us to decide which of the two types of approach is better. This is because we should find that each method does indeed correctly estimate the quantity that it is purporting to quantify. This statement is to some extent weakened by the effects of data being discrete, approximate treatment of nuisance parameters, the definition of the ensemble of experiments for p_0 or the choice of Bayesian priors.

9.3 A real example involving p and B

An intriguing example of the way that p-values and Bayes factors can lead to different conclusions is provided by the early data of the CLAS Collaboration, looking for pentaquarks[20]. Their histogram of the mass values of nK^+ combinations in the reaction $\gamma D \to n\ K^+\ p\ K^-$ has a suggestive looking peak (see fig. 4). Their initial paper[20] states that the probability of this being a statistical fluctuation of the background is equivalent to $5.3 \pm 0.6\ \sigma$, and hence this is strong confirmatory evidence for a pentaquark state.

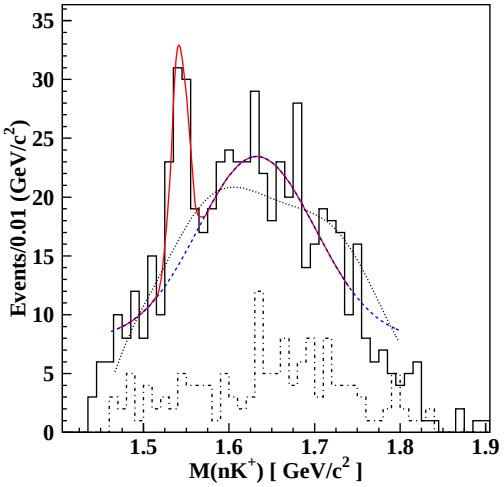

Figure 4: A histogram of the $n\,K^+$ effective mass is plotted for the reaction $\gamma\ D\ \to\ n\ p\ K^+\ K^-$. If a pentaquark decaying into a neutron and a K^+ is produced in these reactions, a narrow peak should appear in this histogram, but if not the distribution should be smooth. The curve is an attempt to deduce this smooth background. Does the histogram provide evidence for a new particle, as opposed to there being a statistical fluctuation from the smooth background, and/or an incorrectly estimated background?

With more data, this peak was not apparent[21]. The CLAS collaboration then reanalysed their data using a Bayesian approach. Even for just the original data sample of fig. 4 which had been claimed to yield strong evidence for a pentaquark, they now concluded that their data was marginally in favour of the hypothesis that there was only smooth background and **no** peak (as compared with the alternative hypothesis of there being a 'pentaquark' state above a smooth background).

These dramatically different conclusions for the frequentist and Bayesian approaches for the data of fig. 4 are potentially very worrying. It is possible

to argue (see, for example, Cousins' comments[22]) with the specifics of the calculated p-value and Bayes factor of references [20] and [21], and so the degree of inconsistency may not actually be as strong as claimed. Also Heinrich has pointed out that there are formidable problems in choosing Bayesian priors which produce reasonable results[15]. Berger[23] has responded by pointing out that frequentist methods often have poorly defined ensembles, which can affect frequentist calculations. It is clear that in analyses of this kind, it is informative to study the sensitivity of the conclusions to the type and details of the approach used.

10 Including Systematics

Various ways of incorporating systematics in analyses have been discussed at the relevant places throughout this note. Here we simply summarise some of the main points for convenience.

Uncertainties in almost any parameter associated with an analysis usually result in a reduction in the significance of any observed effect. Systematic effects are likely to be most important in situations where the background is uncertain, the expected signal gives rise to a wide enhancement rather than a narrow peak, or there is a lot of data so that the statistical uncertainties are small.

In Bayesian methods, systematics are dealt with by assigning them priors, and then integrating over the relevant variable. When the systematic is based on another measurement, this is a relatively straightforward procedure, with the prior being chosen to encapsulate the information from the subsidiary experiment. However, in other cases, little or nothing is known about the systematic effect, and then the choice of prior is much more problematic. (This situation would also create problems for frequentists.) We have also remarked that in choosing between two or more hypotheses, Bayesian methods are particularly sensitive to the choice of prior for parameters which are not common to the different hypotheses.

Methods for incorporating systematics in p-value calculations include:

- Conditioning: In cases where the problem has a certain structure, it may be possible to condition on an ancillary variable, which contains no information about the parameter of interest. Thus if the background is estimated in a subsidiary measurement, it may be possible to condition on the sum of the numbers of counts in the main and the subsidiary experiments, and then to use the binomial distribution to obtain the p-value.

- Plug-in value: The best estimate of the nuisance parameters is used to calculate p, where the nuisance parameters are calculated assuming the null hypothesis.

- Prior predictive value: The p-values are averaged over the nuisance parameters, weighted by their prior distributions. This is in the spirit of the Cousins-Highland approach[24] for incorporating systematics in upper limit calculations.

- Posterior predictive 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 finite number of possible alternative theoretical interpretations.

- Confidence interval: A confidence region of size $1 - \gamma$ 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 have been compared by Demortier [25], while Cranmer [26] and Cousins et al[27] have 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.

10.1 Marginalise or maximise?

It is conventional in likelihood approaches to use profiling i.e. to **maximise** with respect to nuisance parameters. On the other hand, Bayesian methods usually calculate joint posteriors that include the nuisance parameters, and then **integrate** (or 'marginalise') with respect to them. Given that with constant priors the likelihood and the posterior are proportional to each other, it is interesting to consider further the difference between profiling and marginalisation.

If the likelihood is a multi-dimensional, possibly correlated Gaussian, then the two procedures for eliminating the nuisance parameters will lead to the same result. However it is not difficult to think of multi-dimensional functions for which the results can be different. In general, with a function that has a narrow high peak and a lower broad enhancement, profiling will pick out the narrow peak, while marginalisation may be dominated by the lower broad enhancement.

11 Conclusions

Choosing between hypotheses is a non-trivial operation, especially in the presence of nuisance parameters (which are almost always there). In order to make a convincing case for New Physics, it will almost always be insufficient to show that the data are inconsistent with the SM. What is needed is to demonstrate that some other scenario provides a (much) better explanation for the data.

A recurring feature is the need to incorporate nuisance parameters. This is likely to be especially important in the early data-taking with new accelerators and detectors. There are many ways in which this can be done.

The profile likelihood ratio is commonly used, largely because it seems a natural extension of the likelihood ratio which, by the Neyman–Pearson theorem, is optimal for simple hypotheses. For realistic particle physics applications, its distributions for H_0 and for H_1 will need to be determined (almost always by Monte Carlo simulation) in order to assess the significance of its observed value; and this needs to performed for different values of the nuisance parameters.

Bayesian approaches in principle provide a natural setting for determining which hypothesis we should believe, and it is not restricted to comparison of just two hypotheses. However, there can be serious problems in choosing suitable priors for the parameters, and this can create difficulties in interpreting the result.

As usual frequentist approaches only answer questions such as how likely it is to obtain data like ours or more extreme, using various hypotheses, and our deductions about whether we have a discovery have to be based on this. In searches which have several chances of discovering a new effect, we recommend including a 'look elsewhere' factor in the quoted p-value. This should allow for the effective number of opportunities for making a discovery in the performed search.

It is important to decide in advance the details of the technique to be used to assess the significance of any discovery. Using a variety of methods and then selecting the one which gives the most desirable result invalidates any properties of method.

I wish to acknowledge the patience and expertise of David Cox, Martin Crowder, Brad Efron and Steffen Lauritzen 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. Warm thanks are due to Luc Demortier (who also helped produce the diagrams) and to Tom Junk for very useful comments on this note. My understanding of the practical application of statistical techniques has improved considerably as a result of discussions with many experimental Particle Physics colleagues; I especially want to thank the members of the CDF Statistics Committee and Bob Cousins. To all of you, I am most grateful.

References

[1] B. Efron and A. Gous, 'Scales of evidence for model selection: Fisher versus Jeffreys', Symposium on 'Model Selection, Empirical Bayes and related topics' (Nebraska 1999).

[2] Larry Wassserman, 'Bayesian model selection and model averaging', Symposium on Methods for Model Selection (Bloomington 1997).

[3] J. O. Berger, 'Could Fisher Jeffreys and Neyman have agreed on testing?', Fisher Lecture at Joint Statistical Meetings (2001).

[4] Proceedings of PHYSTAT-LHC Workshop at CERN (2007).

[5] L. Demortier, 'Search Procedures' (2009) `http://physics.rockefeller.edu/~luc/talks/SearchProcedures.pdf`

[6] G. Cowan, remark at the PHYSTAT meeting in Durham (reference [29]).

[7] D. R. Cox, 'Statistical significance tests', Br. J. Clin. Pharmacol. **14** (1982) 325.

[8] L. Demortier, 'Some statistical issues in the measurement of the top quark charge', CDF note 9426 (2008).

[9] J. Heinrich, 'Pitfalls of Goodness-of-Fit from Likelihood', ref. [30], page 52.

[10] S. Baker and R. Cousins, Nuclear Instr. and Meth. **A221** (1984) 437.

[11] A. Read, 'Modified frequentist analysis of search results', ref. [28], page 81; 'Presentation of search results - the CL_s method', ref. [29] page 11.

[12] T. Junk, 'Sensitivity, exclusion and discovery with small signals, large backgrounds and large systematic uncertainties', CDF note CDF/DOC/STATISTICS/PUBLIC/8128 (2007).

[13] T. J. Loredo, 'From Laplace to Supernova SN1987a: Bayesian inference in Astrophysics', in 'Maximum Entropy and Bayesian Methods' (Kluwer Academic Publishers, 1990), p81.

[14] R. Trotta, 'Bayes in the sky: Bayesian inference and model selection in Cosmology', Contemporary Physics 49 (2008) 71.

[15] J. Heinrich, 'A Bayes factor example', CDF note 9678 (2009).

[16] L. Demortier, 'The minimum Bayes factor', CDF note 9710 (2009), `http://physics.rockefeller.edu/~luc.memos/bfmin.pdf`

[17] V.E. Edwards, H. Lindman and L. J. Savage, 'Bayes statistical inference for psychological research', Psychological Review **70** (1963) 193.

[18] L. Lyons, 'Comparing two hypotheses', Oxford preprint `http://www-cdf.fnal.gov/physics/statistics/statistics_recommendations.html`

[19] S. S. Wilks, 'The large-sample distribution of the likelihood ratio for testing composite hypotheses', Annals of Math. Stat. **9** (1938) 60.

[20] S. Stepanyan et al, 'Observation of an Exotic $S = +1$ Baryon in Exclusive Photoproduction from the Deuteron', Phys. Rev. Lett. **91** 252001 (2003).

[21] D. G. Ireland et al, 'A Bayesian analysis of pentaquark signals from CLAS data', Phys. Rev. Lett. **100** 052001 (2008).

[22] R. Cousins, Comment on 'Bayesian Analysis of Pentaquark Signals from CLAS Data', http://arxiv.org/abs/0807.1330

[23] J. Berger, private communication.

[24] R. Cousins and V. L. Highland, Nucl Inst and Meth **A320**(1998) 391.

[25] L. Demortier, 'p-values and nuisance parameters', ref. [4], page 23.

[26] K. Cranmer, 'Statistics for LHC: progress, challenges and future', ref. [4], page 47.

[27] R. Cousins, J. Linnemann and J. Tucker, ref. [4], page 11.

[28] 'Workshop on Confidence Limits', CERN Yellow Report 2000-05.

[29] 'Advanced Statistical Techniques in Particle Physics', Durham IPPP/02/39 (2002).

[30] Proceedings of PHYSTAT2003, eConf C030908, SLAC-R-703.