Statistical Techniques in High Energy Physics

Louis Lyons

Nuclear and Astrophysics Laboratory, Keble Rd, Oxford, UK

Abstract. Rather than attempting to cover a wide range of statistical problems, I shall concentrate on four topics: 1) The argument between Bayesians and Frequentists, 2) A paradox in comparing data with two hypotheses; 3) The CLs method used in the search for the Higgs at CERN, 4) The MLBZ method of using data to estimate some of the systematic effects in measuring the W boson's mass.

BAYES VERSUS FREQUENTISM

The Confidence Limits Workshops[1,2] earlier this year brought into sharp focus the differences between the Bayes and Frequentist approaches.

Bayesians start from Bayes' Theorem

$$P(A \text{ and } B) = P(A|B) * P(B) = P(B|A) * P(A) \qquad (1)$$

where P(A|B) means the probability of A happening, given that B has happened. This is completely uncontroversial when used in situations where A and B describe certain events (in the statistical sense) e.g. for a high energy $p\bar{p}$ interaction, A and B are the production of a W boson and a top quark respectively. However Bayesians use eqn (1) for A = hypothesis (or the value of a parameter) and B = data whence (1) can be rewritten

$$P(\text{parameter}|\text{data}) \propto P(\text{data}|\text{parameter}) * P(\text{parameter}) \quad (2)$$

i.e. (posterior prob) $\propto$ (likelihood fn) * (prior). Thus after performing an experiment, the posterior knowledge about a parameter is obtained by combining one's prior knowledge with the likelihood function, as deduced from the experimental data.

To frequentists, this is anathema because they would object to making a probability statement about the value of a physical parameter. For them, the value of as either is between 0.115 and 0.120, or it is not (even if we do not know which), and it does not make any sense to ascribe a probability to this.

Furthermore, to deduce the Bayesians' posterior probability requires a functional form for the prior; there is some arbitrariness in how this should be chosen. It is tempting to try to assign an "uninformative prior", e.g. one which is flat in the parameter. However, flat in p is not the same as flat in p^2 or $\sqrt{p}$ or log(p), and it is not usually clear which is best. A frequentist would want the result of an experiment to be independent of such arbitrary choices.

Bayesians, on the other hand, would interpret P(parameter|data) or P(parameter) not so much as a probability in the classical sense, but more as a degree of belief, and then quote a "credible interval" for the parameter. Concerning priors, they would either justify the use of subjective priors as expressing genuine differences of knowledge of different experimentalists; or they would attempt to find priors with some theoretical justification. They would further argue that the Bayesian approach most closely resembles the way scientists make decisions; in deciding what research to pursue, personal judgments play an important role. Another effect of the prior is that unphysical values of a parameter are excluded; this is not necessarily so in frequentist approaches.

Frequentists construct confidence intervals without invoking P(parameter|data) or P(parameter), and hence do not require a prior. Their method is simply to use P(data|parameter) to construct a probability interval for the data (i.e. the result of the experiment) for each value of p. Thus the shaded band of Fig. 1 shows the

likely result of the experiment for each p. Then for the given result of a particular experiment, the confidence belt for p is given by where a vertical line at the experimental result cuts the shaded band (see Fig. 1).

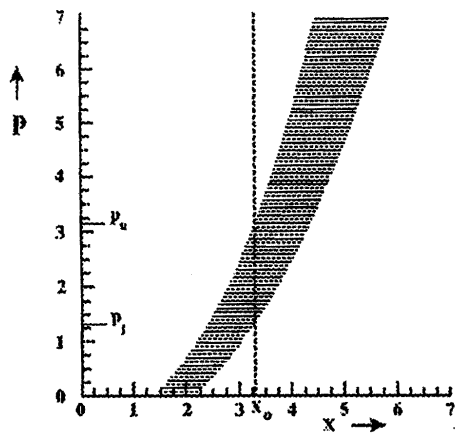

FIGURE 1. Confidence belt. p is a theoretical parameter and x is an observation. For example, p could be the temperature at the centre of the Sun, and x the production rate for events in a solar neutrino detector. For a given p, a band in x is calculated such that the probability of observing x in this range is 90%. As p takes on all possible values, the shaded region is obtained; this gives the likely values of x for any p. For a given experiment observing x_0, the range p_l to p_u contains those values of p for which x_0 was a likely observation.

The frequentist thus produces a statement such as

$$p_l < p < p_u \text{ at 90\% confidence.} \qquad (3)$$

This is based just on the data, without any preconceived ideas about the relative probabilities of different values of p. In eqn (3), the true value of p is unknown but regarded as a constant, while the known p_l and p_u are regarded as random variables; eqn (3) is thus a statement about the probability of the random p_l and p_u containing the unknown p. In a similar statement by Bayesians, p is regarded as a random variable with a probability distribution, and p_l and p_u are constants.

For accurate measurements where the errors are approximately Gaussian (e.g. M_Z, the mass of the Z as determined at LEP, which is 91188±2 MeV), the two approaches give the same results.[3] This is the case where, for the Bayesian, "the data overshadows the prior". The functional form of the prior is then unimportant, because any reasonable prior will be virtually constant over the small mass range of interest. In contrast, in situations where we are dealing with limits, the relevant range extends down to zero, and so here the form of the prior can be very important. This is why the Frequentist-Bayes argument is more relevant in limit situations than for accurate measurements.

Thus Narsky[4] showed how variable upper limits could be, depending on the exact way they are determined. He considered the example of the observation of a given number of events, assumed to be Poisson distributed, when the expected background was b. For six different methods applied to n=3 and b=1.0, upper limits ranged from 0.3 to 3.3. When quoting upper limits, it is thus crucial to explain clearly how they were deduced.

One of the problems apparent from the Workshops was the confusion between P(hypothesis|data) and P(data|hypothesis). This should be clarified by the example of considering a certain unseen person, who is hypothesised to be either male or female. The data is whether or not they are pregnant. For random human beings,

$$P(\text{pregnant}|\text{female}) \sim 3\%$$

whereas P(female|pregnant) is considerably larger.

A specific example of a limit calculation discussed at the Workshops is given in 'CERN Higgs Search' below.

CHOOSING BETWEEN HYPOTHESES

There is a well-known paradox connected with parameter estimation and hypothesis testing.[5]

Imagine that a χ^2 method is being used on a histogram of 100 bins, to estimate the value of a single parameter p. Assume that p_0, the best value of the parameter gives a minimum χ^2 of 90. Then the error σ on p is given in terms of $p_1 = p_0 + \sigma$ the value of p for which χ^2 increases to 91.

Now consider p_2, another value of p, for which $\chi^2 = 115$. Is this value satisfactory?

The probability of $\chi^2 = 115$ for 99 degrees of freedom is not unreasonable. i.e. we would not want to exclude p_2. However, given that σ is such that $\chi^2(p_0+\sigma) = 91$, in the usual parabolic approximation, a χ^2 of 115 corresponds to a 5σ effect i.e p_2 is "completely" excluded. So which are we supposed to believe?

The answer is that, although in general a value of 115 for χ^2_{min}, the minimum value of a χ^2 variable, based on 99 degrees of freedom is satisfactory, in the case where χ^2_{min} is 90, a value of χ^2 (p) of 115 is not.

It is the **difference** in χ^2 of the two hypotheses ($p=p_0$ or $p=p_2$) which is relevant in discriminating between them.

This approach is thus important for reducing "errors of the second kind" i.e. accepting a hypothesis when it is in fact wrong. It can be used for discriminating between whether atmospheric neutrino data is more consistent with ν_μ oscillating to ν_τ or to $\nu_{sterile}$; or which set of parton distributions is consistent with jet production in high energy $p\bar{p}$ collisions. A more elaborate example is provided below.

CERN HIGGS SEARCH

The search for the Higgs at the LEP Collider at CERN is based on the CL_s method.[6] Basically the problem is to try to use the data to distinguish between two hypotheses: A) Background processes produced according to the Standard Model (S.M.) without the Higgs (or with a Higgs that is too heavy to be accessible); or B) S.M. plus Higgs of a certain mass.

For each Higgs mass, the aim of the procedure is to make one of the following choices: i) The data are inconsistent with B at some level (e.g. 5%), and hence a Higgs of that mass is excluded. ii) The data are inconsistent with A at some level (e.g. equivalent to a one-sided 5σ effect) and are more consistent with B. iii) The data cannot either exclude or confirm a Higgs of that mass.

From the data, a test statistic X is constructed. This is in fact the likelihood ratio for the two hypotheses. It involves not only the number of events, which of course is expected to be larger for B than for A, but other kinematic variables e.g. mass of Higgs candidates, whether there are b-hadrons in the possible Higgs decay, etc.

For a given mass Higgs, Monte Carlo simulation is used to predict the expected distributions of X for hypotheses A and B. (Because the production cross-section of a S.M. Higgs of a given mass is well-defined, these predictions involve no free parameters.) Then X from the data is compared with each distribution to decide whether it is possible to favour one hypothesis over the other. For distributions that are well-separated, this is relatively easy. For more realistic situations where there is some overlap (see Fig. 2(a)), a numerical procedure is required.

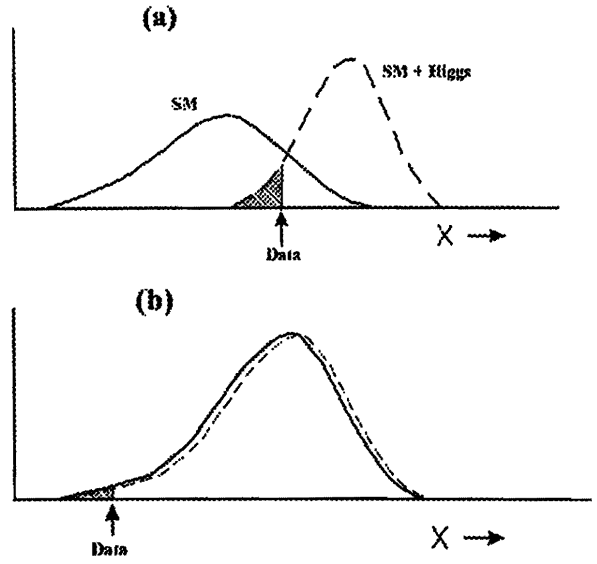

FIGURE 2(a). Predictions from simulation for a particular data statistic X, for Standard Model background only (solid curve) and SM plus Higgs of a particular mass (dashed curve). The shaded area gives CL_{bs}, the fractional area of the dashed curve that is more background-like than the data. CL_b is the corresponding fractional area on the SM curve to the left of the data. Finally CL_s is defined as CL_{bs}/CL_b. **FIGURE 2(b).** For a very heavy Higgs that is barely produced, the solid and dashed curves become almost indistinguishable. If the data fluctuates downwards, CL_{bs} is small, but CL_s is still close to unity. The advantage of CL_s is that it prevents the exclusion of such a Higgs.

In order to see whether exclusion of Higgs is possible for the given value of X for the data, CL_{bs} is defined as the fractional area of the simulated "SM + Higgs" distribution to the left of X (i.e. more background-like). The usual frequentist approach is to exclude this hypothesis if CL_{bs} is less than some preset level (say 5%). However the CERN Higgs group want to avoid the situation in which, for a heavy Higgs which it is barely possible to produce (see Fig. 2(b)), a downward data fluctuation could result in CL_{bs} being small enough to exclude the Higgs, even though the experiment has no sensitivity to it; this would happen at the 5% probability level. So a more conservative approach is adopted by defining

$$CL_s = CL_{bs}/CL_b$$

where CL_b is the fractional area of the "background only" distribution to the left of the data X. Thus CL_s is the ratio of two confidence levels, rather than itself being a confidence level.

The data is then said to exclude the Higgs of that mass if CL_s is below a certain cut e.g. 5%. Since CL_s is forced to be larger than CL_{bs}, the cut on CL_s is

conservative in its coverage i.e. the Higgs hypothesis will be excluded when the Higgs is really there in not more than 5% of experiments. Then in the situation of Fig. 2(b), CL_s is close to unity, and the CL_s method will never be able to exclude a Higgs to which an experiment has no sensitivity.

Although a relatively modest level (95%) is chosen for exclusion, in order to be sure of a discovery claim, a higher degree of confidence is required. The criterion is that CL_b is required to be very close to unity: 1- CL_b is below 5.6 x 10^{-7}. Since the probability of this happening for a random fluctuation of the background is so small, it is not deemed necessary to adopt the analogy with the exclusion procedure, and to divide by 1-CL_{bs}.

Fig. 3 shows CL_s and CL_b for recent data.[7] These exclude Higgs masses up to 113.6 GeV, and there is a signal-like effect in CL_b around 115 GeV, but not at the required discovery level.

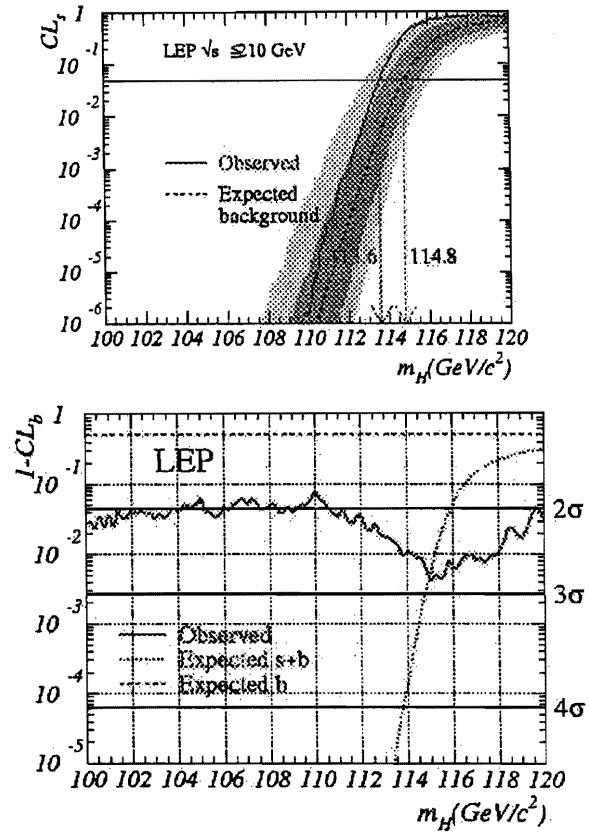

FIGURE 3. Search for the Higgs at LEP. CL_s as a function of M_H is used to see whether it is possible to exclude the Higgs at that mass. The combined data from the four LEP experiments has CL_s (solid curve) below 5% for masses up to 113.6 GeV, and so masses below that are excluded. The median expected value (dashed curve) passes through 5% at 114.8 GeV. In the lower diagram, 1-CL_b is used to see whether there is evidence of the Higgs; the data is shown as the solid curve. The expectation for background only is the horizontal dashed line at 0.5, while that for 'background + Higgs' is the dotted curve. The data shows an effect at slightly below 3σ significance for M_H around 115 GeV. The total luminosity is such that a 5sigma signal is not expected for M_H above 113 GeV.

MLBZ

Systematic errors are much more problematic than statistical ones. Here we discuss a novel approach to trying to estimate realistically some of the systematic errors in the determination of M_W, the mass of the W boson, produced in the reaction

$$e^+e^- \rightarrow W^+W^- \rightarrow 4 \text{ jets} \qquad (4)$$

at the CERN LEP Collider at centre of mass energies around 200 GeV.

Some of the potential sources of error are associated with the estimated jet energies and directions. These include the detector resolution, and the way the fragmentation of quarks and their subsequent hadronisation are described. The traditional method is to use simulation techniques to estimate these, but as ever the question is how realistic these simulations are. The Mixed Lorentz Boosted Z^0s (MLBZ) approach[8] instead uses data to check these effects.

It relies on the fact that the four jets in reaction (4) are rather similar to those of the two jets in

$$e^+e^- \rightarrow Z^0 \rightarrow 2 \text{ jets} \qquad (5)$$

at centre of mass energies around M_Z = 91 GeV. Thus if we take two Z^0 2-jet events, and boost them in opposite directions with a Lorentz boost β corresponding to reaction (4) at a given LEP energy, the event configuration will resemble that for reaction (4), albeit at a slightly higher energy. (Alternatively, the Z^0 jets can be scaled down in energy, so as to correspond to M_W). The same analysis procedure as used to extract M_W from the real 4-jet events is then used on MLBZ events, constructed either from real data or from simulated events of reaction (5). By comparing the mass shifts observed between real and simulated MLBZ events, the reliability of the simulation to predict detector, fragmentation and hadronisation effects can be checked.

Sophistications include reweighting MLBZ events to allow for the different angular distributions of the jets in reactions (4) and (5); to incorporate the finite

width of the bosons in (4), compared with all Z^0 in (5) being produced at the same centre of mass energy; to correct for the fact that Z^0 decays include b jets, while W decays do not; etc.

Of course the MLBZ technique does not include all systematics (e.g. initial state radiation, colour reconnection, etc.) but for those that are, it provides a more direct and powerful way of estimating them. Thus the DELPHI systematic errors calculated by MLBZ have now been explored to a level of precision which is more than a factor of 4 better than was previously possible using simulation.

The method can be extended to deal with aspects of W physics other than just M_W (e.g. triple gauge couplings); and with other reactions e.g.

$e^+e^- \rightarrow W^+W^- \rightarrow$ 2 jets + lepton + neutrino
$e^+e^- \rightarrow Z^0 Z^0$
$e^+e^- \rightarrow Z^0$ + Higgs

CONCLUSION

There continue to be interesting statistical analyses to perform in High Energy Physics. If you are aware of any challenging problems (and especially if you also know how to solve them!), please let me know.

REFERENCES

1. Workshop on Confidence Limits, CERN, 27-28 January 2000, CERN Yellow Report 2000-005.

2. Workshop on Confidence Limits, Fermilab, 27-28 March 2000, http://conferences.fnal.gov/cl2k/.

3. R. Cousins, Am. J. Phys. **63**, 398 (1995).

4. I. Narsky, Nuclear Instruments and Methods **A450**, 444 (2000).

5. L.Lyons, 'Selecting between two hypotheses', Oxford preprint OUNP-99-12.

6. A.L. Reed, ref.1) page 81; W. Murray, ref.2).

7. P. Igo-Kemenes, Talk at CERN LEPC meeting, 3rd Nov 2000; see also http://lephiggs.web.cern.ch/LEPHIGGS/talks/pik_lepc_nov3_2000.ps.

8. N. Kjaer and M. Mulders, 'Mixed Lorentz Boosted Z^0s', DELPHI 2000-51 CONF 366