COMMENTS ON METHODS FOR SETTING CONFIDENCE LIMITS

Robert D. Cousins *
Department of Physics and Astronomy, University of California,
Los Angeles, California 90095, USA

Abstract
I discuss a number of issues which arise when computing confidence limits by frequentist or Bayesian methods. I begin with a reminder why P(hypothesis|data) cannot be determined if the only input is the 'objective' data. I then discuss confidence intervals, with emphasis on the 'unified approach' based on likelihood-ratio ordering, and related methods. A number of issues arise, including conditioning, nuisance parameters, and robustness. For Bayesian methods, important issues are the prior and goodness-of-fit. I conclude with a list of items on which I think physicists from many points of view can agree.

1. PROLOGUE

For most of this talk[1], I assume familiarity with the 'required reading' for this workshop. But first, let's review the root of the problem as I often explain it to students. (Imagine an oral exam.)

Suppose you have a particle ID detector. You take it to a test beam and measure:

- P(counter says π | particle is π) = 90%
- P(counter says not π | particle is π) = 10%
- P(counter says π | particle is not π) = 1%
- P(counter says not π | particle is not π) = 99%

Then you put the detector in your experiment. You select tracks which the detector says are pions.

Question: What fraction of these tracks are pions?

Answer: *Cannot be determined from the given information!*

The missing information is the pion fraction in the particles incident on the detector: the initial P(π). Bayes's theorem then tells us that

P(particle is π | counter says π) $\propto$ P(π) $\times$ P(counter says π | particle is π).

All this makes total sense with the frequentist definition of P. Now suppose you look for a Higgs boson (H) at LEP and you do all the work to know:

- P(H signature | there is H) = 90%
- P(no H signature | there is H) = 10%
- P(H signature | there is no H) = 1%
- P(no H signature | there is no H) = 99%

There is no problem defining these P's with the frequentist definition of P. Then you do the experiment, and you have a Higgs signature.

Question: What is the probability that you found the Higgs?

Answer: *Cannot be determined from the given information!*

The missing information is the analog of P(π): the 'prior' probability that there is a Higgs: P(H). Again Bayes's theorem then tells us that

P(particle is H | H signature) $\propto$ P(H) $\times$ P(H signature | particle is H).

* E-mail address: cousins@physics.ucla.edu

[1] I attempt to preserve the conversational nature of the talk in this writeup.

But what is P(H)? It is problematic to define it with the frequentist definition of P. I think the most compelling definition of P to use here is *subjective* degree of belief.

So suppose with your subjective prior, you compute P(particle is H | H signature) = 98%. Now you must make a *decision* whether or not to announce the discovery of the Higgs.

Question: What decision logically follows from the above?

Answer: *Cannot be determined from the given information!*

The missing information is the *utility function*: How do you personally weigh a) wrongly announcing a discovery, versus b) failing to announce a real discovery. I think this utility function is indisputably *subjective*.

The oral exam concludes: Making a decision requires two subjective inputs: the prior and the utility function.

[An aside: While on the subject of the utility function, I mention that I think it is the preferred place to put *conservativism*, if one so desires. The analog with confidence intervals is the following: if one wants to avoid wrong statements more than 10% of the time, the proper way is not to compute 90% C.L. intervals in some 'conservative' way; the proper way is to compute intervals using a higher C.L.]

I think that this *subjective* Bayesian model of decision making is a good one for a scientific decision-making process[2]. In fact, to focus this workshop's discussion, let's stipulate that the best model for the scientific way to determine P(hypothesis|data) is to define P as degree of belief and invoke a subjective prior[3].

However, I think the question before us today is: How should we experimenters publish the numbers from our experiments? And how should the Particle Data Group (PDG) list them? I think that compromises are inevitable, because it is unlikely that the PDG will be asked to list the 'right' answers: subjective Bayes decisions.

My personal view is that:

- Subjective priors will not be accepted as the basis for reporting of experimental results. Therefore subjective priors are not the answer to today's question.
- 'Non-informative priors do not exist'[4], and the whole 'objective prior' search is not particularly useful.
- We should quote numbers based on the frequentist definition of P.
- That limits us to confidence intervals and the likelihood function.
- That means our published numbers will *not* be P(hypothesis|data) !

This last point is the lesson from the 'oral exam' questions of this prologue: without a prior, you cannot extract P(hypothesis|data). It is critically important to keep this in mind. It follows that what we publish will be a 'halfway house', incomplete but useful if not misinterpreted.

2. FELDMAN–COUSINS AND AFTERMATH

It has been over two years since Gary Feldman and I advocated a 'Unified Approach' [1] to confidence intervals, where unity refers to one-sided limits and two-sided limits. For frequentists, we quantified what a lot of people felt intuitively: The discussion of setting confidence limits cannot be separated from the discussion of setting two-sided intervals. In particular, we showed that 'flip-flopping' between upper limits and two-sided intervals, *based on the observed data*, leads to undercoverage.

Ciampolillo [2] has pointed out that he earlier understood this and found a (different) unified set of confidence intervals.

[2] Due to faulty reasoning, humans may not really act like Bayesians even if they intend to, but that is not my point today.

[3] There are also issues of goodness of fit, which I discuss below.

[4] I do not claim to understand completely this phrase, which is found in the professional statistics literature, but I am sceptical of attempting to represent absence of prior degree of belief.

I think the need for an approach which unifies one-sided limits (typically upper limits) and two-sided limits is now indisputable. (In Bayesian statistics, coverage does not exist, so the issue is different. To be coherent, one must use the same prior for upper limits as for two-sided intervals. For the Poisson mean case, this leads to undercoverage when evaluated by frequentist criteria; see Section 9.)

As noted 'in proof' in the Feldman–Cousins (FC) paper, the 'new' likelihood-ratio ordering principle in our unified approach followed naturally from the classical theory of hypothesis tests in the classic text by Kendall and Stuart, (now Stuart and Ord [3]). In fact, the whole program, including nuisance parameters, is tersely laid out in a page and a half at the beginning of Chapter 23! This is of course good news, since as physicists we would prefer to adopt well-established statistical procedures.

Positive features of this method are:

- The intervals are unified: no flip-flopping.
- The ordering for building the acceptance intervals is based on the likelihood ratio (LR), now known to be the 'standard' ordering in Stuart and Ord.
- It gives an improvement over old classical intervals for both Poisson-with-background and Gaussian-with-boundary cases.
- It can be applied to more general problems. We illustrated this with neutrino oscillations.
- It was immediately applied by several groups, so it is doable.
- One can add nuisance parameters à la Stuart and Ord (see below).
- It can be used to combine experiments.

A drawback for now is that these last two features are not widely known, and have subtleties. They have been implemented approximately (by G. Feldman and A. Geiser) in the NOMAD [4] neutrino oscillation experiment.

Finally, some consider it positive and some consider it negative that the method produces *confidence intervals*. Confidence intervals have a well-defined meaning in terms of the frequentist definition of P. The method does not use a prior, but that also means that the results must not be interpreted as P(hypothesis|data)!

2.1 The 'Karmen problem': mean background 2.8, see no events.

One of the first applications of the unified approach was by the Karmen Collaboration, which saw no events while expecting 2.8 background events. Our recommendations for a situation like this were:

- To educate the world that confidence intervals are not statements about P(hypothesis|data).
- To insist that people show a sensitivity curve [1] if their limit is far from it.

Nonetheless, the most common criticism of the unified approach is that 'it makes no sense' for Karmen to have a tighter upper limit than an experiment with no events and no expected background.

Please don't fall into the following common trap:

1. Assume the confidence intervals are statements about P(parameter|data).
2. Observe that unified confidence intervals violate all sensibility in the Karmen problem when interpreted as P(parameter|data).
3. Conclude that the confidence intervals make no sense.

What makes no sense is assuming that confidence intervals are statements about P(parameter|data). They are statements derived from P(data|parameter) *without invoking a prior*, and hence necessarily cannot be P(parameter|data). In the Karmen problem, the probability of observing no events is less for $b = 2.8$ than it is for $b = 0$. That's what the confidence intervals are reflecting. I come back to this point in discussing *conditioning* below.

2.2 LR ordering 'failures'

It is well-known that maximum likelihood estimation doesn't work in certain cases. That hasn't prevented it from being generally useful. Similarly, exceptions for LR ordering for confidence intervals have been claimed by G. Zech [5], and more recently by G. Punzi [6], and J. Bouchez [7]. It remains to be seen if these represent problems in practice.[5]

2.3 The Roe–Woodroofe modification: conditioning

Of all the post-FC papers, the one I found most enlightening was by Roe and Woodroofe [8]. They invoke a standard idea from the theory of statistics, namely *conditioning*. The idea is to restrict the ensemble used to define frequentist coverage, based on the data observed. Their application stretches the usual conditioning beyond well-known precedents, but gives intervals in the Poisson case which do not depend on the expected background for the $n = 0$ case (zero events observed). For $n = 0$, the idea is that one knows that there are no background events, so the chosen ensemble consists of experiments with no background events rather than the larger ensemble in which the number of background events fluctuates with known mean[6].

Conditioning can make confidence intervals behave more like P(parameter|data). This quantity only exists in Bayesian theory[7], and is proportional to P(data|parameter) evaluated only using the actual data set observed. Confidence intervals are based on P(data|parameter) for all possible data sets in an ensemble, and do not always behave similarly to P(parameter|data). Confidence intervals will in general behave more like P(parameter|data) if the ensemble is restricted to data sets more like the one observed. Traditionally this restriction is made using *ancillary* statistics; see Ref. [8] for more discussion and references.

Owing to the promising, yet somewhat unfounded, appearance of the approach of Roe and Woodroofe, I recently worked out and posted [9] the application of their method, as I understand it, to the other prototype problem of the FC paper: a Gaussian variable near a physical boundary. The result was disappointing: while the upper curve on the confidence belt is moved in the desirable direction, the lower curve on the confidence belt is also moved significantly, in an undesirable manner. We are currently studying the situation further.

3. THE METHOD CALLED THE OLD 'PDG METHOD' OUTSIDE THE PDG

This non-unified method, for upper limits only, is based on the formula

$$1 - \epsilon = 1 - \frac{e^{-(\mu_B+N)} \sum_{n=0}^{n_0} (\mu_b + N)^n / n!}{e^{-\mu_B} \sum_{n=0}^{n_0} \mu_B / n!} \; .$$

Helene [10] derived this result (not in this tidy form) using Bayesian statistics with uniform prior. We emphasize again that this prior is *not* the preferred prior in objective Bayes theory. Attempts have been made to put this formula on a frequentist footing, notably by Zech [11], who was criticized by Highland [12], with a reply by Zech. The issue has to do with the conditional probabilities. Highland showed that a standard conditional probability calculation does not lead to the Helene formula. It turns out that Zech's calculation refers to an ensemble which is known in a Monte Carlo simulation but which is unknowable in

[5] Some of these counter-examples assumed that some points could be excluded from the acceptance region while including some other points with the same LR. This issue was not explicitly addressed in Ref. [1], but Ref. [3] makes clear that all points which 'tie' for the ordering-LR cutoff should be included.

[6] Note added in proof: After the CLW, I realized that the Roe/Woodroofe conditioning was basically the same as that used by Zech [11], and differed from that of Highland [12]. See my writeup for the Fermilab CLW in March, 2000.

[7] In Bayesian theory, it can make sense to talk about the probability associated with a constant of Nature, since probability is defined as degree of belief, not frequency.

experimental data; Highland says, “It is difficult to see what physical experiment this would correspond to”.

As a method for computing upper limits, Helene’s formula *over*covers (more than required by Poisson discreteness) for the usual ensemble. However, as discussed in Section 9, the same uniform prior leads to *lower* limits which *under*cover a Poisson mean. There is no fundamental basis for this formula in the classical theory of confidence intervals, and, as noted above, the uniform prior is not the preferred objective Bayesian prior for a Poisson mean. Hence it is an *ad hoc* adaptation which gives upper limits that some people find to be reasonable.

A. L. Read [13] further discusses this formula and its generalization in one of the required reading articles for this workshop. I believe that one must still ask what is the fundamental basis (in the professional statistics literature) for this method? Does the Neyman–Pearson lemma, which Read cites as the reason his intervals are ‘optimal’, really imply his conclusion? Or is there a leap from event classification to these intervals, especially if flip-flopping is properly treated? In fact, Stuart and Ord [3] cite the same Neyman–Pearson lemma as the justification for the LR ordering principle used by FC.

In my opinion, confidence intervals with extra over-coverage must be justified on grounds of either robustness or conditioning[8].

4. NUISANCE PARAMETERS

Nuisance parameters are parameters such as the detector efficiency, integrated luminosity, mean background, etc., which are not known exactly but must be estimated, even though they are not the parameters of physics interest.

This is an area that could benefit from more work. If one strictly follows the traditional definition of confidence intervals, one must not under-cover for *any* value of the nuisance parameter. The resulting table of intervals typically causes over-coverage for any given value.

Historically, this was an even bigger problem because of the computing resources needed to check coverage for more than a few values of the nuisance parameters; even today, this is a challenge. Therefore, it has been the practice to obtain approximate intervals by covering for estimated values of the nuisance parameters instead of all values [3]. Nowadays, computing is more tractable, so one can check coverage for other values, but it is still typically impossible to obtain an exact solution when there are many nuisance parameters.

Again, a ‘problem’ arises when confidence intervals don’t always behave like P(hypothesis$|$ data). (Because they are *not* P(hypothesis$|$data)!) This occurs in a very simple, common prototype case, which Virgil Highland and I [14] wrote about some years ago: Suppose you see no events, and you have a 10% uncertainty in luminosity. How does the usual 90% C.L. upper limit on the Poisson mean (2.3 before the Unified Approach) change because of the luminosity uncertainty? Surprisingly, the true upper confidence limit is *more restrictive* than if luminosity is perfectly known![9] This seemed so ‘unacceptable’ that we resorted to a Bayesian-inspired technique, namely integrating out the nuisance parameter. This has no justification at all in Neyman’s construction; in fact, it causes over-coverage. Yet, it is very popular in HEP (and was already in use; see the references in Ref. [14]).

Thus, there is food for thought in this problem. It is disturbing that the classical method gives the ‘wrong’ sign to the effect. One of the lessons, however, is that the effect of a 10% uncertainty is quite small, so in many practical cases, this is not really an issue.

Nuisance parameters are straightforward to handle in Bayesian theory *except* it seems that *priors in high dimensions are potentially an issue.* As far as I know, this is not explored well yet in HEP. The professional statistics literature shows that high-dimension priors are not obvious.

[8]Note added in proof: At the time of the CERN CLW, I thought that Highland and Roe/Woodroofe handled conditioning similarly, which is not the case. See my writeup for the Fermilab CLW in March 2000.

[9]This can be demonstrated with a simple Monte Carlo program. Why it happens is briefly described in Ref. [15]

4.1 Related issue: systematic errors

In the frequentist approach, systematic errors are necessarily treated using the frequentist definition of P. This is sometimes conceptually hard to swallow, but doesn't seem to be a problem in practice. (The problem in practice, for both Bayesians and frequentists, is attaching any sensible uncertainty at all to certain theoretical calculations!)

5. PRIORS

For me, the issue is not really 'prior anxiety' [16]. I am perfectly comfortable with subjective priors. However, I do not think that they are the answer to the question of what to publish. To see this, consider some experiments in the field of rare K_L^0 decays, a field in which I worked for a number of years, and which provided the original motivation for my interest in the theory behind upper limit calculations. I have selected three frontier (in their day) experiments which reported results in *Physical Review Letters* regarding searches for, respectively, $K_L^0 \to \mu^+\mu^-$ [17], $K_L^0 \to \mu^\pm e^\mp$ [18], and $K_L^0 \to \pi^0\nu\bar{\nu}$ [19].

In each experiment, the experimenters observed no candidate signal events (after cuts deemed reasonable), and each team calculated its *Single-Event Sensitivity* (SES): that value of the decay branching ratio (BR) for which the experiment would have observed on average one signal event. The known uncertainties in the SES were negligible. So, how is the 90% C.L. upper limit on BR related to the SES? The classical answer (which the experiments in fact reported) is simple: BR$< 2.3\times$SES.

Recall that the subjective Bayes posterior pdf is the product of the prior pdf and the likelihood. The posterior pdf in these three cases depends very much on the experiment, since the priors were so different:

1. A search for $K_L^0 \to \mu^+\mu^-$ [17] had SES of 8×10^{-10}. I think a typical subjective prior pdf at the time very firmly put the believed BR at greater than about 48×10^{-10}. This was because $K_L^0 \to \gamma\gamma$ had been measured, and it was a very plausible QED calculation to add on $\gamma\gamma \to \mu^+\mu^-$, to obtain the so-called 'unitarity bound' on $K_L^0 \to \mu^+\mu^-$. When the experiment saw no events, this subjective prior was so strong that one could still believe, with 90% certainty, that the BR was greater than the unitarity bound, a factor of 6 greater than the SES!
2. A search for $K_L^0 \to \mu^\pm e^\mp$ [18] had SES of 1.4×10^{-11}. When the experiment began, the previous upper limits were several orders of magnitude higher, and there was some plausible beyond-the-Standard-Model speculation that $K_L^0 \to \mu^\pm e^\mp$ might exist within the sensitivity of the experiment. My personal subjective belief gave us a few per cent chance of a discovery. Thus, after seeing nothing, my degree of belief was changed significantly by the experiment, for values of BR above the SES.
3. A search for $K_L^0 \to \pi^0\nu\bar{\nu}$ [19] had SES of 2.5×10^{-5}. Although this was a new experimental range, the Standard Model prediction was many orders of lower ($10^{-10} - 10^{-11}$) and I knew of no plausible way to get a BR as high as the SES. Hence, after this experiment, I believed with 90% certainty that the BR was several orders of magnitude lower than the SES!

These examples show that subjective priors for real experiments can be very different, and that they are *not* uniform in obvious metrics. They really do represent degree-of-belief. Hence there is no 'typical' subjective prior which results in a 'typical' relationship between the SES and the posterior belief. This is why I do not see subjective Bayes statistics as a useful way to communicate experimental data, even though I think it is a good model of how we scientists update our beliefs.

For related reasons, I find objective priors to be not particularly useful, except as calculational tools to get answers whose properties can be studied and justified *post hoc* on other (even frequentist) grounds.

5.1 An under-appreciated advantage of Bayesian statistics

A very nice feature of Bayesian statistics is that it provides an appealing way to formulate a 'sharp hypothesis', one which gives special significance to a particular parameter value. For example, one can formulate a test on $x = 2$ versus $x \neq 2$ in a natural way, using a *subjective* prior with a Dirac δ-function (times a subjective factor) at $x = 2$ and the rest of the probability spread out (according to your degree-of-belief) over $x \neq 2$.

Ironically, this very nice feature of Bayesian statistics is typically lost in 'objective' priors. For me, it's another indication that proposed objective priors throw away too much of the essence of Bayesian statistics.

6. GOODNESS OF FIT

In my opinion, this is a little-known but critical issue for Bayesian statistics. In HEP, we frequently want to test the correctness of the functional form used in a fit.

We recall that the Bayesian posterior obeys the Likelihood Principle: all the information from the experiment is in $\mathcal{L}$ for *your* experiment. The frequentist ensemble does not exist. Therefore, in Bayesian statistics, our usual way of formulating goodness-of-fit does not exist!

As I understand Bayesian statistics, the model error must be incorporated into the prior and $\mathcal{L}$. This appears to be very difficult for the simple question, "Is my chosen functional form a good fit to the data?" In HEP, such issues are still at a very early stage of exploration; see Ref. [20] for an example combining discrepant data.

7. ROBUSTNESS

Robustness (relative insensitivity to departures from assumptions) is an important issue in HEP. The PDG knows that historically, systematic errors are often under-estimated. Therefore they use the famous scale factor S, which has demonstrated robustness at reasonable cost (F. James, private communication.)

Chanowitz [21] has proposed an analogy for confidence limits. I think it is worth investigating doing something similar in the Unified Approach. In the 'Karmen problem', for example, one could inflate the uncertainty on background until the goodness-of-fit (in this case, the probability of obtaining no events) is decent. Then put that uncertainty into the limit calculation.

Whatever one uses, of course, one must be clear on where robustness enters, since this is likely to be a contentious issue.

8. WHAT IS THE ENSEMBLE?

In an interesting chapter entitled "Comparative Statistical Inference", Stuart and Ord [3] note on p. 1227 that "Two of the difficulties facing the frequency approach in practice are the specification of the sample space and the need to ensure random sampling".

HEP is no exception. Specifying the ensemble (the sample space within which frequencies are calculated) has typically not not been a big practical problem in my experience, but it is easy to imagine cases where it can arise. For example, if your experiment sees 77 top-quark events[10], should the imagined ensemble contain experiments which also saw 77 events, or a larger ensemble in which this number undergoes Poisson fluctuations? The issue is once again *conditioning* (encountered in Section 2.3 in the context of the Roe–Woodroofe proposal for modifying the Unified Approach). For an example with references to some literature on the debate, see Ref. [22].

In discovery-oriented experiments, it seems that it will always be a problem (even in Bayesian statistics) to calculate probabilities starting from unusual events. There is the old story (I heard it at-

[10] Harrison Prosper offered this example from the D0 experiment in e-mail circulated before this conference.

tributed to Feynman) about the license plate: "That car's license plate number is GMZ356. Do you realize how unlikely that is?" I seem to recall that early $Z \to e^+e^-\gamma$ events in 1983 had similar issue: what's the ensemble? In practice, the data set analysed at the end of one run is often used to define the hypothesis for the next run. Real signals gain in significance with more running.

Certainly, we can agree that when one quotes coverage, one should define the ensemble used for the coverage calculation, if it is an issue. In many cases, for example NOMAD ν oscillations, this is not a subtlety, it seems.

9. NON-COVERAGE OF BAYESIAN INTERVALS

There is a long history of comparing Bayesian intervals with confidence intervals[11], since the issue of which to use is mitigated to the extent that the numerical answers are the same. Unfortunately, since each type is based on a different definition of P, the math does not ensure that intervals from one realm make sense in the other realm. We have seen above how confidence intervals can 'make no sense' when interpreted as degree of belief. Similarly, Bayesian intervals can 'make no sense' when interpreted according to the frequentist definition of confidence intervals.

As an example [15], let's suppose that one makes a 'measurement' of the mean μ of a Poisson process by performing a single experiment and obtaining $n = 3$ events. The classical central[12] 68% confidence interval for μ is

- (1.37, 5.92). [Central intervals with frequentist coverage.]

The Bayesian intervals depend of course on the prior. If one naively uses a prior uniform in μ, then the 68% credible interval is

- (2.09, 5.92). [Bayesian with prior uniform in μ.]

If one uses one of the 'objective' priors advocated for the Poisson mean by Jeffreys, $P(\mu) \propto 1/\mu$, then the 68% credible interval is

- (1.37, 4.64). [Bayesian with prior $1/\mu$.]

Such Bayesian intervals are shorter, and undercover the unknown true value. Note that the right endpoint of the Bayesian interval with uniform prior is the same as for the frequentist interval, while the left endpoint for the $1/\mu$ prior is the same as for the frequentist interval. This is always true for these priors in this Poisson problem, so that: *90% C.L. upper limits are the 'same' for classical and uniform prior, while 90% C.L. lower limits are the 'same' for classical and* $1/\mu$ *prior.* I am completely convinced that our community's infatuation with uniform prior for the Poisson mean is a consequence of the fact that we are normally interested in *upper* limits! If the nature of our work were such that *lower* limits on Poisson means were the norm, then the $1/\mu$ prior would be the 'obvious' one, and one would even enjoy the luxury of being more consistent with the objective Bayesian literature[13].

10. WHAT MIGHT WE AGREE ON? [14]

I have mentioned many areas in frequentist and Bayesian statistics where there are issues for debate. In spite of the wide range of points of view at this workshop, I think we can agree on a number of statements which are not controversial among people who have learned about them, but which are non-trivial in that I see papers which make assumptions to the contrary.

1. First of all, civility. The debates in the professional statistics community seem to have departed from civility more than one might hope. We physicists have our own role models, in particular Bohr and Einstein in their quantum mechanics debate, for handling serious disagreement.

[11] Ref. [15] has some references.

[12] This assumes that flip-flopping is not an issue. The unified confidence interval is (1.10, 5.30) [1].

[13] However, a prior for μ based on more general 'objective' arguments is yet another one, $P(\mu) \propto 1/\sqrt{\mu}$.

[14] I revised this section slightly after my talk, so that I think it now really does have agreement from a number of knowledgeable speakers with whom I discussed it.

2. The likelihood function $\mathcal{L}$ is not a pdf in the unknown parameters. 'Integrating the likelihood function' is not a concept in either Bayesian or classical statistics: it is not well-defined because one gets a different answer upon reparametrizing the integration variable. (Since $\mathcal{L}$ is not a pdf in the parameters, there is no Jacobian in them to give consistent integrals upon change-of-variable.)
3. Answers based on integrating the posterior pdf with a 'uniform prior' depend on the metric in which the prior is uniform. Uniform priors should be explicitly stated, not hidden.
4. Bayesian intervals typically do not have frequentist coverage. This is not surprising, since the Bayesian formulation makes no reference to an ensemble: it uses the likelihood function for the particular data set observed.
5. Publishing enough information to reconstruct an approximate likelihood function should be strongly encouraged. This allows one to specify one's own prior and calculate a posterior pdf, and it allows approximate classical confidence intervals to be computed.
6. Our usual goodness-of-fit tests do not exist in Bayesian statistics. The Bayesian analog requires a reformulation which extends the space of P to include functional forms.
7. P(hypothesis|data) cannot be calculated without a prior.
8. The confidence interval construction does not use a prior. It uses P(data|theory), and requires the ensemble to be specified. Priors enter when going from P(data|theory) to P(theory|data), which confidence intervals do not do.
9. Regardless of your opinion about priors, a subjective utility function is needed to make a decision, so any argument for totally objective decisions is highly suspicious.
10. If a method is frequentist, one must understand the frequentist coverage. If the coverage differs materially from the stated C.L., then an explanation should be provided. If a method is Bayesian, then it can also be enlightening to look at the frequentist coverage, if only to educate ourselves about the difference between degree-of-belief P and frequentist P!
11. If one uses a method which implicitly or explicitly invokes a prior, then one should understand the sensitivity of the result to the choice of prior.
12. When used without a qualifier, the words 'confidence interval' imply the frequentist definition of P, and at least approximate coverage at the stated C.L. Intervals not having this property should be qualified or called something else: for Bayesian intervals, some prefer 'Bayesian confidence intervals', while others prefer 'credible intervals'.

. . . and remember: All this is irrelevant if you tune the cuts in order to eliminate candidate events in order to get a better limit (unless, of course you are willing to put that tuning into your coverage calculation . . . that leads to loss of power, even if it covers, however).

11. CONCLUSION

In this talk I have tried to highlight some of the most troublesome areas in classical and Bayesian statistics calculations. The LR ordering for confidence intervals, possibly with conditioning added, provides a well-founded general framework for a consistent treatment of one-sided and two-sided intervals. The Bayesian method most closely associated with scientific reasoning, namely using a subjective prior, is hard to imagine as the answer to "What number to publish?" Progress will be made if people use methods with understood properties, and if statements about P(data|parameter) are not interpreted as statements about P(parameter|data).

Acknowledgements

I give great thanks to the co-convenors of this workshop, Fred James and Louis Lyons, for bringing this group of people together for the first time. The local organizers, Fred James and Yves Perrin, provided a perfect environment for the talks, meals, and discussions. This talk reflects many years of off-and-on study with students and colleagues too numerous to list, but most notably Virgil Highland[15], Fred James, and Gary Feldman. This work is supported by UCLA and the U.S. Dept. of Energy.

References

[1] G.J. Feldman and R.D. Cousins, Phys. Rev. **D57** (1998) 3873.

[2] S. Ciampolillo, Il Nuovo Cimento **111** (1998) 1415.

[3] A. Stuart and J.K. Ord, *Kendall's Advanced Theory of Statistics*, Vol. 2, *Classical Inference and Relationship*, 5th Ed. (Oxford University Press, New York, 1991); see also earlier editions by Kendall and Stuart. The LR-ordering principle, including approximate treatment of nuisance parameters, is given at the beginning of Chapter 23 (Chapter 24 in the previous edition).

[4] P. Astier, *et al.*, Phys. Lett. **B453** (1999) 169. Many different decay modes with different proportions of background (each with errors) are combined.

[5] G. Zech, physics/9809035. See also talk at this workshop.

[6] G. Punzi, hep-ex/9912048. See also talk at this workshop.

[7] J. Bouchez, hep-ex/0001036.

[8] B.P. Roe and M.B. Woodroofe, Phys. Rev. **D60** (1999) 053009. See also talk by Woodroofe at this workshop.

[9] R. Cousins, physics/0001031.

[10] O. Helene, Nucl. Instrum. Methods **212** (1983) 319.

[11] G. Zech, Nucl. Instrum. Methods **A277** (1989) 608.

[12] V. Highland, Nucl. Instrum. Methods **A398** (1997) 429, followed by reply by G. Zech.

[13] A.L. Read, DELPHI 97-158 PHYS 737, 29 October 1997,
http://wwwinfo.cern.ch/∼pubxx/www/delsec/delnote/public/97_158_phys_737.ps.gz
See also talk at this workshop.

[14] R. Cousins and V. Highland, Nucl. Instrum. Methods **A320** (1992) 331.

[15] R. Cousins, Am. J. Phys. **63** (1995) 398.

[16] G. D'Agostini, physics/9906048.

[17] A.R. Clark, *et al.*, Phys. Rev. Lett. **26** (1971) 1667.

[18] K. Arisaka, *et al.*, Phys. Rev. Lett. **70** (1993) 1049.

[19] M. Weaver, *et al.*, Phys. Rev. Lett. **72** (1994) 3758.

[20] G. D'Agostini, hep-ex/9910036.

[15] In a widely-read preprint [23] never submitted for publication, Highland gave a critical survey of upper limits methods in 1986.

[21] M. Chanowitz, Phys. Rev. **D59** (1999) 073005.

[22] R. Cousins, Nucl. Instrum. Methods **A417** (1998) 391.

[23] V. Highland, Temple Univ. preprint COO-3539-38 (1986).

Discussion after talk of Bob Cousins. Chairman: Jim Linnemann.

Michael Woodroofe

Again I have really more comments than questions, the first of which is to reinforce the call for civility. I have experienced what the lack of civility can lead to, and you don't want to go there.

R. Cousins

I might add that this is particularly important to us because we are amateurs in statistics, so we are going to make mistakes. We are physicists in our 'day jobs', so when we do statistics let's be kind to each other.

M. Woodroofe

About the reluctance to publish subjective distributions. In the derivation of the Bayesian theory there is an assumption that the person who is writing down the priors is also the person who is incurring the losses or the gains in the utility function. Now that's true in some situations. If you're trying to decide 'what I should do with my life in the next two years, which experiment I should pursue', that's a personal decision and it's true. In other parts of science I think it may not be true. If you're sitting on a panel that's trying to decide which of several different experiments should be funded, you're not paying the losses for that, and I think that's related to the reluctance to publish subjective distributions.

The goodness-of-fit problem for Bayesians is very hard. A simple goodness-of-fit problem is to test whether data is normal, and that problem was solved recently from a Bayesian perspective by Jim Berger. It's a very clever solution, it's a nice solution, it's not easy. It was 1999 when that very basic problem was first worked out, and that's how hard it is.

Harrison Prosper

This flip-flop problem that you solved. Was the problem the fact that people are flip-flopping or was the problem the fact that the ensemble in which this flip-flopping was embedded didn't cover? I can imagine for example, designing an algorithm for limits which allows the experimenter to choose to flip-flop which also covers.

R. Cousins

You could do that, but people certainly were not doing that. You can even imagine an extreme case where you adjust your cuts specifically to get rid of all the candidate signal events you see, and then you do a Monte Carlo of such an ensemble of experiments to see what upper limit should be quoted in order to have coverage. The resulting upper limits are valid in the sense of correct coverage, but have very poor power; in fact the mean upper limit is infinity, as I once mentioned in a *NIM* paper devoted to something better (*NIM* **A337** (1994) 557).

H. Prosper

But the point is that in Neyman's initial paper, he puts no restrictions whatsoever on the ensemble, he simply states "this is what we should satisfy", and so in principle we have complete freedom.

R. Cousins

That's right, and what's happened since Neyman as I understand it, is this business of conditioning. You know, we lump Fisher and Neyman as classical buddies together opposing the Bayesians, but in fact they were at each other's throats because Fisher for instance insisted on conditioning and figuring out what the ensemble is. That's why I quoted Kendall and Stuart. This is a problem you've got to worry about where it matters, and if you get different results depending on what you use for it, I think you should say that too.

H. Prosper

Just one last comment. In the same volume in which Kendall and Stuart described this likelihood ratio test, they also point out that getting rid of the dependence on nuisance parameters is a very difficult problem, so I think even for the case of the likelihood ratio, the calculation of that ratio still depends on those parameters, if the data set size is too small.

R. Cousins

That's right. The advantage we have today is much more computational power, although it can still be insufficient for an exact calculation. Kendall and Stuart make the approximation that you calculate coverage only for values of nuisance parameters equal to their maximum likelihood estimates. With today's computers, one can check coverage for other values of the nuisance parameters, although it is still not practical to do the construction directly in a high-dimensional space.