

TREATMENT OF NUISANCE PARAMETERS IN HIGH ENERGY PHYSICS, AND POSSIBLE JUSTIFICATIONS AND IMPROVEMENTS IN THE STATISTICS LITERATURE

ROBERT D. COUSINS

Dept. of Physics and Astronomy, University of California, Los Angeles, CA 90095, USA

E-mail: cousins@physics.ucla.edu

Nuisance parameters are common in high energy physics (HEP), and various methods are used for incorporating their effects into measurements of physics interest. A survey of some of the professional statistics literature provides justification for and insight into most of the methods commonly used in HEP. There are extensions and refinements whose usefulness could still be explored, especially in higher-dimensional problems.

Keywords: nuisance parameters; systematic uncertainties

1. Introduction

Nuisance parameters appear in virtually every physics measurement of interest because the measuring apparatus must be calibrated, and for all but the simplest apparatus, the calibration technique involves unknowns that are not directly of physical interest. In high energy physics (HEP), the primary measurement nearly always involves counting particle physics interactions of interest known as “events”. The numbers of events with various characteristics are used to make inferences about underlying processes, typically Poisson. For example, suppose that n events from a Poisson interaction process are observed during a measured time interval t , and one wishes to make inferences about the mean interaction rate per unit time, Γ . By a variety of methods discussed in the next section, an interval pertaining to the unknown Poisson mean μ from which n is sampled can be constructed. If the time interval t is known with negligible uncertainty, then an interval pertaining to Γ is obtained by dividing the endpoints of the interval for μ by t .

If the measurement of t itself has non-negligible uncertainty, then t (or some surrogate) becomes a nuisance parameter, and the question arises as to how to incorporate the uncertainty in t into the interval for Γ . Already in this simple example, there is much food for thought. When one adds the common complication of “background” events that mimic the “signal” events of interest, any uncertainty in the mean rate of background events adds another nuisance parameter, and the possibilities proliferate further.

The uncertainties in nuisance parameters often correspond to what we call “systematic uncertainties” in HEP. At the last PhyStat, Sinervo⁵² presented a more careful discussion of this correspondence while advocating a more precise set of definitions of three classes of systematic errors.

In this paper, I survey some representative literature from both the high energy physics and professional statistical communities, and compare and contrast the respective approaches for dealing with nuisance parameters. The ease with which one can follow citations and download papers on the Web resulted in a collection that has a lot of stimulating articles. But while I have taken at least a cursory look at all papers cited and have read a number of them, my study was tightly constrained by the decreasing time I find available to devote to my statistics hobby. Therefore, much of this paper is an annotated bibliography, and I hope that others will be able to pursue these leads.

For context and definiteness, I use the construction of an interval used to characterize the uncertainty in a single unknown parameter of interest. Such intervals are nearly always quoted in experimental HEP papers. (On the other hand, to go beyond intervals, e.g., to explicit decision theory, is rarely if ever done in a formal manner in HEP publications.) I emphasize to statisticians that physicists do not interpret confidence intervals rigidly according to the caricature of “rejecting” or “accepting” the hypothesis, but generally find confidence intervals useful as a way of conveying the results of experiments.

Table 1. 68% C.L. intervals for the mean μ of a Poisson distribution, based on the single observation $n_0 = 3$, calculated by various methods. Only the frequentist intervals avoid under-coverage for all values of μ . The boldface numbers highlight the fact that the frequentist central interval shares the right endpoint with the Bayesian interval with uniform prior, and the left endpoint with the Bayesian interval with $1/\mu$ prior, explaining why neither set of Bayesian intervals covers for all values of μ .

Method	Prior	Interval	Length
rms deviation	–	(1.27, 4.73)	3.46
Bayesian central	1	(2.09, 5.92)	3.83
Bayesian shortest	1	(1.55, 5.15)	3.60
Bayesian central	$1/\mu$	(1.37 , 4.64)	3.27
Bayesian shortest	$1/\mu$	(0.86, 3.85)	2.99
Likelihood ratio	–	(1.58, 5.08)	3.50
Frequentist central	–	(1.37 , 5.92)	4.55
Frequentist shortest	–	(1.29, 5.25)	3.96
Frequentist LR ordering	–	(1.10, 5.30)	4.20

In Sec. 2, I start with a simple problem with no nuisance parameters in order to foreshadow the proliferation of methods and illustrate the first correspondence between frequentist and Bayesian methods. In Sec. 3, I discuss the role of conditioning, which is a concept that I believe deserves more awareness within HEP. In Secs. 4, 5 and 6, I describe methods for incorporating nuisance parameters in, respectively, Bayesian credible intervals, likelihood intervals, and explicitly constructed confidence intervals. I conclude in Sec. 7 with some recommendations given existing tools, and some areas yet to be explored.

2. Intervals for a Poisson mean

To set the stage, we first recall some ways to construct an interval corresponding to a confidence level (or analog) of 68.27% for an unknown Poisson mean μ after a single observation of n events. We take $n = 3$ for definiteness. Table 1, taken from Refs. 21 and 25, gives intervals that we can identify as:

- (1) *estimate of mean and rms deviation*: $n \pm \sqrt{n}$. This is just a crude estimate at small n , and I do not consider it further.
- (2) *credible intervals* constructed by assigning the indicated prior $P(\mu)$ and constructing a Bayesian credible interval, using an auxiliary condition as noted.
- (3) *likelihood intervals* constructed from likelihood ratios, with no integration or other reference to a metric on μ .
- (4) *confidence intervals* constructed from Neyman's construction, with auxiliary conditions specified

as noted.

The frequentist coverage probability of these types of intervals as a function of μ can be easily studied; only the confidence intervals give exact or higher coverage for all values of μ . (Ref. 62 examines the coverage of likelihood intervals.) As noted in Ref. 21, traditionally high energy physicists are rather strict about coverage, in contrast to the attitude expressed by statisticians in Refs. 42 and 63.

By far the most common Bayesian prior for a Poisson mean in HEP is the uniform prior, for which the right endpoint of a central credible interval coincides with that of a frequentist central confidence interval; this makes upper limits identical. On the other hand, left endpoints, and hence *lower* limits, are identical for the $1/\mu$ prior that actually has some motivation in terms of scale invariance. It was advocated by Jeffreys, although the rule for “Jeffreys’ Priors” yields the prior $1/\sqrt{\mu}$. See Refs. 21 and Reid’s respondent’s talk⁷² for further discussion. Reid draws attention to the $1/\sqrt{\mu}$ prior as the more fundamental “matching prior”, and regards the exact matching of the endpoints in Table 1 as essentially an artifact of discreteness.

To such a diverse set of starting points, we add a variety of techniques for coping with nuisance parameters. The recent professional statistics literature seems to be mainly concerned with likelihood and Bayesian methods, while at least some of us in HEP are still interested in confidence intervals that give correct coverage by construction. The various points of view inform each other. The further confidence intervals stray from conditioning or its extreme, the likelihood principle, the more susceptible they are to being deemed irrelevant to the data set at hand; likewise, credible intervals with poor frequentist behavior place the prior under increased scrutiny. In both HEP and some professional statistics literature, performing a Bayesian-style integration over nuisance parameters in an otherwise non-Bayesian method is considered a reasonable thing to try; I think that the ultimate justification (or lack thereof) comes from studying the frequentist properties of the results (although this interpretation can be problematic for some uncertainties). This is the point of view taken by Linnemann in his interesting study⁵¹ of various measures of significance at the previous PhyStat.

3. The Role of Conditioning (or Absence thereof) in HEP

In HEP, it is common to calculate coverage probabilities for confidence intervals (or Type I and II error probabilities) by Monte Carlo simulation using an ensemble of pseudo-experiments that includes all possible data sets that might be obtained according to the experimental procedure. Since our usual procedure is to take data for an amount of “live time” that is well defined (though usually not exactly specified in advance), the number of events obtained in each pseudo-experiment fluctuates according to a Poisson distribution. Consider, however, a situation in which the intrinsic *uncertainty* on the measurement of a parameter θ depends on the total number of events, but in which the number of events itself carries no information about θ . One can argue that the result of a particular experiment should be a confidence interval in which the ensemble used to calculate coverage should consist of pseudo-experiments that all have the same number of events as was actually observed. The argument goes back to Fisher and *conditioning* on an ancillary statistic.

As reviewed by Reid²⁰ and references therein (including notable work by Cox, also speaking at this conference), conditioning on some aspect of the data actually observed has a variety of justifications, including elimination of nuisance parameters. In HEP, conscious conditioning seems to be considered only rarely. To the extent that Bayesian-inspired techniques observe the likelihood principle (as in the case for pure subjective Bayesians), the extreme of conditioning on the actual set of data observed is built in, but I do not know how widespread this is recognized in HEP. Although I have attempted to read some fraction of the vast statistical literature on this topic (including the reviews by Reid in 1995²⁰, by Fraser in 2004⁶⁴, and the discussion in Ref. 30) I still find myself in the state of “a little knowledge is a dangerous thing”. Therefore I will confine my remarks to examples of personal interest, and some pointers to the literature; see also Sec. 5 below. Demortier⁵⁰, another high energy physicist, gave his perspective at the last PhysStat.

3.1. Ratio of Poisson Means

In an example from HEP, an experiment observes x events of one type from Poisson X with unknown

mean μ , and observes y events of another type from (independent) Poisson Y with unknown mean ν . Suppose the physics of interest is in the *ratio* of Poisson means, the single parameter $\lambda = \mu/\nu$. Then either of the individual means, or the sum, can be taken as a nuisance parameter, and we wish to obtain a confidence interval for λ from the data (x, y) in the presence of unknown nuisance parameter. The product of Poisson probabilities can be rewritten as the product of a single Poisson probability with mean $\tau = \mu + \nu$ for the total number of events $Z = X + Y$, and the binomial probability that this total is divided as such with the binomial parameter $\rho = \lambda/(1 + \lambda)$:

$$\begin{aligned} P(x, y) &= \left(\frac{e^{-\mu} \mu^x}{x!} \right) \times \left(\frac{e^{-\nu} \nu^y}{y!} \right) \\ &= \left(\frac{e^{-(\mu+\nu)} (\mu + \nu)^z}{z!} \right) \\ &\quad \times \left(\frac{z!}{x!(z-x)!} \rho^x (1-\rho)^{(z-x)} \right). \end{aligned} \quad (1)$$

That is, rewriting in terms of observables (X, Z) and parameters (λ, τ) :

$$P(x, y; \mu, \nu) = P(z; \mu + \nu) P(x|z; \rho) \quad (2)$$

$$\begin{aligned} P(x, z-x; \lambda\tau/(\lambda+1), \tau/(\lambda+1)) \\ = P(z; \tau) P(x|z; \lambda/(1+\lambda)). \end{aligned} \quad (3)$$

In this form, all the information about λ is in the *conditional* binomial probability for the observed “successes” x , *given* the observed total number of events z . In the words of Reid²⁰, “...it is intuitively obvious that there is no information on the ratio of rates from the total count...”. The same conclusion was reached in our community by James and Roos⁵. Therefore one simply uses x and z to look up a standard confidence interval for ρ , and rewrites it in terms of λ .

3.1.1. Inference about the Total Mean:

Marginalization

Suppose that the parameter of interest and the nuisance parameter are reversed: one desires inference about sum of means $\tau = \mu + \nu$, and the ratio λ is the nuisance parameter! As discussed by Reid²⁰, it is no longer conditioning that is appropriate, but rather *marginalization*, i.e., integrating over a sub-space of the *sample* space. This can be seen from Eq. 2; if we sum over observed x , then the inference on τ is made from the resulting Poisson $P(z; \tau)$.

Thus, this example illustrates the use of both conditioning and marginalization. Both these concepts return repeatedly in modifications to the profile likelihood discussed in Sec. 5. I find it hard to understand, however, how one would be able to develop a general algorithm based on one concept or the other, when this simple example alternates between concepts depending on the parameter of interest.

3.1.2. *Epilogue on the Ratio of Poisson Means*

Many years ago while teaching a seminar on data analysis, I studied the coverage of the confidence intervals in Ref. 5, and found that they not only typically over-covered (as do confidence intervals for a Poisson mean), but that they *always* over-covered by a finite amount! There were *no* combinations of μ and ν for which the set of confidence intervals had coverage even close to the nominal confidence level. This convinced me that there must exist proper subsets of the James/Roos intervals that still covered. A literature search revealed that ratio-of-Poisson-means intervals were derived in an astounding variety of contexts, but that everyone obtained the same intervals, and there was even a theorem by Lehmann and Scheffé to justify the intuitive use of the above factorization. Nonetheless, after playing around with Neyman-like constructions, I found some “improved” intervals, and wrote up the story with all the references²⁶. It was clear that the discreteness of the problem evaded the theorem (as Lehmann had warned).

A problem with some aspects in common (2×2 contingency tables) has been argued about for over 50 years in the statistics literature, with most people coming down on the side of enforcing strict conditioning. Whether or not my intervals (which still over-cover and are shorter by any metric since they are proper subsets of the standard ones) are “improved” or not is a matter of some debate. I tend to conclude that using the statistical fluctuations in the total number of events is a natural and effective way to average out the discreteness, especially in light of the willingness of statisticians to average over discreteness in what seems to me to be a more arbitrary way^{42, 63}. I come back to the construction I used in Sec. 6 below.

3.2. *Non-standard conditioning in HEP on the observed constraint on the number of background events*

In HEP, it has become common in one context to use non-standard conditioning that, as far as I know, has no foundation in the statistics literature. While not requiring a nuisance parameter, I mention it here for completeness, and because the generalization common in HEP does have a nuisance parameter. X and Y are random Poisson variables for (experimentally indistinguishable) signal and background, respectively, and one observes $z = x + y$ from the sum $Z = X + Y$. The mean b of the background Y is known, and one desires a confidence interval on the unknown mean μ of X . This problem has a long history including the paper by Feldman and myself²⁵ that constructs frequentist confidence intervals using the likelihood-ratio ordering in Ref. 30. These intervals cover by construction for the ensemble of all experiments, but they have been criticized for badly violating the likelihood principle³⁴. The most blatant case is when $z = 0$ is observed, in which case one *knows* that for the experiment at hand, there are no background events ($y = 0$). In general, whenever z is observed, one knows that $y \leq z$.

In 1989, Zech¹² calculated upper limits on μ by calculating probabilities conditioned on $y \leq z$; this has been commonly used and extended in other contexts^{22, 29}. (For further perspective on the evolving point of view of Zech on this and other methods, see Ref. 47.) This conditioning on an inequality was proposed independently in 1999 in a modification to Ref. 25 by Roe and Woodroffe (RW)²⁸. However, Zech’s original paper was criticized by Highland²³, and RW was criticized by me³⁹. (Subsequently RW advocated a different technique⁴¹.) That Zech and RW were using the same conditioning escaped me for some time, but I have explained it in detail in Ref. 37, along with Highland’s objections, with which I tend to agree. The conditioning has properties that some find desirable, in particular for upper limits. But for two-sided intervals it leads to a situation in which the intervals cover for the restricted ensemble but not for the unconditional ensemble³⁹.

Read, one of the advocates of a generalized version of this conditioning²², recommended using it for upper limits⁴⁶, and using Ref. 25 when there is a clear signal and there is no issue of interpretation.

It is notable that while most of the vast literature on conditioning seems not to have found its way

into HEP, a non-standard way with no apparent formal justification was invented in HEP and gained a large following in HEP. Given the shaky foundation, caution should be used in any new application.

3.3. *Conditioning in Comparing Simple Hypotheses*

Berger et al.¹⁸ showed that for testing a simple hypothesis against a simple alternative, the Bayesian posterior for equal prior probabilities has a nice frequentist interpretation in terms of error probabilities conditioned on the value of the likelihood ratio statistic actually observed in the data. Dass and Berger⁵⁸ generalized this to certain composite hypotheses. Neither paper seems to be cited much in the statistical literature (except by Berger himself), and a recent review in Ref. 57 is accompanied by spirited and on the whole rather unsympathetic commentary from statisticians.

I actually found Ref. 18 to be somewhat appealing (in the admittedly rare special cases in which we have simple hypotheses), and Ref. 58 to be intriguing. Given the apparent usefulness of conditioning, and the apparent difficulties of conditioning in many of our frequentist techniques in HEP, it would be interesting to see if Berger's point of view could provide some useful inspiration. I note, however, that Reid, the respondent to the present paper, cautions me that part of what I find attractive depends on a certain type of "flat" prior and so may not have good properties in general.

4. Nuisance Parameters in Bayesian Intervals

In the Bayesian world, all the difficulties with nuisance parameters are pushed (where else?) into the prior pdfs for the nuisance parameters. It could be that HEP, with its nearly universal usage of uniform priors, has something substantial to gain from the professional literature, in particular by investigating the so-called reference priors of Bernardo and collaborators¹⁴.

As Bayesians are fond of pointing out, once the priors are specified, turning the crank is intuitive and straightforward: one constructs the posterior pdf as usual and integrates out the nuisance parameters to obtain the marginal posterior pdf for the unknown parameter of interest, and proceeds as from there as

if there had been no nuisance parameters.

Liseo¹⁷ compares a Bayesian analysis based on reference priors (Berger and Bernardo) with the profile likelihood and its modifications (Sec. 5 below), and concludes that "the frequentist coverage properties of the credible sets derived from the reference priors are shown to be better than those computed from the likelihood approach." (A more extensive update is in Ref. 67.) In the Response to the present paper, Reid informs us that "Liseo's comparison of Bayesian analysis methods is somewhat misleading... as it does not use the more accepted likelihood approach...", with reference to her article on this topic.

Berger, Liseo, and Wolpert²⁷ review integrating out nuisance parameters from a point of view somewhat detached from the Bayesian motivation, simply studying the performance and practical issues. Their point of view is unambiguous: in response to a suggestion in the discussion that profile likelihoods be compared to integrated likelihoods as a form of sensitivity analysis, the authors respond that it might provide some assurance if they agree, but if they disagree badly the authors would "simply suspect that it is a situation with a 'bad' profile likelihood."

As advocated by Prosper^{10, 24}, the D0 experiment at Fermilab⁷⁸ has been using Bayesian methods for some time, integrating out the nuisance parameters. This practice has now spread to other collaborations. The statistics committee of the CDF⁷⁹ collaboration at Fermilab has performed a study⁵⁹ of Bayesian elimination of nuisance parameters in upper limit calculations; the associated software is available. Conway, a member of this committee, has separately released a program⁷⁰ for combining different experiments, including correlations. Demortier, another member of this committee, has separately studied^{45, 60} Bayesian techniques, and gives quite an interesting discussion of the prior pdf and the dangers of improper priors, and his recommended solution. At this conference, he has given a nice overview of reference priors, with much food for thought⁷⁴. Also at this conference, Heinrich⁷⁶ has presented an important study of the dangers of uniform priors for multiple background processes.

D'Agostini³⁵ has also been forcefully advocating a Bayesian approach for some time, with less emphasis on frequentist properties.

5. Nuisance Parameters in Likelihood Intervals

A widely used and appreciated parameter-fitting package in high energy physics is MINUIT², written and maintained for several decades by CERN physicist James. The MINUIT manual and the accompanying published paper⁶ describe its method of MINOS for obtaining confidence intervals and regions from likelihood ratios (increments in the negative log-likelihood). It uses Wilks's theorem¹ as applied to the profile likelihood, although until recently⁴⁰, the name profile likelihood was used rarely in HEP. The profile likelihood maximizes the likelihood over the nuisance parameters, separately for each value of the parameter(s) of interest.

Rolke and Lopez⁴⁰ have studied in detail the method of the profile likelihood as applied to the Poisson signal plus background problem in which the background is determined (with some uncertainty). I believe there was some confusion regarding the relationship of this work to MINUIT, that has now been resolved. The paper begins with the formalism of the likelihood ratio test as in Refs. 30, 25, but implements a rather conventional profile likelihood as in the method of MINOS, with an additional patch to improve the performance. Rolke, Lopez, and Conrad⁶⁸ have further studied the performance of the profile likelihood in some of HEP's prototype problems, with encouraging results.

Since MINUIT was first written, there has been quite a bit of study in the professional statistics community of cases in which the simple profile likelihood runs into difficulties, and of ways to overcome them. I am not aware of any of this research being applied routinely in HEP. Already in 1970, Kalbfleisch and Sprott³ surveyed a variety of methods for eliminating parameters from the likelihood function: integrated likelihoods, maximum relative likelihoods, marginal likelihoods, and conditional likelihoods. (The accompanying discussion by a number of luminaries of the day includes this gem from A.W.F. Edwards: "Let me say at once that I can see no reason why it should always be possible to eliminate nuisance parameters. Indeed, one of the many objections to Bayesian inference is that it always permits this elimination.") In 1977, Basu⁴ presented an even longer list and reviewed in detail the marginalizing and conditioning

methods, and worked on a proper definition of nuisance parameter including the Bayesian view.

Barndorff-Nielson, in 1983⁷ and 1986⁸, seems to have triggered a renewed look at the problem from the point of view of speed of asymptotic convergence by studying a "modified profile likelihood" "...with, generally, better inferential properties than the ordinary profile likelihood", and related concepts. He constructed approximate confidence intervals for the parameter of interest that are correct to order $O(n^{-3/2})$.

In 1987, Cox and Reid⁹ proposed transforming the nuisance parameters into a set that is (at least locally) orthogonal to the parameters of interest, in the sense that off-diagonal elements of the information matrix vanish. Then the idea is to condition on the observed values of the nuisance parameters. The result is a formula similar to that of Barndorff-Nielson but able to neglect a term due to the orthogonalization (although thereby losing parameterization invariance). In the discussion, G.A. Barnard also takes the point of view (as did Edwards above) that one should not eliminate nuisance parameters "if the data do not permit it." Given that these methods are quite complex, for me the most interesting question was posed by F. Critchley: "Which values of n are sufficiently sub-asymptotic to make the more elaborate procedures worthwhile and yet sufficiently large to retain enough accuracy in the crucial approximation on which rests the key advantage of parameter orthogonality?" The answer to this question affects whether or not it is worth it to us in HEP to attempt to implement something like this in MINUIT, for example. My concern is that, for very small n that we frequently have in HEP, the asymptotic advantages are not yet apparent.

In the ensuing years, Fraser and Reid¹¹ added additional commentary; McCullagh and Tibshirani¹³ proposed yet another "adjustment" to the profile likelihood; and Cox and Reid¹⁵ added further clarification regarding when the "modifications" gives a real improvement over the vanilla profile likelihood. Severini³¹ also discusses the relationships among the various modified likelihoods and Bayesian methods. At the last PhysStat, Reid and Fraser⁴⁹ provided a useful introduction for non-statisticians, with detailed explanations of examples relevant to HEP.

6. Nuisance Parameters in Frequentist Neyman-like Construction of Confidence Intervals

Traditionally many high energy physicists, including myself, have found confidence intervals to be appealing because probability P is defined in a way we understand and can simulate, and because Neyman taught us how to construct intervals that have the stated coverage (or greater) by construction. There is indeed the issue of educating people that confidence intervals are not the answer to the subjective questions that people want answered, e.g., “How much should I believe the hot new theory given the data in hand, and should I change what I do when I get up in the morning?” I remain optimistic³³ that we can teach people in HEP that $P(\text{data}|\text{theory})$ differs from $P(\text{theory}|\text{data})$, and that decisions require further subjective input about risk tolerance.

In HEP, central confidence intervals and upper confidence limits were for a long time the norm, with the choice of which one to use typically based on the data. It was only in the last decade or so that it became common knowledge in HEP that confidence intervals in general correspond to inverting a hypothesis tests on a parameter, and that the likelihood ratio test is an obvious default test to invert³⁰. The application to prototype cases of interest in the absence of nuisance parameters was worked out by Feldman and myself in 1997–98²⁵, and then we investigated the extension to nuisance parameters, guided by the terse prescription (for an approximate method) in Ref. 30. Except for Feldman’s talk at the Fermilab CLW³⁶, neither this work nor some follow-up work by Feldman has been written up. The initial delay was caused by the realization that one obtains a different answer in the limit the uncertainty on the nuisance parameter goes to zero than that obtained in the absence of a nuisance parameter; this is due to inserting a continuous variable into a discrete problem. Feldman described a patch for this in Ref. 36. As discussed below, for large n where this patch is irrelevant, others have continued and extended this approach.

Fraser, Reid, and Wong⁶¹ argue that the whole approach of confidence intervals is decision-theoretic, and that likelihood-based inference, with ranges of p -values, is a preferred option.

6.1. Full Multi-dimensional Neyman construction

In principle, a brute-force technique is to consider a fine grid in the entire multi-dimensional parameter space, including nuisance parameters, and for each grid point construct an acceptance region of the desired confidence level in the data space. For this one needs an algorithm for ordering the data. Then, for a particular value of the parameter of interest, one takes the union of all the acceptance regions for that value and all values of the nuisance parameters, and proceeds to find confidence regions as usual. This typically leads to confidence intervals or regions that badly over-cover for any particular set of true values of the parameters, in order to cover for all sets.

In practice, I am aware of only a few cases in HEP where this has been attempted^{26, 53, 44}. In the ratio of Poisson means problem described above²⁶, I played around with the ordering and managed to build acceptance regions that were subsets of the standard acceptance regions based on conditioning. But this is a tough (although fun) game that becomes increasingly harder as the number of nuisance parameters increases. I think that practically speaking, using an approximate method and checking the coverage is generally more productive than using the brute-force construction (in which case one will still probably want to check the coverage, to see how badly it over-covers).

In the previous PhyStat, Cranmer⁴⁸ presented a full construction for dealing with the background uncertainty in frequentist hypothesis testing, similar in concept to that in Ref. 26, but using the full generalization of the likelihood ratio ordering in Refs. 25, 30. At this conference⁷³, he compares this method with other methods. This is important work that should be “required reading” for those working on these issues at CERN’s Large Hadron Collider and elsewhere. Related work by Punzi⁷⁷ adds further valuable insight into the full Neyman construction.

6.2. Integrating nuisance Parameters

While participating in a number of experiments looking for “new physics” that we did not find, I encountered the simplest example of the problem discussed in the introduction, namely no event found ($n = 0$), and thus needing an upper confidence limit on Γ in the presence of uncertainty in t . (The symbols Γ and t are typically replaced by more general symbols such

as those for cross section and luminosity.) In 1990, it was common either to ignore the uncertainty in t or to adjust the upper limit on Γ by adjusting \hat{t} by some factor times σ_t .

Using intuition that would make a Bayesian smile, Highland and I averaged upper limits over the pdf for t centered on the measured \hat{t} and obtained well-behaved results¹⁶. F. James explained to us that this was Bayesian averaging grafted on to a frequentist upper limit, but we stayed with it, since a purely frequentist solution had behavior that seemed unlikely to be accepted^{16, 21}. Indeed, such intuitive averaging had already been used in the CDF experiment and elsewhere¹⁶. The important qualitative result was that, for uncertainties of 10% or so in t that were common in that day, the practice of ignoring the uncertainty was a better approximation than adjusting the upper limit by 10% or more. A fully Bayesian treatment with a uniform prior for the Poisson mean μ gave the same upper limit (if one did a sensible thing when the denominator neared zero), and in the cases we tested, the method over-covered for reasons that made sense to us.

More comprehensive coverage tests have been done internally in some collaborations and seem always to find that the method yields upper limits that over-cover (except for an incorrect study that found under-coverage). Blocker and the CDF statistics committee⁶⁹ find the performance of the algorithm in Ref. 16 to be essentially identical to a fully Bayesian technique for setting upper limits, and prefer the latter.

Barlow⁴³ has made available a calculator program with which one can explore results calculated in the spirit of Ref. 16 from n , Γ , t , and in addition the background estimate and its uncertainty.

Conrad et al.⁵⁴, and Tegenfeldt and Conrad⁷¹, studied the properties of integrating out nuisance parameters for background uncertainty as well as luminosity uncertainty in the context of the intervals of likelihood-ratio ordering construction of Ref. 25. The program for performing the calculation is also published⁶⁵ and since updated, including the treatment recommended by Hill⁵⁶ for a pathology in the case of fewer than expected background events. Their conclusions are consistent with the observation in other contexts that such a treatment of nuisance parameters leads to over-coverage for any particular value of nuisance parameters.

Lista⁶⁶ has integrated out a Gaussian uncertainty on the background in the context of the upper limits from non-standard conditioning described in Sec. 3.2.

Cranmer⁷³ has explored what happens if one integrates nuisance parameters out to 5σ significance (!). He finds severe undercoverage. At that level, knowing the form of the pdf for the nuisance parameter becomes a real issue.

7. Conclusions

From the extensive and continuing literature on this topic in both the high energy and statistical communities, it seems clear that more work is necessary before a consensus is attained for even a “convention” that everyone agrees on. As the HEP community seems to be increasingly fond of 5σ significance, this places rather extreme demands on any approximate methods. Regarding what can be tried today, I believe it is worth emphasizing the following.

- As the quotes from Edwards and Barnard above indicate, it may not always be fruitful to eliminate nuisance parameters. In cases where the inference depends strongly on the value of the nuisance parameter, the clearest presentation may be simply to enumerate cases.
- In a completely Bayesian analysis, “turning the crank” within the methodology may be straightforward, but specification of priors is fraught with pitfalls (especially in high dimensions), and interpretation of probability “P” can be a challenge if P is not consistently subjective degree of belief in all the inputs.
- It seems to me that the widespread availability of MINUIT, our long tradition of using it in HEP, and the reasonable frequentist performance of its output combine to make it mandatory that one use the method of MINOS (differences in log of the profile likelihood) on one’s likelihood function while trying out various options. The contours provided by MINUIT give insight into how sensible it is to eliminate the nuisance parameters.
- Already in 2000, Feldman³⁶ outlined the way we interpreted Ref. 30’s prescription to include nuisance parameters in likelihood-ordered Neyman construction, but with the paucity of examples outside of our NOMAD collaboration, this did not become widely known. Now that Cranmer^{48, 73}

and Punzi⁷⁷ have discussed the prescription and its more exact generalization (another way to interpret Ref. 30) in more detail, this situation is much improved. Feldman⁷⁵ and I believe that the Neyman construction using the approximation he presented in 2000 is a scalable, reasonable approach that deserves more study.

- No matter what method is used, the common practice of exploring the frequentist properties of the result should be strongly encouraged.

In addition, from the references and the talks at this conference, some next steps seem to be apparent for further development:

- The performance of reference priors¹⁴, as discussed by Demortier⁷⁴ at this conference, should be explored by those in HEP who advocate a Bayesian approach.
- Conditioning when appropriate should become a part of our conscious thinking, and the pros and cons of restricted and global ensembles should be better understood in our community.
- It would be interesting to explore the consequences of modern modifications to the profile likelihood beyond the examples shown by Reid and Fraser⁴⁹ at the previous PhyStat.

Finally, I end on a note of caution that has its roots in recent work in my current collaboration. If the underlying sources of the nuisance parameters are systematic uncertainties that become quite large, one becomes very sensitive to the details of the pdfs for the nuisance parameters, which can be much more poorly specified than the Poisson process that underlies our statistical uncertainties. In that case, one must be vigilant against blind use of a high-powered algorithm that in the end is not robust in this context, especially when one is applying it in extreme tails such as 5σ significance.

Acknowledgments

I owe a great debt to the many people with whom I have discussed these issues, beginning with the late Virgil Highland, and continuing with Gary Feldman, Fred James, Louis Lyons, Günter Zech, and many others. (Of course, we do not always agree, and the opinions in this paper are my own.) Special thanks go to Louis for organizing another stimulating conference, and to the statisticians who so generously

and graciously help us physicists to understand their work. I particularly thank Nancy Reid for her enlightening comments on my talk and manuscript during and after the conference.

References

1. S.S. Wilks, "The large-sample distribution of the likelihood ratio for testing composite hypotheses", *Annals of Math. Stat.* **9** (1938) 60.
2. F. James, "MINUIT. Function Minimization and Error Analysis," wwwasdoc.web.cern.ch/wwwasdoc/minuit/minmain.html
3. J.D. Kalbfleisch and J.D. Sprott, "Application of likelihood methods to models involving large numbers of parameters," *Jour. Roy. Stat. Soc. Series B* **32**, 175 (1970).
4. D. Basu "On the Elimination of Nuisance Parameters" *JASA* **72**, 355 (1977)
5. F. James and M. Roos, "Errors on Ratios of Small Numbers of Events", *Nuclear Physics* **B172** 475 (1980).
6. F. James, "Interpretation of the shape of the likelihood function around its minimum," *Comput. Phys. Commun.* **20** 29 (1980).
7. O. Barndorff-Nielsen "On a Formula for the Distribution of the Maximum Likelihood Estimates" *Biometrika* **70**, 343 (1983)
8. O. Barndorff-Nielsen "Inference on Full or Partial Parameters Based on the Standardized Signed Log Likelihood Ratio" *Biometrika* **73**, 307 (1986)
9. D.R. Cox, N. Reid "Parameter Orthogonality and Approximate Conditional Inference", *Jour. Roy. Stat. Soc. Series B* **49**, 1 (1987)
10. H. B. Prosper, "Small Signal Analysis In High-Energy Physics: A Bayesian Approach," *Phys. Rev. D* **37**, 1153 (1988); see also D. A. Williams, "Comment On 'Small Signal Analysis In High-Energy Physics: A Bayesian Approach'," *Phys. Rev. D* **38**, 3582 (1988), and reply.
11. D.A.S. Fraser, N. Reid, "Adjustments to Profile Likelihood" *Biometrika* **76**, 477 (1989)
12. G. Zech, "Upper Limits In Experiments With Background Or Measurement Errors," *Nucl. Instr. and Meth.* **A277** 608 (1989).
13. P. McCullagh, R. Tibshirani, "A Simple Method for the Adjustment of Profile Likelihoods" *Jour. Roy. Stat. Soc. Series B* **52**, 325 (1990)
14. See references in Ref. 17.
15. D.R. Cox, N. Reid, "A Note on the Difference Between Profile and Modified Profile Likelihood" *Biometrika* **79**, 408 (1992)
16. R.D. Cousins and V.L. Highland, "Incorporating systematic uncertainties into an upper limit," *Nucl. Instrum. Meth. A* **320**, 331 (1992).
17. B. Liseo "Elimination of Nuisance Parameters with

- Reference Priors” *Biometrika* **80**, 295 (1993); see also
18. J.O. Berger, L.D. Brown, and R.L. Wolpert, “A unified conditional frequentist and Bayesian test for fixed and sequential simple hypothesis-testing”, *Ann. Stat.* **22** 1787 (1994).
 19. V. Innocente and L. Lista, “Evaluation of the upper limit to rare processes in the presence of background, and comparison between the Bayesian and classical approaches,” *Nucl. Instrum. Meth. A* **340**, 396 (1994).
 20. N. Reid, “The Roles of Conditioning in Inference” *Stat. Sci.* **10**, 138 (1995). A more recent, very concise summary of likelihood-based inference is in Ref. 38.
 21. R. D. Cousins, “Why isn’t every physicist a Bayesian?,” *Am. J. Phys.* **63**, 398 (1995).
 22. A.L. Read, “Modified Frequentist Analysis of Search Results (The CL_s Method)”, Workshop on Confidence Limits, CERN (2000). doc.cern.ch/yellowrep/2000/2000-005/p81.pdf;
A.L. Read, “Optimal statistical analysis of search results based on the likelihood ratio and its application to the search for the MSM Higgs boson at $\sqrt{s}=161$ and 172 GeV”, DELPHI collaboration note, 97-158 PHYS 737 (1997).
 23. V. Highland, *Nucl. Instr. and Meth.* **A398** 429 (1997), followed by reply by G. Zech.
 24. P. C. Bhat, H. B. Prosper and S. S. Snyder, “Bayesian analysis of multi-source data,” *Phys. Lett. B* **407**, 73 (1997).
 25. G. J. Feldman and R. D. Cousins, “A Unified approach to the classical statistical analysis of small signals,” *Phys. Rev. D* **57**, 3873 (1998).
 26. R. D. Cousins, “Improved central confidence intervals for the ratio of Poisson means,” *Nucl. Instrum. Meth. A* **417**, 391 (1998).
 27. J.O. Berger, B. Liseo, R.L. Wolpert “Integrated Likelihood Methods for Eliminating Nuisance Parameters” *Stat. Sci.* **14**, 1 (1999)
 28. B. P. Roe and M. B. Woodroffe, “Improved probability method for estimating signal in the presence of background,” *Phys. Rev. D* **60**, 053009 (1999)
 29. T. Junk, “Confidence level computation for combining searches with small statistics”, *Nucl. Instr. and Meth.* **A434** (1999) 435.
 30. A. Stuart, K. Ord, and S. Arnold, *Kendall’s Advanced Theory of Statistics*, Volume 2A, 6th ed., (London:Arnold, 1999), and earlier editions by Kendall and Stuart.
 31. T.A. Severini, “On the Relationship Between Bayesian and Non-Bayesian Elimination of Nuisance Parameters” *Statistica Sinica* **9**, 713 (1999)
 32. J. Conway, “Inclusion of systematic uncertainties in upper limits and hypothesis tests,” Workshop on Confidence Limits, CERN (2000).
 33. R. Cousins, “Comments on methods for setting confidence limits,” Workshop on Confidence Limits, CERN (2000).
 34. G. Zech, “Confronting classical and Bayesian confidence limits to examples,” Workshop on Confidence Limits, CERN (2000). arXiv:hep-ex/0004011.
 35. www-zeus.roma1.infn.it/agostini/prob+stat.html
 36. G. Feldman, “Multiple measurements and parameters in the unified approach,” Workshop on Confidence Limits, Fermilab (2000), conferences.fnal.gov/cl2k/copies/feldman2.pdf.
 37. R.D. Cousins, “Additional comments on methods for setting confidence limits”, Workshop on Confidence Limits, Fermilab (2000), conferences.fnal.gov/cl2k/copies/bcousins2.ps.
 38. N. Reid “Likelihood” *J. Am. Stat. Assoc.* **95**, 1335 (2000)
 39. R. D. Cousins, “Comment on [Improved probability method for estimating signal in the presence of background],” *Phys. Rev. D* **62** (2000) 098301.
 40. W. A. Rolke and A. M. Lopez, “Confidence intervals and upper bounds for small signals in the presence of background noise,” *Nucl. Instrum. Meth. A* **458**, 745 (2001), arXiv:hep-ph/0005187.
 41. B. P. Roe and M. B. Woodroffe, “Setting confidence belts,” *Phys. Rev. D* **63**, 013009 (2001)
 42. L.D. Brown, T.T. Cai, A. DasGupta “Interval Estimation for a Binomial Proportion” *Stat. Sci.* **16**, 101 (2001)
 43. R. Barlow, “A calculator for confidence intervals,” *Comput. Phys. Commun.* **149**, 97 (2002), arXiv:hep-ex/0203002.
 44. D. Nicolo, G. Signorelli, “An application of the strong confidence to the Chooz experiment with frequentist inclusion of systematics,” Conference on Advanced Statistical Techniques in Particle Physics, Durham, England (2002), www.ippp.dur.ac.uk/Workshops/02/statistics/.
 45. L. Demortier, “Bayesian treatments of systematic uncertainties,” Conference on Advanced Statistical Techniques in Particle Physics, Durham, England (2002), www.ippp.dur.ac.uk/Workshops/02/statistics/.
 46. A.L. Read, “Presentation of Search Results — the CL_s Technique,” Conference on Advanced Statistical Techniques in Particle Physics, Durham, England (2002), www.ippp.dur.ac.uk/Workshops/02/statistics/.
 47. G. Zech, “Frequentist and Bayesian confidence limits,” *Eur. Phys. J. C* **4**, 12 (2002), arXiv:hep-ex/0106023.
 48. K. S. Cranmer, “Frequentist hypothesis testing with background uncertainty,” PHYSTAT2003, SLAC (2003). arXiv:physics/0310108.
 49. N. Reid, D.A.S. Fraser, “Likelihood inference in the presence of Nuisance parameters.” PHYSTAT2003, SLAC (2003). arXiv:physics/0312079.

50. L. Demortier, "Constructing ensembles of pseudo-experiments," PHYSTAT2003, SLAC (2003) arXiv:physics/0312100.
51. J. Linnemann, "Measures of significance in HEP and astrophysics," PHYSTAT2003, SLAC (2003) arXiv:physics/0312059.
52. P. Sinervo, "Definition and treatment of systematic uncertainties in high energy physics and astrophysics," PHYSTAT2003, SLAC (2003).
53. G. Punzi, "Including systematic uncertainties in confidence limits," CDF Statistics note (2003).
54. J. Conrad, O. Botner, A. Hallgren and C. Perez de los Heros, "Including systematic uncertainties in confidence interval construction for Poisson statistics," Phys. Rev. D **67**, 012002 (2003), arXiv:hep-ex/0202013. See also Ref. 55.
55. J. Conrad, O. Botner, A. Hallgren and C. P. de los Heros, "Coverage of confidence intervals for Poisson statistics in presence of systematic uncertainties," arXiv:hep-ex/0206034.
56. G. C. Hill, "Comment on 'Including systematic uncertainties in confidence interval construction for Poisson statistics'," Phys. Rev. D **67**, 118101 (2003), arXiv:physics/0302057.
57. J.O. Berger, "Could Fisher, Jeffreys and Neyman Have Agreed on Testing?" Stat. Sci. **18**, 1 (2003)
58. S.C. Dass and J.O. Berger, "Unified Conditional Frequentist and Bayesian Testing of Composite Hypotheses", Scand. J. Statist. **30** 199 (2003).
59. J. Heinrich, C. Blocker, J. Conway, L. Demortier, L. Lyons, G. Punzi, and P. K. Sinervo, "Interval estimation in the presence of nuisance parameters. 1. Bayesian approach," CDF/MEMO/STATISTICS/PUBLIC/7117 (2004), arXiv:physics/0409129.
60. L. Demortier, "A fully Bayesian computation of upper limits for Poisson processes", CDF/MEMO/STATISTICS/PUBLIC/5928 (2004)
61. D. A. S. Fraser, N. Reid and A. C. M. Wong, "Inference for bounded parameters," Phys. Rev. D **69**, 033002 (2004).
62. R. Barlow, "A note on $\Delta \ln L = -1/2$ Errors," arXiv:physics/0403046.
63. M.J. Bayarri, J.O. Berger "The Interplay of Bayesian and Frequentist Analysis" Stat. Sci. **19**, 58 (2004)
64. D.A.S. Fraser, "Ancillaries and Conditional Inference" Stat. Sci. **19**, 333 (2004); see also T.J. diCiccio and M.E. Thompson, "A Conversation with Donald A.S. Fraser," Stat. Sci. **19**, 370 (2004).
65. J. Conrad, "A program for confidence interval calculations for a Poisson process with background including systematic uncertainties: POLE 1.0," Comput. Phys. Commun. **158**, 117 (2004).
66. L. Lista, "Including Gaussian uncertainty on the background estimate for upper limit calculations using Poissonian sampling," Nucl. Instrum. Meth. A **517**, 360 (2004); see also Ref. 19.
67. B. Liseo, "The Elimination of Nuisance Parameters", geostasto.eco.uniroma1.it/utenti/liseo/pub.htm (2004).
68. W. A. Rolke, A. M. Lopez and J. Conrad, "Limits and Confidence Intervals in the Presence of Nuisance Parameters," arXiv:physics/0403059 (v4, 7 Jul 2005).
69. C. Blocker, "Interval Estimation in the Presence of Nuisance Parameters. 2. Cousins and Highland Method," CDF/MEMO/STATISTICS/PUBLIC/7539 (2005).
70. J. Conway, "Calculation of cross section upper limits combining channels incorporating correlated and uncorrelated systematic uncertainties", CDF/PUB/STATISTICS/PUBLIC/6428 (2005)
71. F. Tegenfeldt and J. Conrad, "On Bayesian treatment of systematic uncertainties in confidence interval calculations," Nucl. Instrum. Meth. A **539**, 407 (2005), arXiv:physics/0408039.
72. N. Reid, respondent to this paper, these proceedings.
73. K. Cranmer, "Statistical Challenges of the LHC", these proceedings.
74. L. Demortier, "Bayesian Reference Analysis", these proceedings.
75. G. Feldman, "Concluding Talk", these proceedings.
76. J. Heinrich, "The Bayesian approach to setting limits: what to avoid", these proceedings.
77. G. Punzi, "Ordering algorithms and Confidence Intervals in the presence of nuisance parameters", these proceedings.
78. D0 collaboration, www-d0.fnal.gov.
79. CDF collaboration, www-cdf.fnal.gov.